

Diresidue neural network for the prediction of disulfide connectivity and ligand-bound cysteines

Fabrizio Ferre¹, Peter Clote²

Keywords: Disulfide connectivity, Machine learning, Neural networks, Position specific scoring matrices, ligand-bound cysteines

1 Introduction.

Position-specific scoring matrices (PSSMs), obtained by computing position-dependent log-odds scores from a multiple sequence alignment for a particular class of proteins, have often been used as a protein classification tool. Assuming positional independence, the monoresidue PSSM of a class of proteins is provably the maximum likelihood estimator. Nevertheless, in some cases experimental evidence suggests that protein sequences can be more adequately modeled using diresidue, rather than monoresidue, position-specific scoring matrices [1]. Here we show how to integrate this diresidue signal into a neural network, and we present results obtained by applying our diresidue neural network for the prediction of disulfide connectivity (i.e. which cysteine pairs form a disulfide bond, given an input protein sequence) and to the prediction of which cysteines are bound to a ligand. Cysteine residues play a unique role critical role in determining protein stability and function. Cysteines may be reduced (*free* cysteines) or oxidized; the latter may be involved in a disulfide bond (half-cystine) or instead covalently bonded to a metallic ligand. It is relatively easy to discriminate between half-cystines and free cysteines, and successful efforts have been made for the prediction of ligand-bound cysteines [2], while it is far more difficult to predict the correct disulfide bond topology in a protein; indeed, only a few attempts have been made to solve this problem [3-5]. Nevertheless, the knowledge of cysteine connectivity can be of crucial importance for the understanding of protein function and can reduce greatly the conformational space for protein structure prediction algorithms. We show that our novel approach [6] leads to results that are comparable and in many cases better than the current state-of-the-art methods [2-4].

2 Methods and Results.

We developed a novel neural network approach (the *diresidue* neural network) in which a first hidden layer is connected to the units of the input layer in such way that each hidden layer unit collects the output of input units encoding for a given pair of residues in the input sequence. For the ligand-bound prediction problem, the input is a symmetric window of size $w = 11$ centered about a cysteine. Each window residue is encoded using evolutionary information - i.e. frequencies $f(i, a)$, for each of the 20 amino acids a and each position $1 \leq i \leq w$, obtained from the multiple sequence alignment of homologous proteins. Additionally, secondary structure information is encoded in unary format by the addition of three input units (for example, helix is encoded 1 0 0, coil is 0 1 0 and sheet is 0 0 1). Positive instances are derived from PDB annotations, while negative instances are either free cysteines or half-cystines, or both. The

network architecture thus contains $w \cdot 23$ input units. We designed a first hidden layer containing $\binom{w}{2} = w(w-1)/2$ units,

one for each pair $1 \leq i < j \leq w$ of positions, with connections to input units representing the profile for residues at position i, j and secondary structures at those positions. Thus each of the $w(w-1)/2$ hidden units in the first hidden layer (the *diresidue* layer) is connected to $2(20 + 3) = 46$ input units (Figure 1). A second hidden layer, containing five units, all fully connected with those of the first hidden layer, is then fully connected to the single output unit. We designed this unusual neural network architecture, with the aim of emphasizing the signal that arises when using diresidue position specific scoring matrices [6], i.e. for all windows of length w , for positions $1 \leq i < j \leq w$ and amino acids a, b , we consider the frequency of occurrence of amino acid a in position i when amino acid b is found in position j ; moreover, though there are many hidden units, the training phase is still reasonably fast since the diresidue layer is not fully connected with the input layer. A similar architecture is used for the disulfide connectivity prediction, with the difference that the neural network input is a pair of symmetric windows of size $w = 11$ centered on a cysteine. In this case, the positive dataset

¹ Boston College, Biology Department, Chestnut Hill, MA 02467 (USA)

² Boston College, Biology and Computer Science Departments, Chestnut Hill, MA (USA)

contains all the pairs that are known disulfide bonds, while the negative contains all the other pairs. Results are shown in Figure 2.

