Markovian Domain Signatures: Statistical Segmentation of Protein Sequences - *Award Winning Poster*

Bejerano, G.[1], Seldin, Y.[1], Margalit, H.[2] and Tishby, N.[1]

[1] School of Computer Science and Engineering, Hebrew University of Jerusalem

2 Dept. of Molecular Genetics and Biotechnology, Hadassah Medical School, Hebrew University of Jerusalem

Characterization of a protein family by its distinct sequence domains is crucial for functional annotation and correct classification of newly discovered proteins. Conventional multiple sequence alignment-based methods, such as hidden Markov modeling (HMM), come to difficulties when faced with heterogeneous groups of proteins. However even many families of proteins sharing a common domain contain instances of several other domains, without any common linear ordering. Ignoring this modularity may lead to poor or even false classification and annotation. An automated method that can analyse a group of proteins into the sequence domains it contains is therefore highly desirable.

We apply a novel method to this problem. The method takes as input an unaligned group of protein sequences. It segments them and clusters the segments into groups sharing the same underlying statistics. A variable memory Markov model (VMM) is built using a prediction suffix tree (PST) data structure for each group of segments. Refinement is achieved by letting the PSTs compete over the segments. A deterministic annealing framework infers the number of underlying PST models while avoiding many inferior solutions. We show that regions of conserved statistics correlate well with protein sequence domains, by matching a unique signature to each domain. This is done in a fully automated manner, and does not require or attempt a multiple alignment. Several representative cases are presented. We identify a protein fusion event, refine an HMM superfamily classification into the underlying families the HMM cannot separate, and detect all 12 instances of a short domain in a group of 396 sequences.