BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 23 2007, pages 3147–3154
doi:10.1093/bioinformatics/btm505

Structural bioinformatics

Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure

Jiangning Song[1], Zheng Yuan[2], Hao Tan[3], Thomas Huber[4] and Kevin Burrage[1,2,*]

[1]Advanced Computational Modelling Centre, [2]ARC Centre in Bioinformatics and Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia, [3]Caulfield School of Information Technology, Monash University, Caulfield, East VIC 3145 and [4]School of Molecular & Microbial Sciences and Australian Institute for Bioengineering & Nanotechnology, The University of Queensland, Brisbane, QLD 4072, Australia

Received on June 18, 2007; revised on October 3, 2007; accepted on October 4, 2007

Advance Access publication October 17, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Disulfide bonds are primary covalent crosslinks between two cysteine residues in proteins that play critical roles in stabilizing the protein structures and are commonly found in extracy-toplasmatic or secreted proteins. In protein folding prediction, the localization of disulfide bonds can greatly reduce the search in conformational space. Therefore, there is a great need to develop computational methods capable of accurately predicting disulfide connectivity patterns in proteins that could have potentially important applications.

Results: We have developed a novel method to predict disulfide connectivity patterns from protein primary sequence, using a support vector regression (SVR) approach based on multiple sequence feature vectors and predicted secondary structure by the PSIPRED program. The results indicate that our method could achieve a prediction accuracy of 74.4% and 77.9%, respectively, when averaged on proteins with two to five disulfide bridges using 4-fold cross-validation, measured on the protein and cysteine pair on a well-defined non-homologous dataset. We assessed the effects of different sequence encoding schemes on the prediction performance of disulfide connectivity. It has been shown that the sequence encoding scheme based on multiple sequence feature vectors coupled with predicted secondary structure can significantly improve the prediction accuracy, thus enabling our method to outperform most of other currently available predictors. Our work provides a complementary approach to the current algorithms that should be useful in computationally assigning disulfide connectivity patterns and helps in the annotation of protein sequences generated by large-scale whole-genome projects.

Availability: The prediction web server and Supplementary Material are accessible at http://foo.maths.uq.edu.au/~huber/disulfide

Contact: kb@maths.uq.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

*To whom correspondence should be addressed.

1 INTRODUCTION

Disulfide bonds are primary covalent crosslinks formed between two cysteine residues in the same or different protein polypeptide chains. They play critical roles in stabilizing the protein structures and mediating protein biological functions (Inaba *et al.*, 2006; Kadokura *et al.*, 2004). They are commonly formed in extracytoplasmatic compartments in prokaryotes owing to the oxidizing extracellular environment (Kadokura *et al.*, 2003), while in eukaryotic cells disulfide bonds are formed in the lumen of the endoplasmic reticulum (ER) that provides a sufficiently oxidizing environment to allow for its formation (Sevier *et al.*, 2007). As a well-conserved and stereospecific secondary structural element of a protein, disulfide linkage can impose a substantial distance and angular constraint on the backbone of the protein, thus making an important contribution to the stabilization of protein tertiary structures (Chuang *et al.*, 2003). Disulfide bonds also play critical roles in the protein folding process and help assist proteins to fold into their correct tertiary structures. Statistical analyses and modeling simulations regarding disulfide connectivity have emerged in recent years (Cheek *et al.*, 2006; Thangudu *et al.*, 2005; Thornton, 1981), the majority of which have focused on the analysis of the distribution of disulfide bonds and their specific sequence environments (Abkevich and Shakhnovich, 2000; Gupta *et al.*, 2004; Harrison and Sternberg, 1994; Hartig *et al.*, 2005; van Vlijmen *et al.*, 2004).

Disulfide connectivity patterns give descriptions of the disulfide topology and how the cysteine residues are arranged sequentially to form disulfide bridges. Recent studies have indicated that disulfide connectivity patterns can be applied to efficiently discriminate the structural similarity of protein structures (Chuang *et al.*, 2003) and discover protein structural homologs (Gupta *et al.*, 2004; van Vlijmen *et al.*, 2004). Furthermore, in protein folding prediction, the localization of disulfide bonds can greatly reduce the search in conformational space and help towards the prediction of protein three-dimensional structure (Tsai *et al.*, 2005). Nevertheless, the sequence–structure gap is widening rapidly as a consequence of the large-scale whole-genome projects

(Bairoch and Apweiler, 2000; Berman *et al.*, 2000) and in this context computational methods that could reliably predict protein structure and function given its primary sequence only will continue to be valuable tools either from a computational or from a biological perspective (Thornton, 2001). Therefore, there is a great need to develop computationally efficient methods capable of accurately predicting disulfide connectivity of any protein given only its amino acid sequence. This would have potentially important applications, especially in providing insight into the structure and function of disulfide-rich proteins and in further understanding the role of the disulfide bridge in helping proteins reach their native conformations in the protein folding process.

The prediction of disulfide connectivity in protein has been investigated by a variety of computational methods with the prior knowledge of disulfide-bonding states of cysteines (Baldi *et al.*, 2005; Ceroni *et al.*, 2006; Chen and Hwang, 2005; Chen *et al.*, 2006; Cheng *et al.*, 2006; Fariselli and Casadio, 2001; Fariselli *et al.*, 2002; Ferre and Clote, 2005a,b; Tsai *et al.*, 2005; Vullo and Frasconi, 2004; Zhao *et al.*, 2005). In general, disulfide-predicting approaches fall into two major categories: pattern-wise methods that attempt to predict disulfide connectivity on the basis of the disulfide connectivity patterns and pair-wise methods that are associated with the cysteine residue pairing. Although these two different methods provide a prediction of disulfide connectivity, they both have intrinsic drawbacks. For instance, pair-wise methods rely on extracting pattern-based disulfide-forming sequential features and thus fail to abstract many of the other essential global features, while pattern-wise methods lack the consideration of the obviously important sequential information in the neighboring context of disulfide-bonded cysteine residues. In a recent prediction study, Chen *et al.* (2006) proposed a two-level model to combine both the pattern-wise and pair-wise based methods and achieved a prediction accuracy of 70%. These prediction studies suggest that disulfide connectivity is not only determined by the local protein sequence environment but also depends on the global information of the whole protein. However, although the prediction of disulfide connectivity has reached good accuracy, there is room for further improving the prediction performance.

