Genetics and Molecular Research

Identification of Ca^{2+}-binding residues of a protein from its primary sequence

Z. Jiang, X.Z. Hu, G. Geriletu, H.R. Xing and X.Y. Cao

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

Corresponding author: X.Z. Hu
E-mail: hxz@imut.edu.cn

Genet. Mol. Res. 15 (2): gmr.15027618
Received September 10, 2015
Accepted December 29, 2015
Published May 20, 2016
DOI http://dx.doi.org/10.4238/gmr.15027618

ABSTRACT. Calcium is one of the most abundant minerals in the human body, playing a critical role in many cellular activities by interacting with different calcium ion (Ca^{2+})-binding proteins. Therefore, the correct identification of Ca^{2+}-binding residues is essential for protein functional research. In this study, a new method was developed to predict Ca^{2+}-binding residues from the primary sequence without using three-dimensional information. Through statistical analysis, four kinds of feature parameters were extracted from amino acid sequences: the increment of diversity values of amino acid composition, the matrix scoring values of position conservation, the autocross covariance of physicochemical properties, and the center motif. These features served as input for a support vector machine to predict Ca^{2+}-binding residues. This method was tested on four well-established datasets using a five-fold cross-validation. The accuracies and Matthews correlation coefficients were 75.9% and 0.53 (dataset 1), 79.2% and 0.58 (dataset 2), 77.4% and 0.55 (dataset 3), and 79.1% and 0.58 (dataset 4). Comparative results show that the developed method outperforms previous methods. Based on this study, a web server was developed for predicting Ca^{2+}-binding

residues from any protein sequence, being publically available at http://202.207.29.245/.

Key words: Calcium-binding residues; Increment of diversity; Matrix scoring algorithm; Autocross covariance; Support vector machine

INTRODUCTION

Due to the unprecedented increase of the amount of proteomic data, the annotation of protein function has become a major task in the "post-genome era" (Navarro et al., 2003). In order to carry out their functions, many proteins must bind with their corresponding ligands. Therefore, the identification of ligand binding residues is an important step for protein-function research (Kirberger et al., 2010). Experimental methods to detect binding residues are costly and time-consuming, thus, it is essential to find a rapid theoretical method to identify the ligand binding residues in proteins.

Calcium is an essential element in many vital activities of cellular life. For example, in the regulation of muscle contraction, the binding of calcium ions (Ca^{2+}) to troponin C of the troponin complex leads to a series of conformational changes in the members of the thin and thick filaments (Herzberg et al., 1986). At the same time, Ca^{2+} is the critical component in signal transduction pathways and a switch in ion channels. TMEM16A, a member of the transmembrane protein 16 family, activates chloride channels after binding with Ca^{2+} and affects many important physiological processes, such as electrolyte secretion and smooth muscle excitability (Caputo et al., 2008). Therefore, the identification of Ca^{2+}-binding residues or the prediction of the possible binding residues in proteins will provide assistance for biological mechanism research and drug design.

During the last few decades, many methods have been developed to accurately identify Ca^{2+}-binding sites in proteins. Yamashita et al. (1990) proposed the hydrophobicity contrast function based on the spatial structure of protein to locate the metal binding site in proteins. They were able to calculate the deviation between the predicted site and the real binding site for six calcium binding proteins in PDB (Rose et al., 2015). In 2006, Deng et al. (2006) developed the graphtheory-based and geometry-based approach to detect calcium-binding sites, and got a sensitivity of about 90% for 123 calcium binding proteins. Lu et al. (2012) predicted the metal ion-binding site in proteins by the fragment transformation method. By using the method, they compared 273 known calcium binding proteins with 407 binding templates and achieved 94.1% accuracy with 48.9% true positive rate (Lu et al., 2012). Roy and Zhang (2012) recognized protein-ligand binding site by global structural alignment and local geometry refinement using low-resolution protein structural models, which obtained a good result in the blind test in CASP9 (Schmidt et al., 2011).

The above researchers have shown that a high accuracy can be obtained when using three-dimensional structures. For most proteins, however, we only have the primary structure determined by sequencing technology but no three-dimensional structures. Thus, the identification of calcium-binding residues in a protein from its primary structure is quite important. Metal binding residues were identified in proteins from primary structure by using artificial neural network in Lin's research in 2005.They extracted biological features from amino acid sequence in 2589 Ca^{2+}-binding protein chains of protein set and 1106 chains of

enzyme set. The Ca^{2+}-binding residues were identified by their method with higher than 90% Accuracy under 5-fold cross validation (Lin et al., 2005). In the study by Horst and Samudrala (2010), calcium binding residues were predicted from amino acid sequence by using meta-functional signature. The binding residues were predicted by a regression model derived from five feature parameters including sequence conservation, evolution conservation, amino acid type, neighbor conservation, and physicochemical conservation. Specific training for Ca^{2+}-binding results in 83% area under the receiver operator characteristic curve measured by 10-fold cross validation for parallel sets of 336 and 299 protein chains (Horst and Samudrala, 2010).

