Prediction of the disulphide bonding state of cysteines in proteins using Conditional Random Fields

Watshara Shoombuatong
Faculty of Science,
Bioinformatics Research Laboratory,
Chiang Mai University, 50200, Thailand
E-mail: Satruncmu@hotmail.com

Patrinee Traisathit and Sukon Prasitwattanaseree
Faculty of Science,
Department of Statistics
and Bioinformatics Research Laboratory,
Chiang Mai University, 50200, Thailand
E-mail: patrinee@chiangmai.ac.th
E-mail: sprasitwattanaseree@gmail.com

Chatchai Tayapiwatana
Faculty of Associated Medical Sciences,
Division of Clinical Immunology,
Department of Medical Technology,
Chiang Mai University, 50200, Thailand

and

Biomedical Technology Research Unit,
National Science and Technology Development Agency,
Chiang Mai University, 50200, Thailand
E-mail: asimi002@chiangmai.ac.th

Robert Cutler
15 Prasing Post, Ampur Muang,
Chiang Mai 50200, Thailand
E-mail: cutler121@gmail.com
Jeerayut Chaijaruwanich*

Faculty of Science,
Department of Computer Science
and Bioinformatics Research Laboratory,
Chiang Mai University, 50200, Thailand
E-mail: jeerayut@science.cmu.ac.th
*Corresponding author

Abstract: The formation of disulphide bonds between cysteines plays a major role in protein folding, structure, function and evolution. Many computational approaches have been used to predict the disulphide bonding state of cysteines. In our work, we developed a novel method based on Conditional Random Fields (CRFs) to predict the disulphide bonding state from protein primary sequence, predicted secondary structures and predicted relative solvent accessibilities (all-state information). Our experiments obtain 84% accuracy, 88% precision and 94% recall, using all-state information. However, our results show essentially identical results when using protein sequence and predicted relative solvent accessibilities in the absence of secondary structure.

Keywords: cysteines-bonding state; CRFs; conditional random fields; bioinformatics; machine learning.


Biographical notes: Watshara Shoombuatong graduated with a Bachelor’s Degree in Mathematics from the Faculty of Science at Chiang Mai University in 2006. Now, he is doing Master’s Degree in Bioinformatics Faculty of Science at Chiang Mai University and he is also a member of Bioinformatics Lab, Faculty of Science at Chiang Mai University. His interests are data mining, graphical model, structured model and machine learning.

Patrinee Traisathit received her PhD in Biomathematics from DENIS DIDEROT – Paris VII UNIVERSITY (France). Her current researches focus on growth of HIV-infected children exposed to antiretroviral treatment, tuberculosis and risk of death following initiation of antiretroviral therapy in children, Risk factor for isoniazid-resistant among pulmonary tuberculosis patients.

Sukon Prasitwattanaseree received his PhD in Biomathematics from DENIS DIDEROT – Paris VII UNIVERSITY (France). His current researches focus on applied statistics for bioinformatics researches.

Chatchai Tayapiwatana received his PhD in Biotechnology from Chiang Mai University (Thailand). His current researches focus on Phage display of leukocyte surface molecule for ligand tracing, Improvement of phage display technique for immune cell surface molecule studies, intrabodies for therapeutic means of viral infections and cancers and development of monoclonal antibodies against hapten.
Prediction of the disulphide bonding

Robert Cutler received his PhD in Computational BioPhysics from Vanderbilt University (USA). His current research focuses on understanding the workings of various methods of correcting mutations in genomes, characterisation of phenotypic traits in tropical plants, and development of computational models to better understand genomic function.

Jeerayut Chaijaruwanich received his PhD in Informatique from Universite d’Evry (France). His current researches focus on microarray and protein sequence analysis using machine-learning techniques.

1 Introduction

Disulphide bonds are often found in extracellular and secretary proteins, and involve in bringing the two cysteines into proximity. This process has to be facilitated by Protein Disulphide Isomerases (PDIs) (Bardwell, 2004). PDI consists of four major domains, which is a common characteristic of the reactive sites of PDI domains that assisted in forming disulphide bonds: a universal Cys X-X-Cys pattern in the reactive site (McArthur et al., 2001).

