Keywords: Disulfide Bond, Cysteine, Connectivity Pattern, Support Vector Machine, Behavior Knowledge Space.

Abstract: A disulfide bond, formed by two oxidized cysteines, plays an important role in the protein folding and structure stability, and it may regulate protein functions. The disulfide connectivity prediction problem is to reveal the correct information of disulfide connectivity in the target protein. It is difficult because the number of possible patterns grows rapidly with respect to the number of cysteines. In this paper, we discover some rules to discriminate the patterns with high accuracy in various methods. Then, we propose the pattern-wise and pair-wise BKS (behavior knowledge space) methods to fuse multiple classifiers constructed by the SVM (support vector machine) methods. Furthermore, we combine the CSP (cysteine separation profile) method to form our hybrid method. The prediction accuracy of our hybrid method in SP39 dataset with 4-fold cross-validation is increased to 69.1%, which is better than the best previous result 65.9%.

1 INTRODUCTION

A disulfide bond, also called SS-bond or SS-bridge, is a single covalent bond, and it is usually formed from the oxidation of two thiol groups. 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 candidates. It may be conducive to the solution of the protein structure prediction problem if precise disulfide connectivity information is available.

There are two main ways for solving the DCP problem 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 according to cysteine pairs. The pattern-wise method makes a comprehensive survey of the whole connectivity pattern and it usually ranks the connectivity patterns, so the prediction ability may be limited to the diversity of patterns in a training set.

The pattern-wise DCP task is difficult because the number of possible connectivity patterns grows rapidly with respect to the number of cysteines. The number of possible patterns is given as follows:

$$N = \frac{C_{B}^{2B} \times C_{B}^{2B-2} \times \ldots \times C_{B}^{2}}{B!} = (2B - 1)!!$$  \hspace{1cm} (1)

where \(B\) is the number of disulfide bonds in the protein. For instance, if the oxidized state of each cysteine is known in advance, then \(N = 945\) when \(B = 5\), and \(N\) is up to 10395 when \(B = 6\). Thus, most studies restrict the number of disulfide bonds to be from two to five.

Some statistical analysis (Paul M. Harrison and Michael J. E. Sternberg, 1994; Chih-Hao Lu et al., 2007; Leonid A. Mirny and Eugene I. Shakhnovich, 1996; Rotem Rubinstein and Andras Fiser, 2008) have been applied to the DCP problem. Many researchers tried to solve the problem with machine learning methods such as neural network (NN) (Pierre Baldi et al., 2005; Jianlin Cheng et al., 2006; Piero Fariselli et al., 1999; F. Ferre and P. Cotte, 2005; Pier Luigi Martelli et al., 2002; Alessandro Vullo and Paolo Frasconi, 2004; Castrense Savojardo et al., 2013) and support vector machine (SVM) (Yu-Ching Chen et al., 2004; Yu-Ching Chen and Jenn-Kang Hwang, 2005; P. Frasconi et al., 2002; Jayavardhana Rama G. L. et al., 2005; Hsuan-Liang Liu and Shih-Chieh Chen, 2007; Chih-Hao Lu et al., 2007; Chi-Hung Tsai et al., 2005; Marc Vincent et al., 2008).

Before 2005, many studies (Pierre Baldi et al., 2005; F. Ferre and P. Cotte, 2005) were devoted to the DCP problem, but most of their accuracies are less than 50%. In 2005, Zhao et al. (East Zhao et al., 2005) utilized the global information in a pro-
tein, called cysteine separation profile (CSP), which is the separations among all oxidized cysteines on a protein sequence.

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. (Chih-Hao Lu et al., 2007) call this type of representation of the disulfide pattern as the CP₁ representation. In 2007, Lu et al. further proposed a novel concept of the CP₂ representation which use every two cysteine pairs (four cysteines) as the samples, and applied the genetic algorithm (GA) to the optimization of feature selection.

In 2012, Wang et al. (Chong-Jie Wang et al., 2012) proposed a hybrid model based on SVM and the weighted graph matching (Piero Fariselli and Rita Casadio, 2001), 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 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 related tools and previous works of the DCP problem in Section 2. In Section 3, we describe our hybrid method for solving the DCP problem. Our experimental results are shown in Section 4, and we also compare the prediction accuracy of our method with the previous works. Finally, our conclusion are given in Section 5.

2 PRELIMINARY

In this section, we introduce some background knowledge for this paper, including the Position-Specific Scoring Matrix (PSSM), support vector machine (SVM) and behavior knowledge space (BKS).

