BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 8 2005, pages 1415–1420
doi:10.1093/bioinformatics/bti179

Sequence analysis

Cysteine separations profiles on protein sequences infer disulfide connectivity

East Zhao[1], Hsuan-Liang Liu[2], Chi-Hung Tsai[1], Huai-Kuang Tsai[1], Chen-hsiung Chan[1] and Cheng-Yan Kao[1,*]

[1]Bioinformatics Laboratory, Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan 106 and [2]Department of Chemical Engineering and Graduate Institute of Biotechnology, National Taipei University of Technology, No. 1, Sec. 3, Chung-Hsiao E. Rd., Taipei, Taiwan 10608

Received on July 18, 2004; revised on October 29, 2004; accepted on November 23, 2004
Advance Access publication December 7, 2004

ABSTRACT

Motivation: Disulfide bonds play an important role in protein folding. A precise prediction of disulfide connectivity can strongly reduce the conformational search space and increase the accuracy in protein structure prediction. Conventional disulfide connectivity predictions use sequence information, and prediction accuracy is limited. Here, by using an alternative scheme with global information for disulfide connectivity prediction, higher performance is obtained with respect to other approaches.

Result: Cysteine separation profiles have been used to predict the disulfide connectivity of proteins. The separations among oxidized cysteine residues on a protein sequence have been encoded into vectors named cysteine separation profiles (CSPs). Through comparisons of their CSPs, the disulfide connectivity of a test protein is inferred from a non-redundant template set. For non-redundant proteins in SwissProt 39 (SP39) sharing less than 30% sequence identity, the prediction accuracy of a fourfold cross-validation is 49%. The prediction accuracy of disulfide connectivity for proteins in SwissProt 43 (SP43) is even higher (53%). The relationship between the similarity of CSPs and the prediction accuracy is also discussed. The method proposed in this work is relatively simple and can generate higher accuracies compared to conventional methods. It may be also combined with other algorithms for further improvements in protein structure prediction.

Availability: The program and datasets are available from the authors upon request.

Contact: cykao@csie.ntu.edu.tw

1 INTRODUCTION

A disulfide bond is a strong covalent bond between two cysteine residues in proteins. It plays a key role in protein folding and in determining the structure/function relationships of proteins (Abkevich and Shakhnovich, 2000; Wedemeyer *et al*., 2000; Welker *et al*., 2001). In addition, it is important in maintaining a protein in its stable folded state. A disulfide connectivity pattern can be used to discriminate the structural similarity between proteins (Chuang *et al*., 2003). In protein folding prediction, the knowledge of the locations of disulfide bonds can dramatically reduce the search in conformational space (Skolnick *et al*., 1997; Huang *et al*., 1999). Therefore, a higher performance in predicting disulfide connectivity pattern is likely to increase the accuracy in predicting the three-dimensional (3D) structures of proteins through the reduction of the number of steps during conformational space search.

Generally, the prediction of disulfide connectivity pattern in proteins consists of two consecutive steps. Firstly, the disulfide bonding state of each cysteine residue in a protein is predicted based on its amino acid sequence and evolutionary information using various algorithms, such as neural networks (Fariselli *et al*., 1999; Fiser and Simon, 2000), support vector machines (Chen *et al*., 2004) and hidden Markov models (Martelli *et al*., 2002). Secondly, the location of disulfide bonds is subsequently predicted based on the bonding state of each cysteine residue using algorithms such as Monte Carlo (MC) simulated annealing together with weighted graph matching (Fariselli and Casadio, 2001) and recursive neural networks with evolutionary information (Vullo and Frasconi, 2004). The prediction accuracy of the oxidation state of cysteine residues has reached 90% (Chen *et al*., 2004) and can be used confidently. However, the task of predicting disulfide connectivity remains challenging. The best prediction accuracy ever reported so far is only 44% (Vullo and Frasconi, 2004), in which recursive neural network was used to score connectivity patterns represented in undirected graphs. Such prediction accuracy is still far from being usable, although it is much higher than that by a random predictor.

