Journal of Theoretical Biology 267 (2010) 312–318

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Predicting the state of cysteines based on sequence information

Xuanmin Guang, Yanzhi Guo, Jiamin Xiao, Xia Wang, Jing Sun, Wenjia Xiong, Menglong Li *

College of Chemistry, Sichuan University, Chengdu 610064, PR China

ARTICLE INFO

Article history:
Received 19 February 2010
Received in revised form
16 August 2010
Accepted 1 September 2010
Available online 6 September 2010

Keywords:
Evolution information
Annotation information
Support vector machine
F-score function.

ABSTRACT

A three-stage support vector machine (SVM) was constructed to predict the state of cysteines by fusing sequence information, evolution information and annotation information of protein sequences. The first and second stages were for predicting whether the protein sequences contain disulfide bonds and whether all of the cysteines are involved in disulfide bonds. In the last stage, one SVM was constructed for predicting which cysteines are involved in disulfide bonds, among all these cysteines in proteins. The three SVMs give a good performance and the overall prediction accuracy are 90.05%, 96.36% and 80.00%, respectively, which indicates that the features selected in this work are effective for predicting the state of cysteines. In addition, current methods only paid too much attention to the prediction performance and never showed us how much important the roles of these features played in the prediction. As a result a feature importance measurement designated as *F*-score function was used to evaluate these features. The result shows that among these protein descriptors; evolution information is the most important feature for representing the disulfide-containing proteins. The prediction software and data sets used in this article are freely available at http://cic.scu.edu.cn/bioinformatics/Predict_Cys.zip.

1. Introduction

Cysteines constitute a special class of amino acids among the amino acid residues. The covalent bonds formed by the oxidation cysteines, known as disulfide bridges, can stabilize protein spatial conformation and ensure that protein will perform its biochemical function (Wittrup, 1995). The formation of disulfide bonds between the correct pairs of cysteine residues is a crucial step in the folding pathway (Creighton, 1993, 1995). The capability of predicting the state of cysteines in proteins will be helpful in the study of protein's stability and function (Hogg, 2003; Vieille and Zeikus, 2001). Although we can know its state through crystallographic techniques by producing high-resolution three-dimensional structures of proteins, it is relatively time consuming and expensive. So it is necessary to develop an easy and reliable machine-learning method to predict its states based on protein sequences.

A series of computation approaches have been developed to predict the state of cysteines. These researches broadly occurred along two lines. The first is based on the analysis of the statistical frequency of amino acid residues neighboring the cysteines (Fiser and Simon, 2000; Fiser et al., 1992). The second approach combines the information of the local environment with the machine-learning methods in supervised settings. Muskal et al. (1990) got 80% accuracy by using neural networks and a slide window. Fariselli et al. (1999) constructed a neural network module based on evolutionary information and obtained prediction accuracy of 81%. Fiser and Simon (2000) predicted the state of cysteines based on multiple sequence alignments and the success rate was about 82%. Martelli et al. (2002, 2003) presented an approach based on hidden neural networks and achieved 88% accuracy. Mucchielli-Giorgi et al. (2002) had investigated the efficiency of different descriptors in predicting the state of cysteines. Song et al. (2004a, 2004b) constructed a two-stage classifier based on amino acid composition and a linear discriminator based on dipeptide composition for predicting. Chen et al. (2004) combined local sequence windows and global descriptors for predicting and obtained the best prediction accuracy of 90% and Matthews's correlation coefficient (MCC) 0.77.

Former researches seemed to have paid much attention to the prediction performance, but never showed us how much important the roles of these features have played in the prediction. In this paper, we proposed a three-stage SVM to predict the state of cysteines. The first and second stage was to discriminate proteins, depending on whether all, none, or some of the cysteines in the protein are involved in disulfide bridges. In the last stage, we used a slide window to predict the state of cysteines in proteins which have only some of cysteines involved in disulfide bonds. It is the first time we used annotation

* Corresponding author. Tel.: +86 28 89005151; fax: +86 28 85412356.
E-mail address: liml@scu.edu.cn (M. Li).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.09.002

information and secondary structure information to improve the prediction performance in this three-stage classifier and obtained a prediction accuracy of 90.05%, 96.36% and 80.00%, respectively. At the same time, we used *F*-score to evaluate the importance of each feature for representing the disulfide-containing proteins. The result shows that compared with other used proteins' descriptors, evolution information is a more important feature for representing the disulfide-containing proteins.

