DNA Research 4, 81-90 (1997)

Prediction of Gene Expression Specificity by Promoter Sequence Patterns

Wataru Fujibuchi and Minoru Kanehisa*

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan

(Received 18 March 1997)

Abstract

We present here a heuristic method toward predicting the expression specificity 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 known promoters, such as for housekeeping or tissue-specific genes. Statistically conserved patterns were automatically extracted from a set of unaligned sequences up to 200 bp upstream of the transcription initiation site, by a standard procedure using the Markov chain and binomial distribution models. Furthermore, to obtain signal sequences of optimal lengths we devised a method that combines the multiple alignment and the analysis of the information content (or relative entropy). Groups of related promoters were compiled from the EPD eukaryotic promoter database and the EMBL nucleic acid sequence database. Each promoter was examined for its specificity by linear discriminant analysis to test the validity of the extracted patterns. Our method could correctly discriminate 77.6% of the housekeeping gene promoters and 62.9% of the liver promoters.

Key words: DNA database; sequence analysis; functional prediction; discriminant analysis; housekeeping genes

1. Introduction

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 organisms and different tissues may provide clues to understanding mechanisms of eukaryotic gene expression. It is well known that gene expression is highly regulated by the level of transcription initiation after the cooperative binding of transcription factors to signal sequences. Experimental studies by gel electrophoresis have detected a number of binding sites for transcription factors. However, because transcription initiation involves multiple factors and because collecting sufficient data is labor-intensive, the principles of promoter actions in different regulatory systems are still too complex to be unraveled at the sequence level.

Recently, distinct biological data of promoters have been compiled in several databases. EPD,1 for instance, contains more than 1,000 eukaryotic promoters which are linked to the EMBL nucleotide sequence database entries. It is now possible and practical to perform computational analysis, especially, statistical analysis of sequence data for obtaining new biological insights into promoters.

In the study of nucleic acid sequence analysis, several search methods have been proposed to detect conserved patterns from a set of functionally related sequences. One advantage of the method using the relative entropy2-4 as a discriminator is that it is an easily understandable measure for the degree of pattern conservation. The analysis based on the Markov chain5,6 seems to be successful for searching patterns in view of the nonrandom statistical properties of DNA sequences. None of these methods, however, is suitable to determine the length, i.e., the boundaries, of the conserved patterns, which seems to vary greatly depending on each regulatory signal. The definition of boundaries may better be achieved when all sequences in a data set are known to contain homologous regulatory regions, by studying the information content7-9 of the sequences or by the expectation maximization algorithm10,11

Efforts have been made to organize databases of known signal sequences, TFD,12 for examples, which are compiled from articles reporting the experimental data on transcription signals. There are also associated approaches used to predict promoter sequences and/or promoter classes by utilizing such databases.13,14 However, there has not yet been any study to predict promoter expression specificity, probably because of the shortage of
II. Prediction

![Diagram showing the strategy to predict promoter expression specificity.](image-url)

Figure 1. A schematic illustration of the strategy to predict promoter expression specificity. The first step is the extraction of conserved sequence patterns that characterize a specific group of promoters. The second step is the prediction of whether a given promoter belongs to the specific group based on the discriminant function calculated from the extracted patterns.

registered data for transcription signal sites classified according to the expression variabilities of promoters such as tissue specificity and time dependency. In view of the rapid progress of genome sequencing projects, needs are increasing for computational methods to characterize gene expression specificity by promoter sequence analysis. This paper describes an attempt toward that direction and describes two important aspects: (i) automatic extraction of signal patterns with optimal lengths and (ii) prediction of specificity by linear discriminant functions.

We have previously devised an effective method for automatically identifying possible regulatory signal patterns of optimal lengths from a set of unaligned sequences which are known to be functionally related but not all of which have common homologous regions. The present method is an extension and takes advantage of the Markov chain and information content theories in the past works. Here we also present a prediction method for expression specificity of promoters by making use of the extracted conserved patterns and by applying the linear discriminant function that distinguishes the promoter group of interest from the rest of the promoters.

2. Materials and Methods

2.1. Promoter dataset

A set of promoter sequences was selected from the EPD eukaryotic promoter database and the EMBL nucleotide sequence database (release 41). Out of 1,251 EPD entries we obtained 191 human promoters, which were defined as independent sequences in the promoter homology grouping by EPD and for which at least 200 bp upstream from the site of transcription initiation could be extracted from EMBL. The EPD database also contains the information of expression or regulatory features in the documentation lines. For further analysis, we selected promoter groups that contained enough members with the same expression/regulation features according to the EPD documentation lines. As a result, we identified a group of liver promoters with 36 sequences, a group of housekeeping promoters with 20 sequences, and a group of brain promoters with 9 sequences.

2.2. Pattern extraction and prediction methods

For a given group of promoter sequences, we extracted conserved patterns to see if they are useful in predicting the group. The pattern extraction and prediction schemes are illustrated in Fig. 1 and described in detail below.

