Molecular BioSystems

RSC Publishing

METHOD

ZincExplorer: an accurate hybrid method to improve the prediction of zinc-binding sites from protein sequences†

Cite this: *Mol. BioSyst.*, 2013, **9**, 2213

Zhen Chen,[a] Yanying Wang,[a] Ya-Feng Zhai,[a] Jiangning Song*[bc] and Ziding Zhang*[a]

Received 14th March 2013,
Accepted 26th June 2013

DOI: 10.1039/c3mb70100j

www.rsc.org/molecularbiosystems

As one of the most important trace elements within an organism, zinc has been shown to be involved in numerous biological processes and closely implicated in various diseases. The zinc ion is important for proteins to perform their functional roles. To provide in-depth functional annotation of zinc-binding proteins, an initial but crucial step is the accurate recognition of zinc-binding sites. Motivated by the biological importance of zinc, we propose a new method called ZincExplorer to predict zinc-binding sites from protein sequences. ZincExplorer is a hybrid method that can accurately predict zinc-binding sites from protein sequences. It integrates the outputs of three different types of predictors, namely, SVM-, cluster- and template-based predictors. Four types of zinc-binding amino acids CHEDs (*i.e.* CYS, HIS, ASP and GLU) could be predicted using ZincExplorer. It achieved a high AURPC (Area Under Recall–Precision Curve) of 0.851, and a precision of 85.6% (specificity = 98.4%, MCC = 0.747) at the 70.0% recall for the CHEDs on the 5-fold cross-validation test. When tested on an independent dataset containing 2023 zinc-binding CHEDs and 14 493 non-zinc-binding CHEDs, it achieved about 3–8% higher AURPC in comparison to two other sequence-based predictors. Moreover, ZincExplorer could also identify the interdependent relationships (IRs) of the predicted zinc-binding sites bound to the same zinc ion, which makes it a useful tool for providing in-depth zinc-binding site annotation.

1 Introduction

Zinc is one of the most important and ubiquitous trace elements in microorganisms, plants, and animals.[1] It is reported that zinc is the second most abundant transition metal ion in living organisms, second only to iron.[2,3] Like other types of metal ions, zinc is considerably involved in enzyme catalysis. For instance, zinc is the only metal ion that serves as a cofactor to more than 300 enzymes.[1,4] Many cell processes are regulated by zinc, involved in catalysis and co-catalysis by the enzymes, such as DNA synthesis, normal growth, and brain development.[5,6] Zinc also plays structural roles in a variety of proteins. For example, zinc finger proteins are the largest class of transcription factors in the human genome and their structures are stabilized in the presence of zinc ions.[7,8]

Generally, zinc-binding sites contain several types of amino acids, among which CYS, HIS, GLU and ASP (CHED for short) are the four most abundant, accounting for 96% of all zinc-binding sites. Due to a wide range of functional and structural roles of the zinc ions, identification of the zinc-binding sites is an important step towards our better understanding of the functions of zinc-binding proteins. Accordingly, a number of computational methods have been developed to predict the zinc-binding sites. Passerini *et al.* used a two-stage machine-learning approach to predict all the CYS and HIS of a protein in either of the three states (free, metal bound, or in disulfide bridges) by using the sequence information extracted from the position-specific evolutionary profiles as well as protein length and amino acid compositions.[9] Subsequently, they developed another machine learning method ZincFinder to predict zinc-binding residues and the bonding state of pairs of predicted residues close in sequence.[10] Shu *et al.* integrated an SVM-based predictor and a homology-based predictor into a computational tool called ZincPred to predict four types of zinc-binding CHEDs.

[a] *State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China. E-mail: zidingzhang@cau.edu.cn*
[b] *National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China. E-mail: Jiangning.Song@monash.edu*
[c] *Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC 3800, Australia*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3mb70100j

The input features of ZincPred include the Position Specific Substitution Matrices (PSSM) obtained from PSI-BLAST searching[11] and the conservation score calculated from the PSSM.[12] The two aforementioned methods (ZincFinder and ZincPred) have been considered as representative sequence-based predictors.

The number of high-resolution protein structures deposited in the PDB database[13] is increasing rapidly as a consequence of high-throughput structural genomics efforts, which makes the prediction of zinc-binding sites from protein structure possible. A number of reasonably successful methods[14–18] have been developed by taking advantage of structural information. However, although structure-based methods are generally more accurate than sequence-based counterparts, they have certain limitations. For example, they are only able to predict the zinc-binding sites of the proteins whose structures have been determined but could not be applied to annotate the complete zinc-binding proteomes. To address these issues, in this study, we develop a new hybrid approach called ZincExplorer, which is composed of SVM-, cluster- and template-based predictors to predict the zinc-binding sites in proteins from their amino acid sequences. Similar to most previous studies, our method focuses on prediction of zinc-binding residues CHED, because they account for the majority of all the zinc-binding residues.

2 Methods

In this study, we develop a hybrid approach by training an SVM-based predictor, a cluster-based predictor and a template-based predictor to learn scoring and prediction rules for identifying zinc-binding residues in proteins. All CHEDs of proteins in the datasets were scanned to predict whether they are zinc-binding sites or not. Here we define the zinc-binding CHEDs as positive samples and the non-zinc-binding CHEDs as negative samples. To construct ZincExplorer, the results of the SVM-based predictor and the cluster-based predictor were combined through a simple linear function to produce intermediate prediction results, and the candidate sites were prioritized. Then, the Interdependent Relationship (IR) of the candidate sites was assigned by the template-based predictor. The final prediction results were generated by integrating the intermediate prediction result and the results obtained using the template-based predictor. The complete workflow of our ZincExplorer methodology is illustrated in Fig. 1.

2.1 Dataset

The current dataset used in this study was previously collected by Passerini *et al.* (2006), which we termed the Passerini_dataset. In particular, all the homologous PDB chains in this dataset were filtered to ensure that no pair of chains shared a positive HSSP value.[19] If there was any nitrogen, oxygen or sulfur atom of the residues in a protein chain located within 3 Å to a Zn atom, then the Zn atom would be regarded as binding to the protein chain and the corresponding residues were defined as zinc-binding sites. In our work, we only used the PDB chains containing Zn atoms in the Passerini_dataset and as a result 208 non-redundant PDB chains were kept. We also only considered the biologically significant Zn atoms (*i.e.* Zn3, Zn4 and Zn5, where Zn*x* denotes Zn atoms binding to x residues), because Zn1 and Zn2 are usually located on the surface of the proteins and have no clear biological function.[12] The 999 CHEDs that bind to Zn3, Zn4 or Zn5 in the 208 PDB chains are regarded as positive samples and the rest 7426 CHEDs are taken as negative samples. In order to assess the performance of ZincExplorer, we also used another dataset named the Zhao_dataset, which was originally collected by Zhao *et al.*[17] In order to guarantee an independent test, the PDB chains present in the Passerini_dataset were removed from Zhao_dataset. Finally, 392 PDB chains and 16 516 sites (2023 zinc-binding CHEDs and 14 493 non-zinc-binding CHEDs) were kept in the Zhao_dataset. More details of the Passerini_dataset and the Zhao_dataset are available in the ESI.†

