The Application of Support Vector Machine and Behavior Knowledge Space in the Disulfide Connectivity Prediction Problem

Hong-Yu Chen[1], Kuo-Tsung Tseng[2], Chang-Biau Yang[1(✉)], and Chiou-Yi Hor[1]

[1] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
cbyang@cse.nsysu.edu.tw

[2] Department of Shipping and Transportation Management, National Kaohsiung Marine University, Kaohsiung 81157, Taiwan
tsengkt@nkmu.edu.tw

Abstract. In this paper, we apply support vector machine (SVM) and behavior knowledge space (BKS) to the disulfide connectivity prediction problem. The problem aims to establish the disulfide connectivity pattern of the target protein. It is an important problem since a disulfide bond, formed by two oxidized cysteines, plays an important role in the protein folding and structure stability. The disulfide connectivity prediction problem is difficult because the number of possible patterns grows rapidly with respect to the number of cysteines. We discover some rules to discriminate the patterns with high accuracy in various methods. Then, the pattern-wise and pair-wise BKS methods to fuse multiple classifiers constructed by the SVM methods are proposed. Finally, the CSP (cysteine separation profile) method is also applied to form our hybrid method. We perform some simulation experiments with the 4-fold cross-validation on SP39 dataset. The prediction accuracy of our method is increased to 69.1 %, which is better than the best previous result 65.9 %.

Keywords: Disulfide bond · Cysteine · Connectivity pattern · Support vector machine · Behavior knowledge space

1 Introduction

A *disulfide bond*, also called *SS-bond* or *SS-bridge*, is a single covalent bond which is usually formed from the oxidation of two thiol groups (-SH). The transformation is described as

$$2RSH \rightarrow RS\text{-}SR + 2H^+ + 2e^- \tag{1}$$

where R represents the carbon-containing group of atoms.

In proteins, only the thiol groups of cysteine residues can form the disulfide bonds by oxidation. The goal of the *disulfide connectivity prediction* (DCP) problem is to figure out which cysteine pair would be cross-link from all possible

A. Fred et al. (Eds.): IC3K 2013, CCIS 454, pp. 66–79, 2015.
DOI: 10.1007/978-3-662-46549-3_5

candidates. It may be conducive to the solution of protein structure prediction problem if precise disulfide connectivity information is available.

There are two main ways for connectivity pattern prediction in previous works, pair-wise and pattern-wise. The pair-wise method focuses on the bonding potential of each cysteine pair, and encodes the target based on cysteine pairs. The pattern-wise method makes a comprehensive survey of whole connectivity pattern and usually ranks the connectivity patterns by their possibilities, so the prediction ability may be limited to the diversity of patterns in a training set.

The prediction task is difficult because the number of possible connectivity patterns grows rapidly with respect to the number of cysteines. Most previous studies are limited by the number of disulfide bonds from two to five. It is well known that the number of possible patterns is given as follows:

$$N = \frac{C_2^{2B} \times C_2^{2B-2} \times \ldots \times C_2^2}{B!} = (2B-1)!! \tag{2}$$

where B denotes the number of disulfide bonds in the protein. For instance, if we have known which cysteines are oxidized in advance, $N = 945$ when $B = 5$, and N is up to 10395 when $B = 6$.

Some statistical analyses [1–4] have been applied to the disulfide connectivity prediction problem. Many researchers tried to solve the problem with machine learning methods, such as neural network (NN) [5–10] and support vector machine (SVM) [2,11–17].

Before 2005, many studies [5,8] were devoted to the connectivity prediction, but most of their accuracies are below 50 %. In 2005, Zhao *et al.* [18] utilized the global information of a protein, called *cysteine separation profile* (CSP), which represents the separations among all oxidized cysteines in a protein sequence. In 2007, Lu *et al.* [2] proposed a novel concept of the CP_2 representation, which uses every two cysteine pairs (four cysteines) as one sample, and applied the genetic algorithm (GA) to the optimization of feature selection.

In 2012, Wang *et al.* [19] built a hybrid model based on SVM and the weighted graph matching, with accuracy 65.9 %. They extracted different feature sets depending on whether the number of disulfide bonds in a protein is odd or even. The main difference of the feature sets for the two submodels is the secondary structure information around the oxidized cysteines.

The rest of this paper is organized as follows. We introduce some preliminary knowledge, including support vector machine and behavior knowledge space in Sect. 2. In Sect. 3, we present our hybrid method for solving the DCP problem. The experimental results are given in Sect. 4, and we also describe the performance comparison between our method and some previous works. Finally, our conclusion is given in Sect. 5.