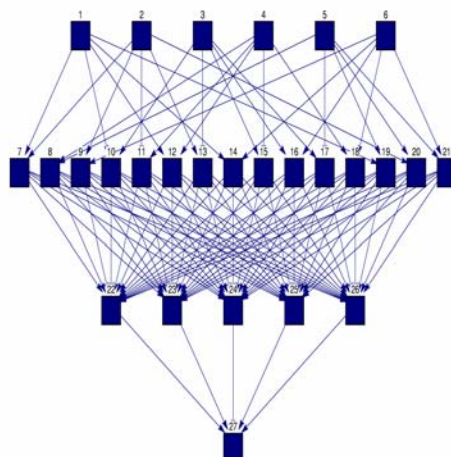


Figure 1: A toy example of the diresidue neural network architecture. Six input units (1,...,6) are connected to the w -choose 2 units of the first hidden layer (7,...,21), called the *diresidue* layer. Each pair of input units is connected to a distinct unit in the diresidue layer. The units of the diresidue layer are then fully connected to the five units (22,...,26) of the second hidden layer, which are fully connected to the single output unit. In the connectivity prediction application, each residue is encoded by 23 input units (20 encoding the evolutionary information and 3 for the secondary structure information), therefore each unit in the diresidue layer is connected $23 + 23 = 46$ input units that code a pair of residues.

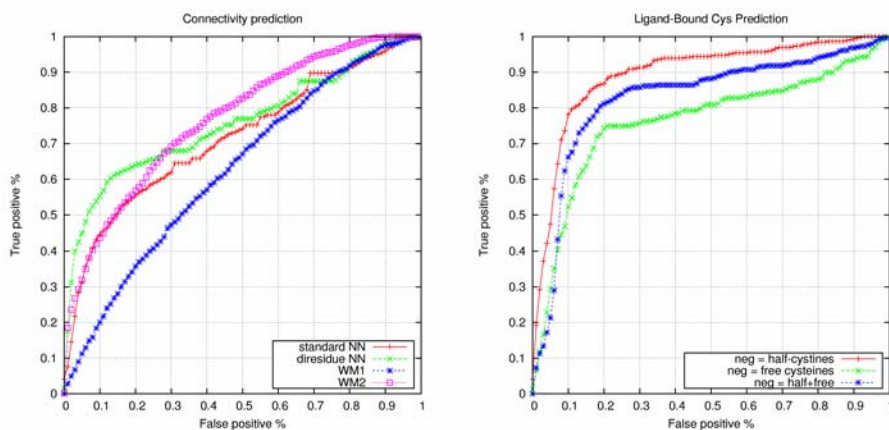


Figure 2: Performance of our method. Left panel: disulfide connectivity. The ROC curves show the performance of the diresidue neural network (green), a standard neural network with the same input and two hidden layers of 5 and 2 units respectively (red), monoresidue weight matrix (blue) and diresidue weight matrix (magenta). Right panel: ligand-bound cysteine prediction. The ROC curves have been obtained using as positive dataset the window content of ligand-bound cysteines, while the negative dataset may be the half-cystines (red), the free cysteines (green) or both (blue). Note that from these results it seems easier to predict ligand-bound cysteines from half-cystines rather than from free cysteines.

References

1. Bulyk, M.L., P.L. Johnson, and G.M. Church, *Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors*. *Nucleic Acids Res*, 2002. **30**(5): p. 1255-61.
2. Passerini, A. and P. Frasconi, *Learning to discriminate between ligand-bound and disulfide-bound cysteines*. *Protein Eng Des Sel*, 2004. **17**(4): p. 367-73.
3. Zhao, E., et al., *Cysteine separations profiles (CSP) on protein sequences infer disulfide connectivity*. *Bioinformatics*, 2004.
4. Vullo, A. and P. Frasconi, *Disulfide connectivity prediction using recursive neural networks and evolutionary information*. *Bioinformatics*, 2004. **20**(5): p. 653-9.
5. Fariselli, P., P.L. Martelli, and R. Casadio. *A neural network based method for predicting the disulfide connectivity in proteins*. in *Knowledge Based Intelligent Information Engineering Systems and Allied Technologies (KES)*. 2002: IOS Press.
6. Ferre, F. and P. Clote, *Disulfide connectivity prediction using secondary structure information and diresidue frequencies*. *Bioinformatics* (in press), 2005.