Protein function prediction using sequence motifs: 
A research proposal

Asa Ben-Hur

Abstract

Protein function prediction, i.e. classification of protein sequences according to their biological function is an important task in bioinformatics. We propose a study of function prediction that uses protein sequence motifs as features; the result will be a system with state-of-the-art performance in function prediction, that provides interpretable results that can help in providing biological insight on the relationships between different classes of proteins. As a proof of concept we will focus on enzymes, which have a well developed system of classification, namely the Enzyme Commission (EC) numbering system.

1 Background

Advances in DNA sequencing are yielding a wealth of sequenced genomes. And yet, understanding the function of the proteins coded by a specific genome is still an unsolved problem. The determination of the function of genes and gene products is performed mainly on the basis of sequence similarity (homology) [1]. Comparing entire proteomes to a database of known sequences leaves the function of a large percentage of genes undetermined: close to 40% of the known human genes do not have a functional classification by homology [2, 3]. Most homology-based methods depend on either global similarity to a known, closely related protein, or homology to a family of known proteins using BLAST, PSI-BLAST, profiles, or HMM methods [4, 5, 6]. The function of genes that have arisen recently by exon shuffling or with no closely related ortholog remains undetermined by these approaches [7].

Motifs on the other hand, represent short, highly conserved regions of proteins. Very often these motifs correspond to functional regions of a protein – catalytic sites, binding sites, structural motifs etc. The presence of such protein motifs often reveals important clues to a protein’s role even if it is not globally similar to any known protein. The motifs for most catalytic sites and binding sites are conserved over much larger taxonomic distances and evolutionary time than the rest of the sequence. However, a single motif is often not sufficient to determine the function of a protein. The catalytic site or binding site of a protein might be composed of several regions that are not contiguous in sequence, but are close in the folded protein structure (for example in serine proteases). In addition, a motif representing a binding site might be common to several protein families that bind the same substrate. Therefore, a pattern
of motifs is required in general to classify a protein into a certain family of proteins. Manually constructed fingerprints are provided by the PRINTS database [8]. In this proposal we suggest an automatic method for the construction of such fingerprints, and to do this in a discriminative manner using tools of supervised learning.

Most motif methods extract conserved elements (blocks) from multiple sequence alignments of groups of proteins. These conserved elements are then represented as either discrete sequence motifs, Position Specific Scoring Matrices (PSSMs), or profiles. In this proposal we focus on discrete ungapped motifs and PSSMs. There are several databases of blocks/motifs that are publicly available, including PROSITE, BLOCKS+, PRINTS, and eMOTIF [9, 10, 8, 11, 12]. The use of a motif database as the basis for constructing a classifier will enable us to compare the merit of different databases and methods for constructing motifs, according to how well the resulting classifier performs.

2 Enzyme Classification

Enzymes represent a substantial fraction of the proteins in the SwissProt database [13], and have a well established system of annotation. The current version of SwissProt contains over 35,000 enzymes. The function of an enzyme is specified by a name given to it by the Enzyme Commission (EC) [14]. The name corresponds to an EC number, which is of the form: n1.n2.n3.n4, e.g. 1.1.3.13 for alcohol oxidase. The first number is between 1 and 6, and indicates the general type of chemical reaction catalyzed by the enzyme; the main categories are oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The remaining numbers have meanings that are particular to each category. Consider for example, the oxidoreductases (EC number starting with 1), which involve reactions in which hydrogen or oxygen atoms, or electrons are transferred between molecules. In these enzymes, n2 specifies the chemical group of the (electron) donor molecule, n3 specifies the (electron) acceptor, and n4 specifies the substrate. The EC classification system specifies over 750 enzyme names, and a particular protein can have several enzymatic activities. Therefore, as a machine learning classification problem, it is not a standard multi-class problem, since each pattern can have more than one class label; this type of problem is sometimes called a “multi-label” problem [15]. However, in preliminary experiments we performed, we found that this multi-label problem can be reduced to a regular multi-class problem by considering a group of enzymes that have several activities as a class by itself; for example, there are 22 enzymes that have EC numbers 1.1.1.1 and 1.2.1.1, and these can be perfectly distinguished from enzymes with the single EC number 1.1.1.1 using a classifier that uses the motif content of the proteins. When looking at the SwissProt database we then found that these two groups are indeed recognized as distinct.
3 Methods

We propose to use the motif composition of a protein to define a similarity measure or “kernel” function that can be used with various kernel based classification methods such as Support Vector Machines (SVM). In a recent study we found that a “bag of motifs” representation of a protein sequence provides state of the art performance in detecting remote homologs [16]. The motif database we plan to use contains over a million discrete sequence motifs [12]. The motif composition vector of a protein sequence is therefore high dimensional, and is also very sparse, much like the bag of words representation of text documents. However, unlike textual data, motifs are highly predictive features: We found that using methods such as the Recursive Feature Elimination (RFE) [17], we could reduce the number of motifs to a few tens at the most, while keeping the same level of classification accuracy. A classifier based on a handful of features has the advantage of interpretability, and can provide biological insight on what differentiates different classes of enzymes from each other. During the development of the system we plan to benchmark various classifiers and feature selection methods.

Recall that protein function prediction is a classification problem with a large number of classes (hundreds of classes in the EC classification alone). This poses a computational challenge when using a two-class classifier such as SVM: the one-against-one method, that appears to perform better than the one-against-rest method of multi-class classification [18], can be too computationally intensive for this application. However, in view of the sparsity of the data in the motif representation, for a given query, only a small number of classes are likely to show any degree of similarity. So in practice only a small number of candidate classes will need to be discriminated against each other.

The motifs in the database we use can be highly redundant in their pattern of occurrence in a group of proteins. Due to this level of redundancy simple feature selection methods that are based on ranking individual features, cannot be used in order to find a small subset of motifs. Wrapper methods such as RFE were checked to be effective in reducing the dimensionality, while maintaining the accuracy of the resulting classifier.

Preliminary results of our experiments will be presented at the workshop on Feature Selection at NIPS2003.

4 Deliverables

- A suite of tools for classification of proteins sequences will be provided as modules for the PyML machine learning environment (current version available at http://cmgm.stanford.edu/~asab/pyml.html).

- A web based interface for performing protein classification. For a given query the tool will show the component motifs in the query, the motifs that contributed to the classification of the motif and the pattern of occurrence of those motifs across different enzyme classes.
References


