Recognizing Ion Ligand–Binding Residues by Random Forest Algorithm Based on Optimized Dihedral Angle

Liu Liu, Xiuzhen Hu*, Zhenxing Feng, Shan Wang, Kai Sun and Shuang Xu

College of Sciences, Inner Mongolia University of Technology, Hohhot, China

The prediction of ion ligand–binding residues in protein sequences is a challenging work that contributes to understand the specific functions of proteins in life processes. In this article, we selected binding residues of 14 ion ligands as research objects, including four acid radical ion ligands and 10 metal ion ligands. Based on the amino acid sequence information, we selected the composition and position conservation information of amino acids, the predicted structural information, and physicochemical properties of amino acids as basic feature parameters. We then performed a statistical analysis and reclassification for dihedral angle and proposed new methods on the extraction of feature parameters. The methods mainly included applying information entropy on the extraction of polarization charge and hydrophilic–hydrophobic information of amino acids and using position weight matrices on the extraction of position conservation information. In the prediction model, we used the random forest algorithm and obtained better prediction results than previous works. With the independent test, the Matthew’s correlation coefficient and accuracy of 10 metal ion ligand–binding residues were larger than 0.07 and 52%, respectively; the corresponding evaluation values of four acid radical ion ligand–binding residues were larger than 0.15 and 86%, respectively. Further, we classified and combined the phi and psi angles and optimized prediction model for each ion ligand–binding residue.

Keywords: binding residues, dihedral angle, information entropy, ion ligands, protein

INTRODUCTION

The protein is the foundation of life and plays an important role in the life activities. Proteins carried out various functions by interactions, such as protein–protein interaction, protein–DNA interaction, protein–RNA interaction, and protein–ion ligand interaction (Lin et al., 2005; Ebert and Altman, 2008; Pan and Shen, 2018; Al-Mugotir et al., 2019; Emamjomeh et al., 2019; Robin et al., 2019). Because more than half of the proteins required binding with ion ligands for functions, research of ion ligand–binding residues on proteins was of great significance. However, it was difficult to accurately predict the ion ligand–binding residues on the protein sequence because of the small size and high versatility of ion ligands.
Current theoretical prediction methods of ligand-binding residues can be roughly classified into sequence-based method and three-dimensional (3D) structure–based method. The experiments showed that the accuracy of the 3D structure–based prediction was higher than that of sequence-based prediction (Yang et al., 2013, 2018). However, the number of proteins with known amino acid sequence was far more than that with known 3D structure. Although the prediction accuracy of sequence-based method is not as satisfactory as 3D structure–based, sequence-based method is still generally used.

In general, there were three key points in predicting the ion ligand–binding residues with theoretical methods: research objects selection, the selection, and extraction of feature parameters, and the selection of algorithms. In recent decades, researchers conducted different studies on binding residues of metal and acid radical ion ligands. In these studies, most researchers were working on the development of features to improve the prediction results of ion ligand–binding residues. Among the feature parameters used to predict ion ligand–binding residues, position conservation information, and composition of amino acids were two commonly used basic feature parameters (Sodhi et al., 2004; Komiyama et al., 2015; Hu et al., 2016a; Liu et al., 2019; Wang et al., 2019). Besides, based on the biological background of interaction between ion ligands and proteins, researchers added physicochemical properties of amino acids, secondary structure, and relative solvent accessibility (RSA) to identify ion ligand–binding residues (Lin et al., 2006; Jiang et al., 2016; Cao et al., 2017; Li et al., 2017). Using these features obtained improved prediction results. In this work, we further improved prediction result by optimizing the extraction method of features. In the previous work, on the extraction of position conservation information of amino acids, some researchers used the position specific score matrix (PSSM) method to extract it (Sodhi et al., 2004; Hu et al., 2016a), whereas the other ones used the position weight scoring algorithm to extract its score values (Liu et al., 2019; Wang et al., 2019). However, the dimension of PSSM is excessively high, which will potentially lead to the overfitting problem, and the dimension of the score values is too low, which will lose a lot of information. Thus, we constructed the position weight matrices to extract the 2L-dimensional position conservation information of amino acids. In terms of extraction of the hydrophilic–hydrophobic information of amino acids, autocross covariance formula was attempted (Jiang et al., 2016). However, the method did not take into account the different number of amino acid species contained in each class of hydrophilic–hydrophobic properties.

| Ligands | Chains | P | N |
|---------|--------|---|---|
| Metal ions | Zn$^{2+}$ | 1428 | 6,408 | 405,113 |
| Cu$^{2+}$ | 117 | 485 | 33,948 |
| Fe$^{2+}$ | 92 | 382 | 29,345 |
| Fe$^{3+}$ | 217 | 1,057 | 68,829 |
| Co$^{2+}$ | 194 | 875 | 55,050 |
| Mn$^{2+}$ | 459 | 2,124 | 156,625 |
| Ca$^{2+}$ | 1237 | 6,789 | 396,957 |
| Mg$^{2+}$ | 1461 | 5,212 | 480,307 |
| Na$^+$ | 78 | 489 | 27,408 |
| K$^+$ | 57 | 535 | 18,777 |
| Acid radical ions | NO$_2^-$ | 22 | 98 | 8,144 |
| CO$_2^-$ | 62 | 316 | 22,766 |
| SO$_4^{2-}$ | 303 | 2,125 | 99,729 |
| PO$_4^{3-}$ | 339 | 2,168 | 112,279 |

<P> is the number of binding residues; <N> is the number of nonbinding residues.

**MATERIALS AND METHODS**

**Dataset**

The dataset of ion ligand–binding residues constructed in this article was from the BioLip database. On the basis of literatures (Hu et al., 2016a; Cao et al., 2017), we selected the protein chains with the length longer than 50 residues and resolution below 3Å and then removed the protein chains whose pairwise sequence identity was higher than 30% by using the CD-HIT software (Li and Godzik, 2006). Inspired by the literature (Hu et al., 2016a), we finally selected the binding residues of 10 metal ion ligands ($\text{Zn}^{2+}, \text{Cu}^{2+}, \text{Fe}^{2+}, \text{Fe}^{3+}, \text{Co}^{2+}, \text{Ca}^{2+}, \text{Mg}^{2+}, \text{Mn}^{2+}, \text{Na}^+, \text{K}^+$) and four acid radical ion ligands ($\text{NO}_2^-, \text{CO}_2^-, \text{SO}_4^{2-}, \text{PO}_4^{3-}$) as research objects.

