Identifying the Sulfate Ion Binding Residues in Proteins

Shao-bo LI, Xiu-zhen HU^{a,*}, Li-xia SUN and Xiao-jin ZHANG

College of Sciences, Inner Mongolia University of Technology, China

^ahxz@imut.edu.cn

*Corresponding author

Keywords: Sulfate Ion Ligand, Binding Residue, Support Vector Machine.

Abstract. Many proteins function execution depends on the process of protein and ligand interact with each other. The identification of ligand binding residues is important for the research of the protein function. The 4442 protein chains with <25% sequence identity and resolution <3.0 Å were analyzed using Ligand Protein Contact database. Our final dataset contained 8112 sulfate ion binding residues (SIBR). We did a statistical analysis on window size as 7 amino acids. Using the amino acid composition, hydropathy information, correlation information and predicted structure information as the characteristic parameter, a Support Vector Machine algorithm for identifying sulphate ion binding residues was proposed. The overall accuracy and Matthew's correlation coefficient achieved 78.5% and 0.571 using the 5 fold cross validation. The Acc and MCC achieved 72.7% and 0.455 by using independent test. In addition, an online web server was established. <http://202.207.30.72:7321/>

Introduction

The purpose of research protein is to understand and annotate the function of protein. In recent years, it is a hot issue to research protein function by using theoretical calculation methods in bioinformatics field. The sulphate ion binding site is a basic form for protein functional annotation. Therefore, the study of sulphate ion binding residue (SIBR) is important [1, 2].

The sulphate ion is an important matter in biological cells, and it takes part in several major processes of cell metabolism [3-5]: the synthesis process of Cysteine, the sulfation process after protein translation, the synthesis process of proteoglycan, the sulfate absorption and decomposition process of plant and others. For example, the sulfation process occurs in the Golgi apparatus, and the sulfate radical is transferred into hydroxyl of tyrosine in peptide chain by catalyzing of enzyme, so it can make many proteins to have certain activity [5]. The identification of SIBR is helpful to study protein cell metabolism processes, and it has very important meaning to design medicine molecules [6].

In 1966, Pardee has researched sulfate-binding protein of salmonella typhimurium by using experimental method observing the interactional mechanism of binding sites with sulfate ion [7]. Richard et al. have tested sulphate ion binding site of proteoglycan in 2002, and they fixed the sites that is interaction with heparan sulfate [8]. In 2003, Tamada has researched sulfation of protein, and modified sulfated silk fibroin by concentrated sulfuric acid. They got silk fibroin sulfated efficiency, and found that the reaction efficiency is not high and is very susceptible to decomposition of the sample [9]. Monigatti et al. have done a statistical analysis on Tyrosine which may be sulfated in 5 proteins, and given the analysis regularity of Tyrosine sulfation in 2002[4]. The previous study of sulphate ion binding site used experimental method or statistical analysis on few proteins. At present, in the post-genomic era, with more

and more amino acid sequences is rapidly growing, as well as the databases of ligand binding sites are existed. It has become possible to identify ligand binding residues in protein by using the theoretical calculation method.

In this paper, we constructed a non-redundant dataset which contained 8112 SIBRs, and used Support Vector Machine (SVM) algorithm to identify the SIBRs. We also established a friendly interface online web identifying server to facilitate researchers. (<http://202.207.30.72:7321/>).

Materials and Methods

The Datasets

A dataset of 16,712 protein chains with <95% sequence identity was downloaded from ASTRAL 1.75(<http://scop.mrc-lmb.cam.ac.uk/scop/>). Blastclust software [10] and PDB were used for obtaining 4442 protein chains with <25% sequence identity and resolution <3.0 Å. We obtained 1251 protein chains which at least contained one sulphate ion binding site by LPC database[11]. It total contained 8112 SIBRs among 2527 sulphate ion binding sites.

The Statistical Analysis of Sequence Segment

Reference to the previous' research for ligand binding residues [12-13]. In this paper, we identified SIBRs also using "sliding window" method. It is adopted a fixed-window size to slid from N-terminal to C-terminal in every sequence, then generate overlap segments. If the central residue of segment was a SIBR, we assigned the segment as positive; otherwise it was assigned as negative. When SIBR located in sequence terminal, we append letter "X" at both terminals to ensure the window sizes. In order to optimize window size, we generated window size are 5, 7, 9, 11, 13, 15 and 17 residues, respectively. Among, the 7 amino acid residues were select as the best window size.

We analyzed the amino acid position conservation of 17 window size by using the WEBLOGO server [14]. It is showed in Fig.1.

Figure1. The position conservative of amino acid in the positive and negative

It can be found from Fig.1 (A) that preference residues of SIBRs are R, K, S, G, E, H, T, D and N in the ninth position.

