The Prediction of Disulfide Bonding Patterns with the Cysteine Labels*

Chun-Yi Fang, Chang-Biau Yang† and Chiou-Yi Hor‡

† Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung 80424, Taiwan

‡ Iron and Steel Research & Development Department
China Steel Corporation
Kaohsiung 81233, Taiwan

***Abstract*—There are four kinds of disulfide prediction problems, including chain classification, bonding state of each cysteine, connectivity of each cysteine pair and disulfide bonding pattern. Among these problems, the prediction of disulfide bonding pattern is the most difficult. In this paper, we propose a novel algorithm to solve the disulfide bonding pattern problem for the proteins that the bonding states of cysteines are given. Our method first predicts the labels of cysteines with global information, and then uses the label prediction results to derive the whole disulfide bonding pattern. Furthermore, to improve the prediction accuracy, we build hybrid methods by combining several methods. As the experimental results show, our hybrid method has prediction accuracy 83.9%, which outperforms 79.8% of the previously best predictor.**

***Index Terms*—protein, disulfide bond, cysteine, support vector machine, behavior knowledge space**

I. Introduction

In the organic chemistry, a disulfide bond is a single covalent bond formed by two sulfhydryl (thiol) groups (-SH) with the redox (oxidation-reduction) reaction between them. The sulfhydryl (thiol) group is a functional group, consisting of a sulfur atom (S) and a hydrogen atom (H). Figure 1 shows the configuration of a disulfide bond, which brings two cysteines together. Each cysteine in the protein has two possible states, either oxidized or reduced. We say a cysteine is in *oxidized state* when its oxidation number is changed by losing an electron (e^-). Otherwise, it is in *reduced state* when it is changed by gaining an electron (e^-).

As well known, a protein is composed of chains of amino acids and the cysteine is one of the twenty standard amino acids. Each protein has its own tertiary (three-dimensional) structure, which is closely related to its functionality. By knowing the three-dimensional structure of a protein, one can more precisely predict the function of the protein. Consequently, in the past, a lot of studies have been devoted to the prediction of disulfide bonds [18], [12], [10], [11], [17], [9], [3], [8].

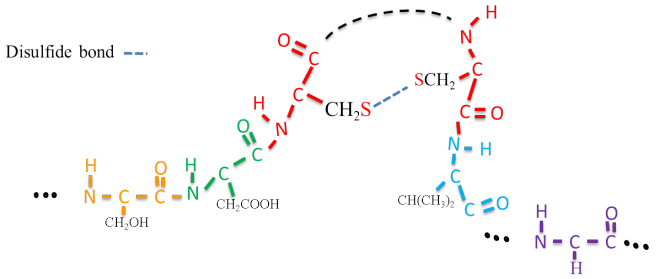

Fig. 1. The configuration of a disulfide bond, where each letter represents an atom.

The disulfide bond prediction problem can be divided into the following four versions.

- *Chain classification:* It determines whether there exist disulfide bonds in a protein sequence.
- *Bonding state of each cysteine:* It determines the bonding state of each cysteine. There are two possible states for a cysteine, either oxidized or reduced.
- *Connectivity of each pair of cysteines:* It is to decide whether a pair of cysteines in a protein are bonded or not. For a protein with exactly D oxidized cysteines, there are C_2^D possible pairs of cysteines, but only $\frac{D}{2}$ pairs are connected together. Note that D should be an even number.
- *Disulfide bonding pattern a protein:* It mainly

*This research work was partially supported by the National Science Council of Taiwan under contract NSC 101-2221-E-110-020-MY3.

†Corresponding author: cbyang@cse.nsysu.edu.tw

focuses on how to predict the whole cysteine connectivity configuration of a protein, under the assumption that the bonding states of all cysteines are known. For a protein with D oxidized cysteines, the number of possible patterns is big as $\frac{C_2^D \times C_2^{D-2} \times ... \times C_2^2}{\frac{D}{2}!} = (D-1)!!$.

According to the literatures, many researchers performed the disulfide bond prediction by making bonding states of cysteines as prior knowledge [18], [4], [10], [17], [3], [9]. By this way, it can efficiently reduce the number of possible bonding pairs and the number of possible bonding patterns to be predicted. Even though, there are still some methods that can handle the problem without the prior knowledge of bonding states [12], [8].

