Cysteine separations profiles on protein secondary structure infer disulfide connectivity

Guantao Chen, Hai Deng, Yufeng Gui, Yi Pan and Xue Wang

***Abstract* — Disulfide connectivity prediction from one chain of protein helps determine protein tertiary structure. The more accuracy of prediction it reaches the more precise three dimensional structures we can obtain through computational methods. Previous methods only use local sequence or secondary structure information or global sequence information or combination of the above descriptors to predict the disulfide bond pattern. Instead of using those descriptors, we take an alternative descriptor of global secondary structure to make prediction, and the highest performance among all pattern-wise methods is obtained. Cysteine separation profiles on protein secondary structure have been used to predict the disulfide connectivity of proteins. The cysteine separation profiles on secondary structure(CSPSS) represent a vector encoded from the sepeartions between any two consecutive cysteine-corresponding positions in a predicted protein secondary structure sequence. Through comparisons of their CSPSS, the disulfide connectivity of a test protein is inferred from a template set. In 4-fold of SP39, any two proteins from different groups share less than 30% sequence identity. The result shows a prediction accuracy (54%), which proves again a disulfide bond pattern is highly related to protein secondary structure.**

***Index Terms*—Cysteine separation profiles, disulfide connectivity, nearest neighbour, secondary structure.**

I. INTRODUCTION

Disulfide bonds are covalent bonds between two non-adjacent cysteine residues in proteins. They determine the major folding of proteins [1]. The more correct prediction on disulfide bonds will lead the more accurate prediction on protein structure. Furthermore, the knowledge of disulfide bonds is very helpful for revealing structure/function relationship. Therefore, a higher performance on the prediction of disulfide connectivity will bring more benefits on the comprehensive study of the proteins with disulfide bonds.

Tsai et al. [6] have categorized all methods into two groups: (1) pattern-wise [6, 7] or (2) pair-wise [2-5]. The prediction accuracies of above methods are around 50% and only the newest method using SVM(Tsai et al., 2005) improve the results to 63%, which shows that this task remains challenging From another angle, the descriptor for predicting performance is a determinant factor.

Fariselli and Casadio [3] calculated the bond probability of two cysteines from local contact potential profile. Zhao et al [7] used global sequence separations as the representation of a protein. Tsai et al [6] also used global sequence separations as one of their descriptors for predicting the bond probabilities of possible pairs of cysteines. Secondary structure has been employed as the descriptor for input coding by Baldi et al. [2] and Ferre and Clote [5]. Instead of using a local window of secondary structures, we extended the window size as the separations of two consecutive cysteines in one protein chain, thus we obtained a global descriptor of secondary structure, cysteine separation profile of secondary structure (CSPSS) similar to the CSP method [6].

Our method derived from the CSP method [6] but created a new descriptor, cysteines separation profiles on secondary structure (CSPSS). It has been demonstrated that proteins with similar fold possibly share similar secondary structure and similar disulfide bond also share similar fold. Therefore, a pattern-wise method with the same schema of the CSP method but using CSPSS was developed. The method encodes the separations among consecutive cysteins of proteins on secondary structure sequence as vectors of four variables including the segment length, the number of coils, the number of sheets, and the number of helixes. Each vector can be viewed as a point in a four-dimensional space and thus each CSPSS can be viewed as a series of points in this space. The prediction is based on the comparisons of CSPSSs from testing and template dataset. The two proteins with the smallest divergence after the comparisons presumably has the same disulfide connectivity pattern. We applied the method on SWISS-PROT 39 (SP39), and 2% improvements on prediction accuracy were obtained than the CSP method, and also found that our correctly predicted patterns included ones from CSP. The results reveal that CSPSS is a better descriptor than CSP for predicting disulfide connectivity.

This work is supported in part by the NSF grant DMS-0500951 and the NSA H98230-04-1-0300 grant to GC, and a grant from the NIH P20 Award to GC (grant # P20 GM065762-01A1).

Guantao Chen is with the Statistics and Mathematics Department, Georgia State University, Atlanta, GA 30303 USA (e-mail: gchen@gsu.edu).

Hai Deng, is with the Department of Computer Science, Georgia State University, Atlanta, GA 30303 USA (corresponding author to provide phone: 404-651-0671; e-mail: hdeng1@student.gsu.edu).

Yufeng Gui is with the Department of Statistics, Wuhan University of Technology, Wuhan, Hubei, China. (guiyufeng@hotmail.com)

Yi Pan is with the Department of Computer Science, Georgia State University, Atlanta, GA 30303 USA (e-mail: pan@cs.gsu.edu).