Fig.1 (A) and Fig.1 (B) showed that it has strong position conservative of in positive and negative set. But in the area of 6-12th position, it has a significant difference of amino acid conservative between positive set and negative set. For example, at the sixth position, the highest frequency of the amino acid is G, L, A, V in positive set; the highest frequency of the amino acid in negative set is L, A, G, V. In the ninth position, the highest frequency of amino acid is R, K, S, G in positive set; the highest frequency is L, A, G, V in negative set.

At the same time, it can be found that in the area of position 6-12th, it has a significant conservative of amino acid Hydropathy. Base on Panek's scale [15], G is a separate class, R, K, E, H are strongly hydrophilic, S, T are weak hydrophilicity. L and A are strong hydrophobic.

Above analysis further proved the rationality of selection 7 amino acid residues as best window size.

We analyzed the composition of amino acid, the statistical results are showed in Fig. 2. From Fig. 2 we can see that the percentage of amino acid composition is obvious different between positive set and negative set. For example, the G, H, R, S, and T in positive set are much higher than the negative set. While the A, D, E, L, and V in the negative set is much higher.

Figure2. Amino acid composition of positive set and negative set

The Selection of Characteristic Parameters

Amino Acid Composition (A).

By the analysis of Fig.2, an amino acid composition is extracted as characteristic parameter. Amino acid composition is represented by a vector of 20 dimensions which are the frequency of amino acid.

Amino Acid Hydropathy Information(S)

We can see from Fig.1 that positive set and negative set have a strong position conservative of amino acid hydropathy, thus the hydropathy component of position is extracted as characteristic parameter. The classification of hydropathy property of 20 amino acids is showed in Table1 [15].

Table 1. Hydropathy property of 20 amino acids

Classification	Amino acids	Classification	Amino acids
Strongly hydrophilic or polar	R, D, E, N, Q, K, H	Proline	P
Strongly hydrophobic	L, I, V, A, M, F	Glycine	G
Weakly hydrophilic or Weakly hydrophobic	S, T, Y, W	Cysteine	C

To avoid over fitting phenomenon we used the scoring matrix function to reduce the dimension. The scoring matrix method can effectively reflect the conservation of the amino acid position. By using hydropathy component of position as basic parameters of scoring matrix function, the values of scoring matrix were as the characteristic parameters. The scoring matrix function [16] can be defined as:

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

The matrix element of position weight matrix is: $m_{i,j} = \log\left(\frac{p_{i,j}}{p_{0,j}}\right)$, the matrix element of position probability matrix is:

$$p_{i,j} = \frac{(n_{i,j} + \sqrt{N_i}/2)}{(N_i + \sqrt{N_i})} \quad (2)$$

Here j is 20 amino acids; n_{ij} is the number of amino acid j at the position i ; N_i is the number of amino acids at the position i ; p_{ij} is the position probability of amino acid j at the position i ; p_{0j} is the background probability; m_{ij} is the matrix element of position weight matrix of amino acid j at the position i ; $m_{i,\min}$ and $m_{i,\max}$ are the maximal and minimal values of position weight at the position i , respectively.

We constructed a standard scoring matrix from the training set, for every testing segment, two scoring matrix values can be obtained.

Correlation Information of Amino Acids (G)

To consider amino acid residues correlation in segment, the frequency of amino acid dipeptide component [17] is used as feature. The description of correlation information of amino acids denoted by AnB ($n=0, 1, 2$). A and B represent 20 amino acids, n represents the gap of amino acids in segment.

The increment of diversity [18] (ID) is an effective identification algorithm. In this work, the value of ID was used for characteristic parameters.

In general, for two sources of diversity in the same space of "s" dimensions $X: \{n_1, n_2, \dots, n_s\}$ and $Y: \{m_1, m_2, \dots, m_s\}$, the increment of diversity is defined by:

$$\begin{aligned} ID(X, Y) &= D(X + Y) - D(X) - D(Y) \\ &= (M + N) \log(M + N) - \sum_i (m_i + n_i) \log(m_i + n_i) - M \log M - N \log N + \sum_i m_i \log m_i + \sum_i n_i \log n_i \end{aligned} \quad (3)$$

The 2 ID values can be obtained from each correlation information. In this way, when $n = 0$ can be obtained two-dimensional vectors, similarly, $n = 1$ and $n = 2$ can be also obtained two-dimensional vectors.

From Fig.1, we can see that the three amino acids in the center of segment have strong position conservative. Thus we also analyzed the triplet component of amino acid in segment center, the most 50 conservative triplets are selected as standard kernel, seen Table 2. It is represented by a two-dimensional vector. If the center position of a segment contains the standard kernel, it is expressed by a vector "1, 0", otherwise it is expressed by a vector "0, 1".

