Published online 8 June 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W503–W507***
doi:10.1093/nar/gkq514

DBCP: a web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding state of cysteines

Hsuan-Hung Lin[1,2] and Lin-Yu Tseng[3,4,*]

[1]Department of Applied Mathematics, National Chung Hsing University, [2]Department of Management Information System, Central Taiwan University of Science and Technology, [3]Institute of Networking and Multimedia and [4]Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan, ROC

Received February 4, 2010; Revised May 11, 2010; Accepted May 21, 2010

ABSTRACT

The proper prediction of the location of disulfide bridges is efficient in helping to solve the protein folding problem. Most of the previous works on the prediction of disulfide connectivity pattern use the prior knowledge of the bonding state of cysteines. The DBCP web server provides prediction of disulfide bonding connectivity pattern without the prior knowledge of the bonding state of cysteines. The method used in this server improves the accuracy of disulfide connectivity pattern prediction (Q_p) over the previous studies reported in the literature. This DBCP server can be accessed at http://120.107.8.16/dbcp or http://140.120.14.136/dbcp.

INTRODUCTION

Disulfide bonds play an important structural role in stabilizing protein conformations. For the protein folding prediction, a correct prediction of disulfide bridges can greatly reduce the search space (1,2). The prediction of disulfide bonding pattern helps, to a certain degree, predict the 3D structure of a protein and hence its function because disulfide bonds impose geometrical constraints on the protein backbones. Some recent research works had shown the close relation between the disulfide bonding patterns and the protein structures (3,4).

In the realm of the disulfide bond prediction, four problems are addressed. The first is the protein chain classification: to classify if the protein contains disulfide bridge(s) or not, the second is the residue classification: to predict the bonding state of cysteines, the third is the bridge classification and the last is the prediction of the disulfide bonding pattern. Over the past years, significant progress has been made on the prediction of the disulfide bonding states (5–8) and the disulfide bonding pattern (9–17). For disulfide bonding pattern prediction, with the exception of the methods proposed by Ferrè and Clote (11, 12) and Cheng *et al.* (15), the others assume that the bonding states are known. The method proposed by Ferrè and Clote (11,12) and Cheng *et al.* (15) can be applied whether the bonding states are known or not.

In this study, the coordinate (X, Y, Z) of the C_α of each amino acid in the protein predicted by MODELLER (18) is used as the feature. The support vector machine (SVM) is then trained to compute the connectivity probabilities of cysteine pairs. The Edmonds–Gabow maximum weight perfect matching algorithm (19) is utilized to find the connectivity pattern.

SYSTEM

The flowchart of this server illustrated by an example is shown in Figure 1.

FEATURE

With the exception of the protein's secondary structure, the features used in the previous studies on disulfide bonding connectivity prediction are protein sequence features and not related to the protein structure. In this study, we propose to use the structure-related feature. The MODELLER (18) is used to predict the coordinate (X, Y, Z) of the C_α of each amino acid in the protein sequence. Having the coordinates, we can compute the Euclidean

*To whom correspondence should be addressed. Tel: 886 4 22874020; Fax: 886 4 22853869; Email: lytseng@cs.nchu.edu.tw