In this work, cysteine separation profiles (CSPs) of proteins are adopted for the prediction of disulfide connectivity. It has been shown that proteins with similar disulfide bonding patterns also share similar folds (Chuang *et al*., 2003; van Vlijmen *et al*., 2004). Theoretical work has suggested that disulfide bonds may stabilize the structures of protein fragments between the connected cysteine residues (Abkevich and Shakhnovich, 2000); therefore, the separations between oxidized cysteine residues may be used in the task of predicting disulfide connectivity. Previous works on disulfide connectivity predictions have used graphs to represent disulfide connection patterns (Fariselli and Casadio, 2001; Vullo and Frasconi, 2004). Protein sequences, contact potentials and evolutionary information have been well used to score various connection patterns. The present approach encodes separations among cysteine residues into the form

*To whom correspondence should be addressed.

Table 1. Number of chains in the datasets, divided according to the number of disulfide bridges (B)

Datasets	$B = 2$	$B = 3$	$B = 4$	$B = 5$	$B = 2 \sim 5$
SP39	150	149	105	45	449
SP39-ID30	92	81	43	28	244
SP39-TEMPLATE	244	198	98	45	585
SP43	124	118	41	35	318

of vectors. The prediction of disulfide connectivity is based on the comparisons of vectors from testing and template dataset, in which similar vectors imply similar connection patterns. The method proposed here is much simpler than graph-based methods, and raises both efficiency and accuracy.

2 SYSTEM AND METHODS

2.1 Datasets

The datasets used to evaluate the predicting power of CSPs were constructed from SwissProt release No. 39 (Bairoch and Apweiler, 2000), including sequences with annotated disulfide bridges. Protein sequences in SwissProt release No. 39 are filtered according to procedures described in two previous works (Fariselli and Casadio, 2001; Vullo and Frasconi, 2004). This dataset is denoted as 'SP39'. Another dataset based on SP39 was also constructed; redundant sequences with pairwise sequence identity of more than 30% were removed. This non-redundant set is denoted as 'SP39-ID30'. SP39-ID30 is used to investigate the effects of sequence identities on the prediction accuracy of CSP.

Another dataset was further constructed to verify the predicting power of CSP. The same filter procedures were applied to sequences in SwissProt release No. 43, where sequences in release 39 were excluded. Thus it is possible to predict proteins newly added to SwissProt database between releases No. 39 and No. 43. This set is denoted as 'SP43'. Redundant sequences with pairwise sequence identity of more than 25% in SP43 were also removed. The template set used to predict disulfide connectivity in SP43 was constructed from SwissProt release 39. Sequences in this set were filtered as in SP39 and SP43, except for the PDB filter. Only sequences sharing less than 30% identity with those in SP43 were kept. This template set is denoted as 'SP39-TEMPLATE'.

The numbers of sequences divided according to the number of disulfide bridges in these datasets are summarized in Table 1.

2.2 Basic assumption

Similar disulfide bonding patterns infer similar protein structures regardless of sequence identity (Chuang *et al*., 2003). Figure 1 shows an example of two proteins with the same disulfide bonding patterns. Tick anticoagulant peptide (serine protease inhibitor, PDB id 1TAP) (Antuch *et al*., 1994) and cacicludine (calcium channel blocker, PDB id 1BF0) (Gilquin *et al*., 1999) exhibit the same disulfide connectivity [1–6, 2–3, 4–5], which means that the first oxidized cysteine is connected with the sixth one, the second with the third, and the forth with the fifth. These two proteins share sequence identity of only 18.2%, but with a C_α root-mean-square deviation (RMSD) of 3.6 Å (Chuang *et al*., 2003). Although the sequence identity is below the twilight zone, the structure and separations among cysteine residues are similar for these two proteins. The residue numbers for cysteines in the two proteins are [5, 15, 33, 39, 55, 59] and [7, 16, 32, 40, 53, 57], respectively. The positions and separations of cysteine residues are similar for these two proteins. It is likely that cysteine separations are related to disulfide connectivity patterns, and through the comparison of CSPs, the disulfide connectivity patterns may be inferred and predicted.