In this article, we introduce a novel approach that requires only the protein's amino acid sequence as input to predict disulfide connectivity with high accuracy. It uses support vector regression (SVR) based on multiple sequence feature vectors and predicted secondary structure by the PSIPRED program. Our method can achieve an overall prediction accuracy of 74.4% and 77.9% using 4-fold cross-validation, tested on the protein and cysteine pair, respectively when averaged over proteins with two to five disulfide bridges. We further assessed the effects of eight different sequence encoding schemes on the prediction performance and compared the prediction accuracy of our method with other disulfide-predicting approaches. The results demonstrated that our method has performed in most cases better than other prediction methods. This proposed approach could be a useful tool in large-scale automated prediction of disulfide connectivity patterns in protein sequences.

2 METHODS

2.1 Datasets

In the present study, in order to objectively compare our method with other available approaches reported previously, we used the same dataset that was originally developed by Fariselli and Casadio (2001). This dataset was extracted from Swiss-Prot 39 release and contains only intrachain disulfide bond annotations that were experimentally verified, whereas the interchain disulfide bonds were not considered and discarded. We selected the protein sequences with at least two and at most five disulfide bonds for the sake of comparison. Every two sequences in the dataset have the pairwise sequence identity <30%. Another two datasets: SP39-template and SP43 are also used in this study. The detailed description of these datasets is given in Supplementary Table 1 available at our website.

We performed the 4-fold cross-validation test to evaluate our method based on this non-homologous dataset. The whole dataset was randomly divided into four subsets of roughly equal size. In each validation step, one subset was selected for testing, while the rest were used as the training dataset. The SP39 dataset as well as the SP39-template and the SP43 dataset with detailed information about the protein Swiss-Prot ID, disulfide connectivity annotation and amino acid sequence have been made available as Supplementary Material at http://foo.maths.uq.edu.au/~huber/disulfide.

2.2 Support vector regression

Support vector machine (SVM) is a widely used machine-learning method based on Statistical Learning Theory and has found increasingly important applications in many aspects of bioinformatics and computational biology, such as microarray data analysis (Brown *et al.*, 2000), protein subcellular localization prediction (Hua and Sun, 2001; Sarda *et al.*, 2005), protein stability change prediction (Capriotti *et al.*, 2005), proline *cis*/*trans* isomerization prediction (Song *et al.*, 2006), protein fold recognition (Chen and Baldi, 2006), disease-associated single point protein mutations (Capriotti *et al.*, 2006) and protein–protein interaction (Bradford and Westhead, 2005; Shen *et al.*, 2007). The concept of SVM was originally proposed by Vapnik and his coworkers (Vapnik, 2000). The basic idea of SVM is to transform the samples into a high-dimensional feature space and construct an Optimal Separating Hyperplane (OSH) that maximizes its distance from the closest training samples.

As one of SVM's two practical modes (the other one is support vector classification, SVC), SVR is a novel machine-learning method that is receiving more and more attention and has been successfully applied in the prediction tasks of protein B-factors (Yuan *et al.*, 2005), residue contact numbers (Ishida *et al.*, 2006; Yuan, 2005), residue-wise contact orders (Song and Burrage, 2006), missing value estimation in microarray data (Wang *et al.*, 2006), MHC peptide binding affinity (Liu *et al.*, 2006; Wan *et al.*, 2006) and solvent accessibility of transmembrane residues (Yuan *et al.*, 2006).

The objective of the regression problem is to estimate an unknown continuous-valued function $y=f(x)$, which is based on a finite number of samples. In the current study, we want to find the relationship function between the protein sequence and disulfide connectivity pattern. In order to achieve this, we use ε-insensitive support vector regression (ε-SVR) (Vapnik, 2000). Let $\{(x_i, y_i)\}$ $(i=1,\ldots,N)$ denote a set of training data, where the feature vector x_i denotes each cysteine–cysteine pair in a protein sequence with N cysteine pairs, and y_i represents its corresponding probability of forming a disulfide bridge.

Thus, the expected function of SVR can be formulated as

$$f(x_i) = \langle W, \Phi(x_i) \rangle + b, \quad (1)$$

where W is the weight defining the solution of the primal formulation, b is the bias, $\Phi(x_i)$ is a non-linear function mapping the input feature into a higher dimensional space, and $\langle W,\Phi(x_i)\rangle$ is the inner product of W and $\Phi(x_i)$. To estimate the function $f(x)$, the optimization problem of SVR can be transformed into the constrained convex optimization problem:

$$\text{minimize} \quad \frac{1}{2}\|W\|^2+C\sum_{i=1}^{M}(\xi_i+\xi_i^*), \tag{2}$$

$$\text{subject to} \quad \begin{cases} f(x_i)-y_i \le \varepsilon+\xi_i \\ y_i-f(x_i) \le \varepsilon+\xi_i^* \\ \xi_i,\xi_i^* \ge 0, \quad i=1,\ldots,M, \end{cases} \tag{3}$$

where C is the regularization parameter that determines the trade-off between the margin and prediction error. Here, ξ_i and ξ_i^* are two positive slack variables used to measure the deviation of samples outside the error tube.

To solve this optimization problem, two Lagrange multipliers were added to the condition equations, and therefore the final regression function can be formulated as

$$f(x)=\sum_{i=1}^{M}(\alpha_i-\alpha_i^*)K(x_i,x)+b, \tag{4}$$

where α_i and a_i^* are Lagrange multipliers to be determined, and the kernel function $K(x_i,x)=\langle\Phi(x_i),\Phi(x)\rangle$, which can take different forms such as the linear kernel function, polynomial kernel function, radial basis kernel function, sigmoid kernel function and a user-defined kernel function. The support vectors are those with corresponding non-zero values of the Lagrange multipliers.