In eukaryotes, such disulphide bonds are key components of growth factors and extracellular domains for various receptors (Horton et al., 1996). The covalent nature of disulphide bonds make such bonding a strong constraint on protein conformations by bringing together sequentially distant regions in proteins. Disulphide bonds also stabilise proteins in a variety of ways. Proteins that contain disulphide bonds usually possess higher thermal and chemical stability (McBride et al., 1992). Disulphide bonds also play important roles in protein-folding pathways by guiding the folding pathway through disulphide-bond-containing intermediates (Vielle and Zeikus, 2001). The functions of various extracellular proteins can be controlled through reduction and formation of one or more disulphide bonds (Hogg, 2003). Many functions of disulphide bonds in proteins are overlapping. Disulphide bond patterns can help provide understanding of structural properties of proteins and possibly identify which family the proteins may belong to, giving important insights into possible biological functions (Chuang et al., 2003). For example, the pheromone-binding proteins straddle water boundaries enabling transport of pheromones past the surface barrier. This transport can only be accomplished owing to a conserved set of cysteine bonds holding hydrophobic and hydrophilic residues in proximity (Munshaw et al., 2004).

The problem of disulphide bond prediction can be subdivided into four related sub-problems, i.e., chain classification, cysteine-bonding state prediction, bonding probability estimation and disulphide connectivity prediction (Tsai et al., 2007; Du, 2007). This paper focused on the cysteine-bonding state prediction in proteins, i.e., each cysteine is divided into two classes: the bonded state or non-bonded state. Owing to the importance of this issue, a wide range of theoretical investigations have emerged in recent years. Muskal et al. (1990) used a sliding window of flanking sequences (the neighbouring amino acids around cysteines). This approach achieved 81% accuracy. Mucchielli-Giorgi et al. (2002) used logistic functions to classify the state.
of each cysteine based on the amino content of their whole proteins and achieved close to 84% accuracy. Martelli et al. (2002a, 2002b) implemented a hybrid system (Hidden Neural Network) that combines Hidden Markov Models (HMMs) and Neural Networks (NNs) based on local and global characteristics of proteins and yielded 88% accuracy. Frasconi et al. proposed a Support Vector Machine (SVM)-based predictor that operated in two states (multi-classifier at protein level and binary classifier at cysteine level). This approach achieved an accuracy of 83.6% (Frasconi et al., 2002). Chen et al. used an SVM based on flanking sequences, global protein information and the cysteine oxidation state pattern. They achieved the highest previously published accuracy of 90%, 77% precision and 91% recall (Chen et al., 2004).

In this paper, we present a novel approach based on CRFs, which are successfully applied to solve the sequence-labelling problem. CRFs have advantages, which integrate both state feature and transition feature between label states. We designed a series of features, which allowed us to determine relative contributions from primary sequence data vs. the additional descriptors (predicted secondary structures and predicted relative solvent accessibilities). Finally, we compare the performance of CRFs with Kernel Support Vector Machines (K SVMs) and Neural Networks (NNs). Our experimental results show that the CRFs model is comparable with other conventional classification methods, which tend to use a smaller window size. Moreover, it can be utilised for any data set, although this data set suffers from unbalance problem.

2 Materials and methods

2.1 Data set

The disulphide bond data set (Frasconi et al., 2002; Chen et al., 2004) (available at http://download.igb.uci.edu/disulfide.txt) used in this work was derived from the Protein Data Bank (PDB) including solvent accessibilities and secondary structures generated by the DSSP programme (Kabsch and Sander, 1983). It can be downloaded freely from http://swift.cmbi.kun.nl/gv/dssp/. Each protein includes the sequence name (pdb code + chain id, line 1), sequence length, the number of bonded cysteines, total number of cysteines (line 2), sequence (line 3), secondary structure (line 4), relative solvent accessibility (line 5: e: exposed, –: buried, determined at a 25% threshold) and disulphide bond information (rest of lines, each line corresponding to one disulphide bond identified by the positions of the cysteine pair). This data set was obtained by further filtering based on the Homology-derived Secondary Structure of Proteins (HSSP) distance (Sander and Schneider, 1991). This data set was randomly split into 10 subsets of roughly the same size. During each 10-fold cross-validation experiment, 9 subsets were used for training and the remaining subset was used for validation. Final results are averaged across the 10 cross-validation experiments. More detail about this data set can be referred from Baldi et al. (Jianlin et al., 2006; Pierre et al., 2004).

