HYPOTHESISCREATOR: A Framework for Knowledge Discovery Application Development

Yoshinori Tamada\textsuperscript{1} \hspace{1cm} Hideo Bannai\textsuperscript{2}
\texttt{tamada@hunter.ss.u-tokai.ac.jp} \hspace{1cm} \texttt{bannai@ims.u-tokyo.ac.jp}
Osamu Maruyama\textsuperscript{3} \hspace{1cm} Satoru Miyano\textsuperscript{2}
\texttt{om@math.kyushu-u.ac.jp} \hspace{1cm} \texttt{miyano@ims.u-tokyo.ac.jp}

\textsuperscript{1} Department of Mathematical Sciences, Tokai University, 1117 Kitakaname, Hiratuka-shi, Kanagawa 259-1292, Japan
\textsuperscript{2} Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
\textsuperscript{3} Graduate School of Mathematics, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka, 812-8581, Japan

\textbf{Keywords:} HYPOTHESISCREATOR, scientific knowledge discovery, view, viewscope, design

1 Introduction

HYPOTHESISCREATOR is a software library designed to assist the knowledge discovery process, by facilitating the development of knowledge discovery systems and computational knowledge discovery experiments. We demonstrate the use of HYPOTHESISCREATOR in developing computational experiments to analyze several kinds of genomic data, showing the diversity of HYPOTHESISCREATOR's applicability and its usefulness.

The implementation of HYPOTHESISCREATOR is based on the idea of viewscope \cite{3, 4}, a generalized form of attribute. The consideration of viewscope allows the seamless integration of various knowledge in databases, newly created attributes, and even experts' intuitions. For a typical user of HYPOTHESISCREATOR, the task consists of designing how to look at the data (the designing of viewscope), and then applying hypothesis generation algorithms to these viewscopes. The current implementation of the library provides numerous viewscopes and several hypothesis generation algorithms, which can be combined to create new viewscopes to suit the user's needs. If the ready-made viewscopes do not suffice, the implementation of new viewscope classes by the user is not so difficult since HYPOTHESISCREATOR is an object oriented class library.

2 Experiments

We are currently applying HYPOTHESISCREATOR to several different kinds of data. The details of the experiments will be reported elsewhere.
2.1 Finding gene regulatory sites

*Transcription factors* influence the transcription of particular genes by binding to specific sites in DNA sequences, called *transcription factor binding sites*. To find putative transcription factor binding sites, we have devised viewscopes for Boolean expressions over pattern matching viewscopes, which enables us to capture the combination of putative sites, and a naive Boolean expression generator. We have applied our approach to the DNA sequences of *S. cerevisiae* and *C. elegans*, and obtained some suggestive results.

2.2 Characterization of subcellular localization signals

Proteins are first synthesized in the cytosol and then carried to specified locations of the cell. The information that determines where the proteins are carried is believed to be encoded in the amino acid sequence. Mitochondrial targeting peptides (mTP), chloroplast targeting peptides (cTP), signal peptide, are some of the known targeting signals. Our aim differs from previous methods which try to characterize these signals [5, 1, 2] in that we try to extract a simple rule or knowledge which may be easily understood. In our experiments, we use an alphabet indexing combined with an approximate string pattern matching, as in [6]. Concerning mitochondrial targeting peptides, we have succeeded in discovering a simple rule which seems to capture biological knowledge about the signals.

3 Discussion

HYPOTHESISCREATOR is free software distributed under the GNU General Public License. It will be available from the HYPOTHESISCREATOR web site. The implementation of new modules into HYPOTHESISCREATOR by many others will further broaden its applicability.

References


---