Prediction of Disulfide Connectivity From Protein Sequences

Yu-Ching Chen[1] **and Jenn-Kang Hwang**[1,2*]
[1] *Institute of Bioinformatics, National Chiao Tung University, Taiwan, Republic of China*
[2] *Department of Biological Science and Technology, National Chiao Tung University, Taiwan, Republic of China*

ABSTRACT **The difficulties in predicting disulfide connectivity from protein sequences lie in the nonlocal properties of the disulfide bridges that involve cysteine pairs at large sequence separation. Though some progress has been recently made in the prediction of disulfide connectivity, the current methods predict less than half of the disulfide patterns for the data set sharing less than 30% sequence identity. In this report, we use the support vector machines based on sequence features such as the coupling between the local sequence environments of cysteine pair, the cysteines sequence separations, and the global sequence descriptor, such as amino acid content. Our approach is able to predict 55% of the disulfide patterns of proteins with two to five disulfide bridges, which is 11–26% higher than other methods in the literature. Proteins 2005;61:507–512.**
© 2005 Wiley-Liss, Inc.

Key words: disulfide connectivity; disulfide patterns; support vector machines

INTRODUCTION

Disulfide bonds are known to play an important structural role in stabilizing protein conformations by reducing the number of unfolded conformations.[1–7] Since disulfide bonds impose geometrical constraints on the protein backbones, the disulfide patterns may well dictate to a certain degree the overall three-dimensional (3D) protein structures. Indeed, recent works[8–11] have shown that the disulfide patterns are closely related to protein structures. There are a number of efforts[12–21] to model disulfide bridges or disulfide-rich systems either from protein sequences or from 3D structures. On the other hand, disulfide bonds are more than just inert structural motifs: It is known that the functions of some secreted soluble proteins and cell-surface receptors depend on the cleavage of their disulfide bonds.[22] Therefore, the knowledge of the disulfide patterns is vital in the study of structure and function of proteins.

Recently, computational biology has made significant progress in the prediction of the bonding states from protein sequences.[23–26] A number of approaches based on neural networks,[23,25] statistical analysis,[24] or support vector machines[26] (SVMs) have been shown to be quite effective in predicting the bonding state of cysteine (around 81–90% prediction accuracy). However, predicting disulfide connectivity from protein sequences remains a challenging problem in computational biology. This is because the disulfide bridges are nonlocal in nature (i.e., though the two cysteines that form the disulfide bridge are close in 3D space, they may be far apart from each other in the sequence). Hence, the prediction of disulfide connectivity requires extracting information about spatial proximity of cysteine pairs from one-dimensional protein sequences. The problem is further complicated by the rapid increase of possible disulfide patterns as the number of disulfide bridges increases. For example, when the number of disulfide bridges is two, there are three possible disulfide patterns; but when the number of disulfide bridges increases to five, the possible number of disulfide patterns rapidly increases to 945. To the best of our knowledge, the first attempt to predict the locations of disulfide bridges directly from protein sequences was done by Fariselli and Casadio.[27] They reduced disulfide connectivity to the graph matching (GM) problem in which the graph vertices are equivalent to the residues of cysteine-forming disulfide bridges, and the weight edges contact potentials. The Monte Carlo (MC) simulated annealing method is used to optimize the weights, and the disulfide bridges are then identified by finding the maximal weight perfect matching. We will refer to this method as MCGM. Fariselli et al.[28] improved their results by using neural networks (NNs) to predict the cysteine pairwise interactions. This method will be referred to here as NNGM. Later, Vullo and Frasconi[29] used an ad hoc recursive neural network (RNN) to predict disulfide connectivity. The performance of RNN is comparable or better than MCGM and NNGM. In general, these approaches predict 29–44% of the disulfide patterns for a data set sharing less than 30% sequence identity, after a four-fold cross-validation procedure. In this report, we use SVMs based on feature vectors such as the coupling between the local sequence environments of cysteine pairs, the cysteine separations, and the amino acid content. Our results compare favorably with those of other approaches.[27–29]

Grant sponsor: National Science Council. Grant sponsor: National Research Program of Genomic Medicine. Grant sponsor: University System of Taiwan and the Veteran General Hospital (joint grant).

