Markovian Domain Fingerprinting: Statistical Segmentation of Protein Sequences

Gill Bejerano
jill@cs.hjui.ac.il
Yevgeny Seldin
seldin@cs.hjui.ac.il
Hanah Margalit*
ahans@md2.huji.ac.il
Naftali Tishby
tishby@cs.hjui.ac.il
School of Computer Science & Engineering and *Hadassah Medical School
The Hebrew University of Jerusalem

Abstract: We present several methods for protein sequence domain detection and classification. Our methods are fully automated and do not require or attempt additional sequence alignment. The methods are based on comparisons between groups of proteins, regardless of domain origin within different genomes. The methods can be applied to the segmentation of a protein sequence into domains and the relative positioning of domains. A new fast method is described for the detection of domain boundaries and domain segmentation. The methods are based on a global alignment of the protein sequence to a database of known domains and domain boundaries. The methods are validated on a large set of known domains and are compared to other methods. The results show that our methods achieve high accuracy and sensitivity in the detection of domain boundaries and domain segmentation.

Biological Motivation (I)
- Proteins are linear molecules, recognizable for early on, activity taking place within the cell.
- Recognizing domain sequences in the proteins is a major step in understanding protein function.

Empirical Results
In [Bejerano & Seldin, 2001, Seldin & Bejerano, 2001], we show how to implement PST-based domain segmentation.

Protein Sequence Segmentation: Domain Boundaries Determination

Protein Family Classification using Variable Memory Models
- Single PST Learning Algorithm Online
- Multiple PST Learning Algorithm Online

Sequence Segmentation
- PST-based methods, using single and multiple PSTs, are used to detect domain boundaries.
- The methods achieve high accuracy and sensitivity in the detection of domain boundaries and domain segmentation.

Detection of a Protein Fusion Event DNA Topoisomerase II

Future Directions
- Analysis of the relationship between protein function and the selected structural alignments
- Application to DNA-sequence analysis

Bibliography
1. 
2. Bejerano & Seldin (2001), Bioinformatics 17(1): 2:1, 3-2:42
5. Bejerano, Seldin, Margalit & Tishby (2001), Bioinformatics 17(1): 2:1, 3-2:42
7. Preparatory Model: Model training data as if the sample sequences all originate from a database model of varying length. Such a model is generated in [Bejerano & Tishby 1996, Preparatory Model], where the problem of variable length is not addressed.