Prediction of Protein–Metal Ion-Binding Sites Using Sequence Homology and Machine-Learning Methods

Zihan Tian [1], Cao Wei [1], Yutaka Moriwaki [1], Tohru Terada [1], Shugo Nakamura [1], Kazuya Sumikoshi [1], Fang Chun [1], and Kentaro Shimizu [1]

Abstract

Metal ions are essential for metalloproteins to perform their catalytic or structural functions. To understand their role in protein function, it is important to identify metal ion-binding sites. Because experimental identification is labor-intensive and time-consuming, computational methods are expected to be used in the prediction of protein–metal ion-binding sites. A range of computational methods have been proposed to predict metal ion-binding sites from protein sequences. In this study, we implemented two methods of predicting metal ion-binding sites for Ca^{2+}, Co^{2+}, Cu^{2+}, Cu^{+}, Fe^{3+}, Fe^{2+}, Hg^{2+}, Mg^{2+}, Mn^{2+}, Ni^{2+}, and Zn^{2+} from amino acid sequences. One is a homology-based method, and the other is a machine-learning method. The homology-based method predicts the binding sites from homologous sequences obtained by a protein–protein basic local alignment search tool (BLASTP) search. The machine-learning method uses a support vector machine with three protein sequence features. Our results showed that the the homology-based method achieved an accuracy of 0.9905 and a specificity of 0.9978, while the machine-learning method showed balanced performance with regard to accuracy, sensitivity, and specificity. Especially, the sensitivity of the machine-learning method was 0.8239, and many metal ion-binding sites were predicted only by the machine-learning method.

Key Words: protein, metal ion, binding site prediction, machine learning, homology search

Significance | Metal ions play an important role for living organisms. We developed two prediction methods for metal ion binding sites of proteins from their amino acid sequences. Our results showed that the machine-learning method showed balanced performance while the sequence similarity-based method achieved a very high specificity.

*Correspondence: Kentaro Shimizu, Associate professor. The Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo.
Email: shimizu@bi.a.u-tokyo.ac.jp

Editor Md. Shamsuddin Sultan Khan, University of Western Sydney. And accepted by the Editorial Board September 01, 2019 (received for review July 21, 2019)

Introduction

Metalloproteins are proteins that can bind one or more metal ions, which are essential for the proteins to perform their catalytic or structural functions (Degtyarenko et al., 2000). Approximately one-third of the structures in the Protein Data Bank (PDB) (Berman et al., 2000) contain at least one metal ion. Metal ions are important for the function of proteins and typically have catalytic, transfer, regulatory, structure, recognition, transcription, and transducer roles. The presence of metal ions in metalloproteins is not only helpful for maintaining spatial stability, but also important for executing the physiological functions of the proteins (Holm et al., 1996, Matthews et al., 2008). With the rapid expansion of protein databases, it has become important to identify metal ion-binding sites in metalloproteins in order to understand their role in protein function. Metal ion-binding proteins are experimentally identified and characterized using nuclear magnetic resonance spectroscopy (Jensen et al., 2005; Zhu et al., 2004), gel electrophoresis (Greenough et al., 2015), metal-affinity column chromatography (Herald et al., 2003), electrophoretic mobility shift assay (EMSA) (Hellman et al., 2007), absorbance spectroscopy (Korshin et al., 2009), and mass spectrometry (Binet et al., 2003). However, most of these methods

Author Affiliation:
[1] The Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo.

Please cite this article:
Zihan Tian, Cao Wei, Yutaka Moriwaki, Tohru Terada, Shugo Nakamura, Kazuya Sumikoshi, Fang Chun, and Kentaro Shimizu (2019). Prediction of Protein–Metal Ion-Binding Sites Using Sequence Homology and Machine-Learning Methods, Advanced Bioinformatics & Chemistry, 1(1), pages 025-034.

require complex steps and specialized equipment, making them labor-intensive and time-consuming. Hence, prediction of metal ion-binding sites using bioinformatics methods will not only be useful for annotation of experimentally uncharacterized proteins, but also will have a significant role in the prediction of protein structure, function, and genome annotation.

Computational methods have been used to predict protein–metal ion-binding sites. Lin et al. 2005 used a neural networks method to predict metal-binding sites in proteins. The target metal ions of Lin's work are Ca^{2+}, Mg^{2+}, K^{+}, and Na^{+}. Kumar, 2017 used the random forest algorithm to predict protein–metal ion–binding sites for multiple metal ions: Ca^{2+}, Co^{2+}, Cu^{2+}, Fe^{3+}, Mg^{2+}, Mn^{2+}, Ni^{2+}, and Zn^{2+}. We compared our method with Kumar's. Many other studies have also predicted specific metal ion–binding sites. For example, Zn^{2+}–binding site predictors (Passerini et al., 2007; Srivastava et al., 2018) used the support vector machine and sequence profile information. Metal ion–binding sites have also be predicted using 3D protein structures (Chen et al., 2013; Lu et al., 2012; Yan et al., 2019; Schymkowitz et al., 2005; Goyal et al., 2008). For example, Chen et al. 2013, searched for a triad of amino acids having ligand atoms within specific distances to predict protein–Zn^{2+}–binding sites while Lu et al. 2012 used the fragments transformation method based on the binding site templates. The approach using 3D protein structures can improve prediction performance; however, 3D protein structures are not always available. Predicting metal ion–binding sites only from sequence information is widely applied. Studies have also identified metal ion–binding histidine and cysteine residues in a protein (Passerini et al., 2006; Haberal et al., 2019). However, since these residues are important in metal ion binding, this approach is different from our study.