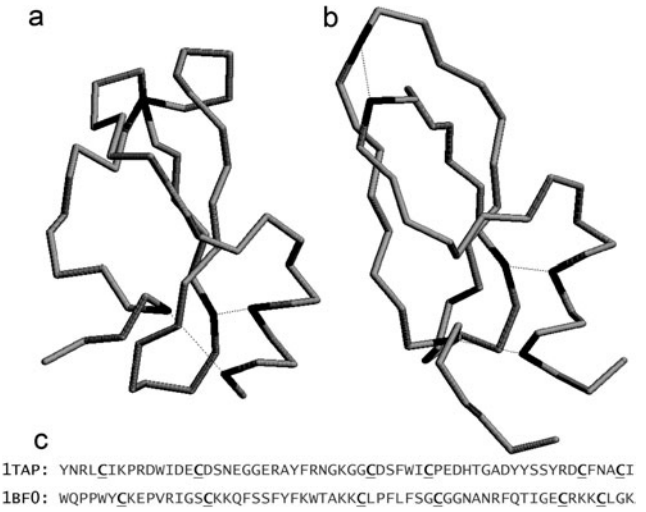

Fig. 1. The structures of two proteins with low sequence identity but sharing the same disulfide bonding patterns: (**a**) anticoagulant protein (PDB id 1TAP), (**b**) calcium channel blocker (PDB id 1BF0), and (**c**) the sequences of the two proteins, with cysteine residues highlighted with bold and underline. Both proteins have three disulfide bonds [1–6, 2–3, 4–5] and BPTI-like structures; the sequence identity is 18.2%.

2.3 CSP and evaluation of prediction accuracy

CSPs contain cysteine separation information. Protein x with n disulfide bonds and $2n$ cysteine residues has a cysteine separation profile (CSP^x) defined as

$$\begin{aligned} CSP^x &= (s_1, s_2, \ldots, s_{2n-1}) \\ &= (C_2 - C_1, C_3 - C_2, \ldots, C_{2n} - C_{2n-1}) \end{aligned}$$

where C_i is the position of ith cysteine residue in the given protein and s_i is the separation between cysteines C_i and C_{i+1}. By this definition, a protein with disulfide bonds will have a CSP.

The divergence, D, between two CSPs is defined as follows:

$$D = \sum_i |s_i^X - s_i^Y|$$

where s_i^X and s_i^Y are the ith separations for CSPs of two different proteins X and Y.

The CSP of a test protein was then compared with all CSPs of template proteins. The disulfide connectivity pattern of the test protein can be predicted as that of the template protein with the most similar CSP, i.e. with the smallest divergence value D. If the divergence D between two CSPs equals 0, the CSPs are termed 'matched profiles', otherwise they are 'mismatched profiles'. If more than one template proteins are matched, one of the templates is randomly selected for the prediction. The ambiguous situations are rare; only less than 2% are observed.

Our method is basically a nearest-neighbor (NN) approach. With only one template for each pattern, our method is essentially a 1-NN approach. We have tried k-NN method in our preliminary investigation. However, the prediction accuracy of k-NN is not significantly better than that of our current approach.

The prediction accuracy of our method was evaluated with Q_p and Q_c values, which are the fraction of proteins with correct disulfide connectivity prediction and are defined as:

$$Q_\mathrm{p} = \frac{C_\mathrm{p}}{T_\mathrm{p},} \qquad Q_\mathrm{c} = \frac{C_\mathrm{c}}{T_\mathrm{c}}$$

where C_p is number of proteins with all the disulfide connectivity correctly predicted; T_p is the total number of test proteins; C_c is the number of disulfide