In this paper, an improved method was proposed to identify the calcium binding residues in proteins from primary sequence. Several features were extracted from the amino acid sequences including the increment of diversity (ID) values of amino acid composition, the matrix scoring (S) values of position conservation, the correlation of the physicochemical properties of residues within the local window calculated by autocross covariance (AC) transformation, and the frequency of the center motif (C). As the best of our knowledge, this is the first time to introduce the autocross covariance and center motif for binding residues prediction of proteins. A support vector machine (SVM) model was established to identify the Ca^{2+}-binding residues in proteins. The experiment was performed on four well-established datasets. First, a new Ca^{2+}-binding proteins dataset (dataset 1) was built with sequence identity below 25% and resolution less than 3Å after filtering from the PDB database. To further test the proposed method, the Ca^{2+}-binding proteins dataset (dataset 2), which was developed by Singh et al. (2012), was downloaded from ccPDB. For comparing with previous research, we also got the two additional datasets (dataset 3, dataset 4) used in the study by Horst and Samudrala (2010). Comprehensive experiment results show that the performance of the proposed method can be gradually improved by the incremental addition of different features.

MATERIAL AND METHODS

Datasets

A new dataset (dataset 1) is constructed for Ca^{2+}-binding residues prediction. First, a dataset of 16,712 proteins with less than 95% sequence identity was downloaded from ASTRAL1.75 of SCOP database. Then a subset of the above dataset containing 4442 protein chains was obtained with sequence identity below 25% and resolution less than 3Å.The length of these protein chains vary from 100 to 1332 residues. The calcium binding residues in these protein chains were annotated by Ligand-Protein Contacts (LPC), and finally, the dataset contains 277 calcium binding protein chains including 1801 Ca^{2+}-binding residues. To further verify the proposed method, the dataset (dataset 2) generated by Singh et al. (2012) was downloaded, which were compiled from Protein Data Bank (PDB) including 973 chains and 3791 binding residues with sequence identity below 25% and resolution less than 3Å. The Ca^{2+}-binding residues in this dataset were also annotated by LPC. To compare with the previous research, two additional datasets (dataset 3, dataset 4) were gained, which were used in the study by Horst and Samudrala (2010). The two parallel datasets (dataset 3, dataset 4) contained 336 and 299 Ca^{2+}-binding chains respectively with sequence identity below 35% and the resolution less than 2.1Å. The Ca^{2+}-binding residues were not given by the researchers; therefore, we annotate the Ca^{2+}-binding residues by LPC for dataset 3 and dataset 4.

Selection of feature parameters

A residue in a sequence binding with a calcium ion is not determined only by the amino acid residue itself but also affects by neighboring residues. Thus, overlapping segments centered at the target residue were generated with different window sizes ranging from 5 to 21 for every Ca^{2+}-binding protein sequence (Chauhan et al., 2009). If the central residue of the segment was a calcium binding residue, then the segment was assigned as binding segment, otherwise it was assigned as non-binding segment. On the basis of calculation results of different window sizes, the optimal window size was selected as 17.

Increment of diversity (ID) value

The diversity information in amino acid sequence can be quantitatively described. Unlike the measure of information in Shannon's theory, the measure of diversity described the overall diversity. Thus, we extracted the feature information from sequence by using Increment of diversity (ID) algorithm and used the ID values as feature parameters for our method.

Increment of diversity (ID) algorithm was a classifier which has been successfully used in identification of protein folds and complex secondary structures in recent years (Laxton, 1978; Li and Lu, 2001; Hu et al., 2010). In the state space of s dimension, the measure of diversity source $X\{n_1, n_2,\ldots, n_s\}$ was:

$$D(X) = N\log_b N - \sum_{i=1}^{S} n_i \log_b n_i \quad \text{(Equation 1)}$$

Here

$$N = \sum_{i=1}^{S} n_i \quad \text{(Equation 2)}$$

For two state space of s dimension, $X\{n_1, n_2,\ldots, n_s\}$ and $Y\{m_1, m_2,\ldots, m_s\}$, the measure of mixed diversity source $X+Y$ was:

$$D(X + Y) = (N + M) \log_b (N + M) - \sum_{i=1}^{S} (n_i + m_i) \log_b (n_i + m_i) \quad \text{(Equation 3)}$$

The increment of diversity between the source of diversity X and Y were:

$$ID(X,Y) = D(X + Y) - D(X) - D(Y) \quad \text{(Equation 4)}$$

Amino acid composition was commonly used in identification of ligand binding residues as kind of important feature information (Ansari and Raghava, 2010). Therefore, we analyzed the amino acid composition in Ca^{2+}-binding segments and non-Ca^{2+}-binding segments in dataset 1. As shown in Figure 1, there was a significant difference between binding and non-binding segments, the aspartic, glutamic, asparagine and isoleucine had larger Ca^{2+}-binding propensities in comparison with other amino acids. Thus, the amino acid composition was selected as the feature information for ID algorithm. The amino acid composition of a segment was a state space of 20 dimensions, for any arbitrary sequence segment two ID values were obtained as feature parameters from Ca^{2+}-binding segments and non-Ca^{2+}-binding segments respectively.