In this study, we implemented and compared two methods to predict metal ion-binding sites for Ca^{2+}, Co^{2+}, Cu^{2+}, Cu^{+}, Fe^{3+}, Fe^{2+}, Hg^{2+}, Mg^{2+}, Mn^{2+}, Ni^{2+}, and Zn^{2+} from amino acid sequences: a homology-based method and a machine-learning method. Homology search is a popular method in which multiple alignment of the homologous sequences enables the prediction of metal ion-binding sites. We implemented the homology-based method, which performs this prediction procedure automatically. In the machine-learning method, a support vector machine was employed, and three protein sequence features were selected as its input. To the best of our knowledge, these two methods have not been directly compared using the same dataset. We therefore compared and discussed the results of prediction for each metal ion.

Materials and methods

We developed two kinds of method to predict metal-binding sites from amino acid sequences only: a homology-based method and a machine-learning method (**Figure 1**). In the homology-based method, BLASTP search (Altschul et al., 1997) was used to identify sequence homology. In the machine-learning method, a support vector machine (SVM) (Boser et al., 1992 was employed with the sequence features amino acid type, position specific scoring matrix (PSSM), and 13 groups of amino acids based on physicochemical properties described below. Basic flows of the homology-based method and the machine-learning method are shown in orange and blue, respectively.

Dataset

All protein structures in PDB files and mmCIF files and sequences in FASTA files were extracted from the website of the PDB database (Berman et al., 2000) using the following criteria: (1) the protein complex contained at least one metal ion; (2) the x-ray protein structure had a resolution less than or equal to 2.5 Å; and (3) the chain length was greater than 50 residues. A total of 11,903 protein structures and 33,271 chains were collected. For the homology-based method, these chains were used as the positive dataset for BLASTP search. For the machine-learning method, these chains were clustered by CD-HIT (Fu et al., 2012) to remove redundancy for sequence identity thresholds of 30%. **Table 1** lists the statistics for each kind of metal ion-binding protein.

Homology-based method

In the homology-based method, we used BLASTP to obtain the homologous sequences of the query sequence in the positive dataset for this method. The homologous sequences were selected based on the E-value; sequences with E-values lower than a threshold value are defined as homologous sequences.

For instance, when 1AH7:A (A chain of PDB structure 1AH7) was input as a query sequence, assuming the binding sites of this sequence were unknown, three sequences of known structures, 1KHO:A, 2WXU:A, and 1OLP:A, were obtained as the homologous sequences. If an aligned residue in one of the homologous sequences was a metal-binding residue, this residue in the query sequence was predicted as a metal-binding site; if not, this residue was predicted as a nonbinding site. **Figure 2** shows an example of this prediction. As shown, 2WXU:A lacks two Zn^{2+}, and the corresponding residues are not defined as binding sites. However, residues of the target aligned to these positions are predicted as binding residues because the other homologs contain Zn^{2+}-binding residues. **Figure 3** shows the superposition of the target and homolog structures. In this example, Zn^{2+}-binding sites (pockets) of the homologs have similar structures and are well aligned.

Machine-learning methods and feature extraction

We used SVM as a machine-learning algorithm. SVM shows good performance and generalization abilities for classification and regression analysis. The radial basis function (RBF) kernel was adopted as a kernel function since it seemed

appropriate to capture the nonlinearity of multiple binding site properties. We used scikit learn (Pedregosa et al., 2011) to implement the SVM. SVM requires a fixed length of the feature vector for training and testing. We extracted three features to train SVM: amino acid type, position specific scoring matrix (PSSM), and side chain type.

Amino acid type

Each amino acid residue was encoded as a vector of 20 elements; an element that corresponds to an amino acid type is one, and the others are zero. The all zero vector represents a spacer for N- and C-terminals of the sequences. Twenty types of amino acids are decoded into a binary pattern.

PSSM

PSSM is the amino acid substitution score for each position in a protein multiple sequence alignment. The PSSMs for each sequence in the dataset were obtained by the PSI-BLAST (Cooper et al., 2004) program with three iterations of search against the nonredundant data in the National Center for Biotechnology Information (NCBI). For each sequence, PSSM was represented by an $N \times 20$ matrix, where N is the length of the amino acid sequence. PSSM scores are generally shown as positive or negative integers.

