Hypothesis Creator for Complete Genome

Osamu Maruyama [1] maruyama@ims.u-tokyo.ac.jp
Hirotaka Seki [2] seki@ims.u-tokyo.ac.jp
Tomoyuki Uchida [3] uchida@cs.hiroshima-cu.ac.jp
Takayoshi Syoudai[2] shoudai@i.kyushu-u.ac.jp
Satoru Miyano[1] miyano@ims.u-tokyo.ac.jp

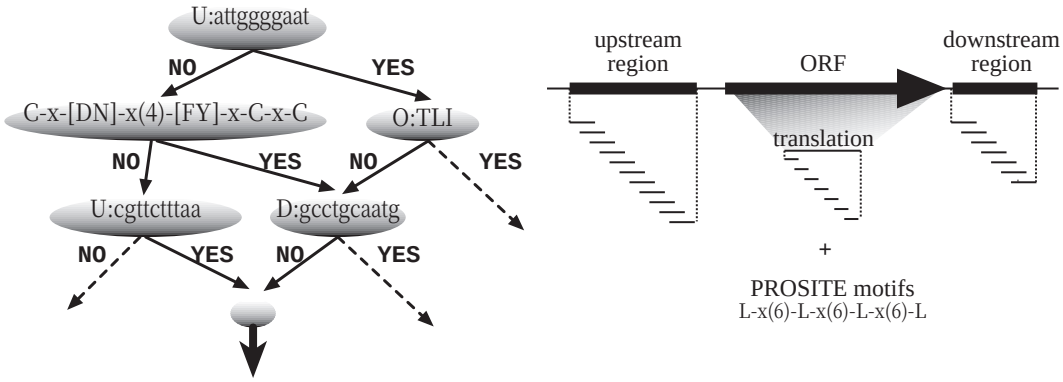
[1] Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

[2] Department of Informatics, Kyushu University 39
6-1 Kasuga-Kouen, Kasuga 816, Japan

[3] Faculty of Information Sciences, Hiroshima City University
151-5, Ozuka, Numata-Cho, Asa-Minami-Ku, Hiroshima 731-31, Japan

1 Introduction

We are developing a knowledge discovery system which creates a hypothesis explaining a collection of ORFs where the hypothesis is represented as a binary decision diagram (BDD) such that nodes of the BDD are assigned various kinds of patterns on DNA sequences and amino acid sequences (see Fig. ?? (a)). The collection of ORFs can be specified by a combination of key words, e.g. `adaptation`, `chaperones` and `transport`. Therefore it would be possible to find some relation between such a key word and patterns on sequences.



(a) Hypothesis represented by a BDD (b) View on an ORF

Figure 1:

The complete DNA sequences of many organisms are being determined by international collaborations, and some of them have already been made public at various ftp sites. For example, the *Escherichia coli* K-12 sequence is compiled with its annotations, which is available at the web site `http://www.genetics.wisc.edu/` [?]. Such a data, which our system utilizes, enable us to find more general and consistent rules which would help us understand *genome*.

2 Hypothesis Creator

The system first gathers ORFs related to given key words by applying text database management system SIGMA [**?**] to various databases including annotation files. Let C denote the collection of ORFs gathered. Next the system produces a BDD consistent with as many ORFs in C as possible where nodes of the BDD are labeled with attributes defined as follows:

2.1 Attributes

For each ORF x of a complete DNA sequence, we specify two regions neighboring on x called the *upstream region* and the *downstream region* of x, respectively (see Fig. **??** (b)). Let k be a constant number. Each subsequence s with length k of the upstream region is used as a pattern of an approximate pattern matching, which is a pattern matching with errors, that is, insertion, deletion and replacement [**?**]. We call s an upstream pattern. The *approximate pattern matching attribute for an upstream pattern s and an ORF x'* is defined as a function returning `YES` if s approximately matches the upstream of x' and `NO` otherwise. Note that users can specify the start and end positions of the upstream region and the subsequence length k to fit their purposes. In the same way we can define the approximate pattern matching attribute for a downstream pattern and an ORF.

We also define two kinds of attributes related to the coding region x. One is the approximate pattern matching between a subsequence of an amino acid sequence and the sequence translated x into. Another is an attribute employing PROSITE motifs patterns.

3 Method of Experiments

In the theoretical aspect of this research, we formulated the problem of creating hypotheses as *data mining for binary decision diagram rules* and showed that the problem is in general NP-complete. This means that there would be no polynomial-time algorithm to solve the problem. We then make BDDs in the following way: First one makes a decision tree T which would be consistent with the ORFs in C by the ID3 algorithm [**?**]. Then T is given to Bryant's algorithm [**?**], which reforms T into an equivalent BDD. As we are now in the process of preliminary experiment, we will report the experimental results at poster site.