We trained and constructed our SVR classifiers based on the Radial Basis Function (RBF kernel), which is given by

$$K(x_i,x)=\exp(-\gamma||x_i-x||^2). \tag{5}$$

There are two parameters needed to be determined in advance to optimize the SVR training. They are the regularization parameter C and the kernel parameter γ. The former is the cost parameter and the latter determines the width of RBF kernel. The selection of the kernel function parameters is an important step for SVR training and testing, as it implicitly determines the structure of the high-dimensional feature space when constructing the OSH. In the present study, we selected the RBF kernel function at $\varepsilon=0.01$, $\gamma=0.01$ and $C=5.0$ to build the SVR models. This combination of parameters has been proven to yield the best performance in our previous studies (Song and Burrage, 2006; Yuan, 2005; Yuan *et al.*, 2005). For the implementation of Vapnik's SVR algorithm, we used the SVM_light (Joachims, 1999) software package.

2.3 Definition of disulfide connectivity pattern

The disulfide connectivity pattern denotes the connectivity arrangement of the oxidized cysteine residue pairs involved in forming disulfide bridges (Chuang *et al.*, 2003; van Vlijmen *et al.*, 2004). For example, a protein with two disulfide bridges has three disulfide connectivity patterns: 1-2_3-4, 1-3_2-4 and 1-4_2-3, where 1, 2, 3 and 4 correspond to the sequential numbering of the disulfide-bonded cysteine residues in a protein, '-'means there is a disulfide bond formed between these two cysteine residues and '_' is used to separate different disulfide bonds. To further illustrate this definition, we highlighted a case protein with four disulfide bonds as an example in Figure 1 (Swiss-Prot ID: MAMB_DENJA in the SP39 dataset), whose disulfide connectivity pattern is defined as 1-3_2-4_5-6_7-8.

For a protein with B disulfide bridges, the number of possible disulfide connectivity patterns N_{ptn} will be:

$$N_{\text{ptn}}=\prod_{i\le B}(2i-1)=\frac{(2B)!}{B!2^B}. \tag{6}$$

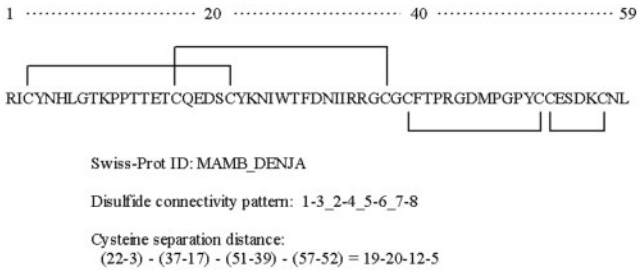

Fig. 1. Example of disulfide connectivity pattern and cysteine separation distance. The black line indicates that there is a disulfide bridge formed between two corresponding cysteines.

For instance, a protein with four and five disulfide bridges would have $N_{\text{ptn}}=7\times5\times3\times1=105$ and $N_{\text{ptn}}=9\times7\times5\times3\times1=945$ disulfide connectivity patterns, respectively. Thus, the number of disulfide connectivity pattern will increase dramatically with the increasing number of disulfide bonds in a protein sequence.

2.4 Sequence encoding schemes

Selecting appropriate sequence encoding schemes is an important step as it determines the quality of feature extraction of SVR models and thus has a significant meaning for the prediction performance.

2.4.1 Multiple sequence feature (MSF) vectors We employed the multiple sequence feature vectors proposed by Chen and Hwang (2005) as the input to our SVR models. They were composed of six sequence feature descriptors: cysteine–cysteine coupling, 20 amino acid compositions, cysteine separation distance, cysteine ordering, protein molecular weight and protein sequence length.

Cysteine–cysteine coupling pair: this vector describes the local sequential environments of two coupled cysteine residues. Numerous previous studies have well established that evolutionary information contained in multiple sequence alignments (MSAs) in the form of position-specific scoring matrices (PSSMs) can significantly improve the overall prediction performance (Ferre and Clote, 2005a,b; Rost and Sander, 1993; Song and Burrage, 2006; Song *et al.*, 2006; Yuan *et al.*, 2005, 2006). This idea was originally proposed and applied by Rost and Sander (1993) in the secondary structure prediction. We ran a three-iteration PSI-BLAST program against the NCBI non-redundant database using a default *E*-value cutoff to obtain these PSSM profiles.

Each disulfide-bonded cysteine residue in a local sequence window was encoded as a vector with 20 elements that represent the probabilities of 20 amino acids occurring at this position. For a cysteine pair forming disulfide connectivity, the PSSM profiles were concatenated using their single cysteine profiles. In this study for all the proteins with different numbers of disulfide bridges, we consistently set up the local window size as 13, because this window size has been demonstrated to lead to the best performance in previous works (Chen *et al.*, 2006; Tsai *et al.*, 2005). Hence in summary, a cysteine–cysteine coupling pair was encoded by a $2\times13\times20=520$-dimensional vector.

Amino acid compositions: these are 20-dimensional vectors and are generally considered as a representation of global information of protein sequence features. The amino acid compositions of the 20 amino acid types are calculated using the following equation:

$$\text{AA}_i=\frac{\sum_{i=1}^{20}n_i}{L} \tag{7}$$

where AA_i is the percentage of residue type i and n_i is the number of residue type i occurring in a protein with sequence length L, respectively. We used the notation A to denote this encoding scheme.

