An Improved Prediction Model for Zinc-Binding Sites in Proteins Based on Bayesian Method

Hui LI, Dechang PI, Chuanming CHEN

Abstract: The zinc ion is the second richest metal ion in organisms. The proteins binding to zinc ions have important biological functions. However, few scholars have integrated the existing tools to predict the zinc-binding sites in proteins. To make up for this gap, this paper combines three well-known prediction tools into an improved model called IBayes_Zinc to predict the zinc-binding sites, and utilizes the advantages of the Bayesian method in handling incomplete or partial missing data. Specifically, the prediction scores of three existing sequence-based prediction tools were adopted, and the missing values were padded, forming an integrated classification tool. Then, the probabilities of positive and negative samples were computed and categorized as the class with higher probabilities. Experiments were conducted on a non-redundant training dataset and an independent testing dataset. The results show that our method surpassed the other three methods by nearly 5–13% in Matthew correlation coefficient (MCC) and outperformed the latter in recall and precision. The research findings promote the detection of zinc-binding sites in protein sequence and the identification of metalloprotein functions.

Keywords: Bayesian; missing; protein prediction; zinc-binding sites

1 INTRODUCTION

The life science in the post-genomic era has found that the interaction between proteins and metal ions is irreplaceable to life activities [1, 2]. The products of the interaction, known as metalloproteins, have important biological functions. The research into this interaction can greatly promote disease treatment and drug development [3, 4]. In the Protein Data Bank (PDB) [5], more than one third of proteins combine with metal ions. Among these ions, the zinc ion is the second richest metal ion in organisms, only behind ferric ion. The binding of the zinc ion to proteins facilitates the catalytic activity and stabilizes the protein structure [6]. As a result, the zinc-binding sites in proteins have become a research hotspot in computational biology.

Traditionally, the zinc-binding sites in proteins are identified by biochemical and biophysical experiments, using mass spectrometry (MS) [7, 8], high-throughput X-ray absorption spectroscopy (HT-XAS) [9], and isothermal titration calorimetry (ITC) [10]. However, these experimental methods are too costly and time-consuming to detect all zinc-binding sites on the level of proteome. To overcome these defects, machine learning algorithms have been developed to predict the zinc-binding sites, such as the Bayesian algorithm, decision tree, support vector machine (SVM), random forest, neural network (NN), regression, and hidden Markov model. These methods can identify the zinc-binding sites in an effective and accurate manner.

The remainder of this paper is organized as follows: Section 2 reviews the existing studies in the related fields; Section 3 presents the materials and methods; Section 4 discusses the results of the experiments; Section 5 puts forward the research conclusions.

2 RELATED WORK

The prediction precision of zinc-binding sites varies with the feature vectors, which are derived from protein attributes and inputted to the prediction algorithms.

Some prediction methods for zinc-binding sites have been established through extraction of protein sequences. For example, Passerini et al. [11] combined the SVM and bidirectional concurrent NN into a two-stage prediction algorithm, which achieved 73% precision and 61% recall in predicting His and Cys sites. Subsequently, they developed ZincFinder [12], an SVM-based automated tool to identify zinc-binding sites. Shu et al. [13] designed ZincPred based on the SVM and homology-based methods, and predicted CHED at 75% precision and 50% recall with this tool. Chen [14] integrated the SVM, cluster and template-based predictors into ZincExplorer, which identified CHED at 86% precision and 70% recall.

Some prediction methods have been proposed based on structural attributes. For instance, Zhao et al. [15] presented 3D template-based metal site prediction (TEMSP), which has a sensitivity of 60% and a selectivity of 80%. He et al. [16] created the mFASD method to recognize varied metal-binding sites, and proved its excellent performance.

There are some other computational prediction tools [17-21] that can determine the binding sites easily. In addition, some prediction tools have been proposed based on multiple types of protein features to promote the prediction precision [22, 23]. However, feature optimization has a limited effect on prediction precision, or proteins have fixed features.