2.2 SVM-based predictor

We used SVM-light (http://svmlight.joachims.org/) as the implementation function of SVM and selected the radial basis function (RBF) as the kernel function. To maximize the performance of SVM, two parameters (*i.e.* the regularization parameter C and the width parameter γ) were preliminarily optimized through a grid search strategy. First, the ranges of parameters C and γ were set as $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{3}]$, respectively. Then, a step size of 2 was assigned for $\log_2 C$ and $\log_2 \gamma$, respectively, which resulted in a total number of $11 \times 10 = 110$ grids. Finally, all the 110 grids were evaluated through 5-fold cross-validation on the Passerini_dataset to determine the optimal parameters ($C = 2.0$ and $\gamma = 0.0078125$). Each CHED of both positive and negative samples was represented by a residue fragment with the CHED located at the center of the fragment. We tested different fragment sizes ranging from 13 to 25 through 5-fold cross-validation on the Passerini_dataset. As a result, the optimal window size of the fragment was assigned as 21. Each fragment was then converted into an input feature vector for SVM training and testing. The feature representation is detailed as follows.

2.2.1 PSSM, RW-GRMTP and Shannon entropy. Previous studies have shown that sequence conservation is an important feature for zinc-binding site prediction.[9,10,12] Zinc-binding sites are generally more conserved than non-zinc-binding sites. It is worth noting that sequence conservation has also been widely used for identifying other functional sites such as catalytic residues.[20–24] To capture the useful information contained in residue conservation as much as possible, we extracted three types of features from the PSI-BLAST profiles. We first obtained the PSSM profile by running PSI-BLAST against the NCBI nr90 database (version as of October 2009) with parameters $-h$ of 0.001 and $-j$ of 3. Then, we extracted the PSSM, the ''relative weight of gapless real matches to pseudocounts'' (RW-GRMTP), and the weighted observed percentages (WOP), respectively. The 20-dimensional PSSM vector represents the log-likelihood of the substitution of 20 amino acids at a specific position,[25] reflecting the conservation level of the corresponding amino acids. The 2-dimensional RW-GRMTP (*i.e.* the last two columns in the PSSM profile) reflects the aligned residue number at that position. Both the 20-dimensional PSSM vector and 2-dimensional

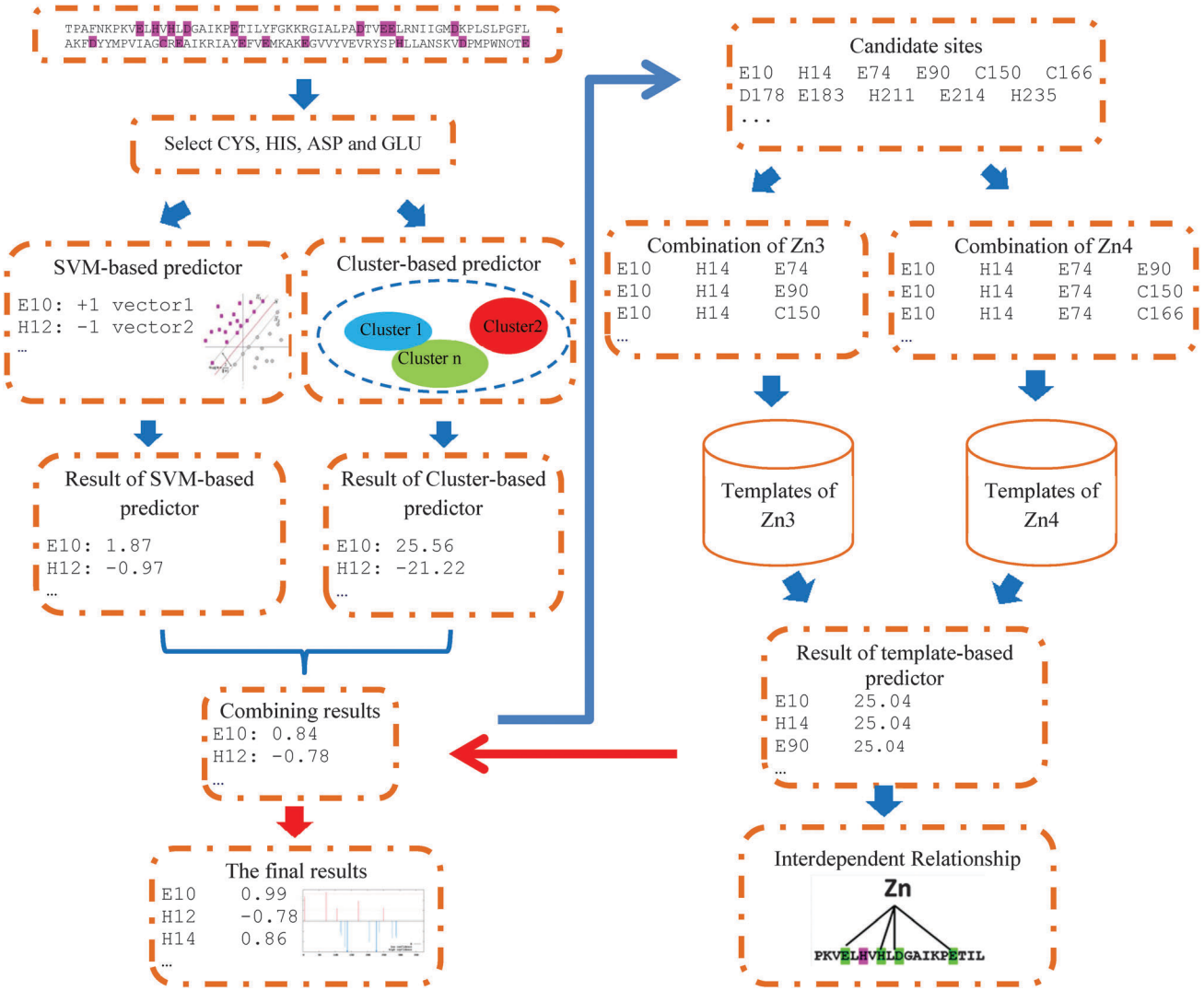

Fig. 1 The complete workflow of the whole prediction procedures of ZincExplorer. First, the results of the SVM-based predictor and the cluster-based predictor were combined; then, the candidate sites were selected based on the combined prediction results; finally, the prediction results obtained by the template-based predictor were fed back to the combined results of the SVM-based and cluster-based predictors to generate the final prediction results. In the meanwhile, the IR between the predicted candidate zinc-binding residues was also established.

RW-GRMTP vector were further normalized by $1/(1 + e^{-x})$. The 20-dimensional WOP vector reflects the frequency distribution of 20 amino acids at that position.[11] Here we converted the vector into a single feature by computing the Shannon entropy:

$$\text{Entropy} = \sum_{i=1}^{20} -p_i \log(p_i) \quad (1)$$

$$p_i = n_i \Big/ \sum_{j=1}^{20} n_j \quad (2)$$

where n_j is the jth element of the WOP vector. Thus, for a fragment of 21 residues, the dimensionality of the features extracted from the PSI-BLAST profile is $23 \times 21 = 483$.

