BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 24 2005, pages 4416–4419
doi:10.1093/bioinformatics/bti715

Sequence analysis

Improving disulfide connectivity prediction with sequential distance between oxidized cysteines

Chi-Hung Tsai[1], Bo-Juen Chen[1], Chen-hsiung Chan[1], Hsuan-Liang Liu[2] and Cheng-Yan Kao[1,3,*]

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan 106, [2]Department of Chemical Engineering and Graduate Institute of Biotechnology, National Taipei University of Technology, Taipei, Taiwan 10608 and [3]Institute for Information Industry, Taipei, Taiwan 106

Received on August 19, 2005; revised and accepted on October 11, 2005
Advance Access publication October 13, 2005

ABSTRACT

Summary: Predicting disulfide connectivity precisely helps towards the solution of protein structure prediction. In this study, a descriptor derived from the sequential distance between oxidized cysteines (denoted as DOC) is proposed. An approach using support vector machine (SVM) method based on weighted graph matching was further developed to predict the disulfide connectivity pattern in proteins. When DOC was applied, prediction accuracy of 63% for our SVM models could be achieved, which is significantly higher than those obtained from previous approaches. The results show that using the non-local descriptor DOC coupled with local sequence profiles significantly improves the prediction accuracy. These improvements demonstrate that DOC, with a proper scaling scheme, is an effective feature for the prediction of disulfide connectivity. The method developed in this work is available at the web server PreCys (prediction of cys–cys linkages of proteins).

Availability: http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/

Contact: cykao@csie.ntu.edu.tw

Supplementary information: Supplementary data, detailed results, tables and information are available at http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/

1 INTRODUCTION

Disulfide bonds, commonly found in extracellular proteins, stabilize folded conformations as they contribute to the stability of the three-dimensional structures with respect to thermodynamics (Wedemeyer *et al*., 2000). Since disulfide bonds impose length and angle constraints on the backbone of a protein, correct prediction of disulfide connectivity can be employed to dramatically reduce the search in conformational space and greatly raise the accuracy for protein structure prediction (Huang *et al*., 1999). Different methods (Fariselli and Casadio, 2001; Fariselli *et al*., 2002; Vullo and Frasconi, 2004) have been developed to predict disulfide connectivity with the prior knowledge of the oxidization states of cysteine residues. These methods can be classified into two categories: (1) patternwise or (2) pairwise. The major difference between them is whether the methodology is developed to deal with alternative disulfide connectivity patterns (Vullo and Frasconi, 2004; Zhao *et al*., 2005) or the relationships between cysteine pairs (Fariselli and Casadio, 2001; Baldi *et al*., 2005; Ferrè and Clote, 2005). This difference decides how the information is encoded. However, the prediction accuracies of these methods are still limited so far (∼50%).

Besides the methodology used, another critical factor determining the predicting performance is the descriptor employed. Fariselli and Casadio (2001) computed residue contact potentials according to the nearest-neighbor residues of bonded cysteines. Secondary structure (Baldi *et al*., 2005; Ferrè and Clote, 2005) and solvent accessibility (Baldi *et al*., 2005) were also used as descriptors to represent input information. All these descriptors only describe the local environments of bonded cysteines. However, a disulfide bridge is a long-range interaction between two linearly distant cysteines. Descriptors containing local information only are insufficient for predicting disulfide connectivity accurately. Therefore, information regarding relationships between cysteines is highly desired.

Harrison and Sternberg (1994) have suggested that sequence separation between bonded cysteines correlates strongly with specific connectivity patterns. Zhao *et al*. (2005) also observed that disulfide connectivity pattern is highly conserved with the same cysteine-separation pattern of oxidized cysteines. Although there have been some attempts (Vullo, 2004; Baldi *et al*., 2005) to take advantage of such information by using descriptors such as positions of cysteines or relative sequence length, no emphasis has been addressed on the effects of these features so far.

In this paper, a descriptor derived from the linear sequence distance between oxidized cysteines (denoted as DOC) was used to demonstrate its power on predicting disulfide connectivity. A pairwise method using support vector machine (SVM) to generate bonding potentials of cysteine pairs was developed. This method was further validated with a dataset derived from Swiss-Prot 39 (SP39), and significant improvements were obtained when the

*To whom correspondence should be addressed.
The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

non-local descriptor DOC coupled with local sequence profiles was applied. These results reveal that DOC is an effective feature in disulfide connectivity prediction. The web interface service of the method proposed in this study for disulfide connectivity prediction is available at http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/

2 METHODOLOGY

2.1 Prediction of the connectivity pattern of disulfide bridges

With prior knowledge of the oxidation states of cysteine residues, a prediction strategy similar to previous studies (Fariselli and Casadio, 2001; Baldi *et al*., 2005; Ferrè and Clote, 2005) was applied. The whole problem was mapped to an undirected complete graph, where oxidized cysteines were considered as vertices and the probabilities of connectivity between cysteine pairs were assigned as the weights of the edges between corresponding vertices. Then the disulfide connectivity pattern can be inferred by solving the maximum weight matching of this graph, which implies maximum probabilities for bonding pairs of this resulting pattern.

