Protein Sequence Classification Through Relevant Sequence Mining and Bayes Classifiers

Pedro Gabriel Ferreira* and Paulo J. Azevedo**

University of Minho,
Department of Informatics,
Campus of Gualtar, 4710-057 Braga, Portugal
{pedrogabriel, pja}@di.uminho.pt

Abstract. We tackle the problem of sequence classification using relevant subsequences found in a dataset of protein labelled sequences. A subsequence is *relevant* if it is frequent and has a minimal length. For each query sequence a vector of features is obtained. The features consist in the number and average length of the relevant subsequences shared with each of the protein families. Classification is performed by combining these features in a Bayes Classifier. The combination of these characteristics results in a multi-class and multi-domain method that is exempt of data transformation and background knowledge. We illustrate the performance of our method using three collections of protein datasets. The performed tests showed that the method has an equivalent performance to state of the art methods in protein classification.

1 Introduction

Concerning data where an order relation between atomic elements occurs, sequence data appears as a natural representation. An important and very useful operation to be done over sequence data is classification. The problem of classifying sequence data is to take a given set of class labelled sequences and build up a procedure to *a posteriori* assign labels to unlabelled sequences (queries). Many examples of the application of this task can be found in a variety of domains. Consider the case of biology/bioinformatics field where given a database of nucleotide sequences (DNA/RNA) or amino-acids sequences. Portions of the former sequences code for the latter through two mechanisms: *transcription* and *translation* [12, 6]. A sequence of amino-acids constitute a protein and is hereafter called as a protein sequence. A possible scenario would be the case where a biologist wants to find the respective family/domain or function of an unclassified sequence, for example a new synthesized protein. This problem is of critical

* Supported by a PhD Scholarship from Fundação para Ciência e Tecnologia, Ministério da Ciência e Ensino Superior of Portugal.
** Supported by POSI/2001/CLASS project, sponsored by Fundação Ciência e Tecnologia and European program FEDER.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 236–247, 2005.

importance due to the exponential growth of newly generated sequence data in the last years, which demands for automatic methods for sequence classification. In the problem of sequence categorization/classification three types of methods can be distinguished:

- The *Direct Sequence Classifiers*, that exploit the sequential nature of data by directly comparing the similarity between sequences. Example of these type of classifiers is the k-Nearest Neighbour. In this method the class label of the k most similar sequences in respect to the query sequence are used to vote on a decision. Sequence similarity can be assessed through a method like FASTA [17] or BLAST [1].
- The *Feature based Sequence Classifiers*, that work by first extracting and model a set of features from the sequences and then adapt those features to accomplish with the traditional techniques, like decision trees, rule based classifiers, SVM's and many others. In [15, 16, 5, 4, 21] we have examples of these type of methods.
- The *Probabilistic Model Classifiers*, that work by simulating the sequence family under consideration. Typical probabilistic classifiers are the simple and k-order Markov Chain [19], Hidden Markon Models [14] and Probabilistic Suffix Trees [11].

Recently Probabilistic Suffix Trees (PSTs) [11] and Sparse Markov Transducers (SMTs) [8] have been applied in the protein classification problem, and have shown superior performance. A PST is essentially a variable length Markov Model, where the probability of a symbol in a sequence depends on the previous symbols. The number of previous considered symbols is variable and context dependent. The prediction of an input sequence is done symbol by symbol. The probability of a symbol is obtained by finding the longest subsequence that appears in the tree and ends just before the symbol. These probabilities are then combined to determine the overall probability of the sequence in respect to a database of sequences. One of the disadvantages of the PSTs is that the conditional probabilities of the symbols rely on exact subsequence matches. In protein family classification this becomes a limitation since substitutions of symbols by equivalent ones is often very frequent. SMTs are a generalization of PSTs that support wild-cards. A wild-card is a symbol that denotes a gap of size one and matches any symbol on the alphabet. In [11] an experimental evaluation has shown that PSTs perform much better than a typical BLAST search and as good as HMM. This is very interesting since the latter approach makes use of multiple alignments and the families are usually defined based on an HMM [9]. Additionally PSTs are a totally automotive method without prior knowledge (multiple alignments or score matrices) or any human intervention. In [8], SMTs have shown to outperform PSTs.

Our motivation to this work is to suggest a robust and adaptable classification method using a straightforward algorithm. We propose a multi-class sequence classification method which can be applied to data in many different domains, in particular to protein sequence data without requiring any type of data transfor-