2 Preliminary

In this section, we introduce some background knowledge used in this paper, including support vector machine and behavior knowledge space.

2.1 Support Vector Machine

Support Vector Machine (SVM) is a machine learning method for classification and regression. It was first introduced by Vapnik [20] in 1999. SVM seeks to create a hyperplane to discriminate different labels of the vectors in the training set and utilizes the model to predict the labels of other data. To discover the discriminative features is the key point for applying SVM. Figure 1 shows an example of the SVM solution with maximum *margin* which means the distance between the hyperplane and the given objects.

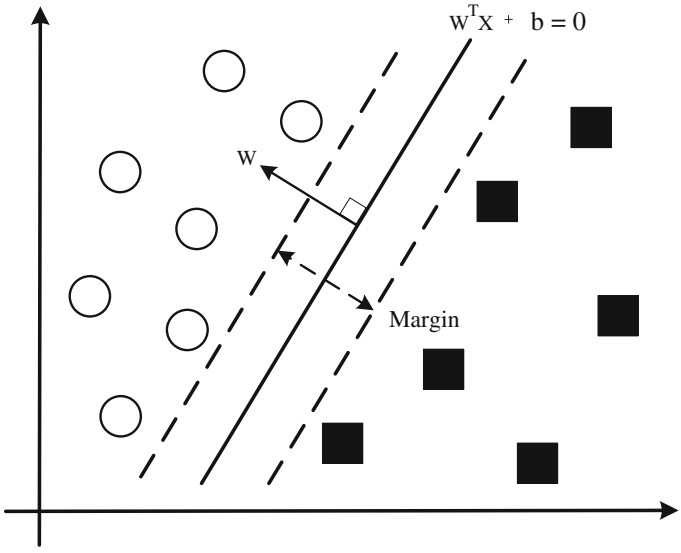

Fig. 1. An example of the SVM solution with maximum margin.

For SVM implementation, we use the `LIBSVM` package [21] which is an easy-to-use tool for *support vector classification* (SVC) and *support vector regression* (SVR). The SVC function classifies the data with their probabilities, and the SVR function generates the regression value of each target data element.

2.2 Behavior Knowledge Space

Behavior knowledge space (BKS) [22] is a kind of method for fusing multiple classifiers. It is a table look-up approach for estimating the probability of every vote combination. Assume there are m classifiers composing an ensemble for a classification task of n labels. The BKS table contains n^m entries, the number of all possible combinations of m classifiers' outputs. And each entry records the distribution of n true labels in the training set.

Table 1 illustrates an example of the BKS table for the 3-label classification problem with two classifiers. The 'C1' and 'C2' represent the predicted

outputs from the two classifiers, and the entries below them are all possible prediction combinations. Cells below 'Real label', 'L1', 'L2', and 'L3', are the distribution of the true labels associated with the predicted label vectors. For example, when 'C1'='L1' and 'C2'='L3', the fused answer should be 'L2' since it is the most possible label. And, if we have 'C1'='L3', 'C2'='L2', the fused answer should go to 'L3'.

Table 1. An example of the BKS table.

Predicted label		Real label		
C1	C2	L1	L2	L3
L1	L1	**23**	8	2
L1	L2	**5**	0	4
L1	L3	2	**7**	1
⋮	⋮	⋮	⋮	⋮
L3	L2	1	1	**5**
L3	L3	1	3	**12**

3 Algorithms for Connectivity Prediction

We observe that the prediction accuracies of Chung *et al.* [23] and Wang *et al.* [19] are 63.5 % and 65.9 %, respectively. It may be hard to find more features with good discrimination capability for a single SVM method in the connectivity prediction. However, we may get better accuracies if we fuse the advantages of the multiple models.

Our method utilizes BKS to fuse the results obtained from SVM models. The features and cysteine-pair representation we adopted are inspired by Wang *et al.* [19] and Lu *et al.* [2]. In addition, we also combine the CSP method [18] to our hybrid method.

3.1 Feature Extraction

In the past, the bonding states of each cysteine pair are usually used to describe the disulfide pattern and used as the samples of SVM. Lu *et al.* [2] call it as the CP_1 representation. Lu *et al.* further proposed a novel concept of the CP_2 representation which use every two cysteine pairs (four cysteines) as the samples. In our method, we adopt the features used by Wang *et al.* [19]. In addition, we encode the CP_2 representation as the *permutation order*, which is also included in our feature set. The definition of the permutation order is given as follows.

