G&I Genomics & Informatics

eISSN 2234-0742
Genomics Informs 2017;15(4):162-169
https://doi.org/10.5808/GI.2017.15.4.162

ORIGINAL ARTICLE

Prediction of Metal Ion Binding Sites in Proteins from Amino Acid Sequences by Using Simplified Amino Acid Alphabets and Random Forest Model

Suresh Kumar*

Department of Diagnostic and Allied Health Sciences, Faculty of Health and Life Sciences, Management and Science University, 40100 Shah Alam, Malaysia

Metal binding proteins or metallo-proteins are important for the stability of the protein and also serve as co-factors in various functions like controlling metabolism, regulating signal transport, and metal homeostasis. In structural genomics, prediction of metal binding proteins help in the selection of suitable growth medium for overexpression's studies and also help in obtaining the functional protein. Computational prediction using machine learning approach has been widely used in various fields of bioinformatics based on the fact all the information contains in amino acid sequence. In this study, random forest machine learning prediction systems were deployed with simplified amino acid for prediction of individual major metal ion binding sites like copper, calcium, cobalt, iron, magnesium, manganese, nickel, and zinc.

Keywords: amino acid sequence, binding sites, machine learning, proteins

Introduction

Amino acids play a central role in the building block of protein. The primary structure of the protein is determined by the arrangement of 20 naturally occurring amino acids. The function of a protein is determined from their amino acids and also they depend upon interaction with cofactors, binding with metal ions and interaction with other proteins. The proteome of all the organism share significant metal ions and metal binding cofactors to carry out its essential function. It has been estimated that approximately 30% of all proteins contain at least one metal. The proteins play a vital role in biological processes and in the stability of the protein by binding with metal ions or metal containing-cofactors [1]. The proteins bind with major metal ions like transition metals, alkali, and alkaline metals. The frequent metal ions that bind with proteins are sodium, copper, iron, magnesium, manganese, potassium, and zinc ions respectively. In *in-vitro* condition, the unfolded polypeptide may are observed to interact with metal ions that direct the polypeptide folding process [2]. Identification of metal binding through experimental procedures like the use of metal ion affinity column chromatography [3, 4], electrophoretic mobility shift assay [5, 6], absorbance spectroscopy [7], gel electrophoresis [8], nuclear magnetic resonance spectroscopy [9-11], and mass spectrometry [3, 12] require tedious steps and specific instruments, making them expensive and may be unsuitable for unknown targets. In this aspect, there is a need for computational predictors of protein binding metal ion in order to reduce time and cost. For example, predictions of protein metal binding ions are useful in structural genomics, to select proper growth medium for overexpression studies and for the easy interpretation of electron density maps. But fortunately, metal-binding ability are encoded in the amino acidic sequences and these primary sequences help in protein structure formation. Through genomic projects various organism genomic sequences have been annotated somehow along with metalloproteins contained in them [1]. Bioinformatics has been extensively used to predict metal-binding ability from amino acid sequences. Various computational methods like artificial neural networks [13], support vector machines [14], decision tree algorithm [15],

Received October 16, 2017; Revised November 16, 2017; Accepted November 16, 2017

***Corresponding author:** Tel: +60-14-2734893, Fax: +60-35-5112848, E-mail: sureshkumar@msu.edu.my

graph theory [16], FoldX force field [17], CHED [18, 19], and geometry algorithm methods [16]. These methods depend upon either only sequence information or the use of both sequence and structure information. However, most of the available prediction methods are either based on the knowledge of the apoprotein structure or restricted to few specific cases, like the metal binding of histidines/cysteines. Most of these methods have been implemented as standalone software or web servers to the research community [15, 20].

Due to the availability of cheap and advancement of sequencing instruments, the sequence of proteins has increased rapidly over when compared to protein structure data. This due to the fact that experimental determining the three-dimensional of protein is difficult and expensive. Through various theoretical and experimental studies, it is proved that minimal set of the amino acid is sufficient for protein folding [21]. The minimal set of representative residues with similar features can be achieved by grouping together the 20 amino acids by clustering. This method is called as reduced or simplified amino acid alphabet. Several simplified amino acid alphabets have been proposed, which have been applied to pattern recognition method in the prediction of protein structure [22], for generation of consensus sequences from multiple alignments, and for protein folding prediction [23]. Various computational predictor has used simplified amino acids to predict the solubility on overexpression, remote homology detection [19], and identify the defensin peptide family [24], effects of cofactors on conformation transition [25], DNA-binding proteins [26], heat shock protein families [27], inter-residue interaction [28], protein adaptation to mutation [29], and protein disorder [30]. In the present study, a random forest algorithm has been deployed to predict metal ion binding protein based on the simplified amino acids proposed by Murphy *et al*. [21].