Side chain type

Amino acids are organic compounds containing amine (-NH_2) and carboxyl (-COOH) functional groups, along with a side chain (R group) specific to each amino acid.

Amino acids can be classified according to the physicochemical properties of their side chains. In this study, we implemented the grouping scheme proposed in (Cooper et al., 2004); according to side chain polarity, amino acids are divided into four groups; according to side chain class, amino acids are divided into nine groups. The definitions of these features are described in **Supplementary material A**. Using this feature, each residue was represented by a vector of 13 elements.

Measurement of performance

In the machine-learning method, we employed five-fold cross-validation. The final performance at each parameter was obtained by averaging the performance of all five test sets. The performance of the SVM model at a particular training parameter was assessed using threshold-dependent parameters, namely accuracy, sensitivity, specificity, precision, and Matthew's correlation coefficient (MCC).

These parameters were calculated using the true positive (TP), true negative (TN), false positive (FP), and false negative (FN), where TP is correctly predicted metal ion-binding amino acids, TN is correctly predicted metal ion-nonbinding amino acids, FP is wrongly predicted metal ion-binding amino acids, and FN is wrongly predicted metal ion-nonbinding amino acids.

Accuracy is the ratio of the correct predictions (TP + TN) to all predictions, and it is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}.$$

Sensitivity is the ratio of correctly predicted binding residues (TP) to actual binding residues (TP + FN), and it is calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}.$$

Specificity is the ratio of correctly predicted nonbinding residues to actual nonbinding residues (TN + FP), and it is calculated as follows:

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

Precision is the ratio of correctly predicted binding residues (TP) to correctly predicted binding and nonbinding residues (TP + FP), and it is calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

MCC is a balanced measurement that is used to assess the effectiveness of the performance, and it was calculated as follows:

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

Results and discussion

Amino acid frequencies of metal ion-binding sites

We calculated the frequencies of 20 types of amino acid residues in the metal ion-binding sites. The frequencies of each amino acid in the binding sites of each metal ion are represented in **Supplementary material B.**

Metal ions are commonly coordinated by nitrogen, oxygen, or sulfur centers belonging to side chains on the amino acid residues of the protein, where metal ions provide empty orbits and amino acids provide electrons. The imidazole substituents of histidine residues, the thiolate substituents of cysteine residues, and the carboxylate groups of aspartic acid and glutamic acid can provide electrons as donor groups. This is consistent with the results of the present study, which showed that metal ions preferentially bind certain residues, namely cysteine, histidine, aspartic acid, and glutamic acid. As for Ca^{2+} and Mg^{2+}, cysteine and histidine do not form cross-links with amino acid residues for constructing a specific structure; therefore, the frequencies of cysteine and histidine are low compared to other ions. Negatively charged residues (aspartic acid and glutamic acid) have high frequencies because of electrostatic interaction with Ca^{2+} and Mg^{2+}. These ions also bind to oxygen atoms of the backbone. The imidazole substituents of histidine residues, the thiolate substituents of cysteine residues, and the carboxylate groups of aspartic acid and glutamic acid can provide electrons as donor groups, and metal ions can provide empty orbits. Nonpolar residues, such as leucine, isoleucine, and valine, as well as less polar amino acids, such as proline and threonine, show no preference for metal coordination. However, they show certain

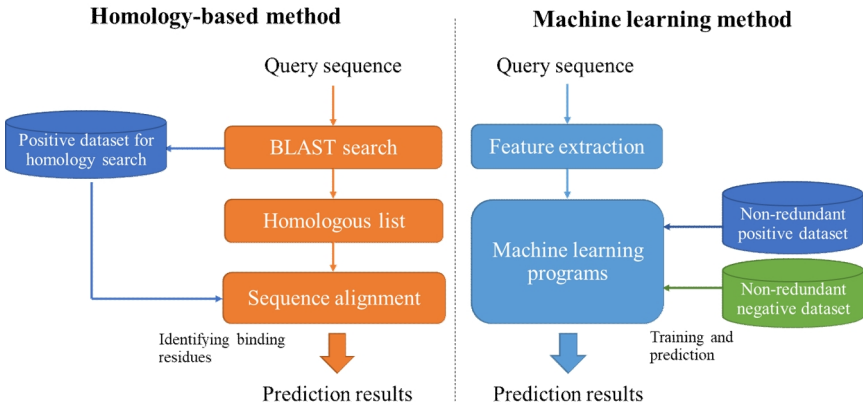

Figure 1 | Homology-based method and machine-learning method

Table 1 | Numbers of proteins and chains in the positive dataset. The metal ion–binding site was defined as residues within 3.5 Å of a metal ion center. These residues were obtained by analyzing PDB structures.