Table 2. Comparison among different disulfide connectivity prediction algorithms

Algorithms	$B = 2$		$B = 3$		$B = 4$		$B = 5$		$B = 2 \sim 5$	
	Q_p (%)	Q_c (%)	Q_p (%)	Q_c (%)	Q_p (%)	Q_c (%)	Q_p (%)	Q_c (%)	Q_p (%)	Q_c (%)
Frequency[a]	58	58	29	37	1	10	0	23	29	32
MC graph-matching[b]	56	56	21	36	17	37	2	21	29	38
NN graph-matching[c]	68	68	22	37	20	37	2	26	34	42
BiRnn-1 sequence[d]	59	59	17	30	10	22	4	18	28	32
BiRnn-1 profile[d]	65	65	46	56	24	32	8	27	42	46
BiRnn-2 sequence[d]	59	59	22	34	18	30	8	24	31	37
BiRnn-2 profile[d]	73	73	41	51	24	37	13	30	44	49
CSP (SP39)[e]	89	89	81	84	81	85	51	60	81	81
CSP (SP39-ID30)[f]	74	74	44	53	26	44	18	31	49	52
CSP (SP43)[g]	71	71	49	58	30	40	28	33	53	53

[a]Prediction accuracy reported by Vullo and Frasconi (2004).
[b]Prediction accuracy reported by Fariselli and Casadio (2001).
[c]Prediction accuracy reported by Fariselli *et al*. (2002).
[d]Prediction accuracy reported by Vullo and Frasconi (2004).
[e]Prediction accuracy of CSP on SP39 with redundant sequences retained.
[f]Prediction accuracy of CSP with redundant sequences removed.
[g]Prediction accuracy of CSP on SP43 using SP39 as template set, with sequence identity less than 30%.

connectivity correctly predicted; and T_c is the total number of disulfide connectivity in test proteins.

3 RESULTS

3.1 Fourfold cross validation

In order to compare with other approaches for disulfide connectivity prediction, similar criteria were used to select our dataset. The same fourfold cross-validation has been applied to our datasets. The SP39 and SP39-ID30 datasets were divided into four subsets, and the disulfide connectivity prediction was repeated four times. For each prediction, one of the four subsets was used as the test set and the other three subsets were put together to form a template set. The final prediction accuracy was averaged over the four prediction results.

Table 2 summarizes the disulfide connectivity prediction results obtained from this study as well as those obtained from the previous works (Fariselli and Casadio, 2001; Vullo and Frasconi, 2004). 'Frequency' is a trivial method, where the prediction is based on most frequently observed pattern in the training set. 'MC graph-matching' and 'NN graph-matching' are both based on a graph representation of disulfide bonding patterns, using Monte Carlo and Neural Networks for pattern recognition, respectively (Fariselli and Casadio, 2001). The results termed BiRnn are obtained from recursive neural networks with sequence and evolutionary information (Vullo and Frasconi, 2004); the disulfide connectivity patterns are also represented using graphs. The prediction results from this work are termed CSP, with dataset noted in the parenthesis. The prediction results are divided according to the number of disulfide bridges.

The average value of Q_p using CSP is 0.81 for SP39. However, redundant sequences were observed in the SP39 dataset. There are 37.4% of matched profiles and 62.6% of mismatched profiles patterns. The number of matched profile patterns is high, and is likely to have resulted from redundant and homologous sequences in the SP39 dataset. The redundancy may have caused over-fitting in SP39, even with fourfold cross-validation. In order to control and test over-fitting, we extracted the sequences with pairwise sequence identities less than 30% from SP39 and then generated another dataset, SP39-ID30. The average value of Q_p ($B = 2 \sim 5$) using CSP is 49% for SP39-ID30. With redundant sequences removed, the fourfold cross-validation prediction accuracy of CSPs is higher than the best results ever reported from previous works.

The prediction accuracies for protein chains with different disulfide bridge numbers are all significantly higher for 'CSP (SP39)'. For proteins with two, four and five disulfide bridges, the prediction accuracies in 'CSP (SP39-ID30)' are higher than other works. The prediction accuracy for proteins with three disulfide bridges is 2% lower than that of 'BiRnn-1 profile', but is still significantly higher than those from other works.