Cysteine separation distance: this is also denoted as DOC (sequence Distance between Oxidized Cysteines) in the literature (Chen *et al*., 2006; Tsai *et al*., 2005). DOC is defined as

$$\mathrm{DOC}(i,j) = \|i-j\| \tag{8}$$

where, i and j represent the two oxidized cysteines that form a disulfide bridge. As indicated by Tsai *et al*. (2005), normalizing the DOC value using the logarithm function can significantly improve the prediction accuracy, when compared with other scaling methods based on either protein sequence length or the maximum DOC value of the whole dataset. Therefore, we also considered incorporating this normalized vector into our SVR models. We used the symbol D to denote this sequence encoding scheme.

Cysteine ordering: this vector describes the sequential order difference between each cysteine pair and was originally suggested by Chen *et al*. (2006). For instance, a protein with three disulfide bridges that are formed between cys 21 and cys 42 (cysordering residues 1 and 4), cys 28 and 60 (cysordering residues 2 and 5), and cys 36 and 66 (cysordering residues 3 and 6) will have the following cysorder: (1/6, 4/6, 2/6, 5/6, 3/6, 6/6) = (0.1667, 0.6667, 0.3333, 0.8333, 0.5000, 1.000). We used the notation O to denote this sequence encoding scheme.

Protein molecular weight (Proweight): our previous work demonstrated that incorporating global features such as protein molecular weight could yield better prediction performance (Song and Burrage, 2006). The normalized Proweight value is given by:

$$y_i = \frac{y_i' - \bar{y}}{\mathrm{SD}}, \tag{9}$$

where y_i is the normalized Proweight value of protein i, y_i' is the raw Proweight value of protein i, $\bar{y}$ is the mean raw Proweight value computed on the whole dataset and SD is the standard deviation based on the whole dataset. The raw Proweight y_i' can be calculated by summing up all its residues using their individual residue molecular weights in a protein i.

Protein sequence length (Prolength): similar to Proweight, the protein sequence length is also a representation of global information of a protein sequence. We encoded this vector into SVR models after the normalization using their respective mean Prolength values and SDs using Equation (9) based on the current dataset. We use L to denote this sequence encoding scheme.

2.4.2 Incorporating predicted secondary structure information We also take into consideration using the predicted probability matrices of secondary structure states from PSIPRED (Jones, 1999) to further enhance the prediction performance, whose output provides the reliability indices for all the three secondary structure states (helix, strand and coil) for each residue in a protein sequence. In an earlier work, both the actual secondary structure annotated by the DSSP program and the predicted secondary structure information obtained by PSIPRED were originally introduced and explored by Ferre and Clote (2005a,b) to infer disulfide connectivity using neural networks. In other prediction studies, the predicted secondary structure by PSIPRED has been proved to lead to a considerable prediction improvement in predicting proline *cis*/*trans* isomerization and residue-wise contact order in proteins (Song and Burrage, 2006; Song *et al*., 2006).

In this study, we applied the PSIPRED algorithm against each protein sequence in the three datasets in order to generate the secondary structure prediction output files and we subsequently extracted the $M \times 3$ matrix from the output file of PSIPRED using a sliding window scheme, where M is the target sequence length centered at the disulfide-bonded cysteines (we adopted $M = 13$ in this study) and 3 is the number of secondary structure types. Therefore, in total, a cysteine–cysteine pair was encoded by a $520 + 78 + 20 + 1 + 1 + 2 + 1 = 623$-dimensional vector.

2.5 Predicting disulfide connectivity patterns

In this study, we have reduced the problem of predicting disulfide connectivity patterns to predicting the disulfide-bonding probability of a cysteine–cysteine pair using SVR by combining their respective sequence profiles with other sequence features. The architecture of our SVR prediction framework is shown in Supplementary Figure 1.

The sequence input vectors of this system consist of seven parts: (1) the PSI-BLAST profiles in the form of PSSMs; (2) the PSIPRED-predicted secondary structure (shortened as PSS); (3) 20 amino acid contents (AA); (4) the normalized protein sequence length (Prolength); (5) the normalized protein molecular weight (Proweight); (6) Cysteine ordering (Cysorder) and (7) the normalized cysteine separation distance (DOC). In particular, we employed a sliding window method to extract the PSI-BLAST profiles and PSIPRED profiles centered at the disulfide-bonded cysteine residue and then formed a cysteine–cysteine coupling pair to concatenate them. We then trained and tested our SVR models based on different combinations of sequence encoding schemes. As a final step, the prediction decisions are made by summing the probabilities of all the possible disulfide connectivity patterns and ranking them according to their respective scores. The disulfide connectivity pattern that has the largest probability score will be predicted as the result.

The prediction problem of disulfide connectivity can be solved by drawing a maximum-weight matching graph whose nodes are disulfide-bonded cysteines and whose edge weight is the potential disulfide-bonding probability of the corresponding cysteine pair. The disulfide connectivity pattern can be assigned and predicted by finding the perfect matching using Edmond's maximum weight matching algorithm (Edmonds, 1965). This prediction strategy has been employed by a number of previous studies in the literature (Cheng *et al*., 2006; Fariselli and Casadio, 2001; Ferre and Clote, 2005a,b; Tsai *et al*., 2005).

We directly solved this difficult prediction problem without exhaustively transforming it into a maximum weight matching problem. On the other hand, by predicting the disulfide-bonding probability of each cysteine pair and then ranking the probability score of every possible disulfide connectivity pattern, our approach has another important advantage over the pattern-wise prediction method, i.e. we considerably reduced the imbalance problem that results from the high positive/negative ratio when adopting the pattern-wise method.

2.6 Performance evaluation

In order to be consistent with the previous studies, we employed the same two assessment measures Q_c and Q_p to evaluate the predictive power of our classifiers (Chen and Hwang, 2005; Lu *et al*., 2007), on the basis of cysteine pair and protein level, respectively.

Q_c (the cysteine pair-based or disulfide bridge-based measure, i.e. the fraction of correctly predicted disulfide bridges in a protein) is given by

$$Q_c = \frac{N_c}{T_c} \tag{10}$$

where N_c is the number of disulfide bridges that are correctly predicted, and T_c is the total number of disulfide bridges in the test dataset.