*Correspondence to: Jenn-Kang Hwang, Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 30050, Taiwan. E-mail: jkhwang@cc.nctu.edu.tw

Received 4 March 2005; Accepted 29 April 2005

Published online 16 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20627

METHODS

Support Vector Machines

The SVM has found many applications[26,30–32] in computational biology and has been shown to be a quite effective machine-learning method. Since this method is quite well known, we give only a brief description of the basic theory behind the SVM. The SVM is basically a binary classifier. Given training vectors x_i, $i = 1,\ldots,l$ and a vector y defined as $y_i = 1$ if x_i is in class I, and $y_i = -1$ if x_i is in the class II. The support vector technique tries to find the separating hyperplane $w^T x_i + b = 0$ with the largest distance between two classes, measured along a line perpendicular to this hyperplane, which is equivalent to solving the following problems:

$$\min_{w,b,\xi} \frac{1}{2}{}^T w + C\left(\sum_{i=1}^{1} \xi_i\right) \text{ and } y_i[(w^T \phi(x_i) + b] \geq 1 - \xi_i. \quad (1)$$

Constraints $y_i[(w^T \phi(x_i)) + b] \geq 1 - \xi_i$ allow that training data may not be on the correct side of the separating hyperplane $w^T x + b = 0$. C is the penalty parameter to be optimized. In practice, the explicit form of $\phi(x)$ is not required, and we only need to calculate the kernel function given by $K(x_i,x_j) \equiv \phi(x_i)^T \phi(x_j)$. We use the radial basis function (RBF) kernel given by $e^{-\gamma\|x_i - x_j\|^2}$ for all the computations, where γ is the kernel parameter. All the SVM calculations are performed using LIBSVM.[33] For SVM training, a few parameters such as the penalty parameter C and the kernel parameter γ of the RBF function must be determined in advance. Choosing optimal parameters for SVMs is an important step in SVM design. In this work, we use cross-validation on different parameters for the model selection.[34]

Data Sets

We followed the same criteria as previous works[27,29] in selecting the sequences from the SWISS-PROT database release No. 39.[35] The constructed data set contains only the sequences with experimentally verified intrachain disulfide bridge annotations, and excludes the sequences whose disulfide bonds are assigned as "probable," "potential," or "by similarity." We consider the sequences with two to five disulfide bridges ($B = 2,\ldots,5$), which account for more than 80% of SWISS-PROT sequences. The final data set contains 482 sequences, of which 168 have two disulfide bonds ($B = 2$), 177 have three ($B = 3$), 95 have four ($B = 4$), and 42 have five ($B = 5$). We further group the sequences into four sets: Each set is selected in such a way that sequence homology among the sets is less than 30%, and the number of sequences of each set is approximately equal. These sets are used for the four-fold cross-validation procedures, as in the previous works.[27,29]

Feature Vectors

The selection of relevant features in large and complex biological data sets significantly affects the effectiveness of the SVM method. We select three types of feature vectors: the coupling between the local sequence environments of cysteine pairs, the cysteine sequence separations, and the amino acid content.