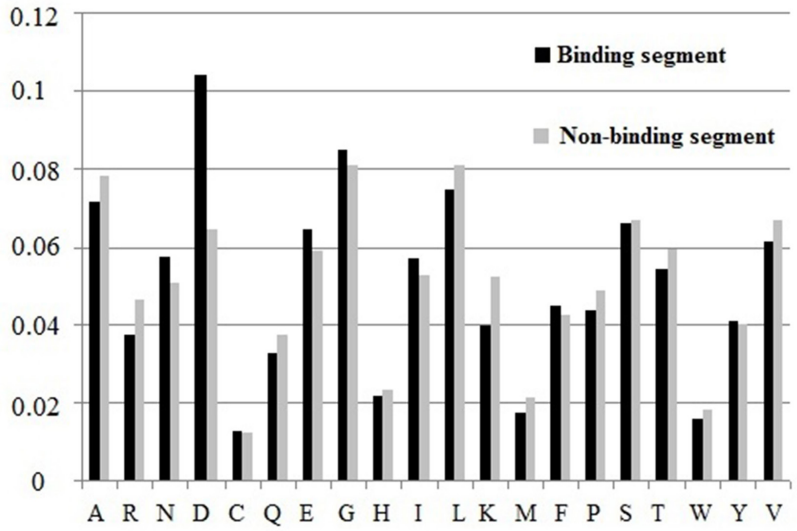

Figure 1. Statistical analysis of amino acid composition in Ca^{2+}-binding and non-Ca^{2+}-binding segments. The letters in the x-axis are the one-letter abbreviation of the natural 20 amino acids and the y-axis represents the average composition of each amino acid in binding and non-binding segments.

Matrix scoring(S) value

Matrix scoring algorithm is a classification algorithm that has been successfully used in the prediction of transcription factor binding sites in genomes and super-secondary structure (Kielbasa et al., 2005; Long and Hu, 2012). Matrix scoring (S) value was given by the following equation:

$$S = \frac{\sum_{i=1}^{L} (m_{i,j} - m_{i,min})}{\sum_{i=1}^{L} (m_{i,max} - m_{i,min})} \quad \text{(Equation 5)}$$

Here:

$$m_{i,j} = \log_e \left(\frac{p_{i,j}}{p_{0,j}}\right) \quad \text{(Equation 6)}$$

$$p_{i,j} = \left(\frac{n_{i,j} + \frac{\sqrt{N_i}}{20}}{N_i + \sqrt{N_i}}\right) \quad \text{(Equation 7)}$$

In the above equation, m_{ij} is j^{th} amino acids weight probabilities at i^{th} position, and $m_{i,max}$, $m_{i,min}$ are the maximum and minimum value at i^{th} position. L is the length of amino acid sequence. p_{ij} is the observed probability of j^{th} amino acids at i^{th} position, and $p_{0,j}$ is background probability of the j^{th} amino acid, respectively. N_i is total number of all amino acids occur at i^{th} position, $n_{i,j}$ is th frequency of the j^{th} amino acids at i^{th} position.

The Matrix scoring (S) value contained the probability of occurrence of each type of amino acid at each position. In the previous research of ligand binding identification, position-specific scoring matrix (PSSM) was commonly used as a measure of residue conversation in a given location. However, for a segment with length of L, the PSSM matrix was a feature parameter with high dimension of 20xL, which may contain much noise as demonstrated by other studies (Yu et al., 2014). Thus, the Matrix scoring (S) algorithm was used in this paper to extract the position conservation of amino acid residues from segments, which has a low dimension feature parameter.

The position conservation of Ca^{2+}-binding segments and non-binding segments was analyzed in dataset 1 by using WEBLOGO software (Schneider and Stephens, 1990). Figure 2 show the example of position conservation with window size of 21. As shown in panel a, there was significant binding preference at the 11th position in Ca^{2+}-binding segments where calcium ions preferred to bind with residue D, E and N. On the other hand, there was no preference at the same position in the non-binding segments in panel b. Besides that, there were also obvious binding differences between binding and non-binding segments at the 4th, 7th, 9th, 15th position. Thus, the position conservation of amino acid residues was selected as the feature information for Matrix scoring (S) algorithm. Based on the training set, the standard position weight matrix was constructed. For arbitrary sequence segment, two Matrix scoring (S) values were obtained as feature parameters.