Permutation Order: This feature implies the order of feature extraction in each cysteine window. For every cysteine-pair combination in the CP_2 representation, we encode the samples in three permutations illustrated in Table 2. For example, C_1-C_3-C_2-C_4 means that the first and third cysteines form a disulfide bond in these four cysteines, and the second and fourth form the other bond. This bond pattern is represented by the feature vector $(0.25, 0.75, 0.5, 1)$.

Table 2. The feature vector of the permutation order.

Permutations	Feature vector
C_1-C_2-C_3-C_4	(0.25, 0.5, 0.75, 1)
C_1-C_3-C_2-C_4	(0.25, 0.75, 0.5, 1)
C_1-C_4-C_2-C_3	(0.25, 1, 0.5, 0.75)

3.2 SVM Method

We implement three SVM models with different feature sets, CP_1F_{521}, CP_1F_{623} and CP_2Label_2, as shown in Table 3. These features are encoded by the segments of every cysteine pair. The cysteine segment is a window centering at a target cysteine. Many previous works [2,4,8,9,12,14,16,17,24–26] also adopted the similar idea of the window approach. Here we set the window size to 13. In other words, $2k + 1 = 13$. So there are 521 features in CP_1F_{521} and 623 features in CP_1F_{623}.

Table 3. The feature sets used in our three models.

Feature	size	M[a]	M[b]	M[c]
Distance of cysteines	1	Y	Y	Y
Cysteine order	2		Y	
Protein weight	1		Y	
Protein length	1		Y	
Amino acid composition	20		Y	
PSSM around cysteine	$(2k+1) \times 20 \times 2$	Y	Y	Y
Secondary structure around cysteine	$(2k+1) \times 3 \times 2$		Y	
Permutation order	4			Y

[a] CP_1F_{521} model.
[b] CP_1F_{623} model.
[c] CP_2Label_2 model.

Table 4. The details of the probability intervals for the BKS method, where B denotes the number of bonds in a target protein.

B	Type of BKS	Probability intervals
2	Pattern-wise	(0, 0.15, 0.2, 0.25, 0.3, 0.35, 0.5, 1)
3	Pattern-wise	(0, 0.25, 0.5, 1)
4	Pair-wise	(0, 0.1, 0.2, 0.3, 0.4, 0.5, 1)
5	Pair-wise	(0, 0.1, 0.2, 0.3, 0.4, 0.5, 1)

3.3 BKS Method

We adopt the concept of the behavior knowledge space (BKS) to fuse the above SVM classifiers. We design two BKS models, pattern-wise BKS and pair-wise BKS, combined with the probability intervals, where the probabilities are obtained from the prediction of SVM classifiers. The details of the probability intervals for the proteins with various number of disulfide bonds are shown in Table 4.

Pattern-Wise BKS Method. After the two classifiers CP_1F_{521} and CP_1F_{623} finish the pattern prediction, the probability of each bonding pattern is obtained. Then, the pattern-wise BKS is constructed according to the prediction probabilities. We adopt the pattern-wise BKS method for the prediction of proteins with two or three bonds. Table 5 illustrates an example of the partial pattern-wise BKS table for 2-bond proteins. For example, in the second row, the probabilities of the predicted pattern 1-1-2-2 for the two classifiers locate in $(0.15, 0.2)$. In this case, 5, 3 and 1 proteins have the true patterns 1-1-2-2, 1-2-1-2 and 1-2-2-1, respectively. Thus, the fused answer is decided to be 1-1-2-2.

We set the threshold of the patterns supported in the pattern-wise BKS table to 2, and reject to give an answer in the case below the threshold. Table 6 shows some examples for 3-bond proteins whose prediction can be corrected by the pattern-wise BKS method.

Table 5. An example of the partial pattern-wise BKS table for 2-bond proteins.

CP_1F_{521}	Interval	CP_1F_{623}	Interval	1-1-2-2	1-2-1-2	1-2-2-1
1-1-2-2	(0.15, 0.2)	1-1-2-2	(0, 0.15)	0	1	0
1-1-2-2	(0.15, 0.2)	1-1-2-2	(0.15, 0.2)	**5**	3	1
1-1-2-2	(0.15, 0.2)	1-1-2-2	(0.2, 0.25)	**4**	0	0
1-1-2-2	(0.15, 0.2)	1-1-2-2	(0.25, 0.3)	0	0	0
1-1-2-2	(0.15, 0.2)	1-1-2-2	(0.3, 0.35)	0	0	0
1-1-2-2	(0.15, 0.2)	1-1-2-2	(0.35, 0.5)	1	0	0
1-1-2-2	(0.15, 0.2)	1-1-2-2	(0.5, 1)	0	0	0

Table 6. Examples for 3-bond proteins corrected by the pattern-wise BKS method.