Cysteine–cysteine coupling

A sequence window of size $2l + 1$ amino acids centered on the cysteine is used to describe the neighboring sequence environment of the cysteine. Evolution information of the protein sequence is included in the window by using the sequence profile generated by PSI-BLAST[36] [i.e., the position-specific substitution matrix (PSSM)]. The use of the PSSM has the advantage of avoiding the time-consuming multiple-sequence alignment procedures. The PSSM of a protein sequence is an $L \times 20$ matrix, where L is the sequence length and 20 is the number of amino acid types (amino acid types are numbered from 1 to 20). The matrix element p_{ij} of the PSSM represents the log-odds score of the ith amino acid of type j. Each 20-element row vector of the PSSM represents the distribution of the occurrences of 20 amino acid types at the specific position. Let $\mathbf{w}_i = (a_{i-l},\ldots,a_{i-1},a_i,a_{i+1},\ldots a_{i+l})$ denote the sequence window of size $2l + 1$ centered around the bonded cysteine at the ith position, where a_k is the kth amino acid. We define a 20-element vector $\mathbf{v}_w = (\nu_1{}^{w_i}, \nu_2{}^{w_i} \ldots \nu_{20}{}^{w_i})$ associated with the sequence window $\mathbf{w}_i$, where $\nu_k{}^{w_i}$ is the PSSM element of the amino acid type k. If the amino acid of a given type occurs more than once within the window, $\nu_k{}^{w_i}$ is the sum of the associated PSSM elements. The coupling between the ith and jth cysteines is computed by $s_{ij} = c'_i \mathbf{v}_{w_j} + c'_j \mathbf{v}_{w_j}$, where c'_k is the PSSM element of cysteine type at the kth row. For a given disulfide pattern, we sum up all the possible cysteine pairs to get $\mathbf{s} = \sum_{ij} \mathbf{s}_{ij}$. We use the symbol S to denote the cysteine–cysteine coupling of disulfide patterns. After preliminary experiment, we set the window size to be 21 for $B = 3$ and 5, 7 for $B = 2$, and 27 for $B = 4$.

Cysteine spacing patterns

For a disulfide protein with n cysteines (i.e., $c_1,c_2,\ldots,c_n$), its disulfide pattern is denoted by $(c_i c_j, c_{i'} c_{j'}, \ldots)$, where $c_i c_j$ designates a disulfide bridge formed between cysteines i and j. For a given disulfide pattern $(c_i c_j, c_{i'} c_{j'}, \ldots)$, there is an associated cysteine spacing pattern given by $(d_{ji}, d_{j'i'}, \ldots)$, where d_{ji} is the sequence spacing between c_i and c_j. An example is given in Figure 1. For a protein with four cysteines, $c_1 c_2 c_3 c_4$, which form two disulfide bonds, there will be three possible disulfide configurations: $C_1 = (c_1 c_2, c_3 c_4)$, $C_2 = (c_1 c_3, c_2 c_4)$, and $C_3 = (c_1 c_4, c_2 c_3)$. The three corresponding cysteine spacing patterns are given by $D_1 = (d_{12}, d_{34})$, $D_2 = (d_{13}, d_{24})$, and $D_3 = (d_{14}, d_{23})$. We use the symbol D to denote the cysteine separation vector.

Amino acid content

Amino acid content has been shown to be a useful global sequence descriptor in fold recognition,[31] and in the prediction of the bonding states of cysteines[26] and protein subcellular localization.[32] Amino acid content is represented by the composition vector $A = (a_1,a_2,\ldots,a_{20})$, where $a_k = n_k/n_0$. Here n_k is the number of occurrences of the amino acid of type k, and n_0 is the total number of amino

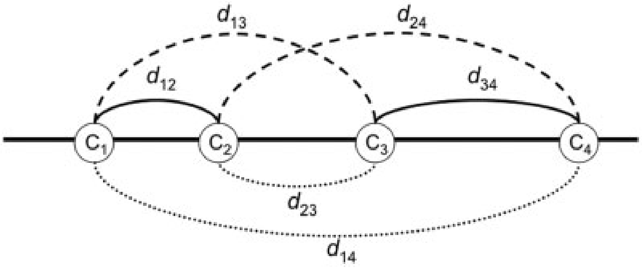

Fig. 1. Example of disulfide patterns consisting of four cysteines $c_1c_2c_3c_4$, which form two disulfide bonds. Three possible disulfide patterns are (c_1c_2,c_3c_4), (c_1c_3,c_2c_4), and (c_1c_4,c_2c_4), where c_ic_j indicates a disulfide bridge between c_i and c_j. And the corresponding cysteine spacing patterns are given by (d_{12},d_{34}) (solid lines), (d_{13},d_{24}) (dashed lines), and (d_{14},d_{23}) (dotted lines).

acids of the query sequence. We will use the notation A to denote the encoding of the amino acid content.

