Bioinformatics, 31(12), 2015, 1938–1944
doi: 10.1093/bioinformatics/btv044
Advance Access Publication Date: 2 February 2015
Original Paper

OXFORD

Structural bioinformatics

mFASD: a structure-based algorithm for discriminating different types of metal-binding sites

Wei He, Zhi Liang*, Maikun Teng and Liwen Niu*

Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, China

*To whom correspondence should be addressed.
Associate Editor: Anna Tramontano

Received on October 15, 2014; revised on January 13, 2015; accepted on January 17, 2015

Abstract

Motivation: A large number of proteins contain metal ions that are essential for their stability and biological activity. Identifying and characterizing metal-binding sites through computational methods is necessary when experimental clues are lacking. Almost all published computational methods are designed to distinguish metal-binding sites from non-metal-binding sites. However, discrimination between different types of metal-binding sites is also needed to make more accurate predictions.

Results: In this work, we proposed a novel algorithm called mFASD, which could discriminate different types of metal-binding sites effectively based on 3D structure data and is useful for accurate metal-binding site prediction. mFASD captures the characteristics of a metal-binding site by investigating the local chemical environment of a set of functional atoms that are considered to be in contact with the bound metal. Then a distance measure defined on functional atom sets enables the comparison between different metal-binding sites. The algorithm could discriminate most types of metal-binding sites from each other with high sensitivity and accuracy. We showed that cascading our method with existing ones could achieve a substantial improvement of the accuracy for metal-binding site prediction.

Availability and implementation: Source code and data used are freely available from http://staff.ustc.edu.cn/~liangzhi/mfasd/

Contact: liangzhi@ustc.edu.cn or hwkobe@mail.ustc.edu.cn

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

1 Introduction

Proteins usually execute their functions by interacting with other components, including macromolecules such as DNA and proteins as well as small cofactors among which metals are most frequently used. Approximately, one-third of all proteins contain at least one metal ion according to the data in Protein Data Bank (PDB; Bernstein *et al.*, 1977; Holm *et al.*, 1996; Matthews *et al.*, 2008). Binding of metal ions is known to have profound effect on the overall protein conformation, where metals contribute significantly to proteins' stability and biochemical activity. On one hand, metals are often directly involved in chemical reactions by serving as redox centers or electrophilic reactants in catalysis and helping to bring reactive groups into correct orientations for reactions (Lin *et al.*, 2005). On the other hand, metals play a significant role in signal transduction and conformational changes of proteins by switching from state to state since they have different valence states (Williams *et al.*, 1985). Thus, identifying metal-binding sites could be a key step in understanding the function mechanisms of metal-binding proteins.

Experimental approaches to detect whether a protein binds a specific metal ion include atom absorption spectroscopy (Reed and

Poyner 2000), metal-affinity column chromatography (Herald *et al.*, 2003), mass spectrometry (Binet *et al.*, 2003) and so on. Although, these methods can provide definite evidence about the presence of a specific metal in a protein, they all need professional equipment and complicated steps thus are cost-expensive and time-consuming. Therefore, it is necessary to develop computational tools to predict metal-binding sites with high accuracy and sensitivity.

Various methods have been proposed in literature to predict metal-binding sites. Among them, some try to predict metal-binding sites based on sequence information and others depend on 3D structure information to achieve the goal. For example, a method using the sequence profiles of known metal-binding domains deposited in PDB has been reported and applied to prediction of copper binding proteins (Andreini *et al.*, 2009). Fold-X force field (Schymkowitz *et al.*, 2005) and CHED (Babor *et al.*, 2008) are two representatives for structure-based algorithms. The former predicts the spatial position of a metal ion in a protein by searching in a library that contains the spatial positions of metal ions relative to the corresponding ligating atoms. The latter predicts metal-binding sites based on statistically derived geometric constrains of different metal-binding sites. There are also many other methods. To name a few, TEMSP is specialized in zinc-binding sites prediction by recognizing residue triads specific for zinc sites (Zhao *et al.*, 2011). SitePredict is able to identify metal-binding sites as well as small ligand pockets by using a random forest classifier (Babor *et al.*, 2008). The fragment transformation method (Lu *et al.*, 2012) can predict metal-binding sites by searching a template library of metal-binding residues. Recently, a web server called CheckMyMetal (CMM; Zheng *et al.*, 2014) has been reported to support sophisticated evaluation of metal-binding sites in macromolecules structures by using a set of parameters derived from 7350 metal-binding sites of 2304 high resolution structures.