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

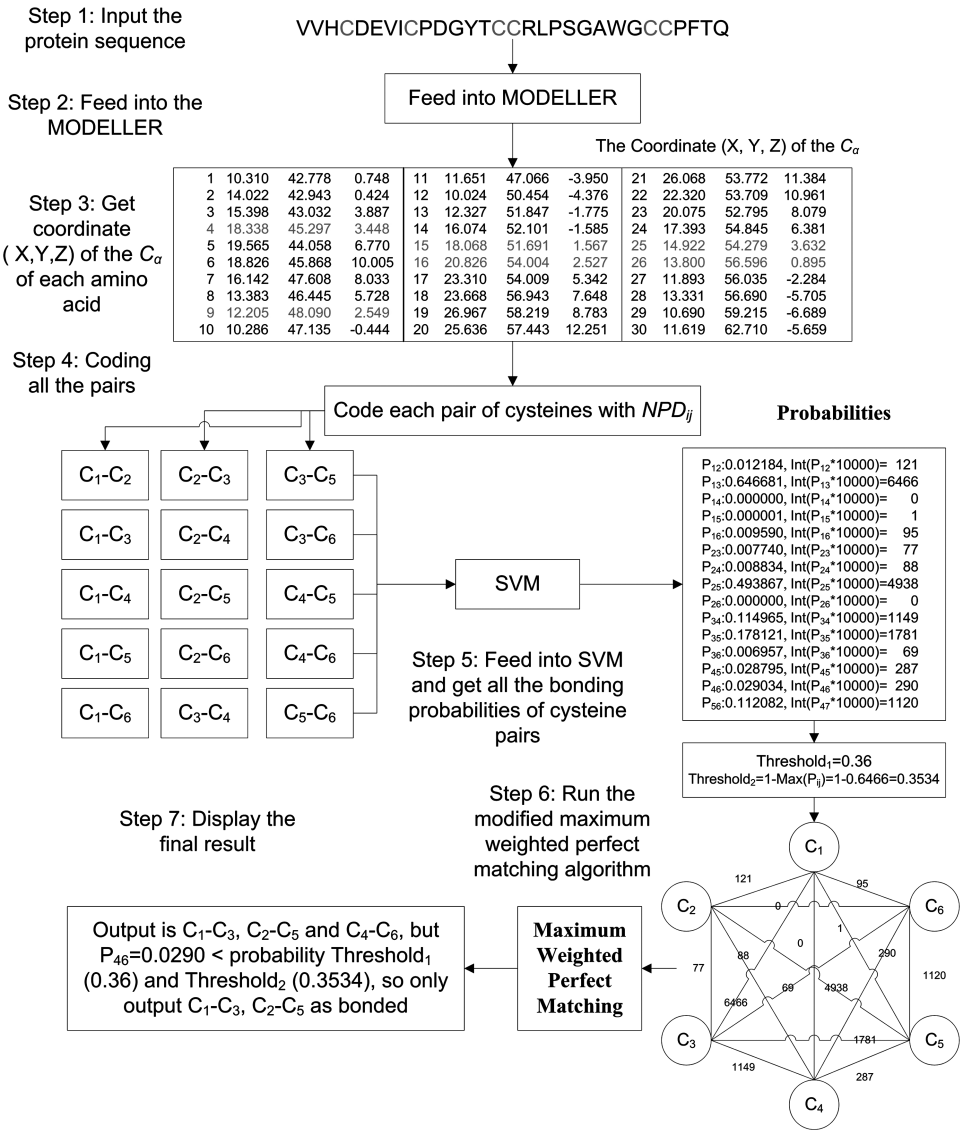

Figure 1. The flowchart of the DBCP illustrated by an example.

distance D_{ij} between the amino acid at the i-th position and the amino acid at the j-th position. We further extend the definition of Euclidean distance to the pair distance (PD). Let the positions of cysteine i and cysteine j be P_i and P_j, respectively. The PD between cysteine i and cysteine j is defined to be a vector $(D_{Pi-[w/2],\ Pj-[w/2]}, \ldots, D_{Pi-1,\ Pj-1}, D_{Pi,\ Pj}, D_{Pi+1,\ Pj+1}, \ldots, D_{Pi+[(w-1)/2],\ Pj+[(w-1)/2]})$ that contains w Euclidean distances, where w is the window size. If we have k cysteines in the protein, there are as many as $C_2^k = k(k-1)/2$ cysteine pairs. Since most cysteine pairs will not constitute a disulfide bond, by examining $D_{Pi,\ Pj}$ of the cysteine pairs that constitute a disulfide bond, we set a threshold of value 15 for $D_{Pi,\ Pj}$. In other words, if $D_{Pi,\ Pj}$ is >15, this pair of cysteines will not be considered as a candidate that may have a disulfide bond. In order to make the values proper to be input to the SVM, which is -1 to 1, each component of the vector PD is normalized by the equation $(D_{ij} - 7.5)/7.5$. The resultant vector is called the normalized PD (NPD) and is the input to the SVM.

