Large-Scale Prediction of Disulphide Bond Connectivity

Pierre Baldi , Jianlin Cheng
{pfbaldi,jianlinc}@ics.uci.edu

Alessandro Vullo
alessandro.vullo@ucd.ie

Schoolof Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425

Computer Science Department
University College Dublin
Dublin, Ireland

Cysteine is a semiessential proteinogenic amino acid with the formula HOOC–CH(NH_2)–CH_2–SH. The formation of covalent links among cysteine (Cys) residues with disulphide bridges is an important and unique feature of protein folding and structure. Disulphide bridges may link distant portions of a protein sequence, providing strong structural constraints in the form of long-range interactions.

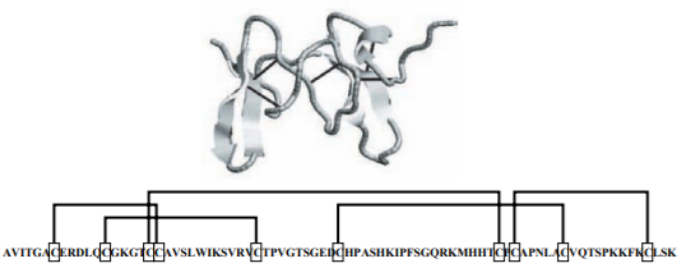

Figure 1: Structure (top) and connectivity pattern (bottom) of intestinal toxin 1, PDB code 1IMT. Disulphide bonds in the structure are shown as thick lines.

Thus prediction/knowledge of the disulphide connectivity of a protein is important and provides essential insights into its structure and possibly also into its function and evolution.This paper introduces a new method for the prediction of disulphide bond connectivity. In this paper 2-Dimensional Recursive Neural Network (2D-RNN, [1]) is used to predict disulphide connectivity in proteins starting from their primary sequence and its homologues. The output of 2D-RNN are the pairwise probabilities of the existence of a bridge between any pair of cysteines. Candidate disulphide connectivities are predicted by finding the maximum weight perfect matching.

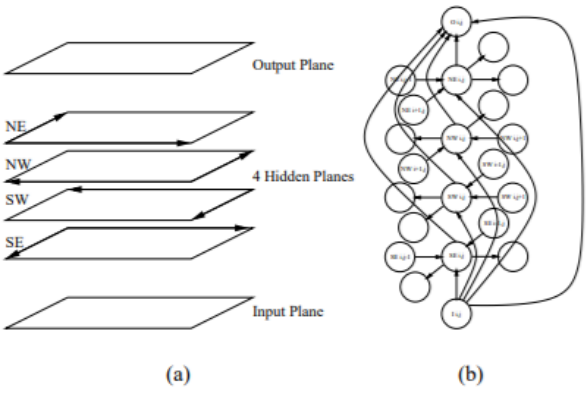

Figure 2: (a) General layout of a 2D-RNN for processing two-dimensional objects such as disulphide contacts, with nodes regularly arranged in one input plane, one output plane, and four hidden planes. (b) Connections within a vertical column (i, j) of the directed graph.

Here the underlying directed graph for disulphide connectivity has six 2D-layers: input, output, and four hidden layers (Figure 2(a)). Vertical connections, within an (i, j) column, run from input to hidden and output layers, and from hidden layers to output (Figure 2(b)). In each one of the four hidden planes, square lattice connections are oriented towards one of the four cardinal corners.

In a disulphide contact map prediction, the (i, j) output represents the probability of whether the i-th and j-th cysteines in the sequence are linked by a disulphide bridge or not. This prediction depends directly on the (i, j) input and the four-hidden units in the same columnn. Hence, using weight sharing across different columns, the model can be summarized by 5 distinct neural networks in the form :-

$$\begin{pmatrix} O_{ij} = N_o(I_{ij}, H_{i,j}^{NW}, H_{i,j}^{NE}, H_{i,j}^{SW}, H_{i,j}^{SE}) \\ H_{i,j}^{NE} = N_{NE}(I_{i,j}, H_{i-1,j}^{NE}, H_{i,j-1}^{NE}) \\ H_{i,j}^{NW} = N_{NW}(I_{i,j}, H_{i+1,j}^{NW}, H_{i,j-1}^{NW}) \\ H_{i,j}^{SW} = N_{SW}(I_{i,j}, H_{i+1,j}^{SW}, H_{i,j+1}^{SW}) \\ H_{i,j}^{SE} = N_{SE}(I_{i,j}, H_{i-1,j}^{SE}, H_{i,j+1}^{SE}) \end{pmatrix} \quad (1)$$

where N represents the NN parameterization.

Here 2D-RNN architectures is trained using the SP39 data set to compare with other published results. Then the performance is evaluated using the precision P (P=TP/(TP+FP) with TP = true positives and FP = false positives) and recall R (R=TP/(TP+FN) with FN = false negatives).

In some cases, results are substantially better. For instance, in the case of 3 bonded cysteines, the precision reaches 0.61 and 0.51 at the pair and pattern levels, whereas the best similar results reported in the literature are 0.51 (pair) and 0.41 (pattern).

Analysis of the prediction results shows that there is a relationship between the sum of all the probabilities in the graph (or the output layer of the 2D-RNN) and the total number of bonded cysteines. Using this, the total number of bonded cysteines is estimated using linear regression and rounding off, making sure that the total number is even and does not exceed the total number of cysteines in the sequence.

For the last set of experiments, No assumption is made regarding the knowledge of the bonding state and 2D-RNN approach is applied to all the cysteines (both bonded and not bonded) in each sequence. The number of bonds, the bonding state, and connectivity pattern are predicted using one predictor. Experiments are run both on SP39 (4-fold cross validation) and SP41 (10-fold cross validation).

This paper presents a complete system for disulphide connectivity prediction in cysteine-rich proteins. Assuming knowledge of cysteine bonding state, the method outperforms existing approaches on the same validation data. The results also show that the 2D-RNN method achieves good recall and accuracy on the prediction of connectivity pattern even when the bonding state of individual cysteines is not known. Differently from previous approaches, this method can be applied to chains with K > 5 bonds also and yields good predictions of the total number of bonds, as well as of the bonding states and bond locations. Training can take days but once trained predictions can be carried on a proteomic or protein engineering scale.

References : -

[1] P. Baldi and G. Pollastri. *The principled design of large – scale recursive neural network architectures–dag-rnns and the protein structure prediction problem*. Journal of Machine Learning Research, 4:575–602, 2003.

[2] A. Vullo and P. Frasconi. *Disulfide connectivity prediction using recursive neural networks and evolutionary information*. Bioinformatics, 20:653–659, 2004.