Performance Assessment

To evaluate the performance of the classifiers, we use two assessment of measures[27,29]: Q_c, a cysteine pair-based measure of the fraction of the correctly predicted disulfide bridges, and Q_p, a pattern-based measure of the fraction of proteins whose global disulfide pattern is correctly predicted. Q_p is the more stringent performance index. Specifically, they are defined as

$$Q_c = \frac{1}{N_c}\sum_{i=1}^{N_c}\delta_{c_i} \tag{2}$$

$$Q_p = \frac{1}{N_p}\sum_{i=1}^{N_p}\delta_{p_i}, \tag{3}$$

where δ_{c_i} is defined for the i^{th} disulfide bridge as

$$\delta_{c_i} = \left\{\begin{array}{l} 1,\text{ if the } i^{\text{th}} \text{ predicted disulfide} \\ \quad\text{bridge is correct} \\ 0,\text{ if the } i^{\text{th}} \text{ predicted disulfide} \\ \quad\text{bridge is incorrect} \end{array}\right\}$$

and N_c is the total number of disulfide bridges. Similarly, δ_{p_i} is defined for the ith disulfide proteins as

$$\delta_{p_i} = \left\{\begin{array}{l} 1,\text{ if the predicted connectivity pattern} \\ \quad\text{of the } i^{\text{th}} \text{ protein is correct'} \\ 0,\text{ if the predicted connectivity pattern} \\ \quad\text{of the } i^{\text{th}} \text{ protein is incorrect'} \end{array}\right.$$

and N_p is the total number of disulfide proteins.

RESULTS AND DISCUSSION

Table I summarizes the performances of SVMs based on various encodings. We also list the results computed from the random predictor, referred to as R, as the reference of the base performance. The Q_p and Q_c of the random predictor are given by $1/(2B - 1)!!$ and $1/(2B - 1)$, respectively.[27] In general, the pattern-based Q_p is lower than the disulfide bridge-based Q_c, since the former counts only those proteins whose complete disulfide patterns are correctly predicted. In the case of $B = 2$, both D and S classifiers perform similarly (67%). However, it is interesting to note that the much simpler A classifier, which uses only global sequence information of amino acid content, gives fairly good results (61%). In the case of $B = 3$, the differences in the predictive performance among the classifiers start to show themselves. The D classifier performs significantly better, and, in terms of the more stringent Q_p, it is 16% and 7% higher than A and S, respectively. Note that the D encoding does not contain any information about the explicit amino acid sequence other than the cysteine separations. This is consistent with previous works[10,11] indicating that disulfide patterns and cysteine separations are closely related to each other and that disulfide patterns can be effectively used to detect remote homologues undetectable by the sequence alignment methods. In the case of $B = 4$ and 5, the prediction accuracies of the SVMs, though significantly better than those of the random predictor, are not yet practical at present. The poor results for these cases are due to the relatively smaller number of the reliably annotated proteins with higher number of disulfide bridges in the data set (see the Methods section on data sets). However, the situation is expected to improve when more structures are available in the future. On the other hand, when comparing the results of the D classifiers with those of the random predictor R, we found that, the ratios of Q_p between D and R are 28 and 120 for $B = 4$ and 5, respectively, indicating that the SVM is still effective in these cases.

We have previously shown in many biological applications[26,31,32] that using multiple-feature vectors can improve on the performance of the SVM classifiers based on a single-feature vector type. We selected the following linear combinations: $D + w_AA$, $D + w_SS$, and $D + w_AA + w_SS$, where w_d is the weight associated with the d encoding. After preliminary experiment, we set the weights to be $w_A = 1$ and $w_s = 0.001$. For the sake of simplicity, we will use the simpler notations $D + A$, $D + S$, and $D + A + S$, with the understanding that w_A and w_S are omitted from the notations. Table II compares the performances of the