The interaction between the protein and ion ligand was related to both binding residues and their surrounding residues. Thus, we used the sliding window method to cut the protein chain into the corresponding sequence segments. If sequence segment center was an ion ligand–binding residue, it was defined as positive segment; otherwise, it was defined as the negative segment. In order to make each amino acid residue on the protein chains appear in the center of the sequence segment, we added...
According to the description in the literatures (Sodhi et al., 2004; Hu et al., 2016a,b; Cao et al., 2017; Li et al., 2017; Cui et al., 2019) and the statistical analysis of the position conservation in metal ion ligand–binding segments and nonbinding segments done by our group (Cao et al., 2017), we selected the amino acid position conservation information as a feature parameter. Because the dimension of PSSM (20\(\times\)L) that is commonly used to extract position conservation information is excessively high, we constructed position weight matrices to extract position conservation information of amino acids (Kel et al., 2003; Gao and Hu, 2014).

The Selection and Extraction of Feature Parameters

**Amino Acid Composition and Position Conservation Information**

Our prestudy showed that amino acid composition of metal ion ligand–binding segments was different from amino acid composition of nonbinding segments (Cao et al., 2017). Besides, the literatures (Ansari and Raghava, 2010; Jiang et al., 2016; Li et al., 2017) also showed the amino acid composition was an important feature in the recognition work of ligand-binding residues. Therefore, the amino acid composition was selected as a feature parameter in this article.

### Equation 1

\[
m_{ij} = \ln\left(\frac{p_{ij}}{p_{0j}}\right)
\]

### Equation 2

\[
p_{ij} = \frac{(n_{ij} + \sqrt{N_j})}{(N_i + \sqrt{N_j})}
\]

In the above equation, \(i\) represents position, and \(j\) runs for 20 different amino acids and virtual residue "X". \(p_{0j}\) represents the background probability of the amino acid \(j\), and \(p_{ij}\) represents the probability of the amino acid \(j\) at the \(i^{th}\) position. \(n_{ij}\) represents the frequency of amino acid \(j\) at the \(i^{th}\) position. \(N_j\) is total number of amino acids at the \(i^{th}\) position. \(q\) represents the total number of categories, where \(q = 21\). \(L\) is the length of the amino acid sequence segments. Taking \(\text{Mn}^{2+}\) ligand as an example, the position weight matrices constructed for its positive and negative training samples are shown in Figures 1, 2, respectively. For arbitrary amino acid sequence...
segment, such as “ACFPQSW,” according to the corresponding position weight matrices, we can get a 14-dimensional amino acid position conservation information: \((-1.8245, -1.9236, -1.9109, -1.9971, -1.822, -1.9681, -3.808, -1.8407, -1.8452, -1.9978, -1.9378, -1.9129, -1.9378, -1.9378, -1.9378)\). Finally, the 2L-dimensional amino acid position conservation information was used as a feature parameter.

### The Predicted Structural Information

It is well-known that structure of proteins directly determines their function, but not all proteins have experimentally measured 3D structure information. Because the secondary structure, RSA, and dihedral angle can reflect the local structure information of the protein, we selected above information as basic features (Hu et al., 2016a; Cao et al., 2017). In this article, the predicted secondary structure, RSA, and dihedral angle were obtained by ANGLOR software (Yang Zhang Lab; https://zhanglab.cimb.med.umich.edu/ANGLOR/) (Wu and Zhang, 2008).

The secondary structure types that were predicted by ANGLOR software included α-helix (H), β-strand (E), and coil (C). In previous studies, its composition and position conservation information were already used to predict ion ligand–binding residues and obtained great prediction results (Hu et al., 2016a,b; Cao et al., 2017; Li et al., 2017). Therefore, composition and 2L-dimensional position conservation features were extracted from secondary structure as feature parameters.
Liu et al. Optimized Dihedral Angle

FIGURE 4 | Distribution of psi angle for binding and nonbinding residues of Mn$^{2+}$ ligand. The abscissa represents 24 classification intervals; the ordinate represents the percentage of predicted psi angle that appears in each interval.

FIGURE 5 | Polarization charge classifications of amino acids.

FIGURE 6 | Hydrophilic–hydrophobic classifications of amino acids.

For the predicted RSA, its threshold value 0.25 was usually chosen to indicate whether the residue was exposed (RSA > 0.25) or buried (RSA < 0.25). In the literature (Cao et al., 2017), the four-classification of RSA was used as the basic feature parameter, and prediction results were obviously improved. Therefore, we selected four-classification of RSA in the literature (Cao et al., 2017) as a basic feature parameter and extracted its composition and 2L-dimensional position conservation information as predicted feature parameters. The four-classification of RSA in the literature (Cao et al., 2017) was as follows:

\[
\begin{align*}
I, & \quad x \in (0, 0.2] \\
J, & \quad x \in (0.2, 0.45] \\
M, & \quad x \in (0.45, 0.6] \\
N, & \quad x \in (0.6, 0.85]
\end{align*}
\]

TABLE 2 | Prediction results without adding dihedral angle information.