However, most of the methods mentioned above are designed to discriminate between metal-binding sites and non-metal-binding sites instead of different types of metal-binding sites. For example, CHED can be used to predict metal-binding sites for a wide range of metal ions, but the predicted binding sites can bind any other metal ions in theory. Methods predicting one or two specific types of metal ions, like TEMSP (for Zn) and FEATURE (for Zn and Ca; Ebert and Altman, 2008), only tell the sites binding these metal ions from those not binding these metals. The CMM web server can provide alternative metal suggestions for users when suspecting incorrect metal assignment, but it does not show specific functionality to discriminate different metal-binding sites. In other words, no effective methods are available to discriminate metal-binding sites that bind different types of metal ions. SitePredict has tried to distinguish different metal-binding sites, however, its performance seems not satisfied and needs to be improved (Babor *et al.*, 2008). There may be two main reasons why it is hard to discriminate different types of metal-binding sites. First, the cavity holding a metal ion is often very small so that only limited information is included. It is worth mentioning that the types of residues involved in metal binding are rather restricted, often concentrating upon CYS, HIS, ASP and GLU (Golovin *et al.*, 2005; Auld *et al.*, 2001). This fact makes it difficult to discriminate metal-binding sites of different types of metals especially on residue level. Second, the surroundings of metal ions are similar to each other due to the similarity of properties of metal ions. Some experiments have proved that the binding sites for one specific type of metal can also bind other type of metals (Bock *et al.*, 1999). It is known that one of the most prevalent Ca^{2+} binding motifs, the EF-hand motif, can also bind Mg^{2+} in some cases (Lewit-Bentley and Rety, 2000), thus making it hard to tell one from another.

In this work, we proposed a novel algorithm based on 3D structure information called mFASD, which can discriminate different types of metal-binding sites at atom level and is effective in distinguishing proteins that bind different types of metal ions. In mFASD, a metal-binding site was modeled as an atom set, called functional atom set (FAS), and the associated local chemical environment, which is similar to the minimal functional site proposed by Andreini *et al.* (2011, 2013). Each atom in a FAS was considered to be in contact with the metal ion situated in the metal-binding site. Considering that the binding environments for different types of metals are different at atom level, it is possible to discriminate FASs that bind different types of metals based on proper distance measure. The algorithm achieves this goal by introducing a distance metric called functional atom set distance (FASD) that is to be explained in detail in the next section. mFASD attained good performance in distinguishing most types of metal-binding sites from each other. We exemplified potential applications of mFASD by (i) cascading it with CHED and SitePredict, respectively, which leaded to a substantial improvement of metal-binding site prediction; (ii) applying it to predict the metals contained in two protein structures solved by our labs thus providing useful information for structural biologists.

2 Methods

2.1 Description of the algorithm

The workflow of the algorithm is shown in Figure 1. Briefly, we presented a metal-binding site as a set of functional atoms that were considered to contact the metal ion. For each atom in the FAS, we described its local chemical environment by integrating information of its chemical property and the chemical properties of atoms in its neighborhood, based on which a distance measure between two functional atoms was defined. We transformed the problem of comparing two metal-binding sites as comparing two corresponding FASs. We calculated the FASD between two FASs by all-versus-all comparison between the functional atoms in the two FASs. Two FASs were classified as the same metal-binding type, if their distance was smaller than a predefined threshold. For each type of metal-binding site, we constructed a reference FAS set. To determine whether a metal-binding site could bind a certain type of metal, individual comparison of the query FAS with the reference FASs were combined to generate the final decision by taking a simple majority vote.

Step 1: Extraction of functional atoms from a metal-binding protein

We defined atoms situated within a distance of r from the metal ion as functional atoms (also known as donor atoms). In this work, a value of 3.5 Å was adopted for r by referring to Lu's work (Lu *et al.*, 2012). Therefore, a metal-binding site of a metalloprotein could be extracted and presented as a functional atom set or a FAS. It was worth mentioning that only those sites containing single metal ions were considered here. For example, among the several different types of Fe-containing sites, namely single iron ions, ion-sulfur clusters and heme groups, only the first type was covered in the present study.

Step 2: Extraction of local chemical environment of a functional atom

We captured the characteristics of local chemical environment of a functional atom F from two aspects, the chemical property of F itself and the chemical properties of atoms in the neighborhood of F. We used the atom type of F as the descriptor of its chemical property