3.2 Handout prediction of new sequences from SP43

We further validate CSP on a new dataset, SP43, which contains new sequences not seen in SwissProt release 39. We use SP39-TEMPLATE as the template set to predict disulfide connectivity patterns of new sequences in SP43. The pairwise identities of sequences in the template set and SP43 are less than 30%, with template sequences sharing higher identities with those in SP43 being removed. The overall prediction accuracy in SP43 dataset is 53%, which shows significant improvement over the prediction on the other dataset, SP39. The prediction results for SP43 are listed in Table 2. For proteins with three, four and five disulfide bridges, the prediction accuracies in the SP43 dataset are higher than those obtained with fourfold cross-validation in SP39-ID30 dataset. This implies that increasing even the number of non-redundant templates may improve the prediction accuracy of CSP.

3.3 Examples

Three examples of CSP matching are listed in Table 3. These examples are taken from the SSDB database (Chuang *et al*., 2003). The CSPs for template and query protein sequences, as well as their divergence score D, disulfide connectivity patterns and sequence

Table 3. Examples of CSP

Template PDB id	Template CSP	Query PDB id	Query CSP	Disulfide connectivity pattern	Divergence (D)	Sequence identity (%)
1TAP	(10, 18, 6, 16, 4)	1BF0	(9, 16, 8, 13, 4)	[1–6, 2–3, 4–5]	8	18.2
1GPS	(11, 6, 4, 10, 7, 2, 4)	1BRZ	(12, 6, 4, 11, 10, 2, 3)	[1–8, 2–5, 3–4, 6–7]	6	18.8
1TN3	(10, 17, 75, 16, 8)	1C3A:A	(11, 17, 72, 17, 8)	[1–2, 3–6, 4–5]	6	17.7

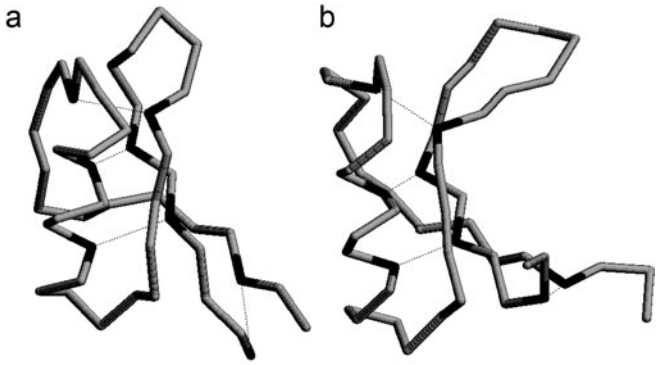

Fig. 2. The structures of (**a**) thionin, toxic arthropod protein (PDB id 1GPS), (**b**) brazzein, thermostable sweet-tasting protein (PDB id 1brz), and (**c**) their sequences with cysteine highlighted. The divergence score D between these two protein sequences is 6. Both proteins have disulfide connectivity [1–8, 2–5, 3–4, 6–7] and their sequence identity is 18.8%.

identities, are shown in Table 3. In the three examples, the divergence scores are all smaller than 10, implying that they share similar disulfide positioning and connectivity patterns. The sequence identities in the three examples are all lower than 20%, thus structure similarity from sequence homology can be ruled out.