| Ligand | L | Sn (%) | Sp (%) | Acc (%) | MCC |
|--------|---|--------|--------|---------|-----|
| Zn$^{2+}$ | 7 | 90.3 | 88.8 | 89.6 | 0.791 |
| Cu$^{2+}$ | 13 | 85.8 | 91.3 | 88.6 | 0.772 |
| Fe$^{2+}$ | 9 | 89.3 | 90.6 | 89.9 | 0.798 |
| Fe$^{3+}$ | 9 | 82.6 | 89.4 | 86.0 | 0.722 |
| Co$^{2+}$ | 11 | 76.8 | 86.2 | 81.5 | 0.632 |
| Mn$^{2+}$ | 7 | 77.8 | 88.3 | 82.1 | 0.644 |
| Co$^{3+}$ | 9 | 71.7 | 76.5 | 74.1 | 0.483 |
| Mg$^{2+}$ | 9 | 71.6 | 82.2 | 76.9 | 0.541 |
| Na$^+$ | 9 | 71.8 | 71.8 | 71.8 | 0.436 |
| K$^+$ | 11 | 75.7 | 64.7 | 70.2 | 0.406 |
| NO$_3^-$ | 13 | 78.6 | 73.5 | 76.0 | 0.521 |
| CO$_2^-$ | 15 | 72.2 | 76.9 | 74.5 | 0.491 |
| SO$_4^{2-}$ | 13 | 74.1 | 73.0 | 73.6 | 0.472 |
| PO$_4^{3-}$ | 13 | 76.1 | 78.2 | 77.1 | 0.543 |

The above values were calculated on downsampling negative samples.

Protein backbone dihedral angle specifies the backbone conformation of protein and is important for describing the local conformation of amino acids. Therefore, the dihedral angle information was selected as a basic feature parameter to predict the binding residues of the ion ligands. In a previous prediction work on binding sites of ligand–proteins, Chen et al. used 20 regions of the Ramachandran plot to calculate the value of propensity for ligand binding (Chen and Xu, 2016); Cui
Liu et al. used the real values of phi and psi angles to predict the ligand-binding residues on proteins (Cui et al., 2019). However, prediction results obtained by the above two methods did not achieve expectation. Therefore, the phi and psi angles of amino acid residues on protein chains binding with ion ligands were statistically analyzed and reclassified in this article. The degree of phi and psi angles predicted by ANGLOR software retained one decimal point, and the value range of phi and psi angles all was $[-180^\circ, 180^\circ]$. To simplify statistics, every 15th degree was divided into an interval; $[-180^\circ, 180^\circ]$ was divided into 24 intervals. Then, we performed statistical analysis for the phi and psi angles. Taking Mn$^{2+}$ ligand as an example, the distributions of phi and psi angles for binding and nonbinding residues are shown in Figures 3, 4, respectively.

In Figures 3, 4, the abscissa represents 24 classification intervals; the ordinate represents the percentage of predicted phi/psi angle appearing in each interval. We chose thresholds based on the difference of the predicted phi/psi angle between the binding and nonbinding residues. The threshold of phi angle was defined by the function $g(x)$, and the threshold of the psi angle was defined by the function $h(x)$. As can be seen from Figure 3, the difference of phi angle on the Mn$^{2+}$ ligand-binding residues and nonbinding residues was mainly concentrated in two different intervals. Therefore, $g(x)$ was defined as follows:

$$g(x) = \begin{cases} A, & x \in [-180^\circ, -75^\circ] \\ B, & x \in [-75^\circ, 180^\circ] \end{cases}$$

(4)

Similarly, it can be seen from Figure 4 that the difference of psi angle on the Mn$^{2+}$ ligand-binding residues and nonbinding residues was mainly concentrated in three intervals. Therefore, $h(x)$ was defined as follows:

$$h(x) = \begin{cases} A, & x \in [-180^\circ, 15^\circ] \\ B, & x \in [15^\circ, 135^\circ] \\ C, & x \in [135^\circ, 180^\circ] \end{cases}$$

(5)

In summary, the reclassification information of phi and psi angles was taken as the basic feature parameters, and its composition and the 2L-dimensional position conservation information were extracted as the feature parameters.

**Physicochemical Properties of Amino Acids**

Because the ion ligand was charged, it was easy to bind to residues with opposite charge. Therefore, we selected the polarization charge information of amino acids as a basic feature parameter. The 20 amino acids were divided into three categories in this article (Taylor, 1986), as shown in Figure 5.

Actually, ion ligand mainly interacted with the amino acids exposed on the surface of the protein pocket. These amino acids generally showed hydrophilicity, so the hydrophilic–hydrophobic property of amino acids was also selected as a basic feature parameter. The 20 amino acids were divided into six categories according to hydrophilic–hydrophobic property (Pánek et al., 2005), as shown in Figure 6.
It can be seen from Figure 5 that the number of uncharged amino acids was many times more than charged amino acids, which made the information of charged amino acids unavailable. To solve this problem, information entropy was applied to extract the polarization charge information. Similarly, in the hydrophilic–hydrophobic classification of amino acids, P, G, and C were respectively divided into one category; the number of amino acids in these categories was far less than that in the other three categories. So, information entropy was also used on the extraction of hydrophilic–hydrophobic information.

Information entropy was proposed by Shannon (1948) to describe the uncertainty of information sources. Although the signal is uncertain, it can be measured according to the probability of its occurrence. For the state space $X: \{x_1, x_2, \ldots, x_q\}$, $n_j$ is the occurrence number of $x_j$ ($j = 1, 2, \ldots, q$), and information entropy was defined as follows:

$$H(x) = -\sum_{j=1}^{q} p_j \log_2 p_j$$ (6)

$$p_j = \frac{n_j + \sqrt{N}}{N + \sqrt{N}}$$ (7)

where, $N = \sum_{j=1}^{q} n_j$, $x_j$ represents hydrophilic–hydrophobic (or polarization charge) classifications of amino acids and virtual residue “X”. $p_j$ represents the probability of the information symbol $x_j$. When $x_j$ represents hydrophilic–hydrophobic classifications of amino acids, $q = 7$, whereas when it represents polarization charge classifications of amino acids, $q = 4$. For arbitrary sequence segment, we can get the one-dimensional (1D) information entropy value according to Equation (6). Therefore, we used the 1D information entropy value as a feature.