Autocross covariance (AC) value

The interaction between amino acid residues could influence the formation of Ca^{2+}-binding region. The autocross covariance (AC) transformation is employed to calculate the interaction of residues along the sequences (Wold et al., 1993). The autocross covariance (AC) has been successfully adopted by many researchers for the prediction of protein folds, proteins interaction predictions (Deng et al., 2009; Feng and Hu, 2014). As we know, this is the first time to bring AC into the identification of ligand binding residues. The AC transformation is defined as follows:

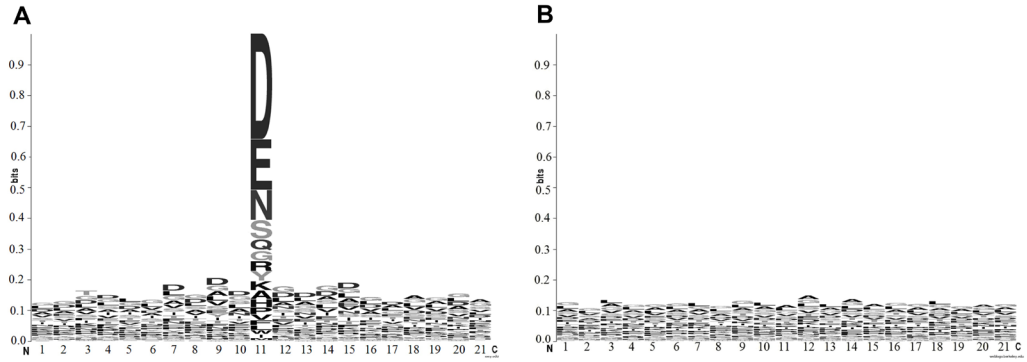

Figure 2. Position conservation of amino acid residues in Ca^{2+}-binding and non-Ca^{2+}-binding segments: The x-axis represents 21 positions in Ca^{2+}-binding and non-binding segments and the y-axis represents the conservation of amino acid in every position per residue type, with the height of each letter corresponding to the occurrence probability of the corresponding amino acid (according to one-letter abbreviations) in that position.

$$AC_{j,d} = \frac{1}{N-d}\sum_{i=1}^{N-d}(P_{j,i} - \frac{1}{N}\sum_{i=1}^{N}P_{j,i}) \times (P_{j,(i+d)} - \frac{1}{N}\sum_{i=1}^{N}P_{j,(i+d)}) \quad \text{(Equation 8)}$$

Here, d is the distance between two residues, N is the length of each segment, j represents j^{th} property of amino acid. $P_{j,i}$, $P_{j,(i+d)}$ are the j^{th} property of the amino acid at i^{th} position and $(i+d)^{th}$ position, respectively. Based on the definition, the AC values measured the correlation of the same property along the amino acid residue sequences.

The previous study showed that the ligand binding region was influenced by physicochemical property of amino acid residues (Zhou et al., 2013). Thus, we selected hydrophobicity P_1, hydrophilic P_2 and polarity P_3 as the input properties of residues for AC transformation. The values of hydrophobicity, hydrophilicity, and polarity for 20 native amino acids are shown in Table 1 (Kawashima and Kanehisa, 2000). Since the window size of each segment is 17 in this study, the distance index d goes from 1 to 16, and every segment got 16 AC values for each physicochemical property, and totally 48 AC values are extracted as feature parameters.

Table 1. Values of hydrophobicity, hydrophilicity, and polarity for 20 amino acids.

Amino acid	P_1	P_2	P_3	Amino acid	P_1	P_2	P_3
Ala	0.25	-0.5	8.1	Met	0.26	-1.3	5.7
Cys	0.04	-1	5.5	Asn	-0.64	2	11.6
Asp	-0.72	3	13	Pro	-0.07	0	8
Glu	-0.62	3	12.3	Gln	-0.69	0.2	10.5
Phe	0.61	-2.5	5.2	Arg	-1.76	3	10.5
Gly	0.16	0	9	Ser	-0.26	0.3	9.2
His	-0.4	-0.5	10.4	Thr	-0.18	-0.4	8.6
Ile	0.73	-1.8	5.2	Val	0.54	-1.5	5.9
Lys	-1.1	3	11.3	Trp	0.37	-3.4	5.4
Leu	0.53	-1.8	4.9	Tyr	0.02	-2.3	6.2

Center motif (C)

Taking the microenvironment of Ca^{2+}-binding residues into consideration, the three contiguous amino acid residues in the center of every Ca^{2+}-binding segments, named as the center motif, were statistically analyzed in dataset 1. According to the statistical results, we defined the center motifs, which occurred 5 or more times as "Preferred Motifs", occurred 2-4 times as "Normal Motif", occurred 1 time as "Rare Motif". Therefore, each sequence segment was assigned with a 3-dimension vector. For any query segment, if its center motif belonged to Preferred Motifs, then this segment was assigned a feature vector as "1 0 0", Normal Motif as "0 1 0", Rare Motif as "0 0 1", and if its center motif did not exist in our statistical results, than the query segment would be assigned as "0 0 0".