To solve the problem, several machine learning algorithms have been integrated to the existing prediction tools. Considering the availability of sequence-based information of existing proteins and the restrictions on the acquisition of structural data, References [12-14] explore three well-known prediction tools based on sequence features, with the aim to promote the prediction precision and robustness. Reference [24] applies linear regression to integrate the three tools and achieves good prediction results. But the integrated method only fills a special value to the attribute values that are lost. Later, Reference [25] establishes a prediction model for zinc-binding sites based on Bayesian method, which is insensitive to missing data, requires a few estimation parameters, and boasts a high classification efficiency. Then, an improved prediction

model for zinc-binding sites was developed based on Bayesian algorithm to better fill and process the missing attribute values, drawing on the ability of the Bayesian decision theory to handle incomplete or partially missing data.

3 MATERIALS AND METHODS
3.1 Dataset

The dataset proposed by Zhao et al. [15] (Zhao_dataset) was taken as the training dataset. This is a unified open standard dataset popular among researchers. To prevent over-fitting, some redundant protein sequences were removed, leaving 392 protein chains. These reserved protein chains include 2,023 zinc-binding sites and 14,493 non-zinc-binding sites.

To verify the precision and robustness of our method, the zinc-binding protein complexes, which were submitted to the PDB in recent years, were adopted as the testing dataset. To remove the unreliable, repetitive, short and non-binding chains, the upper limit of structure resolution was set to 2.5 Å, and the peptide chains with homology higher than 70% and sequence redundancy less than 20% were eliminated through X-ray diffraction. Finally, the testing dataset (Final_dataset) include 213 randomly selected protein chains, including 1,017 zinc-binding sites and 10,148 non-zinc-binding sites.

3.2 Improved Prediction Model Based on Bayesian Method

Firstly, each residue of every protein chain was predicted by ZincExplorer, ZincFinder, and ZincPred, respectively, and the results were recorded as a vector of three fractional values x_i ($i = 1, 2, 3$), with $x_i \in [0, 1]$. In Reference [24], some sites in the protein sequence have no predictive score. Considering the extremely small ratio of actual binding sites, the predictive scores of these sites were set to 0, i.e. these sites were considered as non-binding by default.
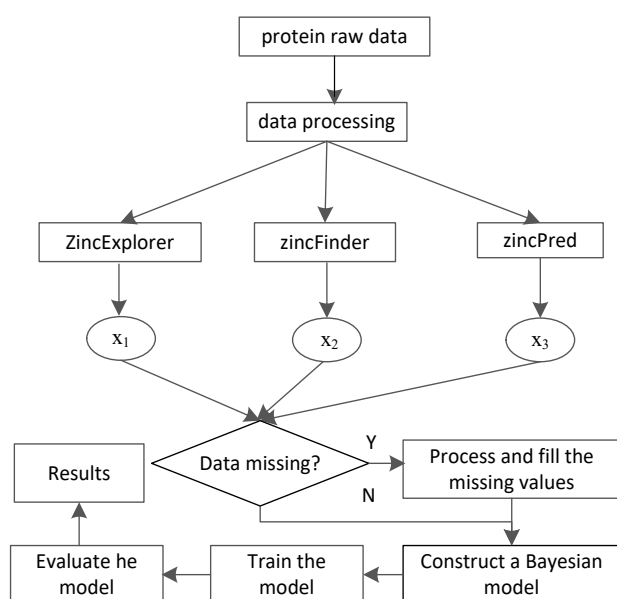

Figure 1 Framework of IBayes_Zinc model

Since different attributes are independent, three scores were used as the attribute variables of the Bayesian

classifier, whose advantages were fully utilized to tackle the missing values of some attributes.

The class variable cutoff was defined to reflect whether a site is zinc-binding or not. Then, the classification model IBayes_Zinc was constructed based on the independence of attributes and class variable. The model framework is shown in Fig. 1.