Table 1 shows cysteine predicted secondary structure frequencies (%), which consist of both half-cysteine (bonding state of cysteine) and free-cysteine (non-bonding state of cysteine) in the selected data set, which often occurs in coil structure regions (49.1%).
Table 1  Cysteine secondary structure frequencies (%)

<table>
<thead>
<tr>
<th>Secondary structure</th>
<th>All residue</th>
<th>Non-bond</th>
<th>Bond</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheet</td>
<td>31.7</td>
<td>31.2</td>
<td>31.8</td>
</tr>
<tr>
<td>Coil</td>
<td>47.6</td>
<td>39.2</td>
<td>49.1</td>
</tr>
<tr>
<td>Helix</td>
<td>20.7</td>
<td>29.6</td>
<td>19.1</td>
</tr>
</tbody>
</table>

Frequencies of secondary structure, computed on the whole dataset.

2.2 Conditional Random Field used for labelling sequence data

To predict disulphide bonding state of cysteines, we denote this problem as a sequence labelling task. Primary sequence data and the additional descriptors were treated as sequence data. Cysteine bonding stated based on protein primary sequence were labelled as in a bonded or non-bonded state using CRFs. CRFs have advantages which integrate both state feature and transition feature between label states. Both CRFs and HMMs suit to label sequence, differing from the probability solution formulation. HMMs obtain the target label sequence \( y \) by maximising the joint probability from \( x \) and \( y \) (Lafferty et al., 2001), but HMMs cannot use long distance features.

A CRF is a form of undirected graphical model (Wallach, 2002), which may be viewed as Markov random field (Jordan, 2004). Formally, \( G = (V, E) \) is defined as an undirected graph such that there is a node \( v \in V \) corresponding to each of the random variables representing an element \( Y_i \) of \( Y \), including each random variable \( Y_i \) obeys the Markov property with respect to \( G \).

Conditional Random Fields (CRFs) are described through the concept of graphical models. Lafferty et al. define CRFs as the probability of a particular label sequence \( y = (y_1, y_2, \ldots, y_n) \) given observation sequence \( x = (x_1, x_2, \ldots, x_n) \) to be a normalised product of potential functions (Lafferty et al., 2001). This is represented in form

\[
\text{exp} \left[ \sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_j \mu_j s_j(y_i, x, i) \right]
\]

where \( t_j(y_{i-1}, y_i, x, i) \) is a transition feature function of the space of all possible observation sequence and the labels at positions \( i \) and \( i - 1 \) in the label sequence; \( s_j(y_i, x, i) \) is a state feature function of label sequence at position \( i \) and the observation sequence. The index \( j \) in \( t_j \) and \( s_j \) is feature serial number to present different feature. Parameters \( \lambda_j \) and \( \mu_j \) correspond with feature \( t_j \) and \( s_j \), respectively, and they are learned of maximising the conditional likelihood of the training data.

When we define each feature function \( E = (x, i) \), which is constructed as a set of real-valued features of the observation sequence to express some types of amino acid of the training data. An example of such a feature is

\[
E(x, i) = \begin{cases} 
1, & \text{if the observation at position } i \text{ is amino acid type A} \\
0, & \text{otherwise}
\end{cases}
\]

Each feature function takes on the value of one of these real-valued observation features \( E = (x, i) \), if the current state takes on particular value. All feature functions are therefore real-valued. An example of such a state feature function is
For state feature function, it can be simplified by writing
\[ s_j(y_i, x, i) = s_j(y_{i-1}, y_i, x, i) \] (4)

and
\[ F_j(y, x) = \sum_{j=1}^{s} f_j(y_{i-1}, y_i, x, i) \] (5)

where each \( f_j(y_{i-1}, y_i, x, i) \) is either a state function \( s_j(y_{i-1}, y_i, x, i) \) or a transition function \( t_j(y_{i-1}, y_i, x, i) \). This allows the probability of a label sequence \( y \) given an observation sequence \( x \) to be written as
\[ p(y | x, \lambda) = \frac{1}{Z(x)} \exp \left( \sum_j \lambda_j F_j(y, x) \right) \] (6)

where \( Z(x) \) is the normalizing factor. More detail about CRFs can be referred from Lafferty et al.