Methods

Dataset construction

All the protein sequences were downloaded from the UniProt database [31] available at http://www.uniprot.org/. The downloaded sequences, annotated as metal containing, were grouped into eight subsets. Each of the subsets, containing one of the metal species viz., calcium, cobalt, copper, iron, magnesium, manganese, nickel, and zinc was considered to be metal-containing while all other entries were considered to be metal-free. Redundancy among the amino acid sequences was removed by clustering analysis using the cd-hit program [32] with the threshold of 50% level of percentage of identity, analogous by the UniRef 50 list [33] available in the UniProt database.

This resulted in eight data sets containing 186 calcium-containing proteins, 69 cobalt-containing proteins, 215 copper-containing proteins, 315 iron-containing proteins, 961 magnesium-containing proteins, 386 manganese-containing proteins, 74 nickel-containing proteins, and 1,716 zinc-containing proteins. All proteins containing calcium, cobalt, copper, magnesium, manganese, nickel, or zinc were then subtracted from the UniRef50 list, resulting in a collection of non-metalloproteins. The workflow of dataset construction is shown in Fig. 1. The problem of the imbalanced dataset can be solved as proposed by Cohen *et al*. [34]. Firstly, they pre-processes the data to re-establish class balance (either by upsizing the minority class or downsizing the majority class). Secondly, they modify the learning algorithm itself to copy with imbalanced data. In this study, we pre-processed the data which contains a balanced set of metal and non-metal ions. For this construction, non–metallo-proteins datasets sequences were randomly selected in order to have balanced set of metal and non-metal binding proteins for each metal ion, respectively.

Feature extraction by simplified amino acid alphabets

In order to investigate the effect of a particular class of amino acids on metal ion binding, the 20 amino acids were grouped into various classes based on certain common properties and the composition of the reduced sets of amino acids was considered. Feature extraction is done using the simplified amino acid alphabet. It estimates that reduced alphabets containing 10–12 letters can be used to design foldable sequences for a large number of protein families. This estimate is based on the observation that there is little loss of the information necessary to pick out structural homologs in a clustered protein sequence database when a suitable reduction of the amino acid alphabet from 20 to 10 letters is made.

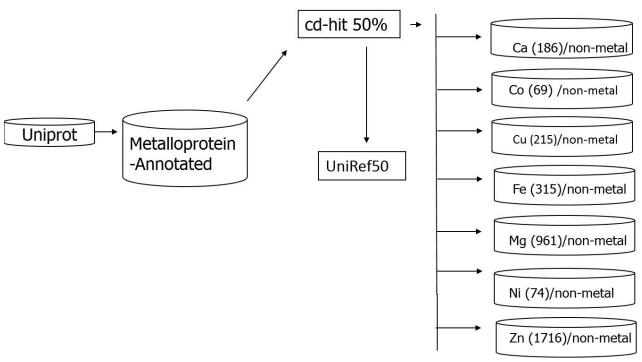

Fig. 1. Construction of dataset used for prediction.

A simplified amino acid alphabet of 18 characters was used (Table 1). It is based on three independent amino acid classifications.

Conformational similarity

Conformational similarity indices are proposed by Chakrabarti and Pal [28] based on different residues are computed using the distribution of the main-chain and side-chain torsion angles and values have been used to cluster amino acids in proteins. In this method, the conformational similarity of the 20 amino acids based on torsion angles, which contains seven clusters ([CMQLEKRA], [P], [ND], [G], [HWFY], [S], and [TIV]) are used to represent variables.

BLOSUM 50 substitution matrix

The BLOSUM-50 matrix is proposed by Cannata *et al*. [35]. The matrix is deduced from amino acid pair frequencies in aligned blocks of a protein sequence database and is widely used for sequence alignment and comparison. The BLOSUM 50 matrix that they group together on the basis of the possibility of foldable structures and consists of the clusters: [P], [KR], [EDNQ], [ST], [AG], [H], [CILMV], and [YWF].