2.2.2 *k*-Spaced amino acid pair composition. The composition of k-spaced amino acid pairs (CKSAAP) has been successfully used in many prediction tasks.[26–29] It describes the short-range interactions of residues within a sequence or a sequence fragment. In this work, k = 0, 1, 2, 3, 4 and 5 were taken into account. For each k, there are 400 possible residue pairs. Therefore, the total dimensionality of CKSAAP is 2400. Here, we proposed a simplified method to reduce the dimensionality of CKSAAP. We computed the occurring frequency for each amino acid pair based on all the positive samples of the Zhao_dataset (or based on all the positive samples of the Passerini_dataset when performing the independent test). Only those amino acid pairs having a frequency larger than a threshold value of 0.006 were selected. Note that the optimal threshold was obtained by testing different values from 0.001 to 0.01 with a step size of 0.001 through 5-fold cross-validation. Finally, 51 and 64 amino acid pairs were selected from the Passerini_dataset and the Zhao_dataset, respectively. To avoid potential over-fitting, the 64 amino acid pairs obtained from the Zhao_dataset were used for the 5-fold cross-validation tests. For sequence

fragment encoding, we used a 64-dimensional vector labeled with values 0 or 1. If a sequence fragment contained a given amino acid pair, the corresponding element would be 1. Otherwise, the corresponding element would be 0.

2.2.3 CHED type. In order to discriminate the zinc-binding site types, the centered CHEDs were encoded for training the classifiers. More specifically, a 4-dimensional vector was used to encode the CHED type, for example, "C" was encoded as "1000", "H" was encoded as "0100", "D" was encoded as "0010", while "E" was encoded as "0001".

2.2.4 The feature vector dimensionality of the SVM-based predictor. To construct the SVM-based predictor, all the feature encodings mentioned above were concatenated together to form a large feature vector. In summary, the total dimensionality of the feature vector is 551 in the 5-fold cross-validation tests, including 420-dimensional PSSM, 42-dimensional RW-GRMTP, 21-dimensional Shannon entropy, 4-dimensional CHED type as well as 64-dimensional CKSAAP selected from the Zhao_dataset.

To avoid any potential over-estimation in the independent test, the employed CKSAAP encoding should not be selected from the Zhao_dataset. Thus, we used 51-dimensional CKSAAP selected from the Passerini_dataset to train the model for the independent test. Finally, only 538-dimensional feature vectors were used in the independent test.

2.3 Cluster-based predictor

Because most of the zinc-binding sites are highly conserved, we hypothesize that zinc-binding sites along with their surrounding residues could be divided into a limited number of clusters that have similar evolutionary characteristics. Based on this hypothesis we were able to predict whether a given sample was a zinc-binding site or not by calculating the similarity between the given sample and the existing clusters. Training of this novel cluster-based predictor contains the following four major steps.

Step 1: divide all the positive samples in the training data into groups

All the positive samples in the training data were divided into four groups according to their central amino acid types (*i.e.* CYS, HIS, GLU and ASP groups), since the evolutionary characteristics may be diverse for different zinc-binding residues. In the proposed cluster-based predictor, we used the same searching strategy as mentioned in the SVM-based predictor to find the optimal window size of samples. At last, the optimal window size was set as 15.

Step 2: calculate the similarity matrix

The similarity between any two positive samples in a group can be measured by the PICASSO3 score[30] between the two corresponding profiles, which is defined as:

$$S(a,b)=\sum_{i=1}^{15}\left(\sum_{j=1}^{20}\left(a_{ij}\log\left(\frac{b_{ij}}{f_j}\right)+b_{ij}\log\left(\frac{a_{ij}}{f_j}\right)\right)\right) \quad (3)$$

where a and b are the profiles of two sequence fragments in a group, while f is the background frequency of 20 amino acid types. $a_{i,j}$ is calculated as:

$$a_{i,j}=e^{M_{i,j}\lambda_u}f_j \quad (4)$$

where M denotes the PSSM vector, and λ_u is the standard ungapped Lambda value in the PSSM file of profile a. Likewise, $b_{i,j}$ can be derived from the PSSM profile of b. To facilitate the following clustering step, the element values in the similarity matrix are further modified as follows:

$$\mathrm{Matrix}(a,b)=\begin{cases}S(a,b) & \text{if } (S(a,b)\geq 0)\\ 0 & \text{if } (S(a,b)<0)\end{cases} \quad (5)$$

Step 3: clustering

The Markov Cluster Algorithm (MCL) (http://micans.org/mcl/) was employed to divide the samples of a group into clusters with the inflation parameter $-I$ set at 6.

Step 4: compute the prediction score of the test sample

For a query sample U, suppose its central amino acid type is CYS, the pseudocode of the prediction procedure can be described as follows:

for each cluster C_i in the CYS-centered group
{
  for each sample S_j in C_i
  {
    calculate the similarity score $S(U, S_j)$;
    Score_i+ = $S(U, S_j)$;
  }
  Score_i/ = $N(C_i)$; #$N(C_i)$ is the sample number in cluster C_i
}
$\mathrm{Score}_{\mathrm{cluster}}(U)$ = max{Score_i}
return $\mathrm{Score}_{\mathrm{cluster}}(U)$;

where $\mathrm{Score}_{\mathrm{cluster}}(U)$ is the result of the cluster-based method.

2.4 Linear combination between the SVM-based and cluster-based predictors

We first rescaled the prediction scores of the SVM-based and cluster-based predictors to [0, 1] by using the function $y = 1/(1+e^{-x})$. Then, we combined the output scores resulting from these two predictors using the following linear formula:

$$\mathrm{Score}_{\mathrm{SVM+cluster}}=\alpha\times\mathrm{Score}_{\mathrm{SVM}}+(1-\alpha)\times\mathrm{Score}_{\mathrm{cluster}} \quad (6)$$

To determine the optimal value of α, we tested the performance of $\mathrm{Score}_{\mathrm{SVM+cluster}}$ using different α values, ranging from 0.0 to 1.0 at an interval of 0.05. In this study, the optimal value of α was assigned to 0.8.

2.5 Template-based predictor

The Zn atom needs to bind at least three residues in order to play its functional role.[1] The residues that coordinate with the same Zn atom have an Interdependent Relationship (IR), which are denoted 3 or 4-residues and are jointly involved in the coordination of the same Zn ion.[31] It is possible to detect IRs of zinc-binding residues in a protein using the known IRs in the training data of Zn3 or Zn4 as templates. Furthermore, the obtained IR information could be used in turn to improve the prediction accuracy.

Here we propose a template-based predictor. A Zn3/Zn4 template is defined as three residues (or four residues) that

are coordinated by the same Zn atom and their surrounding residues. The window size was set as 15 in this study. The procedures for implementing the algorithm are described as follows:

(a) Extract all the Zn3/Zn4 templates that belong to the same PDB chain in the training data.