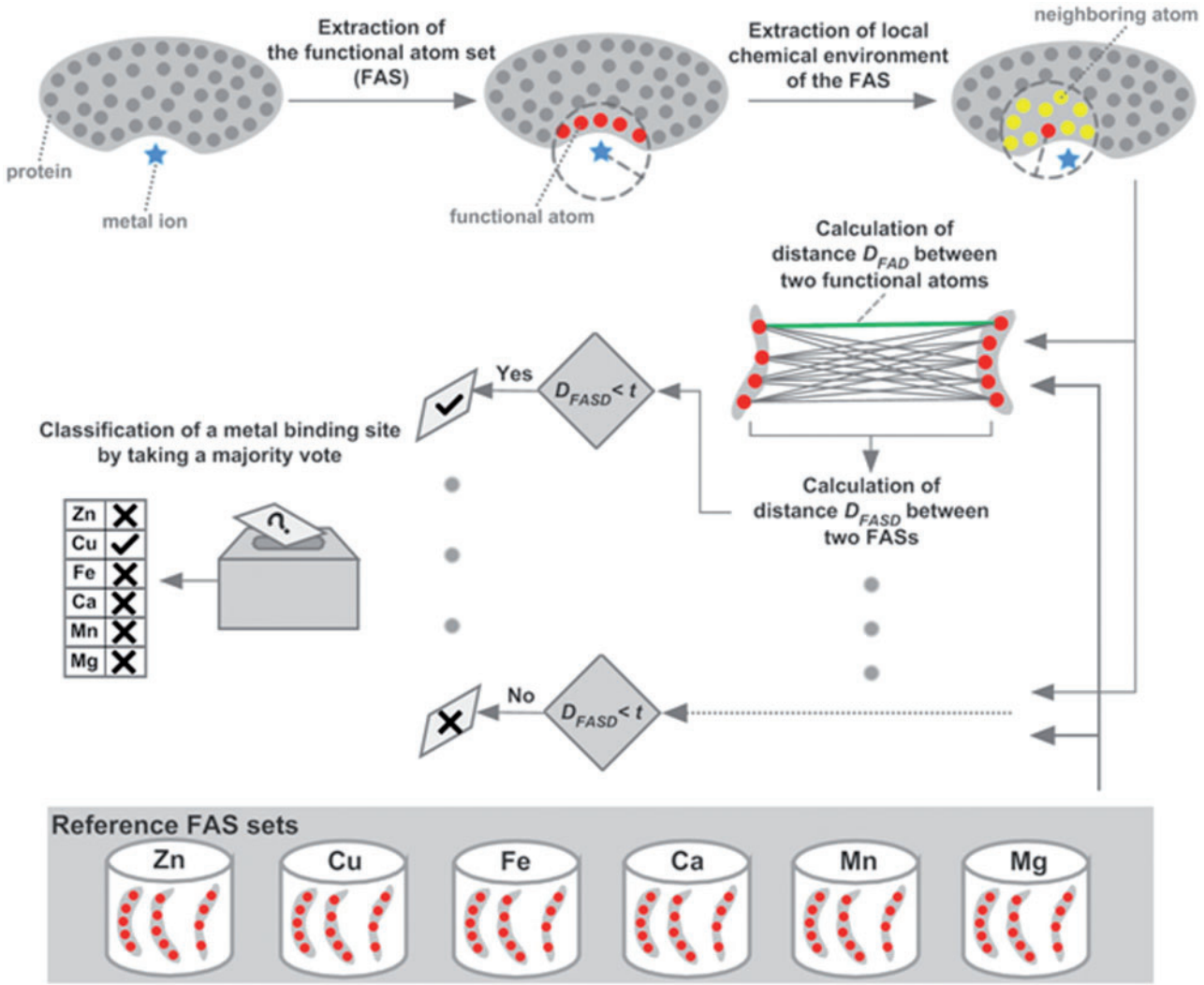

Fig. 1. Workflow of the mFASD algorithm

directly. To describe the neighborhood of *F*, we adopted a coarse-grained description. We classified atoms into six categories according to their chemical properties: (i) hydrophibic, (ii) acceptor, (iii) donor, (iv) hydrophobic, (v) aromatic and (vi) neutral (Sobolev *et al.*, 1996). In this way, the chemical property of a neighboring atom of *F* could be expressed as its chemical category. We defined that if an atom *A* is within a distance of 5.0 Å to *F*, it is considered to interact with *F*. The *F*–*A* interactions reflect the chemical properties of atoms in the neighborhood of *F*. We presented each *F*–*A* interaction as a tuple (*i*, *j*) of the corresponding chemical categories, where *i* and *j* denote the possible chemical category values of *F* and the interacting atom *A*, respectively. Since there are six chemical categories, the distribution of the number of *F*–*A* interactions could be described using a 36-dimensional vector

$$(C_{1,1}, C_{1,2}, C_{1,3}, C_{1,4}, C_{1,5}, C_{1,6}, C_{2,1}, \ldots, C_{6,6}),$$

where

$$C_{i,j} = \begin{cases} 0, & if\ chemical\ category\ of\ F \neq i \\ q, & if\ chemical\ category\ of\ F = i \end{cases},$$

q is the number of atoms of chemical category *j* that interact with *F*.

