Disulfide Bond Prediction using Neural Network and Secondary Structure Information

Ouyan Shi
Faculty of Basic Medicine
Tianjin Medical University
Tianjin, China

Huiyun Yang
Department of Biomedical Engineering
Tianjin Medical University
Tianjin, China

Chunquan Cai
Department of Neurosurgery
General Hospital of Tianjin Medical University
Tianjin, China

Jing Yang
Faculty of Basic Medicine
Tianjin Medical University
Tianjin, China

Xin Tian
Department of Biomedical Engineering
Tianjin Medical University
Tianjin, China
Corresponding Author: tianx@tijmu.edu.cn

Abstract—**In protein-folding prediction, the location of disulfide bonds can strongly reduce the search in the conformational space. Therefore the correct prediction of the disulfide connectivity starting from the protein residue sequence may also help in predicting its 3D structure. In this paper, we describe a method to predict disulfide connectivity in a protein given only the amino acid sequence, using neural network, and given input of symmetric flanking regions of N-terminus and C-terminus cystines augmented with residue secondary structure (helix, sheet, and coil) as well as evolutionary information. 252 protein sequences were selected from the SWISS-PROT database. From the results of 4-fold cross validation, we find that merging protein secondary structure allows us to obtain significant prediction accuracy improvements.**

Keywords-disulfide bonds; neural network; protein secondary structure

I. Introduction

Correctly predicting the disulfide bond topology in a protein is of crucial importance for the understanding of protein function and can be of great help for tertiary prediction methods.

A disulfide bond is formed by the oxidative linkage of two cysteines through their thiol groups. In proteins some cysteines, called cystines, are oxidized and the others are called free cysteines.

A necessary step to the prediction of disulfide connectivity is the prediction of the disulfide bonding state of cysteine in proteins. This has been tackled before [1, 2] and recently refined [3, 4]. The methods presently available discriminate between free and bonded state of cysteine with a high accuracy (about 80%) starting from the residue chain.

In 1999, Fariselli et al. [1] designed a jury of neural networks, trained on flanking sequence information in neighborhoods of oxidized versus reduced cysteines. Their algorithm obtained an accuracy of 71%; when additionally trained on flanking evolutionary information (i.e. multiple sequence alignments of homologous proteins) the accuracy improved to 81%. Fiser and Simon [2] used multiple sequence alignments in a different manner to obtain an accuracy of 82%. Mucchielli-Giorgi et al. [3] used a combination of perceptrons, trained on sets of proteins homogeneous in terms of their amino acid content, to obtain an accuracy of 84%. In the same year, Martelli et al. [4] used a hybrid hidden Markov model and neural network system, reaching 88% accuracy.

The question then poses as to also whether cysteine pairing in the bond is endowed with characteristic marks that can be captured and used to univocally establish the disulfide connectivity in proteins. In this work we will complement the prediction of disulfide bonding state by focusing on the prediction of the disulfide connectivity, i.e. which, if any, pairs of cystines form a bond in a given protein sequence.

Beginning with the earlier observation that there is a bias in the secondary structure preference of free cysteines and cystines [5], we develop a BP neural network to learn amino acid environments constituting the window contents of a symmetric region centered at partner cystines. The inputs of the neural network are the symmetric flanking residues about both cystines of a potential disulfide bond, along with the secondary structure of the residues and PSI-BLAST-determined evolutionary information (PSSM). Finally, apply Ed

Supported by the National Natural Science Foundation of China (30770545) and the Natural Science Foundation of Tianjin (07JCYBJC17100)

978-1-4244-1748-3/08/$25.00

Rothberg's program (wmatch, http://elib.zib.de/pub/Packages/mathprog/matching/weighted) of the Edmonds–Gabow maximum weight matching algorithm [6] to assign disulfide bond partners, given the weighted complete graph, whose nodes are cystines and whose weights are values output from the neural network.

