Poster H-55

HMM-based subsequence feature map for Protein Classification and Remote Homology Detection

Authors:
Omer Sinan Sarac (*Dept. of Computer Engineering, Middle East Technical Univeristy*)
Volkan Atalay (*Dept. of Computer Engineering, Middle East Technical Univeristy*)
Rengul Cetin-Atalay (*Dept. of Molecular Biology and Genetics, Bilkent Univeristy*)

Short Abstract: This study represents a feature mapping for protein sequence classification that uses the models of conserved regions in proteins. An kmeans-like algorithm with an HMM map is used to cluster the subsequences and learn the models of these regions from a given family of proteins in an unsupervised manner.

Long Abstract:
HMM-based subsequence feature map for Protein Classification and Remote Homology Detection

1. Introduction

Classification of proteins into functional or structural classes according to their primary sequences is a fundamental problem in computational biology. Standard alignment-based methods for classification and homology detection perform poorly when there is no significant sequence similarity. Recently, discriminative methods that explicitly model the differences between positive and negative examples were shown to be more accurate when sequence similarity is low (Jaakkola et al., 2000; Li and Noble, 2003; Hou et al., 2003; Ben-hur and Brutlag, 2003; Leslie et. al, 2004). Most of these methods use support vector machines (SVMs) combined with an appropriate string kernel or feature space mapping. The main bottleneck/issue in classification of proteins according to their primary sequences is to find a kernel or a feature mapping that captures the information hidden in the important discriminative regions of the given sequences. Since, functionally important regions (catalytic sites, binding sites, structural motifs) are conserved over much wider taxonomic distances than the sequences themselves, conserved subsequences among different protein sequences are strong indicators of functional or structural similarity. This idea is especially important in remote homology detection and protein classification tasks and exploited in the literature (Leslie et al., 2004; Ben-hur and Brutlag, 2003).

We propose a feature map that takes into account the information coming from the subsequences of a protein. We decompose given sequences into fixed-length subsequences and cluster similar subsequences. A mapping can then be defined as the distribution of the subsequences of a new sample over these clusters. This study is an extension over the P2SL system which obtained results comparable or better than the existing methods for subcellular localization (Atalay and Cetin-Atalay 2005). P2SL used fixed length predefined partitioning of protein sequences to train a self-organizing feature map. Here, we make use of a set of Hidden Markov Models (HMMs) to model subsequences of a given family of

proteins. In contrast to profile-HMM methods (Krogh et al., 1994) that attempt to find a single HMM which is capable to generate whole sequences of a family of proteins, we have multiple HMMs each of which is sensitive to different short subsequences.

2. Method

A vector of generative models is used as a feature map. In the feature space, each sequence is represented by the distribution of its subsequences over these generative models. A single model is a fully connected HMM of 20 states, one state for each amino acid. Observation probabilities are set fixed using an amino acid similarity matrix to incorporate biochemical properties of amino acids. In this way, we are able to model biologically acceptable substitutions in the protein sequences. Transition probabilities are initialized randomly and they are to be learned in an unsupervised manner. Main focus here is not the subsequences themselves but the HMM models that would generate them. HMMs are ergodic and do not impose any length on the subsequences so setting a different length l for subsequences do not effect the model of the HMM. This approach takes into account all the subsequences of a protein whether it is highly conserved or not. The assumption here is that, most discriminative subsequences among a family of proteins are more likely to be frequent than irrelevant subsequences when we feed a set of protein sequences to the model. One can feed protein sequences from several different classes or families. Given enough representative for each class, discriminative subsequences of each class are expected to dominate during the training of the feature map. Once trained successfully, this feature map can be used with any numerical classification tools such as SVM for any one of the classes fed as training data.

The training of the HMMs and the clustering of subsequences of the proteins in the training set are performed simultaneously with a k-means like algorithm. Subsequences are assigned to K HMMs of the map according to their likelihood values. A subsequence is assigned to the HMM that produces the highest likelihood value. The parameters of the HMMs are, then, updated with the assigned subsequences using the Baum-Welch algorithm. This approach is similar to the well-known k-means clustering algorithm only this time the cluster centers are defined as HMMs. After the training of the K HMM models, a K dimentional feature space representation of a protein sequence can be defined as the summation of the likelihood values of its subsequences produced by its assigned HMM. Attention must be paid to the training phase of the HMMs because the convergence of the algorithm to a local optimal clustering is not guaranteed. In the standart k-means algorithm with the Euclidian distance measure, the surface defined by the points that have the same distance to a specific point is convex. In the probability space defined by HMMs however, such a surface is not necessarily convex.

3. Results and Future Work

In order to test the performance of the proposed method, we prepared an artificial dataset that consist of 400 training sequences that belong to two different classes and 100 test sequences for each class. First we chose two sets among GPCRMGR motifs extracted from PRINTS database. Each set is associated with one of the classes and each set has two different types of motifs. Members of a class are random amino acid sequences with one or both of the associated motifs embedded in random positions. Motif lengths are ranging from

13 to 30, with 0 to 8 possible mutations in each sample. Lengths of the final sequences are ranging from 130 to 220. We observed that the mapping combined with a SVM perfectly classified the test samples.

The next step is to test the algorithm with SCOP benchmark dataset and then, to apply to real life problems. The generated HMM models should also be analyzed if they became models for discriminative subsequences for a specific family.