(b) Select the candidate sites for each protein in the test set. Only the sites predicted from the SVM- and cluster-based predictors with a relatively high prediction score (*i.e.* $\text{Score}_{\text{svm+cluster}} > 0.19$) were selected. The purpose of this filtering is to reduce the computational complexity of the template-based predictor. It was estimated that many non-zinc-binding sites were filtered and more than 95% of the zinc-binding sites were still retained after this filtering.

(c) Enumerate all the combinations of 3-residue (or 4-residue) groups of the candidate sites in the tested protein.

(d) Calculate the similarity score between a 3-residue group (or a 4-residue group) in the tested protein and all the Zn3 (or Zn4) templates in the training data. The procedure is generally time consuming, which is detailed in the ESI.†

(e) Extract all the 3-residue (or 4-residue) groups in the tested protein whose similarity score is larger than 0. Then select the 3-residue (or 4-residue) group that has the highest similarity score and remove the 3-residue (4-residue) groups that have identical residues with the selected 3-residue (or 4-residue) group.

(f) Repeat the above step (e) on the unselected groups until no other 3-residue (or 4-residue) groups can be further selected or removed.

Finally, the remaining residue groups are regarded as the predicted IRs in the tested protein, which can be used to refine the zinc-binding site prediction. Briefly, all the sites of the selected 3-residue or 4-residue groups form a sample set Ω. The final prediction score of a potential zinc-binding site S can be defined as follows:

$$\text{Score}_{\text{Final}} = \begin{cases} \text{Score}_{\text{SVM+cluster}} & \text{if}(S \notin \Omega) \\ \text{Score}_{\text{SVM+cluster}} + \delta & \text{if}(S \in \Omega) \end{cases} \quad (7)$$

where δ is the reliability parameter. We varied δ from 0.05 to 0.40 at an interval of 0.05 to benchmark the performance of $\text{Score}_{\text{Final}}$ on the Passerini_dataset and the optimal δ was set to 0.1.

2.6 Performance assessment

Four measurements [*i.e.* precision, recall, specificity and Matthew correlation coefficient (MCC)] are used to evaluate the prediction performance, which are, respectively, defined as:

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \quad (8)$$

$$\text{Recall} = \text{Sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (9)$$

$$\text{Specificity} = \frac{\text{tn}}{\text{tn} + \text{fp}} \quad (10)$$

$$\text{MCC} = \frac{\text{tp} \times \text{tn} - \text{fp} \times \text{fn}}{\sqrt{(\text{tp} + \text{fp}) \times (\text{tp} + \text{fn}) \times (\text{tn} + \text{fn}) \times (\text{tn} + \text{fp})}} \quad (11)$$

where tp, fp, fn and tn represent the true positives, false positives, false negatives and true negatives, respectively. Due to the unbalanced dataset of the positive and negative samples, the Recall–Precision Curve (RPC) which plots precision as a function of recall for all the possible thresholds is also used to evaluate the performance of our method, ZincExplorer, as it is suitable for dealing with unbalanced samples.[32] Furthermore, the overall performance of ZincExplorer is also quantified by the corresponding Area Under the Recall–Precision Curve (AURPC). The closer an AURPC value is to 1, the better the performance of a prediction method is.

3 Results and discussion

3.1 Performance of ZincExplorer

Through 5-fold cross-validation tests on the Passerini_dataset, ZincExplorer achieved an AUPRC of 0.851, and a precision of 85.6% (specificity = 98.4%, MCC = 0.747) at a recall of 70% (Fig. 2). As a hybrid method, the performance of ZincExplorer's component predictors was also assessed individually. As a result, the SVM-based predictor alone achieved an AURPC of 0.813, and a 79.1% precision (specificity = 97.5%, MCC = 0.711) at a recall of 70.0% (Fig. 2). In contrast, the cluster-based predictor alone only yielded an AURPC of 0.664. After combining the SVM-based and cluster-based predictors, the resultant predictor attained an AUPRC of 0.840 (Z-test,[33] p-value = 4.28×10^{-12}) and the corresponding precision at the 70% recall control was 84.6% (specificity = 98.3%, MCC = 0.743). Although the template-based predictor alone cannot be used to predict the zinc-binding site

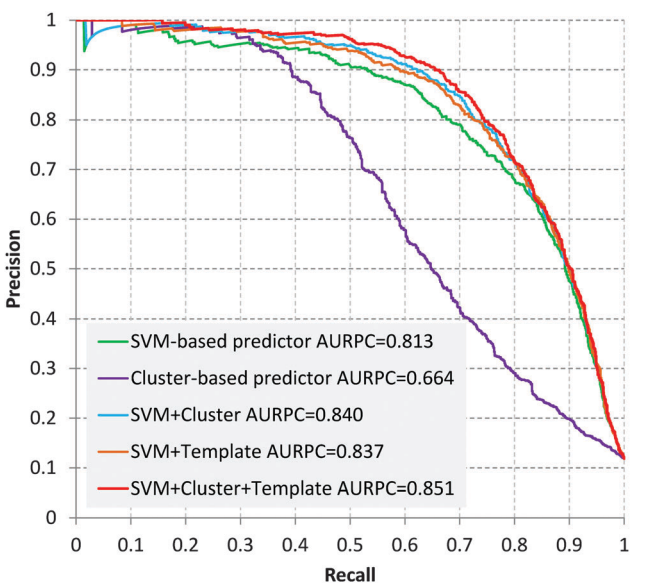

Fig. 2 Recall–precision curves for component predictors of ZincExplorer based on the Passerini_dataset. The performance of all the predictors was evaluated using the 5-fold cross-validation tests.

directly, it is indispensable for ZincExplorer to achieve the best performance.