2. Materials and methods

2.1. Data sets

In this paper, the data set was mainly from Martelli et al. (2002). The data were taken from the crystallographic data of the Brookhaven protein data bank. Disulfide bond assignment was based on the annotation in protein data bank (PDB). These proteins which existed in the SWISS-PORT were used. Proteins with possible chain breaks were excluded. Sequence identity between each pair of sequences was less than 25%. After this procedure, 957 protein sequences were achieved. First, all these sequences were divided into two classes depending on whether they had disulfide bonds and we called them "None" and "Have", respectively. Then these "Have" sequences were divided into two classes depending on whether all the cysteines have been formed into disulfide bonds, so they were named "Mix" and "All". At last, a slide window was used to discriminate each oxidation cysteine from reduced cysteine in the "Mix" sequences. The whole process can be seen from Fig. 1 and the details about the data set can be seen from Supplementary 1.

2.2. Feature extraction

To construct a module which can better distinguish oxidation cysteines from reduced cysteines, we extracted a number of features such as sequence information, evolution information, annotation information and secondary structure information.

2.2.1. Amino acid composition

Amino acid composition has been shown to be the widest used global sequence descriptor in the area of protein classification (Chen et al., 2004; Chou, 2001; Guo et al., 2006a, 2006b; Mucchielli-Giorgi et al., 2002; Shen and Chou, 2007; Shen et al., 2007; Xiao et al., 2009; Zeng et al., 2009; Zhou et al., 2008). Amino acid composition is defined as a 20-dimensional vector which consists of the occurrence frequencies of the 20 typical amino acids. Giving a protein P, its amino acid composition can be expressed as a vector in a 20-D (dimensional) space, as given by

$$\left.\begin{array}{l} p_i = n_i/L \quad (i = 1,2,\ldots,20) \\ P = [p_1, p_2, p_3, \ldots p_i, \ldots, p_{20}]^T \end{array}\right\} \tag{1}$$

Here p_i is the percentage of residue i and n_i is the number of residue i occurring in protein P. L is the length of the protein P.

2.2.2. Auto covariance variables

In order to represent the sequence information more effectively, two physicochemical properties of amino acids were selected to capture the truly specific disulfide bond information of proteins. The two physicochemical properties were hydrophobicity (Wimley and White, 1996) and polarity (Radzicka and Wolfenden, 1988) which have shown to be important features for proteins with disulfide bonds. The value of two physicochemical properties for each amino acid was normalized. Auto covariance (AC) variables were used to transform these proteins into matrices (Guo et al., 2008, 2006a; Xiao et al., 2009; Zeng et al., 2009). Here, lag is the distance between an amino acid residue and its neighbor a certain number of residues away. The AC variables are calculated according to Eq. (2), where j represents one descriptor, i is the position in the sequence X, $X_{i,j}$ represents the value of the chosen features of the ith residue, n is the length of the sequence X and lag the value of the lag

$$AC_{lag,\,j} = \frac{1}{n-lag}\sum_{i=1}^{n-lag}\left(X_{i,\,j} - \frac{1}{n}\sum_{i=1}^{n} X_{i,\,j}\right) \times \left(X_{(i+lag),\,j} - \frac{1}{n}\sum_{i=1}^{n} X_{i,\,j}\right) \tag{2}$$

In this way, the number of AC variables, D can be calculated as $D = lg \times P$, where P is the number of descriptors and lg is the maximum lag (lag = 1, 2…lg). So a protein is transformed into a $2 \times lg$ dimensional vector by AC.

2.2.3. Annotation information and the number of cysteines

Most of proteins with disulfide bonds existed in periplasm of prokaryotic cell or organelles of eukaryotic. Tessier et al. (2004) found that signal peptides and subcellular locations were the primary descriptors which must be considered for predicting the state of cysteines. We obtained subcellular locations and signal peptides information from the SWISS-PROT database (version 54.4). For a protein sequences, if there existed annotation "CC -! - SUBCELLULAR LOCATION" or "FT SIGNAL" we encoded this protein with 1, otherwise 0. In 683 "None" protein sequences, 160 sequences have signal peptides and 356 sequences have subcellular location annotation. In 198 "All" protein sequences, 159 sequences have signal peptides and 135 sequences with subcellular location annotation. Among these 76 "Mix" protein sequences, the number of sequences with signal or subcellular location annotation is 40 and 38. These could be seen from Supplementary 2. The number of cysteines in proteins was primary information for disulfide bonds, especially in the second stage to discriminate "All" from "Mix". The protein with odd cysteines was encoded with 1, otherwise 0.