Predicted Secondary Structure Information (SS)

We also extract predicted secondary structure (SS) information as characteristic parameters. The predicted secondary structure are obtained by using the PSIPRED software[19], and they are represented by a vector of 3 dimensions which are the frequency of predicted secondary structure α -helix, β -sheet and coil.

Table2. The most 50 conservative triplets

<i>GKS</i>	<i>LRR</i>	<i>LRE</i>	<i>ERL</i>	<i>LRL</i>	<i>GKT</i>	<i>VRL</i>	<i>KST</i>	<i>LRP</i>	<i>GRG</i>
LSG	TGG	LRT	GRV	ERA	SRK	LRS	SRG	TRS	VRR
GKG	AGG	GRF	PSG	SGT	IRD	TRG	IRA	GTG	GRL
LKG	TKG	LKE	LKK	KRG	IEL	GSG	ISG	DRL	NRI
LRA	VGK	LRI	PRL	VRE	SRL	ATG	GGT	HHH	ARL

Support Vector Machine Algorithm

The SVM algorithm has been widely used for prediction of protein structure and function [1, 2, 20]. SVM has been compiled into software packages. In this work, we used the libsvm-3.0 package (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>), and the radial kernel function is used. Inputting the characteristic parameters into SVM for training set, it will be scaled. And then, through the grid search the optimal C and gamma values were obtained. Lastly, a classifier was constructed to identify testing set.

Performance Measure

We use the following standard measures which are widely used in ligand binding residues research [12, 13]: sensitivity (S_n), specificity (S_p), Matthew's correlation coefficient (MCC), overall accuracy of identification (Acc), calculating formula as follow:

$$S_n = \frac{TZ}{TZ + FF} \times 100\% \quad (4)$$

$$S_p = \frac{TZ}{TZ + FZ} \times 100\% \quad (5)$$

$$MCC = \frac{(TZ \times TF) - (FZ \times FF)}{\sqrt{(TZ + FZ)(TZ + FF)(TF + FZ)(TF + FF)}} \quad (6)$$

$$Acc = \frac{(TZ + TF)}{TZ + FZ + TF + FF} \times 100\% \quad (7)$$

Here TZ and TF denote the number of correctly identified SIBR and non- SIBR, respectively. FF denotes the number of the SIBR that are identified as non- SIBR; FZ denotes the number of non- SIBR that are identified as SIBR.

In general, it is used 5-fold cross validation test to evaluation the performance of identification ligand binding residue. We also used the 5-fold cross validation test. In addition, the independent test can reflect the differences between theory and the practice, in independent test, the sequences in the training set are not occurs in the training set. We used independent test for the first time.

Results and Discussion

The Results Using 5-fold Cross-validations Test

For the positive set contained 8112 segments of window size 7, 174185 segments in negative set. Because the segments of positive set are far less than negative sets and the proportion is 1:21. We referred to the identification of ligand binding residues [12-13]. Equal numbers of negative set were randomly select as an identification negative set.

We used the amino acid composition, hydropathy information, correlation information and predicted structure information as the characteristic parameter, input

in SVM algorithm for identifying sulphate ion binding residues. The identifying result is showed in Table 3 ($c=16$, $g=0.03125$).

Table3. The identify results of SIBR using the 5-fold cross-validation

Parameters (dimension)	Sn(%)	Sp(%)	MCC	Acc(%)
A (20)	60.3	63.8	0.261	63.0
A+S (22)	66.1	67.7	0.346	67.3
A+S+G (30)	74.8	74.5	0.493	74.5
A+S+G+SS (33)	79.7	77.9	0.571	78.5

The Results Using Independent Test

The dataset contains 1251 protein chains. 1000 protein chains which contain segments were as training set, the remaining 251 protein chains contain segments were as independent testing set.

The training set included positive set segment 6076, and negative set segment 130465; testing set included 2036 positive set segments, and 43720 negative set segments.

In this section, we used amino acid composition, hydropathy information, correlation information and predicted structure information together as characteristic parameters, and randomly selected identification negative set as equal as positive segment from negative set. The results of SIBR are shown in Table 4. The Acc and MCC values of were 72.7% and 0.455.