2.1 Position-specific Scoring Matrix

Position-Specific Scoring Matrix (PSSM) (Stephen F. Altschul et al., 1997), also called profile, is a scoring matrix derived from a group of aligned protein sequences. It represents the similarity of residues in every specific position of a query sequence (a target protein) according to the alignment result of the query sequence and the others (probes) in database. Basically, PSSM is a matrix of size $N \times 20$, where $N$ denotes the length of a query sequence and every residue in the query sequence contains a 20-element vector. The 20-element vector respectively represents the scores of 20 standard amino acids which are substituted for the position-specific residue of the query sequence.

2.2 Support Vector Machine

Support vector machine (SVM) is a machine learning method for classification and regression. It was first introduced by Vapnik (Vladimir N. Vapnik, 1999) in 1999. SVM seeks to create a hyperplane to discriminate different labels of the data elements (vectors) in the training set and utilizes the model to predict the labels of target data elements. Each vector is considered as a point in the feature space, and each dimension of the coordinates represents one kind of features. To discover the discriminative features is the key point of SVM.

For SVM implementation in this paper, we use the LIBSVM package (Chih-Chung Chang and Chih-Jen Lin, 2011), 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.

2.3 Behavior Knowledge Space

Behavior knowledge space (BKS) (Sarunas Raudys and Fabio Roli, 2003) is a method for fusing multiple classifiers. It builds a look-up table for estimating the posterior probabilities and every combination of votes. 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 a 3-label classification problem with two classifiers. The ‘C1’ and ‘C2’ represent the outputs from the two classifiers, and the entries below them show all nine 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 predicted answer of the ensemble is ‘L2’ since it is the most possible label. As another example, if we have ‘C1’ = ‘L3’ and ‘C2’ = ‘L2’, the answer goes to ‘L3’.

3 ALGORITHMS FOR CONNECTIVITY PREDICTION

The prediction accuracies of Chung et al. (Wei-Chun...
3.1 Feature Extraction

We follow the features used by Wang et al. (Chong-Jie Wang et al., 2012) and add the new feature permutation order for the model of CP2 representation. The definition of 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 CP2 representation, we encode the samples in three permutations illustrated in Table 2. For example, C1-C2-C3-C4 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 (0.25,0.75,0.5,1).

3.2 SVM Method

We implement three SVM models with different features, CP1F521, CP1F623 and CP2Label2. Table 3 shows the feature set used in each model. 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 (Yu-Ching Chen and Jenn-Kang Hwang, 2005; Guantao Chen et al., 2006; Chao-Chun Chuang et al., 2003; F. Ferre and P. Clote, 2005; David T Jones, 1999; Jayavardhana Rama G. L. et al., 2005; Chih-Hao Lu et al., 2007; Pier Luigi Martelli et al., 2002; Rotem Rubinstein and Andras Fiser, 2008; Chi-Hung Tsai et al., 2005; Marc Vincent et al., 2008) also adopted the similar idea of the window approach. Here we set the window size to 13.

3.3 BKS Methods

We adopt the BKS concept to fuse the classifiers mentioned above. We design two BKS models, pattern-wise BKS and pair-wise BKS, combined with the probability intervals. The probability intervals for predicting different proteins are illustrated in Table 4.

The pattern-wise BKS is constructed from the combinations of the predicted pattern probabilities of two classifiers, CP1F521 and CP1F623. The pattern-wise BKS method is used for the prediction of proteins with 2 or 3 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 both 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 predicted answer would be 1-1-2-2. We set the threshold of the pattern support in the pattern-wise BKS table to 2, and reject to
Table 4: The probability intervals for BKS methods with various numbers of bonds, denoted by $B$.

<table>
<thead>
<tr>
<th>$B$</th>
<th>Type of BKS</th>
<th>Probability intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Pattern-wise</td>
<td>$(0, 0.15, 0.2, 0.25, 0.3, 0.35, 0.5, 1)$</td>
</tr>
<tr>
<td>3</td>
<td>Pattern-wise</td>
<td>$(0, 0.15, 0.2, 0.25, 0.3, 0.35, 0.5, 1)$</td>
</tr>
<tr>
<td>4</td>
<td>Pair-wise</td>
<td>$(0, 0.1, 0.2, 0.3, 0.4, 0.5, 1)$</td>
</tr>
<tr>
<td>5</td>
<td>Pair-wise</td>
<td>$(0, 0.1, 0.2, 0.3, 0.4, 0.5, 1)$</td>
</tr>
</tbody>
</table>

give an answer in the case below the threshold. Table 6 shows some real examples for 3-bond proteins whose prediction can be corrected by the pattern-wise BKS method, while the original predictions made by the classifiers are wrong.