**The RF Algorithm**

The RF algorithm is a machine learning algorithm proposed by Breiman (2001). It is composed of multiple independent decision trees, each of which is a classifier. The basic principle is to integrate weak classifiers into one strong classifier, and the final result is determined by voting. The RF algorithm has been successfully applied to the prediction of β-hairpin and protein...
fold (Jia and Hu, 2011; Chen et al., 2014; Feng and Hu, 2014). The advantage of the RF algorithm is that fewer parameters need to be adjusted (Kandaswamy et al., 2010), in comparison with other algorithms, when the dimension is relatively high, the overfitting problem of RF algorithm is not so serious. The RF algorithm has two important parameters, one is the $m$ ($m = \sqrt{M}$, and $M$ is the number of the feature initially selected), which is the size of a random feature subset for splitting the nodes; the other is the $k$ that is the number of decision trees in the RF, generally $k = 500$ (Liaw and Wiener, 2002). In this article, we established our prediction model using the 4.6–12 RF algorithm package of R software version 3.4.3 (Vienna, Austria; https://www.R-project.org/).

### The Validation Methods and Evaluation Metrics

#### The Validation Methods

The constructed model was tested by the 5-fold cross-validation and independent test, which were commonly used on ligand–binding residues predictions (Sodhi et al., 2004; Lin et al., 2005; Hu et al., 2016a,b; Jiang et al., 2016; Cao et al., 2017; Li et al., 2017).

In the 5-fold cross-validation, the dataset was randomly divided into five subsets with same size. Each subset was regarded as test dataset in turn, and the rest of the four were accordingly as training dataset. After prediction process was repeated five times, the average value of five results was adopted as the final result.

In the independent test, the protein chains binding to each ion ligand were divided into two parts: one part accounted for 80% of the total protein chains, which was the training dataset for training the model; the other part accounted for 20% of the total protein chains, which was the independent test dataset for testing the model. Because the protein chains that generated the independent test dataset and the training dataset were different, the data of independent test dataset and training dataset were independent.

#### The Evaluation Metrics

The following four measures were used to evaluate the prediction performance of ion ligand–binding residues: sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthew’s correlation coefficient (MCC) (Hu et al., 2016a,b; Lin et al., 2016; Cao et al., 2017; Li et al., 2017). These measures were defined as:

$$
Sn = \frac{TP}{TP + FN} \times 100 \%
$$

$$
Sp = \frac{TN}{TN + FP} \times 100 \%
$$

### Tables

#### TABLE 3 | Prediction results with adding dihedral angle information by 5-fold cross-validation.

| Ligand | $\mathcal{L}$ | Sn (%) | Sp (%) | Acc (%) | MCC |
|--------|--------------|--------|--------|---------|-----|
| Zn$^{2+}$ | 7 | 93.0 (99.8) | 93.2 (99.5) | 93.1 (99.7) | 0.862 (0.993) |
| Cu$^{2+}$ | 13 | 87.8 (85.5) | 93.4 (87.1) | 90.6 (86.3) | 0.814 (0.926) |
| Fe$^{2+}$ | 9 | 90.3 (91.9) | 90.1 (90.7) | 90.2 (91.3) | 0.804 (0.826) |
| Fe$^{3+}$ | 9 | 86.4 (88.9) | 92.3 (88.7) | 89.4 (87.8) | 0.789 (0.756) |
| Ca$^{2+}$ | 11 | 86.1 (80.8) | 88.2 (85.1) | 87.1 (83.0) | 0.743 (0.680) |
| Mn$^{2+}$ | 7 | 84.9 (82.1) | 89.6 (84.7) | 87.3 (83.2) | 0.747 (0.664) |
| Mg$^{2+}$ | 9 | 94.8 (71.3) | 85.5 (79.1) | 90.2 (74.8) | 0.807 (0.502) |
| Na$^+$ | 9 | 88.1 (82.2) | 76.3 (76.2) | 82.2 (78.4) | 0.649 (0.586) |
| K$^+$ | 11 | 89.3 (77.3) | 71.0 (83.2) | 80.2 (80.3) | 0.614 (0.607) |

The above values were calculated on downsampling negative samples. The values in brackets are the 5-fold cross-validation results obtained by SVM in literature (Cao et al., 2017).

#### TABLE 4 | Data of the training dataset and independent test dataset.

| Ligand | Training dataset | Independent test dataset |
|--------|------------------|-------------------------|
| | Chains | $P$ | $N$ | Chains | $P$ | $N$ |
| Zn$^{2+}$ | 1,142 | 5,145 | 321,161 | 286 | 1,263 | 83,952 |
| Cu$^{2+}$ | 93 | 377 | 27,548 | 24 | 108 | 6,400 |
| Fe$^{2+}$ | 73 | 301 | 23,824 | 19 | 81 | 5,521 |
| Fe$^{3+}$ | 173 | 859 | 54,945 | 44 | 198 | 13,884 |
| Co$^{2+}$ | 155 | 707 | 44,300 | 39 | 168 | 10,750 |
| Mn$^{2+}$ | 367 | 1,685 | 124,543 | 92 | 439 | 32,082 |
| Ca$^{2+}$ | 989 | 5,256 | 312,876 | 248 | 1,533 | 84,081 |
| Mg$^{2+}$ | 1,168 | 4,069 | 384,265 | 293 | 1,143 | 95,942 |
| Na$^+$ | 62 | 408 | 22,411 | 16 | 81 | 4,957 |
| K$^+$ | 45 | 410 | 14,882 | 12 | 125 | 3,895 |
| NO$^-$ | 17 | 76 | 6,218 | 5 | 22 | 1,926 |
| CO$_2^-$ | 49 | 252 | 18,066 | 13 | 64 | 4,700 |
| SO$_4^{2-}$ | 242 | 1,751 | 79,164 | 61 | 374 | 20,565 |
| PO$_4^{3-}$ | 271 | 1,730 | 90,786 | 68 | 438 | 21,493 |

$P$ is the number of binding residues; $N$ is the number of nonbinding residues.