Proteins	Real patterns	CP_1F_{521}	CP_1F_{623}	Predicted by BKS
CXOA_CONMA	1-2-3-1-2-3	1-2-1-3-2-3	1-2-1-3-2-3	1-2-3-1-2-3
HST1_ECOLI	1-2-3-1-2-3	1-2-1-3-2-3	1-2-1-3-2-3	1-2-3-1-2-3
HCYA_PANIN	1-1-2-2-3-3	1-1-2-2-3-3	1-1-2-3-3-2	1-1-2-2-3-3
CXOB_CONST	1-2-3-1-2-3	1-2-1-3-2-3	1-2-1-3-2-3	1-2-3-1-2-3

Pair-Wise BKS Method. The pattern-wise BKS method is not suitable for the prediction of all proteins. The number of all possible combinations of patterns grows rapidly with respect to the number of bonds, so the number of the training samples is relatively not enough. We propose the pair-wise BKS method for the prediction of proteins with four or five bonds. The pair-wise BKS table records the ratio of the pairs truly bonded or not in various probability intervals from the two classifiers, CP_1F_{521} and CP_2Label_2. Table 7 shows an example of the partial pair-wise BKS table for 5-bond proteins. For every cysteine pair, we advisably adjust the original probability from CP_1F_{521} method according to the ratio of the truly bonded pairs in the pair-wise BKS table. Table 8 illustrates the adjustment rules. Eventually, the predicted pattern is derived from the top N maximum weighted graph matching by the adjusted weighted matrix until the matching pattern belongs to a real pattern in PDB dataset according to our statistics, where the probabilities obtained by SVM classifiers are input as the edge weights in graph matching.

Table 7. An example of the partial pair-wise BKS table for 5-bond proteins.

Pairs from CP_1F_{521}	Pairs from CP_2Label_2	Truly bonded	Not bonded
(0.3, 0.4)	(0, 0.1)	0	0
(0.3, 0.4)	(0.1, 0.2)	0	1
(0.3, 0.4)	(0.2, 0.3)	6	5
(0.3, 0.4)	(0.3, 0.4)	4	6
(0.3, 0.4)	(0.4, 0.5)	6	6
(0.3, 0.4)	(0.5, 1)	1	12

Table 8. The adjustment rules for the pair-wise BKS method.

Truly bonded	Not bonded	Adjustment ratio
0	0	1
[1, 10)	0	4
≥ 10	0	8
0	[1, 10)	0.25
0	≥ 10	0.125
x	x	1
[x, 2x)	x	2
$\geq$ 2x	x	4
x	[x, 2x)	0.5
x	$\geq$ 2x	0.25

3.4 Hybrid Method

Instead of a large amount of features used by the SVM method, Zhao *et al.* [18] adopted only one feature, CSP (cysteine separations profile), to achieve nearly 50 % accuracy in the dataset with insufficient information. The CSP of protein x with $2n$ oxidized cysteines (n disulfide bonds) is defined as

$$CSP_x = (\delta 1, \delta_2, \ldots, \delta_{2n-1}) = (\rho_2 - \rho_1, \rho_3 - \rho_2, \ldots, \rho_{2n} - \rho_{2n-1}) \tag{3}$$

where ρ_i denotes the sequence position of the ith oxidized cysteine in the protein and δ_i denotes the separation distance between oxidized cysteines i and $i+1$.

The divergence (D) of two CSPs for two proteins x and y is defined [18] as follows:

$$D = \sum_{i=1}^{i=2n-1} |\delta_{x,i} - \delta_{y,i}|. \tag{4}$$

It has been shown that the CSP is an important global feature for the disulfide connectivity prediction, so we also combine the CSP method into our hybrid method. Figure 2 exhibits the flow chart of our work. Our hybrid method for predicting the disulfide connectivity pattern is described as follows.

Algorithm. The hybrid method.
Input: A protein sequence and the bonding states of its all cysteines.
Output: The predicted disulfide connectivity pattern.
Case 1: For a 2-bond or 3-bond protein.