The structures and sequences of these examples are illustrated in Figures 1–3. The first example is shown in Figure 1. Tick anticoagulant peptide (serine protease inhibitor, PDB id 1TAP) (Antuch *et al*., 1994) and cacicludine (calcium channel blocker, PDB id 1BF0) (Gilquin *et al*., 1999) have a divergence score $D = 8$; their disulfide connectivity pattern is [1–6, 2–3, 4–5]. Example 2 is illustrated in Figure 2. Thionin (toxic arthropod protein, PDB id 1GPS) (Bruix *et al*., 1993) and brazzein (thermostable sweet-tasting protein, PDB id 1brz) (Caldwell *et al*., 1998) share 18.8% sequence identity. Their divergence score D is 6, and the disulfide connectivity pattern is [1–8, 2–5, 3–4, 6–7]. The third example (Fig. 3), C-type lectin carbohydrate recognition domain of human tetranectin (PDB id 1TN3) (Kastrup *et al*., 1998) and flavocetin-A from Habu snake venom (PDB id 1C3A:A) (Fukuda *et al*., 2000) also have a divergence score of $D = 6$. Their sequence identity is 17.7% and the connectivity pattern is [1–2, 3–6, 4–5]. For all proteins, the oxidized cysteine residues are indicated in black. Cysteine residues on sequences are highlighted in bold and underline. In each case, the cysteine residues are positioned in similar sites along the sequence, and the separations among these cysteine residues are nearly identical.

4 DISCUSSIONS

The number of possible disulfide connectivity patterns increases rapidly with the number of disulfide bridges. For a protein with n disulfide bridges ($n * 2$ oxidized cysteines), the number of possible disulfide connectivity patterns N_p can be formulated as follows:

$$N_p = \frac{\binom{2n}{2}\binom{2n-2}{2}\binom{2n-4}{2}\cdots\binom{2}{2}}{n!} = (2n-1)!! = \prod_{i\leq n}(2i-1).$$

Table 4 lists the number of possible disulfide connectivity patterns for proteins with different disulfide bridge numbers. The use of CSPs may be obscure at first, since the rapidly increasing number of patterns cannot be covered exhaustively. However, the observed numbers of patterns in PDB peak at five disulfide bridges, and decline afterward. Only 45 patterns are observed for protein chains with five disulfide bridges, as opposed to the possible 945 patterns expected. These results imply that the disulfide connectivity pattern of a protein sequence can be predicted from a limited set of templates.

One limitation of our approach is that a pattern not presented in the training set cannot be predicted correctly. Other machine-learning approaches have to enumerate all possible patterns to obtain a prediction with the maximum score (Vullo and Frasconi, 2004); therefore it is possible to correctly predict a pattern never seen in the training set. However, evaluation of all possible patterns is expensive (Vullo and Frasconi, 2004); our approach can achieve comparable prediction performance in a much simpler and faster algorithm.

The prediction accuracies for protein chains with different divergence coverage are shown in Figure 4. The divergence coverage means that a profile matches with a divergence score smaller than or equal to that specified. For example, divergence coverage 5 means profiles matched with a divergence score $\leq$5. Prediction results of the three datasets are illustrated in Figure 4. As can be seen, when divergence coverage is 0, which means the profiles are 'matched profiles', the prediction accuracy is 100% for all datasets. The prediction accuracies become lower as divergence coverage increases. For divergence coverage 50, the prediction accuracy is slightly higher than the overall accuracy. Thus divergence coverage can be used as an index for adoption of CSP or other machine-learning approaches to predict disulfide connectivity. However, these divergence scores are not normalized according to the number of disulfide bridges and the lengths of protein sequences. Several complex factors should be considered in the normalization of divergence score; this is one of the objectives currently undertaken in our group. Sequences with low divergence coverage in a dataset (e.g. 5 for Q_p 0.8) can be predicted by CSP proposed in this work with high accuracy; otherwise, the