Step3: Calculation of distance between two functional atoms
According to the description of the local chemical environment of a functional atom described above, we defined the distance between two functional atoms F_1 and F_2 as

$$D_{FAD}(F_1, F_2) = w \times d_F + (1 - w) \times d_N$$

where

$$d_F = \begin{cases} 0, & if\ atom\ type\ of\ F_1 = atom\ type\ of\ F_2 \\ 1, & if\ atom\ type\ of\ F_1 \neq atom\ type\ of\ F_2 \end{cases}$$

$$d_N = 1 - \sum_{i,j=1}^{6} min\left(C_{i,j}^{F_1}, C_{i,j}^{F_2}\right) / \sum_{i,j=1}^{6} max\left(C_{i,j}^{F_1}, C_{i,j}^{F_2}\right)$$

The functional atom distance (FAD), D_{FAD}, consists of two parts, d_F and d_N. d_F accounts for the atom type difference between the two functional atoms. If the two atom types are the same, d_F is zero, else d_F equals 1. d_N quantifies the difference of *F*–*A* interaction distributions between the two functional atoms. d_N ranges from zero to one. The minimal value is reached, when the two functional atoms have the same distributions of *F*–*A* interactions, suggesting the high similarity of their neighborhood. In contrast, the maximal value of d_N indicates that the two functional atoms have completely

different distributions of *F–A* interactions and therefore distinct neighborhood. Parameter w was introduced to control the ratio of contribution of the functional atom itself to that of its neighborhood and was determined through training.

Step 4: Calculation of distance between two FASs

To compare two FASs S_1 and S_2, we constructed a complete bipartite graph. Each vertex of the graph is a functional atom. Each edge represents comparison between two functional atoms from S_1 and S_2 respectively, and is weighted using the corresponding D_{FAD} value that quantifies the distance between the two functional atoms. We defined the distance between S_1 and S_2, called functional atom set distance or FASD, as follows

$$D_{FASD}(S_1, S_2) = \sum_{F_i \in S_1} min_{F_j \in S_2}(D_{FAD}(F_i, F))/|S_1|,$$

where the numerator sums up the minimal D_{FAD} from each functional atom in S_1 to those in S_2, and the denominator is the number of functional atoms in S_1. Therefore, D_{FASD} reflects the average minimal D_{FAD} from S_1 to S_2.

Step 5: Classification of a metal-binding site

Based on FASD, the similarity between two metal-binding sites could be evaluated by calculating the D_{FASD} between their corresponding FASs. Two metal-binding sites are regarded to be of the same metal-binding type, if their D_{FAD} is lower than a threshold t. The proper value of t was determined through training.

To determine whether a query protein could bind a specific type of metal M, we constructed a reference FAS set R^M for the corresponding metal and compared the FAS Q of the query protein against the reference set R^M. Individual comparison between the query FAS Q and the reference FASs in R^M were then combined to generate the final decision by taking a simple majority vote. Concretely, if more than half of the FASs in the reference set R^M were judged to be of the same metal-binding type as the query FAS Q, the query protein was determined to be a member of M-binding proteins, otherwise its membership was rejected.

To construct the reference FAS sets for different types of metals, we downloaded non-redundant protein structures from PDBSelect25 (Griep *et al.*, 2010), the sequence identity of which was less than 25%. We kept structures containing at least one of the following six types of metals, i.e. Cu, Fe, Mg, Mn, Zn and Ca. FASs were then extracted from these structures. Six reference FAS sets, one for each metal type, were constructed by separating FASs binding different types of metals into different sets, namely R^{Zn}, R^{Fe}, R^{Cu}, R^{Mn}, R^{Ca} and R^{Mg}.

As described above, FASs with explicitly bound metal ions could be extracted from protein structures directly. However, a query protein might have no clear clue of metal binding, for example, a newly solved crystal structure without any bound metal ions. In this case, there are two means to extract candidate FASs from the query protein. First, one can construct a FAS by searching cavities that are suitable to hold a metal ion. Second, there are many existing methods for metal-binding site prediction, some of which provide details of residues that might bind a metal ion. The side-chain atoms of these predicted residues could be assigned to the FAS of the query protein.

