Identification of Transmembrane Domains by Decision Trees over Regular Patterns

Setsuo Arikawa[†] Satoru Kuhara[‡] Satoru Miyano[†] Ayumi Shinohara[†]
Takeshi Shinohara[††]

As a primary structure, a protein is expressed as a sequence from twenty symbols called amino acids. We propose a new systematic method for identifying a class of amino acid sequences based on their features as strings and report the performance of our method for transmembrane domain identification.

A regular pattern is a string of the form $\alpha_0 x_1 \alpha_1 \cdots x_n \alpha_n$, where $x_1, ..., x_n$ are distinct variable symbols and $\alpha_0, ..., \alpha_n$ are strings over constant symbols. It defines strings obtained by substituting arbitrary strings of constant symbols to the variables. Although regular patterns themselves can be used to describe some features of amino acid sequences, in some cases it is impossible for any single regular pattern to classify amino acid sequences precisely. Our classification method uses a decision tree whose internal nodes are labeled with several regular patterns. For the construction of a decision tree, we use the idea of Quinlan's ID3 algorithm since experiences tell us that the ID3 algorithm usually constructs a small size tree although no mathematical performance evaluation has not yet been given.

Using the database PIR, we made experiments with protein data on our system. The experiments show that the idea of combining decision trees and regular patterns is quite successful. When we use an indexing method known as hydropathy index, our system produced some decision trees with drastic performance. For example, it found, from randomly chosen 10 positive and 10 negative examples, a decision tree with only a few nodes whose accuracy is more than 90% for 689 positive data and 19256 negative data. From a view point of searching motifs, it should be also mentioned that our system made a new discovery that suggests the importance of negative examples.

As a theoretical foundation, we establish a relation to PAC-learnability. We define the class $DTRP(d, k)$ of languages defined by decision trees of depth at most d over regular patterns with at most k variables. We prove that $DTRP(d, k)$ is polynomial-time learnable for any fixed d, $k \geq 0$. Our learning algorithm for this class finds a minimum depth decision tree and runs in polynomial time. But the running time is exponential with respect to the constants d and k and cannot be used for practical use. It supplies a theoretical evidence.

[†] Research Institute of Fundamental Information Science, Kyushu University 33, Fukuoka 812, Japan.
[‡] Graduate School of Genetic Resources Technology, Kyushu University 46, Fukuoka 812, Japan.
[††] Department of Artificial Intelligence, Kyushu Institute of Technology, Iizuka 820, Japan.