INPUT

The inputs to the DBCP include three parts:

(1) Query name is an optional name for the user to distinguish his or her queries. If a user sends more than one query and selects sending the results by the e-mail, the order in which he or she receives the results may not match the order in which he or she sends the queries.
(2) The sequence of a protein (plain sequence without the header, the maximum size is 9000 residues) may be input in the input box. If one presses the [Sample Data] button, the sample input sequence will be displayed on the protein sequence input box.

(3) The email address (If 'send the result by Email' is checked, otherwise, the results will be displayed on the web page).

METHODOLOGY

(1) Run BLAST to get the template sequence of the input sequence. The parameters of BLAST are set as follows: the Expectation value (E) threshold for saving hits is set to a very large value 10 000 and the database is set to pdb that contains sequences derived from the 3D structure records from the Protein Data Bank. If the E-value of the template sequence is >10 or the template sequence shares identity $<25\%$ to the input sequence, instead of going to Step 2, we use the method previously developed by us (20) to predict the disulfide bonding pattern.
(2) Align the input sequence and the template sequence.
(3) Feed the alignment file into MODELLER and run the procedure to evaluate the model of the input sequence using the template sequence.
(4) Get the coordinate (X, Y, Z) of the C_α (α Carbon) of each residue.
(5) Coding each cysteine pair as the NPD, this will be the input to the SVM.
(6) Feed the coding file into the SVM to predict the bonding probability of each cysteine pair with the trained model. The multiple trajectory search (21) is tightly integrated with the SVM training. For more details, please refer to the Supplementary Data on the DBCP web server.
(7) Coding the input file with the probabilities from the SVM output and using the modified weighted perfect matching algorithm to get the first level disulfide bonding connectivity.
(8) Justify the first level disulfide bonding connectivity with the thresholds to get the final result.
(9) Display the result on the web page or send the result to the user.

In Step 1, if the E-value of the template sequence is >10 or the template sequence shares identity $<25\%$ to the input sequence, a previously proposed method (20) is used for prediction. In this method, the position-specific scoring matrix, the normalized bond lengths, the predicted secondary structure of protein and the physicochemical properties index of the amino acid were used as features. The multiple trajectory search and the SVM training were tightly integrated to train the predictor. For more details, please refer to (20).

ENVIRONMENT

The DBCP web server is free and open to all users and there is no login requirement. This prediction software was implemented using C language and the server-side scripting language PHP, and it employed the web page on the Apache web server.

OUTPUT

In this subsection, we introduce the results of the DBCP, as listed below:

(1) Job ID: the id assigned by the server for this query job.
(2) Query name: the same as that was input by the user.
(3) SEQUENCE: the input protein sequence.
(4) Positions of the cysteines.
(5) E-value: the E-value of the template sequence found by the BLAST. If the E-value is >10, the sequence identity and the E-value will be marked in red to indicate a warning.
(6) Identity: the identity between the template sequence and the input sequence. This value is provided by the MODELLER. If the template sequence shares $<25\%$ identity with the input sequence, the sequence identity and the E-value will be marked in red to indicate a warning.
(7) Probability: the prediction probability of each cysteine pair.
(8) Metal binding site score: a rough estimation that the cysteine pair may be involved in the metal binding site.
(9) Positions of oxidized cysteines.
(10) Predicted disulfide bonding connectivity pattern.
(11) Predicted positions of cysteines in metal binding site: this is only a rough prediction and users should consult other methods or web services for more accurate prediction of cysteines involved in metal binding sites (22).

EVALUATION OF WEB SERVER

We found four web sites that provided the prediction of the disulfide bonding connectivity pattern without prior knowledge of bonding state of cysteines (12,14–16). Cheng *et al*. (15) tested their prediction method by a 10-fold cross validation on the data set SPX (15). As a comparison, we also tested our method by a 10-fold cross validation on the same data set, and the results were shown in the Supplementary Data. The method proposed by Song *et al*. (16) can process only protein sequences that have less than 12 cysteines. Therefore, we conducted a test to compare our method only with the other three methods. We took 56 protein sequences from the SWISS-PROT database release no. 56.3 that are neither in the SWISS-PROT release no. 39 nor in the data set SPX, this set of sequences is denoted as 'SP56NS'. The prediction accuracies of our method and the other three methods on this data set are shown in Table 1.