#### TABLE 5 | Prediction results of the independent test.

| Ligand | Sn (%) | Sp (%) | Acc (%) | MCC |
|--------|--------|--------|---------|-----|
| Zn$^{2+}$ | 92.2 (94.1) | 90.7 (84.3) | 90.7 (84.4) | 0.326 (0.2528) |
| Cu$^{2+}$ | 88.0 (91.7) | 93.9 (82.9) | 93.8 (83.0) | 0.399 (0.2458) |
| Fe$^{2+}$ | 79.0 (90.1) | 93.7 (73.6) | 93.5 (73.9) | 0.339 (0.1708) |
| Fe$^{3+}$ | 72.7 (87.9) | 94.3 (72.7) | 94.0 (72.9) | 0.316 (0.1584) |
| Co$^{2+}$ | 75.6 (73.2) | 87.6 (82.3) | 87.4 (82.2) | 0.329 (0.1760) |
| Mn$^{2+}$ | 72.9 (76.5) | 91.9 (79.8) | 91.7 (79.8) | 0.262 (0.1599) |
| Ca$^{2+}$ | 51.1 (59.5) | 88.7 (79.2) | 88.1 (78.9) | 0.163 (0.1251) |
| Mg$^{2+}$ | 74.6 (50.2) | 81.8 (81.9) | 81.7 (81.6) | 0.150 (0.0871) |
| Na$^+$ | 54.3 (33.3) | 72.8 (78.2) | 72.5 (77.5) | 0.076 (0.0348) |
| K$^+$ | 87.2 (45.6) | 51.2 (62.8) | 52.3 (62.3) | 0.133 (0.0301) |
| NO$^-$ | 88.4 | 95.4 | 95.3 | 0.377 |
| CO$_2^-$ | 70.3 | 86.7 | 86.5 | 0.189 |
| SO$_4^{2-}$ | 50.8 | 88.4 | 87.7 | 0.158 |
| PO$_4^{3-}$ | 75.6 | 88.8 | 88.5 | 0.272 |

The values in brackets are the independent test results obtained by SVM in literature (Cao et al., 2017).
\[
\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (10)
\]
\[
\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (11)
\]

where TP represents the number of correctly predicted ion ligand–binding residues, TN represents the number of correctly predicted nonbinding residues, FP represents the number of nonbinding residues predicted as binding residues, and FN represents the number of binding residues predicted as nonbinding residues.

### RESULTS AND DISCUSSION

#### The Predicted Results of 5-Fold Cross-Validation

Because of the heavy imbalance in the original dataset, we took the number of positive segments as the standard and randomly sampled the equal number of negative segments 10 times, which generated 10 negative subsets. For each negative subset and the positive set, we extracted related feature parameters. Then, with five-fold cross-validation, we combined the feature parameters and inputted them into the RF algorithm. In this way, the progress was repeated 10 times, and the average result was taken as the result of one subset. Finally, we averaged the above results of 10 subsets as the prediction results.

#### The Prediction Results Obtained Without Adding Dihedral Angle Information

The composition and 2L-dimensional position conservation information of amino acids, secondary structure, RSA, and information entropy of polarization charge and hydrophilic–hydrophobic were used as feature parameters, and then we inputted them into the RF algorithm to predict the ion ligand–binding residues, and the predicted results obtained by the 5-fold cross-validation are shown in Table 6.

As can be seen from Table 6, comparing to the results of Ca\(^{2+}\), Mg\(^{2+}\), Na\(^{+}\), K\(^{+}\) ligands, the results of transition metal ions Zn\(^{2+}\), Cu\(^{2+}\), Fe\(^{2+}\), Fe\(^{3+}\), Co\(^{2+}\), and Mn\(^{2+}\) were better, with MCC value greater than 0.630, Acc value > 81%, Sn value > 76%, and Sp value > 86%. This is because transition metal ion binding residues have a preference for conservation. For example, Zn\(^{2+}\) ligand–binding residues prefer to use C, H, D, E, and so on; Cu\(^{2+}\) ligand–binding residues prefer to use C, H, D, E, and so on; Fe\(^{2+}\) ligand–binding residues prefer to use C, H, D, E, and so on; Fe\(^{3+}\) ligand–binding residues prefer to use C, H, D, E, and so on; Co\(^{2+}\) ligand–binding residues prefer to use C, H, D, E, and so on; Mn\(^{2+}\) ligand–binding residues prefer to use C, H, D, E, and so on.