2.2.1. Extraction of conserved segments

In order to search for a putative functional fragment which may play a biological role, we assume that it is a conserved pattern in the stochastic sense below:

\[
\text{an } l\text{-base pattern of } f\% \text{ probability allowing up to } s\% \text{ substitution.}
\]

where a 6-bp pattern \((l = 6)\) is searched against the dataset of sequences each of which is 200 bp long. The probability \(f\) is the rarity calculated approximately by
using the relative entropy based on the binomial distribution model.\textsuperscript{17} If we know beforehand the probability \( p \) for finding a pattern within one sequence, the probability of finding \( k \) occurrences of the same pattern in a set of \( N \) sequences is given by:

\[
I = \frac{1}{1 - r} \left\{ \frac{1}{2pa (1 - a)N} \right\} e^{-NH(a,p)},
\]

where \( r = p (1 - a) / \{ a (1 - p) \} \), \( a = k/N \), and \( H \) is the relative entropy:

\[
H(a,p) = a \cdot \log(a/p) + (1 - a) \cdot \log \{ (1 - a) / (a - p) \}
\]

as long as \( 0 < p < a < 1 \) and \( N \) is large. Assuming that the patterns are Poisson distribution, the probability \( p \) that a pattern is found at least once in one sequence is given by:

\[
p = 1 - \exp\{-E_W(L - l + 1)\},
\]

where \( L \) and \( l \) are the lengths of the sequence and the pattern, respectively, and:

\[
E_W = \frac{E(X_1X_2) \cdot E(X_2X_3) \cdots E(X_{l-1}X_l)}{E(X_2) \cdot E(X_3) \cdots E(X_{l-1})}
\]

is the expected frequency of observing pattern \( W \) in the set of \( N \) sequences. This is calculated by assuming the first order Markov chain according to Stückle et al.\textsuperscript{5} and Pesole et al.,\textsuperscript{6} and \( E(X_iX_j) \) and \( E(X_i) \) are the frequencies of dinucleotides and mononucleotides, respectively, observed in all of the \( N \) sequences. By taking base substitutions into account, which are allowed up to \( s\% \), the expectation values of \( E_W \) for all substitution patterns are added to \( E_W \).

Once conserved 6-bp patterns are found, the actual locations of the patterns are then examined in the original dataset, from which the patterns were extracted, allowing again up to \( s\% \) base substitutions. If the locations found appear consecutively one after another, that is, one fragment overlaps another with 5 bp being shared, those fragments are combined into one segment. By this process, a set of conserved segments with various lengths is obtained.

2.2.2. Length refinement process

In the next step, each segment found is in turn taken for length refinement as follows.

\begin{enumerate}
\item \textbf{step 1:} start with the current segment that has the longest length
\item \textbf{step 2:} select similar segments from all other segments
\item \textbf{step 3:} align all similar segments against the current segment by pairwise alignment
\item \textbf{step 4:} truncate the current segment by information content of the block formed
\item \textbf{step 5:} go to step 1
\end{enumerate}

Let us suppose that one segment is selected for refinement of its length. First, all other segments above a given threshold of local homology are aligned against this segment by pairwise alignment without gaps. The local homology score of two segments is defined by:

\[
S_{ij} = \max_{i,j,k} \left\{ \sum_{i,j} e_{ij} \right\}, \quad (k \geq l),
\]

\[
e_{ij} \begin{cases} 
1 & (\text{base } i = j) \\
-1 & (\text{otherwise})
\end{cases}
\]

where \( e_{ij} \) is the single base similarity score between position \( i \) of one segment and position \( j \) of the other. \( S_{ij} \) is the locally maximal homology score for the partial sequences of length \( k \) when two segments are aligned without any gaps. If the base substitution rate in the \( k \)-base segment is equal to or below the substitution parameter, \( s\% \), which is the same value used for extracting fragments, the aligned segment is retained. Thus, the segment being considered is multiply aligned with a number of locally homologous segments of varying lengths. By referring to the original sequences in the dataset, each aligned segment is extended up to the same length of the segment being considered. The resulting alignment block is converted into the base frequency matrix that has elements of four times the length of the block.

Next, to extract a significantly conserved portion of the block, the information content is calculated using the base frequency matrix. We adopt the definition in terms of the decrease of Shannon’s entropy following the work of Schneider et al.\textsuperscript{18} and Iijima et al.\textsuperscript{9} Thus, the information content of a given column in the aligned block is:

\[
I = - \sum_b f_b \cdot \log_2(f_b) - \left\{ \sum_b p_b \cdot \log_2(p_b) \right\},
\]

\[
p_b = \frac{n_b + 1}{N + 4},
\]

where \( f_b \) is the frequency of base \( b \) in the entire sequences of the dataset, \( n_b \) is the frequency of base \( b \) in the column, and \( N \) is the number of aligned sequences.