2.2.4. Position specific scoring matrix

Position specific scoring matrix (PSSM) was created by using PSI-BLAST to search the Swiss-Prot for multiple sequence alignment against the protein samples. It was widely used to represent the protein samples for predicting proteins. Numerous previous researches have proved that evolutionary information can significantly improve the overall prediction performance for predicting the state of cysteines (Fariselli et al., 1999; Martelli et al., 2003; Song et al., 2007). Three iterations of PSI-BLAST were carried out at a cut-off E-value of 0.001. For a protein sequence P with L amino acid residues, PSSM is obtained according to the following equation (Shen and Chou, 2007):

$$P_{PSSM} = \begin{bmatrix} P_{1\to1} & P_{1\to2} & \cdots & P_{1\to j} & \cdots & P_{1\to20} \\ P_{2\to1} & P_{2\to1} & \cdots & P_{2\to j} & \cdots & P_{2\to20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i\to1} & P_{i\to2} & \cdots & P_{i\to j} & \cdots & P_{i\to20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{L\to1} & P_{L\to2} & \cdots & P_{L\to j} & \cdots & P_{L\to20} \end{bmatrix} \tag{3}$$

In Eq. (3), where $i \to j$ describes ith amino acid residue of the protein sequence P being mutated to amino acid type j in the biology evolution process and $P_{i\to j}$ is the score of this mutation. Here the numerical codes 1, 2, 3,…, 20 represent the single character of ordered 20 native types of amino acids. For getting

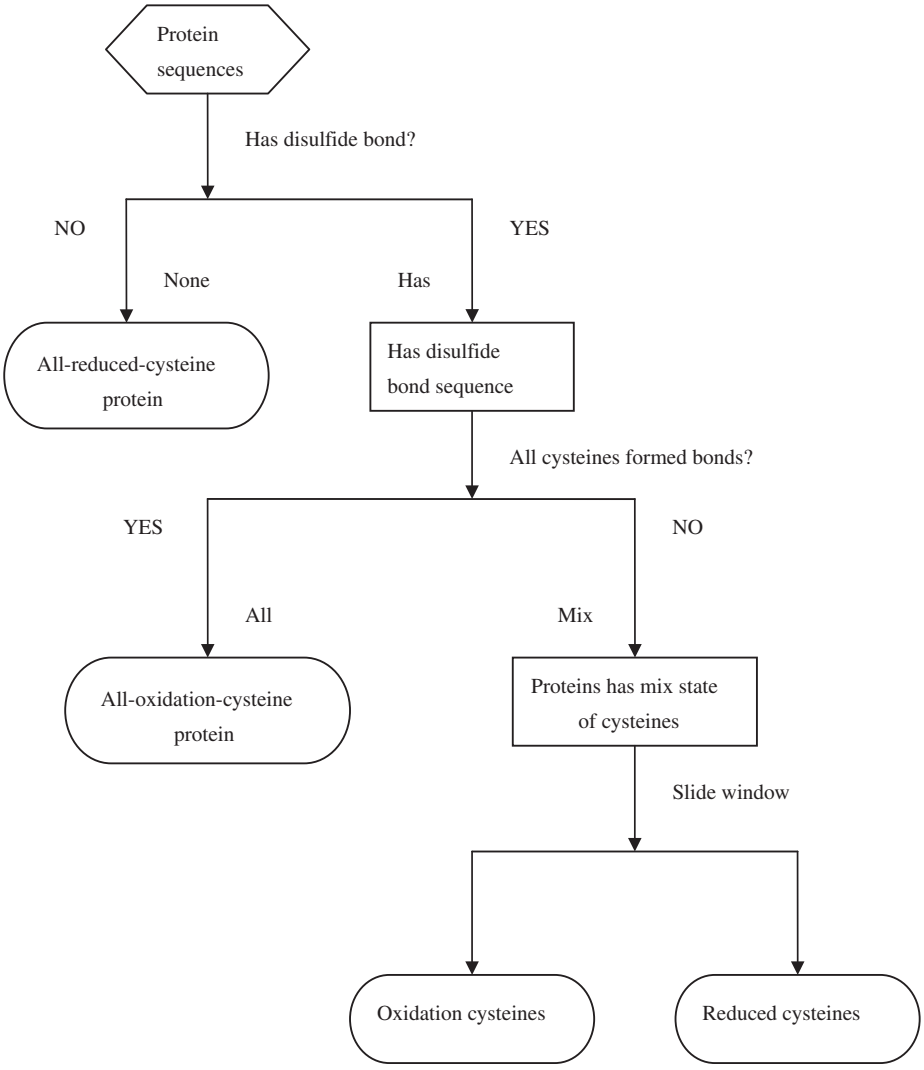

Fig. 1. The architecture of the three stage classifier prediction system.