Several disulfide bonds prediction methods are based on machine learning methods. In 2005, Zhao *et al.* [18] presented an analytic method to predict disulfide bonding patterns. They assumed that the less difference between two *cysteine separation profiles* (CSPs) [5] [18] there is, the more similar disulfide bonding pattern between two proteins it has. In the same year, Rama *et al.* [12] took each pair of oxidized cysteines as a sample and used *support vector machine* (SVM) [15] to predict the connectivity of the two cysteines. In 2006, Cheng *et al.* [4] used recursive neural networks to predict the connectivity probability of each pair of cysteines. Then, a *graph matching* [6] algorithm is invoked to convert the connectivity probabilities of all possible pairs into the connectivity pattern. Here, graph matching is applied on a weighted undirected graph, where its vertices represent the oxidized cysteines, and the weight of each edge is the connectivity probability between two oxidized cysteines. In 2007, Lu *et al.* [11] took every four oxidized cysteines as a sample for the SVM, and they invoked a feature selection method, based on the genetic algorithm, to predict the connectivity pattern.

In 2012, Wang *et al.* [17] proposed hybrid methods, which predicted odd and even numbers of disulfide bonds separately by using two different models. Chen *et al.* [3] also proposed a hybrid method which fuses CSP and SVM by means of the *behavior knowledge space* (BKS) [13] method to predict the disulfide pattern. Lin and Tseng [9] proposed the multiple trajectory search to adjust the parameters of the SVM. In 2013, Lin *et al.* [8] used the three-dimension coordinates of amino acids, produced by MODELLER[14], as features for an ensemble of SVMs to predict the cysteine connectivity. Then, the *multiple trajectory search* (MTS) [9] is adopted to determine the best parameters of the ensemble classifier. Finally, a modified graph matching algorithm is adopted to get the disulfide bonding pattern. The method achieves a high classification accuracy even though the bonding information is not given in advance.

The rest of this paper is organized as follows. In Section II, we will introduce the dataset and features used in this paper. In Section III, we will propose our method. In Section IV, we will show our experiment results. In Section V, we will give our contribution and finally draw a conclusion.

II. Dataset and Features

A. Datasets and Performance Evaluation

SP39 [16] is a widely used dataset for disulfide experiments. In this paper, we also adopt it for a fair comparison with previous works. *SP39* dataset contains 446 proteins, in which the number of disulfide bonds ranges between 2 and 5.

In our experiments, we apply the leave-one-protein-out cross-validation scheme. We use $Q_p = C_p/T_p$ to evaluate the prediction accuracy, where C_p is the number of proteins with correctly predicted disulfide pattern and T_p is the totally number of testing proteins. Besides, we also use the sensitivity to evaluate the performance of the prediction of cysteine labels. The definition of the *sensitivity* or *true positive* rate (TPR) is given by $TPR = \frac{TP}{TP+FN}$, where TP, TN and FN are true positive, true negative and false negative numbers, respectively.

B. Feature Extraction

Table I shows the features used in this paper. For a protein P_x with D cysteines, the features associated with a selected pair of cysteines C_i^x and C_j^x is given as follows.

DCR (Distance of Cysteine Residues): We normalize the position distance between two cysteines by $\log_2(1 + \frac{|C_i^x - C_j^x|}{|P_x|})$, where $|P_x|$ denotes the length of P_x.

TABLE I
THE FEATURES USED BY VARIOUS MODELS (DESCRIBED IN SECTION III), WHERE w DENOTES THE WINDOW SIZE, Y/N DENOTES WHETHER THE FEATURE IS ADOPTED OR NOT AND, D IS THE NUMBER OF CYSTEINES.

Features	Size	CP_1F_{521}	CP_1F_{623}	CP_bF_{521}	$CPbF_{623}$	CL
DCR^a	1	Y	Y	Y	Y	N
Cysteine residues order	2	N	Y	N	Y	N
Protein weight	1	N	Y	N	Y	N
Protein length	1	N	Y	N	Y	N
Amino acid composition	20	N	Y	N	Y	N
PSSM^b	$(2w+1)\times 20\times D$	Y	Y	Y	Y	Y
Secondary structure^c	$(2w+1)\times 3\times D$	N	Y	N	Y	N

[a]Distance of cysteines
[b]PSSM centering at the cysteine with window size $2w+1$
[c]Secondary structure centering at the cysteine with window size $2w+1$

Cysteine Residues Order: We normalize the position indices of the two selected cysteines as $\frac{i}{D}$ and $\frac{j}{D}$.