II. System and Methods

A. The Protein Data Sets

We selected 252 protein sequences from the SWISS-PROT database having at least two and at most five (i.e. 4–10 cysteines in the protein forming disulfide bonds) intra-chain disulfide bonds, and for which structural data were available in the Protein Data Bank (PDB) [7]. These protein sequences were divided into four groups of the same size approximately in order to perform 4-fold cross validation experiments.

B. Performance Measures

Given an even number of cysteines believed to form disulfide bonds, the problem is to determine the correct connectivity pattern among all the possible alternatives.

Throughout the following sections, P and N represent a training file of positive and negative examples, respectively, of sequence length $2w$, e.g. two 15-mers corresponding to the symmetric cystine-centered size w=2n+1=15 window contents of cystines (i.e. the n residues N-terminal and C-terminal to each cystine, where n=7). Let P denote the pairs of window contents for all the cystines involved in an intra-chain bond, and let N denote the corresponding set of possible pairs of cystines that are not intra-chain disulfide bonds. True positive predictions occur when a cystine pair with a known bond is correctly predicted as such, while false negative predictions occur when known disulfide bonds are predicted not to be such. Accordingly, a true negative is a cystine pair correctly predicted to not form a disulfide bond, while a false positive is a pair of cystines that is not a bond though predicted as such. Letting TP, TN, FP and FN denote, respectively, the number of true positives, true negatives, false positives and false negatives, recall the definitions of sensitivity (Sn), specificity (Sp) and Matthew's correlation coefficient (Mcc):

$$Sn = \frac{TP}{TP + FN}\,, \qquad (1)$$

$$Sp = \frac{TN}{TN + FP}\,, \qquad (2)$$

$$Mcc = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}\,. \qquad (3)$$

Finally, two indexes can also be used [8]: Qp and Qc. For a protein p Qp is defined as:

$$Qp = \delta(correct\ pattern, predicted\ pattern)\,, \qquad (4)$$

where δ (x, y) is 1 if and only if the predicted pattern coincides with the correct pattern. Alternatively, Qc is defined as:

$$Qc = \frac{number\ of\ correctly\ predicted\ pairs}{number\ of\ possible\ pairs}\,. \qquad (5)$$

The two indexes are equally suited and complimentary for measuring the accuracy of the prediction: Qp is a measure of the predictive performance on each protein (either 1 or 0) and can be averaged over the number of predicted proteins to give a global measure of the accuracy of the method. Qc quantifies the accuracy of the method based on the number of pairs correctly predicted with respect to the total number of possible pairs.

C. Neural Network Model

The amino acid environment of cystines shows peculiar sequence characteristics that allow the discrimination between cystines and free cysteines using machine learning [1, 2]. Moreover, the secondary structure conformation assumed by the cysteines and their neighboring residues is remarkably different when comparing disulfide-bonded versus free cysteines [5]. Tab. I, П and Ш show the secondary structure conformation frequencies detected in the dataset and computed using DSSP annotations. From the Tab. I, we can see that disulfide bonds in the selected data set predominantly occur in coil structures (47.39%). Tab. П shows the relative frequency of secondary structures flanking the N-terminus and the respective C-terminus cystine in a disulfide bond in symmetric size 15 window. A secondary structure is assigned to each cystine seven N-terminal and seven C-terminal residues using a majority decision (i.e. counting which secondary structure of each group of seven residues is prevalent). Note the remarkable asymmetry of the coil–sheet (8.97%) and sheet–coil (13.93%) frequencies. Considering the secondary structure of pairs of cystines known to form a disulfide bond, some combinations are preferred (Tab. Ш).