Hydrophobicity

The hydrophobicity scale by Rose *et al*. [36] is correlated to the average area of buried amino acids in globular proteins. This results in a scale which is not showing the helices of a protein, but rather the surface accessibility. It is based on the hydrophobicity scale which consists of the following cluster: [CFILMVW], [AG], [PH], [EDRK], and [NQSTY].

Table 1. The 18 variables, obtained by merging three simplified alphabets of amino acid residues used to represent protein sequences

Variable	Residues
V1	CMQLEKRA
V2	P
V3	ND
V4	G
V5	HWFY
V6	S
V7	TIV
V8	CFILMVW
V9	AG
V10	PH
V11	EDRK
V12	NQSTY
V13	FWY
V14	CILMV
V15	H
V16	ST
V17	EDNQ
V18	KR

Random forest predictions

Random forest is a classification algorithm [37] that uses an ensemble of tree-structured classifiers. The random forest is a popular algorithm that has been used in designing computational predictors for various biological problems. Random forest is an ensemble learning method for classification. The random forest classifies a new object with an input vector, the input vector is predicted by each decision tree in the forest. Each tree provides a classification with votes and the class with most votes will be output as the predicted class. It is implemented by using Weka package [38, 39]. To ensure that parameter estimation and model generation of random forest is completely independent of the test data, a nested cross-validation procedure is performed. Nested cross-validation [40] means that there is an outer cross-validation loop for model assessment and an inner loop for model selection. In this study, the original samples are randomly divided into $k = 10$ parts in the outer loop. Each of these parts is chosen one by one for assessment, and the remaining nine of 10 samples are for model selection in the inner loop where a type of cross-validation using the so-called out-of-bag samples is performed.

Measurement of classifier's performance

When the predictor was focused on the problem of distinguishing proteins containing a certain type of metal ion from proteins that do not contain any type of metal, it is important that both sets contain the same number of proteins; otherwise, several figures of merit that are commonly used to monitor the prediction reliability would be seriously biased. The reliability of the predictions was monitored with the following quantities. If a protein of type 1 must be distinguished from a protein of type 2, a prediction was considered to be a true-positive if type 1 was correctly predicted; it was considered to be a true-negative if type 2 was correctly predicted; it was considered to be a false-negative if a type 1 protein was predicted to be a type 2 protein; and it was considered to be a false-positive if a type 2 protein was predicted to be a type 1 protein. Consequently, the following figures of merit, the sensitivity, the specificity, the accuracy, the Mathews correlation are computed [41] as shown in the Eq. (1) below.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN},$$

$$Sn = \frac{TP}{TP + FN},$$

$$Sp = \frac{TN}{TN + FP}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \quad (1)$$

Results and Discussion

By using a simplified amino acid alphabet based on three independent amino acid classifications, amino acid cluster variables were obtained. Conformational similarity contains seven clusters: [CMQLEKRA], [P], [ND], [G], [HWFY], [S], and [TIV]. BLOSUM 50 substitution matrix contain [P], [KR], [EDNQ], [ST], [AG], [H], [CILMV], and [YWF]. The hydrophobicity scale contains [CFILMVW], [AG], [PH], [EDRK], and [NQSTY]. Out of 20 amino acid clusters, cluster [P] and [AG] which are present in more than one simplified alphabet were considered only once and these results in 18 variables (Table 1). The 18 variables are represented with percentage of occurrence as follows.

Table 2. Overall prediction performance of the classifier in predicting individual metal ion binding sites