Support vector machine (SVM)

Support vector machine (SVM) is a machine learning algorithm proposed by Vapnik, which has a good performance on classification of small samples based on the principles of structural risk minimization (Cortes and Vapnik, 1995). We established our identification model by using libsvm-3.17 package based on C-SVC and radial basis function (RBF), *c* and *g* were set to the default value (Chang and Lin, 2011). There would be an overfitting problem in the training process, when dimension of input vector is too high. Thus, we reduced the dimension of input vector by using ID algorithm and Matrix scoring algorithm to advance the learning ability and generalization ability of SVM.

The 5-fold cross validation and evaluation metrics

The proposed method was tested by 5-fold cross validation, which was commonly used in the prediction of ligand binding residues (Chauhan et al., 2010). The dataset was randomly divided into five sets. One set was used for testing and the remaining four sets were used for training. This process was repeated five times in such a way that each set was used once for testing. The final performance was obtained by averaging the performance of five sets.

Several evaluation metrics are used to evaluate the proposed method. Acc is the percentage of correctly identified calcium binding and non-binding residues. MCC is Matthews's Correlation Coefficient, a measure of the quality and balance of binary classification. $S_{n(Ca)}$, $S_{n(non\text{-}Ca)}$ are the sensitivity of Ca^{2+}-binding and non-binding residues. $S_{p(Ca)}$ and $S_{p(non\text{-}Ca)}$ are the specificity of Ca^{2+}-binding and non-binding residues.

$$Acc = \frac{(TP + TN)}{TP + TN + FP + FN} \times 100\% \quad \text{(Equation 9)}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad \text{(Equation 10)}$$

$$S_{n(Ca)} = \frac{TP}{TP + FN} \times 100\% \quad \text{(Equation 11)}$$

$$S_{p(Ca)} = \frac{TP}{TP + FP} \times 100\% \quad \text{(Equation 12)}$$

$$S_{n(non\text{-}Ca)} = \frac{TN}{TN + FP} \times 100\% \quad \text{(Equation 13)}$$

$$S_{p(non\text{-}Ca)} = \frac{TN}{TN + FN} \times 100\% \quad \text{(Equation 14)}$$

Where TP is the number of correctly identified Ca^{2+}-binding residues, TN is the number of correctly identified non-binding residues, FP is the number of non-binding residues identified as binding residues and FN is the number of binding residues wrongly identified as non-binding residues.

RESULTS AND DISCUSSION

Comparison results with different features

In order to study the effects of the individual feature on the identification of calcium binding residues, the feature parameters were gradually added to the input vectors for SVM. Based on five cross-validation on dataset 1, when only one feature parameter, ID values of amino acid composition, was used, An Acc of 62.4% and MCC of 0.25 was obtained (Table 2). After adding the feature parameter of matrix scoring values of position conservation, the identification result has been significantly improved with Acc value of 74.5% and MCC value of 0.49,which showed that position conservation has a great effect on the identification of Ca^{2+}-binding residues. When the AC values of physicochemical property was added, the performance was further improved where the Acc was increased to 75.1% and MCC was 0.50. Finally, the center motif was added after ID values, matrix scoring values and AC values, the best result was obtained with Acc value of 75.9% and MCC value of 0.53. The above results show that, the position conservation of the binding residues and the correlation of the adjacent residues are the key factors of ligand binding identification.

Table 2. Performance of the proposed method on dataset 1 by using different features.

Features	$S_{n(Ca)}$	$S_{n(non\text{-}Ca)}$	$S_{p(Ca)}$	$S_{p(non\text{-}Ca)}$	Acc	MCC
ID	68.1%	56.7%	61.1%	64.0%	62.4%	0.25
ID+S	70.1%	78.9%	76.8%	72.5%	74.5%	0.49
ID+S+AC	73.7%	76.4%	75.7%	74.4%	75.1%	0.50
ID+S+AC+C	74.9%	77.3%	76.2%	75.3%	75.9%	0.53

Results in different datasets

The proposed method for the Ca^{2+}-binding residues prediction was further tested on dataset 2, dataset 3 and dataset 4 with the same combination of feature parameters. The 5-fold cross validation result showed that best performance was achieved on dataset 2 (Table 3) with

Acc of 79.2% and MCC of 0.58. On the dataset 3 and dataset 4, the Acc and MCC value were 77.4% and 0.548 (Table 4), 79.1% and 0.581 (Table 5), respectively. The result on the three datasets showed the same growth trend of feature combination, which verified the universality and the stability of the proposed method.

