

Application of Language Technologies in Biology: Feature Extraction and Modeling for Transmembrane Helix Prediction

Madhavi Ganapathiraju
Carnegie Mellon University
<http://www.cs.cmu.edu/~madhavi>

Dissertation Abstract
May 2007

This thesis provides new insights into the application of algorithms developed for language processing towards problems in mapping of protein sequences to their structure and function, in direct analogy to the mapping of words to meaning in natural language. While there have been applications of language algorithms previously in computational biology, most notably hidden Markov models, there has been no systematic investigation of what are appropriate word equivalents and vocabularies in biology to date. In this thesis, we consider amino acids, chemical vocabularies and amino acid properties as fundamental building blocks of protein sequence language and study n-grams and other positional word-associations and latent semantic analysis towards prediction transmembrane helices.

First, a toolkit referred to as the Biological Language Modeling Toolkit has been developed for biological sequence analysis through amino acid n-gram and amino acid word-association analysis. N-gram comparisons across genomes showed that biological sequence language differs from organism to organism, and has resulted in identification of genome signatures.

Next, we used a biologically well established mapping problem, namely the mapping of protein sequences to their secondary structures, to quantitatively compare the utility of different fundamental building blocks in representing protein sequences. We found that the different vocabularies capture different aspects of protein secondary structure best. Finally, the conclusions from the study of biological vocabularies were used, in combination with the latent semantic analysis and signal processing techniques to address the biologically important but technically challenging and unsolved problem of predicting transmembrane segments.

This work led to the development of TMpro, which achieves reduced transmembrane segment prediction error rate by 20-50% compared to previous state-of-the-art methods. The method is a novel approach of analyzing amino-acid property sequences as opposed to analyzing amino acid sequences: following our work, it has already been applied towards protein remote homology detection and protein structural type classifications by others.