If some sites in a protein sequence have no predictive score, then some attribute values must be missing. In the case that all attribute values are missing, these data should not be used to compute probability. In fact, the probability computation depends on the number of attribute values appearing in the training samples, rather than the total number of the training samples. In the case that some attribute values are missing, the missing values should be padded. On this basis, a Bayesian model can be constructed to train and predict the binding sites.

Let $x = (x_1, x_2, x_3)$ be a sample eigenvector containing three fractional sample eigenvalues, where $x_j \begin{cases} < cutoff & \text{the sample with feature } j \text{ is negative} \\ < cutoff & \text{the sample with feature } j \text{ is positive} \end{cases}$. Note that $j \in [1, 3]$, and $cutoff$ is the class variable.

Let Ω be the total sample space and N be the number of total samples. Obviously, N is the sum of the number of positive samples N_1 and the number of negative samples N_2, as well as the sum of the number of samples predicted to be positive n_1 and the number of samples predicted to be negative n_0. Then, n_1 can be divided to the number of correctly predicted positive samples n_{11} and falsely predicted positive samples n_{10}, while n_0 can be divided to the number of correctly predicted negative samples n_{00} and the number of falsely predicted negative samples n_{01}.

It is assumed that event A_1 indicates that samples are positive, event A_0 indicates that samples are negative, and $A_1 \cup A_0 = \Omega$; event $B(j)_1$ indicates that the samples with feature j(j=1, 2, 3) are predicted as positive, and event $B(j)_0$ indicates that the samples with feature j are predicted as negative; event $A_1 B(j)_1$ represents that the samples are positive, and the prediction is true; event $A_1 B(j)_0$ represents that the samples are positive, but the prediction is false; event $A_0 B(j)_1$ represents that the samples are negative, and the prediction is false; event $A_0 B(j)_0$ represents that the samples are negative, and the prediction is true.

Based on the Bayesian theory, the probability of sample X to be predicted as positive can be expressed as:

$$P_1 = P(A_1 \mid X) = \frac{\prod_{j=1}^{3} P(A_1, x_j)}{P(X)P(A_1)^2}, \tag{1}$$

where, $P(A_1, x_j) = \begin{cases} P(A_1, B(j)_1) \text{ if } x_j \geq cutoff \\ P(A_1, B(j)_0) \text{ if } x_j < cutoff \end{cases}$.

Similarly, the probability of sample X to be predicted as negative can be expressed as:

$$P_0 = P(A_0 \mid X) = \frac{\prod_{j=1}^{3} P(A_0, x_j)}{P(X)P(A_0)^2}, \tag{2}$$

where, $P(A_0, x_j) = \begin{cases} P(A_1, B(j)_1) \text{ if } x_j \geq cutoff \\ P(A_1, B(j)_0) \text{ if } x_j < cutoff \end{cases}$.

The pseudo-codes of the IBayes_Zinc model is given below:

Algorithm 1.1 IBayes_Zinc

Input: The training dataset D of protein sequences, the total number of samples N, the number of positive samples N_1, the number of negative samples N_2, and the class variable *cutoff*.

Output: Whether a site is zinc-binding

1 Begin

2 Initialize D;

3 Preprocess D;

4 Set up and run ZincExplorer;

5 Set up and run ZincFinder;

6 Set up and run ZincRed;

7 Predict x_1 with ZincExplorer;

8 Predict x_2 with ZincFinder;

9 Predict x_3 with ZincRed;

10 Identify the missing data (x_1, x_2, x_3); // Process the missing attribute values

11 Compute $P_{A1} = N_1/N$;

12 Compute $P_{A0} = N_2/N$;

13 Compute $n_1 = \sum_N \sum_{i=1}^3 x_i \geq cutoff$;

14 Compute $n_0 = \sum_N \sum_{i=1}^3 x_i < cutoff$;

15 Compute n_{11};

16 Compute n_{10};

17 Compute n_{01};

18 Compute n_{00};

19 For ($j=1; j\leq3; j++$)

20 $P_1 = P_{A1}$;