### 2.3 Predicting disulphide bonding state of cysteines based on CRFs

Here, the bonding states of cysteines are labelled by CRFs. The label set for states is \( S = \{1, 0\} \), where 1 represents the bonded state and 0 represents the non-bonded state, and the space of all possible sequence of observations are \( X = (x_1, x_2, x_3) \), which are represented as observation information composed of protein primary sequence \( x_1 \), predicted secondary structures \( x_2 \), and predicted relative solvent accessibilities \( x_3 \).

Given a sequence of observation \( x_i = (y_{i1}, y_{i2}, \ldots, y_{im}) \) \((i = 1, 2)\), the most probable label sequence \( y = (y_1, y_2, \ldots, y_n) \) \((y_i \in S)\) is obtained by CRFs. For each \( y \) in the label sequence, we define the conditional probability distribution \( p(y | X) \) for the linear-chain CRFs illustrated in Figure 1 as:

**Figure 1** The system of disulphide bonding state prediction

\[
p(y|x) = \frac{1}{Z(x)} \exp \left( \sum_j \lambda_j F_j(y, x) \right)
\]
Given a sequence of observations $X$, we would like to get the most probable label sequence $y$ from $\hat{Y} = \arg\max_y P(y | X)$ (Durbin et al., 1998).

### 2.4 Definition of features

The features of CRFs include state and transition features. We define several types of state feature based on the notion of binding sites of PDIs (Bardwell, 2004; McArthur et al., 2001). Ruddock et al. (2000) and Pirneskoski et al. (2004) found that binding sites of PDIs would have a special primary substrate-binding region, and its region provide some significant information, which can help to identify state of cysteines. Hence, flanking sequences, which refer to the neighbouring amino acids around cysteines, are the significant information to identify bonding state of cysteines, as well as predicted secondary structure and predicted solvent accessibility play important roles (Ferre and Clote, 2005). Therefore, we would like to design a series of features, which allowed us to determine relative contributions from primary sequence data vs. the additional descriptors (predicted secondary structures and predicted relative solvent accessibilities) for testing CRF methodology. These features also provided insight into the degree to which amino acid sites positioned at increasing lengths from the cysteine amino acid contribute to the binding state.

#### 2.4.1 Pair site neighbours feature

The first CRF feature (Figure 2) generated conditional probabilities based on the correlation of adjacent amino acid sites as follows:

$$s_i(y_i, X, i) = \begin{cases} 1 & \text{if } y_i = 1 \text{ and } X = (x_i, x_{i+1}) \text{ are particular types of amino acids} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

**Figure 2** Graphical representations of pair site neighbours feature when using 2 windows ($N = 2$)

#### 2.4.2 All pair sites feature

Second CRF feature (Figure 3) included all possible pairs of correlations between pairs of amino acids on length scales as follows

$$s_i(y_i, X, i) = \begin{cases} 1 & \text{if } y_i = 1 \text{ and } X = (x_{i-1}, x_i, x_{i+1}, \ldots) \text{ are particular types of amino acids} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$
2.4.3 Linearly independent full feature

Third CRF feature (Figure 4) included that a minimal set of linearly independent correlated sites was developed that contain all of the paired amino acids by cut-off amino acid correlations, which are not linearly dependent as follows:

\[ s_i(y_i, X, t) = \begin{cases} 
1 & \text{if } y_i = 1 \text{ and linearly independent correlated of } x \text{ which are particular types of amino acids} \\
0 & \text{otherwise} 
\end{cases} \quad (9) \]

