Prediction of Transcriptional Control by Promoter Specificity Index for Conserved Sequence Patterns

Wataru Fujibuchi and Minoru Kanehisa

We have developed a prediction method for expression specificities in the transcriptional process, which is known to be regulated in large part by promoter sequences, by observing the appearance of conserved sequence patterns in a group of promoters, such as for brain, liver, and house-keeping genes. Related promoters in the same group were compiled from EPD eukaryotic promoter database and an index(PSI, "Promoter Specificity Index") to represent the group specificity of each pattern was calculated. Each promoter was examined for its specificity to test the validity of these indices constructed from the rest of the promoters in our dataset. Currently, our system could discriminate 40 to 50 % of human promoters with 11 to 17 % of false positive rate.

Keywords: Transcriptional control/ Expression specificity/ Signal sequence/ Relative entropy/ Binomial distribution/ Markov chain/ Information content

Eukaryotic genes are expressed under complex regulatory systems that depend on time, place, and other environmental factors. Thus, detecting sequence differences of eukaryotic promoters in different tissues may provide clues to understanding mechanisms of eukaryotic gene expression. It is well known that the gene expression is highly regulated by the level of transcription initiation after the cooperative binding of transcription factors to signal pattern sequences. Experimental approach for collecting sufficient data, however, is a laborious work because the transcription initiation involves multiple factors, and the principles of promoter actions are still too complex to be unraveled.

We have previously developed a new method [1] for automatically identifying possible regulatory signal patterns with optimal lengths from a set of unaligned sequences which are known to be functionally related but not all of which have common homologous regions. It takes the advantages of the Markov chain, relative entropy, and information content theories. Here we present a prediction method for expression specificity of promoters as an application of these extracted conserved patterns.

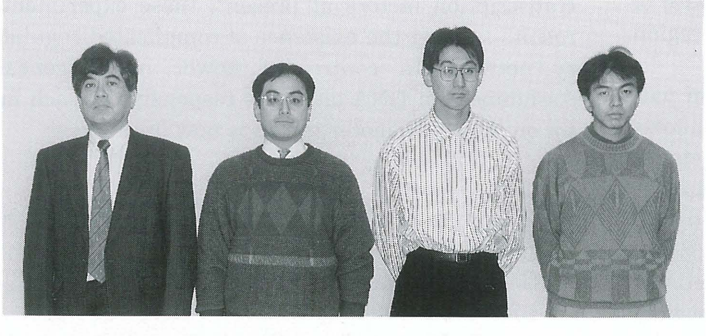
Homo sapiens promoters (-200 to -1) were collected from the EMBL nucleic acid database Release 41.0 according to the EPD [2] entries. We obtained 191 independent (non-homologous) promoters, including

MOLECULAR BIOLOGY AND INFORMATION —Biological Information Science—

Scope of research

The following is the current major activities of this laboratory.

- *Characterization of amino acid sequence motifs from protein structure and sequence databases.*
- *Characterization of nucleotide sequence motifs around promoter, translation initiation and splice sites.*
- *Construction of new databases that describe molecular interactions, such as signal transduction and metabolic pathways.*
- *Modeling three-dimensional structures of RNAs, DNAs and proteins, and their interactions.*
- *Development of database systems and computational tools to support genome research. They are made freely available all over the world via the Internet (http://www.genome.ad.jp/).*


KANEHISA AKIYAMA GOTO FUJIBUCHI

Professor
KANEHISA, Minoru (D Sc)
Associate Professor
AKIYAMA, Yutaka (D Eng)
Instructors
GOTO, Susumu (D Eng)
FUJIBUCHI, Wataru
Students
OHKUBO, Zenmei (DC); OGATA, Hiroyuki (DC); TOMII, Kentarou (DC); SUZUKI, Kenji (MC); KIHARA, Daisuke (MC); TSUKAMOTO, Nobuo (MC); BONO, Hidemasa (MC); HATTORI, Masahiro (MC); KAWASHIMA, Shuichi (MC); PARK, Keun-joon (MC); FUKUSHIMA, Nobuhiro RF, D Sc)

9, 36 and 20 active promoters in the brain, liver and house-keeping genes, respectively.