2.1.1 SVM In this work, SVM was employed to predict the potential of connectivity between cysteines. SVM has been applied broadly within the field of computational biology to pattern-recognition problems and is a promising technique for data classification (Vapnik, 1998). Given data $x_1, \ldots, x_1$, we set their labels, y_i, as +1 if x_i is in class 1 and as −1 if x_i belongs to class 2. Then with these training data, SVM solves an optimization problem for binary classification:

$$\min_{\omega,b,\xi} \frac{1}{2}\omega^{\mathrm{T}}\omega + C\sum_{i=1}^{l}\xi_i \quad \text{and} \quad y_i\left(\omega^{\mathrm{T}}\varphi(x_i)+b\right) \geq 1-\xi_i, \quad \xi_i \geq 0, \tag{1}$$

where x_i is mapped to a higher dimensional space by the function φ; ξ_i is the training error allowed and C is the cost of error. Moreover, SVM can further be solved to approximate posterior class probability $P(y_i = 1 \mid x_i)$ with a sigmoid function (Platt, 2000):

$$P(y_i = 1 \mid f_i) = \frac{1}{1+\exp(Af_i + B)}, \tag{2}$$

where A and B are parameters and $f_i = \omega^{\mathrm{T}}\varphi(x_i) + b$. Using (2), we can infer the bonding probability for each pair of cysteines. The software LIBSVM (Chang and Lin, 2000), a library for SVMs, was adopted in our experiments.

2.1.2 Data encoding Two descriptors were mainly considered to encode input data for the SVM: (1) local sequence profiles (evolutionary information) around target cysteines from multiple sequence alignments and (2) the linear DOC.

We generated sequence profiles by performing multiple sequence alignments with the widely used program PSI-BLAST (Altschul *et al*., 1997). For each cysteine pair Cys(i, j), profiles were extracted using a window centered at cysteines i and j. The window size indicates the scope of vicinity of the target cysteine and determines how much information is provided for our models. In our experiments, the window size was set to 13, and the values of elements in the profiles were scaled to [0, 1].

For a cysteine pair with sequence indexes i and j, the corresponding DOC is defined as follows:

$$\mathrm{DOC}(i,j) = \|i - j\|. \tag{3}$$

Since scaling approaches may affect the performance of SVM, three scaling schemes for DOC were tested:

(1) $\mathrm{DOC_L}$, DOC normalized with the protein sequence length L.

(2) $\mathrm{DOC_{max}}$, DOC normalized with the maximum value of the whole dataset.

(3) $\mathrm{DOC_{log}}$, DOC values normalized with the logarithm function.

2.1.3 Maximum weight matching Features were encoded with respect to each pair of cysteines, and SVM models were trained with these data to generate posterior probabilities that indicate the potential of connectivity between cysteine pairs. After the bonding probability of each cysteine pair was produced by SVM models, an implementation of Gabow's algorithm (Gabow, 1973), wmatch (Rothberg, http://elib.zib.de/pub/Packages/mathprog/matching/weighted/), was used to find the maximum weight matching. Finally, the matching with maximum weight was transformed to the corresponding disulfide connectivity pattern.

2.2 Evaluation criteria

Our models were evaluated by Q_{p} and Q_{c} which are defined as follows:

$$Q_{\mathrm{p}} = \frac{C_{\mathrm{p}}}{T_{\mathrm{p}}}, \quad Q_{\mathrm{c}} = \frac{C_{\mathrm{c}}}{T_{\mathrm{c}}}, \tag{4}$$

where C_{p} is the number of proteins whose connectivity patterns are correctly predicted; T_{p} is the total number of proteins in the test set; C_{c} is the number of disulfide bridges correctly predicted and T_{c} is the total number of disulfide bridges in test proteins.

3 IMPLEMENTATION AND RESULTS

3.1 Dataset

In order to compare our method with the approaches reported previously (Vullo and Frasconi, 2004; Baldi *et al*., 2005), the same dataset extracted from SP39 (Bairoch and Apweiler, 2002) was employed. The same filtering procedure (Fariselli and Casadio, 2001) was applied to ensure only high quality and experimentally verified intra-chain disulfide bridge annotations were included. For cross-validation, this dataset was further divided into four subsets so that each of the two shared sequence homology $\leq$30%.