Q_p (the protein-based measure, i.e. the fraction of proteins whose disulfide connectivity patterns are all predicted correctly) is given by

$$Q_p = \frac{N_p}{T_p} \tag{11}$$

where N_p is the number of proteins whose disulfide connectivity patterns are correctly predicted, and T_p is the total number of proteins in the test dataset.

The results obtained in this study were evaluated using a 4-fold cross-validation procedure, i.e. the dataset was randomly divided into four

Table 1. Prediction accuracies in terms of Q_p and Q_c (%) based on different sequence encoding schemes

Sequence encoding schemes	$B=2$		$B=3$		$B=4$		$B=5$		Overall	
	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
L	80.8	80.8	63.0	70.3	76.8	83.3	44.5	62.2	70.4	75.1
S	71.2	71.2	37.0	49.1	52.6	61.2	17.8	37.8	50.5	55.7
L+S	81.4	81.4	63.0	69.8	78.8	84.6	44.5	61.2	71.1	75.3
L+S+A	81.4	81.4	63.0	69.8	78.8	84.6	44.5	60.4	71.1	75.2
L+S+A+W+H	80.8	80.8	63.7	71.2	78.8	84.6	44.5	61.2	71.1	75.6
L+S+A+W+H+O	78.8	78.8	63.0	70.1	78.8	84.6	44.5	61.0	70.2	74.8
L+S+A+W+H+D+O	85.9	85.9	67.1	72.8	79.8	84.8	46.8	62.7	74.4	77.6
L+S+A+W+H+D	86.5	86.5	67.1	72.6	78.8	84.8	46.8	64.0	74.4	77.9

This result was drawn based on the SP39 dataset using 4-fold cross-validation.

groups, with each group containing roughly equal numbers of protein sequences. Each group was singled out in turn as the testing dataset, while the remaining proteins in other groups were used as the training dataset.

3 RESULTS

3.1 Effects of different sequence encoding schemes

Unless otherwise stated, we refer to the sequence encoding schemes in the following studies based on PSI-BLAST profile, PSIPRED-predicted secondary structure, amino acid composition, protein molecular weight, protein sequence length, cysteine ordering and cysteine separation with respect to the distance of two oxidized cysteines, as 'L', 'S', 'A', 'W', 'H', 'O' and 'D', respectively. We carried out an extensive investigation based on eight different combinations of sequence encoding schemes in order to evaluate their corresponding prediction performance. The performance comparison of these eight different sequence encoding schemes is presented in Table 1.

3.1.1 The benchmark prediction accuracy using local sequence only in the form of PSI-BLAST profiles As can be seen from Table 1, if using PSI-BLAST profiles only, our SVR classifier based on MSAs alone could provide a benchmark overall prediction accuracy of 70.4% and 75.1% that was evaluated by Q_p and Q_c measures, respectively. This observation is consistent with numerous previous prediction studies that PSI-BLAST profiles in the form of PSSMs contain important evolutionary information for accurately predicting disulfide connectivity patterns (Chen and Hwang, 2005; Chen *et al*., 2006; Ferre and Clote, 2005b; Lu *et al*., 2007).

3.1.2 Improved prediction by the predicted secondary structure The SVR classifier based on the predicted secondary structure information alone could provide a benchmark overall prediction accuracy of 50.5% and 55.7% evaluated by Q_p and Q_c measures, respectively. Furthermore, when combining 'L' and 'S' sequence encoding schemes, i.e. adopting 'L+S' scheme, SVR could achieve an overall prediction accuracy of $Q_p=71.1\%$ and $Q_c=75.3\%$, respectively. In contrast, Ferre and Clote (2005b) also developed a predictor based on a neural network and secondary structure information by PSIPRED. Their method achieved the best prediction accuracy of $Q_p=49\%$. Please note that this accuracy was achieved based on a different dataset originally prepared by Vullo and Frasconi (2004). However, both studies have indicated that considering secondary structure information can lead to a significant performance improvement, since statistical analyses have revealed that there is a distinct bias in the secondary structure preferences of disulfide-bonded and non-disulfide-bonded cysteines (Ferre and Clote, 2005b). The improvement on the prediction performance is a reflection of this preference.

3.1.3 DOC can significantly improve the prediction performance The prediction accuracy of our SVR classifier can be further improved by incorporating other informative sequence features such as cysteine separation distance in terms of the DOC value. This finding is consistent with the observations from Tsai *et al*. (2005) and Chen *et al*. (2006), who found that using a normalized DOC value has a significant influence on the prediction performance improvement. Furthermore, if using an 'L+S+A+W+H+D+O' sequence encoding scheme, our method could reach an overall prediction accuracy of $Q_p=74.4\%$ and $Q_c=77.6\%$, respectively.

It is also worth noting that the inclusion of some global sequence features with respect to amino acid composition and cysteine ordering does not necessarily lead to performance improvement in our studies. For example, compared with $Q_c=75.3\%$ obtained using 'L+S' scheme, 'L+S+A' has a decreased prediction accuracy of $Q_c=75.2\%$. This is also the case when we look at the prediction performance based on 'L+S+A+W+H+D+O' and 'L+S+A+W+H+D'. Although the former has more informative feature vectors, the latter has a Q_c accuracy higher by 0.3%.

3.2 Comparison with other approaches

Supplementary Table 2 shows the prediction comparison with other disulfide connectivity approaches. Although this comparison in some ways is misleading because some prediction studies are performed on different datasets, our SVR approach can provide at least comparable or much better prediction accuracy compared with most of other prediction algorithms. Moreover, for proteins with $B=2$, our SVR method provides an improvement of 0.8% higher in Q_p and Q_c than any other method.