Metal ion	Number of proteins	Number of chains	Number of nonredundant chains
Ca^{2+}	2,626	6,861	1,246
Co^{2+}	260	603	181
Cu^{2+}	268	644	131
Cu^{+}	52	111	33
Fe^{3+}	413	1,167	215
Fe^{2+}	167	545	97
Hg^{2+}	152	318	119
Mg^{2+}	3,782	11,600	1,423
Mn^{2+}	883	2,750	437
Ni^{2+}	498	1,125	353
Zn^{2+}	2,802	7,547	1,478
Total	11,903	33,271	5,713

```
Target     (1AH7_A)   1  WSAEDKHKEGVNSHLWIVNRAIDIMS----RNTTLVKQDRVAQLNEWRTELENGIYAADYENPYYDNSTFASHFYDPDNGKTYIP---------FAKQAK   87
                         *          *                                                *                *

Homolog 1  (1KHO_A)   1  W---DGKADGTGTHAMIATQGVTILENDLSSNEPEVIRNNLEILKQNMHDLQLGSTYPDYDKNAYD--LYQDHFWDPDTDNNFTKDSKWYLSYSIPDTAE   95
Homolog 2  (2WXU_A)   1  W---DGKIDGTGTHAMIVTQGVSILENDLSKNEPESVRKNLEILKENMHELQLGSTYPDYDKNAYD--LYQDHFWDPDIDNNFSKDNSWYLAYSIPDTGE   95
Homolog 3  (1OLP_A)   1  W---DGKEDGTGTHSVIVTQAIEMLKHDLSKDEPEAIRNDLSILEKNLHKFQLGSTFPDYDPNAYS--LYQDHFWDPDTDHNFTQDNKWYLSYAVPDNAE   95

Target     (1AH7_A)  88  ETGAKYFKLAGESYKNKDMKQAFFYLGLSLHYLGDVNQPMHAANFTNLSYPQGFHSKYENFVDTIKDNYKVTDGNGYWNWKGTNPEEWIHGAAVVAKQDY  188
                                                    *   *   *                 *   *

Homolog 1  (1KHO_A)  96  SQIRKFSALARYEWKRGNYKQATFYLGEAMHYFGDADTPYHAANVTAVDSPG--HVKFETFAEDRKDQYKIN-TTGSKTNDAFYSNILTNEDFNSWSKEF  192
Homolog 2  (2WXU_A)  96  SQIRKFSALARYEWQRGNYKQATFYLGEAMHYFGDIDTPYHPANVTAVDSAG--HVKFETFAEERKEQYKIN-TAGCKTNEAFYTDILKNKDFNAWSKEY  192
Homolog 3  (1OLP_A)  96  SQTRKFATLAKNEWDKGNYEKAAWYLGQGMHYFGDLNTPYHAANVTAVDSPG--HVKFETYAEERKDTYRLD-TTGYNTDDAFYKDTLKNDNFNEWSKGY  192
```

Figure 2 | Example of the homology-based method.

frequencies because we define a metal ion–binding site not as interacting residues but as neighboring residues within 3.5 Å of a metal ion center.

Performance of the homology-based method

In the homology-based method, we used different E-value thresholds: 0.0001. 0.001, 0.01, and 0.1. The results are shown in **Supplementary material C**. The number of true positive cases increases and the number of true negative cases decreases as the E-value threshold increases. The accuracy depends on the number of true positives and true negatives, and in many metal ions, except for Ca^{2+}, Mg^{2+}, and Ni^{2+}, the increase in true positives is larger than the increase in true negatives, and thus the accuracy is best for the E-value threshold 0.1. We set this rather high E-value threshold because the sequences of metalloproteins for each metal ion are not so similar as can be detected by BLASTP although some kinds of motifs may exist for binding sites. **Table 2** shows the results of homology-based prediction for each of 11 metal ions when the E-value threshold is 0.1. The last column, "All," shows the performance for all 11 metal ions (not all metal ions in natural proteins). Metal ions bound to some specific protein families have high sensitivity. In terms of accuracy, the homology-based method performed excellently, with an overall accuracy of 0.9905. It should be noted that with this method, the amount of negative data, which was easier to predict, was much larger than the amount of positive data. For instance, the sensitivity was 0.3544 and the specificity was 0.9961 when an E-value of 0.1 was used to predict Ca^{2+} binding sites, which means that only 35.44% of actual binding sites (positive data) were correctly predicted. Since substantial numbers of nonbinding sites were correctly predicted, the accuracy became higher.

Performance of the machine-learning method

An SVM classifier with RBF kernel has at least two parameters that need to be tuned for good performance: the cost parameter C, which determines the misclassification penalty; and the gamma parameter γ, which is used in the RBF kernel function. We used grid search for obtaining optimal values of γ and C to train the SVM models. As a result, we set γ and C to 0.07 and 100, respectively, to train the SVM models.