21 $P_0 = P_{A0}$;

22 Compute $P_{11\_j} = P(A_1, B(j)_1) = n_{11}/N$;

23 Compute $P_{01\_j} = P(A_1, B(j)_0) = n_{01}/N$;

24 Compute $P_{10\_j} = P(A_0, B(j)_1) = n_{10}/N$;

25 Compute $P_{00\_j} = P(A_0, B(j)_0) = n_{00}/N$;

22 If $x_j \geq cutoff$

23 $P_1 = P_1 * P_{11\_j}/P_{A1}$;

24 $P_0 = P_0 * P_{10\_j}/P_{A0}$;

25 Else

26 $P_1 = P_1 * P_{01\_j}/P_{A1}$;

27 $P_0 = P_0 * P_{00\_j}/P_{A0}$;

28 End if

29 End for

20 If $P_1 > P_0$

21 Output it is a binding site.

22 Else

23 Output it is a non-binding site.

24 End if

25 End

In Algorithm 1.1, the decision is made based on the probability for each sample. If $P_1 \geq P_0$, the sample is determined as positive; Otherwise, the sample is determined as negative.

3.3 Evaluation Criteria

The following indices were selected to evaluate the effect of the proposed model:

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \qquad (5)$$

where, recall is a quantity descriptor equaling the proportion of the number of correctly predicted positive samples to the number of positive samples; precision is a quality descriptor equaling the proportion of the number of correctly predicted positive samples to the number of samples predicted to be positive; $MCC \in [-1, 1]$ (Matthew correlation coefficient) is an indicator of the correlation between the data and a feature classification (if $MCC=1$, the prediction is completely correct; if $MCC=0$, the prediction is random; if $MCC=-1$, the prediction is completely contradictory).

4 RESULTS AND DISCUSSION

4.1 Performance Analysis

Based on the Zhao_dataset, the tools ZincExplorer, ZincFinder, ZincPred and IBayes_Zinc were tested and evaluated separately. In the whole interval of *cutoff*, the MCC and recall of IBayes_Zinc were higher than those of the other three methods. As for precision, IBayes_Zinc exceeded ZincExplorer and ZincPred, but stayed slightly below ZincFinder.

To further verify the precision of these tools, the sum of Precision and Recall was calculated. As shown in Table 1, IBayes_Zinc was still superior to the three contrastive methods.

Table 1 Results of four tools on Zhao_dataset

Index	MCC	Precision	Recall	Precision+Recall
IBayes_Zinc	0.7256	86.3%	67.9%	1.542
ZincExplorer	0.5269	72.9%	63.1%	1.361
ZincFinder	0.6811	91.1%	57.1%	1.482
ZincPred	0.6006	77.9%	59.8%	1.377

Thus, IBayes_Zinc outshined the other three methods in overall performance. When recall was 70% (the general level for actual recognition of zinc-binding sites), the precision of IBayes_Zinc was 2-15% higher than that of the other three methods. To sum up, IBayes_Zinc is more accurate than the other methods in predicting Zinc-binding sites.

4.2 Performance Analysis on an Independent Testing Dataset

The precision and robustness of the four tools were further compared on Final_dataset, an independent testing dataset. The results show that IBayes_Zinc outputted a higher MCC than that of the other three methods, across the interval of *cutoff*, except for 0.2-0.28. Since a small *cutoff* is not usually adopted to judge an instance, the exception does not affect the overall performance of these methods.