However, the pattern-wise BKS method is not suitable for the prediction of every protein. 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 then adopt the pair-wise BKS method for the prediction of proteins with 4 or 5 bonds. The pair-wise BKS table records the numbers of the truly bonded pairs and non-bonded pairs in various probability intervals from two classifiers, $CP_{1,F_{521}}$ and $CP_{2,Label_2}$. 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 $CP_{1,F_{521}}$ method according to the ratio of the truly bonded pairs in the pair-wise BKS table. As the experimental results show in the next section, we get better prediction accuracies if we adopt different methods to solve the DCP problem with different numbers of bonds.

### 3.4 The Hybrid Method

Instead of large amount of features used by the SVM method, Zhao et al. (East Zhao et al., 2005) adopted only one feature, CSP (cysteine separations profile), to achieve nearly 50% accuracy in the insufficient dataset. The CSP of protein $x$ with $2n$ oxidized cysteines ($n$ disulfide bonds) is defined as

$$CSP_x = (\delta_1, \delta_2, \ldots, \delta_{2n-1})$$

where $\delta_i$ denotes the sequence position of the $i$th oxidized cysteine in the protein and $\delta_i$ 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 (East Zhao et al., 2005) as follows:

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

It shows that the CSP is an important global feature for the DCP problem. Thus, we also combine the CSP method to our hybrid method. Our hybrid method for predicting the disulfide connectivity pattern is described as follows.

**Algorithm: Hybrid method for DCP.**

**Input:** A protein sequence and the bonding states of all cysteines in it.

**Output:** The predicted disulfide connectivity pattern.

**Case 1:** For a 2-bond or 3-bond protein.

- **Step 1.1:** If the query meets the threshold in the pattern-wise BKS method for fusing the results of $CP_{1,F_{521}}$ and $CP_{1,F_{623}}$, report this pattern as the predicted pattern.
- **Step 1.2:** If the minimum divergence obtained by the CSP search is less than or equal to the threshold, report this pattern as the predicted pattern.
- **Step 1.3:** For the remaining, take the original maximum weighted pattern from the $CP_{1,F_{521}}$ method as the predicted result.

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

- **Step 2.1:** If the minimum divergence obtained by the CSP search is less than or equal to the threshold, report this pattern as the predicted pattern.
- **Step 2.2:** Apply the pair-wise BKS method to fusing the results of $CP_{1,F_{521}}$ and $CP_{2,Label_2}$. And then report the answer.

### 4 EXPERIMENTAL RESULTS

In this section, we present the dataset used in our experiments and performance evaluation criteria of the DCP problem. We also show the experimental results.

#### 4.1 Dataset and Performance Evaluation

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. Table 8 illustrates the summary of SP39 dataset. This dataset was first used by Vullo and Frasconi (Alessandro Vullo and Paolo Frasconi, 2004), and it contains 446 proteins with 2 to 5 disulfide bonds, derived from the SWISS-PROT release no. 39. We also use the same way as Wang et al.’s (Chong-Jie Wang et al., 2012) to divide SP39 dataset into 4 disjoint subsets.
Table 5: An example of the partial pattern-wise BKS table for 2-bond proteins.

<table>
<thead>
<tr>
<th>CP1F521 Interval</th>
<th>CP1F623 Interval</th>
<th>1-1-2-2</th>
<th>1-2-1-2</th>
<th>1-2-2-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.15, 0.2)</td>
<td>(0.15, 0.2)</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(0.15, 0.2)</td>
<td>(0.2, 0.25)</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0.15, 0.2)</td>
<td>(0.25, 0.3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0.15, 0.2)</td>
<td>(0.3, 0.35)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0.15, 0.2)</td>
<td>(0.35, 0.4)</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Real patterns</th>
<th>CP1F521</th>
<th>CP1F623</th>
<th>Predicted by BKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CXOA_CONMA</td>
<td>1-2-3-1-2-3</td>
<td>1-2-1-3-2-3</td>
<td>1-2-3-1-2-3</td>
<td></td>
</tr>
<tr>
<td>HST1_ECOLI</td>
<td>1-2-3-1-2-3</td>
<td>1-2-1-3-2-3</td>
<td>1-2-3-1-2-3</td>
<td></td>
</tr>
<tr>
<td>HCYA_PANIN</td>
<td>1-1-2-2-3-3</td>
<td>1-1-2-3-3-2</td>
<td>1-1-2-3-3-2</td>
<td></td>
</tr>
<tr>
<td>CXOB_CONST</td>
<td>1-2-3-1-2-3</td>
<td>1-2-1-3-2-3</td>
<td>1-2-3-1-2-3</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Pairs from CP1F521</th>
<th>Pairs from CP2Label2</th>
<th>Truly bonded</th>
<th>Not bonded</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.3, 0.4)</td>
<td>(0.1, 0.1)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0.3, 0.4)</td>
<td>(0.1, 0.2)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0.3, 0.4)</td>
<td>(0.2, 0.3)</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>(0.3, 0.4)</td>
<td>(0.3, 0.4)</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>(0.3, 0.4)</td>
<td>(0.4, 0.5)</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>(0.3, 0.4)</td>
<td>(0.5, 1)</td>
<td>1</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 8: The summary of SP39 dataset, where $B$ denotes the number of disulfide bonds.