Metal	Sensitivity	Specificity	Mathews correlation	Accuracy
Ca	0.769	0.739	0.507	0.754
Co	0.884	0.823	0.708	0.853
Cu	0.746	0.815	0.563	0.781
Fe	0.772	0.740	0.512	0.756
Mg	0.766	0.714	0.481	0.740
Mn	0.729	0.647	0.378	0.688
Ni	0.945	0.869	0.817	0.907
Zn	0.740	0.640	0.382	0.690

$$pc_{aa,i} = 100\frac{n_{aa,i}}{nres_i}, \quad (2)$$

The percentage of occurrence $pc_{aa,i}$ of the amino acid *aa* in the i_{th} protein was computed for each of the 20 types of amino acids in each protein as per Eq. (2). The protein sequences represented by the amino acid percentage of occurrence using 18 variables were employed with random forest algorithm using Weka suite. The metallo-proteins were identified using all the 18 variables with high accuracy ranging from 69% for zinc and 90% for nickel (Table 2). Moreover, prediction performance was studied by feature selection method by removing one variable at a time and maintaining the highest value in performance indices. Measurements are removed until there is an unacceptable degradation in system performance. The use of feature selection method will eliminate alphabets which are irrelevant or redundant features, and thus it improves the accuracy of the learning algorithm. To select an optimal subset of variables, we first analyzed how individual attributes from the initial set of 18 variables, contributed to predictive accuracy. For feature selection, we employed the wrapper approach as it uses the learning algorithm to test all existing feature subsets. The wrapper method will use a subset of features to train the model. Based on the inferences, the feature can be added or removed to improve the accuracy of the learning algorithm. We used a backward feature elimination, by starting with the full set and deleting attributes one at a time for searching the feature space [42, 43].

The specific steps of the wrapper approach followed in this study.

(1) Partitioning the data with 10-fold cross-validation (k = 10).

(2) On each cross-validation training set, the learning machine was trained by using all 18 variables, to produce a ranking of the variables according to the importance. The cross-validation test set predictions were recorded.

(3) Then the variables are removed which are least

Table 3. Feature selection of variables in improving the performance of copper ion prediction against proteins that lack metal ions

Variable removed	Average sensitivity	Average specificity	Average accuracy	Average Mathews correlation
None	0.746	0.815	0.781	0.563
AG	0.762	0.809	0.786	0.571
CMQLEKRA	0.794	0.804	0.799	0.599
NQSTY	0.779	0.814	0.796	0.593
EDNQ	0.796	0.797	0.796	0.592
CFILMVW	0.785	0.803	0.794	0.588
TIV	0.785	0.798	0.792	0.583
PH	0.774	0.801	0.788	0.576

important one by one and another learning machine was trained based on remaining variables, the cross-validation test set predictions were once again recorded. This step is repeated by removing each variable until at small number remain.

(4) Aggregate the predictions from all 10 cross-validation test sets and compute the aggregate accuracy at each step down in a number of variables.

By the following the above steps, feature selection of variables was done by wrapper approach employing random forest machine learning algorithm. Based on aggregate accuracy, the important variables for copper ion prediction are PH variable and least preferred variables are AG and CMQLEKRA (Table 3). Based on Table 3, it is understood that removing PH variable decrease the accuracy of the classifier whereas removing AG and CMQLEKRA improves the accuracy of the classifier. For calcium ion prediction, the least important variable is P and EDNQ; removing these variable improves the performance of the classifier (Table 4). Similarly, for cobalt ion prediction, the variable CILMV is the least preferred variable as it affects the performance of the classifier (Table 5). For iron ion prediction, removing variable CFILMVW improves the performance of the classifier (Table 6). For magnesium, ion prediction variable ST and ND are least preferred variables (Table 7). For manganese ion prediction, removing variable FWY improves

Table 4. Feature selection of variables in improving the performance of calcium ion prediction against proteins that lack metal ions

Variable removed	Average sensitivity	Average specificity	Average accuracy	Average Mathews correlation
None	0.769	0.738	0.754	0.507
P	0.783	0.758	0.770	0.541
EDNQ	0.788	0.751	0.770	0.541
EDRK	0.796	0.758	0.777	0.554
PH	0.785	0.756	0.770	0.541
CILMV	0.801	0.754	0.777	0.556
AG	0.790	0.749	0.770	0.539
CFILMVW	0.789	0.765	0.777	0.554
NQSTY	0.785	0.767	0.776	0.552
CMQLEKRA	0.780	0.765	0.772	0.545

Table 5. Feature selection of variables in improving the performance of cobalt ion prediction against proteins that lack metal ions

Variable removed	Average sensitivity	Average specificity	Average accuracy	Average Mathews correlation
None	0.884	0.823	0.853	0.708
CILMV	0.903	0.842	0.872	0.747
CFILMVW	0.899	0.837	0.868	0.737
ND	0.894	0.828	0.861	0.724
EDNQ	0.884	0.833	0.858	0.717
PH	0.894	0.847	0.870	0.741
ST	0.903	0.837	0.870	0.742
NQSTY	0.860	0.833	0.846	0.693