The mean MCC of IBayes_Zinc was 0.5369 in the whole interval, about 13%, 5%, and 9% higher than ZincExplorer, ZincFinder, and ZincPred, respectively. The performance index curves are shown in Fig. 2 below.
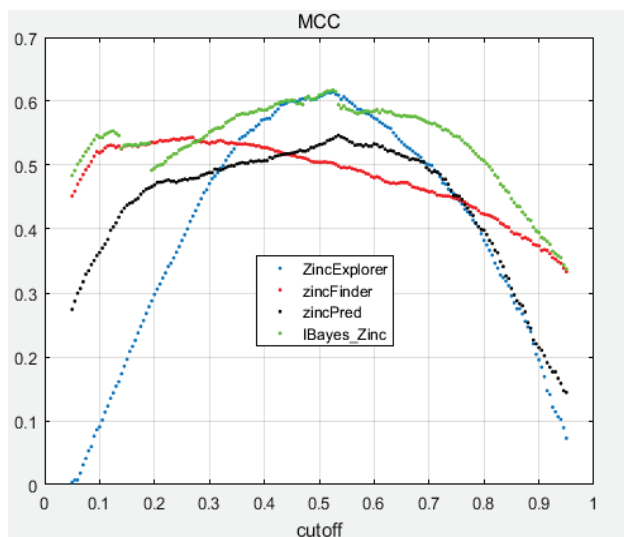
Figure 2 The MCC curves of the four tools on Final_Dataset

Table 2 Results of the four tools on Final_Dataset

Index	MCC	Precision	Recall	Precision+Recall
IBayes_Zinc	0.5369	75.67%	45.6%	1.213
ZincExplorer	0.4055	64.58%	49.95%	1.145
ZincFinder	0.483	81.7%	33.99%	1.157
ZincPred	0.442	67.28%	42.86%	1.101

Statistical analysis of the other performance indices (Tab. 2) shows that the mean precision of IBayes_Zinc reached 75.67%, nearly 11% and 8% higher than that of ZincExplorer and ZincPred, respectively, and about 6% less than that of ZincFinder. The mean recall of IBayes_Zinc was 45.6%, nearly 12% and 3% higher than that of ZincFinder and ZincPred, respectively, and 4% lower than that of ZincExplorer. IBayes_Zinc also outperformed the other methods in the sum of precision and recall. Overall, IBayes_Zinc is superior to the other three methods.

4.3 Case Study

Se-Met-Ampd derivative is a protein complex (RSCB:2Y28) with a crystal structure that contains three protein chains A, B, and C. In this paper, 2Y28 (Chain A), denoted as 2Y28_A, is selected as the visual instance to verify the predictive power of IBayes_Zinc.
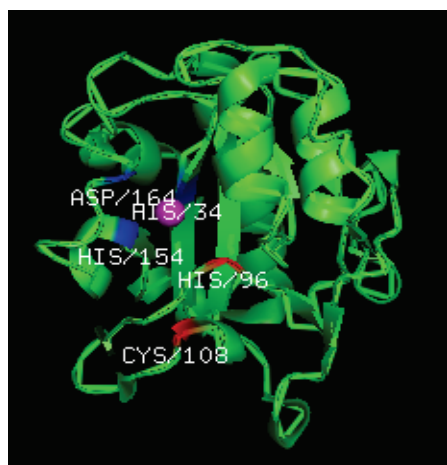

Figure 3 Visualization of the prediction results of the four tools on zinc-binding sites in 2Y28_A

Fig. 3 displays the visualization results of zinc-binding sites identified by four different tools, namely, IBayes_Zinc, ZincExplorer, ZincFinder and ZincPred. In 2Y28_A, there are three residues, HIS/34, HIS/154, and ASP/164 (marked in blue) binding to zinc ions (marked as magnetic balls). IBayes_Zinc correctly identified all of the three zinc-binding sites. ZincExplore also predicted the three sites, but falsely predicted a positive site CYS/108 (marked in red). ZincFinder only identified two sites, HIS/154 and ASP/164, failing to find HIS/34. ZincPred predicted the three sites, but falsely predicted the positive site HIS/96 (marked in red).

Tab. 3 compares the prediction results of the four tools on zinc-binding sites in 2Y28_A.