<table>
<thead>
<tr>
<th>Number of proteins</th>
<th>Number of cysteines</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B = 2$</td>
<td>156</td>
</tr>
<tr>
<td>$B = 3$</td>
<td>146</td>
</tr>
<tr>
<td>$B = 4$</td>
<td>99</td>
</tr>
<tr>
<td>$B = 5$</td>
<td>45</td>
</tr>
<tr>
<td>$B = 6$</td>
<td>446</td>
</tr>
<tr>
<td>Oxidized</td>
<td>2742</td>
</tr>
<tr>
<td>Total</td>
<td>4401</td>
</tr>
</tbody>
</table>

Table 9 shows the $Q_p$ of our methods compared with previous works in SP39 dataset. The accuracies of the three SVM models are derived from the patterns with the maximum weighted graph matching (Piero Fariselli and Rita Casadio, 2001). However, we find that it is hard to improve the accuracy by only one single SVM model. BKS can play a supporting role in our method. Although the performance of CP2Label2 is not better than CP1F521 or CP1F623, CP2Label2 provides the effect for pair-wise BKS since CP2Label2 represents another concept of pair extraction. As one can see in the table, with the help of BKS fusing methods, the accuracy is improved to 65.9%.

Furthermore, when the divergence of CSP is low, the prediction confidence is also high. Thus, we set the applicable thresholds of CSP to pick out the patterns as predicted results. Here, we set the threshold of CSP to 0, 5, 10, and 15 for proteins with 2 to 5 bonds, respectively. Eventually, the prediction accuracy of our hybrid method with SVM, BKS and CSP reaches 69.1%, a great improvement compared with the previous results.

4.2 Results

In the CP1F521 method, combined by SVM and the maximum weighted graph matching (Piero Fariselli and Rita Casadio, 2001), we find that the prediction accuracy is very high when the probability of the predicted pattern is greater than or equal to 0.5 (half). Thus, before performing our method, the answer is settled down for these predictions.

Table 9 shows the $Q_p$ of our methods compared with previous works in SP39 dataset. The accuracies of the three SVM models are derived from the patterns with the maximum weighted graph matching (Piero Fariselli and Rita Casadio, 2001). However, we find that it is hard to improve the accuracy by only one single SVM model. BKS can play a supporting role in our method. Although the performance of CP2Label2 is not better than CP1F521 or CP1F623, CP2Label2 provides the effect for pair-wise BKS since CP2Label2 represents another concept of pair extraction. As one can see in the table, with the help of BKS fusing methods, the accuracy is improved to 65.9%.

Furthermore, when the divergence of CSP is low, the prediction confidence is also high. Thus, we set the applicable thresholds of CSP to pick out the patterns as predicted results. Here, we set the threshold of CSP to 0, 5, 10, and 15 for proteins with 2 to 5 bonds, respectively. Eventually, the prediction accuracy of our hybrid method with SVM, BKS and CSP reaches 69.1%, a great improvement compared with the previous results.
5 CONCLUSIONS

According to the study of Wang et al. (Chong-Jie Wang et al., 2012), which focuses SVM models on varied features, and the concept of different cysteine-pair representations proposed by Lu et al. (Chih-Hao Lu et al., 2007), we do many integrated experiments, whose results are not all shown in this paper. However, the improvement of the pure SVM method is not so significant although the SVM method is still relatively good among the various methods. Some studies (Bo-Juen Chen et al., 2006; Yu-Ching Chen, 2007) 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 deeply depends on the pattern varieties in the template set.

In this paper, we first gather some statistics about the disulfide bonds, and have successfully found some rules to discriminate the patterns with high accuracy in various methods. Then, we adopt the pattern-wise and pair-wise BKS methods to fuse multiple SVM models. In addition, the CSP search method is also invoked in our method. As the experimental results show, we think that the hybrid method is one of the good ways to increase the prediction accuracy in the DCP problem.

In the future, we may apply 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 on the inter-bond template dataset to explore more possibilities of sub-pattern development.

ACKNOWLEDGEMENTS

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

**REFERENCES**