TABLE I. Cysteine Secondary Structure Frequencies(%)

Secondary Structure	All residues	Cystines	Free cysteines
Helix	34.38	22.84	36.09
Sheet	27.24	29.77	32.65
Coil	38.38	47.39	31.26

TABLE II. Secondary Structure of Cystines Neighbors

Secondary Structure	Frequency(%)
H-H	11.07
H-E	1.27
H-C	11.45
E-H	1.34
E-E	6.87
E-C	13.93
C-H	9.99
C-E	8.97
C-C	35.11

TABLE III. SECONDARY STRUCTURE OF DISULFIDE BONDS FREQUENCIES (%)

N-terminal secondary structure	C-terminal secondary structure	Percentage expected	Percentage detected
H	H	5.22	6.39
H	E	6.80	9.81
H	C	10.82	9.72
E	H	6.80	3.71
E	E	8.86	9.21
E	C	14.11	8.31
C	H	10.82	9.59
C	E	14.11	19.05
C	C	22.46	24.04

The expected frequencies for pairs of secondary structures, one for each cystine, assuming independence of each cysteine, are computed as the product of corresponding frequencies from Tab. I. The detected frequency is computed using DSSP annotations. For example, in 19.05% of the cases in the dataset the N-terminal cystine is in coil conformation, while the C-terminal is in sheet conformation (this is the value reported in the 'Percentage detected' column). Since the frequency of coil cystine in the dataset is 0.4739, and the frequency of sheet cystine is 0.2977 (as reported in Tab. I), one can expect the frequency of bonds, in which one cystine is a coil and the other a sheet, to be 0.4739 × 0.2977 = 0.1411 (14.11%). This is the 'expected' frequency, which is different from the detected frequency; moreover, the frequency of the bonds in which the N-terminal cystine is in sheet conformation and the C-terminal is in coil conformation is remarkably different (8.31%).

Therefore, we explored the possibility of using sequence and secondary structure information to infer the protein disulfide connectivity.

Standard feed-forward back-propagation network architecture with a single hidden layer was used. A window of 15 amino acid residues (*w*=15) was found to be optimal. In the input encoding, given two-size *w* windows centered at N-terminus and respective C-terminus cystines. For each residue in this window, 20 units were used for the scores in the PSSM (ran PSI-BLAST, against the non-redundant (NR) database, three iterations). To include secondary structure information, we extracted DSSP secondary structure annotations of each of the 2*w* residues, and we added to the evolutionary encoding vectors, $2w \times 3$ additional binary inputs [9] (e.g. H was encoded as 100, E as 010, C as 001).

The resulting inputs to our neural network consist of $2w \times 23$ units. And the output unit is unique. The final disulfide connectivity is obtained by running the wmatch program. To the graph, whose nodes are the putative cystines and whose edges, which join pairs of nodes, are weighted by the output of the neural network. Of several architectures tested, 50 units of hidden layer showed the best performance.

The dataset of positive examples contain all the disulfide bonds annotated in the DSSP files. The negative dataset contain all possible cystine pairs of each sequence that are not disulfide bonds.

To implement the neural network by MATLAB software, the training function is 'traingdx', transfer function are 'tansig' and 'logsig', training epochs is 2000, training goal is 0.001.

III. RESULTS AND DISCUSSION

The measures of only PSSM encoding and combined secondary structure with PSSM encoding are summarized in Tab. Ⅳ and Tab. Ⅴ. We selected 10,25,30,35,40,50,60 hidden units respectively.

TABLE IV. DISULFIDE CONNECTIVY PREDICTION PERFORMANCE OF PSSM ENCODING

Hidden Units	Sn (%)	Sp (%)	Mcc	Qc (%)	Qp (%)
10	57.40	95.22	0.5757	88.68	40.08
25	54.85	94.53	0.5386	87.67	39.29
30	60.08	95.33	0.5996	89.23	44.44
35	58.55	95.31	0.5856	88.95	39.68
40	58.29	95.28	0.5834	88.88	38.89
50	59.44	95.15	0.5896	88.97	43.25
60	58.42	95.04	0.5785	88.70	42.06

We can see that the accuracy is highest when using 30 hidden units.

