Understanding Signal Sequences with Machine Learning

Jean-Luc Falcone[1,⋆], Renée Kreuter, Dominique Belin[2], and Bastien Chopard[1]

[1] Département d'informatique, Université de Genève, 1211 Genève 4, Switzerland
[2] Département de Pathologie et d'Immunologie, Université de Genève, Switzerland

Abstract. Protein translocation, the transport of newly synthesized proteins out of the cell, is a fundamental mechanism of life. We are interested in understanding how cells recognize the proteins that are to be exported and how the necessary information is encoded in the so called "Signal Sequences". In this paper, we address these problems by building a physico-chemical model of signal sequence recognition, using experimental data. This model was built using *decision trees*. In a first phase the classifier were built from a set of features derived from the current knowledge about signal sequences. It was then expanded by feature generation with *genetic algorithms*. The resulting predictors are efficient, achieving an accuracy of more than 99% with our wild-type proteins set. Furthermore the generated features can give us a biological insight about the export mechanism. Our tool is freely available through a web interface.

1 Introduction

1.1 Signal Sequences

Proteins synthesized in the cell must be transported to the correct cellular compartment so that they can achieve their role. This process is called protein targeting and is a fundamental aspect of cell protein metabolism [1]. For instance blood plasma proteins and polypeptidic hormones must be delivered to the extracellular space. We are interested in the secretion pathway, which involves the targeting and transport of the proteins out of the cell. The protein complex (called *translocon*) which exports the proteins varies from one species to another.

All the proteins that must be exported, carry a particular region of conserved function, the *signal sequence* (SS) or *signal peptide*, located in N-terminal extremity. The length varies slightly from 10 to 50 amino-acids (AA). The protein is exported before folding and the SS is usually cleaved after the export. The precise location where the cleavage occurs is called the cleavage site.

The most interesting feature of SS is their inter- and intra-species variability. Their sequence as well as their length vary. Thus, they do not carry any systematic consensus. However, three properties have been proposed as distinguishing

⋆ Supported by the Swiss National Science Foundation.

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 57–67, 2007.

features of SS [2]: (i) They begin with an N-terminal region which includes one or several positively charged lysine or arginine residues. This region is called the *N-Region*. (ii) Following the N-Region, SS contain a stretch of hydrophobic AA forming the so-called *H-Region*. (iii) In the majority of secreted proteins there is a third region, the *C-Region*, located between the H-Region and the cleavage site. It carries a weak consensus recognized by the leader-peptidase.

The above properties are too vague to easily determine whether or not a protein will be secreted. The hypothesis that these three regions, as defined above, characterize an exported protein is based on observations made on known SS. However there are no experiments that confirm that these three properties are sufficient and/or relevant for the recognition process. To address the problem of correctly discriminating secreted proteins from the other ones (cytosolic), artificial intelligence techniques have been considered.

1.2 Computer Predictions of Signal Sequences

Many methods for the recognition of SS have been proposed. For all these methods, the predictors are built by an algorithm based on supervised learning techniques. The SignalP method is currently considered as the best classification method [3] and is the most widely used. It consists of two feed-forward neural networks [4]. The first is trained to recognize the SS itself; the second recognizes the cleavage site.

SignalP cannot help us understand the physico-chemical properties recognized by the translocon. The problem is intrinsic to the nature of classical neural networks which does not allow the retrieval of high-level symbolic rules. Another weakness resides in the fact that SignalP uses the existence of a valid cleavage site as a strong classification criterion. Although it is true that most SS include such a site, there are proteins like ovalbumin which are exported but lack a cleavage site. Furthermore mutated SS are poorly recognized by existing predictors. Therefore there is a need to develop new approaches with better prediction scores on such proteins and which gives better insight into the mechanisms at work.

1.3 Decision Trees

In this paper we propose a novel approach based on *decision tree* classifiers to understand how SS are recognized. Decision trees are classification programs that can classify objects according to properties (or features). Different algorithms exist to build such trees from a list of properties and a set of example objects representative of the different classes. As it is the case with neural networks, the learning strategy is supervised, i.e. based on a training set of sequences. The output of this algorithm is a tree in which the non-terminal nodes are evaluations based on the properties characterizing the objects being classified. The leaves of the tree (the terminal nodes) are the possible classes.

An important advantage of decision tree building algorithms is that only the properties necessary for the classification are retained. The most discriminant properties appear at the root of the tree and the less discriminant ones are near