the $L \times 20$ scores of the P_{PSSM} in Eq. (3), PSI-BLAST (Schaffer et al., 2001) is used to search the Swiss-Prot database (version 54.4) for multiple sequence alignment against the protein P. Then, the value of $P_{i\to j}$ is standardized by Eq. (4), as given below

$$p_{i\to j} = \left({p^o}_{i\to j} - 1/20 \sum_{k=1}^{20} {p^o}_{i\to k} \right) \Big/ (p_{\max,i} - p_{\min,i}) \quad (i = 1,2,3\ldots L; j = 1, 2, 3,\ldots,20) \tag{4}$$

where $P^o_{i\to j}$is the original scores generated by PSI-BLAST, $P_{i\to j}$ is a normalized value. However, as different proteins have different length L, the matrices of the PSSM descriptor in Eq. (3) would have different number of rows. In order to obtain a uniform matrix for protein sequences with different lengths, Eq. (5) was used to convert the PSSM of protein P to a uniform vector

$$P_{PSSM} = \left[\overline{P}_1 \quad \overline{P}_2 \quad \cdots \quad \overline{P}_j \quad \cdots \quad \overline{P}_{20} \right]^T (j = 1,2,\ldots,20) \tag{5}$$

where T is the transpose operator, and

$$\overline{P}_j = \frac{1}{L} \sum_{i=1}^{L} P_{i\to j} \quad (i = 1,2,\ldots,L;\ j = 1,2,\ldots,20) \tag{6}$$

where $\overline{P}_j$ is the average score of all the amino acid residues mutated to type j amino acid in the evolution procession in protein P. Furthermore, in order to predict the state of cysteines in the "Mix" proteins, a slide window incorporating the evolutionary information from upstream and downstream neighbors was used. For an amino acid residue a_i in sequence position a_i, we used a feature vector P_i to represent it. l is an odd number which stands for the size of sliding window

$$P_i = [p_{[a_i-(w-1)/2]},\ldots,p_{[a_i]},\ldots,p_{[a_i+(w-1)/2]}] \tag{7}$$

where $p_{[a_i]}$ is the ith PSSM row of the residue a_i. If the window extends beyond the sequence, $(w-1)/2$ zero vectors of dimension 20 are appended on empty positions before the first and after the last residue of a PSSM profile. The profile added a sliding window is defined as the standard PSSM profile (Cheng et al., 2008).

2.2.5. Secondary structure information

Secondary structure information has been widely used in predicting the disulfide bonds (Ferre and Clote, 2005a, 2005b; Song et al., 2007). Ferre and Clote (2005b) reported that secondary structure information can lead to a marked improvement in predicting the state of cysteines. The secondary structure information of these 76 'Mix' proteins was predicted through the PSIPred (http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py?form=psipred) (Jones, 1999). The secondary structure information encoded in unary format by the addition of three input units, Helix was encoded 1 0 0, coil was 0 1 0 and sheet was 0 0 1. Though

the slide window, each cysteine was substituted by a $1 \times l$ vector and l was the window size.

2.3. Support vector machine

SVM has been shown to be quite an effective machine-learning method in computational biology (Ceroni et al., 2003; Chou and Cai, 2002; Chou and Shen, 2006; Guo et al., 2006a, 2006b; Zhou et al., 2008). Vapnik (1998) has given a full description about how to use SVM to do classification. In this paper, the classifications were implemented by libsvm 2.88 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). Details about the libsvm can be seen in the web (Fan et al., 2005, 2008). A radial basic function (RBF):$e^{-r||x_i-x_j||^2}$ was chosen as the kernel function, where r is the parameter. In the training process, two parameters, the regularization parameter C and the kernel width parameter γ were optimized by using a grid search approach within a limited range.

2.4. Performance evaluation and feature importance measures

2.4.1. Performance measures

The performance of the modules constructed in this study was evaluated using a fivefold cross-validation technique (Chou and Shen, 2007). In this technique, the dataset was randomly divided into five equally sized sets. One set was used for testing and the remaining four were for training. A SVM model built using the training set was trained and validated by fivefold cross-validation, and the performance of this model was evaluated by the corresponding test set. In order to minimize the dependence of the method on the dataset, the process of random selection of training set and test set was repeated five times. Thus five training sets and five test sets were prepared, so five models were generated. We used the average accuracy as the last result. Four parameters—sensitivity, specificity, accuracy and Matthews's correlation coefficient (MCC) (Matthews, 1975) were used to evaluate the method's performance. They are defined by the following formulations:

$$Sensitivity = TP/(TP+FN) \tag{8}$$

$$Specificity = TN/(TN+FP) \tag{9}$$

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \tag{10}$$

$$MCC = (TP \times TN - FP \times FN)/\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)} \tag{11}$$

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative, respectively.