In order to determine an optimal window size, we used the PSSM features of metal ion-binding proteins with window sizes of 9 to 19 to train SVM models. The results are shown in **Supplementary material D**. As can be seen in the results, for Ca^{2+} and Mg^{2+}, the accuracy increased from window size 9 to window size 15; for Co^{2+}, Cu^{+}, Fe^{3+}, Hg^{2+}, Mn^{2+}, and Zn^{2+}, the accuracy increased up to window size 13; for Cu^{2+} and Fe^{2+}, the accuracy increased up to window size 11; for Ni^{2+}, the best performance was shown at window size 9. More than half of the models performed best at window size 13, and the accuracy of the others did not change much in the range between their own optimal window size and window size 13. Therefore, we selected 13 residues as the optimal window size of all 11 types of metal ion-binding proteins. **Table 3** shows the performance of the machine-learning (SVM) method with the PSSM feature. The accuracy was 0.8017 and the MCC was 0.61 overall, and the performance was best for Cu^{+}, with an accuracy of 0.8846 and an MCC of 0.77. It is interesting that the accuracy for Cu^{+} was the worst in the homology-based method. **Table 4** shows the performance of the machine-learning method with the PSSM, the amino acid type, and the side chain type features. The accuracy increased to 0.8336 and the MCC increased to 0.67 overall. The performance was best for Zn^{2+}, with an accuracy of 0.8901 and an MCC of 0.78.

Comparison of the two methods

We compared the results of the homology-based method and the machine-learning method. **Figure 4** shows the number of chains predicted by the two methods for each metal ion, and **Table 5** shows the number of unpredictable chains in the homology-based method. (In the homology-based method, we used the E-value threshold of 0.01.) For 11 types of metal ion-binding protein, the numbers of chains predicted by the SVM method were larger than those predicted by the homology-based method, and only 78% of chains hit their homologous sequences. For instance, the Cu^{+} binding sites of three chains (PDB IDs: 4BZ4-A, 4MAI-A, 5FJE-B) were successfully predicted by the SVM method and were not predicted by the homology-based method. The SVM method can better predict Ca^{2+}-, Mg^{2+}-, and Zn^{2+}-binding chains compared to the homology-based method. Since Ca^{2+}- and Mg^{2+}-binding proteins exist in various families, the sensitivity of the homology-based method is low. Zn^{2+}-binding sites can often be represented as motifs, and their sequence features tend to be local. The SVM method can recognize these features, while the homology-based method cannot align these sequence patterns. **Table 5** summarizes the performances of the homology-based method and the machine-learning method. The high accuracy and low sensitivity of the homology-based method were caused by the inequality between negative and positive data. By contrast, the SVM method, which predicted with a balanced performance of accuracy, sensitivity, and specificity, was more effective.

Comparison with other work

We also compared our results with those of Kumar's method (Kumar et al., 2017), which used simplified amino acid alphabets and a random forest model (**Table 6**). Our method using an SVM model targeted the prediction of three types of metal ion-binding protein (Cu^{+}-, Fe^{2+}-, and Hg^{2+}-binding proteins) which are not available in Kumar's method. As for Hg^{2+} ion, however, our dataset contains Hg^{2+} ions for the soaking in the X-Ray crystal structure analysis, and the results are not clear. When the same types of metal ion-binding proteins were compared, a considerable increase in the accuracy of the Zn^{2+}-, Mn^{2+}-, and Fe^{3+}-binding

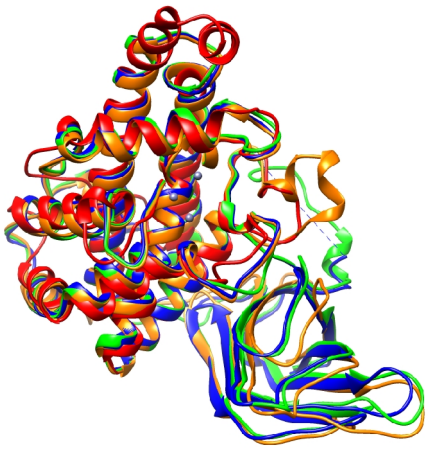

Figure 3 | Superposition of the target and homolog structures in the example given in Figure 2. The target structure 1AH7:A is in red, while the three homolog structures 1KHO:A, 2WXU:A, and 1OLP:A are in blue, green, and orange, respectively. Zn2+ of the target structure are represented as spheres in gray.

Table 2 | Performance of the homology-based method. MCC, Matthew's correlation coefficient.

Metal ion	Accuracy	Sensitivity	Specificity	Precision	MCC
Ca^{2+}	0.9873	0.3544	0.9961	0.5580	0.4386
Co^{2+}	0.9898	0.1566	0.9995	0.7778	0.3463
Cu^{2+}	0.9915	0.4486	0.9982	0.7616	0.5807
Cu^{+}	0.9872	0.2266	0.9997	0.9355	0.4570
Fe^{3+}	0.9922	0.4977	0.9986	0.8226	0.6365
Fe^{2+}	0.9921	0.3228	0.9993	0.8358	0.5166
Hg^{2+}	0.9893	0.0174	0.9999	0.7778	0.1153
Mg^{2+}	0.9931	0.3671	0.9975	0.5139	0.4310
Mn^{2+}	0.9929	0.3407	0.9992	0.7950	0.5177
Ni^{2+}	0.9903	0.1072	0.9993	0.5943	0.2497
Zn^{2+}	0.9896	0.4177	0.9983	0.7915	0.5706
All	0.9905	0.3591	0.9978	0.6563	0.4812

Table 3 | Performance of the SVM model using the PSSM feature. SVM, support vector machine; PSSM, position specific scoring matrix; MCC, Matthew's correlation coefficient.