- Step 1.1: Apply the maximum weighted graph matching algorithm to derive the pattern from the CP_1F_{521} method for 2-bond proteins (the CP_1F_{623} method for 3-bond proteins). If the normalized weight of one pattern is greater than or equal to 0.5, report this pattern as the predicted pattern.
- Step 1.2: If the condition meets a predefined threshold in the pattern-wise BKS method, report this pattern as the predicted pattern.
- Step 1.3: If the minimum divergence obtained by the CSP search is less than or equal to a predefined threshold, report this pattern as the predicted pattern.
- Step 1.4: For the remaining, take the original maximum weighted pattern from the CP_1F_{521} method for 2-bond proteins (the CP_1F_{623} method for 3-bond proteins) as the predicted result.

Case 2: For a 4-bond or 5-bond protein.

- Step 2.1: Apply the maximum weighted graph matching algorithm to derive the pattern from the CP_1F_{521} method. If the normalized weight of one pattern is greater than or equal to 0.5, report this pattern as the predicted pattern.
- Step 2.2: If the minimum divergence obtained by the CSP search is less than or equal to a predefined threshold, report this pattern as the predicted pattern.
- Step 2.3: Adjust the original weight (probability) of each pair from the CP_1F_{521} method according to the pair-wise BKS table and report the pattern derived from weighted graph matching algorithm as the predicted result.