Table 3 Prediction results of the four tools on zinc-binding sites in 2Y28_A

Index	IBayes_Zinc	ZincExplorer	ZincFinder	ZincPred
HIS/34	Y	Y	Y	N
HIS/154	Y	Y	N	Y
ASP/164	Y	Y	Y	N
False Positive Sites	None	CYS/108	None	HIS/96

Note: Y means the residue is zinc-binding; N means the residue is non-zinc-binding; None means the absence of falsely predicted positive site.

5 CONCLUSION

The proteins binding to zinc ions have important biological functions, making it critical to identify the zinc-binding sites in proteins. Considering the availability of protein sequence information, this paper develops a robust and precise prediction model (IBayes_Zinc) for zinc-binding sites based on the Bayesian method, which can effectively handle incomplete or missing data. Specifically, the prediction scores of three existing sequence-based prediction tools were adopted, and the missing values were padded, forming an integrated classification tool. Then, the probabilities of positive and negative samples were computed, and categorized as the class with higher probabilities. Experiments were conducted on a non-redundant training dataset and an independent testing dataset. The results show that our method surpassed the other three methods by nearly 5%–13% in MCC and outperformed the latter in recall and precision.

6 REFERENCES

[1] Eisenberg, D., Marcotte, E. M., Xenarios, I., & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, 405(6788), 823-826. https://doi.org/10.1038/35015694

[2] Ciechanover, A. (1998). The ubiquitin-proteasome pathway: on protein death and cell life. *Embo J*, 17(24), 7151-7160. https://doi.org/10.1093/emboj/17.24.7151

[3] Xu, D., Jalal, S. I., Sledge, G. W., & Meroueh, S. O. (2016). Small-molecule binding sites to explore protein-protein interactions in the cancer proteome. *MolBiosyst*, 12(10), 3067-3087. https://doi.org/10.1039/C6MB00231E

[4] Kenyon, J. C., Prestwood, L. J., &Lever, A. M. (2015). A novel combined RNA-protein interaction analysis distinguishes HIV-1 Gag protein binding sites from structural change in the viral RNA leader. *Sci Rep*, 5(1), 14369. https://doi.org/10.1038/srep14369

[5] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer-Jr, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem*, 80(2), 319-324. https://doi.org/10.1111/j.1432-1033.1977.tb11885.x

[6] Chasapis, C. T., Loutsidou, A. C., Spiliopoulou, C. A., Stefanidou, M. E. (2012). Zinc and human health: an update. *Arch Toxicol*, 86(4), 521-534. https://doi.org/10.1007/s00204-011-0775-1

[7] Bettmer, J. (2005). Metalloproteomics: a challenge for analytical chemists. *Anal BioanalChem*, 383(3), 370-371. https://doi.org/10.1007/s00216-005-3404-0

[8] Sanz-Medel, A. (2005). From metalloproteomics to heteroatom-tagged proteomics. *Anal BioanalChem*, 381(1), 1-2. https://doi.org/10.1007/s00216-004-2907-4

[9] Lobinski, R. & Moulin, C. (2006). Ortega R. Imaging and speciation of trace elements in biological environment. *Biochimie*, 88(11), 1591-1604. https://doi.org/10.1016/j.biochi.2006.10.003

[10] Guo, M., Li, H. D., Shao, S., Lu, X. W., & Fan, W. X. (2015).Study on the interaction of zinc ion binding with human serum albumin using isothermal titration calorimetry. *Journal Chemical Society of Pakistan*, 37(3), 452-457.

[11] Passerini, A., Punta, M., Ceroni, A., Rost, B., & Frasconi, P. (2006). Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins-Structure Function and Bioinformatics*, 65(2), 305-316. https://doi.org/10.1002/prot.21135

[12] Passerini, A., Andreini, C., Menchetti, S., Rosato, A., & Frasconi, P. (2007). Predicting zinc binding at the proteome level. *Bmc Bioinformatics*, 8(1), 39. https://doi.org/10.1186/1471-2105-8-39

[13] Shu, N. J., Zhou, T. P., & Hovmoller, S. (2008). Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics*, 24(6), 775-782. https://doi.org/10.1093/bioinformatics/btm618