Metal ion	Accuracy	Sensitivity	Specificity	Precision	MCC
Ca^{2+}	0.7593	0.7100	0.8087	0.7882	0.5213
Co^{2+}	0.7931	0.7818	0.8033	0.7818	0.5851
Cu^{2+}	0.7842	0.7474	0.8211	0.8068	0.5299
Cu^{+}	0.8846	0.8519	0.9200	0.9200	0.7719
Fe^{3+}	0.8613	0.8765	0.8466	0.8466	0.7231
Fe^{2+}	0.8613	0.8750	0.8493	0.8358	0.6802
Hg^{2+}	0.6790	0.6941	0.6623	0.6941	0.2231
Mg^{2+}	0.7761	0.7149	0.8331	0.7997	0.5529
Mn^{2+}	0.8467	0.8480	0.8454	0.8423	0.6933
Ni^{2+}	0.7629	0.7566	0.7688	0.7566	0.4852
Zn^{2+}	0.8372	0.8237	0.8511	0.8514	0.6749
All	0.8017	0.7492	0.8537	0.8354	0.6064

Table 4 | Performance of the SVM model using combined features. SVM, support vector machine; MCC, Matthew's correlation coefficient.

Metal ion	Accuracy	Sensitivity	Specificity	Precision	MCC
Ca^{2+}	0.7726	0.7382	0.8070	0.7932	0.5465
Co^{2+}	0.8103	0.7909	0.8279	0.8056	0.6194
Cu^{2+}	0.7789	0.7368	0.8211	0.8046	0.5599
Cu^{+}	0.8846	0.8519	0.9200	0.9200	0.7719
Fe^{3+}	0.8642	0.8824	0.8466	0.8475	0.7290
Fe^{2+}	0.8321	0.8657	0.8000	0.8056	0.6664
Hg^{2+}	0.6049	0.6000	0.6104	0.6296	0.2101
Mg^{2+}	0.7840	0.7358	0.8289	0.8003	0.5680
Mn^{2+}	0.8500	0.8649	0.8355	0.8366	0.7005
Ni^{2+}	0.7990	0.8307	0.7688	0.7734	0.6000
Zn^{2+}	0.8901	0.9410	0.8374	0.8570	0.7837
All	0.8336	0.8239	0.8432	0.8388	0.6673

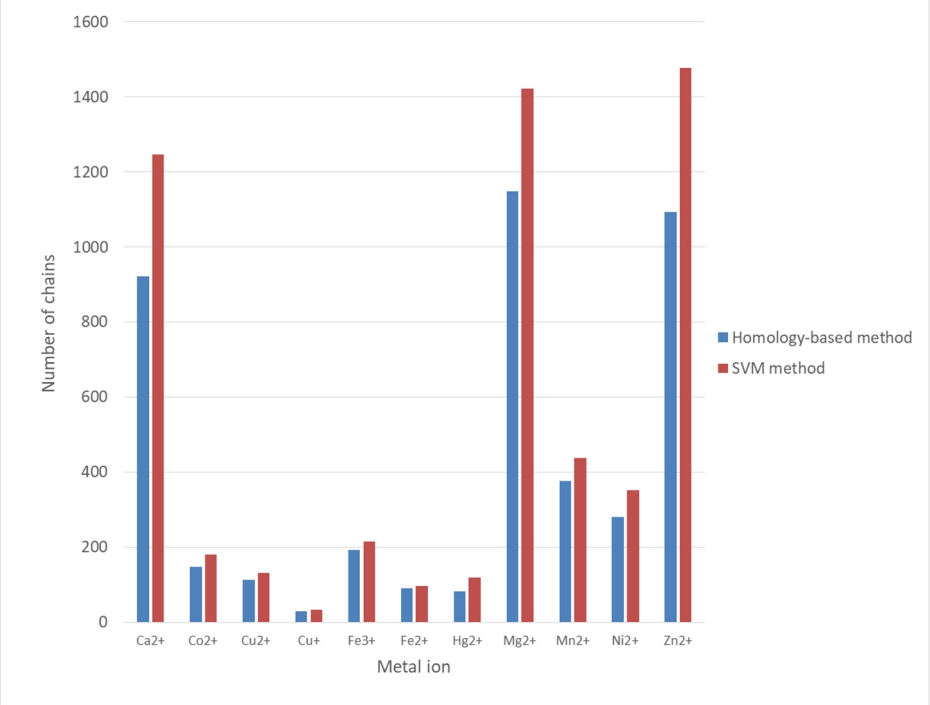

Figure 4 | Numbers of chains predicted by the homology-based method and the support vector machine (SVM) method

Table 5 | Performance of the homology-based method and the support vector machine (SVM) method. MCC, Matthew's correlation coefficient.

