Markovian Domain Fingerprinting:
Statistical Segmentation of Protein Sequences

Gill Bejerano   **Yevgeny Seldin**   **Hanah Margalit***   **Naftali Tishby**
jill@cs.huji.ac.il   seldin@cs.huji.ac.il   hanah@md2.huji.ac.il   tishby@cs.huji.ac.il

School of Computer Science & Engineering and *Hadassah Medical School
The Hebrew University of Jerusalem

Abstract: We present a novel method for protein sequence domain detection and classification. Our method is fully automated and does not require or attempt a multiple sequence alignment. It handles well heterogeneous multi-domain groups regardless of domain ordering within different proteins. The method constructs unique domain signatures through clustering regions of *conserved statistics*. Examples pinpoint domain boundaries, detect a protein fusion event, refine an HMM superfamily classification and find all 12 instances of a domain in 396 unaligned multi-domain sequences.

Biological Motivation (I)

- Proteins are Linear molecules, responsible for nearly all activity taking place within every living cell.
  - Reading the linear sequence is easy.
  - Determining or deducing the fold is hard.
  - What then can be said of the function?

  Observation: Proteins with similar sequences usually fall into similar folds, and the fold itself is crucial for preserving function.

  Definition: Protein family – a group of proteins sharing sequence similarities (and hopefully also origin, fold and function).

  Challenge: Group protein sequences into meaningful families.

A folded (linear) protein.

Protein Family Classification using Variable Memory Models

Input: a subset of related sequences (some known family members).

Objective: Generate a model that can discriminate between (previously unseen) family members and non-related proteins.

Current state of the art: profile Hidden Markov Models.

Disadvantages:
- Known hardness-of-training results, with implications on data requirements, training and prediction speed.
- The need for a multiple alignment of the training data.
- Essentially very short memory, affecting sensitivity.

Proposition: Model the training data as if the sample sequences all originate from a *Markov model of varying (high) order*.

Such a model was presented in [**Ron, Singer & Tishby '96, Mach. Learn.**], where predictive contexts of variable lengths are collected during training, into an efficient data structure representation named *Prediction Suffix Tree (PST)*.

Given a PST model T, a query sequence $x_1...x_m$ is scored through:

$$P^T(x_1...x_m) = \Pi_i P(x_i \mid x_1...x_{i-1}) \approx \Pi_i P(x_i \mid \mathbf{max\_suf}(x_1...x_{i-1}))$$

where max_suf($x_1...x_{i-1}$) is the longest suffix of $x_1...x_{i-1}$ memorized during training. E.g.:

A PST over the alphabet {a,b,c,d,r}

- Node labels represent memorized suffixes, and adjacent vectors are the associated next symbol distributions.
- Max_suf is found by traversing down the tree until leaf/stuck.

so P^T(braad) = P^T(b) P^T(r|b) P^T(a|br) P^T(a|bra) P^T(d|braa)
= P^T(b|λ) P^T(r|λ) P^T(a|r) P^T(a|bra) P^T(d|a) = 0.2 ·0.2 ·0.6 ·0.1 ·0.2

Single PST Learning Algorithm Outline

(Intuitive version:)
- Initialize a PST with a single root node (annotated by the overall per-symbol distribution).
- Go over all subsequences $\sigma_1...\sigma_k$ of length k=1..L, which are found (*anywhere*) within the training sequences in sufficient quantities.
  - For each, ask whether the distribution of the next symbol after $\sigma_1\sigma_2...\sigma_k$ significantly differs from that after $\sigma_2...\sigma_k$.
  - If so, add node $\sigma_1...\sigma_k$, and all the nodes on its path from the root, into the PST.
- Finally, smooth vectors such that no prediction yields probability zero.

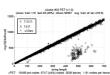E.g., b is the only predictive suffix extension to ra here:

brad......drad..........
..........crab......drab....
rab..................brad...

Empirical Results

In [**Bejerano & Yona, Bioinform.'01**] PST models are used to reconstruct all Pfam (v1.0) families from Swissprot (v33).

Pfam is a database of profile HMMs trained in a semi-automatic manner, from quality seed alignments, to model protein domain families.

The new method outperformed gapped-BLAST and was comparable to the curated Pfam HMMs as well as SAM (v2.2) and HMMER (v2.1).

Performance of a PST for the Neuro-transmitter-gated ion-channels, plotting the minus log likelihoods of all Swissprot sequences vs. their lengths. O marks family seed members, X previously unseen family members, and + the rest of Swissprot. Clear separation can be seen between + and O,X .

Advantages over the HMM approach include:
- Faster training and prediction run-time ([**Apostolico & Bejerano, J. Comp. Biol.'00**] show how to implement PSTs in linear time).
- The lack of need for a multiple alignment or human supervision.
- Resistant to over-fitting related error (deeper nodes seldom in use).