Xue Wang, is with the Department of Computer Science, Georgia State University, Atlanta, GA 30303 USA (e-mail: xwang14@student.gsu.edu).

1-4244-0134-8/06/$20.00 © 2006 IEEE

II. Methodology

A. Basic assumption

Similar protein secondary structure patterns possibly imply similar protein tertiary structure patterns and lead to similar disulfide connectivity pattern because a disulfide bonding pattern is one aspect of protein tertiary structure. For example, the structures of two proteins (PDB id 1TAP and PDB id 1BF0) exhibit the same disulfide connectivity pattern (1-6, 2-3, 4-5) (the number is the index which indicates the occurrence order of cysteines in one protein chain) but share only 18.2% sequence identity.

CSPSS

CSPSS contains cysteine separation information about protein secondary structure. Protein *x* with *n* disulfide bonds and 2*n* cysteine residues has a cysteine separation profile (*CSPSS*^{x}) defined as

$$\mathrm{CSPSS}^X = (S_1, S_2, \dots, S_{2n\text{-}1}) \tag{1}$$

where S_i is the vector of secondary structure sequence between the ith cysteine and the (i+1)th cysteine. S_i can be represented as a segment of secondary structure sequence annotated by PSIPRED server (http://bioinf.cs.ucl.ac.uk/psipred/psiform.html). Therefore, every vector S_i is a character string each position of which is one of three possible symbols(C, H, E) representing secondary structure. To make it numeric, we extract some digital information from each vector. Therefore, S_i is transformed as

$$S_i = (l_i, C_i, E_{i,,} H_i) \tag{2}$$

A. where l_i is the length of the vector, C_i is the number of C, H_i is the number of H and E_i is the number of E.

For example, S_i can be represented as a segment of secondary structure (ss) sequence annotated by PSIPRED server, i.e., a protein chain(AMCI_APIME) with 56 amino acids (aa) and 10 cysteines with the pattern (1-7, 2-5, 3-6, 4-10, 8-9),

aa:
EECGPNEVFNTCGSACAPTCAQPKTRICTMQCRIGCQ**C**
QEGFLRNGEGACVLPENC

ss:
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC**CC**
CCCCCEECCCCCCCCCCCCC

S_1: [CCCCCCCCCC], …, S_8: [**CCCCCCCEECCCCCC**], S_9: [CCCCCCC].

After the encoding the segments of secondary structure according to the equation (2), S_1, S_8, S_9 are numberic values as following:

S_1: [CCCCCCCCCC] → [10, 10, 0, 0]

S_8: [CCCCCCCEECCCCCC] → [15, 13, 2, 0]

S_9: [CCCCCCC] → [7, 7, 0, 0]

The divergence, *D*, between two CSPSS is defined as follows

$$D = \sum_i dist(S_i^X - S_i^Y) = \sum_i \left(\sqrt{(l_i^X - l_i^Y)^2 + (C_i^X - C_i^Y)^2 + (E_i^X - E_i^Y)^2 + (H_i^X - H_i^Y)^2}\right) \tag{3}$$

where S_i^X and S_i^Y are the *i*th separations for CSPSS of two different proteins *X* and *Y*.

The CSPSS of a test protein was then compared with all CSPSS of template proteins. The disulfide connectivity pattern of the test protein can be predicted as that of the template protein with the most similar CSPSS, i.e. with the minimum divergence value *D*. If more than one minimum value meets for one test protein, one of the template patterns will be chosen. In fact, this very rare situation did not occur in the experiments.

III. Measurement

The prediction accuracy of our method was also evaluated with *Q*p, which is the fraction of proteins with correct disulfide connectivity pattern prediction and is defined as:

$$Q_p = \frac{C_p}{T_p} \tag{4}$$

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

IV. Results

A. Dataset

In order to compare our method to the methods published in 2005, the dataset from SWISS-PROT named SP39 was adopted for method validation. To avoid the influence of sequence homology, the dataset was divided into four groups to guarantee that each two proteins from different groups have a sequence identity less than 30%.

The numbers of sequences according to the bridges are displayed in Table 1.

Table 1. Number of chains according to the the number of disulfide bridges (B)

Dataset	B=2	B=3	B=4	B=5	B=2...5
SP39	156	146	99	45	446

B. Cross-validation of SP39

In order to compare with other methods for disulfide connectivity prediction, same criteria were applied on selecting our dataset. Also the same fourfold cross-validation has been applied on our dataset. Even the selection of four subsets is same as the method from Baldi et al. [2]. The SP39 were divided into four subsets each of which has four balanced groups according to the bridges. It is worth pointing out that any test protein pattern can only be predicted from template proteins with same bridges from our method, which can be inferred from the methodology section.

