Mining Association Rules from Signals found in Mammalian Promoter Sequences

Gen Shibayama gengen@ims.u-tokyo.ac.jp
Kenji Satou ken@ims.u-tokyo.ac.jp
Toshihisa Takagi takagi@ims.u-tokyo.ac.jp

Human Genome Center, Institute of Medical Science, The University of Tokyo
Shiroganedai, Minato-ku, Tokyo 108, Japan

Abstract

To find associations among large amount of genome data, we implemented a data mining algorithm developed by Houtsma et al. As the result of a computer experiment about signals found in mammalian promoter sequences, the system generated association rules with high accuracy and large coverage.

Introduction

In the research area of database, studies on *knowledge discovery* or *data mining* have been active in these years. Besides flexible and high-speeded search of large amount of genome data, such methods for automated knowledge acquisition from them are desired under the situation that they have been increasing exponentially.

In 1993, Agrawal et al.[1] proposed a novel framework of data mining and applied it to sales data obtained from a large retailing company. From customer transaction data, each of which consists of a set of items bought together in a transaction, their system generates all the significant association rules among items. For example, it might generate a rule which means *"If a customer buys bread and butter then the customer buys milk too in 90% of the cases."* As for application to genome data, this method is promising in the following two points: 1) low computation cost which implies high applicability to large amount of data, and 2) since examples are not needed, there is a possibility to point out unnoticed associations.

As the first step of applying this method to genome data, we tried to find associations among signals found in mammalian promoter sequences.

System and Methods

In carrying out this experiment, we adopted the set-oriented algorithm for mining association rules proposed by Houtsma et al.[4] which performs Agrawal's algorithm by iterating SQL

芝山元, 佐藤賢二, 高木利久: 東京大学医科学研究所ヒトゲノム解析センター, 〒108 東京都港区白金台 4-6-1

queries. An experimental system was implemented on SUN SPARCserver 690MP by using perl scripts and SYBASE with SQL interface.

To prepare the data for mining, 356 mammalian promoter sequences were chosen from 1250 sequences in Eukaryotic Promoter Database Release 43[2]. Signals on these promoter sequences were searched for by using Sigscan 3.0[5] with Ghosh's Transcription Factors Database 7.0[3]. As the result of scanning, 370 kinds of signals were found in the mammalian promoter sequences.

The importance of association rules are evaluated in two criteria called *support* and *confidence factor* in this framework of data mining. The former means the coverage of the association rule, and the latter means the accuracy of it in the coverage. The system receives minimum values on these criteria from a user, and uses minimum support for pruning. In this experiment, minimum support and minimum confidence factor were set to 200 sequences and 80%, respectively. As the result of mining, we obtained 505 association rules of the following form:

$$Conf\ \%\ (Sup\ \text{sp.}):\ SIG_1,...,SIG_{n-1} \Rightarrow SIG_n$$

where $Conf$, Sup and SIG_i $(1 \leq i \leq n)$ are confidence factor, support and signal name.

So far, we have implemented the algorithm and seen it works with realistic genome data. We are considering as the next step applying the algorithm to the specific domain of promoter sequences, that is, to restrict the promoter sequences to those of genes expressed in some specific tissue. It may reveal association rules of signals specific to such promoters. Further, we are considering to utilize such rules for predicting the location where a gene would be expressed. Another application of our system would be restricting promoters to those of genes expressed in some embryonic stage, which can help determine at which embryonic stage a gene would be expressed.

Acknowledgement

This work was supported in part by a Grant-in Aid for Scientific Research on Priority Areas, "Genome Informatics," from the Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Agrawal,R., Imielinski,T. and Swami,A.: *Mining Association Rules between Sets of Items in Large Databases*, ACM SIGMOD, pp.207-216 (1993).

[2] Bucher,P. and Trifonov,E.N.: *Compilation and analysis of eukaryotic POL II promoter sequences*, Nucl. Acids Res., 14, pp.10009-10026 (1986).

[3] Ghosh, D.: NAR, 18, pp.1749-1756 (1990).

[4] Houtsma,M. and Swami,A.: *Set-Oriented Mining for Association Rules in Relational Databases*, ICDE'95, pp.25-33 (1995).

[5] Prestridge,D.S.: *SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements*, CABIOS, 7, pp.203-206 (1991).