3.2 Cross-validation of SP39

Table 1 lists the accuracies of 4-fold cross-validation performed with the dataset SP39 for our model along with the results reported previously. Using sequence profiles only, our SVM models obtained a Q_{P} of 59%, which is better than those obtained in previous works. This may benefit from the generality of SVM, which avoids overfitting during the training process. Another reason for the improvement is the enlarging of window size when extracting sequence profiles. We tried to use different window sizes to build SVM models, and the accuracy of the predictions is shown in Figure 1. The overall Q_{P} increases with enlarging window size and peaks at 13, which was adopted in this work. Using the same window size of 5 as used by Vullo and Frasconi (2004) and Baldi *et al*. (2005), similar accuracy of 52% was also obtained using our method.

Moreover, when DOC was used, the prediction accuracy was further improved. To explore the effects of scaling schemes on DOC, three scaling functions were considered: $\mathrm{DOC_L}$, $\mathrm{DOC_{max}}$

Table 1. Results of cross-validation on the data extracted from SP39

Methods	$B = 2$		$B = 3$		$B = 4$		$B = 5$		$B = 2$–5	
	Q_p (%)	Q_c (%)	Q_p (%)	Q_c (%)	Q_p (%)	Q_c (%)	Q_p (%)	Q_c (%)	Q_p (%)	Q_c (%)
MC graph-matching[a]	56	56	21	36	17	37	2	21	29	38
NN graph-matching[b]	68	68	22	37	20	37	2	26	34	42
BiRnn-2 profile[c]	73	73	41	51	24	37	13	30	44	49
2D-Rnn profile[d]	74	74	51	61	27	44	11	41	49	56
dNN2[e]	62	—	40	—	55	—	26	—	49	—
CSP	72	72	54	66	33	50	18	36	52	58
SVM profile	76	76	53	62	48	62	44	60	59	65
SVM profile + DOC_{log}	79	79	53	62	55	70	58	71	63	70

[a]Fariselli and Casadio (2001).
[b]Fariselli *et al*. (2002).
[c]Vullo and Frasconi (2004).
[d]Baldi *et al*. (2005).
[e]Ferrè and Clote (2005), only results of Q_p are available.

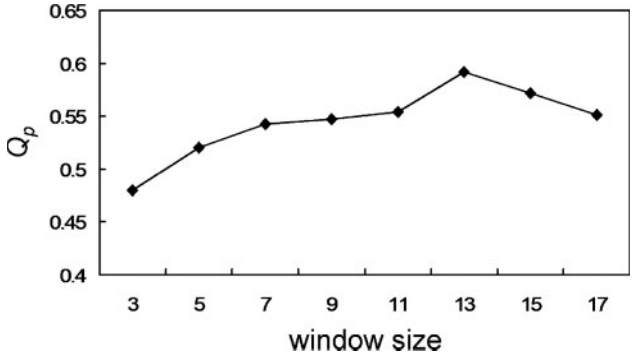


Fig. 1. The accuracy (Q_p) of predictions using different window sizes to extract sequence profiles on the dataset SP39.

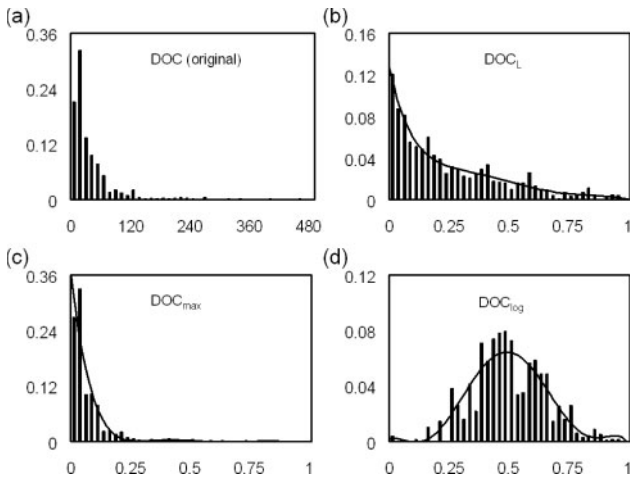


Fig. 2. Histogram of the fraction of chains versus (**a**) the original distribution of DOC without normalization, (**b**) DOC_L, (**c**) DOC_{max} and (**d**) DOC_{log} in the dataset SP39.

and DOC_{log}. The trend of DOC between cysteine bonding pairs in dataset SP39 is shown in Figure 2a, and the distributions of DOC_L, DOC_{max} and DOC_{log} are also shown in Figure 2b–d, respectively. As can be seen, DOC_{max} remains the distribution of the DOC since the scaling is simply performed by dividing the distance with a fixed value. On the other hand, the originally skewed distribution of DOC becomes close to a normal distribution after logarithm function was applied, and the distribution of DOC_L becomes blurred due to the variation of sequence lengths.