2.4.2. Feature importance measures

Current classification seemed to have paid too much attention to the prediction performance, but never showed us how much important the roles of these features have played in the prediction. In this paper, a novel conditional feature importance strategy was used to evaluate these features. This strategy was implemented by using the F-score function of the libsvm tool (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/fselect/fselect.py). This feature selection tool is a simple python script. F-score is a technique which can measure the discriminability of two sets of real numbers. For a training vectors $x_k(k=1, \ldots, m)$, n_+ and n_- represent the number of positive and negatives instances, respectively, then the F-score of the ith feature is defined as

$$F(i) = \left((\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2\right) \Big/ \left[\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2/(n_+-1) + \sum_{k=1}^{n_-}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2/(n_--1)\right] \tag{12}$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ are the average of the ith feature of the whole, positive, and negative data sets, respectively; $x_{k,i}^{(+)}$ and $x_{k,i}^{(-)}$ represent the ith feature of the kth positive and negative instance, respectively. The larger the F-score is, more important the feature is. The details can be found in Chen's book (Chen and Lin, 2006).

3. Results and discussions

3.1. Selecting optimal parameters

The parameter lg represents the maximum distance between two considered residues. For different datasets, different lgs were chose to get the best predicting performance. The maximal possible lg cannot be larger than the length of the shortest sequence in the dataset. In this paper, six different lgs ($lg=0, 1, 5, 10, 15, 20, 25$) were selected to build SVM modules for the first and second stage prediction. Prediction performance can be seen from Fig. 2. As seen from the curve, when the lg was 5, the prediction accuracy was 82.72% which was the highest among these lgs. So the optimal lg was 5.

3.2. Prediction of proteins with disulfide bond

In the first stage, SVM modules were developed to discriminate "Has" from "None" proteins. The performance of the modules based on different features has been shown in Table 1. The performance of SVM module obtained with AC was similar to which obtained with the PSSM and the SVM module based on annotation information and the number of cysteines also obtained a satisfactory accuracy of 80.10%. The hybrid approach increased the overall accuracy compared with the individual modules. The best results were obtained with highest accuracy of 90.05% and MCC of 0.75 based on the whole features, when $C=2$ and $\gamma=0.125$. These results indicate that all these features are important for representing the disulfide-containing proteins, but we cannot see how important these features are. So a measurement named F-score function was used. The F-score of each vector can be seen in Fig. 3a. The vectors from 1 to 30 represent the AC feature, from 31 to 50 are the PSSM feature and the last three are the annotation information and cysteines information. It is easy to see that signal peptides vector (53) has the highest F-score among these vectors, which indicates that signal peptide is a strong descriptor for proteins with disulfide bonds. This finding

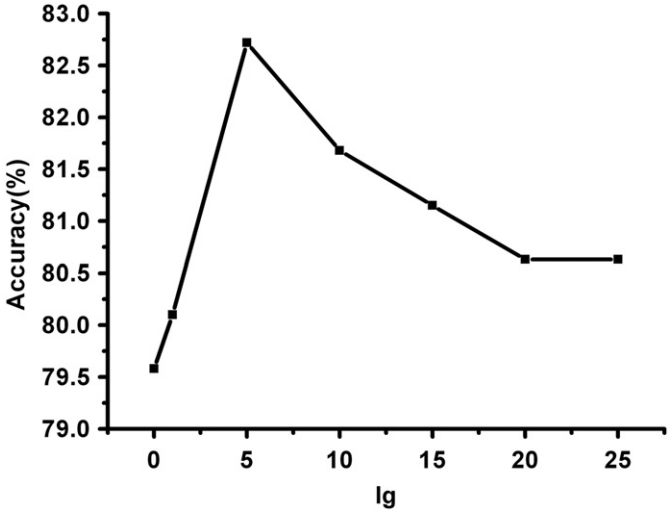

Fig. 2. The prediction accuracy of the method with auto covariance of different lgs respectively.

Table 1
The performance of various modules used in predicting proteins with disulfide bond.

Approach	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
AC (A)	82.72	53.36	93.38	0.56
PSSM(B)	81.15	52.72	92.64	0.51
Annotation and Cysteines(C)	80.10	60.00	88.24	0.50
Hybrid1(A+B)	85.86	74.54	90.44	0.65
Hybrid2(A+C)	83.77	56.36	93.38	0.55
Hybrid3(B+C)	84.82	68.52	91.24	0.62
Hybrid4(A+B+C)	90.05	76.36	95.59	0.75

AC: Auto covariance variables, MCC: Matthew's correlation coefficient.