Since the present version of the web server was trained by using the data set SPX, we also took 50 sequences from the SWISS-PROT database release no. 56.3 that are neither in the SWISS-PROT release no. 39 nor in the data set SPX. Furthermore, the pairwise sequence identity of these 50 sequences and the sequences in SPX is $<25\%$. This set of sequences is denoted as 'SP56NS_25'.

Table 1. Comparison of the prediction accuracies on the data set SP56NS

Number of bonds	Number of sequences	DBCP (%)	Dipro (15) (%)	DiANNA (12) (%)	DISULFIND (14) (%)
2	10	50	10	10	0
3	10	80	40	10	30
4	10	70	20	0	0
5	10	60	10	0	0
6–9	16	50	0	0	n.a
All	56	60.7	14.3	3.6	5.4

Table 2. Comparison of the prediction accuracies on the data set SP56NS_25

Number of bonds	Number of sequences	DBCP (%)	Dipro (15) (%)	DiANNA (12) (%)	DISULFIND (14) (%)
2	10	60	20	20	0
3	10	50	30	0	30
4	10	50	30	0	10
5	10	50	0	0	0
6–14	10	30	0	0	n.a
All	50	48	16	4	8

Table 3. The prediction accuracy on the data set CHK25

Number of bonds	Number of sequences	Qp (%)	Qc (%)
2	15	53.3	76.7
3	11	54.5	69.7
4	3	33.3	33.3
6–9	3	0	43.5
All	32	46.9	61.2

The prediction accuracies of our method and the other three methods on this data set are shown in Table 2.

For checking the prediction accuracy when the input sequence has low identity to the overall set of PDB proteins, we took 32 sequences from the SWISS-PROT database release no. 56.3, where either the sequence shares identity <25% to the template sequence found by the BLAST or the *E*-value of the template sequence is >10. This set of sequences is denoted as 'CHK25'. The prediction accuracy of DBCP for this data set is shown in Table 3.

LIMITATIONS

(1) The prediction accuracy may degenerate if the template sequence found by the BLAST has an *E*-value >10 or the identity of the template sequence and the input sequence is <25%.
(2) The web server is designed aiming to predict the disulfide bonding connectivity pattern of a sequence that does not have cysteines involved in the metal binding sites. The prediction accuracy will degenerate if the input sequence contains cysteines involved in the metal binding sites. The metal binding site score is provided only to indicate possible metal binding sites. If this score is >0.5, users are strongly suggested using other methods or web services to more accurately predict the cysteines that are involved in the metal binding sites. The Metal Detector server (http://metaldetector.dsi.unifi.it/) is one of such web services.

CONCLUSION

A web-based application system called the DBCP is provided for the prediction of the disulfide bonding connectivity pattern without the prior knowledge of the bonding state of cysteines. In previous research works, without the prior knowledge of the bonding state of cysteines, to the best of our knowledge, the best accuracy of disulfide connectivity pattern prediction (Q_p) and that of disulfide bridge prediction (Q_c) are 51% and 52%, respectively, on the data set SPX with 10-fold cross validation. The method used in this server improved the prediction accuracies on the same test data set SPX to 84.4% (Q_p) and 94.6% (Q_c) with 10-fold cross validation. The comparison of the prediction accuracy of the DBCP with that of three other state-of-the-arts web services on the data sets SP56NS and SP56NS_25 also reveals that the DBCP outperforms the other three methods.

If the template sequence found by the BLAST has an *E*-value >10 or the identity of the template sequence and the input sequence is <25%, another method previously proposed by us is used for prediction. In this case, the prediction accuracy may slightly degenerate. Since the DBCP is designed aiming to predict the disulfide bonding connectivity pattern of a sequence that does not have cysteines involved in the metal binding sites, for protein sequences that contain cysteines involved in the metal binding sites, other methods that can predict both the disulfide bonds and the metal binding sites will be more suitable for prediction. The high metal binding site score (e.g. >0.5) indicates that there may be cysteines involved in the metal binding sites. In this case, users are strongly suggested using other methods in addition to the DBCP and conclude the prediction result based on the results of all methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors like to thank the anonymous reviewers for pointing out the problems of proteins lacking homology and sequences containing cysteines involved in metal binding sites. The comments of all anonymous reviewers have improved the quality of the paper.