Table 6. Feature selection of variables in improving the performance of iron ion prediction against proteins that lack metal ions

Variable removed	Average sensitivity	Average specificity	Average accuracy	Average Mathews correlation
None	0.772	0.740	0.756	0.512
NQSTY	0.778	0.731	0.754	0.509
S	0.786	0.727	0.757	0.514
PH	0.786	0.724	0.755	0.511
CMQLEKRA	0.785	0.720	0.753	0.507
CFILMVW	0.787	0.734	0.761	0.523
AG	0.790	0.720	0.755	0.511
TIV	0.780	0.725	0.753	0.507
HWFY	0.790	0.735	0.762	0.525

Table 7. Feature selection of variables in improving the performance of magnesium ion prediction against proteins that lack metal ions

Variable removed	Average sensitivity	Average specificity	Average accuracy	Average Mathews correlation
None	0.766	0.714	0.740	0.481
ST	0.779	0.714	0.746	0.494
ND	0.774	0.720	0.747	0.494
NQSTY	0.767	0.717	0.742	0.485
S	0.772	0.711	0.742	0.484
HWFY	0.770	0.716	0.743	0.487
PH	0.777	0.709	0.743	0.487
CMQLEKRA	0.775	0.708	0.741	0.484

Table 8. Feature selection of variables in improving the performance of manganese ion prediction against proteins that lack metal ions

Variable removed	Average sensitivity	Average specificity	Average accuracy	Average Mathews correlation
None	0.729	0.647	0.688	0.378
FWY	0.731	0.717	0.734	0.474
EDNQ	0.741	0.656	0.698	0.398
CMQLEKRA	0.750	0.647	0.698	0.399
AG	0.750	0.643	0.697	0.396
S	0.739	0.660	0.700	0.400

Table 9. Feature selection of variables in improving the performance of nickel ion prediction against proteins that lack metal ions

Variable removed	Average sensitivity	Average specificity	Average accuracy	Average Mathews correlation
None	0.945	0.869	0.907	0.817
EDRK	0.950	0.887	0.918	0.838
G	0.931	0.892	0.917	0.824
NQSTY	0.923	0.887	0.905	0.810
ST	0.941	0.878	0.909	0.821
EDNQ	0.936	0.865	0.900	0.803
FWY	0.918	0.860	0.889	0.780
HWFY	0.931	0.865	0.898	0.800
TIV	0.927	0.869	0.898	0.797

Table 10. Feature selection of variables in improving the performance of zinc metal ion prediction against proteins that lack metal ions

Variable removed	Average sensitivity	Average specificity	Average accuracy	Average Mathews correlation
None	0.740	0.640	0.690	0.382
HWFY	0.751	0.638	0.695	0.391
CMQLEKRA	0.750	0.636	0.692	0.386
AG	0.747	0.638	0.693	0.388
ST	0.743	0.644	0.693	0.389
EDNQ	0.743	0.636	0.689	0.381

the accuracy of the classifier (Table 8). For nickel ion prediction, variable EDRK is the least preferred one (Table 9). For zinc ion prediction, the least preferred variable is HWFY (Table 10).

For example, cobalt metal binding protein can be discriminated from non-metal ions with all 18 variables with the accuracy of 85% (Fig. 2). It can be seen that, on removing variable V14 (CILMV) from the subset, the accuracy of the predictor improves from 85% to 87%. After removing of variables V8 (CFILMVW), V3 (ND), V17 (EDNQ), V10

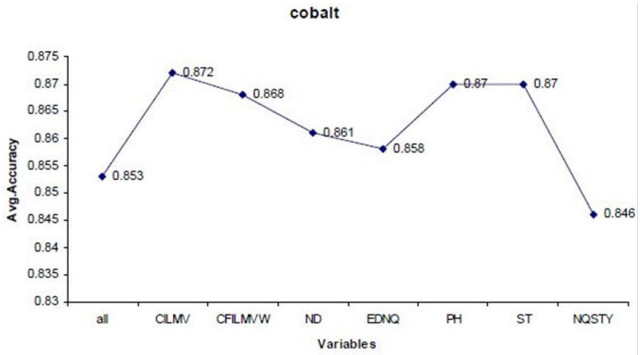

Fig. 2. The performance graph of the Random forest classifier using feature selection (10-fold cross validation for cobalt ion prediction).