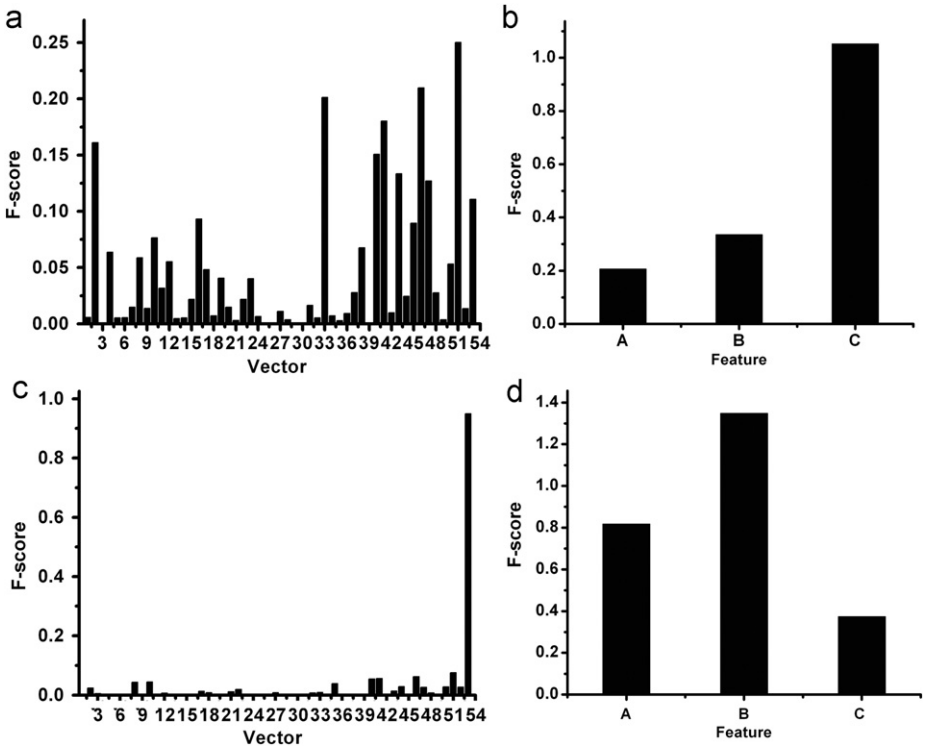

Fig. 3. The value of *F*-score for different vectors and features in the first and second predicting stage. A: Auto covariance variables. B: PSSM. C: Annotation information and the number of cysteines.

was consistent with the observations drawn by Tessier et al. (2004) that signal peptide is a strong descriptor for the disulfide-bonding state in proteins. Vectors in PSSM also obtained high *F*-scores. Fig. 3b shows the importance of these three features. The *F*-score of PSSM, AC, annotation information and the number of cysteines were 1.35, 0.82 and 0.37, respectively. Through annotation information and the number of cysteines are less important, they are still useful variables. These *F*-score values are consistent with the performance of Table 1. So *F*-score function was successful in evaluating these features. From these observations, we can draw a conclusion that evolution information contained more information for proteins with disulfide bond.

3.3. Prediction of "Mix" and "All" proteins

To classify proteins depending on whether its all cysteines have formed into disulfide bonds, the SVM modules at this second stage were constructed based on different features and the prediction results are summarized in Table 2. The hybrid approach combining all these features as input obtained highest overall accuracy of 96.36%, 0.91 MCC ($C=128$, $\gamma=0.000122$). The *F*-score of each vector has been shown in Fig. 3c. The last vector represented the number of cysteines and which obtained the highest *F*-score of 0.95. The reason is that the number of cysteines in "All" proteins is always even but in 'Mix', the number is always odd. So the model only based on the number of cysteines can effectively discriminate "All" from "Mix" proteins. The importance of each feature has been shown in Fig. 3d. Annotation information and the number of cysteines have the highest *F*-score of 1.05. The value of *F*-score for PSSM and AC are 0.34 and 0.21. This could explain the reason why hybrid SVM module based on PSSM and AC cannot improve performance at this stage. From these two stage results, we can draw a conclusion that *F*-score is able to evaluate the features.

3.4. Predicting the state of cysteines in "Mix" proteins

In predicting the state of each cysteine in "Mix" proteins, a slide window was used in this stage. The slide window centered on the cysteine residue was based on its local amino acids composition, evolution information and secondary structure information. In order to get a better prediction performance, annotation information and the number of cysteines were also added. A preliminary test was done to determine the optimal

Table 2
The performance of various modules constructed in classifying "Mix" from "All" proteins.

Approach	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Annotation and Cysteines(C)	90.91	67.78	100	0.77
Hybrid1(A+C)	91.18	71.88	100	0.81
Hybrid2(B+C)	92.73	73.33	100	0.82
Hybrid3(A+B+C)	96.36	87.50	100	0.91

A: Auto covariance variables, B: PSSM.

Table 3
The performance of different window sizes used in predicting the state of cysteines in "Mix" proteins.