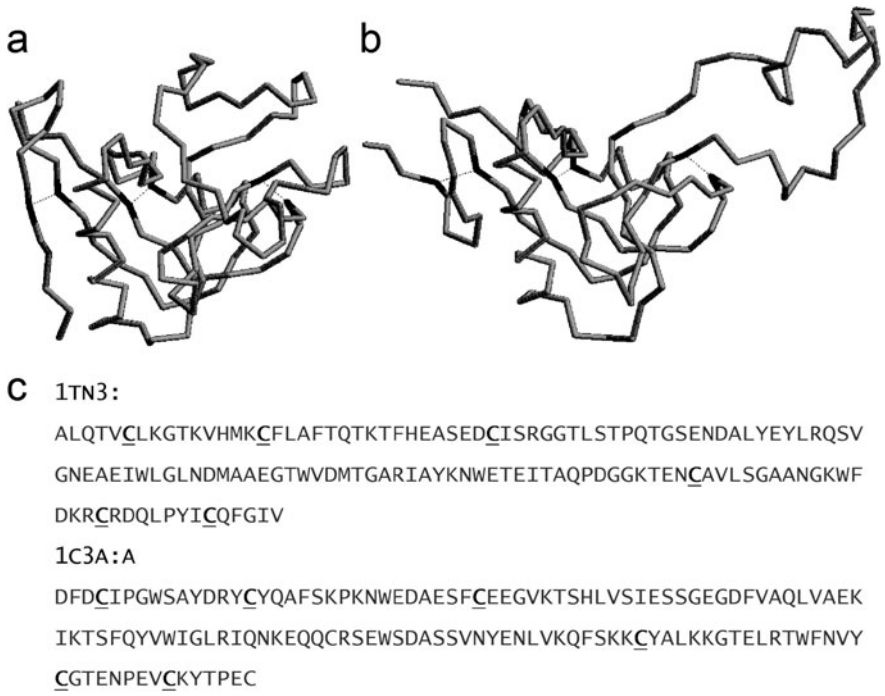

Fig. 3. Structures of (**a**) C-type lectin carbohydrate recognition domain of human tetranectin (PDB id 1TN3), (**b**) flavocetin-A from Habu snake venom (PDB id 1C3A:A), and (**c**) their sequences with cysteine residues highlighted. The divergence score D between these two protein sequences is 6. Both proteins have disulfide connectivity [1–2, 3–6, 4–5] and their sequence identity is 17.7%.

Table 4. Number of possible disulfide connectivity patterns (N_p) for protein chains with different disulfide bridge numbers

Number of disulfide bridges (B)	N_p^a	Observed N_p^b
$B = 2$	3	3
$B = 3$	15	15
$B = 4$	105	43
$B = 5$	945	45
$B = 6$	10 395	29
$B = 7$	135 135	14

[a]Number of possible disulfide connectivity patterns.
[b]Observed number of disulfide connectivity patterns. Statistics obtained from the SSDB database (http://www.e106.life.nctu.edu.tw/∼ssbond/) (Chuang *et al*., 2003).

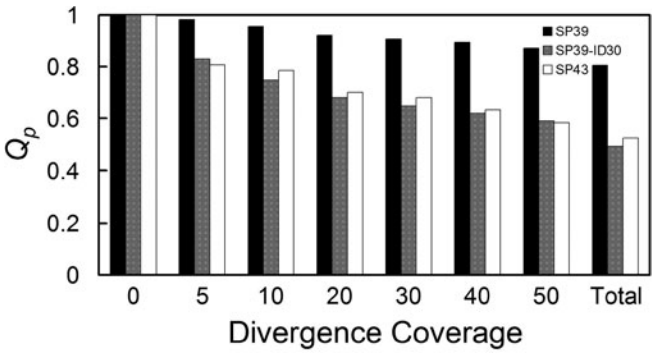

Fig. 4. Prediction accuracy of the datasets with various divergence coverage, which means that a profile matches with divergence score smaller than or equal to that specified. The prediction accuracy (Q_p) is higher with lower divergence coverage. With $D \leq 50$ the prediction accuracy is still slightly higher than the overall Q_p. See text for details.

connectivity patterns of the other sequences in the same dataset can be elucidated by neural networks (Vullo and Frasconi, 2004), support vector machines or other machine-learning approaches.