The prediction accuracies of 59 and 61% were obtained by using the scaling function DOC_L or DOC_{max}. On the other hand, the highest prediction accuracy of 63% was obtained by using the scaling function DOC_{log}, which was selected to build our SVM models for disulfide connectivity prediction. These results suggest that the scaling of DOC can affect its contribution to our models. With a proper scaling function, DOC can enhance the performance of SVM models.

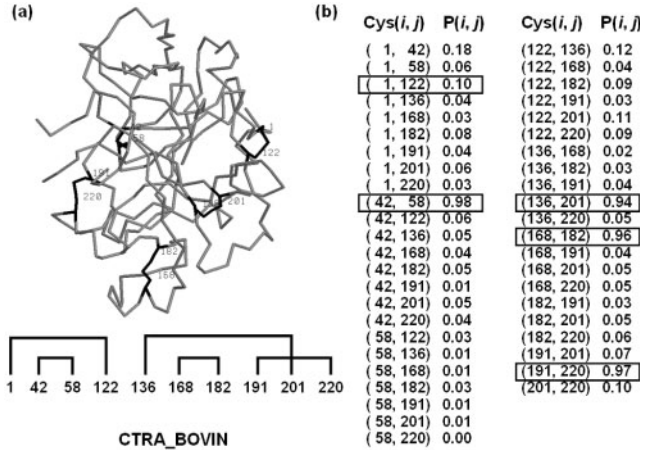
3.3 PreCys (prediction of cys–cys linkages in proteins) web server

The PreCys server (at http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/) provides the service of disulfide connectivity prediction by the method developed in this work. In addition, a simple CSP search can also be accessed on the website. This server provides two SVM models built from Swiss-Prot releases 39 and 47. With the sequence and the positions of oxidized cysteines (optional) input, the bonding probabilities of cysteine pairs and the final connectivity pattern can be generated. Additional experimental results and the chain lists used can be found at this website.

4 DISCUSSION AND CONCLUSION

There are two major categories for the methods of disulfide connectivity prediction. The 'patternwise' approaches take the whole protein as a unit directly and rank alternative connectivity patterns (Vullo and Frasconi, 2004). They can easily include global information, such as the sequence length, amino acid contents or the positions of all cysteines. On the other hand, the 'pairwise' methods (Baldi *et al*., 2005; Ferrè and Clote, 2005) lack the overview of the whole protein and are usually limited to the scope of local environments of cysteines.

However, the patternwise methods often suffer from the problem of insufficient data, especially when the number of disulfide bonds increases. For proteins with five disulfide bonds, there are some



Cys(*i*, *j*)	P(*i*, *j*)	Cys(*i*, *j*)	P(*i*, *j*)
(1, 42)	0.18	(122, 136)	0.12
(1, 58)	0.06	(122, 168)	0.04
(1, 122)	0.10	(122, 182)	0.09
(1, 136)	0.04	(122, 191)	0.03
(1, 168)	0.03	(122, 201)	0.11
(1, 182)	0.08	(122, 220)	0.09
(1, 191)	0.04	(136, 168)	0.02
(1, 201)	0.06	(136, 182)	0.03
(1, 220)	0.03	(136, 191)	0.04
(42, 58)	0.98	(136, 201)	0.94
(42, 122)	0.06	(136, 220)	0.05
(42, 136)	0.05	(168, 182)	0.96
(42, 168)	0.04	(168, 191)	0.04
(42, 182)	0.05	(168, 201)	0.05
(42, 191)	0.01	(168, 220)	0.05
(42, 201)	0.05	(182, 191)	0.03
(42, 220)	0.04	(182, 201)	0.05
(58, 122)	0.03	(182, 220)	0.06
(58, 136)	0.01	(191, 201)	0.07
(58, 168)	0.01	(191, 220)	0.97
(58, 182)	0.03	(201, 220)	0.10
(58, 191)	0.01		
(58, 201)	0.01		
(58, 220)	0.00		

Fig. 3. (**a**) The structure and the connectivity pattern of disulfide bridges and (**b**) the bonding potential $P(i, j)$ for each cysteine pair cys(i, j) generated by SVM model for chymotrypsinogen A (PDB id 1HJA). Selected bonding pairs are boxed.

