Critical Review

Topic : Large-Scale prediction of Disulphide Bond Connectivity

Name : Prince Gupta

Roll No. : B19BB033

Cysteine is a semi- essential proteinogenic amino acid. The formation of covalent links among cysteine (Cys) residues with disulphide bridges is an important and unique feature of protein folding and structure. Disulphide bonds are important in stabilizing the native state of proteins. Disulphide bridges may link distant portions of a protein sequence, providing strong structural constraints in the form of long-range interactions. 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.

In this paper, a 2-D recursive neural network (RNN) was used for scoring undirected graphs that represent connectivity patterns by their similarity to the correct graph. Each vertex in the graph contains a description of the local environment of the bonded cysteine. Specifically, 20-element vectors corresponding to multiple alignment profile in a local window around each cysteine were used. During the prediction stage, the score computed by the RNN was used to exhaustively search the space of all possible candidate graphs. Comparative experiments on the SP39 dataset show improved results. For the dataset containing two to five bonds, Qp score (fraction of correctly assigned proteins) of 44% and Qc score (recall or sensitivity) of 49% were obtained.

Disulphide Connectivity Prediction for Bonded Cysteines : Here it was assumed that the bonding state of the cysteines are known. 2D-RNN was trained using the SP39 data set to compare with other published results.

The performance was then 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).

Disulphide Connectivity Prediction from scratch : Here it was not assumed that the bonding state of the cysteines are known and the 2D-RNN was

applied to all the cysteines(bonded or non-bonded) in the sequence. 2D-RNN was trained using both the the SP39 data set (4-fold cross validation) and SP41(10-fold cross validation).

When it was assumed that the bonded state of the cysteines is known, it 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.

➔ Pros of the paper :-

1. 2-D RNN used in the research can process inputs of any lengths.

2.The RNN model used can remember each information throughout the time which is very helpful in any time series predictor.

3. The model size does not increase with the increase in size of the input.

4. The weights can be shared across the time steps.

5. It shows improved accuracy in predicting disulphide bonds over already existing approaches.

➔ Cons of the paper :-

1. Due to its recurrent nature, the computation is a slow process.

2.Training with a RNN model can be difficult.

3. The method does not achieve good accuracy on proteins with more than 5 cysteine bonds.

4. Training the model can take days.

➔ Technical Suggestions :-

1. We can use feature selection to improve the performance of the 2-D Recursive Neural Network. We can try to remove only some features (pairwise profile correlation, pairwise mutual information, residue type, and protein information features) once a time to test how they affect the

prediction accuracy. However, we cannot clearly state if the improvement is due to the random variation or due to the removal of the features. But at least, these features were not essential or being compensated by other similar features.

2. For an input feature vector associated with a pair of residues, we can use Support Vector Machines (SVMs) to predict if the two residues are in contact (positive) or not (negative). SVMs provide a non-linear classifier model by non-linearly mapping the input vectors into a feature space and using linear methods for classification in the feature space. Thus SVMs, and more generally kernel methods, attempt to combine the advantages of both linear and nonlinear methods by first embedding the data into a feature space equipped with a dot product and then using linear methods in the feature space to perform classification or regression tasks based on the Gram matrix of dot products between data points.

➔ Future Improvements :-

1. We can develop a classifier using kernel methods to distinguish proteins containing disulphide bridges from proteins with no disulphide bridges. This will in turn reduce the time taken for the model to train on the data set.

2.We can also take into account the effect of additional input information such as secondary structure of the protein to be analysed and the accessibility of the solvent.

3. We can also use the information of predicted cysteines contacts in 3D protein structure prediction for maximum benefits.

4. We could search for a new, larger and better training set so that we could train the model more effectively.