(PH), and V16 (ST), the accuracy values are in the range from 86% to 87%. There is a drastic decrease in accuracy of the classifier by removing the variable V12 (NQSTY) to 84%. No further reduction of the set was possible, as the performance of random forest classifier dropped if any further attributes were eliminated. It can be seen that accuracy of prediction of metal binding proteins can be improved (e.g., calcium from 74% to 77%, cobalt from 83% to 85%, and nickel from 69% to 77%) by elimination of certain noisy features, up to certain limit and further improvement is then impossible. According to this backward strategy of feature selection, it can be observed that the prediction performance can be slightly improved. Some common variables rejected are V14 (CILMV) in calcium and cobalt, V8 (CFILMVW) in copper and iron.

In this work, a new random forest based approach is developed combining hybrid feature of simplified amino acid alphabets for prediction of metal ion binding sites of iron, copper manganese, magnesium, nickel, calcium, cobalt, and zinc from amino acid sequence data. The result indicates that the random forest model has a high prediction accuracy in predicting metal ion binding sites. These metal binding prediction methods are helpful to avoid the selection of 'impossible' targets in structural biology and proteomics.

ORCID: Suresh Kumar: http://orcid.org/0000-0001-5682-0938

Acknowledgments

The author acknowledges Department of Diagnostic and Allied Health Sciences, Faculty of Health and Life Sciences, Management & Science University, Shah Alam, Selangor Darul Ehsan, Malaysia for providing necessary infrastructure facility to carry out this research.

References

1. Andreini C, Bertini I, Rosato A. A hint to search for metalloproteins in gene banks. *Bioinformatics* 2004;20:1373-1380.
2. Clapp LA, Siddons CJ, Whitehead JR, VanDerveer DG, Rogers RD, Griffin ST, *et al*. Factors controlling metal-ion selectivity in the binding sites of calcium-binding proteins: the metal-binding properties of amide donors. A crystallographic and thermodynamic study. *Inorg Chem* 2005;44:8495-8502.
3. Kaur-Atwal G, Weston DJ, Green PS, Crosland S, Bonner PL, Creaser CS. On-line capillary column immobilised metal affinity chromatography/electrospray ionisation mass spectrometry for the selective analysis of histidine-containing peptides. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007; 857:240-245.
4. Feng S, Pan C, Jiang X, Xu S, Zhou H, Ye M, *et al*. Fe3+ immobilized metal affinity chromatography with silica monolithic capillary column for phosphoproteome analysis. *Proteomics* 2007;7:351-360.
5. Osborn MT, Herrin K, Buzen FG, Hurlburt BK, Chambers TC. Electrophoretic mobility shift assay coupled with immunoblotting for the identification of DNA-binding proteins. *Biotechniques* 1999;27:887-890, 892.
6. Smith MF Jr, Delbary-Gossart S. Electrophoretic mobility shift assay (EMSA). *Methods Mol Med* 2001;50:249-257.
7. Korshin G, Chow CW, Fabris R, Drikas M. Absorbance spectroscopy-based examination of effects of coagulation on the reactivity of fractions of natural organic matter with varying apparent molecular weights. *Water Res* 2009;43:1541-1548.
8. Nigg PE, Pavlovic J. Characterization of multi-subunit protein complexes of human MxA using non-denaturing polyacrylamide gel-electrophoresis. *J Vis Exp* 2016;(116):e54683.
9. Jensen MR, Petersen G, Lauritzen C, Pedersen J, Led JJ. Metal binding sites in proteins: identification and characterization by paramagnetic NMR relaxation. *Biochemistry* 2005;44: 11014-11023.
10. Rondeau P, Sers S, Jhurry D, Cadet F. Sugar interaction with metals in aqueous solution: indirect determination from infrared and direct determination from nuclear magnetic resonance spectroscopy. *Appl Spectrosc* 2003;57:466-472.
11. Zhu D, Herbert BE, Schlautman MA, Carraway ER. Characterization of cation-pi interactions in aqueous solution using deuterium nuclear magnetic resonance spectroscopy. *J Environ Qual* 2004;33:276-284.
12. Butler M, Cabrera GM. A mass spectrometry-based method for differentiation of positional isomers of monosubstituted pyrazine N-oxides using metal ion complexes. *J Mass Spectrom* 2015;50:136-144.
13. Lin CT, Lin KL, Yang CH, Chung IF, Huang CD, Yang YS. Protein metal binding residue prediction based on neural networks. *Int J Neural Syst* 2005;15:71-84.
14. Passerini A, Punta M, Ceroni A, Rost B, Frasconi P. Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins* 2006;65:305-316.
15. Lippi M, Passerini A, Punta M, Rost B, Frasconi P. MetalDetector: a web server for predicting metal-binding sites and