patterns that only have one instance in the dataset. These patterns are not likely to be predicted correctly by patternwise methods because there is not enough information for model training. For example, the connectivity patterns of the protein chains CTRA_BOVIN (PDB: 1HJA, pattern: [1–4, 2–3, 5–9, 6–7, 8–10], Fig. 3) and UROK_HUMAN (PDB: 1LMW, pattern: [1–3, 2–4, 5–9, 6–7, 8–10]) only appear once in the dataset SP39. The patternwise method CSP fails to predict the disulfide connectivity of these chains, because no template is available for the patterns to be predicted. On the other hand, our pairwise SVM models can still predict their connectivity correctly, since the pattern can be assembled by the bonding pairs predicted.

In addition, the imbalance situation between the positive and negative data differs for pairwise and patternwise methods. As to a protein with B disulfide bonds, the positive/negative ratio is $1{:}(2B - 2)$ for pairwise encoding. However, for the patternwise encoding, the imbalance is more severe, since there is only one correct pattern among the $(2B - 1)!!$ generated entries. Taking $B = 5$ for an example, the positive/negative ratio is only 1:8 in pairwise encoding. With the same bond number B in patternwise encoding, there are 945 entries where the positive/negative ratio is 1:944. Such severe imbalance can bias the learning process and result in poor models. Due to the insufficiency of data and the severe imbalance issue of patternwise encoding, we adopted the pairwise approach in our method.

In this paper, we developed a method to predict disulfide connectivity based on SVMs. The non-local descriptor DOC describing the distance between oxidized cysteines was proposed to encode additional information for our input. For the dataset SP39, the prediction accuracy can be improved significantly with the combination of local sequence profiles and the non-local descriptor DOC. The significant improvement on prediction accuracies against previous approaches is because of the following reasons. First, SVMs can avoid over-fitting problems commonly seen in neural networks and other machine learning methods. Second, we explored the local environments of oxidized cysteines and found the optimum window size with best Q_p values. Third, the non-local descriptor DOC_{log} also contributes to the prediction accuracies. Our method achieved an accuracy of 63% in dataset SP39 when DOC was used, which outperforms other previous approaches. Consistent improvements were also obtained on other datasets, detailed results can be found in the Supplementary data. These results imply that the formation of disulfide linkages between cysteines is determined not only by the local information of cysteines but also by the relationships between them. The descriptor DOC contains important information about the relationships between oxidized cysteines and is an effective feature for predicting disulfide connectivity accurately. This descriptor can be additionally applied to other problems where the knowledge of disulfide bridges is required. The web interface of our program is provided on the PreCys website. The results from our method may be useful for advanced studies in protein structure prediction, protein structure modeling and protein engineering.

ACKNOWLEDGEMENTS

We would like to thank Jianlin Cheng for generously sharing datasets and useful comments and Shih-Chieh Chen for enlightening discussion. Funding to pay the Open Access publication charges for this article was provided by the Institute for Information Industry.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bairoch,A. and Apweiler,R. (2000) The Swiss–Prot protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*., **28**, 45–48.

Baldi,P., Cheng,J. and Vullo,A. (2005) Large-scale prediction of disulphide bond connectivity. In Saul,L.K., Weiss,Y. and Bottou,L. (eds), *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, pp. 97–104.

Chang,C.-C. and Lin,C.-J. (2000) LIBSVM: introduction and benchmarks. *Technical Report*, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

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

Fariselli,P., Riccobelli,P. and Casadio,R. (2002) A neural network based method for predicting the disulfide connectivity in proteins. In Damiani,E., Jain,L.C., Howlett,R.J. and Ichalkaranje,N. (eds), *Knowledge based intelligent information engineering systems and allied technologies (KES 2002)*. IOS Press, Amsterdam, 1, pp. 464–468.

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

Gabow,H.N. (1973) Implementation of algorithms for maximum matching on nonbipartite graphs. Phd Thesis, Stanford University, CA.

Harrison,P.M. and Sternberg,M.J.E. (1994) Analysis and classification of disulphide connectivity in proteins. *J. Mol. Biol*., **244**, 448–463.

Huang,E.S. *et al*. (1999) *Ab initio* fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J. Mol. Biol*., **290**, 267–281.

Platt,J. (2000) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Smola,A.J., Bartlett,P.L., Schölkopf,B. and Schuurmans,D. (eds), *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, pp. 61–74.

Rothberg,E. (1985) wmatch: a C Program to solve maximum weight matching.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York, NY.

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

Wedemeyer,W.J. *et al*. (2000) Disulfide bonds and protein folding. *Biochemistry*, **39**, 4207–4216.

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