Predicting disulfide bond partners

F. Ferrè[1] and P. Clote[2]

Keywords: disulfide bond, neural network, machine learning.

1 Introduction

Disulfide bonds (covalently bonded sulfur atoms from nonadjacent cysteine residues) play a critical role for protein functionality and in stabilizing the protein structure. A number of relatively good algorithms have been developed to determine whether a cysteine is *reduced* (sulfur occurring in reactive sulfhydryl group SH) or *oxidized* (sulfur covalently bonded)[3], reaching 88% accuracy [5]. Despite this success there has been little progress in the problem of determining whether two half-cystines form a disulfide bond with each other – the *disulfide bond partner prediction* problem. In [1] a neural network is used to predict the probability of a disulfide bond between two half-cystines, using flanking sequence information, and subsequently, maximum weight matching is applied to pair those most likely partners.

Starting from the observation that there is a bias in the secondary structure preferences of free cysteines and half-cystines, we develop a neural network to learn disulfide bond preferences of both amino acid residues and secondary structure assignment of the symmetric flanking regions centered at partner half-cystines. Considering the secondary structure of pairs of half-cystines known to form a disulfide bond, some combinations are preferred, presumably indicating a sort of structural complementarity. This novel approach, as calibrated using receiver operating characteristic (ROC) curves [2], shows a marked improvement over previous work of Fariselli and Casadio [1]. Our final stand-alone program uses a neural network on the symmetric flanking residues about both cysteines of a potential disulfide bond, along with the PsiPred-determined secondary structure of the residues and PsiBlast-determined evolutionary information.

2 Methods and Results

We built a database by extracting flanking residues from the symmetric window of size w centered at each half-cystine in each mono-chain peptide domain from the nonredundant collection PDBselect25 [3], using DSSP to determine the cysteine oxidation state. Given two size w windows centered at an N- resp. C-terminus half-cystines, we then extracted DSSP [4] secondary structure annotations for each of the $2w$ residues; subsequently we ran PsiBlast to produce a profile, consisting of frequencies $f(i,a)$, for each of the 20 amino acids a and each position $1 \leq i \leq 2w$, obtained from the multiple sequence alignment of homologous proteins. The resulting input to our neural network consisted of $2w \cdot 20$ frequencies, along with $2w \cdot 3$ additional binary inputs, which latter encode in unary the secondary structure (H, C, E) of each of the $2w$ residues. When training the neural network, we used output value of 1 for an input corresponding to a valid disulfide bond, as determined by DSSP, and 0 for a pair of half-cystine flanking regions for incorrectly paired half-cystines. Altogether, there were $O(N)$ many positive [resp. $O(N^2)$ many negative] training examples.

[1]Department of Biology, Boston College, Chestnut Hill, MA 02467, ferref@bc.edu

[2]Departments of Biology and Computer Science (courtesy appointment), Boston College, Chestnut Hill, MA 02467, clote@bc.edu

[3]Disulfide-bonded cysteines are known as *half-cystines*, while reduced cysteines are also called *free* cysteines.

For the resulting neural network, trained with evolutionary information and secondary structure preferences, we tested a variety of possible network architectures. Of those tested, two architectures showed the best results in 20-fold cross-validation experiments with our database (see discussion above) using a window size of $w = 11$ residues. The first architecture had one hidden layer with two units, while the second had two hidden layers with 5 and 2 units, respectively. See Figure 1 for a summary of the statistics of our neural network, as well as a ROC curve comparison of our method with that of Fariselli and Casadio [1]. For the latter, we parsed the Fariselli-Casadio CONPRED neural network scores for likelihood of disulfide bond formation, without using their additional application of maximum weight matching.

Accuracy	76.58%
True positive rate%	81.05%
Falso positive rate %	26.57%
Correlation coefficient	53.66%
Sensitivity	81.05%
Specificity	73.43%

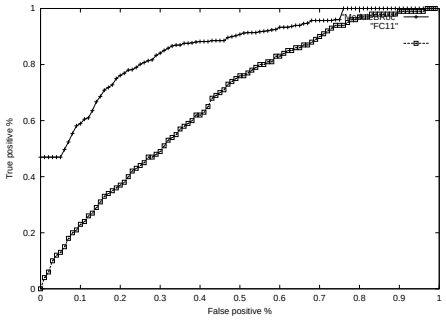

Figure 1: *(i)* Performance of the neural network disulfide connectivity prediction using secondary structure and evolutionary information. *(ii)* ROC curve for our method, described in this paper, compared with that of CONPRED – our method is the upper curve. Window size for both algorithms is $w = 11$.

After training, our prediction software works as follows. Given an input peptide along with user-designated half-cystine positions, our program uses PSIBLAST to obtain a profile and PSIPRED to predict secondary structure for the flanking residues. Our software then calls the neural network described in this paper.

References

[1] P. Fariselli and R. Casadio. Prediction of disulfide connectivity in proteins. *Bioinformatics*, 17(10):957–964, 2001.

[2] M. Gribskov and N.L. Robinson. The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers Chem.*, 20:25–34, 1996.

[3] U. Hobohm and C.Sander. Enlarged representative set of protein structures. *Proteins Science*, 3:522, 1994.

[4] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[5] P.L. Martelli, P. Fariselli, L. Malaguti, and R. Casadio. Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. *Protein Eng.*, 15(12):951–953, 2002.