2.2 Performance assessment

To evaluate the capability of our method to discriminate the binding sites of P from those of N, where P denotes a certain metal type and N denotes any metal type that is different from P or a set of metal types excluding P, we prepared a FAS set $R = \{R^P, R^N\}$. For each FAS R_i in R, we trained the classifier of metal type P using the rest FASs in R and determined whether R_i can bind P. Suppose R_i was drawn from R^P. If R_i was determined to bind P, then R_i was counted as a true positive (TP), otherwise a false negative. Suppose R_i was drawn from R^N. R_i was counted as a false positive (FP), if R_i was determined to bind P, otherwise a true negative. True positive rate (TPR) or sensitivity and false positive rate (FPR) or selectivity were then calculated according to the following equations:

$$\mathrm{TPR} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN}}$$

$$\mathrm{FPR} = \frac{\#\mathrm{FP}}{\#\mathrm{FP} + \#\mathrm{TN}}$$

Based on the computed (FPR, TPR) pairs under different values of t, an ROC curve could be plotted. An AUC score was obtained by integrating the area under the corresponding ROC curve. Generally, a better algorithm has a higher AUC score. The highest AUC score of 1.0 is reached, when a method could correctly tell all the binding sites of P from those of N. For a random classifier, its AUC score equals 0.5, indicating no capability of discriminating different types of metal-binding sites.

3 Results

3.1 Determination of parameters

In the mFASD algorithm, two parameters are adjustable. The first one is the w in D_{FAD}, which determines the ratio of contribution of the functional atom itself to that of its neighborhood. The second is the threshold t for determining whether two FASs are of the same metal-binding type. To obtain appropriate values of these two parameters for the classifier of a certain metal type M, we scanned the support of the parameters and evaluated the performance of the classifier for discriminating M-binding sites from other non-M-binding sites under different parameter values (Section 2).

Here, we exemplified the procedure of parameter determination for zinc-binding site classifier. We first prepared the FAS set $R = \{R^P, R^N\}$, where P = Zn and N = {Fe, Cu, Mn, Ca, Mg}. Then, we scanned the parameter w from 0.1 to 0.9 with a step of 0.1. For every value of w, the AUC score was computed. It was found that the AUC score reached three peaks 0.890, 0.878 and 0.881, when w was equal to 0.2, 0.5 and 0.8, respectively (Fig. 2a). The high AUC scores indicated that our method performed well in discriminating zinc-binding sites from other metal-binding sites. We further investigated the corresponding ROC curves (Fig. 2b). Apparently, when w was equal to 0.8, the highest sensitivity (TPR) could be achieved with the selectivity (FPR) kept at a certain low level (usually 0.05). Taking this factor into account, 0.8 was finally taken as the value for parameter w to classify zinc-binding and non-zinc-binding sites. The value 0.3 of t corresponding to the point (0.7, 0.05) of (TPR, FPR) was then taken as a suitable choice of t. For the other five types of metals, the values for w and t were determined in the same way.

3.2 Discrimination of different metal-binding sites

After the parameters were determined, we also evaluated the performance of mFASD to discriminate binding sites of two different metal types (Section 2). As shown in Table 1, binding sites of most metal types could be discriminated from each other with an AUC score higher than 0.7. For instance, for proteins binding Zn, our method could distinguish them from proteins binding other five