2.5 Implementation of Conditional Random Fields and performance

We used the CRF’s software, named CRF++ (Kudo, 1999), which is a simple, customisable, and open-source implementation of CRFs for segmenting/labelling sequential data implemented by Taku Kudo. CRF++ was designed as a generic resource, which can be applied to a variety of natural language processing tasks. It can be downloaded freely from http://crfpp.sourceforge.net/. The values of parameters were set by default. Moreover, we used the srl-eval.pl (Carreras and Marquez, 2005), which is an evaluation programme for the CoNLL-2005 Shared Task, which is open-source implementation implemented by Xavier Carreras and Lluis Marquez. This programme outputs performance measures based on accuracy, precision and recall. It can be downloaded freely from http://www.lsi.upc.es/~srlconll/srl-eval.pl
3 Results and discussion

3.1 Results and discussion of CRFs

In Table 2, we compare the experimental results of primary sequence of different feature and window sizes. Each of the three major rows is a different feature. The first feature (N) is a simple feature which generated conditional probabilities based on the correlation of adjacent amino acid sites distant from the cysteine amino acid in question is denoted as Section 2.4.1. To account for possible additional biochemical constraints between non-adjacent amino acids, our second feature (A) included all possible correlations between pairs of amino acids surrounding the cysteine in question as is shown in Section 2.4.2. The third and last feature (L) is composed of minimal sets of linearly independent correlated sites that contain all of the paired amino acid correlations in (A) denoted as Section 2.4.3. Each row is categorised for different window sizes between two and four amino acids distant. For example, the S2 feature uses the correlation of pair site neighbours and includes the two upstream and two downstream amino acids from the cysteine in question. Each of three major columns report accuracy, precision, recall, and $Q_p$. All methods are measured according to the evaluation of residue labelling (or classification) based on the following quantities:

- TP is the number of true positives, which cysteine residues correctly classified as bonding state
- TN is the number of true negatives, which cysteine residues correctly classified as non-bonding state
- FP is the number of false positives, which non-bonding state of cysteine residues are incorrectly classified as bonding state of cysteine residues
- FN is the number of false negatives, which bonding state of cysteine residues are incorrectly classified as non-bonding state of cysteine residues
- $P$ is the total number of protein sequences whose all disulphide bonds are correctly predicted
- $N$ is the total number of protein sequences tested.

Then we used the following measures to evaluate the labelling (and classification) performance:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (11)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (12)
\]

\[
Q_p = \frac{P}{N}. \quad (13)
\]
Accuracy, precision, and recall are all used to measure the performance of labelling or classification of interface residues.

In Table 2, the first major row reports that groups of amino acid sites are considered individually for different window sizes. We use three window sizes (2–4), with a result of about 80% accuracy when using primary sequence. Since the second feature generates conditional probabilities based on all pairs of amino acids within the specified region, the improvement in prediction result reaches 83% accuracy, which is higher than first feature when using the four window sizes shown in Table 2. With respect to these first two features, these results show that there are additional biochemical constraints on at least some combination of amino acids, which are not immediately adjacent (A). For the previous selected feature, it takes into account amino acid correlations, which are not linearly dependent to maximise the overall predictive ability of the method. The prediction results improve using the last feature (L), and reach a maximum 84.4% accuracy when using only primary sequence (as is represented in Section 2.4.3). Because this feature (L) eliminates redundant correlations, the overall size of the feature is reduced and the prediction accuracy, recall and precision are improved when using a four-length window, but not for window lengths larger than 4. On the other hand, when we focus on the correctly predicted all-bonding state for the same protein, these results demonstrate that it decreases to 59.0%.