3.2 The power of a linear combination between the SVM- and cluster-based predictors

The highlight of the current work is to combine the SVM- and cluster-based predictors into the computational framework of ZincExplorer. It is well known that SVM is suitable for dealing with binary classification tasks. For some prediction problems, however, it is more appropriate to cluster the samples into several subgroups rather than two classes only. The cluster-based predictor was designed to address this issue. For example, 19 zinc-binding sites of CYS in the Passerini_dataset were grouped into one cluster. All of these 19 sites shared a CxxC motif at the same position (Fig. 3), which exemplified that the cluster-based predictor could effectively group the zinc-binding sites with a similar sequence pattern together and was thus suitable for the prediction. Therefore, it is understandable that we could achieve a better performance by combining the SVM- and cluster-based predictors. To demonstrate this, we further compared the performance of the SVM- and cluster-based predictors for four zinc-binding residue types CYS, HIS, GLU and ASP, respectively (Table 1). It can be seen that the prediction performance of CYS was clearly the best among the four amino acid types, with the AURPC ranging between 0.927 and 0.947. We can also see that, on average, there was an accuracy increase when comparing the performance of the combined SVM-based predictor and the cluster-based predictor (SVM + cluster) with that of the individual predictor (SVM) (Table 1). This tendency holds across all the four major zinc-binding types of residues. The reason why the prediction accuracy of CYS is much higher than the other three types might be due to the relatively high abundance of zinc-binding CYS in the training data. In addition, SVM has an excellent ability to find comprehensive prediction rules that can cover most cases of CYS and identify the zinc-binding CYS from all candidate CYSs, with a sole prerequisite that abundant and unbiased training data must be available in order to infer the rules. With respect to the other three amino acid types, especially GLU and ASP, their sample size is so small in nature that it is difficult for SVM to find complete and unbiased rules to better distinguish them. For the cluster-based predictor, although its discriminating ability is not as good as SVM, it has an advantage of not being overly dependent on the sample size. For a query sample, it can usually detect a cluster as its neighbor. If the similarity score between the query sample and the neighboring cluster was larger than the prediction cutoff, the query sample would be predicted as zinc-binding. For CYS, HIS, GLU and ASP, there was a performance improvement when combining the SVM-based predictor and the cluster-based predictor. These observations suggest that the strategy of combining the SVM-based and cluster-based predictors could make use of their respective advantages of the two algorithms and lead to a significant performance improvement. It is worth noting that the strategy of the cluster-based predictor has been successfully applied to other prediction tasks such as the prediction of protein phosphorylation sites.[34,35] Considering the fact that many bioinformatics classification tasks are not two-class problems, we expect that this powerful integration strategy between the SVM- and cluster-based predictors may serve as an effective framework to address many other diverse classification tasks in the fields of bioinformatics and computational biology.

Table 1 Performance comparison of the SVM-based, cluster-based and template-based predictors on the Passerini_dataset for zinc-binding CHEDs. The performance was evaluated by the AURPC values

Zinc-binding type	SVM	SVM + cluster	SVM + template	SVM + cluster + template
CYS	0.927	0.941	0.939	0.947
HIS	0.751	0.798	0.793	0.815
GLU	0.168	0.277	0.221	0.298
ASP	0.356	0.461	0.470	0.516
ALL	0.813	0.840	0.837	0.851

In theory, the cluster-based predictor is similar to the *k*-nearest-neighbor (KNN) approach. We further tested the KNN method in our work. The results showed that the introduction of KNN achieved considerably lower prediction accuracy in comparison to the use of the cluster-based predictor (data not shown). Indeed, there is still a clear methodology difference between these two methods. The neighbor number *k* for KNN is fixed, while the neighbor number is variable in the clustering-based method. The above computational experiment suggests that the variable neighbor number makes more sense than the fixed neighbor number in our work.

3.3 The power of integration with the template-based predictor

In order for a Zn atom to play its functional role, it has to coordinate with three or four (even more) zinc-binding residues.

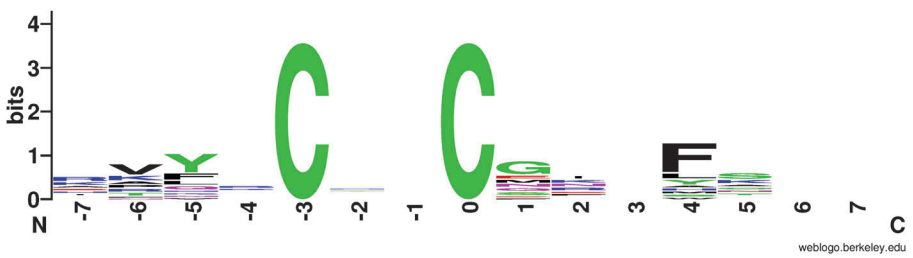

Fig. 3 Sequence logo of a representative zinc-binding site cluster containing 19 zinc-binding sites of CYS. All of these 19 sites shared a CxxC motif at the same position. The sequence logo was prepared using the web server http://weblogo.berkeley.edu/logo.cgi.

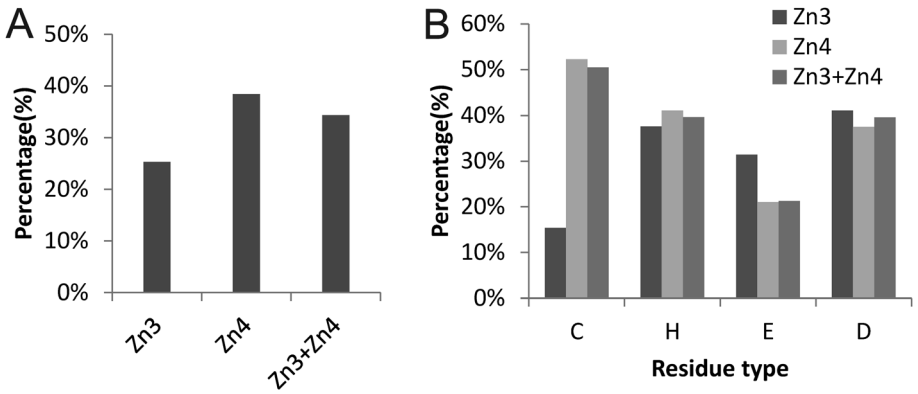

Fig. 4 The IR prediction accuracy of the template-based predictor through 5-fold cross-validation on the Passerini_dataset. The prediction accuracy was defined as: accuracy = $N_{predicted}/N_{all}$, where $N_{predicted}$ is the number of Zinc3/Zinc4 residue groups correctly identified by the template-based predictor, and N_{all} is the total number of the corresponding Zinc3/Zinc4 groups in the Passerini_dataset. (A) The overall accuracy values; (B) the accuracy values measured at the residue level.

As such, the IR of zinc-binding residues that bind to the same Zn atom can be effectively employed to improve the predictive performance of our method (Fig. 2). The candidate sites were selected based on the intermediate prediction result of the combined SVM-based and cluster-based predictors. Then the output of the template-based predictor was fed back to the intermediate result to augment the final prediction. Meanwhile, the IR of the predicted zinc-binding sites was also inferred. Evaluated by 5-fold cross-validation tests on the Passerini_dataset, the template-based predictor achieved an accuracy of 25.3% for Zn3 and 38.4% for Zn4 (Fig. 4A), respectively. We further calculated the corresponding accuracy values at the residue level. As shown in Fig. 4B, for Zn3 ASP (41.1%) was the best predicted residue, followed by HIS (37.6%) and GLU (31.4%), whereas CYS (15.4%) was predicted worst. For Zn4, CYS (52.3%) was predicted with the best performance, followed by HIS (41.1%) and ASP (37.5%), whereas GLU (21.1%) was predicted worst. Although the template-based predictor's performance in identifying the IR of predicted zinc-binding sites is still not perfect, it provides more comprehensive zinc-binding site information for further experimental validation. In addition, the predicted IR information is also useful for the 3D structure prediction of the query protein. To the best of our knowledge, there is only one existing method, MetalDetector,[31] which was designed to provide the IR information of the predicted zinc-binding residues. However, MetalDetector failed to exploit this to improve the performance of the predictor and could only predict two types of zinc-binding residues, *i.e.* CYS and HIS, rather than four major types.