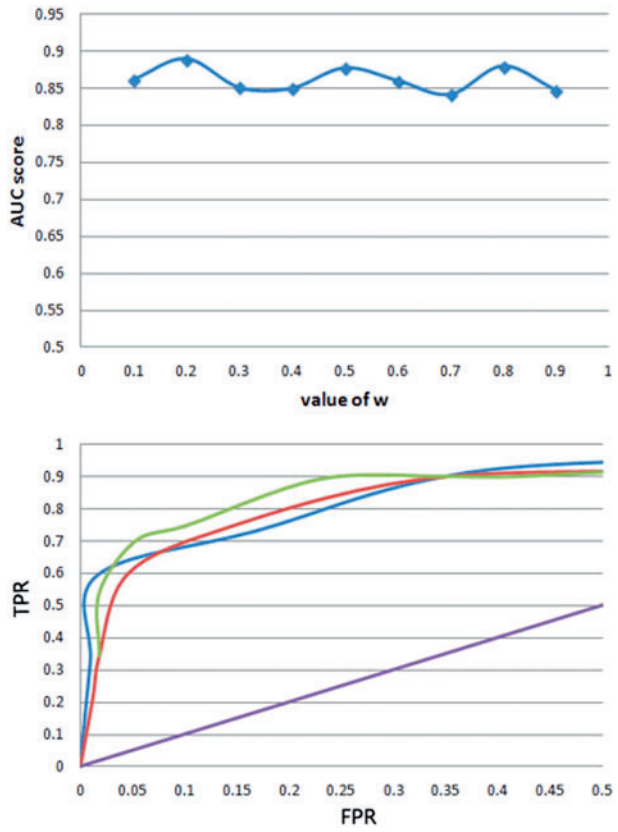

Fig. 2. Determination of parameters *w* and *t* for zinc-binding site classifier. (a) The AUC score under different value of parameter *w*. Three AUC score peaks were achieved when *w* was equal to 0.2, 0.5 and 0.8, respectively. (b) The ROC curves corresponding to the three AUC score peaks (red, $w=0.2$; blue, $w=0.5$; green, $w=0.8$). The purple line is the reference line for a random classifier with AUC score of 0.5. The algorithm performed best in terms of high TPR and low FPR when parameter *w* was 0.8. Under this situation, 0.3 was taken as a suitable choice of parameter *t*, which corresponded to the point with a high TPR of 0.7 and a low FPR of 0.05

Table 1. Performance of mFASD for discriminating different types of metal-binding sites

P	*N*	AUC
Zn	Fe	0.831
Zn	Cu	0.722
Zn	Mn	0.886
Zn	Ca	0.937
Zn	Mg	0.906
Fe	Zn	0.753
Fe	Cu	0.822
Fe	Mn	0.549
Fe	Ca	0.716
Fe	Mg	0.827
Cu	Zn	0.692
Cu	Fe	0.841
Cu	Mn	0.911
Cu	Ca	0.868
Cu	Mg	0.973
Mn	Zn	0.826
Mn	Fe	0.506
Mn	Cu	0.764
Mn	Ca	0.674
Mn	Mg	0.793
Ca	Zn	0.904
Ca	Fe	0.908
Ca	Cu	0.917
Ca	Mn	0.859
Ca	Mg	0.783
Mg	Zn	0.720
Mg	Fe	0.841
Mg	Cu	0.734
Mg	Mn	0.566
Mg	Ca	0.448

The non-symmetric discrimination performance resulted from the difference of positive set and negative set. For example, for Zn–Fe AUC, Zn is the positive set and Fe is the negative, while for Fe-Zn AUC, Fe is the positive set and Zn is the negative set. The phenomenon was consistent with the report of SitePredict (Babor et al., 2008)

metal types Fe, Cu, Mn, Ca and Mg very well with AUC scores of 0.831, 0.722, 0.886, 0.937 and 0.906, respectively. Another frequently used metal type, Ca, could also be identified from other five metal types Fe, Cu, Mn, Zn and Mg with high AUC scores of 0.904, 0.908, 0.917, 0.859 and 0.783, respectively.

However, our method failed to achieve satisfied discrimination in some cases. For example, the AUC score of discriminating Ca binding sites from Mg binding sites was close to 0.5. Discrimination between Fe and Mn binding sites and discrimination between Zn and Cu binding sites were also difficult as indicated by the AUC scores lower than 0.7. As discussed in the Section 1, some proteins that bind one type of metal could also bind other types of metals. Although, our method performed well for discriminating most binding sites of different metal types from each other, it could not handle proteins with versatile binding sites that could bind more than one type of metal.

3.3 Applications (I)

In this section, we discussed how our method could be applied to improving the accuracy of metal-binding site prediction of existing methods. Most algorithms in literature showed no or poor capability to discriminate different types of metal-binding sites, while our method could reduce FPs of these algorithms by identifying proteins that were wrongly predicted to bind a certain metal type. In addition, for approaches just predicting the presence of a metal-binding site in a protein, our method could give further information about whether the predicted site could bind a specific metal. Here, we took SitePredict and CHED, two state-of-the-art methods in literature, as examples to illustrate the cascading of our method with existing algorithms.

SitePredict was proposed to predict metal-binding sites and other small ligand binding sites by using a random forest classifier (Babor *et al.*, 2008). We randomly chose 100 structures containing different types of metal ions from PDB that had been clustered using CD-Hit program (Li *et al.*, 2006) at 30% sequence identity (Supplementary Table S1). Zinc-binding sites were predicted for these proteins using SitePredict. 30 structures were predicted to bind Zn, 14 of which were wrong (Supplementary Table S2). We delivered all the predictions to mFASD. For the wrongly predicted Zn binding sites, 1, 1, 2, 4 and 6 were in fact found to bind Fe, Mg, Mn, Cu and Ca, respectively. It was found that eleven FPs could be excluded by mFASD except three (1G8K_A, 1PU4_A and 1D6U_A) (Supplementary Table S2). We carefully inspected the local configurations of the predicted metal-binding sites of these three proteins. Both 1PU4_A and 1D6U_A contain a Cu in the predicted sites, which consist of three HISs to coordinate with the metal (Fig. 3a, b) just like many