Table4. The identifying results using independent test ($c=32$ $g=0.001656$)

Sn(%)	Sp (%)	MCC	Acc(%)
74.7	70.8	0.455	72.7

The identifying result using 5-fold cross-validations test is showed in table 3, when the 20-dimension amino acid component (A) is as the characteristic parameter input into the SVM. The Acc was only 63.0%, and the MCC was 0.261. When adding the 2-dimension amino acid hydropathy information(S), the identifying effect was batter, and the values of Acc and MCC can achieve 67.3% and 0.346. With the adding of 8-dimension correlation information of amino acids (G), the identifying effect further improved. The Acc and MCC can achieve 74.5% and 0.493, respectively. The sensitivity and specificity have also increased greatly, and the value is more than 74.4%. After that we further added the 3-dimension predicted secondary structure (SS), the best identifying results were obtained by using the composite vector A+S+G+SS. The Acc and MCC can achieve 78.5% and 0.571, and the value of sensitivity and specificity is more than 77.3%. Apparently, it is found that the identifying results will be improved with the adding of effective feature information.

Because the independent test is stricter, the effect was a little bit worse. However, it is closer to the reality.

Conclusion

SVM is a convex optimization problem, thus a local optimal solution is the global optimal solution. It is an effective method to identify the small sample. SVM algorithm can effectively syncretize useful parameters. SVM algorithm will produce over fitting phenomenon when the characteristic parameters of higher dimension are

used. In this paper, we used the increment of diversity and the scoring matrix to reduce the dimension in the characteristic parameter.

Acknowledgement

This work was supported by Natural Science Foundation of China (31260203), Science Foundation of Inner Mongolia at China (Grant No. 2016MS0378) and the "CHUN HUI" Plan of Ministry of Education

References

- [1] H. Xiuzhen, D. Qiwen, Y. Jianyi, Z. Yang. Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transfers. *Bioinformatics*. 32(2016)3260-3269.
- [2] H. Xiuzhen, W. Kai, D. Qiwen. Protein ligand-specific binding residue predictions by an ensemble classifier. *BMC Bioinformatics*. 17 (2016)470 DOI 10.1186/s12859-016-1348-3.
- [3] L. Thomas and M. Michael. Cloning of a cDNA encoding ATP sulfurylase form *Arabidopsis thaliana* by functional expression in *Saccharomyces cerevisiae*. *Plant Physiol*. 105 (1994) 897-902.
- [4] F. Monigatti, E. Gasteiger, A. Bairoch and E. Jung. The sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics*. 18 (2002)769-770.
- [5] T. Y. Lee, W. C. Chang, H. D. Huang and P. C. Hsu. A comprehensive resource for integrating and displaying protein post-translational modifications. *Comput. Chem*. 30 (2009)2526-2537.
- [6] A. T. Laurie and R. M. Jackson. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr Protein Pept*. 7 (2006)395-406.
- [7] Pardee. Purification and Properties of a Sulfate-binding Protein from *Salmonella typhimurium*. Received for publication. *J Biol Chem*. 24 (1966)5886-5892.
- [8] G. W. Richard, H. Christoph, W. Stefan and J. E. Turnbull. Identification of the Heparan Sulfate Binding Sites in the Cellular Prion Protein. *J Biol Chem*. 277 (2002)18421-18430.
- [9] Y. Tamada. Sulfation of silk fibroin by sulfuric acid and anticoagulant activity. *Appl Polym Sci*. 87 (2003)2377-2382.
- [10] F. Stephen, Altschul and L. Thomas. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25 (1997) 3389-3402.
- [11] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola and M. Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*. 15 (1999)327-332.
- [12] J. Chauhan, N. Mishra and G. Raghava. Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. *BMC Bioinformatics*. 11(2010)301-308.

- [13] J. Chauhan, N. Mishra and G. Raghava. Identification of ATP binding residues of a protein from its primary sequence. *BMC Bioinformatics*.10 (2009)434 doi:10.1186/1471-2105-10-434.
- [14] G. E. Crooks, G. Hon and J. M. Chandonia. WebLogo: A sequence logo generator. *Genome Research*.14 (2004)1188-1190.
- [15] J. Panek, I. Eidhammer and R. Aasland. A new method for identification of protein (sub) families in a set of proteins based on hydropathy distribution in proteins. *PROTEINS: Structure, Function, Bioinformatics*.58 (2005)923-934.
- [16] A. E. Kel, E. GoBling and I. Reuter. MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*.13 (2003)3576-3579.
- [17] F. M. Li and Q. Z. Li. Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. *Amino Acids*.34 (2008)119-125.
- [18] Q. Z. Li and Z. Q. Lu. The Prediction of the Structural Class of Protein: Application of the Measure of Diversity. *Theory Biol*.213 (2001)493-502.
- [19] M. Kumar, M. Bhasin and N. K. Natt. BhairPred: prediction of β -hairpins in a protein from multiple alignment information using ANN and SVM techniques., *Nucleic Acids Research*.33(2005)154-159.
- [20] X. Z. Hu and Q. Z. Li. Using support vector machine to predict β -turns and γ -turns in proteins. *Comput Chem*.29 (2008)1867-1875.