3.4 The web server of ZincExplorer

To facilitate the research community, we have implemented an online web server of our ZincExplorer method, which is freely accessible at http://protein.cau.edu.cn/ZincExplorer. It accepts a query sequence in the RAW or the FASTA format. The Passerini_dataset was used as the training dataset for training the prediction models of this web server. The prediction outputs include residue position, prediction score, zinc-binding annotation and IRs. It is estimated that a prediction score of ≥ 0.4 corresponds to a false positive rate of $\leq 2.5\%$, *i.e.*, 97.5% confidence level. A four-CPU Dell Linux machine with 16 GB of main memory hosts the web server. The computational time required for processing a query sequence of 600 amino acids is usually no more than ten minutes. For proteome-wide applications, users are strongly recommended to download the stand-alone version and run it locally.

3.5 Comparison with other methods

We further compared our developed tool ZincExplorer with two other state-of-the-art sequence-based tools ZincFinder[10] and ZincPred[12] based on the independent test dataset, *i.e.* the Zhao_dataset. For making a comparison, the stand-alone versions of ZincFinder and ZincPred were, respectively, downloaded from their websites. We then evaluated the prediction performance of zinc-binding sites using these three tools at both the overall level and individual CYS, HIS, GLU and ASP levels, respectively. In particular, as both ZincFinder and ZincPred cannot provide prediction scores for certain CHEDs in some protein sequences, we considered these CHEDs as predicted negatives and accordingly assigned their prediction scores as 0 for the sake of facilitating the comparison.

As a result, the performance comparison clearly demonstrates that ZincExplorer has outperformed the other two sequence-based tools. ZincExplorer reached an AUPRC of 0.907, which represents 8% and 3% increase than ZincPred (AURPC = 0.823; *Z*-test, *p*-value = 0) and ZincFinder (AURPC = 0.873; *Z*-test, *p*-value = 3.23×10^{-9}), respectively (Fig. 5A). At the 70% recall control, ZincExplorer achieved a precision of 94.3% (specificity = 99.4%, MCC = 0.792), while ZincPred and ZincFinder achieved a precision of 82.2% (specificity = 97.8%, MCC = 0.729) and 91.6% (specificity = 99.1%, MCC = 0.778), respectively.

To provide further insights into the predictive power of the three tools, we further examined their performance on predicting zinc-binding CYS, HIS, GLU and ASP separately (Table 2 and Fig. 6). We can see that, in general, CYS was predicted with the overall best performance by all the predictors, followed by HIS and GLU. Among the four types of zinc-binding

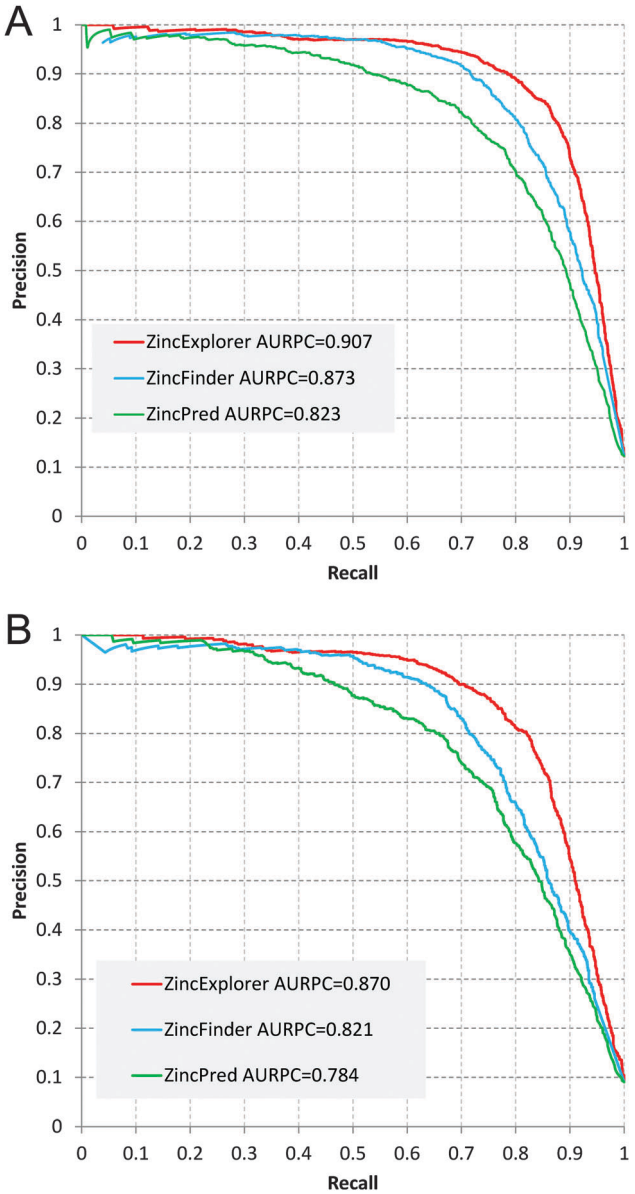

Fig. 5 Performance comparison of the three prediction tools (ZincExplorer, ZincFinder and ZincPred) based on the Zhao_dataset (A) and Zhao_dataset_sub (B).

residues, ASP was predicted worst. As mentioned above, because CYS is the most abundant form of zinc-binding residues and the machine learning predictors like SVM tend to better learn the underlying rules of sample classification, given large amounts of available data, all these three tools performed well in terms of the CYS zinc-binding type prediction. Accordingly, the AURPC values of ZincExplorer, ZincPred and ZincFinder have reached 0.973, 0.960 and 0.960 (Table 2), respectively. Comparatively, ZincExplorer performed slightly better than ZincPred and ZincFinder for predicting CYS. And compared with the other two tools, ZincExplorer also exhibited a great improvement in predicting HIS, GLU and ASP. For example, it achieved 21.3% and 25.9% higher in AURPC than ZincPred and ZincFinder for predicting GLU and 2.7–16% higher in AURPC for predicting HIS and ASP, respectively. The consistently improved performance by ZincExplorer across the four types of zinc-binding residues implies the advantage of incorporating the cluster-based predictor into the SVM-based machine learning framework. To make an objective and fair comparison, we also compared ZincExplorer with ZincPred and ZincFinder by only evaluating those predicted CHEDs for which both ZincPred and ZincFinder yielded valid prediction outputs, and similar performance observations were obtained (see ESI† for detailed results).

To provide a more rigorous method of comparison, we also built a subset from the Zhao_dataset by removing the sequences sharing >20% sequence identity with any sequence in the Passerini_dataset. In the resulting subset (*i.e.* Zhao_dataset_sub), there were 257 PDB chains and 13 789 sites (1244 zinc-binding CHEDs and 12 545 non-zinc-binding CHEDs). Again, ZincExplorer outperformed ZincPred and ZincFinder on the Zhao_dataset_sub dataset, although all these three methods showed less powerful performance. The AUPRC of ZincExplorer was 0.870, while the AUPRC of ZincPred and ZincFinder was 0.784 and 0.821, respectively (Fig. 5B).