Biological Motivation (II)

Observation: Proteins are composed of distinct organizational units, usually connected by relatively unstructured linker regions.

Definition: Protein domain – a subsequence of a protein, which can fold independently into a compact stable 3-D structure.

- Different domains are often associated with different functions.
- A typical domain is of length 50-350 amino-acids (aa), and a protein may have from one up to several dozen domains.
- The domains are usually the more conserved parts in a protein family.

Colors represent different domains in a single multi-domain protein.

Challenge: segment proteins into domains and group instances.

Sequence Segmentation

Observation: A single PST model for a protein family appears to distinguish between more and less conserved regions *within* family members, in correspondence to domains and linker regions, respectively.

Hypothesis: Can we separate this PST into several models, each capturing only a conserved segment (hopefully a domain) of the protein sequence?

Competitive Learning approach: Iteratively split the original PST such that its sons have to compete over the training data. The ones that better predict a segment get a bigger handle over it.

In [**Seldin, Bejerano & Tishby, Icml'01**] we embed this approach in a Deterministic Annealing framework that tries to infer the correct number of underlying PST models. This is done by gradually increasing a resolution parameter β governing the hardness of data assignment to models.

PST prediction, symbol by symbol, over a family member with two distinct domains, a Zinc Finger C4 domain (aa 79-154) and a ligand binding domain (aa 190-413).

Segmentation Algorithm Outline

Input: Set of unaligned sequences.
- Grow a single PST over the data.
- Split each PST into two replicas and perturb copies anti-symmetrically.
- Repeat until convergence:
  - Repartition (soft) the data prop. to the performance of the new models.
  - Retrain the models on the data given its new reweighing.
- Remove emptied models, and split.
- When the number of models stops growing, increase β.
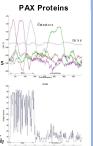- Terminate when β reaches β_fin.

Output: Resulting PST models.

Protein Sequence Segmentation: Domain Boundaries Determination

In [**Bejerano et al., Bioinform.'01**] we apply the novel segmentation alg. to groups of multi-domain proteins. Each group shares a single domain. We find that in the protein context two types of models result:

Detector models - peek in specific regions and perform below average elsewhere.
Noise models - perform averagely over *all* the data. These are easily discarded.

Right: Top - the unsupervised segmentation of a protein induced by our models, compared to the correct domain assignment shown above the X axis. Bottom - the same protein segmented by an alignment-based method (Clustal X). The second, homeobox domain, appearing in only 1/2 the family members is found above, and lost in the noise below.
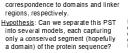
PAX Proteins

Refinement of an HMM classification
Glutathione S-Transferases

- Five classes of GST proteins are known: alpha, mu, pi, sigma, theta.
- Sequence similarities between the different classes are very high. In particular the sigma and theta classes are not well defined.
- S-crystallin refractory lens proteins probably lack the GST catalytic activity but show a high degree of sequence similarity to the GSTs.
- Pfam (v5.4) contains a single HMM model for all GST classes and the S-crystallins together, as class dissimilarities are too subtle for it.
- Our algorithm yields four additional signatures on this group, of:
  1. S-crystallins.
  2. Alpha + pi classes.
  3. Putative theta sub-class.
  4. All 12 out 396 unaligned sequence (3%) where the GST domain is followed by the elongation factor 1 gamma (EF1G) domain.

Detection of a Protein Fusion Event
DNA Topoisomerase II

Above - schematic description of the relationship between a protein in higher organisms (1) and the two proteins that apparently fill together the same role in lower organisms (2,3). Right - results of segmenting a heterogenous set containing all known proteins belonging to the groups of (1-3). Bottom left - signatures of the two domains in (2). Bottom right - the signature of the domain in (3). Top - A representative of the domain in (1) showing the correct signatures, one from (2) and one from (3).

Future Directions
- Analysis of the relationship between protein fold & function and the acquired statistical signatures.
- Application to DNA sequence analysis.

Bibliography
1. Ron, Singer & Tishby (1996). *Machine Learning* 25, 117-149.
2. Bejerano & Yona (2001). *Bioinformatics* 17(1), 23-43.
3. Apostolico & Bejerano (2000). *Journal of Computational Biology* 7(3/4), 381-393.
4. Seldin, Bejerano & Tishby (2001). *Proc. 18th International Conference on Machine Learning (ICML-01)*, 513-520.
5. Bejerano, Seldin, Margalit & Tishby (2001). *Bioinformatics* 17(10), 927-934, and *3rd Georgia Tech - Emory International Conference on Bioinformatics*.

Papers available via http://www.cs.huji.ac.il/~jill