TABLE I. Performance of the SVMs Based on a Single-Feature Vector Type

	$B = 2$		$B = 3$		$B = 4$		$B = 5$		$B = 2...5$	
Method	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
R	0.33	0.33	0.06	0.20	0.01	0.14	0.001	0.11	0.14	0.20
A	0.61	0.61	0.38	0.51	0.13	0.20	0.07	0.27	0.39	0.42
S	0.67	0.67	0.47	0.60	0.17	0.24	0.12	0.32	0.45	0.48
D	0.67	0.67	0.54	0.64	0.28	0.39	0.12	0.30	0.50	0.54

TABLE II. Performances of the SVMs Based on Multiple-Feature Vectors

	$B = 2$		$B = 3$		$B = 4$		$B = 5$		$B = 2...5$	
Method	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
$D + A$	0.74	0.74	0.54	0.64	0.28	0.39	0.12	0.30	0.52	0.55
$D + S$	0.71	0.71	0.60	0.66	0.30	0.41	0.12	0.30	0.54	0.55
$D + S + A$	0.74	0.74	0.61	0.69	0.30	0.40	0.12	0.31	0.55	0.57

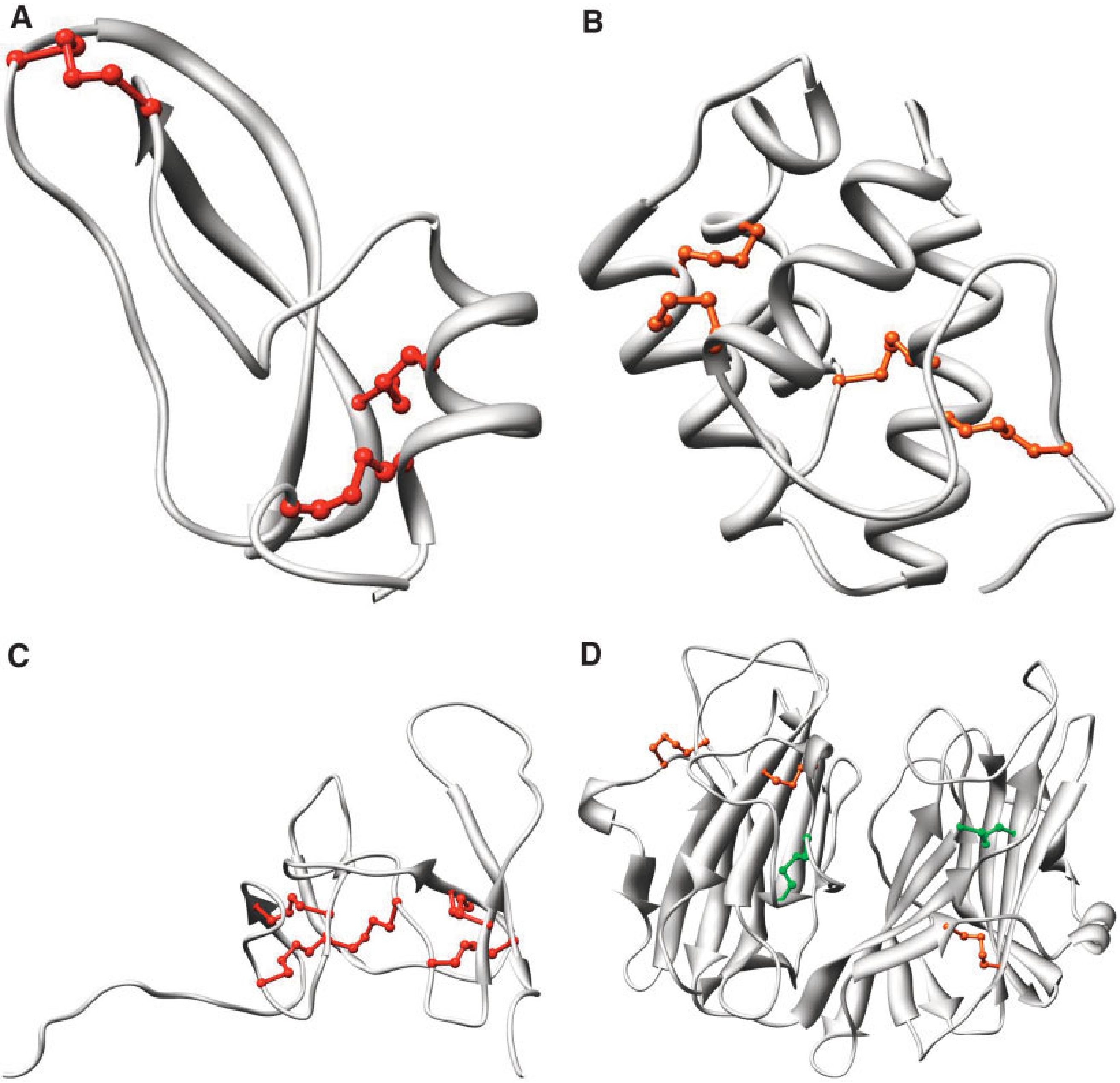

Fig. 2. The ribbon models of (**A**) the bovine pancreatic trypsin inhibitor (1tpa:I), (**B**) the nonspecific lipid transfer protein (1afh), (**C**) porcine pancreatic procolipase (1pcn), and (**D**) peptidylglycine α-hydroxylating monooxygenase (1phm). The disulfide bonds are represented in the ball-and-stick model. The correctly predicted disulfide bridges are in red, while the incorrectly predicted ones are in green. The molecular images were generated by UCSF Chimera.[41]

TABLE III. Comparison of Predictive Performances of Different Approaches to Predict Disulfide Connectivity

	$B = 2$		$B = 3$		$B = 4$		$B = 5$		$B = 2..5$	
Method	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
MCGM[27]	0.56	0.56	0.21	0.36	0.17	0.37	0.02	0.21	0.29	0.38
NNGM[28]	0.68	0.68	0.22	0.37	0.20	0.37	0.02	0.26	0.34	0.42
RNN[29]	0.73	0.73	0.41	0.51	0.24	0.37	0.13	0.30	0.44	0.49
This work	0.74	0.74	0.61	0.69	0.30	0.40	0.12	0.31	0.55	0.57

SVMs based on the multiple feature vectors. As expected, the SVMs based on the multiple-feature vectors in general perform better than those based on a single-feature vector type.