Protein Length: The feature is given by $\frac{|P_x|}{|P_x|_{Max}}$, where $|P_x|_{Max}$ is the maximum length of the proteins in the dataset.

Amino Acid Composition: The feature represents the frequency that each kind of amino acid appears. We normalized $\frac{AA_t^x}{|P_x|}$, where AA_t^x is the number of occurrences of amino acid t in P^x

PSSM around the Cysteine Residues: The feature is normalized as $(PSSM_{i,j} - PSSM_{min})/PSSM_{Max}$, where $PSSM_{i,j}$, $PSSM_{min}$ and $PSSM_{Max}$ are the score of element in (i,j) (row i and column j), the maximum score and minimum score in the PSSM table [1], respectively.

Secondary Structure around the Cysteines: The feature is generated by PSIPRED [7], which produces a probability matrix with size $|P_x|\times|S|$, where S is composed of the three substructures, *alpha helices*, *beta sheets* and *coils*.

III. ALGORITHMS FOR THE DISULFIDE BONDING PATTERN PREDICTION

In this paper, we propose a novel model, *cysteine label prediction* (CL), for increasing the accuracy of disulfide bonding pattern prediction. And we also design one hybrid method based on the CL method.

A. Cysteine Label Prediction

We first define some notations. P_x denotes a protein sequence consisting of elements in Σ, where Σ denotes the set of twenty standard amino acids, and $P_{x,i}$ denotes the ith element of P_x. Let B and D be the number of disulfide bonds and the number of oxidized cystines, respectively, where $B=\frac{D}{2}$. Besides, C_j^x and L_j^x denote the index of an oxidized cysteine and its bond label (bond index), respectively, for P_x in ascending order. The bond label indicates which disulfide bond a cysteine belongs to. To be more specific, if there is a disulfide bond between two different cysteines C_a^x and C_b^x, their L_a^x and L_b^x will be the same. Otherwise, L_a^x and L_b^x will be different if they share no disulfide bond. By this way, one can gather information of all disulfide bonds in P_x and get the disulfide bonding pattern, '$L_1^x - L_2^x - L_3^x \ldots L_D^x$'. For example, suppose that there are three disulfide bonds in P_y where C_1^y is connected to C_2^y, C_3^y to C_5^y, and C_4^y to C_6^y. Then the disulfide bonding pattern of P_y is represented by '1-1-2-3-2-3'. After defining the notations, we describe our cysteine label prediction as follows.

Algorithm: Cysteine Label Prediction.

Input: (a) A training set, consisting of primary sequences of proteins and the bonding pattern of all oxidized cysteines; and (b) A testing set, consisting of primary sequences of proteins.

Output: Probability associated with each label for each oxidized cysteine of every protein in the testing set.

Step 1 (Assign cysteine labels): According to the disulfide bonding pattern of each protein in the training set, label each cysteine with its bond index.

Step 2 (Arrange training dataset): Collect the proteins with i disulfide bonds ($2i$ oxidized

cysteines) into a data subset PD_i, where $i = 2, 3, 4$ or 5. All of the proteins in the training set and testing set are in the same subset (PD_i).

Step 3 (Extract features): For each protein, we extract features from the potential values of the PSSM matrix centering at each oxidized cysteine with window size $2 \times w + 1$. The total number of features is $20 \times D \times (2 \times w + 1)$.

Step 4 (Train classifiers): According to the position indices of the oxidized cysteines in the training set, we further divide each data subset PD_i into $2i$ subsets, $PD_{i,1}$, $PD_{i,2}$, ..., $PD_{i,2i}$, where $PD_{i,j}$ consists of all jth cysteins in all proteins. Note that the labels of the cysteines in $PD_{i,j}$ may be different, ranging between label 1 and label $\min\{j, B\}$. Then, for each $PD_{i,j}$, we train an SVM model with `LIBSVM` package [2] to predict the probability of each cysteine label in the testing set.