[14] Chen, Z., Wang, Y. Y., Zhai, Y. F., Song, J., & Zhang, Z. (2013). ZincExplorer: an accurate hybrid method to improve the prediction of zinc-binding sites from protein sequences. *Molecular Biosystems,* 9(9), 2213-2222. https://doi.org/10.1039/c3mb70100j

[15] Zhao, W., Xu, M., Liang, Z., Ding, B., Niu, L. W., Liu, H. Y., & Teng, M. K. (2011).Structure-based de novo prediction of zinc-binding sites in proteins of unknown function. *Bioinformatics*, 27(9), 1262-1268. https://doi.org/10.1093/bioinformatics/btr133

[16] He, W., Liang, Z., Teng, M. K., Niu, L. W. (2015). mFASD: a structure-based algorithm for discriminating different types of metal-binding sites. *Bioinformatics*, 31(12), 1938-1944. https://doi.org/10.1093/bioinformatics/btv044

[17] Liu, Z., Wang, Y., Zhou, C. H., Xue, Y., Zhao, W., & Liu, H. Y. (2014). Computationally characterizing and comprehensive analysis of zinc-binding sites in proteins. *BiochimBiophysActa*, 1844(1), 171-180. https://doi.org/10.1016/j.bbapap.2013.03.001

[18] Ebert, J. C. & Altman, R. B. (2008). Robust recognition of zinc binding sites in proteins. *Protein Sci*, 17(1), 54-65. https://doi.org/10.1110/ps.073138508

[19] Babor, M., Gerzon, S., Raveh, B., Sobolev, V., & Edelman, M. (2008). Prediction of transition metal-binding sites from apo protein structures. *Proteins*, 70(1), 208-217. https://doi.org/10.1002/prot.21587

[20] Goyal, K. & Mande, S. C. (2008). Exploiting 3D structural templates for detection of metal-binding sites in protein structures. *Proteins*, 70(4), 1206-1218. https://doi.org/10.1002/prot.21601

[21] Bordner, A. J. (2008). Predicting small ligand binding sites in proteins using backbone structure. *Bioinformatics*, 24(24), 2865-2871. https://doi.org/10.1093/bioinformatics/btn543

[22] Zheng, C., Wang, M., Takemoto, K., Akutsu, T., Zhang, Z. D., & Song, J. N. (2012).An integrative computational framework based on a two-step random forest algorithm improves prediction of zinc-binding sites in proteins. *PLoS One*, 7(11), e49716. https://doi.org/10.1371/journal.pone.0049716

[23] Lu, C. H., Lin, Y. F., Lin, J. J., & Yu, C. S. (2012). Prediction of Metal Ion-Binding Sites in Proteins Using the Fragment Transformation Method. *Plos One*, 7(6), e39252. https://doi.org/10.1371/journal.pone.0039252

[24] Li, H., Pi, D. C., Wu, Y. L., & Chen, C. M. (2017).Integrative Method Based on Linear Regression for the Prediction of Zinc-Binding Sites in Proteins. *IEEE Access*, 5, 14647-14656. https://doi.org/10.1109/ACCESS.2017.2731872

[25] Li, H., Pi, D. C., Zhang, L. H., & Hong, L. (2018). Prediction of Zinc-Binding Sites in Proteins Based on Bayesian Method. *Fresenius Environmental Bulletin*, 27(1), 430-435.

Contact information:

Hui LI, Associate Professor
(Corresponding author)
(1) School of Software Engineering,
Jinling Institute of Technology, Jiangsu 211169, China
(2) Center for Intelligent Computer Human Interaction,
Nanjing Institute of Big Data, Jiangsu, 211169 China
(3) College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Jiangsu, 211169 China
E-mail: alicehh_ok@126.com

Dechang PI, Professor
College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Jiangsu, 211169 China
E-mail: nuaacs@126.com

Chuanming CHEN, Associate Professor
College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Jiangsu, 211169 China
E-mail: ccm_0@163.com