Figure 2 shows some typical examples of the predictions by the $D + A + S$ classifier. Figure 2(A) shows the case of $B = 3$, 1tpa:I,[38] which is a bovine pancreatic trypsin inhibitor; Figure 2(B), the case of $B = 4$, 1afh,[39] a nonspecific lipid transfer protein; and Figure 2(C), the case of $B = 5$, 1pcn,[40] a porcine pancreatic procolipase. In these cases, the disulfide bridges are all perfectly predicted. The number of incorrectly predicted disulfide bridges, if any, will be either greater than or equal to 2, since one incorrectly predicted disulfide bridge will necessarily give rise to another one. An example is given in Figure 2(D). The observed and the predicted disulfide patterns of 1phm[41] (peptidylglycine α-hydroxylating monooxygenase) are [1–6,2–4,3–5,7–10,8–9] and [1–6,2–4,*3–8*,7–10,*5–9*], respectively (the incorrect predictions are in italics). Hence, in the case of $B = 2$, the cysteine pair-based measure Q_c of a protein is either 0 or 1, while in the case of $B = 3$, Q_c is 1, $\frac{1}{3}$, or 0.

Comparison with other methods: Table III compares the results of the $D + A + S$ with those of other methods. The $D + A + S$ is the only method that gives the overall prediction accuracy above 50% ($Q_p = 0.55$ and $Q_c = 0.57$), while the other methods give 0.29–0.44 in Q_p and 0.38–0.49 in Q_c. In the case of $B = 2$, the $D + A + S$ and the RNN give similar prediction accuracies. In the case of $B = 3$, the $D + A + S$ outperforms other approaches by 20–40% in Q_p and by 14–30% in Q_c. In the case of $B \geq 4$, the $D + A + S$ is better than or comparable with those of other methods; however, since the sample size of these cases is relatively small, it is not easy to draw a conclusion of statistical significance.