Table 3. Performance of the proposed method on dataset 2 by using different features.

Parameter	$S_{n(Ca)}$	$S_{n(non\text{-}Ca)}$	$S_{p(Ca)}$	$S_{p(non\text{-}Ca)}$	Acc	MCC
ID	69.7%	58.6%	58.9%	62.9%	66.1%	0.29
ID+S	75.5%	80.6%	79.6%	76.7%	78.1%	0.56
ID+S+AC	78.0%	78.8%	78.6%	78.2%	78.4%	0.57
ID+S+AC+C	79.5%	78.9%	79.0%	79.4%	79.2%	0.58

Table 4. Performance of the proposed method on dataset 3 by using different features.

Parameter	$S_{n(Ca)}$	$S_{n(non\text{-}Ca)}$	$S_{p(Ca)}$	$S_{p(non\text{-}Ca)}$	Acc	MCC
ID	67.2%	58.3%	61.7%	64.0%	62.8%	0.25
ID+S	70.0%	80.4%	78.2%	72.9%	75.2%	0.50
ID+S+AC	74.0%	77.9%	77.0%	75.0%	75.9%	0.51
ID+S+AC+C	75.5%	79.3%	78.5%	76.4%	77.4%	0.55

Table 5. Performance of the proposed method on dataset 4 by using different features.

Parameter	$S_{n(Ca)}$	$S_{n(non\text{-}Ca)}$	$S_{p(Ca)}$	$S_{p(non\text{-}Ca)}$	Acc	MCC
ID	70.3%	58.6%	62.9%	66.4%	64.4%	0.29
ID+S	72.3%	82.1%	80.2%	74.8%	77.2%	0.54
ID+S+AC	76.5%	79.5%	78.9%	77.2%	78.0%	0.56
ID+S+AC+C	78.1%	80%	79.6%	78.5%	79.1%	0.58

Comparison with previous research

The proposed method was compared with the method developed by Horst and Samudrala (2010) on the same datasets (dataset 3, dataset 4). The 10-fold cross validation was adopted to identify the Ca^{2+}-binding residues as used by Horst and Samudrala (2010). They reported the best performance with area under the receiver operator characteristic curve (ROC) of 83% measured by 10-fold cross validation for parallel sets of 336 (dataset 3) and 299 (dataset 4) protein chains. The proposed method obtained the Acc of 76.4% and ROC value of 83.6% on dataset 3 (Figure 3A), and Acc of 77.3% and ROC value of 84.8% on dataset 4 (Figure 3B). The comparison of ROC value showed that the proposed method outperforms the method of Horst and Samudrala (2010).

In the research by Horst and Samudrala (2010), their method was independently validated on dataset 3 and dataset 4 respectively, that is, one dataset is used for training, the other dataset is used for test. The proposed method is also independently validated in the same way and the results are shown in Figure 4. The ROC in dotted line were achieved by our method and the solid line is the identification result in study by Horst and Samudrala (2010) in which As shown in the figure, the performance of the proposed method was better than their method.

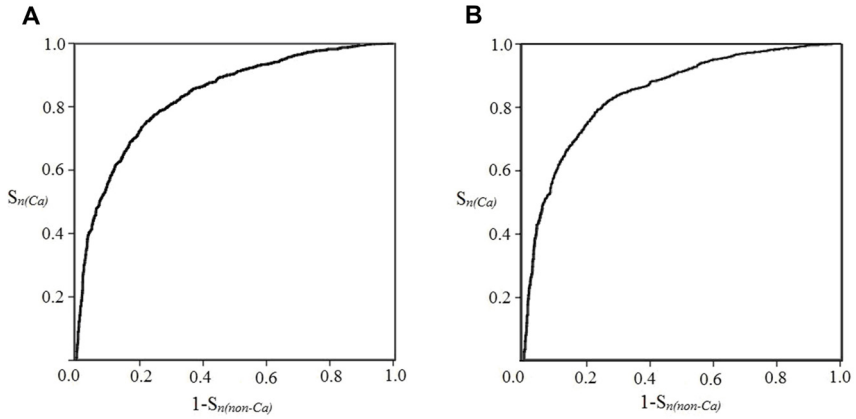

Figure 3. Receiver operating characteristic (ROC) curve for the identification of Ca^{2+}-binding residues using the ten-fold cross validation on dataset 3 (**A**) (ROC value of 83.6%) and dataset 4 (**B**) (ROC value of 84.8%).