### Table 6 | Comparison of independent test results obtained by RF, SVM, and ANN.

| Ligands | Methods | Sn (%) | Sp (%) | Acc (%) | MCC  |
|---------|---------|--------|--------|---------|------|
| Zn\(^{2+}\) | RF      | 92.2   | 90.7   | 90.7    | 0.326|
|         | SVM     | 93.4   | 88.6   | 88.7    | 0.298|
|         | ANN     | 90.4   | 86.2   | 86.2    | 0.260|
| Cu\(^{2+}\) | RF      | 88.0   | 93.9   | 93.8    | 0.399|
|          | SVM     | 96.3   | 84.4   | 84.6    | 0.275|
|          | ANN     | 87.0   | 93.0   | 92.9    | 0.371|
| Fe\(^{2+}\) | RF      | 79.0   | 93.7   | 93.5    | 0.333|
|          | SVM     | 96.3   | 80.1   | 80.3    | 0.224|
|          | ANN     | 82.7   | 91.2   | 91.1    | 0.296|
| Fe\(^{3+}\) | RF      | 72.7   | 94.3   | 94.0    | 0.316|
|          | SVM     | 83.8   | 83.6   | 83.6    | 0.210|
|          | ANN     | 78.8   | 87.2   | 87.0    | 0.225|
| Co\(^{2+}\) | RF      | 75.6   | 87.6   | 87.4    | 0.229|
|          | SVM     | 79.2   | 88.1   | 88.0    | 0.247|
|          | ANN     | 77.4   | 84.2   | 84.1    | 0.203|
| Mn\(^{2+}\) | RF      | 72.9   | 91.9   | 91.7    | 0.262|
|          | SVM     | 79.3   | 86.5   | 86.4    | 0.217|
|          | ANN     | 81.5   | 76.9   | 77.0    | 0.158|
| Ca\(^{2+}\) | RF      | 51.1   | 88.7   | 88.1    | 0.163|
|          | SVM     | 67.0   | 77.6   | 77.4    | 0.149|
|          | ANN     | 69.2   | 66.5   | 66.6    | 0.097|
| Mg\(^{2+}\) | RF      | 74.6   | 81.8   | 81.7    | 0.150|
|          | SVM     | 77.2   | 78.9   | 78.8    | 0.141|
|          | ANN     | 70.2   | 79.9   | 79.8    | 0.129|
| K\(^{+}\)   | RF      | 87.2   | 51.2   | 52.3    | 0.133|
|          | SVM     | 80.0   | 70.1   | 70.4    | 0.187|
|          | ANN     | 75.2   | 67.4   | 67.7    | 0.156|
| Na\(^{+}\)  | RF      | 54.3   | 72.8   | 72.5    | 0.076|
|          | SVM     | 64.2   | 67.1   | 67.1    | 0.083|
|          | ANN     | 44.4   | 80.4   | 79.8    | 0.078|
| NO\(_3\)^{−} | RF     | 86.4   | 95.4   | 95.3    | 0.377|
|           | SVM     | 59.1   | 96.2   | 95.8    | 0.284|
|           | ANN     | 86.4   | 77.9   | 78.0    | 0.162|
| CO\(_3\)^{2−} | RF     | 70.3   | 86.7   | 86.5    | 0.189|
|           | SVM     | 76.6   | 83.9   | 83.8    | 0.188|
|           | ANN     | 95.3   | 65.8   | 66.2    | 0.147|
| SO\(_3\)^{2−} | RF     | 50.8   | 88.4   | 87.7    | 0.158|
|           | SVM     | 83.4   | 62.2   | 62.6    | 0.124|
|           | ANN     | 95.5   | 25.1   | 26.4    | 0.063|
| PO\(_4\)^{3−} | RF     | 75.6   | 88.8   | 88.5    | 0.272|
|           | SVM     | 88.1   | 72.0   | 72.4    | 0.185|
|           | ANN     | 93.2   | 75.7   | 76.1    | 0.221
residues prefer to use H, C, E, and so on (Cao et al., 2017). It was visualized in Figure 7.

The Prediction Results Obtained by Adding Dihedral Angle Information

The dihedral angle has an important influence on local structure of the protein backbone. The literatures (Chen and Xu, 2016; Hu et al., 2016a,b; Cui et al., 2019) showed that the dihedral angle played an important role in predicting the ligand binding sites on proteins. Therefore, composition feature and 2L-dimensional position conservation feature extracted from the phi and the psi angles were added to predict the binding residues of ion ligands. The flowchart of the proposed method was displayed in Figure 8. And the results obtained by the five-fold cross-validation are shown in Table 3.

As shown in Table 3, prediction results were improved after adding dihedral angle information; the MCC values of Ca2+, Na+, K+, SO4 2−, and PO4 3− ligands increased by at least 20 percentage points; the MCC values of Mn2+, Co2+, Mg2+, and CO3 2− increased by at least 10 percentage points; and the values of Acc, Sn, and Sp of these nine ion ligands were also significantly improved. The above values of other ion ligands were also slightly improved. The results showed that the ion ligand–binding residues were sensitive to the information of the reclassified dihedral angle. Namely, the information of the reclassified dihedral angle was effective for identifying ion ligand–binding residues.

The Comparison of the Predicted Results of 10 Metal Ion–Ligand Binding Residues

Because the datasets of metal ion ligand–binding residues used in this article and the optimal window length selected for each metal ion were the same as used in literature (Cao et al., 2017), the prediction results of 10 metal ion ligands obtained in this article were compared with prediction results obtained by SVM in literature (Cao et al., 2017).

In order to facilitate comparison, prediction results obtained by SVM were putted into brackets in Table 3. Apparently, the predicted results of Zn2+, Fe2+, and Cu2+ were lower than those obtained by SVM, but prediction results of the other seven metal ion ligands obtained by RF algorithm were all more accurate than those obtained by SVM. In particular, the MCC values of Ca2+ and Mg2+ were increased more than 20 percentage points comparing to SVM.

The Prediction Results of Independent Test

In order to test the practicability of the model established in this article, independent test was performed. The independent test datasets of 14 ion ligands are shown in Table 4.

In the independent test, composition and 2L-dimensional position conservation information of the amino acids, secondary structure, RSA, and dihedral angle, as well as information entropy of polarization charge and hydrophilic–hydrophobic, were used as feature parameters and inputted into the RF algorithm to predict ion ligand–binding residues. The independent test results are shown in Table 5.

The predicted results of Zn2+, Cu2+, Fe2+, Fe3+, Co2+, Mn2+, NO2 −, and PO4 3− were improved, MCC values were higher than 0.220, the values of Acc and Sp were higher than 87%, and the Sn values were higher than 72%. The result of Cu2+ ligand was the highest, in which the MCC value was 0.399, and the values of Sn, Sp, and Acc were higher than 87%. The obtained results of Ca2+, Mg2+, Na+, K+, CO3 2−, and SO4 2− ligands were lower, the MCC values were lower than 0.200, in which the predicted result of Na+ was the lowest, with the MCC value of 0.076 and the Acc value of 73.5%.

Because the independent test datasets of 10 metal ion ligands constructed in this article were the same as those used in the literature (Cao et al., 2017), the results obtained by independent test in this article were compared with those obtained by SVM in the literature (Cao et al., 2017). The results of the independent test in the literature (Cao et al., 2017) are shown in the brackets of Table 5.

Comparing MCC value, it is obvious that the independent test results obtained by the RF algorithm were better than those obtained by the SVM. In terms of values of the Acc and Sp, the RF algorithm achieved better predicted results of metal ion ligands, except for Na+ and K+ ligands. As for the Sn value, the prediction results of Co2+, Mg2+, Na+, and K+ ligands were better than those obtained by SVM, whereas other ligands were slightly lower. In general, the model constructed in this article was practical on the metal ion ligand–binding residues prediction.

