TCS-BIOINFO
4709-4671-8155

Highly Trained Neural Network Predictors

Dilip Antony Joseph[1], Vidyasagar, M[2] and Sharmila Mande[2]
[1]Indian Institute of Technology, Madras, [2]Tata Consultancy Services
dilip@peacock.iitm.ernet.in

Artificial Neural Network based predictors have proven to be very effective in protein secondary structure prediction. The neural network is able to predict the state of the central residue of a window of amino acids. These predictors have obtained prediction accuracies of over 75%. The number of neurons in a predictor is often more than 20000. To effectively train a network with such a large number of changeable parameters requires a very large training set. The predictor developed here attempts to use a large amount of the available secondary structure data in training.

A standard feed forward neural network was used in the predictor. The input to the neural network consisted of a window of 15 amino acids [3]. Each amino acid in the window was represented by the 20 numbers obtained from the PSIBLAST [1, 2] profile of the sequence. The network classifies the central residue of the window as either in the Alpha Helix, Beta Strand or Coil state. A second network was used to 'clean up' the secondary structure sequence produced by the first network. The training set consisted of over 6500 protein sequences from the PDB SELECT [4, 5] database, which gave 1436264 input-output pairs for training the network. The effectiveness of the above training set is diminished by the similarity (up to 90 %) between some of the sequences in the set. However, this training set did give better prediction accuracies than the networks trained on a smaller number of sequences. Nine neural networks trained independently on the same training set (randomly shuffled) were constituted into a jury. This also led to a small increase in prediction accuracy.

It has been observed that the larger and more varied the training set; the better is the prediction accuracy. As more and more structural data becomes known in the future, it is important to include those sequences in the training set. However, retraining the whole network is a time consuming exercise. The effectiveness of retraining the network with only the new sequences is studied.

A jury consisting of highly trained predictors along with specialized alpha and beta strand predictors can effectively increase the prediction accuracies.

1. Altschul S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* ***25*** *(17),* 3389-3402
2. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292*** *(2),* 195-202.
3. B. Rost and C. Sander. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc. Natl. Acad. Sci. USA, Vol.**90**, 7558-7562
4. U. Hobohm, M. Scharf, R. Schneider, C. Sander (1992) Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Science **1**, 409-417
5. U. Hobohm and C. Sander (1994) Enlarged representative set of protein structures, Protein Science **3**, 522