B. The Pairwise Method with Label Predictions

We regard that cysteines with identical predicted cysteine labels should belong to the same disulfide bond. In this sense, we convert the predicted probabilities of cysteine labels into a pairwise manner for the subsequent prediction of disulfide bonding patterns.

Algorithm: Pairwise Method.

Input: Predicted label probabilities of all oxidized cysteines in one protein.

Output: The disulfide bonding pattern of the protein.

Step 1 (Convert to pairwise probability): The bonding probability weight of two cysteines C_i and C_j in protein P_x is given by $P(C_i, C_j) = \sum_{q=1}^{B} Prob(i, q) \times Prob(j, q)$, where $1 \leq i < j \leq D$ and q denotes disulfide bond label.

Step 2 (Generate the bonding pattern): Construct a complete and undirected graph for a protein, where the vertex set consists of the oxidized cysteines and the weight of each edge is the bonding probability weight between two cysteines. Then, a graph matching algorithm [6] is invoked to find the most possible pattern for the protein.

C. The Hybrid Method

According to our preliminary experimental results, no single method predicts the bonding patterns with all different numbers of disulfide bonds accurately. Thus, we design the hybrid methods, which combine four different methods with different features and training datasets. The methods include CP_1F_{521} [3], CP_1F_{623} [3], $CPbF_{521}$ and $CPbF_{623}$. The latter two methods are similar to the former two, except that the training datasets in the latter two involve only the proteins with b bounds when b-bound proteins are predicted. All of these methods are pairwise methods which use each cysteine pair as a sample to predict the boding probability of each cysteine pair. Here, we use two different hybrid methods for proteins with different disulfide bond numbers. The pairwise BKS method handles the proteins with four or five disulfide bonds and the pattern-wise method handles proteins with two or three disulfide bonds.

Algorithm: The pairwise BKS method

Input: The bonding probabilities of all cysteine pairs in PD_4, with 4 bonds (or PD_5, with 5 bonds).

Output: A modified disulfide bonding pattern.

Step 1 (Original method): Adopt the CP_1F_{521} [3] or CP_1F_{623} [3] method to predict the disulfide bonding pattern. For each edge, if the normalized weight of a protein is greater than a predefined global threshold ($\theta_1 = 0.5$ in this paper), the result is exported as the final result.

Step 2 (The CSP method): Adopt the CSP method to get the disulfide bonding pattern. Here, another global threshold θ_2 is used to choose between the CSP method and the original method. If the divergence of the protein is less than θ_2, we will adopt CSP's result.

Step 3 (The pairwise BKS method): Conduct the pairwise BKS method to modify the disulfide pairwise bonding probabilities according to the difference between the corresponding BKS entries [3]. Then a graph matching algorithm [6] is utilized to obtain the new disulfide bonding pattern.

Algorithm: The pattern-wise BKS method

Input: The disulfide bonding patterns of all proteins

in PD_2, with 2 bonds (or PD_3, with 3 bonds).

Output: A modified disulfide bonding pattern.

Step 1 (Original method): Same as Step 1 in the pairwise BKS method.

Step 2 (The pattern-wise BKS method): Assign possible patterns as the BKS entries, and use the training data to fill in the BKS table. For cases that the values of the corresponding BKS entries are empty, the next step is proceeding. Otherwise, stop here.

Step 3 (The CSP method): Assign the result of the CSP method as the final result.

IV. Experimental Results

In our study, problems on the prediction of disulfide connecting patterns include large numbers of possible patterns, insufficient training samples and imbalanced numbers of positive and negative samples. These issues totally account for the low prediction accuracy. The CL method is a label-wise method which takes the global information of cysteines into consideration. In this way, the CL method provides a different aspect other than the conventional pairwise method, which only considers the information of two cysteines. To achieve an acceptable accuracy of the CL method and its hybrid methods, we finally set the window size to $2w + 1$, where $w = 2$. In addition, we compare our prediction accuracies with previous works by means of the leave-one-protein-out cross-validation.