Comparison With SVM and ANN Algorithms

It is objective to compare our proposed methods with previous model using the same dataset. From Tables 3, 5, we found that most of the prediction results of this work were better than those of previous work (Cao et al., 2017).

In order to test the performance of our proposed method, we further made a comparison between RF and SVM and ANN algorithms on the same dataset, feature parameters, classification strategy, and evaluation methods. We inputted the

| Ligand | Model | Sn (%) | Sp (%) | Acc (%) | MCC |
|--------|-------|--------|--------|---------|-----|
| Zn2+  | ①    | 92.9   | 93.5   | 93.2    | 0.865|
| Cu2+  | ①    | 87.8   | 93.4   | 90.6    | 0.814|
| Fe2+  | ②    | 90.3   | 90.1   | 90.2    | 0.804|
| Fe3+  | ②    | 87.3   | 92.1   | 89.7    | 0.796|
| Co2+  | ③    | 87.2   | 91.0   | 89.1    | 0.782|
| Mn2+  | ③    | 86.3   | 90.9   | 88.6    | 0.773|
| Ca2+  | ①    | 94.8   | 85.5   | 90.2    | 0.807|
| Mg2+  | ①    | 88.5   | 86.3   | 87.4    | 0.749|
| K+    | ③    | 90.3   | 75.9   | 83.1    | 0.669|
| Na+   | ③    | 88.1   | 76.3   | 82.2    | 0.649|
| NO2 − | ③    | 75.5   | 78.6   | 77.0    | 0.541|
| CO3 2−| ③    | 81.3   | 84.8   | 83.1    | 0.662|
| SO4 2−| ③    | 94.0   | 90.3   | 92.1    | 0.843|
| PO4 3−| ③    | 92.3   | 92.5   | 92.4    | 0.849|
same feature parameters extracted in this article into SVM and ANN algorithms to identify ion ligand–binding residues. The results obtained by the independent test are shown in Table 6. As seen, the prediction results of RF algorithm were best, although the results of Na\(^+\) and K\(^+\) ligands were slightly lower. At the same time, we can see that, when using the SVM on the same dataset, the prediction results obtained by selecting the feature parameters of this article were also better than those obtained in literature (Cao et al., 2017).

The Optimization of Model

Because the reclassified phi and psi angles have a good effect on the prediction of ion ligand–binding residues, we further classified the phi and psi angles in the other way to optimize the prediction model for each ion ligand. At first, we classified the phi and psi angles by distribution. Therefore, we tried to classify the phi and the psi angles according to peak value. Taking Mn\(^{2+}\) ligand as an example, according to Figure 3, phi angle was divided into four categories and defined by function \(f(x)\):

\[
f(x) = \begin{cases} 
A, x \in [-180°, -105°] \\
B, x \in (-105°, -75°] \\
C, x \in (-75°, -60°] \\
D, x \in (-60°, 180°]
\end{cases}
\]

According to Figure 4, psi angle was divided into three categories and defined by function \(p(x)\):

\[
p(x) = \begin{cases} 
A, x \in [-180°, -15°] \\
B, x \in (-15°, 75°] \\
C, x \in (75°, 180°]
\end{cases}
\]

The two classifications of phi angle and the two classifications of psi angle were combined, respectively (Figure 9). Their composition and 2L-dimensional position conservation information were extracted as the characteristic parameters.

The composition and 2L-dimensional position conservation information of amino acids, secondary structure, RSA, dihedral angle, and information entropy of polarization charge and hydrophilic–hydrophobic were used as characteristic parameters and inputted into the RF algorithm to predict ion ligand–binding residues. The optimal prediction model and the corresponding prediction result of each ion ligand obtained by the five-fold cross-validation are shown in Table 7.

As seen, not all ion ligand–binding residues were sensitive to the first combination way of phi and psi angles. For example, the optimal prediction model of Mn\(^{2+}\) ligand corresponded to the second combination way of phi and psi angles, and the optimal prediction model of Zn\(^{2+}\) ligand corresponded to the third combination way of phi and psi angles. It can be seen that the different reclassification and combination ways of phi and psi angles have an important impact on the prediction of ion ligand–binding residues. Therefore, when using the reclassified dihedral angle information to predict ion ligand–binding residues, different classifications and combination ways of dihedral angle (phi and psi angles) should be considered to optimize the prediction model of binding residues of each ion ligand.

CONCLUSION

Many proteins perform their functions by interacting with ion ligands. To illustrate the protein functions, it is a significant work to recognize the ion ligand–binding residues. In this article, based on optimized dihedral angle, we predicted the 14 ion ligand–binding residues from the BioLip database by RF algorithm and obtained improved results. During the progress, the dihedral angle information was statistically analyzed and reclassified. Besides, the new extraction methods of feature parameter were proposed, in which the position weight matrices were constructed to extract the 2L-dimensional position conservation features; the polarization charge information and hydrophilic–hydrophobic information of amino acids were extracted by using information entropy. These changes in extraction methods improved the predicted results of ion ligand–binding residues. In particular, the reclassification information of the dihedral angle has significantly improved the prediction results of the ion ligand–binding residues, indicating that the reclassification information of the dihedral angle is an important feature parameter for the identification of the ion ligand–binding residues. By classifying and combining phi and psi angles, we optimized the prediction model for each ion ligand–binding residue. Thus, with different classification standards and combined methods of the dihedral angle (phi and psi angles), it can further improve the prediction results of ion ligand–binding residues.