5 CONCLUSIONS

In this work, we have shown that cysteine separation profiles (CSPs) can be used in predicting disulfide connectivity patterns based on the hypothesis that proteins with similar cysteine separations in sequences may have similar disulfide bonding patterns. The prediction accuracy of CSP proposed in this study is higher than those obtained by other approaches. The handout prediction of new sequences in SP43 dataset can reach 53%. The method mentioned here is extremely simple; therefore the computation time is minimum compared to other methods. The rationale behind our method is completely different from previous studies using sequence and evolutionary information. Our method suggests that topology itself may be an important factor in disulfide connectivity, as it has been proposed by theoretical study (Abkevich and Shakhnovich, 2000) and observations in structure databases (Chuang *et al*., 2003). Although many efforts have been made to predict the disulfide connectivity patterns, current prediction accuracy is limited around 50%. However, by combining CSP and other algorithms proposed previously

(Fariselli and Casadio, 2001; Vullo and Frasconi, 2004), it is possible to further improve the prediction accuracy. The use of predicted disulfide connectivity patterns in *ab initio* protein structure prediction and other applications would become more reliable in the foreseeable future.

ACKNOWLEDGEMENTS

The authors thank National Science Council of Taiwan for financial support (project number NSC-93-3112-B-002-022).

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.

Antuch,W., Guntert,P., Billeter,M., Hawthorne,T., Grossenbacher,H. and Wuthrich,K. (1994) NMR solution structure of the recombinant tick anticoagulant protein (rtap), a factor Xa inhibitor from the tick *Ornithodoros moubata*. *FEBS Lett.*, **352**, 251–257.

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

Bruix,M., Jimenez,M.A., Santoro,J., Gonzalez,C., Colilla,F.J., Mendez,E. and Rico,M. (1993) Solution structure of gamma 1-H and gamma 1-P thionins from barley and wheat endosperm determined by 1H-NMR: A structural motif common to toxic arthropod proteins. *Biochemistry*, **32**, 715–724.

Caldwell,J.E., Abildgaard,F., Dzakula,Z., Ming,D., Hellekant,G. and Markley,J.L. (1998) Solution structure of the thermostable sweet-tasting protein brazzein. *Nat. Struct. Biol.*, **5**, 427–431.

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.

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.

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

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.

Fariselli,P., Martelli,P.L., & Casadio,R. (2002). A neural network base method for prediction the disulfide connectivity in proteins. In Damiani,E. *et al*., eds. *Knowledge based Intelligent Information Engineering Systems and Allied Technologies KES 2002*, vol. 1. IOS Press, pp. 464–468.

Fiser,A. and Simon,I. (2000) Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics*, **16**, 251–256.

Fukuda,K., Mizuno,H., Atoda,H. and Morita,T. (2000) Crystal structure of flavocetin-a, a platelet glycoprotein Ib-binding protein, reveals a novel cyclic tetramer of c-type lectin-like heterodimers. *Biochemistry*, **39**, 1915–1923.

Gilquin,B., Lecoq,A., Desne,F., Guenneugues,M., Zinn-Justin,S. and Menez,A. (1999) Conformational and functional variability supported by the BPTI fold: Solution structure of the Ca2+ channel blocker calcicludine. *Proteins*, **34**, 520–532.

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.

Kastrup,J.S., Nielsen,B.B., Rasmussen,H., Holtet,T.L., Graversen,J.H., Etzerodt,M., Thogersen,H.C. and Larsen,I.K. (1998) Structure of the c-type lectin carbohydrate recognition domain of human tetranectin. *Acta Crystallogr. D Biol. Crystallogr.*, **54**, 757–766.

Martelli,P.L., Fariselli,P., Malaguti,L. and Casadio,R. (2002) Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy. *Protein Sci.*, **11**, 2735–2739.

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.

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.

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

Wedemeyer,W.J., Welker,E., Narayan,M. and Scheraga,H.A. (2000) Disulfide bonds and protein folding. *Biochemistry*, **39**, 4207–4216.

Welker,E., Wedemeyer,W.J., Narayan,M. and Scheraga,H.A. (2001) Coupling of conformational folding and disulfide-bond reactions in oxidative folding of proteins. *Biochemistry*, **40**, 9059–9064.