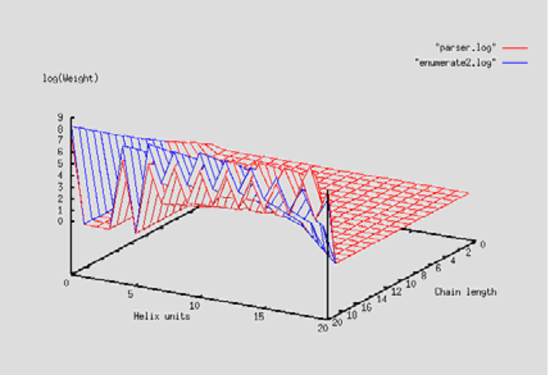
EIA-0205456: “Language, Learning, and Modeling Biological Sequences”
University of Pennsylvania
Investigators: Aravind Joshi, Fernando Pereira , Mark Liberman, John Lafferty (CMU), Ken Dill (UCSF), David Roos, Sampath Kannan, Lyle Ungar and David Searls (GSK)

Website: http://www.ircs.upenn.edu/sequences.html
Our overall goal is the application of natural language processing (NLP) and machine learning techniques for modeling biological sequences, such as certain long range dependencies and folded structures, for example.



This graph shows the agreement between the parser (red) and exact enumeration on a square lattice (blue) for computing the probability distribution of the helicity of chains of varying length. The parser approximates the exact enumeration fairly closely, and runs many times faster for longer chain lengths.

The main projects addressing our major goal are 1. Develop new techniques for integrating grammatical and probabilistic information. 2. Develop, integrate and evaluate grammatical, probabilistic, and approximate counting methods for fold prediction in secondary and tertiary structures of biomolecules. 3. Develop and evaluate probabilistic exponential models for gene finding, in particular, genes for apicoplast-targeted proteins in eukaryotic pathogens of the phylum Apicomplexa.

- More specifically, we have begun work on the modeling of alpha helices both by combining grammatical models and weights associated with the grammatical model that are derived from the energy considerations (partition function). The graphic above gives some details of the approximation of the exact enumeration by the parser. The grammatical model involved here describes self contacts in the helices. We have begun extending this study for contact patterns that are not describable by context-free grammars but can be described by tree-adjoining grammars.

- In many traditional approaches to machine learning, a target function is estimated using labeled data, which can be thought of as examples given by a "teacher" to a "student." Labeled examples are often, however, very time consuming and expensive to obtain, as they require the efforts of human annotators, who must often be quite skilled. For instance, obtaining a single labeled example for protein shape classification, which is one of the grand challenges of biological and computational science, requires months of expensive analysis by expert crystallographers. The problem of effectively combining unlabeled data with labeled data is therefore of central importance in machine learning. We have developed an approach to this problem based on a random field model, have established theoretical bounds on accuracy, and have demonstrated very promising performance for image and text classification tasks. We are currently investigating ways of extending these techniques for modeling biological sequences.

- In collaboration with the Genetics Department at Penn we are building a new gene finder based on conditional random fields that integrates a wide range of features, including coding potential, transcription, translation, and splicing signals, and EST-derived evidence of alternative splicing. We use a flexible feature representation that allows other sources of evidence to be incorporated as they become available. Preliminary results on standard test sets are promising.

All of the software developed in this project will be available under open-source license.