Table 2. The experimental results of CRFs based on primary sequence of different window size and features

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>Qp (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N2</td>
<td>82.3</td>
<td>85.3</td>
<td>95.7</td>
<td>57.5</td>
</tr>
<tr>
<td>N3</td>
<td>80.9</td>
<td>87.0</td>
<td>91.0</td>
<td>50.9</td>
</tr>
<tr>
<td>N4</td>
<td>79.7</td>
<td>87.3</td>
<td>89.1</td>
<td>47.2</td>
</tr>
<tr>
<td>A2</td>
<td>79.5</td>
<td>87.2</td>
<td>88.9</td>
<td>46.7</td>
</tr>
<tr>
<td>A3</td>
<td>81.6</td>
<td>87.5</td>
<td>91.3</td>
<td>52.7</td>
</tr>
<tr>
<td>A4</td>
<td>82.9</td>
<td>87.3</td>
<td>93.5</td>
<td>57.1</td>
</tr>
<tr>
<td>L2</td>
<td>82.6</td>
<td>86.7</td>
<td>94.0</td>
<td>57.0</td>
</tr>
<tr>
<td>L3</td>
<td>83.6</td>
<td>87.0</td>
<td>94.7</td>
<td>58.2</td>
</tr>
<tr>
<td>L4</td>
<td>84.1</td>
<td>87.0</td>
<td>95.6</td>
<td>59.0</td>
</tr>
</tbody>
</table>

Table 3, we compare experimental results of CRFs based on all state information of different window sizes and features. The result of the simplest feature of CRFs is 79% accuracy, 87% precision, and 87% recall, which remains about the same as using primary sequence. However, when using the second feature (A), the increase in precision mainly comes from improvement of recall. This result shows that relative correlation of cysteine combine with predicted secondary structure and predicted relative solvent accessibilities, which are concerned with the bonding or non-bonding state of the cysteines. When using the linearly independent third feature (L), the result performance increases dramatically reaching 90% precision and 94% recall.
Table 3  The experimental results of CRFs based on all state information (primary sequence, predicted secondary structure, and predicted solvent accessibility) of different window size and features

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>Qp (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N2</td>
<td>81.1</td>
<td>86.5</td>
<td>92.2</td>
<td>53.3</td>
</tr>
<tr>
<td>N3</td>
<td>79.0</td>
<td>87.4</td>
<td>88.0</td>
<td>46.1</td>
</tr>
<tr>
<td>N4</td>
<td>79.0</td>
<td>87.7</td>
<td>87.5</td>
<td>45.8</td>
</tr>
<tr>
<td>A2</td>
<td>77.8</td>
<td>87.3</td>
<td>86.4</td>
<td>42.0</td>
</tr>
<tr>
<td>A3</td>
<td>81.9</td>
<td>87.8</td>
<td>91.4</td>
<td>52.4</td>
</tr>
<tr>
<td>A4</td>
<td>82.9</td>
<td>87.8</td>
<td>92.4</td>
<td>55.4</td>
</tr>
<tr>
<td>L2</td>
<td>81.9</td>
<td>87.0</td>
<td>92.5</td>
<td>53.9</td>
</tr>
<tr>
<td>L3</td>
<td>82.4</td>
<td>87.3</td>
<td>92.8</td>
<td>55.9</td>
</tr>
<tr>
<td>L4</td>
<td>84.4</td>
<td>89.6</td>
<td>94.2</td>
<td>60.1</td>
</tr>
</tbody>
</table>

The main factors that influence the results as presented in our experiment rely on differences in protein sequence information, different window sizes, and different combinations of site correlations examined in each feature. By varying these parameters, we find the feature that can best predict the disulfide bonding state of cysteines in addition to shedding some light on the possible areas of biological importance leading to these results. The first feature (N4) obtains a 79.7% and 79.0% accuracy when we use the primary sequence and all information, respectively. The second feature obtains accuracy close to 82.9% for primary sequence and all information, trained with a 4-length window. This is a demonstration of the ability of additional state information to build a richer feature for local dependencies of the bonding state. The prediction ability of this experiment is higher than for all of the previous selected features. Finally, we employ feature Section 2.4.3 for prediction. This choice proves to be successful, attaining the best performance, with an accuracy, precision and recall as high as 84.4% accuracy, 89.6% precision and 94.2% recall (Table 3) using a 4-length window and all-state information as the descriptors (but not for window length more than 3). These results demonstrate the capability of this feature to accurately predict the cysteine-bonding status of residues at a level previously unattainable. However, when we focus on the correctly predicted all-bonding state in the same protein, this result decreases to 60.1%.

Table 5 reports the comparison of false positive and false negative (as shown in each row) using a 4-window size as representative feature with different descriptors.