In order to explore these issues further, we note that Lu and colleagues have recently developed a computational method that used a genetic algorithm (GA) to optimize the feature selection process and obtained an overall prediction accuracy 73.9% using 4-fold validation on a non-homologous dataset of 482 sequences (Lu *et al*., 2007). This is a state-of-art prediction performance up to date. Although the prediction performance of the GA algorithm showed improvement over our approach (in the case of $B=3$, their results are 7.5% higher in Q_p accuracy and 7.1% higher in Q_c accuracy than our method, respectively. While in the case of $B=5$, their results are 0.8% higher in Q_p accuracy and 7.4% higher in Q_c accuracy, respectively), our method provides at least a comparable or competitive prediction performance when compared with other prediction algorithms.

Moreover, to further validate our SVR approach, we employed the more rigorous and objective independent hold-out test adopted previously by Zhao *et al*. (2005) and Chen *et al*. (2006) using the SP39-template as the training dataset and the SP43 dataset as the independent testing dataset. The sequences in these two datasets share sequence identity lower than 30%. The prediction comparison results are shown in Supplementary Table 3. The overall prediction accuracy (Q_p) of the SVR approach is 3% and 9% higher than the pair-wise SVM and CSP methods, respectively, but 1% lower than the two-level SVM method. Altogether, these results indicate that the SVR approach provides at least a comparable prediction performance in comparison with other methods, suggesting that the SVR approach is a useful machine-learning method and should be a powerful tool in accurately predicting disulfide connectivity patterns by using appropriate multiple sequence feature vectors.

3.3 Case study

For a better understanding of the significance of the Q_c and Q_p measures used in this study, we highlighted four representative cases as an elucidation, which can be seen from Figure 2A, B, C and D, respectively. They are the heat-stable enterotoxinII (Swiss-Prot ID: HSTI_ECOLI and PDB ID: 1EHS), plasma retinol-binding protein (Swiss-Prot ID: RETB_PIG and PDB ID: 1AQB), Lysozyme C (Swiss-Prot ID: LYC_MELGA and PDB ID: 135L) and Type-2 ice-structuring protein (Swiss-Prot ID: ANP_HEMAM and PDB ID: 135L).

In the first three cases, SVR based on the sequence encoding scheme 'L' successfully predicted all of their actual disulfide connectivity patterns. However, in the last case of the type-2 ice-structuring protein, its disulfide connectivity pattern was predicted as 1-10_2-3_4-6_5-8_7-9 by using the 'L' sequence encoding scheme. The first two disulfide bridges in this protein are wrongly predicted as 1-10 and 2-3, respectively, which is presented in light black in Figure 2D. However, after adopting the 'L+S+A+W+H+D' scheme, the original pattern 1-2_3-10_4-6_5-8_7-9 is correctly predicted.

3.4 Prediction web server

We have implemented an online prediction web server (available at http://foo.maths.uq.edu.au/~huber/disulfide) for predicting disulfide connectivity patterns in proteins, which employed the methodology used in this study. The web server has an easy-to-use interface and accepts a single protein amino acid sequence in the form of the one-letter FASTA format. In addition, two SVR models are provided for users' options that are built based on the SP39 dataset and the SP39-template dataset, respectively. After the prediction task for the query sequence is accomplished, users will immediately receive a web link by email that points to a temporary webpage containing the prediction results.

4 DISCUSSION

Accurately predicting disulfide connectivity patterns could provide important information towards the prediction of protein structure. This study presented a prediction framework that uses both multiple sequence feature vectors and PSIPRED-predicted secondary structure information, aiming to provide some deep insights into the sequence–structure relationship between protein primary sequence and its disulfide connectivity patterns. We believe that functionally conserved disulfide connectivity patterns are encoded by protein sequence. We have investigated the effects of different sequence encoding schemes on the prediction performance of disulfide connectivity. We have analyzed the extent to which disulfide connectivity patterns can be correctly identified by using different knowledge of sequence encoding schemes as the input to our SVR approach. We find that of different sequence encoding schemes to SVR predictors, the PSI-BLAST profiles in the form of position-specific scoring matrices, the predicted secondary structure using the PSIPRED program and the normalized DOC value are three of the most important features that are very likely to result in significant performance improvement.

Our approach presented here bears three key advantages over current approaches that could account for the reported prediction success in this study. First, we trained and tested our SVR models using protein sequences as a whole dataset, instead of training and testing them separately using the subgroups with different disulfide bridges. This strategy is different from most previous SVM-based studies that built their SVM predictors by training them based on protein subgroups with the same number of disulfide bridges (Chen and Hwang, 2005; Lu *et al*., 2007). Therefore, for the relatively small numbers of proteins with $B=4$ and $B=5$, our strategy makes them more adequately represented and more adequately trained after input into the SVR models. Second, the combination of all these three sequence encoding schemes, i.e. multiple sequence vectors especially in terms of DOC and predicted secondary structure by PSIPRED can considerably improve the prediction performance. This suggests that as expected, the inclusion of more informative and complementary sequence features do have an important impact on the prediction accuracy. However, it must be pointed out that sequence features in terms of amino acid composition and cysteine ordering in this study have no significant effect on the prediction accuracy. The reasons for this are not clear and need further investigation, but are beyond the scope of this article. Third, our approach can greatly reduce the imbalance problem by predicting the disulfide-bonding probability of each cysteine pair and subsequently ranking the probability score of every possible disulfide connectivity pattern, thus avoiding the exhaustive transformation into a