Moreover, we also compared the result of our template-based predictor with the MetalDetector approach by selecting the Zn3 and Zn4 templates comprising CYS and HIS, since the MetalDetector approach could only predict the two residue types. We installed the stand-alone version of MetalDetector in our local machine and did the prediction for the whole Passerini_dataset. For Zn3 and Zn4 groups, our template-based predictor achieved an accuracy of 84.0% and 47.4%, while MetalDetector yielded an accuracy of 32.0% and 32.2%, respectively (Fig. 7). These results indicate that our template-based predictor outperformed MetalDetector considerably.

Table 2 Performance comparison of the three prediction tools, ZincExplorer, ZincFinder and ZincPred on the Zhao_dataset. The performance was evaluated using the precision, specificity and MCC values at the 70% recall control as well as the AUPRC values

	ALL		CYS		HIS		GLU		ASP	
Tool	Precision/specificity/MCC	AURPC	Precision/specificity/MCC	AURPC	Precision/specificity/MCC	AURPC	Precision/specificity/MCC	AURPC	Precision/specificity/MCC	AURPC
ZincExplorer	94.3%/99.4%/0.792	0.907	98.3%/98.3%/0.680	0.973	76.8%/95.4%/0.679	0.798	6.3%/89.8%/0.192	0.421	11.5%/91.6%/0.262	0.402
ZincPred	82.2%/97.8%/0.729	0.823	96.6%/96.3%/0.660	0.960	64.0%/91.4%/0.594	0.745	6.1%/89.2%/0.187	0.208	6.7%/84.9%/0.186	0.375
ZincFinder	91.6%/99.1%/0.778	0.873	97.6%/97.4%/0.671	0.960	62.2%/90.8%/0.581	0.732	6.4%/92.8%/0.171[a]	0.162	6.5%/88.5%/0.161[a]	0.242

[a] The precision, specificity and MCC values were generated based on a recall value of 50%.

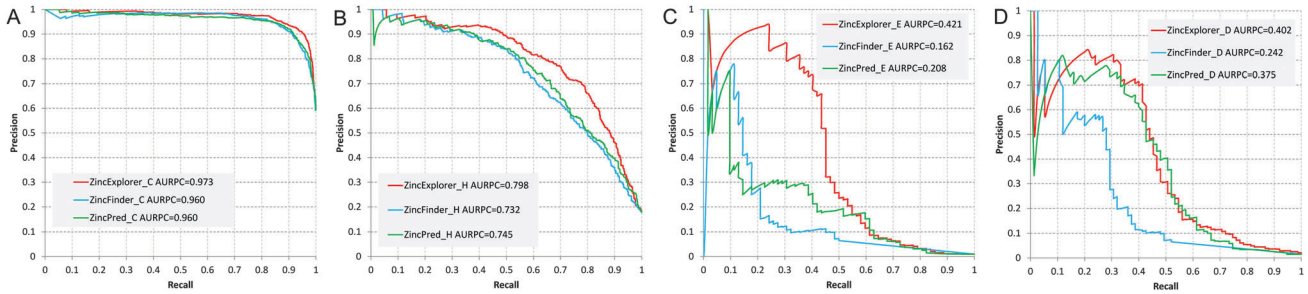

Fig. 6 Performance comparison of ZincExplorer, ZincFinder and ZincPred for predicting zinc-binding CYS, HIS, GLU and ASP, respectively, based on the Zhao_dataset. Recall–precision curves for CYS, HIS, GLU and ASP were plotted in panels A, B, C and D, respectively. In terms of AUPRC, our method improved the prediction significantly (*Z*-test, *p*-value < 0.05) for all the CHEDs with the only exception of the comparison with ZincPred for ASP (*Z*-test, *p*-value = 0.273).

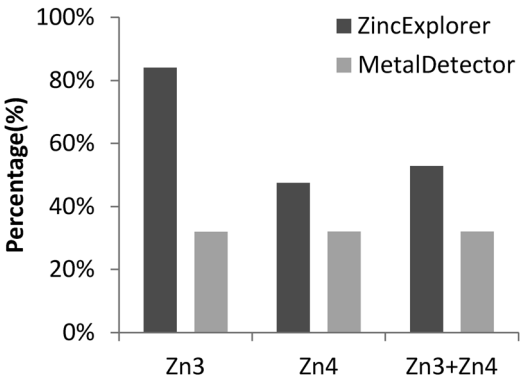

Fig. 7 Performance comparison of the template-based predictor with MetalDetector on the Zn3 or Zn4 templates. The comparison was based merely on two residue types (CYS and ASP) in the Passerini_dataset, since MetalDetector could only predict these two residue types.

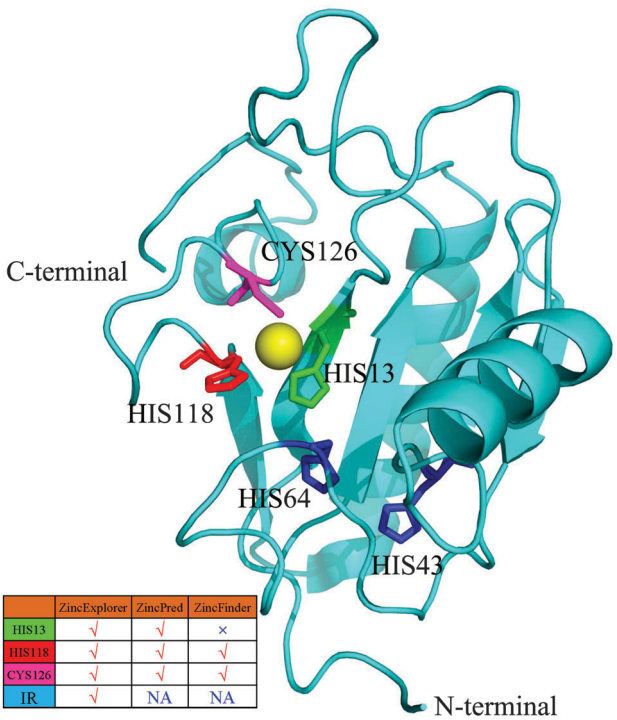

Fig. 8 The prediction performance of the three sequence-based tools (ZincExplorer, ZincPred and ZincFinder) for the prediction of zinc-binding sites in the bacteriophage T7 lysozyme (PDB ID: 1LBA). There are three zinc-binding residues, *i.e.* HIS_13, HIS_118 and CYS_126, which coordinate with the same zinc ion. As can be seen, ZincExplorer correctly predicted all the three residues and accurately inferred the IR between them. As a comparison, ZincPred also identified all the three residues but identified two false zinc-binding sites (*i.e.* HIS_43 and HIS_64). ZincFinder identified two of the three residues. All the residue positions highlighted in the graph are their sequence positions.

To further illustrate the predictive power of ZincExplorer, we performed a case study of bacteriophage T7 lysozyme, as shown in Fig. 8. The structure of the bacteriophage T7 lysozyme (PDB entry: 1LBA) contains a zinc atom, bound directly to three residues (HIS_13, HIS_118 and CYS_126). ZincExplorer correctly predicted all the three zinc-binding sites and accurately identified the IR of these three residues. In contrast, ZincPred also identified all of the three residues as well as predicted two false positive binding sites (*i.e.* HIS_43 and HIS_64). In the case of ZincFinder, it correctly predicted HIS_118 and CYS_126, but failed to detect HIS_13.