TABLE 2. COMPARISON AMONG DIFFERENT DISULFIDE CONNECTIVITY PREDICTION ALGORITHMS

	B=2	B=3	B=4	B=5	B=2…5
Methods	Qp(%)	Qp(%)	Qp(%)	Qp(%)	Qp(%)
2D-RNN [a]	74	51	27	11	49
DiANNA[b]	62	40	55	26	49
CSP[c]	72	54	33	18	52
CSPSS	72	58	37	18	***54***
SVM[d]	79	53	55	71	63

[a]Reported by Baldi in 2005 [2]
[b]Reported by Ferre and Clote in 2005 [5]
[c]Reported by Zhao et al. in 2005 [8]
[d]Reported by Tsai et al. in 2005 [6]

Table 2 lists the accuracies of four-fold cross-validation performed with the dataset SP39 for our method along with some results reported previously. Here we only list the results published in 2005 because the results before 2005 show the accuracy is up to 46%. Baldi et al. used 2-Dimensional Recursive Neural Network (2D-RNN, [2]) to predict disulphide connectivity in proteins starting from their primary sequence and its homologues. The outputs of 2D-RNN are the pair-wise probabilities of the existence of a bridge between any pair of cysteines. Finally, the weighted matching algorithm is applied on the graph with all edges/possibilities between any two vertices/cysteins. A *diresidue* Neural Network (DiANNA) [5] is trained to recognize pairs of bonded half-cystines given input of half-cystines symmetric flanking regions. The network is trained using disulfide bonds information derived from high-quality protein structures. the data are encoded with respect to cysteine pairs. Zhao et al. [8] simply adopted the descriptor of cysteine separation profile and used the nearest neighboring method. For the SVM model[6], the features encoded are the information extracted from profile and distances between oxidized cysteines (DOC). After the data are encoded, the SVM model is used to predict bonding probabilities for each cysteine pair. Finally, the problem is transformed into a maximum weight matching problem and solved to find the final bonding pattern for a protein.

Using CSPSS, the nearest neighboring method obtained a *Q*p of 54%, which is better than those obtained in previous methods except the SVM model. The reason for the improvement is the consideration of global secondary structure pattern. We also found that our results for correct predictions contain ones from the CSP method, which validated that our descriptor includes the information of the CSP descriptor.

V. DISCUSSION AND CONCLUSION

There are two major categories for the descriptors of disulfide connectivity prediction:

(1) The global descriptors [6] such as sequence length, the positions of all cysteines and (2) local descriptors [2, 5, 7, 8] such as secondary structure and residue contact potential.

We designed a new descriptor CSPSS which should be classified into the global descriptor, gave an example of how to use this descriptor with a pattern-wise method mentioned before and obtained a second-to-top performance. The SVM model [6] for predicting disulfide connectivity benefits from the combination of the global descriptor (DOC) and local descriptor (sequence profile) and is the state-of-art method. Therefore we will work on the combination of CSPSS with local descriptors using different methods such as the pair-wise method SVM to improve the results.

ACKNOWLEDGMENT

We thank Jianlin Cheng for providing datasets and Guiyu Feng for helpful discussion. Hai Deng thanks Drs. Charles Derby, Yanqing Zhang and Raj Sunderrraman's help.

REFERENCES

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

[2] 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 NeuralInformation Processing Systems 17*. MIT Press, Cambridge, MA, pp. 97-104.

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

[4] Fariselli, P., Riccobelli, P., and Casadio, R. (2002) A neural network based method for predicting the disulfide connectivity in proteins. In Damiani E et al. (eds), *Knowledge based intelligent information engineering systems and allied technologies (KES 2002).*IOS Press, Amsterdam, **1**, pp. 464-468.

[5] Ferrѐ, F. and Clote, P. (2005) Disulfide connectivity prediction using secondary structure information and diresidue frequencies. *Bioinformatics*, **21**, 2336-2346.

[6] Tsai, C., Chen, B-J, Chan, C-H, Tsai, Liu, H-L, and Kao, C-Y (2005) Improving disulfide connectivity prediction with sequential distance between oxidized cysteins. *Bioinformatics,* **21**, 4416 - 4419.

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

[8] Zhao, E., Liu, H-L, Tsai, C-H, Tsai, H-K, Chen, C-H, and Kao, C-Y (2005) Cysteine separations profiles on protein sequences infer disulfide connectivity. *Bioinformatics*,**21**, 1415-1420.