disulfide bridges in proteins from sequence. *Bioinformatics* 2008;24:2094-2095.
16. Deng H, Chen G, Yang W, Yang JJ. Predicting calcium-binding sites in proteins: a graph theory and geometry approach. *Proteins* 2006;64:34-42.
17. Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc Natl Acad Sci U S A* 2005;102:10147-10152.
18. Chen Z, Wang Y, Zhai YF, Song J, Zhang Z. ZincExplorer: an accurate hybrid method to improve the prediction of zinc-binding sites from protein sequences. *Mol Biosyst* 2013;9: 2213-2222.
19. Levy R, Edelman M, Sobolev V. Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates. *Proteins* 2009;76:365-374.
20. Passerini A, Lippi M, Frasconi P. MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence. *Nucleic Acids Res* 2011;39:W288-W292.
21. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng* 2000;13:149-152.
22. Parisi G, Echave J. Structural constraints and emergence of sequence patterns in protein evolution. *Mol Biol Evol* 2001;18: 750-756.
23. Tainer JA, Roberts VA, Getzoff ED. Metal-binding sites in proteins. *Curr Opin Biotechnol* 1991;2:582-591.
24. Zuo Y, Lv Y, Wei Z, Yang L, Li G, Fan G. iDPF-PseRAAAC: a web-server for identifying the defensin peptide family and subfamily using pseudo reduced amino acid alphabet composition. *PLoS One* 2015;10:e0145541.
25. Lu MF, Xie Y, Zhang YJ, Xing XY. Effects of cofactors on conformation transition of random peptides consisting of a reduced amino acid alphabet. *Protein Pept Lett* 2015;22:579-585.
26. Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, *et al*. iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PLoS One* 2014;9:e106691.
27. Feng PM, Chen W, Lin H, Chou KC. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal Biochem* 2013;442:118-125.
28. Chakrabarti P, Pal D. The interrelationships of side-chain and main-chain conformations in proteins. *Prog Biophys Mol Biol* 2001;76:1-102.
29. Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. *Eur Biophys J* 2007; 36:1059-1069.
30. Weathers EA, Paulaitis ME, Woolf TB, Hoh JH. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett* 2004;576:348-352.
31. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;45:D158-D169.
32. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22:1658-1659.
33. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH; UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;31:926-932.
34. Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. *Artif Intell Med* 2006;37:7-18.
35. Cannata N, Toppo S, Romualdi C, Valle G. Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. *Bioinformatics* 2002;18:1102-1108.
36. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. *Science* 1985;229:834-838.
37. Zheng C, Wang M, Takemoto K, Akutsu T, Zhang Z, Song J. An integrative computational framework based on a two-step random forest algorithm improves prediction of zinc-binding sites in proteins. *PLoS One* 2012;7:e49716.
38. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. *Bioinformatics* 2004;20:2479-2481.
39. Smith TC, Frank E. Introducing machine learning concepts with WEKA. *Methods Mol Biol* 2016;1418:353-378.
40. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006;7:91.
41. Sahiner B, Chan HP, Hadjiiski L. Classifier performance prediction for computer-aided diagnosis using a limited dataset. *Med Phys* 2008;35:1559-1570.
42. Liu H, Jiang H, Zheng R. The hybrid feature selection algorithm based on maximum minimum backward selection search strategy for liver tissue pathological image classification. *Comput Math Methods Med* 2016;2016:7369137.
43. Mandal M, Mukhopadhyay A, Maulik U. Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC. *Med Biol Eng Comput* 2015;53:331-344.