Although we have attempted to make a fair performance comparison between different predictors, an entirely fair comparison is hard to achieve due to methodological and dataset differences. Considering that these three predictors were developed using different datasets, one may argue that the performance difference of these three methods could be caused by different training sets. In this work, the NCBI nr90 database (version as of October 2009) was used to extract sequence features, which should be newer than the corresponding database used in ZincPred and ZincFinder. Thus, one may also further argue that the improved performance of ZincExplorer may benefit from an updated version of the NCBI nr90 database. We hope some standard training/testing datasets as well as benchmarking frameworks will be available in the field of zinc-binding site prediction in future. Thus, different prediction methods can be more reliably and unbiasedly benchmarked.

4 Conclusion

In this study, in order to accurately identify zinc-binding sites in proteins, we developed an effective prediction tool termed ZincExplorer, which combines an SVM-based predictor, a cluster-based predictor and an *ad hoc* template-based predictor.

When evaluated based on the 5-fold cross-validation tests on a non-redundant dataset of 208 PDB chains, it achieved an AURPC of 0.851 for all CHEDs. When tested on an independent dataset, ZincExplorer clearly outperformed the other two sequence-based tools ZincPred and ZincFinder, especially for predicting the zinc-binding types HIS, GLU and ASP. Furthermore, ZincExplorer can not only predict zinc-binding sites from sequence information but also infer the IRs of the predicted candidate sites that bind to the same zinc ion. The latter feature makes it an attractive tool for facilitating the identification of coordinating zinc-binding sites and the final prediction of the 3D structure of the protein. Built upon an effective combination of the three different component predictors, ZincExplorer was shown to be able to provide a significantly improved performance for predicting HIS, GLU and ASP despite their limited abundance. We believe that ZincExplorer can be applied as a powerful tool to perform *in silico* proteome-wide prediction of zinc-binding proteins, which will greatly aid the current efforts for annotating the zinc proteome.

Acknowledgements

This work was supported by grants from the National Key Basic Research Project of China (2009CB918802), the Hundred Talents Program of the Chinese Academy of Sciences (CAS), the National Natural Science Foundation of China (61202167 and J1103520), Tianjin Municipal Science & Technology Commission (10ZCKFSY05600) and the National Health and Medical Research Council of Australia (NHMRC) (490989). JS is an NHMRC Peter Doherty Fellow and a Recipient of the Hundred Talents Program of CAS.

References

1 C. T. Chasapis, A. C. Loutsidou, C. A. Spiliopoulou and M. E. Stefanidou, *Arch. Toxicol.*, 2012, **86**, 521–534.
2 J. E. Coleman, *Annu. Rev. Biochem.*, 1992, **61**, 897–946.
3 M. Vasak and D. W. Hasler, *Curr. Opin. Chem. Biol.*, 2000, **4**, 177–183.
4 L. Rink and P. Gabriel, *Proc. Nutr. Soc.*, 2000, **59**, 541–552.
5 E. Mocchegiani, M. Muzzioli and R. Giacconi, *Trends Pharmacol. Sci.*, 2000, **21**, 205–208.
6 M. Stefanidou, C. Maravelias, A. Dona and C. Spiliopoulou, *Arch. Toxicol.*, 2006, **80**, 1–9.
7 J. C. Ebert and R. B. Altman, *Protein Sci.*, 2008, **17**, 54–65.
8 R. Tupler, G. Perini and M. R. Green, *Nature*, 2001, **409**, 832–833.
9 A. Passerini, M. Punta, A. Ceroni, B. Rost and P. Frasconi, *Proteins*, 2006, **65**, 305–316.
10 A. Passerini, C. Andreini, S. Menchetti, A. Rosato and P. Frasconi, *BMC Bioinf.*, 2007, **8**, 39.
11 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
12 N. Shu, T. Zhou and S. Hovmoller, *Bioinformatics*, 2008, **24**, 775–782.
13 F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *Eur. J. Biochem.*, 1977, **80**, 319–324.
14 K. Goyal and S. C. Mande, *Proteins*, 2008, **70**, 1206–1218.
15 M. Babor, S. Gerzon, B. Raveh, V. Sobolev and M. Edelman, *Proteins*, 2008, **70**, 208–217.
16 A. J. Bordner, *Bioinformatics*, 2008, **24**, 2865–2871.
17 W. Zhao, M. Xu, Z. Liang, B. Ding, L. Niu, H. Liu and M. Teng, *Bioinformatics*, 2011, **27**, 1262–1268.
18 C. Zheng, M. Wang, K. Takemoto, T. Akutsu, Z. Zhang and J. Song, *PLoS One*, 2012, **7**, e49716.
19 B. Rost, *Protein Eng.*, 1999, **12**, 85–94.
20 T. Zhang, H. Zhang, K. Chen, S. Shen, J. Ruan and L. Kurgan, *Bioinformatics*, 2008, **24**, 2329–2338.
21 J. A. Capra and M. Singh, *Bioinformatics*, 2007, **23**, 1875–1882.
22 J. D. Fischer, C. E. Mayer and J. Soding, *Bioinformatics*, 2008, **24**, 613–620.
23 N. V. Petrova and C. H. Wu, *BMC Bioinf.*, 2006, **7**, 312.
24 L. Han, Y. J. Zhang, J. Song, M. S. Liu and Z. Zhang, *PLoS One*, 2012, **7**, e41370.
25 D. T. Jones, *J. Mol. Biol.*, 1999, **292**, 195–202.
26 K. Chen, L. A. Kurgan and J. Ruan, *BMC Struct. Biol.*, 2007, **7**, 25.
27 K. Chen, L. A. Kurgan and J. Ruan, *J. Comput. Chem.*, 2008, **29**, 1596–1604.
28 Y. Z. Chen, Y. R. Tang, Z. Y. Sheng and Z. Zhang, *BMC Bioinf.*, 2008, **9**, 101.
29 Z. Chen, Y. Z. Chen, X. F. Wang, C. Wang, R. X. Yan and Z. Zhang, *PLoS One*, 2011, **6**, e22930.
30 D. Mittelman, R. Sadreyev and N. Grishin, *Bioinformatics*, 2003, **19**, 1531–1539.
31 A. Passerini, M. Lippi and P. Frasconi, *Nucleic Acids Res.*, 2011, **39**, W288–W292.
32 J. P. Zhang, E. Bloedorn, L. Rosen and D. Venese, Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on, 2004.
33 X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez and M. Muller, *BMC Bioinf.*, 2011, **12**, 77.
34 F. F. Zhou, Y. Xue, G. L. Chen and X. Yao, *Biochem. Biophys. Res. Commun.*, 2004, **325**, 1443–1448.
35 Y. Xue, J. Ren, X. Gao, C. Jin, L. Wen and X. Yao, *Mol. Cell. Proteomics*, 2008, **7**, 1598–1608.