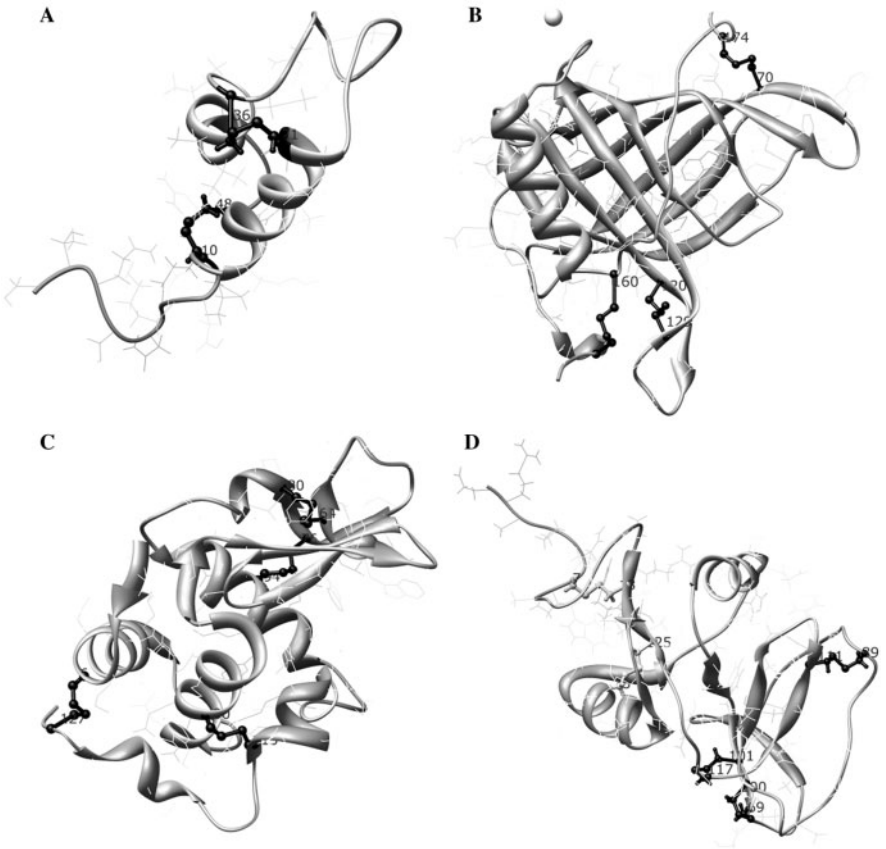

Fig. 2. Four prediction examples of proteins with 2, 3, 4 and 5 disulfide bridges by using our SVR-based approach: (**A**) heat-stable enterotoxin II (Swiss-Prot ID in the SP39 dataset: HSTI_ECOLI, and PDB ID: 1EHS) with disulfide connectivity pattern: 1-4_2-3; (**B**) plasma retinol-binding protein (Swiss-Prot ID in the SP39 dataset: RETB_PIG, which was renamed to RETBP_PIG in release 46.1, and PDB ID: 1AQB) with disulfide connectivity pattern: 1-5_2-6_3-4; (**C**) lysozyme C (Swiss-Prot ID in the SP39 dataset: LYC_MELGA, which was renamed to LYSC_PHACO in release 46.1, and PDB ID: 135L) with disulfide connectivity pattern: 1-8_2-7_3-5_4-6 and (**D**) Type-2 ice-structuring protein (Swiss-Prot ID in the SP39 dataset: ANP_HEMAM, which was renamed to LYSC_PHACO in release 46.1, and PDB ID: 135L) with disulfide connectivity pattern: 1-2_3-10_4-6_5-8_7-9. Disulfide bridges are represented using ball-and-stick models and their corresponding disulfide-bonded cysteine positions are denoted by numbers in gray. The correctly predicted disulfide bridges are shown in dark black, while the incorrectly predicted disulfide bonds are presented in light black. These three-dimensional molecular images were rendered using UCSF Chimera package (Pettersen *et al*., 2004).

maximum weight matching problem. In addition, it takes 429 CPU seconds to predict and enumerate all the disulfide connectivity patterns for the SP43 dataset with 313 proteins using the SVR model built on the SP39-template dataset as the training dataset.

Our prediction results have clearly demonstrated that the proposed method can substantially improve disulfide connectivity assignment accuracy, especially for those protein sequences that are distantly related in terms of sequence homology, and will be very useful in disulfide connectivity annotation of previously uncharacterized sequences generated by high-throughput large-scale genome sequencing projects. However, it must be noted that for proteins with five disulfide bridges, our SVR method did not perform well. One possible reason for this is that the relatively small dataset size makes it less likely to be represented when building SVR models. Therefore, effort in effectively representing the under-represented protein numbers should be of some assistance. On the other hand, we believe that the mutual information of coupled candidate cysteine residues in protein sequence that contains conserved evolutionary information of forming disulfide connectivity patterns has important effects on the reliable assignment of disulfide connectivity. This should be taken into consideration and be incorporated into the prediction models when attempting to develop appropriate sequence-based methods in the future.

5 CONCLUSION

In this article, we developed a novel approach based on SVR that uses multiple sequence feature vectors and predicted secondary structure by PSIPRED. In particular, we used SVR to predict the disulfide-bonding probability with respect to each cysteine–cysteine pair in a protein for the sake of ranking all possible disulfide connectivity patterns, thus directly solving this difficult prediction problem without transforming into a maximum weight matching one. Our method achieved an overall prediction accuracy of $Q_p = 74.4\%$ and $Q_c = 77.9\%$,

respectively. This score was averaged on the proteins with two to five disulfide bridges with the prediction performance comparing favorably to other algorithms in the literature. Our work provides a complementary method to the current prediction algorithms used in computationally identifying disulfide connectivity patterns in disulfide-rich proteins and is a further step towards automatic annotation of protein sequences generated by large-scale genome sequencing projects.

ACKNOWLEDGEMENTS

The authors would like to thank Mr Bo-Juen Chen (Columbia University) and Dr Jianlin Cheng (University of Central Florida) for generously providing the datasets used in this investigation. This work was supported by the grants from Australian Research Council (ARC). K.B. gratefully acknowledges support of the Australian Research Council for the award of a Federation Fellowship.

Conflict of Interest: none declared.

REFERENCES

Abkevich,V.I. and Shakhnovich,E.I. (2000) What can disulfide bonds tell us about protein energetics, function and folding: simulations and bioinformatics analysis. *J. Mol. Biol.*, **300**, 975–985.