TABLE V. DISULFIDE CONNECTIVY PREDICTION PERFORMANCE OF PSSM +SECONDARY STRUCTURE

Hidden Units	Sn (%)	Sp (%)	Mcc	Qc (%)	Qp (%)
10	57.40	95.73	0.5885	89.10	39.60
25	56.51	94.77	0.5570	88.15	39.30
30	60.33	95.87	0.6152	89.72	40.68
35	58.42	95.33	0.5858	88.95	42.46
40	59.69	95.73	0.6064	89.50	40.87
50	62.50	95.84	0.6307	90.07	46.03
60	57.40	95.20	0.5757	88.66	38.89

From the above table, we can see that the use of secondary structure information leads to a clear improvement. The architecture of 50 hidden units shows the best performance. The comparison of only PSSM encoding and combined protein secondary structure with PSSM encoding in the same 50 hidden units is shown in Fig. 1.

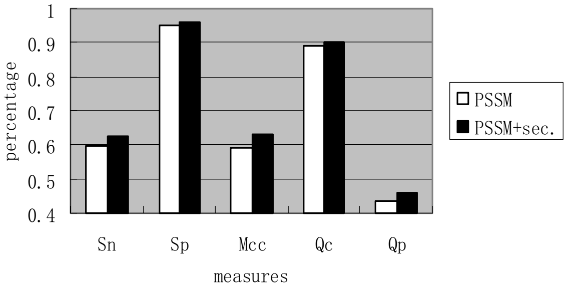

Figure 1. Comparison of different input encoding (50 hidden units)

The inclusion of secondary structure information leads to a marked improvement as well as the inclusion of the 20 frequencies obtained in a multiple sequence alignment for each given residue of the window. This is known as incorporating evolutionary information and has been shown to substantially increase the accuracy of neural networks for protein secondary structure prediction; similar improvements obtained using evolutionary information in predicting cysteine oxidation state and disulfide connectivity have been demonstrated[1,8].

IV. Conclusions

We have proposed and tested a novel machine learning method for predicting disulfide connectivity patterns in proteins. Performance is better. In addition, our model guarantees a significant decrease in training time. One obvious direction for further study is to combine cysteine bonding state predictors with a pairing algorithm like the one presented in this paper, in order to build a complete predictor of disulfide bonds.

Disulfide bonds can also be seen as a special (and important) case of residue contacts. Therefore it may be important to compare and combine predictors of disulfide bonds with predictors of contact maps whose performance is improving but still appears unsatisfactory for long ranged interactions.

Acknowledgment

The authors thank Prof. Zhaoyue Wang, Tianjin Medical University, for her constructive advice.

References

[1] P.Fariselli, P.Riccobelli and R.Casadio, “Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins”, Proteins, Vol.36, pp. 340–346, August 1999.

[2] A.Fiser and I.Simon, “Predicting the oxidation state of cysteines by multiple sequence alignment”,Bioinformatics,Vol.16, pp.251-256,March 2000.

[3] M.H. Mucchielli-Giorgi, S.Hazout,and P.Tufféry, “Predicting the disulfide bonding state of cysteines using protein descriptors”, Proteins, Vol.46,pp.243–249, February 2002.

[4] PL.Martelli, P.Fariselli, L.Malaguti, and R.Casadio, “Prediction of the disulfide bonding state of cysteines in proteins with hidden nerual networks”, Protein Eng, Vol.15,pp.951-953,December 2002.

[5] MT.Petersen,PH. Jonson, and SB.Petersen, “Amino acid neighbours and detailed conformational analysis of cysteines in proteins”, Protein Eng.,Vol.12, pp. 535–548, July 1999.

[6] HN.Gabow, “An efficient implementation of Edmond's algorithm for maximum matching on graphs”, Journal of the ACM, Vol.23, pp. 221-234, April 1976.

[7] HM.Berman et al. “The Protein Data Bank”,. Acta Crystallogr. D, Vol.58,pp.899–907, June 2002.

[8] A.Vullo and P.Frasconi, “Disulfide connectivity prediction using recursive neural networks and evolutionary information”, Bioinformatics, Vol.20, pp. 653–659, May 2004.

[9] F.Ferrè and P.Clote, “DiANNA: a web server for disulfide connectivity prediction”, Nucleic Acids Res,Vol.33, pp.w230-w232, July 2005.