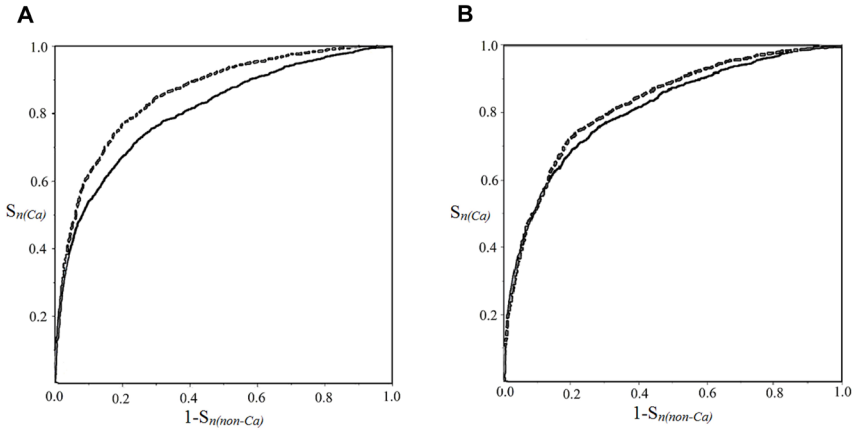

Figure 4. Receiver operating characteristic (ROC) for identification of Ca^{2+}-binding residues using an independent validation on dataset 3 (**A**) and dataset 4 (**B**).

Web server

Based on the above study, a web server has been developed for predicting Ca^{2+}-binding residues in a protein, and is freely available at http://202.207.29.245/. Users can simply submit protein sequence in the form of FASTA to obtain the prediction result of Ca^{2+}-binding residues by our method.

CONCLUSION

A large amount of protein sequence data has been produced in recent years through sequencing technology or DNA translation. However, most of the proteins do not have a determined three-dimensional structure. Thus, the identification of the Ca^{2+}-binding residues

in proteins from its primary structure is very important. In this study, we developed a method to identify the Ca^{2+}-binding residues in proteins from amino acid sequence. A SVM classification model has been built by using ID values of amino acid composition, matrix scoring values of position conservation, AC values of physicochemical and the center motif. We obtained a good identification result and increased tendency on dataset 1 by adding the feature parameter combination, especially the matrix scoring values of position conservation has a great influence on the performance. To the best of our knowledge, this study was the first to introduce AC transformation into the identification of ligand binding residues for extracting information on the physicochemical property from amino acid sequences, which can improve the accuracy and balance the sensitivity and specificity effectively. The features of ID values and Matrix scoring values were also used for the first time in the identification of Ca^{2+}-binding residues, which can significantly reduce the dimension of feature parameters. Another Ca^{2+}-binding protein dataset (dataset 2) was used to further examine the proposed method. A good performance and same increasing tendency of feature combination are achieved. Based on the same datasets and 10-fold cross validation policy, the proposed method was compared with previous research. Our method obtained a better identification result based on ROC evaluation. A publically available web server has been built to identify the calcium binding residues in a protein, which will provide certain help for related research. For the further research, we will combine our method with the spatial structure for the better identification of calcium binding residues.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (#30960090, #31260203), the “CHUN HUI” Plan of Ministry of Education, and the Talent Development Foundation of Inner Mongolia.

REFERENCES

Ansari HR and Raghava GP (2010). Identification of NAD interacting residues in proteins. *BMC Bioinformatics* 11: 160. http://dx.doi.org/10.1186/1471-2105-11-160

Caputo A, Caci E, Ferrera L, Pedemonte N, et al. (2008). TMEM16A, a membrane protein associated with calcium-dependent chloride channel activity. *Science* 322: 590-594. http://dx.doi.org/10.1126/science.1163518

Chauhan JS, Mishra NK and Raghava GP (2009). Identification of ATP binding residues of a protein from its primary sequence. *BMC Bioinformatics* 10: 434. http://dx.doi.org/10.1186/1471-2105-10-434

Cortes C and Vapnik V (1995). Support-vector networks. *Mach. Learn.* 20: 273-297. http://dx.doi.org/10.1007/BF00994018

Chang CC and Lin CJ (2011). LIBSVM: a library for support vector machines. *ACM TIST* 2: 27.

Chauhan JS, Mishra NK and Raghava GP (2010). Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. *BMC Bioinformatics* 11: 301. http://dx.doi.org/10.1186/1471-2105-11-301

Deng H, Chen G, Yang W and Yang JJ (2006). Predicting calcium-binding sites in proteins - a graph theory and geometry approach. *Proteins* 64: 34-42. http://dx.doi.org/10.1002/prot.20973

Deng L, Guan J, Dong Q and Zhou S (2009). Prediction of protein-protein interaction sites using an ensemble method. *BMC Bioinformatics* 10: 426. http://dx.doi.org/10.1186/1471-2105-10-426

Feng Z and Hu X (2014). Recognition of 27-class protein folds by adding the interaction of segments and motif information.