FUNDING

National Science Council of ROC (contract number NSC98-2221-E005-049-MY3, partial); Ministry of Education, Taiwan, ROC under ATU plan (partial); Central Taiwan University of Science and Technology (grant CTU99-P-33). Funding for open access charge: National Chung Hsing University.

Conflict of interest statement. None declared.

REFERENCES

1. Skolnick,J., Kolinski,A. and Ortiz,A.R. (1997) MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.*, **265**, 217–241.
2. Huang,E.S., Samudrala,R. and Ponder,J.W. (1999) Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J. Mol. Biol.*, **290**, 267–281.
3. Chuang,C.C., Chen,C.Y., Yang,J.M., Lyu,P.C. and Hwang,J.K. (2003) Relationship between protein structures and disulfide-bonding patterns. *Proteins*, **55**, 1–5.
4. van Vlijmen,H.W.T., Gupta,A., Narasimhan,L.S. and Singh,J. (2004) A novel database of disulfide patterns and its application to the discovery of distantly related homologs. *J. Mol. Biol.*, **335**, 1083–1092.
5. Fariselli,P., Riccobelli,P. and Casadio,R. (1999) Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. *Proteins*, **36**, 340–346.
6. Fiser,A. and Simon,I. (2000) Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics*, **16**, 251–256.
7. Martelli,P.L., Fariselli,P., Malaguti,L. and Casadio,R. (2002) Prediction of the disulfide-bonding state of cysteines in proteins with hidden neural networks. *Protein Eng.*, **15**, 951–953.
8. Chen,Y.C., Lin,Y.S., Lin,C.J. and Hwang,J.K. (2004) Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. *Proteins*, **55**, 1036–1042.
9. Fariselli,P. and Casadio,R. (2001) Prediction of disulfide connectivity in proteins. *Bioinformatics*, **17**, 957–964.
10. Vullo,A. and Frasconi,P. (2004) Disulfide connectivity prediction using recursive neural networks and evolutionary information. *Bioinformatics*, **20**, 653–659.
11. Ferrè,F. and Clote,P. (2005) Disulfide connectivity prediction using secondary structure information and diresidue frequencies. *Bioinformatics*, **21**, 2336–2346.
12. Ferre,F. and Clote,P. (2006) DiANNA 1.1: An extension of the DiANNA web server for ternary cysteine classification. *Nucleic Acids Res.*, **34**, W182–W185.
13. Chen,B.J., Tsai,C.H., Chan,C.H. and Kao,C.Y. (2006) Disulfide connectivity prediction with 70% accuracy using two-level models. *Proteins*, **64**, 246–252.
14. Ceroni,A., Passerini,A., Vullo,A. and Frasconi,P. (2006) DISULFIND: a Disulfide Bonding State and Cysteine Connectivity Prediction Server. *Nucleic Acids Res.*, **34**, W177–W181.
15. Cheng,J., Saigo,H. and Baldi,P. (2006) Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins*, **62**, 617–629.
16. Song,J., Yuan,Z., Tan,H., Huber,T. and Burrage,K. (2007) Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. *Bioinformatics*, **23**, 3147–3154.
17. Rubinstein,R. and Fiser,A. (2008) Predicting disulfide bond connectivity in proteins by correlated mutations analysis. *Bioinformatics*, **24**, 498–504.
18. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 799–815.
19. Gabow,H.N. (1973) Implementation of algorithms for maximum matching on nonbipartite graphs. *Ph.D Thesis*. Stanford University, CA.
20. Lin,H.H. and Tseng,L.Y. (2009) Prediction of disulfide bonding pattern based on support vector machine with parameters tuned by multiple trajectory search. *WSEAS Trans. Comput.*, **9**, 1429–1439.
21. Tseng,L.Y. and Chen,C. (2008) Multiple trajectory search for large scale global optimization. *Proceedings of 2008 IEEE Congress on Evolutionary Computation, CEC'08*, Hong Kong, 2008, 3052–3059.
22. Lippi,M., Passerini,A., Punta,M., Rost,B. and Frasconi,P. (2008) MetalDetector: a web server for predicting metal binding sites and disulfide bridges in proteins from sequence. *Bioinformatics*, **24**, 2094–2095.