Table II shows the prediction accuracies of various single methods. Though the accuracies of both CP_1F_{521} and CP_1F_{623} are better than the CL method, the latter one presents discriminative results. Thus the hybrid methods involving the CL method usually get good accuracies. Table III shows the prediction accuracies of various hybrid methods. The accuracies of models that combine CL are always higher than $CPbF_{521}$ and $CPbF_{623}$ for all hybrid methods. The most accurate prediction is shown in the last row. The accuracy is 83.9%, which outperforms 79.8% of the previously best predictor with CP_1F_{623}. The experimental result shows that the CL method uses smaller training dataset and fewer features (not shown here) to get a comparable or better accuracy to the conventional pairwise methods.

V. Conclusion

For the prediction of disulfide bonding pattern, most methods adopt either pattern-wise or pairwise approaches. However, pattern-wise methods may be limited if the training dataset is too small. That is, for each pattern, there should be enough samples for training. The number of possible disulfide bonding patterns increases rapidly as the number of disulfide bonds increases. Pairwise methods might also suffer from the data imbalance problem. It is because numbers of the positive (connected) and negative (unconnected) samples are not equal, especially when the number of disulfide bonds increases. In addition, the pairwise method also does not consider global information.

To overcome the above problems, we propose the CL method that incorporates pairwise methods to predict the disulfide bonding pattern with the global information of oxidized cysteines. The CL method is to start from the prediction of cysteine label, and to end up at the cysteine bonding pattern. The prediction of cysteine label alleviates the data imbalance problem by making each label of cysteines as a training sample instead of the connectivity of two cysteines. With the help of the leave-one-protein-out cross-validation scheme, we can have more samples for training.

Though the prediction accuracies of the CL method are not better than other single methods, such as CP_1F_{521} and CP_1F_{623}, the CL method plays a discriminative role in the hybrid method because it provide a different viewpoint of the prediction. As the experimental results show, the hybrid methods with CL always have more accurate results. The hybrid method with the most accurate prediction has accuracy 83.9%, which outperforms 79.8% of the previously best predictor.

In the future, we may adopt our CL method in other datasets, and we may also use a larger dataset to train classifiers with a larger window size.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, Vol. 25, No. 17, pp. 3389–3402, 1997.

[2] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

TABLE II
THE ACCURACIES (%) OF VARIOUS SINGLE METHODS WITH THE LEAVE-ONE-PROTEIN-OUT CROSS-VALIDATION.

$Methods$	B = 2	B = 3	B = 4	B = 5	B = 2 ...5
CSP	81.4	76.7	75.8	48.9	75.3
CP_1F_{521}	89.7	76.7	80.8	51.1	79.6
CP_1F_{623}	90.4	76.7	79.8	53.3	79.8
$CPbF_{521}$	75.6	47.9	54.6	31.1	57.4
$CPbF_{623}$	73.1	51.4	58.6	37.8	58.5
CL	79.5	78.8	75.8	37.8	74.2

TABLE III
THE ACCURACIES (%) OF HYBRID METHODS WIH CSP, CP_1F_{521}, $CPbF_{623}$, $CPbF_{623}$, CL, $CPbF_{521}$ AND $CPbF_{623}$.

$Method$	B = 2	B = 3	B = 4	B = 5	B = 2 ...5
$CPbF_{521} + CSP$ (θ_1=0.5)	82.7	76.7	75.8	48.9	75.8
$CPbF_{623} + CSP$ (θ_1=0.5)	80.1	76.7	75.8	48.9	74.2
CL + CSP (θ_1=0.5)	**79.5**	**78.8**	**76.8**	**44.4**	**73.5**
$BKS(CP_1F_{521}, CPbF_{521})$	89.7	76.7	75.8	55.6	78.9
$BKS(CP_1F_{521}, CPbF_{623})$	89.4	76.7	76.7	53.3	78.8
$\mathbf{BKS(CP_1F_{521}, CL)}$	**89.7**	**76.7**	**79.8**	**55.6**	**79.8**
$BKS(CP_1F_{623}, CPbF_{521})$	90.4	76.7	76.7	55.6	79.4
$BKS(CP_1F_{623}, CPbF_{623})$	90.4	76.7	76.7	57.8	79.6
$\mathbf{BKS(CP_1F_{623}, CL)}$	**90.4**	**76.7**	**80.8**	**62.2**	**80.9**
CSP+$BKS(CP_1F_{521}, CPbF_{521})$	89.7	81.5	78.8	62.2	81.8
CSP+$BKS(CP_1F_{521}, CPbF_{623})$	89.7	81.5	79.8	62.2	82.0
CSP+BKS($\mathbf{CP_1F_{521}}$, CL)	**89.7**	**82.2**	**81.8**	**62.2**	**82.7**
CSP+$BKS(CP_1F_{623}, CPbF_{521})$	90.4	82.9	79.8	66.7	83.2
CSP+$BKS(CP_1F_{623}, CPbF_{623})$	90.4	82.9	79.8	68.9	83.4
CSP+$BKS(CP_1F_{623}, CL)$	**90.4**	**83.6**	**80.8**	**68.9**	**83.9**