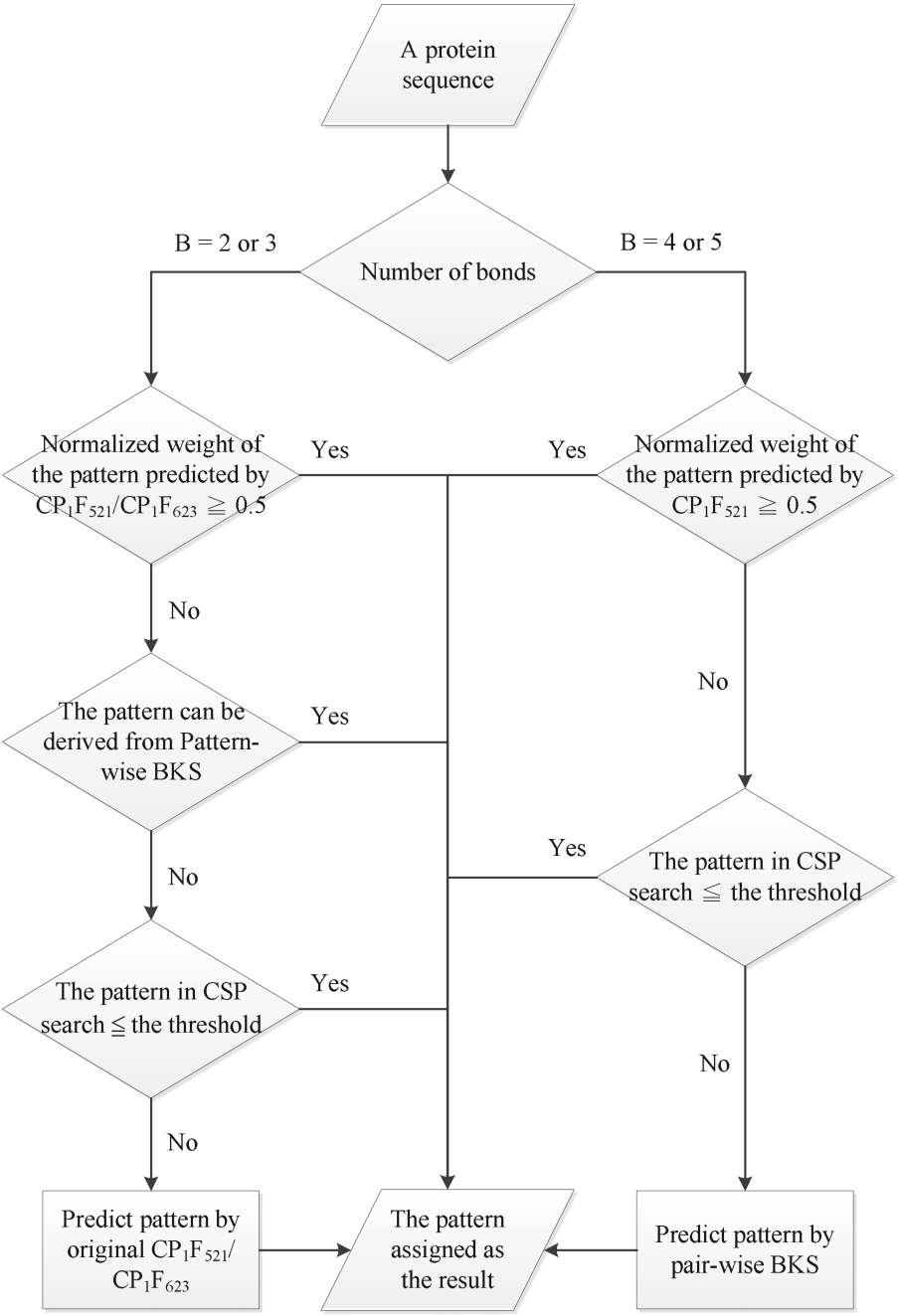

Fig. 2. The system flow chart of our method.

4 Experimental Results

In this section, we introduce our testing dataset and performance evaluation criteria of the disulfide connectivity prediction. In addition, we show the experimental results of various methods.

4.1 Dataset

For the fair comparison of the prediction accuracy with previous works, we use SP39 dataset, which is the same dataset adopted in some previous works, for our training and testing. Table 9 illustrates the summary of SP39 dataset. This dataset contains 446 proteins with two to five disulfide bonds, derived from the SWISS-PROT release no. 39. It was first used by Vullo and Frasconi [10]. We also use the same way as Wang *et al.*'s [19] to divide SP39 dataset into four subsets for the 4-fold cross-validation. The sequence identity of proteins between any two subsets is less than 30 %.

Table 9. The summary of SP39 dataset.

	Number of proteins by the number of bonds					Number of cysteines	
	$B = 2$	$B = 3$	$B = 4$	$B = 5$	$B = 2 \cdots 5$	Oxidized	Total
SP39[a]	156	146	99	45	446	2742	4401

[a] Defined by Vullo and Frasconi [10].

4.2 Performance Evaluation

The definition of k-fold cross-validation is given as follows. A dataset D is divided into k subsets $D_1, D_2, \ldots, D_k$, which are disjoint to each other. Each time, we take a subset D_i, $1 \leq i \leq k$, as the testing set and use the other $k-1$ subsets for training. Repeat this procedure k times until each subset is tested once. Here, we adopt the 4-fold cross-validation. For the measurement of the performance in connectivity pattern prediction, the accuracy is calculated as follows:

$$Q_p = \frac{C_p}{T_p}, \tag{5}$$

where C_p denotes the number of proteins whose connectivity patterns are correctly predicted, and T_p is the total number of proteins for testing.

4.3 Results

In the CP_1F_{521} method, combined by the SVM method with the maximum weighted graph matching [27], we discover that the prediction accuracy is very high when the normalized weight of one predicted pattern is greater than or equal to 0.5 (half). Table 10 shows the ratio and accuracies of these patterns. In other words, the confidence of such prediction is very high. Thus, in Step 1.1 or Step 2.1 of our method, the answer is settled down for these predictions.

Table 10. The accuracy and ratio of the predicted patterns, whose normalized weights are greater than or equal to 0.5 (half) by CP_1F_{521} method in SP39 dataset.

	$B = 2$	$B = 3$	$B = 4$	$B = 5$	$B = 2 \cdots 5$
Accuracy (Q_p)	100	93.0	93.9	92.9	96.0
Ratio in the dataset	37.8	29.5	33.3	31.1	33.4

Table 11. The accuracy and ratio of the pattern-wise BKS method in SP39 dataset when the threshold is set to 2.

	$B = 2$	$B = 3$
Accuracy (Q_p)	94.0	95.7
Ratio in the dataset	53.2	31.5

Table 12. The Q_p and ratio of the divergence of the CSP in SP39 dataset.

	$B = 2$		$B = 3$		$B = 4$		$B = 5$	
CSP	Q_p	Ratio	Q_p	Ratio	Q_p	Ratio	Q_p	Ratio
0	100	13	100	5	100	2	N/A	0
≤ 5	90	38	96	34	100	24	100	11
≤ 10	78	54	77	57	84	31	100	13
≤ 15	71	67	72	66	73	37	100	13
≤ 20	73	71	72	66	52	55	62	29
≤ 25	71	76	71	66	48	59	53	33

We use the BKS as a supporting role in our method. For the two kinds of the BKS, the performance of the pattern-wise BKS is more effective than the pair-wise BKS. Table 11 illustrates the details of the pattern-wise BKS method when we set the threshold to 2 for 2-bond and 3-bond proteins.

