Prediction of Promoter Expression Specificity by Conserved Sequence Patterns

Wataru Fujibuchi[1]
wataru@kuicr.kyoto-u.ac.jp

Minoru Kanehisa[1]
kanehisa@kuicr.kyoto-u.ac.jp

[1]Institute for Chemical Research, Kyoto University
Uji, Kyoto 611, Japan

Abstract

We present here a prediction method for expression specificities of promoters by observing the appearance of conserved sequence patterns in a group of promoters, such as liver, brain, and housekeeping. Related promoters in the same group were compiled from EPD [1] database and an index to represent the group specificity of each pattern was calculated. Each promoter was examined for its specificity by the collection of these indices constructed from the rest of the promoters in our dataset. Currently, our system could discriminate 50 % of human liver promoters with 11% false positive rate. The distribution profile of scores also suggested that the liver promoter group may be divided into two or more subgroups.

1 System Overview

Homo sapiens promoters (-200 to -1) were collected from the EMBL nucleic acid database Release 41.0 according to the EPD entries. We obtained 191 independent (non-homologous) promoters, including 36 active promoters in the liver. Conserved sequence patterns were extracted based on the binomial distribution model, in which we can approximately calculate the probability P of finding a pattern K or more times. Given the number of sequences N, the probability of finding a pattern in one sequence p, and the ratio $a = K/N$, the following equation holds for $0 < p < a < 1$:

$$P \sim \frac{1}{1-r}\left(\frac{l}{\sqrt{2\pi a(1-a)N}}\right)e^{-NH},$$

where $H = a\cdot\log(a/p) + (1-a)\cdot\log\{(1-a)/(1-p)\}$ and $r = p\cdot(1-a)/\{a(1-p)\}$ [2].

Taking into account the various length of patterns, we have developed a method of defining sequence blocks by merging the fixed-length fragments (see [3] for detail). The degree of finding those blocks in specific promoters is defined by the index *PSI*:

$$PSI = \frac{const. + Fg}{const. + Fr},$$

where

Fg : *fraction of sequences containing a pattern in a given group*,

Fr : *fraction of sequences containing a pattern in the rest of the groups.*

In order to examine the predictive ability of this index, a test promoter is checked in turn whether it has any specificity by the following score, which is calculated from a set of indices defined from the rest of the promoters in the dataset.

$$score = \sum_{found\ patterns} PSI$$

[1]藤渕　航，金久　實：京都大学化学研究所，〒 611 京都府宇治市五ケ庄

2 Result and Perspective

Figure 1 is the result of score distributions for liver specific promoters and the rest of the promoters, which show a difference in the two distribution profiles. If the threshold is set to be Z=1, which can discriminate 50% of liver-specific promoters, the false positive rate is 11% ($= \frac{17}{191-36}$). Note that the

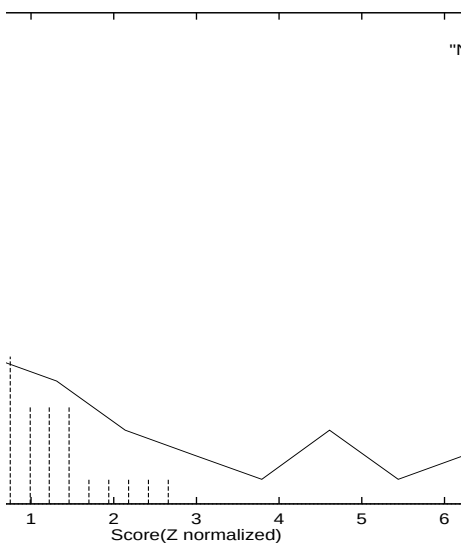

Figure 1: The plot of score distribution between liver and non-liver promoters.

profile for liver is spread wider, which may suggest that the liver-specific promoters can be divided into subgroups.

We are now developing an integrated prediction system including not only the liver but other promoter groups, such as brain and house keeping.

Acknowledgement

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Area 'Genome Informatics' from The Ministry of Education, Science, Sports and Culture in Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] P. Bucher and E. N. Trifonov, "Compilation and analysis of eukaryotic Pol II promoter sequences", *Nucl. Acids Res.*, Vol. 14, pp. 10009-10026, 1986.

[2] R. Arratia and L. Gordon, "Tutorial on large deviations for the binomial distribution", *Bull. Math. Biol.*, Vol. 51, pp. 125-131, 1989.

[3] W. Fujibuchi and M. Kanehisa, "Construction of a functional word dictionary for primate promoter sequences", *Proc. Genome Informatics Workshop IV*, pp. 275-282, 1993.