Method	Accuracy	Sensitivity	Specificity	Precision	MCC
Homology-based	0.9905	0.3591	0.9978	0.6563	0.4812
SVM	0.8336	0.8239	0.8432	0.8388	0.6673

Table 6 | Performance of Kumar's method and our method

Metal ion	Kumar's method		Our method	
	Accuracy	Sensitivity	Accuracy	Sensitivity
Ca^{2+}	0.754	0.769	0.773	0.738
Co^{2+}	0.853	0.884	0.810	0.791
Cu^{2+}	0.781	0.746	0.779	0.737
Cu^{+}	-	-	0.885	0.852
Fe^{3+}	0.756	0.722	0.864	0.882
Fe^{2+}	-	-	0.832	0.866
Hg^{2+}	-	-	0.605	0.600
Mg^{2+}	0.740	0.766	0.784	0.736
Mn^{2+}	0.688	0.729	0.850	0.865
Ni^{2+}	0.907	0.945	0.799	0.831
Zn^{2+}	0.690	0.740	0.890	0.941

proteins was observed. The disparity between the accuracy of the Ca^{2+}-, Co^{2+}-, Cu^{2+}-, and Mg^{2+}-binding proteins and the accuracy of Kumar's method was not apparent. For Ni^{2+}-binding proteins, the performance of our method was unsatisfactory.

Conclusions

Metal ions preferentially bind certain amino acid residues, and many sequence motifs are known in the metalloproteins. In this study, the homology-based method achieved higher accuracy and specificity compared to the machine-learning (SVM) method, while the machine-learning method showed balanced performance with regard to accuracy, sensitivity, and specificity. Especially, the sensitivity of the machine-learning method was high, and we found that it could predict some metal ion-binding sites that were not predicted by the homology-based method and achieved a balanced performance of accuracy, sensitivity, and specificity. We can conceive the following integration of the machine-learning and homology-based methods. Since the machine-learning method achieves high sensitivity, we first obtain candidates of metal ion–binding sites using the machine-learning method. Then, if the homology-based method predicts that the candidates are nonbinding sites, we discard them. This result can be reliable for very high specificity of the homology-based method. In addition, if the homology-based method predicts metal ion–binding sites when homologous sequences are hit by a BLASTP search with low *E*-values, the result can also be used as candidates of metal ion–binding sites. A detailed design of the integration would be part of future research.

Author Contributions

K. S. (Kentaro Shimizu) designed the study and made critical revision of the article for important intellectual content. Z. T. implemented the prediction programs and wrote the initial draft of the manuscript. C. W. contributed to data collection and interpretation. Y. M. assisted the implementation of the prediction programs. T. T. contributed to analysis and interpretation of data. S. N. and K. S. (Kazuya Sumikoshi) assisted in the preparation of the manuscript. F. C. assisted the statistical analysis and machine learning. All authors approved the final version of the manuscript, and agree to be accountable for all aspects of the work.

Acknowledgment

The authors would like to thank Yan Chen for valuable comments. This research is supported by Basis for Supporting Innovative Drug Discovery and Life Science Research from the Japan Agency for Medical Research and Development, and the Uehara Memorial Foundation.

Competing financial interests

The author(s) declare no competing financial interests.

Supplementary Information

All standard and non-standard abbreviation in PDF file. Please download.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

https://doi.org/10.1093/nar/25.17.3389

PMid:9254694 PMCid:PMC146917

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. 28, 235-242.
https://doi.org/10.1093/nar/28.1.235
PMid:10592235 PMCid:PMC102472

Binet, M.R.B., Ma, R., McLeod, C.W., Poole, R.K. (2003). Detection and characterization of zinc-and cadmium-binding proteins in Escherichia coli by gel electrophoresis and laser ablation-inductively coupled plasma-mass spectrometry. Anal. Biochem. 318, 30-38.
https://doi.org/10.1016/S0003-2697(03)00190-8

Boser, B.E., Guyon, I.M., Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. Proc. the fifth Annual Workshop on Computational Learning Theory. ACM. 25, 144-152.
https://doi.org/10.1145/130385.130401

Chen, Z., Wang, 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. Mol. Biosyst. 9, 2213-2222.
https://doi.org/10.1039/c3mb70100j
PMid:23861030

Cooper, G.M., Hausman, R.E. (2007). The cell: Molecular approach. ASM Press, Washington, D.C.

