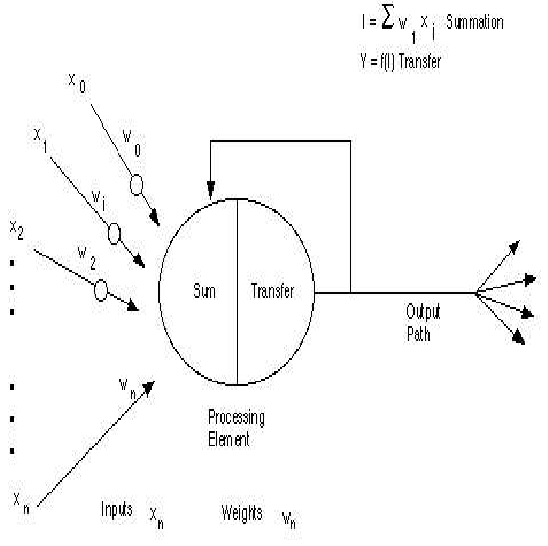
Protein Secondary Structure Prediction using Neural Networks

Arpit Ghoting

What is a Neural Network ?

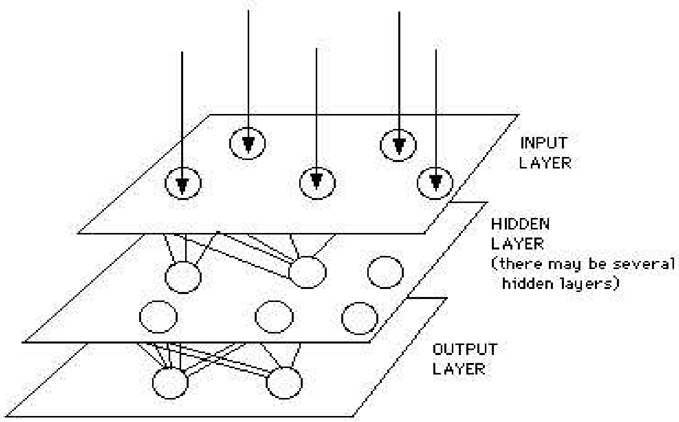
- Neural Network is a system loosely modeled on the human brain. It is an attempt to simulate within specialized hardware or sophisticated software, the multiple layers of simple processing elements called neurons. Each neuron is linked to certain of its neighbors with varying coefficients of connectivity that represent the strengths of these connections. Learning is accomplished by adjusting these strengths to cause the overall network to output appropriate results.

The Artificial Neuron



Note that various inputs to the network are represented by the mathematical symbol, x(n). Each of these inputs are multiplied by a connection weight, these weights are represented by w(n). In the simplest case, these products are simply summed, fed through a transfer function to generate a result, and then output.

Communication and types of connections



Layers

The figure above shows, the neurons are grouped into layers. The input layer consist of neurons that receive input form the external environment. The output layer consists of neurons that communicate the output of the system to the user or external environment. There are usually a number of hidden layers between these two layers; the figure above shows a simple structure with only one hidden layer.

When the input layer receives the input its neurons produce output, which becomes input to the other layers of the system. The process continues until a certain condition is satisfied or until the output layer is invoked and fires their output to the external environment.

- Inter-layer connections

  There are different types of connections used between layers, these connections between layers are called inter-layer connections.

- Fully connected: Each neuron on the first layer is connected to every neuron on the second layer.

- Partially connected: A neuron of the first layer does not have to be connected to all neurons on the second layer.

- Feed forward: The neurons on the first layer send their output to the neurons on the second layer, but they do not receive any input back form the neurons on the second layer.
- Bi-directional: There is another set of connections carrying the output of the neurons of the second layer into the neurons of the first layer.

Learning

The brain basically learns from experience. Neural networks are sometimes called machine learning algorithms, because changing of its connection weights (training) causes the network to learn the solution to a problem. The system learns new knowledge by adjusting these connection weights.

The training method usually consists of one of three schemes:

- Unsupervised learning :The hidden neurons must find a way to organize themselves without help from the outside. This is learning by doing.

- Reinforcement learning: This method works on reinforcement from the outside. The connections among the neurons in the hidden layer are randomly arranged, then reshuffled as the network is told how close it is to solving the problem. Reinforcement learning is also called supervised learning, because it requires a teacher. The teacher may be a training set of data or an observer who grades the performance of the network results.

- Back propagation: This method is proven highly successful in training of multilayered neural nets. The network is not just given reinforcement for how it is doing on a task. Information about errors is also filtered back through the system and is used to adjust the connections between the layers, thus improving performance.

Applications

- Prediction: Uses input values to predict some output. e.g. predict secondary protein structures, pick the best stocks in the market, predict weather, identify people with cancer risk.
- Classification: Use input values to determine the classification. e.g. is the input the letter A, is the blob of the video data a plane and what kind of plane is it.
- Data association: Like classification but it also recognizes data that contains errors. e.g. not only identify the characters that were scanned but identify when the scanner is not working properly.
- Data Conceptualization: Analyze the inputs so that grouping relationships can be inferred. e.g. extract from a database the names of those most likely to by a particular product.
- Data Filtering: Smooth an input signal. e.g. take the noise out of a telephone signal.