Window size (l)	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
5	70.77	64.29	75.68	0.40
9	73.85	71.43	75.68	0.47
13	80.00	78.57	81.08	0.59
17	72.31	60.71	81.08	0.43
21	67.69	60.71	72.97	0.34

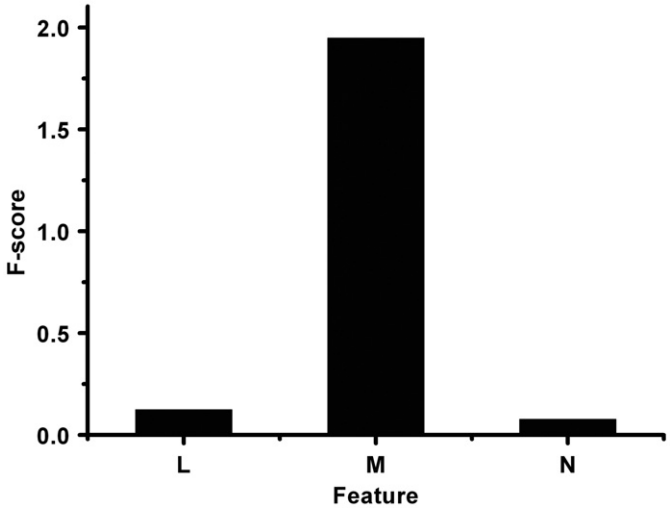

Fig. 4. The value of *F*-score for different features in predicting the state of cysteines in "Mix" proteins. L: Annotation and the number of cysteines. M: slide PSSM. N: Secondary structure information.

window size l by measuring the prediction accuracy of the various window sizes l from 5 to 21. The performance of different window sizes can be seen from Table 3. The best prediction performance was obtained with an accuracy of 80.00% and MCC 0.59 ($C=2$, $\gamma=0.03125$) when the window size $l=13$. The value of *F*-score for each feature has been shown in Fig. 4. The slide PSSM gives the highest *F*-score of 1.948, annotation information and the number of cysteines has the second highest *F*-score of 0.1634. Although Ferre et al. reported that secondary structure information can improve the prediction of the state of cysteines; *F*-score of secondary structure information in this stage are relatively low. This indicated that local amino acid composition, annotation information and number of cysteines can only help increase prediction accuracy slightly. From these observations, the slide PSSM makes significant contribution to classification accuracy. Cysteines tend to be more conserved in proteins when they form disulfide bridges. From PSSM, we can found the conserved value of each cysteine, so it is an important attribute for representing the disulfide-containing proteins. Although there is a phenomenon that the oxidation of cysteines exhibits obvious cooperatively and the state of cysteines in "Mix" proteins is hard to discriminate (Song et al., 2004a), this efficient method provides a comparative prediction.

4. Conclusions

In the present study, we have developed an efficient three-stage classification model to predict the state of cysteine by using SVM methods based on AC, PSSM, annotation information and global descriptors such as the number of cysteines in protein sequences. Our method achieved an overall prediction accuracy of 90.05%, 96.36% and 80.00%, respectively. At the same time we introduced *F*-score to measure the feature's importance in each stage successfully. Through the *F*-score of each feature, we found that PSSM contained much information about disulfide bond protein. Our study has also shown that the slide PSSM as the local sequence environment of cysteine was an important descriptor for representing the disulfide-containing proteins. Overall such a method will be a good supplementary tool for cysteines studies.

Acknowledgements

The authors would like to thank Martelli for sharing the dataset. We would like to thank the anonymous reviewers for their patient review and constructive suggestions. This work was supported by the National Natural Science Foundation of China (Nos. 20905054, 20972103).

Appendix A. Supplementary materials

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2010.09.002.

References

Ceroni, A., Frasconi, P., Passerini, A., Vullo, A., 2003. Predicting the disulfide bonding state of cysteines with combinations of kernel machines. J. VLSI. Signal Process. Syst. 35, 287–295.
Chen, Y.C., Lin, S.C., Lin, C.J., Hwang, J.K., 2004. Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. Proteins 55, 1036–1042.
Chen, Y.W., Lin, C.J., 2006. Combining SVMs with various feature selection strategies.. Springer.
Cheng, C.W., Su, E.C.Y., Hwang, J.K., Sung, T.Y., Hsu, W.L., 2008. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. BMC Bioinformatics 9, S6.
Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43, 246–255.
Chou, K.C., Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. 277, 45765–45769.
Chou, K.C., Shen, H.B., 2006. Predicting protein subcellular location by fusing multiple classifiers. J. Cell. Biochem. 99, 517–527.
Chou, K.C., Shen, H.B., 2007. Recent progress in protein subcellular location prediction. Anal. Biochem. 370, 1–16.
Creighton, T., 1993. Proteins: Structures and Molecular Properties, New York.