Conserved 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 p of finding a pattern in one sequence, and the ratio $a = K/N$, the following equation holds for $0 < p < a < 1$:

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

where $r = p(1-a)/\{a(1-p)\}$ [3], and H is the relative entropy between the observed frequency a and the expected frequency p calculated from the background nucleotides by assuming the first order Markov chain:

$$H = a\log(a/p)+(1-a)\log\{(1-a)/(1-p)\}.$$

Taking into account the various lengths of patterns, we have developed a method of defining sequence blocks by merging the fixed-length fragments (see [1] for detail) based on the information content theory.

Once conserved block patterns for each promoter group are found, the degree of those specificities are determined by comparing relative conservation rate between the group and the outside of the group, which is defined by the index PSI [4]:

$$\mathrm{PSI} = \log\frac{\mathrm{const}+Fg}{\mathrm{const}+Fr}$$

where Fg and Fr are, respectively, the fractions of sequences containing a pattern in a given group and 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.

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

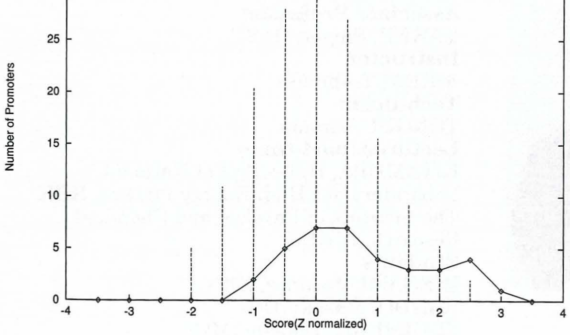


Figure 1. The plot of score distribution profile for both liver and non-liver promoters.

Table 1. A Summery of total prediction rate for brain, liver and house-keeping gene promoter expression.

Promoter(#)	Correct Answer(%)	False Positive(%)
Brain(9)	44.4	16.1
Liver(36)	41.7	15.9
House-keeping(20)	50.0	11.6

The name of promoter group, the number of sequences, and the rates of both correct answers and false positives are shown. The values in the case where the threshold is set to be Z ≥ *1.0* are shown here.

Figure 1 is an example of score distributions for liver specific promoters and the rest of the promoters, which shows a difference in the two distribution profiles. If the threshold is set to be $Z \geq 1.0$, which can discriminate 41.7% of liver-specific promoters, the false positive rate is 15.9%. Note that the profile for the liver is spread wider, which may suggest that the liver-specific promoters can be divided into subgroups.

The summary of total prediction rate is shown in Table 1. The result also indicated that there was no gene preference for giving a correct answer (data not shown).

The PSI indices may give us an important clues for understanding biological processes. For example, it shows a pattern 'TGCCCA' is specifically conserved in the liver promoter group(~42% of the promoters). It closely resemble to the known consensus patterns of liver-specific factor ('TG[A/G][A/C]CC' and a portion of 'TGGTTATN[A/T]TCNNCA') in the TFD transcription factor database[5]. It is, however, still unregistered in the database.

As the genome sequencing projects proceed, there will be more pressing needs for understanding signals that may play important roles in gene regulation and expression. The method presented here is a step forward toward extracting and applying biological knowledge from rapidly expanding sequence data.

Acknowledgment

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

References

1. Fujibuchi W and Kanehisa M, *Proc. Genome Informatics Workshop IV*, 275-282 (1993).
2. Bucher P and Trifonov E.N., *Nucl. Acids Res.*, **14**, 10009-10026 (1986).
3. Arratia R and Gordon L, *Bull. Math. Biol.*, **51**, 125-131 (1989).
4. Fujibuchi W and Kanehisa M, *Proc. Genome Informatics Workshop 1995*, 106-107 (1995).
5. Ghosh D, *Nucl. Acids Res.*, **18**, 1749-1756 (1990).