Because our selected data set has a higher number of half-cysteine than number of free-cysteine as represented in Table 1, it leads to a higher prior probability of half-cysteine than free-cysteine. In the same way, we can see that each descriptor shows that the false positives are higher than false negatives as shown in Table 4. This means that the feature prediction of the bonding state of the cysteine has a tendency to predict a bonding state more than a non-bonding one (or overprediction). Therefore, we would like to reduce the overfitting problem from our selected data set by choosing the information or descriptor, which predominantly corresponded to the prediction of the
disulphide bonding state, which guarantees that it can be applied to other data sets as shown in Table 5 (Seq = primary sequence, SS = Predicted Secondary Structure and RS = Predicted Relative Solvent Accessibility).

Table 4  Comparison number of false positive and false negative (%) with different descriptors when using 4 windows sizes

<table>
<thead>
<tr>
<th></th>
<th>Seq</th>
<th>Seq + SS</th>
<th>Seq + RS</th>
<th>All state information</th>
</tr>
</thead>
<tbody>
<tr>
<td>False positive</td>
<td>11.6</td>
<td>12.6</td>
<td>11.6</td>
<td>10.7</td>
</tr>
<tr>
<td>False negative</td>
<td>3.9</td>
<td>4.1</td>
<td>4.1</td>
<td>4.6</td>
</tr>
</tbody>
</table>

Table 5  Comparison of measures from different descriptors when using a 4 length window and the full linearly independent feature (L)

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>Qp (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq</td>
<td>84.1</td>
<td>87.0</td>
<td>95.6</td>
<td>59.0</td>
</tr>
<tr>
<td>Seq + SS</td>
<td>83.3</td>
<td>86.5</td>
<td>95.1</td>
<td>59.8</td>
</tr>
<tr>
<td>Seq + RS</td>
<td>84.3</td>
<td>87.5</td>
<td>95.2</td>
<td>61.3</td>
</tr>
<tr>
<td>All state information</td>
<td>83.9</td>
<td>87.8</td>
<td>94.2</td>
<td>60.1</td>
</tr>
</tbody>
</table>

In Table 5, we compare different descriptors where all of descriptors are trained using only a four-window size, for the third feature (L). The best performance is 84% accuracy, 89% precision and 95% recall, but these results occur in different descriptor combinations. However, when using all-state information as descriptor, the experimental result achieved close to 87% precision and 94% recall, whereas the result of experimental of the bonding state prediction is close to 88% precision and 95% recall, which is close to our best performance, when using protein sequence and predicted relative solvent accessibilities in the absence of secondary structure. Moreover, this selected information has 84% accuracy, which is better than using all-state information. This result shows that when using only protein sequence and predicted relative solvent accessibilities, its descriptor is adequate to classify the bonding state of cysteine.

3.2 Performance of CRFs vs. other classification methods

Kernel Support Vector Machines (K SVMs) and Neural Networks (NNs) were selected to compare against our method, because these are well-known models to approach the classification task. The Neural Network Package in RGui was used as the implementation. This package is nnet (Ripley et al., 2009), which can be downloaded freely from http://cran.r-project.org/web/packages/e1071/index.html. The values of size (number of units in the hidden layer), rang (Initial random weights), decay (parameter for weight decay) and maxit (maximum number of iterations) were set as 2, 0.1, 5e-4 and 6000, respectively. The package of KSVMs in RGui is kernlab (Karatzoglou et al., 2009), which can be downloaded freely from http://cran.r-project.org/web/packages/kernlab/index.html. The KSVM contains C-svc as a valid option, Radial Basis kernel and cost of constraints violation as 20. To compare with other methods for prediction of disulphide bonding state, we used the same criteria to analyse our selected data.
Table 6 compares results based on different descriptors and methods. When using the primary sequence as a descriptor, we obtain accuracy, precision and recall of CRFs that is better than the other methods. In this case, other methods have a problem of window length of training dataset. Furthermore, when using primary sequence and predicted relative solvent accessibilities, we obtain 84% accuracy and 87% precision, which is better than NN and KSVMs. Other descriptors, both precision and recall of KSVMs, are better than our features. However, the accuracy and precision of CRFs is similar to KSVMs.