DATA AVAILABILITY STATEMENT

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

LL performed the experiments and wrote the paper. XH designed the experiments and analyzed the results. ZF, SW, SX, and KS gave guidance on the writing of the paper. All authors read and approved the final manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (61961032), Natural Science Foundation of the Inner Mongolia of China (2019BS03025).
REFERENCES

Al-Mugotir, M., Kolar, C., Vance, K., Kelly, D. L., Natarajan, A., and Borgstahl, G. E. O. (2019). A simple fluorescent assay for the discovery of protein-protein interaction inhibitors. Anal. Biochem. 569, 46–52. doi: 10.1016/j.ab.2019.01.010

Ansari, H. R., and Raghava, G. P. S. (2010). Identification of NAD interacting residues in proteins. BMC Bioinformatics 11:160. doi: 10.1186/1471-2105-11-160

Breiman, L. (2001). Random forests. Mach. Learn. 45, 5–32. doi: 10.1023/A:1010933404324

Cao, X., Hu, X., Zhang, X., Gao, S., Ding, C., Feng, Y., et al. (2017). Identification of metal ion binding sites based on amino acid sequences. PLoS ONE 12:e0183756. doi: 10.1371/journal.pone.0183756

Chen, C., and Xu, S. (2016). Improving the performance of the PLB index for ligand-binding site prediction using dihedral angles and the solvent-accessible surface. Sci. Rep. 6:33232. doi: 10.1038/srep33232

Chen, P., Huang, J., and Gao, X. (2014). LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone. BMC Bioinformatics 15(Suppl.15):S4. doi: 10.1186/1471-2105-15-S15-S4

Cui, Y., Dong, Q., Hong, D., and Wang, X. (2019). Predicting protein-ligand binding residues with deep convolutional neural networks. BMC Bioinformatics 20:93. doi: 10.1186/s12859-019-2672-1

Ebert, J. C., and Altman, R. B. (2008). Robust recognition of zinc binding sites in enzymes by using ensemble classifier based on scoring SVM. Hans J. Comput. Biol. 4, 1–11. doi: 10.12677/HJCB.2014.41001

Hu, X., Dong, Q., Yang, J., and Zhang, Y. (2016a). Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transfe rals. Bioinformatics 32, 3260–3269. doi: 10.1093/bioinformatics/btw396

Hu, X., Wang, X., and Dong, Q. (2016b). Protein ligand-specific binding residue predictions by an ensemble classifier. BMC Bioinformatics 17:470. doi: 10.1186/s12859-016-1348-3

Jia, S., and Hu, X. (2011). Using random forest algorithm to predict β-hairpin motifs. Protein Pept. Lett. 18, 609–617. doi:10.2174/092966011795222777

Jiang, Z., Hu, X., Gerslétu, G., Xing, H., and Cao, X. (2016). Identification of Ca(2+)-binding residues of a protein from its primary sequence. Genet. Mol. Res. 15, 1–13. doi:10.4238/gmr.15027618

Kandaswamy, K. K., Chou, K. C., Martinetz, T., Möller, S., Suganthan, P. N., Sridharan, S., et al. (2010). AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J. Theor. Biol. 270, 56–62. doi: 10.1016/j.tib.2010.03.037

Kel, A. E., Goßling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003). MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 31, 3576–3579. doi: 10.1093/nar/gkg585

Komiyama, Y., Banno, M., Ueki, K., Saad, G., and Shimizu, K. (2015). Automatic generation of bioinformatics tools for predicting protein–ligand binding sites. BMC Bioinformatics 32, 901–907. doi:10.1093/bioinformatics/btv593

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, S., Hu, X., Sun, L., and Zhang, X. (2017). “Identifying the sulfate ion binding residues in proteins,” in International Conference on Biomedical and Biological Engineering (Guilin). 4, 209–216. doi: 10.2991/ibbe-17.2017.34

Liaw, A., and Wiener, M. (2002). Classification and regression by random forest. R News 2, 18–22.

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

Lin, H., Han, L., Zhang, H., Zheng, C., Xie, B., Cao, Z., et al. (2006). Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. BMC Bioinformatics 7(Suppl.5):S13. doi: 10.1186/1471-2105-7-S5-S13

Lin, Y. F., Cheng, C. W., Shih, C. S., Hwang, J. K., Yu, C. S., and Lu, C. H. (2016). MIIB: metal ion-binding site prediction and docking server. J. Chem. Inf. Model. 56, 2287–2291. doi: 10.1021/acs.jcim.6b00407

Liu, L., Hu, X., Feng, Z., Zhang, X., Wang, S., Xu, S., et al. (2019). Prediction of acid radical ion binding residues by K-nearest neighbors classifier. BMC Mol. Cell Biol. 20:52. doi: 10.1186/s12860-019-02338-8

Pan, X., and Shen, H. B. (2018). Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. Bioinformatics 34, 3427–3436. doi: 10.1093/bioinformatics/bty364

Panek, J., Eidhammer, I., and Aasland, R. (2005). A new method for identification of protein (sub)structures in a set of proteins based on hydropathy distribution in proteins. Proteins 58, 923–934. doi: 10.1002/prot.20356

Robin, OZ, Siriam, K. K., and Westerlund, F. (2019). A nanofluidic device for real-time visualization of DNA-protein interactions on the single DNA molecule level. Nanoscale 11, 2071–2078. doi: 10.1039/C8NR09023H

Shannon, C. E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. 27, 623–656. doi:10.1002/j.1538-7305.1948.tb00917.x

Sodhi, J. S., Bryson, K., McGuffin, L. J., Ward, J. J., Wernisch, L., and Jones, D. T. (2004). Predicting metal-binding site residues in low-resolution structural models. J. Mol. Biol. 342, 307–320. doi: 10.1016/j.jmb.2004.07.019

Taylor, W. R. (1986). The classification of amino acid conservation. J. Theor. Biol. 119, 205–218. doi: 10.1016/S0022-5193(86)80073-3

Wang, S., Hu, X., Feng, Z., Zhang, X., Liu, L., Sun, K., et al. (2019). Recognizing ligand-binding sites by SMO algorithm. BMC Mol. Cell Biol. 20:52. doi: 10.1186/s12860-019-0237-9

Wu, S., and Zhang, Y. (2008). ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. PLoS ONE 3:e3400. doi: 10.1371/journal.pone.0003400

Yang, J., Roy, A., and Zhang, Y. (2013). Biolip: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res. 41, D1096–D1103. doi: 10.1093/nar/gkt966

Yang, J., Roy, A., and Zhang, Y. (2018). Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 29, 2588–2595. doi: 10.1093/bioinformatics/bty144

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Liu, Hu, Feng, Wang, Sun and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.