Our hybrid method is combined by the SVM method and the BKS method. In addition, we also take the CSP method into consideration. Table 12 shows the Q_p and ratio of the divergence of the CSP. Take 3-bond proteins as an example. There are 34 % proteins with CSP search less than or equal to 5, and the Q_p of these proteins reaches up to 96 %. According to the observation, we set the applicable thresholds of CSP to pick out the patterns as results. Here, we set the threshold of CSP to 0, 5, 10, and 15 for proteins with two to five bonds, respectively.

Table 13 shows the Q_p of our methods and some previous works in SP39 dataset. The accuracies of the three SVM models are derived from the patterns with the maximum weighted graph matching. However, we find that it is hard to improve the accuracy by one single SVM model. Although the performance of CP_2Label_2 is not better than CP_1F_{521} or CP_1F_{623}, CP_2Label_2 provides the effect for pair-wise BKS since CP_2Label_2 represents another concept of pair extraction. Eventually, the prediction accuracy of our hybrid method with SVM and BKS reaches 65.9 %, and up to 69.1 % combined with CSP method.

Table 13. The Q_p of our methods and previous works in SP39 dataset.

Method	$B = 2$	$B = 3$	$B = 4$	$B = 5$	$B = 2 \cdots 5$
CSP[a]	72.4	54.1	33.3	17.8	52.2
Wang's method[b]	84.0	60.3	55.6	44.4	65.9
CP_1F_{521}	84.0	53.4	55.6	46.7	63.9
CP_1F_{623}	78.2	60.3	53.5	44.4	63.5
CP_2Label_2	75.0	49.3	52.5	40.0	58.1
CP_1F_{521} + BKS	84.0	56.8	55.6	55.6	65.9
CP_1F_{521} + BKS + CSP	84.0	64.4	57.6	57.8	69.1

[a]Proposed by Zhao *et al.* [18].
[b]Proposed by Wang *et al.* [19].

5 Conclusion

According to the study of Wang *et al.* [19], which focuses SVM models on varied features, and the concept of different cysteine-pair representations proposed by Lu *et al.* [2], we do many integrated experiments. However, the improvement of the pure SVM methods is not so significant although the SVM method is still relatively better, compared with other machine learning methods. Some studies [28,29] combine the SVM method with CSP or sequence alignment to raise the accuracy. The key step of the CSP method and the sequence alignment method is to search for a good template set. However, the accuracy of these two methods depends on the pattern varieties in the template set.

We think that the design of hybrid methods is the trend in the disulfide connectivity prediction problem. In this paper, we gather some statistics about the disulfide bonds, and have successfully found some rules to discriminate the patterns with high accuracy in several methods. Furthermore, we adopt the pattern-wise and pair-wise BKS methods to fuse multiple SVM models, and use the predicted patterns from the original SVM method for the rest of proteins.

In the future, we may examine our hybrid method to other datasets, and explore more methods for fusing multiple classifiers, such as the weighted majority vote. We may try the CSP method with the inter-bond template dataset to explore more possibilities of the development with the concept of subpatterns.

Acknowledgements. This research work was partially supported by the National Science Council of Taiwan under contract NSC 100-2221-E-242-003.

References

1. Harrison, P.M., Sternberg, M.J.E.: Analysis and classification of disulphide connectivity in proteins: the entropic effect of cross-linkage. J. Mol. Biol. **244**(4), 448–463 (1994)

