Daniel Standage

BCB 569, Fall 2011

Term Project Proposal

The earliest methods for predicting protein secondary structure involved analyzing windows of consecutive amino acids to classify the residue at the center of the window in terms of its secondary structure. These methods were able to achieve about 50% accuracy, a huge improvement over random guessing but still leaving a lot of room for improvement. Subsequently, secondary structure prediction was the first bioinformatics problem to which a large-scale machine learning method was successfully applied. Neural networks typically have an accuracy of between 60-70% when applied to secondary structure prediction problems. Furthermore, by incorporating profile information (obtained by aligning the protein sequence to the sequences of homologous), machine learning methods can reach accuracy levels around 75% (Cheng *et al.,* 2009).

For my term project, I would like to explore the secondary structure prediction problem using machine learning methods. I see several possible ways of exploring novel approaches to this problem.

* Most machine learning methods for secondary structure prediction involve neural networks or support vector machines. It would be interesting to see how other machine learning methods compare to these.
* It would be interesting to explore a consensus data mining approach similar to that described in (Sen *et al.*, 2006) that combines two or more successful machine learning approaches to create a more accurate hybrid method.
* It would be interesting to see if we can identify additional features of the protein sequences that will improve the performance of these classifiers on secondary structure prediction problems.

I do not expect to explore each of these areas in depth given the short amount of time allotted for the term project. Likely, I will choose one area (or maybe two) as my primary focus after some preliminary exploration.

I will obtain data for my project (protein sequences and secondary structure classifications) from the Protein Data Bank (http://www.pdb.org/). For model training I will use Weka (Hall *et al.,* 2009), which has implementations for a wide variety of machine learning models as well as a graphical interface and a Java API.

Chang, J., A. N. Tegge, P. Baldi (2009) Machine Learning Methods for Protein Structure Prediction. *IEEE Reviews in Biomedical Engineering*, **1**, 41-49.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009) The WEKA data mining software: an update. SIGKDD Explorations, **11**(1), 10-18.

Sen, T.Z., H. Cheng, A. Kloczkowski, R.L. Jernigan (2006) A Consensus Data Mining secondary structure prediction by combining GOR V and Fragment Database Mining, Protein Science, **15**(11):2499-2506.