[3] H. Y. Chen, C. B. Yang, K. T. Tseng, and C. Y. Hor, "The disulfide connectivity prediction with support vector machine and behavior knowledge space," *Proc. of International Conference on Knowledge Discovery and Information Retrieval*, Vilamoura, Algarve, Portugal, pp. 112–118, Sept. 19-22, 2013.

[4] J. Cheng, H. Saigo, and P. Baldi, "Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching," *PROTEINS: Structure, Function, and Genetics*, Vol. 62, pp. 617–629, 2006.

[5] C. C. Chuang, C. Y. Chen, J. M. Yang, P. C. Lyu, and J. K. Hwang, "Relationship between protein structures and disulfide-bonding patterns," *PROTEINS: Structure, Function, and Genetics*, Vol. 53, pp. 1–5, 2003.

[6] R. Ed, "A C program to solve maximum weight matching," 1999. Software available at http://elib.zib.de/pub/Packages/mathprog/.

[7] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology*, Vol. 292, No. 2, pp. 195–202, 1999.

[8] H. H. Lin, J. C. Hsu, Y. N. Hsu, R. H. Pan, Y. F. Chen, and L. Y. Tseng, "Disulfide connectivity prediction based on structural information without a prior knowledge of the bonding state of cysteines," *Computers in Biology and Medicine*, Vol. 43, pp. 1941–1948, 2013.

[9] H. H. Lin and L. Y. Tseng, "Prediction of disulfide bonding pattern based on a support vector machine and multiple trajectory search," *Information Sciences*, Vol. 199, pp. 167–178, 2012.

[10] H. L. Liu and S. C. Chen, "Prediction of disulfide connectivity in proteins with support vector machine," *Journal of the Chinese Institute of Chemical Engineers*, Vol. 38, No. 1, pp. 63–70, 2007.

[11] C. H. Lu, Y. C. Chen, C. S. Yu, and J. K. Hwang, "Predicting disulfide connectivity patterns," *PROTEINS: Structure, Function, and Genetics*, Vol. 67, pp. 262–270, 2007.

[12] G. L. J. Rama, A. P. Shilton, M. M. Parker, and M. Palaniswami, "Prediction of cystine connectivity using svm," *Bioinformation*, Vol. 1, No. 2, pp. 69–74, 2005.

[13] S. Raudys and F. Roli, "The behavior knowledge space fusion method: Analysis of generalization error and strategies for performance improvement," *In Proc. Int. Workshop on Multiple Classifier Systems (LNCS 2709*, pp. 55–64, Springer, 2003.

[14] A. Sali and T. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *Journal of Molecular Biology*, Vol. 1, pp. 779–815, 1993.

[15] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1999.

[16] A. Vullo and P. Frasconi, "Disulfide connectivity prediction using recursive neural networks and evolutionary information," *Bioinformatics*, Vol. 20, No. 5, pp. 653–659, 2004.

[17] C. J. Wang, C. B. Yang, C. Y. Hor, and K. T. Tseng, "Disulfide bond prediction with hybrid models," *Proc. of the 2012 International Conference on Computing and Security (ICCS12)*, Ulaanbaatar, Mongolia, July, 2012.

[18] E. Zhao, H. L. Liu, C. H. Tsai, H. K. Tsai, C. H. Chan, and C. Y. Kao, "Cysteine separations profiles on protein sequences infer disulfide connectivity," *Bioinformatics*, Vol. 21, No. 8, pp. 1415–1420, 2005.