Protein Secondary Structure Prediction by Merged Hidden Markov Models

Student name: Christian A. Cumbaa
(M.Math student in computer science)
University of Waterloo, Waterloo, Ontario, Canada
Advisor: Dr. Forbes J. Burkowski
http://www.math.uwaterloo.ca/~ccumbaa

Project Overview

A protein molecule is a linear chain of amino acid residues, which typically folds into a complex, globular shape in its native solvent environment. The protein folding problem is that of determining the native three-dimensional (tertiary) structure of a protein molecule given only its amino acid sequence and its environment. The importance of the protein folding problem springs from the huge amount of genetic sequence data currently available and the many ongoing whole-genome sequencing projects. Determining the shape of a protein whose amino acid sequence is encoded in a gene sequence is an intermediate step on the path to understanding the function of an organism. Knowing the structure of a target protein is also crucial to rational drug design.

The protein folding problem is hard. Determining protein structure by experimental observation is an expensive and time-consuming process. Solving the structure by molecular dynamics simulation is not yet computationally feasible. Machine learning methods have therefore been developed in order to predict tertiary structure, but none are very successful.

A simpler, but related, problem is that of predicting protein secondary structure. Within a protein molecule, segments of amino acid residues align into regular substructures such as α helices, β sheets and coils. Secondary structure prediction is the assignment of α, β, and coil labels to each residue in a molecule. The best machine learning methods achieve a maximum success rate of about 75%, depending on the similarity of the target protein to proteins with known structure. These methods include sequence alignment, statistical methods, neural networks, and hidden Markov models (HMMs).

HMMs are a common tool for biological sequence analysis. A HMM is a probabilistic model that generates sequences by a series of random transitions between internal states and a random emission of sequence units after each transition.

We are developing a pattern-based, statistical approach to protein secondary structure prediction. Our model of protein folding assumes that protein structure is governed by short patterns (motifs) in the amino acid sequence. Each pattern exerts a local influence, represented by probability distributions over the structure space, on the underlying structure of a protein molecule. When two or more patterns occur in overlapping regions of a sequence, their structural influences combine to form a new, unified influence (probability distribution). We use a limited form of HMM to model the structural influences.

Our prediction method has a preprocessing step and a prediction step. The preprocessing step finds the patterns and calculates their structural influences. To find the patterns, a training set containing protein sequences with known secondary structure is searched using a pattern discovery algorithm. Next, for each pattern, a smaller training set is assembled from the underlying structures at each occurrence of the pattern. This smaller set is used to train a HMM representing the structural influence of the pattern on a protein.

The prediction step is applied to a target sequence. The target sequence is searched for occurrences of patterns found in the preprocessing step. The HMMs for each pattern are then combined using a special merging operation developed for this study, to create a single HMM describing the probability distribution over all possible underlying structures for the target sequence.

Our current research focuses on theoretical justification of the HMM merging step, selection of Bayesian prior probabilities in the HMM learning step, and methods for eliminating bias from sequences in the Protein Data Bank, the source of our training data.

Acknowledgement

This research is supported by a grant from Communications and Information Technology Ontario (CITO).