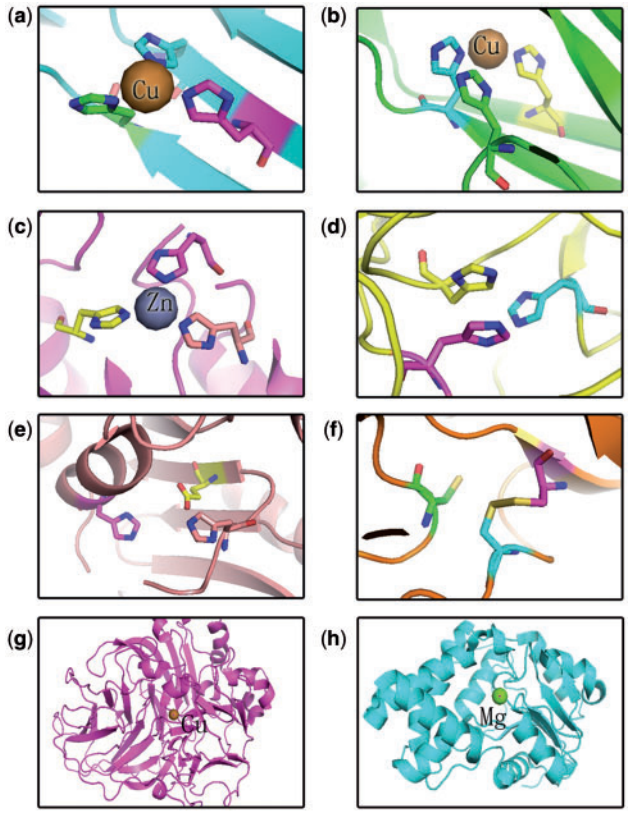

Fig. 3. Metal-binding sites predicted by SitePredict and CHED and protein structures taken for example rendered by PyMOL. (a) and (b) The zinc-binding sites predicted by SitePredict in 1PU4_A and 1D6U_A, respectively. Both sites consist of three HISs and were predicted to bind Zn by SitePredict but were found to bind Cu in fact. (c) The zinc-binding site predicted by SitePredict in 1M68_A is very similar to those in (a) and (b). (d) Although nothing is present in the zinc-binding site predicted by SitePredict in 1G8K_A, this site is of high probability to bind Zn due to the similarity of its local conformation to other zinc-binding sites. (e) and (f) The zinc-binding sites predicted by CHED in 1F6B_A and 1LFG_A, respectively. (g) The structure and its binding metal of 3KW7. (h) The structure and its binding metal of 3OPX

zinc-binding sites for instance 1M68_A (Fig. 3c). So these two sites also had the potential to bind a zinc ion. In contrast, 1G8K_A does not bind any metal ions in the predicted site of the PDB structure. However, its local configuration is very similar to the above two binding sites (Fig. 3d). Therefore, there is a high probability that the predicted site could bind Zn as well. For the rest 16 TP s predicted by SitePredict, 15 were determined as zinc binding correctly by mFASD (Supplementary Table S2). Apparently, our method dramatically reduced the FPs of SitePredict for predicting zinc-binding sites and made metal-binding site prediction more accurate.

CHED could be applied to predicting intra-chain metal-binding sites for transition metals. Here, we showed how our method could improve the prediction of CHED by taking zinc-binding sites as an example. We randomly selected 90 structures containing different metals from PDB that had been clustered using CD-Hit program at 30% sequence identity (Supplementary Table S3). Metal-binding sites in these proteins were predicted using CHED, which gave 115 metal-binding sites. Among these predicted metal-binding sites, 51 were confirmed to bind Zn and 64 were found to bind other metals or nothing in the structures (Supplementary Table S4). We processed these predictions with mFASD. Fifty out of the 51 Zn-binding sites were correctly identified indicating a high sensitivity of our method (Supplementary Table S4). For the rest 64 predictions that did not bind Zn in their structures, 50 were correctly excluded by mFASD. We inspected the 14 predictions not filtered out by our method (Supplementary Table S4). Only four binding sites were observed to bind other metals in their structures. Although, nothing was bound in their structures, the remaining 10 binding sites might also be able to bind Zn, since their local configurations are very similar to those binding Zn. For example, the predicted metal-binding site in 1F6B_A contains two HIS and one ASP just as many Zn-binding sites did (Fig. 3e). Also, the predicted binding site in 1LFG_A consists of three CYSs that often appear in zinc motif (Fig. 3f). Our method provided further information on the predicted metal-binding sites generated by CHED.