CONCLUSION

Though the SVM is known to be a powerful machine learning method, due to the complexity of biological data, the identification and selection of relevant biological features become an important issue in the applications of SVMs to biological problems. In this work, we tested SVMs in the prediction of disulfide connectivity using biological features characteristic of disulfide bridges. Our results indicate that both cysteine–cysteine sequence couplings and cysteine separations are important features in predicting disulfide connectivity. This is consistent with the previous studies[10,11] indicating that a close relationship exists between cysteine separations and disulfide patterns, and that such a relationship can be utilized to identify the remote homologs undetectable by sequence alignments. We showed that the SVM based on the cysteine separations give the best predictive performance among the SVMs based on the single-feature vector. We also showed that the SVMs based on the multiple-feature vectors outperform those based on the single-feature vector. At present, our method may be useful in the prediction of disulfide bridges, especially in the cases of $B = 2$ and 3. As for the cases of higher number of disulfide bridges (i.e., $B \geq 4$), our approach is expected to be applicable when there are more reliably annotated disulfide proteins available in the future.

ACKNOWLEDGMENTS

We are grateful for both hardware and software support from the Structural Bioinformatics Core Facility at National Chiao Tung University. We would like to thank Professor Chih-Jen Lin for helpful discussion concerning SVMs.

REFERENCES

1. Anfinsen CB, Scheraga HA. Experimental and theoretical aspects of protein folding. Adv Protein Chem 1975;29:205–300.
2. Clarke J, Fersht AR. Engineered disulfide bonds as probes of the folding pathway of barnase: increasing the stability of proteins against the rate of denaturation. Biochemistry 1993;32:4322–4329.
3. Harrison PM, Sternberg MJ. Analysis and classification of disulphide connectivity in proteins: the entropic effect of cross-linkage. J Mol Biol 1994;244:448–463.
4. Abkevich VI, Shakhnovich EI. What can disulfide bonds tell us about protein energetics, function and folding?: simulations and bioninformatics analysis. J Mol Biol 2000;300:975–985.
5. Clarke J, Hounslow AM, Bond CJ, Fersht AR, Daggett V. The effects of disulfide bonds on the denatured state of barnase. Protein Sci 2000;9:2394–2404.
6. Wedemeyer WJ, Welker E, Narayan M, Scheraga HA. Disulfide bonds and protein folding. Biochemistry 2000;39:4207–4216.
7. Yokota A, Izutani K, Takai M, Kubo Y, Noda Y, Koumoto Y, Tachibana H, Segawa S. The transition state in the folding-unfolding reaction of four species of three-disulfide variant of hen lysozyme: the role of each disulfide bridge. J Mol Biol 2000;295:1275–1288.
8. Harrison PM, Sternberg MJ. The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. J Mol Biol 1996;264:603–623.
9. Mas JM, Aloy P, Marti-Renom MA, Oliva B, Blanco-Aparicio C, Molina MA, de Llorens R, Querol E, Aviles FX. Protein similarities beyond disulphide bridge topology. J Mol Biol 1998;284:541–548.
10. Chuang CC, Chen CY, Yang JM, Lyu PC, Hwang JK. Relationship between protein structures and disulfide-bonding patterns. Proteins 2003;53:1–5.
11. van Vlijmen HW, Gupta A, Narasimhan LS, Singh J. A novel database of disulfide patterns and its application to the discovery of distantly related homologs. J Mol Biol 2004;335:1083–1092.
12. Pabo CO, Suchanek EG. Computer-aided model-building strategies for protein design. Biochemistry 1986;25:5987–5991.
13. Hazes B, Dijkstra BW. Model building of disulfide bonds in

proteins with known three-dimensional structure. Protein Eng 1988;2:119–125.