2. Lu, C.-H., Chen, Y.-C., Yu, C.-S., Hwang, J.-K.: Predicting disulfide connectivity patterns. Proteins Struct. Funct. Genet. **67**, 262–270 (2007)
3. Mirny, L.A., Shakhnovich, E.I.: How to derive a protein folding potential? a new approach to an old problem. J. Mol. Biol. **264**(5), 1164–1179 (1996)
4. Rubinstein, R., Fiser, A.: Predicting disulfide bond connectivity in proteins by correlated mutations analysis. Bioinformatics **24**(4), 498–504 (2008)
5. Baldi, P., Cheng, J., Vullo, A.: Large-scale prediction of disulphide bond connectivity. In: Saul, L., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems, vol. 17, pp. 97–104. MIT Press, Cambridge (2005)
6. Cheng, J., Saigo, H., Baldi, P.: Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. Proteins Struct. Funct. Genet. **62**, 617–629 (2006)
7. Fariselli, P., Riccobelli, P., Casadio, R.: Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins Struct. Funct. Genet. **36**, 340–346 (1999)
8. Ferre, F., Clote, P.: Disulfide connectivity prediction using secondary structure information and diresidue frequencies. Bioinformatics **21**(10), 2336–2346 (2005)
9. Martelli, P.L., Fariselli, P., Malaguti, L., Casadio, R.: Prediction of the disulfide-bonding state of cysteines in proteins at 88 % accuracy. Protein Sci. **11**, 2735–2739 (2002)
10. Vullo, A., Frasconi, P.: Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics **20**(5), 653–659 (2004)
11. Chen, Y.-C., Lin, Y.-S., Lin, C.-J., Hwang, J.-K.: Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. Proteins Struct. Funct. Genet. **55**, 1036–1042 (2004)
12. Chen, Y.-C., Hwang, J.-K.: Prediction of disulfide connectivity from protein sequences. Proteins Struct. Funct. Genet. **61**, 507–512 (2005)
13. Frasconi, P., Passerini, A., Vullo, A.: A two-stage svm architecture for predicting the disulfide bonding state of cysteines. In: Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, pp. 25–34 (2002)
14. Jayavardhana Rama, G.L., Shilton, A.P., Parker, M.M., Palaniswami, M.: Prediction of cystine connectivity using svm. Bioinformation **1**(2), 69–74 (2005)
15. Liu, H.-L., Chen, S.-C.: Prediction of disulfide connectivity in proteins with support vector machine. J. Chin. Inst. Chem. Eng. **38**(1), 63–70 (2007)
16. Tsai, C.-H., Chen, B.-J., Chan, C.-H., Liu, H.-L., Kao, C.-Y.: Improving disulfide connectivity prediction with sequential distance between oxidized cysteines. Bioinformatics **21**(24), 4416–4419 (2005)
17. Vincent, M., Passerini, A., Labbe, M., Frasconi, P.: A simplified approach to disulfide connectivity prediction from protein sequences. BMC Bioinform. **9**(1), 20 (2008)
18. Zhao, E., Liu, H.-L., Tsai, C.-H., Tsai, H.-K., Chan, C.-H., Kao, C.-Y.: Cysteine separations profiles on protein sequences infer disulfide connectivity. Bioinformatics **21**(8), 1415–1420 (2005)
19. Wang, C.-J., Yang, C.-B., Hor, C.-Y., Tseng, K.-T.: Disulfide bond prediction with hybrid models. In: Proceedings of the 2012 International Conference on Computing and Security (ICCS 2012), Ulaanbaatar, Mongolia, July 2012
20. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1999)
21. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines (2001). http://www.csie.ntu.edu.tw/cjlin/libsvm

22. Raudys, S., Roli, F.: The behavior knowledge space fusion method: analysis of generalization error and strategies for performance improvement. In: Windeatt, T., Roli, F. (eds.) MCS 2003. LNCS, vol. 2709, pp. 55–64. Springer, Heidelberg (2003)
23. Chung, W.-C., Yang, C.-B., Hor, C.-Y.: An effective tuning method for cysteine state classification. In: Proceedings of National Computer Symposium, Workshop on Algorithms and Bioinformatics, Taipei, Taiwan, 27–28 November 2009
24. Chen, G., Deng, H., Gui, Y., Pan, Y., Wang, X.: Cysteine separations profiles on protein secondary structure infer disulfide connectivity. In: 2006 IEEE International Conference on Granular Computing, pp. 663–665, May 2006
25. Chuang, C.-C., Chen, C.-Y., Yang, J.-M., Lyu, P.-C., Hwang, J.-K.: Relationship between protein structures and disulfide-bonding patterns. Proteins Struct. Funct. Genet. **53**, 1–5 (2003)
26. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292**(2), 195–202 (1999)
27. Fariselli, P., Casadio, R.: Prediction of disulfide connectivity in proteins. Bioinformatics **17**(10), 957–964 (2001)
28. Chen, B.-J., Tsai, C.-H., Chan, C.-H., Kao, C.-Y.: Disulfide connectivity prediction with 70 % accuracy using two-level models. Proteins Struct. Funct. Genet. **64**, 246–252 (2006)
29. Chen, Y.-C.: Prediction of Disulfide Connectivity from Protein Sequences. Ph.D. dissertation, National Chiao Tung University, Hsinchu, Taiwan (2007)