3.4 Applications (II)

3KW7 (Fig. 3g) and 3OPX7 (Fig. 3h) were two protein structures solved by other members of our labs, which contains Cu and Mg respectively. Here, we took them as examples to show how mFASD could be applied to predict the metals situated in the potential metal-binding sites of newly determined structures thus providing useful information for structural biologists. As described in the Section 2, we constructed the FAS by manually extracting atoms within 3.5 Å to the candidate position of the metal ion. We then applied mFASD to predict the metals binding to the FAS. For 3KW7, only Cu was predicted to be present at the metal-binding site, which was consistent with the experimental result. For 3OPX, both Mg and Ca were predicted to be able to bind the FAS, although only Mg was found in the structures. However, as we mentioned above, Mg and Ca are hard to discriminate in some cases and proteins binding Ca could also bind Mg (Lewit-Bentley and Rety, 2000). The accurate predictions given by mFASD could help a structural biologist to infer the functions and mechanisms of the studied proteins.

4 Conclusions

In this work, a novel method was proposed to discriminate different types of metal-binding sites at atom level. The method compared two metal-binding sites using a metric called FASD, which evaluated the similarity of atoms in contact with the metal ions in terms of their local chemical environment. Despite some exceptions, our method could discriminate most types of metal-binding sites from each other very well. Since higher structure resolution brings more accurate structural information which is important for metal-binding site prediction, structures with higher resolution are expected to have better performance. Also, mFASD does not limit to experimental structures and should be applicable to homology models with reasonable accuracy especially at the potential metal-binding sites. As described in the method, the local environment of a metal-binding site is encoded as the distribution of chemical categories in the neighborhood of the FAS. Despite of some loss of information, this coarse-grained description endows mFASD with the robustness to moderate structure flexibility. From the methodological point of view, the strategy adopted by mFASD could also be extended to comparison of small ligand binding pockets, protein-protein interaction interfaces and protein-DNA/RNA binding regions.

Acknowledgements

The authors are grateful to Dr Zhongliang Zhu and Dr Yongxiang Gao for helpful discussion and valuable advices. the authors also thank other members of the lab for assistance during the study.

Funding

This work is supported by grants to L.N from the Ministry of Science and Technology (2011CBA00804, 2011CB917200), the National Natural Science Foundation of China (31170726); by grant to Z.L. from the Ministry of Science and Technology (2012CB917200, 2014CB910600).

Conflict of Interest: none declared.

References

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

Andreini,C. *et al.* (2011) Minimal functional sites allow a classification of zinc sites in proteins. *PLoS One*, **6**, e26325.

Andreini,C. *et al.* (2013) MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.*, **41**, D312–D319.

Auld,D.S. (2001) Zinc coordination sphere in biochemical zinc sites. *Biometals*, **14**, 271–313.

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.

Binet,M.R. *et al.* (2003) Detection and characterization of zinc- and cadmium-binding proteins in *Escherichia coli* by gel electrophoresis and laser ablation-inductively coupled plasma-mass spectrometry. *Anal. Biochem.*, **318**, 30–38.

Bock,C.W. *et al.* (1999) Manganese as a replacement for magnesium and zinc: functional comparison of the divalent ions. *J. Am. Chem. Soc.*, **121**, 7360–7372.

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

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

Griep,S. *et al.* (2010) PDBSelect1992–2009 and PDBfilter-select. *Nucleic Acids Res.*, **2009**, gkp786.

Golovin,A. *et al.* (2005) MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins*, **58**, 190–199.

Herald,V.L. *et al.* (2003) Proteomic identification of divalent metal cation binding proteins in plant mitochondria. *FEBS Lett.*, **537**, 96–100.

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

Lewit-Bentley,A. and Rety,S. (2000) EF-hand calcium-binding proteins. *Curr. Opin. Struct. Biol.*, **10**, 637–643.

Li,W. *et al.* (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

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

Lu,C.H. *et al.* (2012) Prediction of Metal Ion-Binding Sites in Proteins Using the Fragment Transformation Method. *PLoS One*, **7**, e39252.

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

Reed,G.H. and Poyner,R.R. (2000) Mn2+ as a probe of divalent metal ion binding and function in enzymes and other proteins. *Met. Ions Biol. Syst.*, **37**, 183–207.

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.

Sobolev,V. *et al.* (1996) Molecular docking using surface complementarity. *Proteins Struct. Funct. Bioinf*, **25**, 120–129.

Williams,R.J.P. (1985) The symbiosis of metal and protein functions. *Eur. J. Biochem.*, **150**, 213–248.

Zhao,W. *et al.* (2011) Structure-based *de novo* prediction of zinc-binding sites in proteins of unknown function. *Bioinformatics*, **27**, 1262–1268.

Zheng,H. *et al.* (2014) Validation of metal binding sites in macromolecular structures with the CheckMyMetal web server. *Nat. Protoc.*, **9**, 156–170.