Introduction

Basics

- Secondary structures are grouped into 3 main categories (1) alpha - helix, (2) Beta - sheets, (3) Coil
- Approach: Divide and Conquer
- Divide: This step can be done by designing separate models for each of the different classes to be recognised.
- Conquer: This step can be carried out by combining the individual experts using another neural network.
- A lot of interesting work has been done on predicting secondary structures, and over the last 10-20 years the methods have gradually increased in accuracy. This improvement is partly due to the increased number of reliable structures from which rules can be extracted and partly due to the improvement in the methods.

Aim

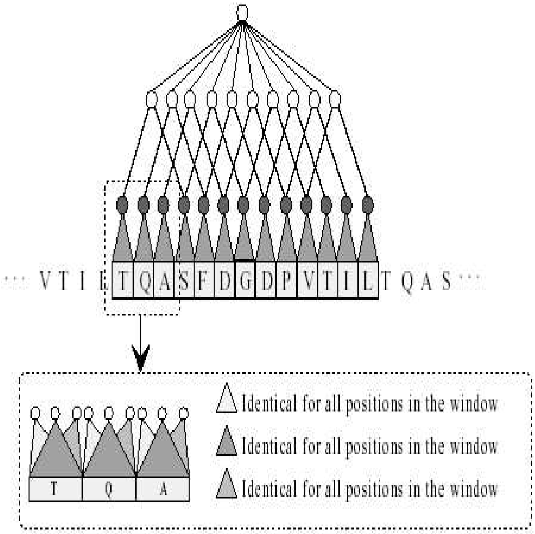
This paper aims to get as good predictions as possible from single sequences i.e. only the amino acid sequence of the considered protein is used as input.

- This work has three stages:
- Individual networks were designed for prediction of the three structures.
- Instead of using only one network for each type of structure, an ensemble of 5 networks was used for each structure.
- These ensembles of single structure networks were combined by another neural network to obtain a three state prediction.

Single Structure Prediction

Adaptive encoding of amino acids

- As in most of the existing methods, the secondary structure of the jth residue Rj is predicted from a window of amino acids.
- Usually the amino acids are encoded by 21 binary numbers, such that each number corresponds to one amino acid. The last number corresponds to a space, and is used to indicate the end of a protein. This type of encoding is called orthogonal encoding.
- This type of encoding is highly redundant, since 21 symbols can be encoded in 5 bits.It is also possible to let the network choose the best encoding of the amino acids.This adaptive encoding scheme is also known as Local Encoding.
- For each window position the 20 inputs are connected to 20 X M hidden units by 20 X M weights.
- The set of weights corresponding to one window position is identical to those used for all other window positions.
- More precisely if the weight from input j to hidden unit i is called W(k)ij, then W(k)ij = W(l)ij for all k and l.


··· V T I E T Q A S F D G D P V T I L T Q A S ···
T
Q
A
Identical for all positions in the window
Identical for all positions in the window
Identical for all positions in the window

Structured Networks

- A residue in an alpha helix is hydrogen bonded to t he fourth residue above and to the fourth residue below in the primary sequence,and it takes 3.6 amino acids to make a turn in an alpha - helix.

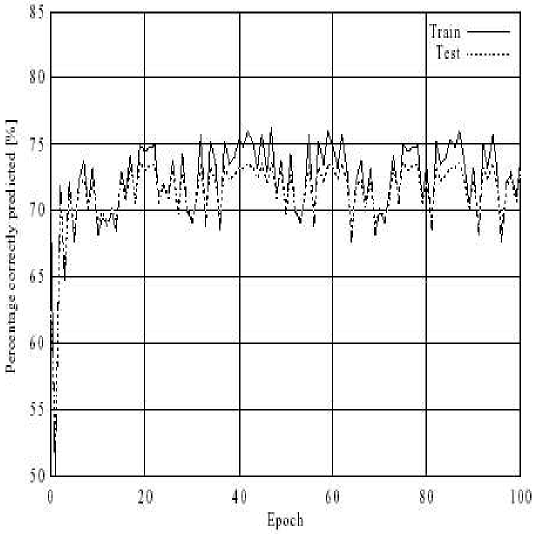
- It is likely that this periodic structure is essential for the characterization of an alpha - helix.

- These characteristics are all of local nature and can therefore be easily be built into a network that predicts helices from windows of the amino acid sequence.

- In contrast to helices, Beta - sheets and coils do not have such a locally described periodic structure. Therefore, the sheet and coil networks only use the local encoding scheme.

Results

The following figure shows the result of training the structured alpha network and then testing it.



The following observations can be made:

- The network gives reliable estimates of prediction accuracy on new proteins not in the training set used for developing the method.
- The observed fluctuations are mostly due to the use of a balanced training set of negative examples (non - helix) used in each training epoch.

Combining Single - Structure Predictions

- To combine the single structured predictions a neural network is used.
- The network takes a prediction of 15 single structure predictions of helix, sheet and coil as input and is fully connected to the output via 10 hidden units.
- In addition to combining the submodels this network acts as a filter i.e it has a tendency to eliminate unrealistic predictions and it results in more realistic secondary structure segments.

Conclusion

- The use of specialised models for protein secondary structure have been investigated.
- By using ensembles of specialised neural networks for predicting each of the three secondary structures over-fitting was avoided.
- One of the features of the single - structured networks were an adaptive encoding of the amino acids.