Using Neural Networks for Prediction of Secondary Structure in Proteins

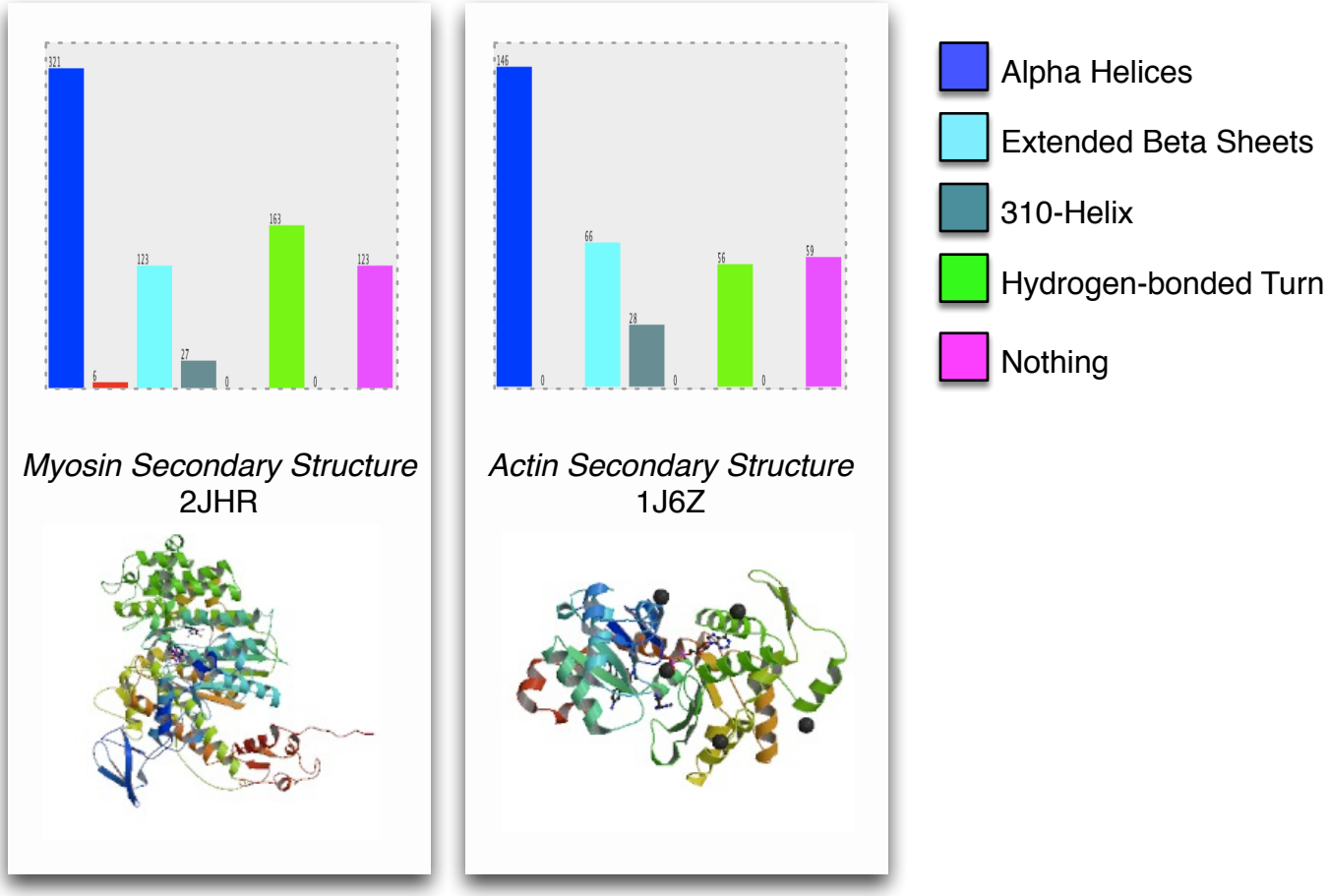
Bio 465 Lab 4
Duane Johnson
Mar 5, 2009

Introduction

The molecules formed by the machinery of living organisms are extremely complex. In order to make sense of the properties and functions of these molecules, some shape patterns have been identified which form the substructures of one class of molecules, namely proteins. These substructures, known as *secondary structure,* form patterns of helices, sheets, and strands. In order to minimize the work done in the lab, it is useful to attempt to predict these structures using neural network algorithms. This report outlines one such attempt.

Procedure

Sharma et al. [1] used a variety of simple protein families to configure and train a neural network for the purpose of secondary structure prediction. Based on their property of having each of the varying kinds of secondary structures, their training data included the 4 protein families: actin, myosin, phytochrome and ribonuclease. In order to find a reasonable training sample, I used protein sequences in the actin and myosin families, namely the Myosin-2 Motor Domain 2JHR, and the Rabbit Actin 1J6Z:



The knowledge analysis tool Weka was used to as a workbench environment simulating a perceptron. The actin and myosin protein data were split into overlapping sequences of 13 residues so that a "window" of values could then be fed through the perceptron. Each possible amino acid type was represented by a node in the perceptron, which in turn was mapped to each possible secondary structure state. The final output is the cumulative weighted decisions based on the input values and their affect on the 2-layer perceptron within Weka.

The Stride web tool [2] and the makearff.pl script [3] were used to prepare the PDB data for Weka.

Results

The results of the training are expressed in the table below.

	Myosin Trained	**Actin Trained**
10-Fold Tested	37.1 %	38.9 %
Tested Against The Other	27.6 %	33.4 %

The tests for accuracy are of two kinds: "fold" means that data from the training set was put aside and used after training for testing; "against the other" means that, for example, the perceptron was trained using myosin and then tested against actin (and vice versa).

In order to achieve the above results, parameters with a training time of 500, a hidden layer value of 2, and a learning rate of 0.3 and momentum of 0.2 were used.

Conclusion

While the results were not accurate enough to be particularly useful in predicting secondary structure, they were nevertheless robust enough to be used to test against their own sequences and that of at least one other protein.

The actin-trained perceptron was slightly more accurate than the myosin-trained perceptron which may be attributed to the more balanced secondary structure levels in the actin source data (extended beta sheets and hydrogen-bonded turns were nearly equal in actin, whereas the myosin levels were slightly off).

Providing longer sequences of protein data or adding more than 2 levels of hidden layers may have improved the overall results, but these variations were not made during this study.

References

1. Application of Neural Networks for Protein Sequence Classification, Sharma et. al. 2004
2. Stride Web Interface (http://webclu.bio.wzw.tum.de/cgi-bin/stride/stridecgi.py)
3. Bio465 Script (http://dna.cs.byu.edu/bio465/Labs/makearff.txt)