BioMed Res. Int. 2014: 262850. http://dx.doi.org/10.1155/2014/262850

Herzberg O, Moult J and James MN (1986). A model for the Ca^{2+}-induced conformational transition of troponin C. A trigger for muscle contraction. *J. Biol. Chem.* 261: 2638-2644.

Horst JA and Samudrala R (2010). A protein sequence meta-functional signature for calcium binding residue prediction. *Pattern Recognit. Lett.* 31: 2103-2112. http://dx.doi.org/10.1016/j.patrec.2010.04.012

Hu XZ, Li QZ and Wang CL (2010). Recognition of β-hairpin motifs in proteins by using the composite vector. *Amino Acids* 38: 915-921. http://dx.doi.org/10.1007/s00726-009-0299-7

Kawashima S and Kanehisa M (2000). AAindex: amino acid index database. *Nucleic Acids Res.* 28: 374. http://dx.doi.org/10.1093/nar/28.1.374

Kielbasa SM, Gonze D and Herzel H (2005). Measuring similarities between transcription factor binding sites. *BMC Bioinformatics* 6: 237. http://dx.doi.org/10.1186/1471-2105-6-237

Kirberger M, Wang X, Zhao K, Tang S, et al. (2010). Integration of Diverse Research Methods to Analyze and Engineer Ca-Binding Proteins: From Prediction to Production. *Curr. Bioinform.* 5: 68-80. http://dx.doi.org/10.2174/157489310790596358

Laxton RR (1978). The measure of diversity. *J. Theor. Biol.* 70: 51-67. http://dx.doi.org/10.1016/0022-5193(78)90302-8

Li QZ and Lu ZQ (2001). The prediction of the structural class of protein: application of the measure of diversity. *J. Theor. Biol.* 213: 493-502. http://dx.doi.org/10.1006/jtbi.2001.2441

Lin CT, Lin KL, Yang CH, Chung IF, et al. (2005). Protein metal binding residue prediction based on neural networks. *Int. J. Neural Syst.* 15: 71-84. http://dx.doi.org/10.1142/S0129065705000116

Long HX and Hu XZ (2012). Prediction β-hairpin motifs in enzyme protein using three methods. In 2012 Eighth International Conference on Natural Computation (ICNC). Chongqing, 570-574.

Lu CH, Lin YF, Lin JJ and Yu CS (2012). Prediction of metal ion-binding sites in proteins using the fragment transformation method. *PLoS One* 7: e39252. http://dx.doi.org/10.1371/journal.pone.0039252

Navarro JD, Niranjan V, Peri S, Jonnalagadda CK, et al. (2003). From biological databases to platforms for biomedical discovery. *Trends Biotechnol.* 21: 263-268. http://dx.doi.org/10.1016/S0167-7799(03)00108-2

Rose PW, Prlić A, Bi C, Bluhm WF, et al. (2015). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.* 43: D345-D356. http://dx.doi.org/10.1093/nar/gku1214

Roy A and Zhang Y (2012). Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure* 20: 987-997. http://dx.doi.org/10.1016/j.str.2012.03.009

Singh H, Chauhan JS, Gromiha MM and Raghava GP; Open Source Drug Discovery Consortium (2012). ccPDB: compilation and creation of data sets from Protein Data Bank. *Nucleic Acids Res.* 40: D486-D489. http://dx.doi.org/10.1093/nar/gkr1150

Schneider TD and Stephens RM (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18: 6097-6100. http://dx.doi.org/10.1093/nar/18.20.6097

Schmidt T, Haas J, Gallo Cassarino T and Schwede T (2011). Assessment of ligand-binding residue predictions in CASP9. *Proteins* 79 (Suppl 10): 126-136. http://dx.doi.org/10.1002/prot.23174

Wold S, Jonsson J, Sjörström M, Sandberg M, et al. (1993). DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Anal. Chim. Acta* 277: 239-253. http://dx.doi.org/10.1016/0003-2670(93)80437-P

Yamashita MM, Wesson L, Eisenman G and Eisenberg D (1990). Where metal ions bind in proteins. *Proc. Natl. Acad. Sci. USA* 87: 5648-5652. http://dx.doi.org/10.1073/pnas.87.15.5648

Yu DJ, Hu J, Yan H, Yang XB, et al. (2014). Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble. *BMC Bioinformatics* 15: 297. http://dx.doi.org/10.1186/1471-2105-15-297

Zhou Y, Xue S and Yang JJ (2013). Calciomics: integrative studies of Ca^{2+}-binding proteins and their interactomes in biological systems. *Metallomics* 5: 29-42. http://dx.doi.org/10.1039/C2MT20009K