randomly partitioned into m sets $POS_0, \ldots, POS_{m-1}$ of the same size. BONSAI Garden is initialized by starting B_i on (POS_i, NEG) with $TRASH_i$ empty for $0 \leq i \leq m-1$. When a hypothesis (T_i, ψ_i) is produced by B_i, the sequences in POS_i misclassified by (T_i, ψ_i) shall be put into $TRASH_i$. POS_i is updated with the positive examples correctly classified by (T_i, ψ_i). Each BONSAI repeats this process under the control of *Gardener*.

The *Gardener* of BONSAI Garden is watching the behaviors of BONSAI's $B_0, \ldots, B_{m-1}$ and executes the following task. *Gardener* will find two BONSAI's, say B_i and B_j, which have finished one job cycle with hypotheses (T_i, ψ_i) and (T_j, ψ_j) and determine which hypothesis is smaller. Suppose that (T_i, ψ_i) is smaller than (T_j, ψ_j). Then *Gardener* will move the positive examples in POS_j accepted by both hypotheses to POS_i. Thus BONSAI with a smaller hypothesis will get more sequences while BONSAI with a larger hypothesis will loose sequences. Another task of *Gardener* is to distribute the positive examples in the trash boxes to $POS_0, \ldots, POS_{m-1}$ in some way.

3 Experiments by BONSAI Garden

Experimental results by BONSAI Garden under various conditions will be presented at the workstation session. We just enumerate the items of data below:

Signal peptides: As positive examples, the signal peptide sequences beginning with a Methionine and of length at most 32 are collected. For the negative examples, we take N-terminal regions of length 30 obtained from complete sequences that have no signal peptide and begin with a Methionine. In total, 3,796 positive examples and 27,952 negative examples are collected from GenBank.

Eukaryotic promoters: A positive example is a DNA sequence of length 200 locating in the upstream of the transcription site. A negative example is a DNA sequence of length 200 which encodes a protein. In total, 1,170 positive examples and 15,611 negative examples are collected from EMBL. *E. Coli* promoters are also collected in the same way.

Random sequences: Randomly generated amino acid sequences of length from 35 to 45 are used as positive and negative examples. The number of positive examples is 5,000 and the number of negative examples is 20,000.

The data on α-helices and ORF's are also collected for BONSAI Garden for experiments. The details are omitted.

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] 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.

[3] T. Shoudai, M. Lappe, S. Miyano, A. Shinohara, T. Okazaki, S. Arikawa, T. Uchida, S. Shimozono, T. Shinohara, and S. Kuhara, "BONSAI Garden: parallel knowledge discovery system for amino acid sequences," *Proc. Third International Conference on Intelligent Systems for Molecular Biology*, pp. 359–366, 1995.

Experiments by BONSAI Garden

Naohiro Furukawa [1] Erika Tateishi [1] Takayoshi Shoudai [2] Satoru Miyano [3]
{furukawa, tateishi, shoudai, 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

BONSAI [2] is a knowledge discovery system for DNA and amino acid sequences. From positive and negative examples of sequences, it finds a small hypothesis (T, ψ) of a decision tree T over regular patterns [1] and an alphabet indexing ψ [2] that classifies the positive and negative examples with high accuracy. The knowledge about the sequences is extracted as the hypothesis (T, ψ). The name of "BONSAI" comes from the fact that the knowledge (the nature) is expressed as a small tree T (a decision tree over regular patterns) in harmony with an alphabet indexing ψ (a container).

BONSAI has revealed its high ability in discovering knowledge from uniform sequences such as transmembrane domain sequences [2]. However, if data is a mixture of several kinds of sequences, i.e., a hodgepodge of sequences, BONSAI cannot show its full ability. BONSAI Garden [3] was designed to cope with such situation. The target of BONSAI Garden is that the system should be able to handle a hodgepodge of data and/or noisy data so that it classifies the data into some number of classes of sequences and simultaneously finds for each of these classes a hypothesis which explains the sequences in the class. This paper presents a brief overview of BONSAI Garden and describes computational experiments by BONSAI Garden scheduled for the workstation session.

2 BONSAI Garden

BONSAI Garden consists of a coordinator program called a *Gardener* and some number of BONSAI's. The *Gardener* and BONSAI's run over a network in parallel.

Let POS and NEG be the collections of positive and negative examples. BONSAI Garden with m BONSAI's runs as follows: Let B_i be the ith BONSAI for $0 \leq i \leq m-1$. B_i is equipped with a trash box $TRASH_i$ which shall contain misclassified positive examples. First, POS is

[1]古川　直広，立石恵利佳: 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、 〒812 福岡市中央区六本松 4–2–1
[3]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