Degtyarenko, K. (2000). Bioinorganic motifs: towards functional classification of metalloproteins. Bioinformatics. 16, 851-864.
https://doi.org/10.1093/bioinformatics/16.10.851
PMid:11120676

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 28, 3150-3152.
https://doi.org/10.1093/bioinformatics/bts565
PMid:23060610 PMCid:PMC3516142

Goyal, K., Mande, S.C. (2008). Exploiting 3D structural templates for detection of metal-binding sites in protein structures. Proteins. 70, 1206-1218.
https://doi.org/10.1002/prot.21601
PMid:17847089

Greenough, L., Schermerhorn, K.M., Mazzola, L., Bybee, J., Rivizzigno, D., Cantin, E., Slatko, B.E., Gardner, A.F. (2015). Adapting capillary gel electrophoresis as a sensitive, high-throughput method to accelerate characterization of nucleic acid metabolic enzymes. Nucleic Acids Res. 44, e15-e15.
https://doi.org/10.1093/nar/gkv899
PMid:26365239 PMCid:PMC4737176

Haberal, İ., Oğul, H. (2019). Prediction of Protein Metal Binding Sites Using Deep Neural Networks. Mol. Inform. 38, e1800169.
https://doi.org/10.1002/minf.201800169
PMid:30977960

Hellman, L.M., Fried, M.G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat. Protoc. 2, 1849.
https://doi.org/10.1038/nprot.2007.249
PMid:17703195 PMCid:PMC2757439

Herald, V.L., Heazlewood, J.L., Day, D.A., Millar, A.H. (2003). Proteomic identification of divalent metal cation-binding proteins in plant mitochondria. FEBS Lett. 537, 96-100.
https://doi.org/10.1016/S0014-5793(03)00101-7

Holm, R.H., Kennepohl, P., Solomon, E.I. (1996). Structural and functional aspects of metal sites in biology. Chem. Rev. 96, 2239-2314.
https://doi.org/10.1021/cr9500390
PMid:11848828

Jensen, M.R., Petersen, G., Lauritzen, C., Pedersen, J., Led, J.J. (2005). Metal binding sites in proteins: identification and characterization by paramagnetic NMR relaxation. Biochemistry. 44, 11014-11023.
https://doi.org/10.1021/bi0508136
PMid:16101285

Korshin, G., Chow, C.W.K., Fabris, R., Drikas, M. (2009). Absorbance spectroscopy-based examination of effects of coagulation on the reactivity of fractions of natural organic matter with varying apparent molecular weights. Water Res. 43, 1541-1548.
https://doi.org/10.1016/j.watres.2008.12.041
PMid:19131089

Kumar, S. (2017). Prediction of metal ion binding sites in proteins from amino acid sequences by using simplified amino acid alphabets and random forest model. Genomics Inform. 15, 162-169.
https://doi.org/10.5808/GI.2017.15.4.162
PMid:29307143 PMCid:PMC5769865

Lin, C.T., Lin, K.L., Yang, C.H., Chung, I.F., Huang, C.D., Yang, Y.S. (2005). Protein metal binding residue prediction based on neural networks. Int. J. Neural Syst. 15, 71-84.
https://doi.org/10.1142/S0129065705000116
PMid:15912584

Lu, C., Lin, Y., Lin, J., Yu, C. (2012). Prediction of Metal Ion-Binding Sites in Proteins Using the Fragment Transformation Method. PLoS ONE. 7, e39252.
https://doi.org/10.1371/journal.pone.0039252
PMid:22723976 PMCid:PMC3377655

Matthews, J.M., Loughlin, F.E., Mackay, J.P. (2008). Designed metal-binding sites in biomolecular and bioinorganic interactions. Curr. Opin. Struct. Biol. 18, 484-490.
https://doi.org/10.1016/j.sbi.2008.04.009
PMid:18554898

Passerini, A., Andreini, C., Menchetti, S., Rosato, A., Frasconi, P. (2007). Predicting zinc binding at the proteome level. BMC Bioinformatics. 8, 39.
https://doi.org/10.1186/1471-2105-8-39
PMid:17280606 PMCid:PMC1800866

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. 65, 305-316.
https://doi.org/10.1002/prot.21135
PMid:16927295

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V.,

Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12, 2825-2830.

Schymkowitz, J.W.H., Rousseau, F., Martins, I.C., Ferkinghoff-Borg, J., Stricher, F., Serrano, L. (2005). Prediction of water and metal binding sites and their affinities by using the Fold-X force field. Nucleic Acids Res. 102, 10147-10152.
https://doi.org/10.1073/pnas.0501980102
PMid:16006526 PMCid:PMC1177371

Srivastava, A., Kumar, M. (2018). Prediction of zinc binding sites in proteins using sequence derived information. J. Biomol. Struct. Dyn. 36, 4413-4423.
https://doi.org/10.1080/07391102.2017.1417910
PMid:29241411

Yan, R., Wang, X., Tian, Y., Xu, J., Xu, X., Lin, J. (2019). Prediction of zinc-binding sites using multiple sequence profiles and machine learning methods. Molecular Omics. 15, 205-215.
https://doi.org/10.1039/C9MO00043G
PMid:31046040

Zhu,D., Herbert, B.E., Schlautman, M.A., Carraway, E.R. (2004). Characterization of cation-π interactions in aqueous solution using deuterium nuclear magnetic resonance spectroscopy. J. Environ. Qual. 33, 276-284.
https://doi.org/10.2134/jeq2004.2760
PMid:14964382