Creighton, T., 1995. Disulfide-coupled protein-folding pathways. Phil. Trans. R. Soc. London B Biol. Sci. 348, 5–10.
Fan, R.E., Chen, P.H., Lin, C.J., 2005. Working set selection using order information for training SVM. J. Mach. Learn. Res. 6, 1889–1918.
Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. Liblinear: a library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874.
Fariselli, P., Riccobelli, P., Casadio, R., 1999. Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins 36, 340–346.
Ferre, F., Clote, P., 2005a. Disulfide connectivity prediction using secondary structure information and diresidue frequencies. Bioinformatics 21, 2336–2346.
Ferre, F., Clote, P., 2005b. DiANNA: a web server for disulfide connectivity prediction. Nucleic Acids Res. 33, W230–W232.
Fiser, A., Simon, I., 2000. Predicting the oxidation state of cysteines by multiple sequence alignment. Bioinformatics 16, 251–256.
Fiser, A., Cserzo, M., Tudos, E., Simon, I., 1992. Different sequence environments of cysteines and half cystines in proteins application to predict disulfide forming residues. FEBS Lett. 302, 117–120.
Guo, Y.Z., Yu, L.Z., Wen, Z.N., Li, M.L., 2008. Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences. Nucleic Acids Res. 36, 3025–3030.
Guo, Y.Z., Li, M.L., Lu, M.C., Wen, Z.N., Huang, Z.T., 2006a. Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform. Proteins 65, 55–60.
Guo, Y.Z., Li, M., Lu, M., Wen, Z., Wang, K., Li, G., Wu, J., 2006b. Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform. Amino Acids 30, 397–402.
Hogg, P.J., 2003. Disulfide bonds as switches for protein function. Trends Biochem. Sci. 28, 210–214.
Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.
Martelli, P.L., Fariselli, P., Malaguti, L., Casadio, R., 2002. Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. Protein Eng. 15, 951–953.
Martelli, P.L., Fariselli, P., Malaguti, L., Casadio, R., 2003. Prediction of the disuffide-bonding state of cysteines in proteins at 88% accuracy. Protein Sci. 11, 2735–2739.
Matthews, B.W., 1975. Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta. 405, 442–451.
Mucchielli-Giorgi, M.H.M., Hazout, S., Tuffery, P., 2002. Predicting the disulfide bonding state of cysteines using protein descriptors. Proteins 46, 243–249.
Muskal, S.M., Holbrook, S.R., Kim, S.H., 1990. Prediction of the disulfide-bonding state of cysteine in proteins. Protein Eng. 3, 667–672.
Radzicka, A., Wolfenden, R., 1988. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry 27, 1664–1670.
Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F., 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29, 2994–3005.
Shen, H.B., Chou, K.C., 2007. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng. Des. Sel. 20, 561–567.
Shen, H.B., Yang, J., Chou, K.C., 2007. Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33, 57–67.
Song, J.N., Li, W.J., Xu, W.B., 2004a. Cooperativity of the oxidization of cysteines in globular proteins. J. Theor. Biol. 231, 85–95.
Song, J.N., Wang, M.L., Li, W.J., Xu, W.B., 2004b. Prediction of the disulfide-bonding state of cysteines in proteins based on dipeptide composition. Biochem. Biophys. Res. Commun. 318, 142–147.
Song, J.N., Yuan, Z., Tan, H., Huber, T., Burrage, K., 2007. Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. Bioinformatics 23, 3147–3154.
Tessier, D., Bardiaux, B., Larre, C., Popineau, Y., 2004. Data mining techniques to study the disulfide-bonding state in proteins: signal peptide is a strong descriptor. Bioinformatics 20, 2509–2512.
Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.
Vieille, C., Zeikus, G.J., 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol. Mol. Biol. Rev. 65, 1–43.
Wimley, W.C., White, S.H., 1996. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat. Struct. Biol. 3, 842–848.
Wittrup, K.D., 1995. Disulfide bond formation and eukaryotic secretory productivity. Curr. Opin. Biotechnol. 6, 203–208.
Xiao, R.Q., Guo, Y.Z., Zeng, Y.H., Tan, H.F., Pu, X.M., Li, M.L., 2009. Using position specific scoring matrix and auto covariance to predict protein subnuclear localization. J. Biomed. Sci. Eng. 2, 51–56.
Zeng, Y.H., Guo, Y.Z., Xiao, R.Q., Yang, L., Yu, L.Z., Li, M.L., 2009. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J. Theor. Biol. 259, 366–372.
Zhou, X.B., Chen, C., Li, Z.C., Zou, X.Y., 2008. Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. Amino Acids 35, 383–388.