Our study shows that the disulphide bonding state of cysteine is determined by protein primary sequence, predicted secondary structures and predicted relative solvent accessibilities. Moreover, the idea of the correlation of amino acid sites distant from the cysteine amino acid is also exploited to improve the result performance (Li et al., 2006). This idea may be useful in glycosylation site prediction (Pirneskoski et al., 2004). Since, the character of each type of glycosylation can occur in a recognisable motif. Further research into glycosylation could provide information into protein function, protein folding, metabolism, transport, etc. (Caragea et al., 2007).

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Methods</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary sequence</td>
<td>NN</td>
<td>68.5</td>
<td>86.4</td>
<td>73.5</td>
</tr>
<tr>
<td></td>
<td>KSVMs</td>
<td>82.1</td>
<td>85.4</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>CRFs</td>
<td>84.1</td>
<td>87.0</td>
<td>95.6</td>
</tr>
<tr>
<td>Primary sequence +</td>
<td>NN</td>
<td>78.4</td>
<td>88.9</td>
<td>85.4</td>
</tr>
<tr>
<td>Predicted secondary structure</td>
<td>KSVMs</td>
<td>84.2</td>
<td>86.9</td>
<td>94.1</td>
</tr>
<tr>
<td></td>
<td>CRFs</td>
<td>83.3</td>
<td>86.5</td>
<td>95.1</td>
</tr>
<tr>
<td>Primary sequence + predicted</td>
<td>NN</td>
<td>78.8</td>
<td>87.2</td>
<td>87.5</td>
</tr>
<tr>
<td>Relative solvent accessibilities</td>
<td>KSVMs</td>
<td>83.3</td>
<td>85.8</td>
<td>96.6</td>
</tr>
<tr>
<td></td>
<td>CRFs</td>
<td>84.3</td>
<td>87.5</td>
<td>95.2</td>
</tr>
<tr>
<td>All state information</td>
<td>NN</td>
<td>79.3</td>
<td>88.2</td>
<td>87.4</td>
</tr>
<tr>
<td></td>
<td>KSVMs</td>
<td>79.1</td>
<td>87.6</td>
<td>95.4</td>
</tr>
<tr>
<td></td>
<td>CRFs</td>
<td>83.9</td>
<td>87.8</td>
<td>94.2</td>
</tr>
</tbody>
</table>

4 Conclusion

We have proposed a novel method based on CRFs to predict the disulphide bonding state from protein primary sequence, predicted secondary structures and predicted relative solvent accessibilities performance on the PDB. Table 5 summarises the performance comparison of the predicted disulphide bonding state of the cysteines of different descriptors when using only a four-window size and third feature (L) as selected feature. Tables 2 and 3 show that residue dependencies can improve prediction accuracy compared with adjacent amino acid based on correlated information. Furthermore, an increased window size over the \( N = 2 \) length features previously examined provides
additional valuable information and improves performance in all cases. After that, we would like to eliminate the overfitting problem from our selected data set by choosing the information, which predominantly corresponded to the prediction of the disulphide bonding state. Finally, we compare CRFs with KSVMs and NN. Table 6 shows that when using primary sequence as descriptor and just using 4-window sizes, it leads to the prediction performance of CRFs is close to 96% recall. This result is close to the best score of recall when combined primary sequence and predicted relative solvent accessibilities were used as the descriptors in the KSVMs method.

Acknowledgements

We thank Professor R. Hewett for helpful comments and suggestion on this manuscript. Moreover, this project was supported by The Graduate School Chiang Mai University, Biomedical Engineering Center, Faculty of Engineering, Chiang Mai University, the Center for Innovation in Chemistry: Postgraduate and Research Programme in Chemistry (PERCH-CIC), and the National Center for Genetic Engineering and Biotechnology, Thailand.

References


Prediction of the disulphide bonding


Ripley, B., Hornik, K. and Gebhardt, A. (2009) nnet (Feed-forward Neural Networks and Multinomial Log-Linear Models); software available at http://cran.r-project.org/web/packages/nnet/index.html