14. Sowdhamini R, Srinivasan N, Shoichet B, Santi DV, Ramakrishnan C, Balaram P. Stereochemical modeling of disulfide bridges: criteria for introduction into proteins by site-directed mutagenesis. Protein Eng 1989;3:95–103.
15. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
16. Huang ES, Samudrala R, Ponder JW. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. J Mol Biol 1999;290:267–281.
17. Mas JM, Aloy P, Marti-Renom MA, Oliva B, de Llorens R, Aviles FX, Querol E. Classification of protein disulphide-bridge topologies. J Comput Aided Mol Des 2001;15:477–487.
18. Dani VS, Ramakrishnan C, Varadarajan R. MODIP revisited: re-evaluation and refinement of an automated procedure for modeling of disulfide bonds in proteins. Protein Eng 2003;16:187–193.
19. Kong L, Lee BT, Tong JC, Tan TW, Ranganathan S. SDPMOD: an automated comparative modeling server for small disulfide-bonded proteins. Nucleic Acids Res 2004;32:W356–W359.
20. Vinayagam A, Pugalenthi G, Rajesh R, Sowdhamini R. DSDBASE: a consortium of native and modelled disulphide bonds in proteins. Nucleic Acids Res 2004;32:D200–D202.
21. Thangudu RR, Vinayagam A, Pugalenthi G, Manonmani A, Offmann B, Sowdhamini R. Native and modeled disulfide bonds in proteins: knowledge-based approaches toward structure prediction of disulfide-rich polypeptides. Proteins 2005;58:866–879.
22. Hogg PJ. Disulfide bonds as switches for protein function. Trends Biochem Sci 2003;28:210–214.
23. Fariselli P, Riccobelli P, Casadio R. Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins 1999;36:340–346.
24. Fiser A, Simon I. Predicting the oxidation state of cysteines by multiple sequence alignment. Bioinformatics 2000;16:251–256.
25. Martelli PL, Fariselli P, Malaguti L, Casadio R. Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. Protein Eng 2002;15:951–953.
26. Chen YC, Lin YS, Lin CJ, Hwang JK. Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. Proteins 2004;55:1036–1042.
27. Fariselli P, Casadio R. Prediction of disulfide connectivity in proteins. Bioinformatics 2001;17:957–964.
28. Fariselli P, Riccobelli P, Casadio R. A neural network-based method for predicting the disulfide connectivity in proteins. In: Damiiani E, Jain LC, Howlett RJ, Ichalkaranje N, editors. Knowledge based intelligent information engineering systems and allied technologies (KES 2002). Vol. 1. Amsterdam: IOS Press; 2002. p 464–468.
29. Vullo A, Frasconi P. Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics 2004;20:653–659.
30. Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 2001;308:397–407.
31. Yu CS, Wang JY, Yang JM, Lyu PC, Lin CJ, Hwang JK. Fine-grained protein fold assignment by support vector machines using generalized n-peptide coding schemes and jury voting from multiple-parameter sets. Proteins 2003;50:531–536.
32. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 2004;13:1402–1406.
33. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines, 2001. Available online (http://www.csie.ntu.edu.tw/~cjlin/libsvm).
34. Duan K, Keerthi SS, Poo AN. Evaluation of simple performance measures for tuning SVM hyperparameters. Neurocomputing 2003;51:41–59.
35. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
37. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605–1612.
38. Marquart M, Walter J, Deisenhofer J, Bode W, Huber R. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. Acta Crystallogr B 1983;39:480–490.
39. Gomar J, Petit MC, Sodano P, Sy D, Marion D, Kader JC, Vovelle F, Ptak M. Solution structure and lipid binding of a nonspecific lipid transfer protein extracted from maize seeds. Protein Sci 1996;5:565–577.
40. Breg JN, Sarda L, Cozzone PJ, Rugani N, Boelens R, Kaptein R. Solution structure of porcine pancreatic procolipase as determined from 1H homonuclear two-dimensional and three-dimensional NMR. Eur J Biochem 1995;227:663–672.
41. Prigge ST, Kolhekar AS, Eipper BA, Mains RE, Amzel LM. Amidation of bioactive peptides: the structure of peptidylglycine alpha-hydroxylating monooxygenase. Science 1997;278:1300–1305.