Baldi,P. *et al.* (2005) Large-scale prediction of disulphide bond connectivity. In Saul,L.K. *et al.* (ed.), *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, pp. 97–104.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bradford,J.R. and Westhead,D.R. (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.

Brown,M.P.S. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Capriotti,E. *et al.* (2005) Predicting protein stability changes from sequences using support vector machines. *Bioinformatics*, **21** (Suppl. 2), ii54–ii58.

Capriotti,E. *et al.* (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, **22**, 2729–2734.

Ceroni,A. *et al.* (2006) DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res.*, **34**, W177–W181.

Cheek,S. *et al.* (2006) Structural classification of small, disulfide-rich protein domains. *J. Mol. Biol.*, **359**, 215–237.

Chen,B.J. *et al.* (2006) Disulfide connectivity prediction with 70% accuracy using two-level models. *Proteins*, **64**, 246–252.

Chen,Y.C. and Hwang,J.K. (2005) Prediction of disulfide connectivity from protein sequences. *Proteins*, **61**, 507–512.

Cheng,J. and Baldi,P. (2006) A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, **22**, 1456–1463.

Cheng,J. *et al.* (2006) Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins*, **62**, 617–629.

Chuang,C.C. *et al.* (2003) Relationship between protein structures and disulfide-bonding patterns. *Proteins*, **53**, 1–5.

Edmonds,J. (1965) Paths, trees, and flowers. *Can. J. Math.*, **17**, 449–467.

Fariselli,P. and Casadio,R. (2001) Prediction of disulfide connectivity in proteins. *Bioinformatics*, **17**, 957–964.

Fariselli,P. *et al.* (2002) A neural network based method for predicting the disulfide connectivity in proteins. In Damiani,E. *et al.* (ed.), *Knowledge Based Intelligent Information Engineering Systems and Allied Technologies (KES 2002)*. Vol. 1, IOS Press, Amsterdam, pp. 464–468.

Ferre,F. and Clote,P. (2005a) DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res.*, **33**, W230–W232.

Ferre,F. and Clote,P. (2005b) Disulfide connectivity prediction using secondary structure information and diresidue frequencies. *Bioinformatics*, **21**, 2336–2346.

Gupta,A. *et al.* (2004) A classification of disulfide patterns and its relationship to protein structure and function. *Protein Sci.*, **13**, 2045–2058.

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

Hartig,G.R. *et al.* (2005) Intramolecular disulphide bond arrangements in nonhomologous proteins. *Protein Sci.*, **14**, 474–482.

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Inaba,K. *et al.* (2006) Crystal structure of the DsbB-DsbA complex reveals a mechanism of disulfide bond generation. *Cell*, **127**, 789–801.

Ishida,T. *et al.* (2006) Potential for assessing quality of protein structure based on contact number prediction. *Proteins*, **64**, 940–947.

Joachims,T. (1999) Making large-scale SVM learning practical. In Schölkopf,B. *et al.* (eds.), Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge, MA. http://svmlight.joachims.org/

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Kadokura,H. *et al.* (2003) Protein disulfide bond formation in prokaryotes. *Annu. Rev. Biochem.*, **72**, 111–135.

Kadokura,H. *et al.* (2004) Snapshots of DsbA in action: detection of proteins in the process of oxidative folding. *Science*, **303**, 534–537.

Liu,W. *et al.* (2006) Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics*, **7**, 182.

Lu,C.H. *et al.* (2007) Predicting disulfide connectivity patterns. *Proteins*, **67**, 262–270.

Pettersen,E.F. *et al.* (2004) UCSF Chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612. http://www.rbvi.ucsf.edu/chimera/

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Sarda,D. *et al.* (2005) pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinformatics*, **6**, 152.

Sevier,C.S. *et al.* (2007) Modulation of cellular disulfide-bond formation and the ER redox environment by feedback regulation of Ero1. *Cell*, **129**, 333–344.

Shen,J. *et al.* (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl Acad. Sci. USA*, **104**, 4337–4441.

Song,J. and Burrage,K. (2006) Predicting residue-wise contact orders in proteins by support vector regression. *BMC Bioinformatics*, **7**, 425.

Song,J. *et al.* (2006) Prediction of *cis*/*trans* isomerization in proteins using PSI-BLAST profiles and secondary structure information. *BMC Bioinformatics*, **7**, 124.

Thangudu,R.R. *et al.* (2005) Native and modeled disulfide bonds in proteins: knowledge-based approaches toward structure prediction of disulfide-rich polypeptides. *Proteins*, **58**, 866–879.

Thornton,J.M. (1981) Disulphide bridges in globular proteins. *J. Mol. Biol.*, **151**, 261–287.

Thornton,J.M. (2001) From genome to function. *Science*, **292**, 2095–2097.

Tsai,C.H. *et al.* (2005) Improving disulfide connectivity prediction with sequential distance between oxidized cysteines. *Bioinformatics*, **21**, 4416–4419.

van Vlijmen,H.W. *et al.* (2004) A novel database of disulfide patterns. *J. Mol. Biol.*, **335**, 1083–1092.

Vapnik,V. (2000) *The nature of statistical learning theory*. Springer, New York.

Vullo,A. and Frasconi,P. (2004) Disulfide connectivity prediction using recursive neural networks and evolutionary information. *Bioinformatics*, **20**, 653–659.

Wan,J. *et al.* (2006) SVRMHC prediction server for MHC-binding peptides. *BMC Bioinformatics*, **7**, 463.

Wang,X. *et al.* (2006) Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, **7**, 32.

Yuan,Z. (2005) Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics*, **6**, 248.

Yuan,Z. *et al.* (2005) Prediction of protein B-factor profiles. *Proteins*, **58**, 905–912.

Yuan,Z. *et al.* (2006) Predicting the solvent accessibility of transmembrane residues from protein sequence. *J. Proteome Res.*, **5**, 1063–1070.

Zhao,E. *et al.* (2005) Cysteine separations profiles on protein sequences infer disulfide connectivity. *Bioinformatics*, **21**, 1415–1420.