HAKKE: Automatic Predictor Generator for Sequences

Naohiro Furukawa [1] Takayoshi Shoudai [2] Ayumi Shinohara [3] Satoru Miyano [3]
{furukawa, shoudai, ayumi, miyano}@rifis.kyushu-u.ac.jp

[1] Department of Information Systems, Kyushu University, Kasuga 816, Japan
[2] Department of Physics, Kyushu University, Fukuoka 810, Japan
[3] Research Institute of Fundamental Information Science, Kyushu University, Fukuoka 812, Japan

1 Introduction

Databases of DNA and amino acid sequences compile many sequences where some regions of specific functions or structures are indicated as segments. For example, α-helices, β-sheets, transmembrane domains and signal peptides are specified by positions on the sequences. In this paper we call such specified regions as *marked regions*.

For each family of sequences with marked regions of a specific type, some programs have been devolved for predicting such marked regions on unknown sequences. For example, Chou-Fasman's method [2] is well known for predicting the α-helices and β-sheets. The hydropathy plot of Kyte and Doolittle [3] predicts transmembrane domains on amino acid sequences of membrane proteins. Instead of designing such programs independently for each family of sequences, we have developed an automatic predictor generator HAKKE by using the knowledge discovery system BONSAI [1, 4] which finds hypotheses from positive and negative examples of sequences. Given a sample collection of sequences with marked regions, HAKKE produces automatically a C program called a *predictor* that may predict marked regions on unknown sequences.

2 Method

BONSAI is a knowledge discovery system for DNA and amino acid sequences by exploiting the PAC-learning theory and the local search technique. BONSAI receives positive and negative examples as inputs and produces a pair of an alphabet indexing and a decision tree over regular patterns as a hypothesis. An alphabet indexing is a mapping from the alphabet describing sequences to a smaller alphabet. A decision tree over regular patterns is a decision tree whose

[1]古川　直広　　：九州大学総合理工学研究科情報システム学専攻、〒816 春日市春日公園 6–1
[2]正代　隆義　　：九州大学理学部物理学科、〒810 福岡市中央区六本松 4–2–1
[3]篠原　歩，　宮野　悟　：九州大学理学部基礎情報学研究施設、〒812 福岡市東区箱崎 6–10–1

internal nodes are labeled with regular patterns for classification. The details are found in [1, 4]. According to computational experiments on transmembrane domains and signal peptides, BONSAI has succeeded in discovering very useful and simple hypotheses with high accuracy.

HAKKE mainly consists of the following two stages: First, HAKKE creates from positive and negative examples for BONSAI from a collection F of sequences with marked regions. A positive example is a marked region of a sequence in F. A negative example is a sequence which has no overlap with any marked regions. The length of a negative example is set to be about the average length of all positive examples. HAKKE runs BONSAI on these positive and negative examples for finding a hypothesis (T, ψ) of a decision tree over regular patterns and an alphabet indexing that explains the positive and negative examples with high accuracy.

The second stage of HAKKE is to create a predictor by using the hypothesis (T, ψ). The predictor has four parameters w (*window size*), t (*threshold*), w_p (*positive weight*) and w_n (*negative weight*). Given a sequence $x = x[1] \cdots x[m]$, the predictor scans x from left to right through the window of size w, i.e., $x_i = x[i] \cdots x[i+w-1]$ for $1 \leq i \leq m - w + 1$. For each x_i, if (T, ψ) accepts x_i, then w_p is added to the positions of $x[i], \ldots, x[i+w-1]$. Otherwise, w_n is added. The predictor marks the positions with value greater than or equal to the threshold t that constitute the predicted marked regions. HAKKE searches these parameters so that the accuracy of the predictor attains the highest value for the collection F. Then HAKKE produces a C program implementing this predictor.

"HAKKE" is a kind of divination which comes from the principles of Yin and Yang (the positive and negative principles) and the system was named by following a famous proverb about HAKKE "the prediction may or may not come true"

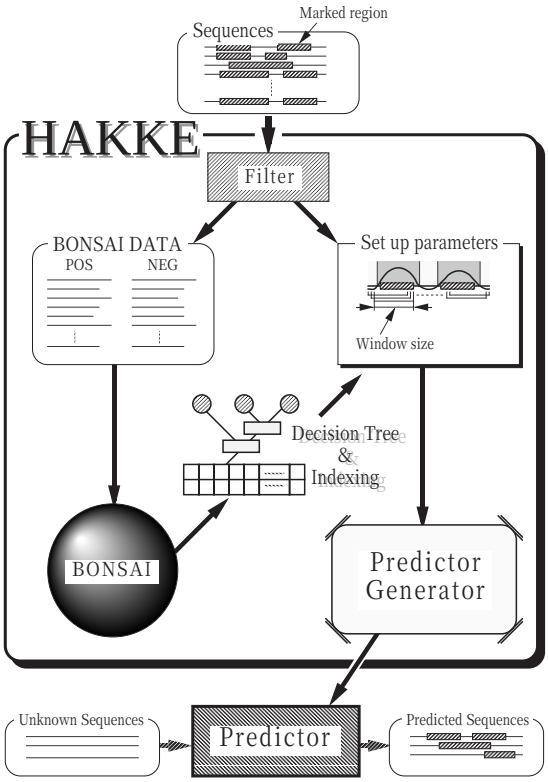

Fig 1. HAKKE

References

[1] S. Arikawa, S. Miyano, A. Shinohara, S. Kuhara, Y. Mukouchi and T. Shinohara, "A machine discovery from amino acid sequences by decision trees over regular patterns", *New Generation Computing*, **11**, pp. 361–375, 1993.

[2] P. Y. Chou and G. D. Fasman, "Prediction of the secondary structure of proteins from their amino acid sequence", *Advances in Enzymology*, **47**, pp. 45–147, 1978.

[3] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein", *J. Mol. Biol.*, **157**, pp. 105–132, 1982.

[4] S. Shimozono, A. Shinohara, T. Shinohara, S. Miyano, S. Kuhara and S. Arikawa, "Knowledge acquisition from amino acid sequences by machine learning system BONSAI", *Trans. Information Processing Society of Japan*, **35**, pp. 2009–2018, 1994.