We consider the portion of the block that satisfies the following two conditions to be significant and only this portion is retained for further analysis.

\begin{enumerate}
\item \( I_{\text{term}} > I_{\text{thres}} \)
\item \( \sum_{i=1}^{L_s} I_i > 0.5 \)
\end{enumerate}
1. Six-base pattern search

2. Generation of conserved segments

3. Length refinement

A detailed illustration of the signal extraction process. After finding conserved patterns of a fixed length (6 bp), adjacent patterns are merged to form conserved segments of varying lengths. Then the length refinement of each segment is performed by making the multiple alignment and calculating the information content.

Here $I_{\text{term3}}$ is the mean information content of the terminal three bases, $I_{\text{term}}$ is the information content of the terminal base, $I_{\text{thresh}}$ is the user-defined threshold of the information content, and $L_s$ is the length of the segment. The segment is truncated by trimming the terminal bases until it satisfies the two conditions given above. If the segment is shortened to less than 3 bp, it is excluded from the segment pool being considered.

During the refinement process, the segments are first sorted and grouped according to length. Once the segment is shortened to less than 3 bp, it is excluded from the segment pool being considered. If the segment is shortened to less than 3 bp, it is excluded from the segment pool being considered. If the segment is shortened to less than 3 bp, it is excluded from the segment pool being considered.

To avoid order dependency, this move is done at the end of the refinement for the same length group. All base frequency matrices and information contents are also calculated at a time for the same length group before shortening of the blocks. A schematic view of the pattern extraction and refinement processes is shown in Fig. 2.

2.2.3. Linear discriminant function

Once conserved sequence patterns are found for a given group of promoters, a linear discriminant function is derived for distinguishing those promoters from the rest of the promoters. The discrimination is based on the relative conservation rate of each pattern within the group against outside the group. The formulation, which is known as Bayes' decision rule is defined by:

$$g(x) = w_0 + w_1x_1 + \cdots + w_rx_r,$$

$$w_0 = \sum_{k=1}^{r} \log \left( \frac{p_k(0|c_1)}{p_k(0|c_2)} \right) + \log \left( \frac{P_1}{P_2} \right),$$

$$w_k = \log \left( \frac{p_k(1|c_1)}{p_k(1|c_2)} \right) / \left( \frac{p_k(0|c_1)}{p_k(0|c_2)} \right), \quad k = 1, \ldots, r,$$

where $p_k(1|c)$ and $p_k(0|c)$ are the conditional probabilities where pattern $k$ appears or does not appear in promoter group $c$, and $P_1/P_2$ is the ratio of the sample sizes of the two groups of the training data. Then, a test promoter can be checked by calculating $g(x)$, where $x_k$ is 1 when the pattern appears in the test promoter, otherwise $x_k$ is 0. The test promoter is considered to belong to group $c$ when $g(x) \geq 0$, and to not belong to $c$ when $g(x) < 0$.

To evaluate the predictive ability of the discriminant function, we introduce a cross-validation mechanism in a manner similar to that of Prestridge, where promoter sequences known to belong to a given group are divided into several sets and each set is, in turn, put outside of the training data and used as the test data for prediction. In our analysis, six sets each containing six sequences were made for liver promoters, and four sets each containing five sequences were made for housekeeping gene promoters. The negative data were prepared from the promoters outside the group. Because the total number of promoters was 191, we prepared six sets with 25 sequences each for negative liver promoters and four sets each containing five sequences were made for housekeeping gene promoters.

The prediction rate is calculated by the following equation:

$$\text{Prediction Rate} (%) = \frac{T_p(\%)+T_n(\%)}{200},$$

where

$$T_p = \frac{\text{recognized positives}}{\text{all positive data}} \times 100(\%)$$

and

$$T_n = \frac{\text{recognized negatives}}{\text{all negative data}} \times 100(\%).$$
Figure 3. The base compositions in the 200-base range upstream from the transcription initiation site are shown in color for liver and housekeeping promoters. The color codes are: red for A, yellow for T, green for G and blue for C.

Note that the prediction rate is first calculated separately for the positive and negative data and then combined with equal weights, which is intended to avoid possible overestimation of the prediction rate from the larger number of negative data.

3. Results and Discussion

3.1. Base compositions of different promoter groups

The liver promoters and housekeeping gene promoters that we used for analysis are listed in Table 1. The genes were classified according to the EPD documentation lines. Note that a liver gene does not necessarily mean liver-specific expression of the gene, but that the gene is known to be expressed in liver and possibly in other tissues. Figure 3 shows in color the base compositions of liver and housekeeping promoters. The coloring scheme is: red for A, yellow for T, green for G, and blue for C. The figure apparently indicates the differences of base compositions between the two groups of promoters. The promoters for housekeeping genes are relatively GC-rich (green and blue), while the TATA box (~ -35) appears well conserved in liver promoters.

3.2. Parameter optimization and prediction accuracy

There are three parameters in our method that need to be adjusted:

- $f$: the probability of observing a pattern against the background sequences
- $s$: the rate of base substitutions allowed
- $I_{thres}$: the threshold of the information content for trimming terminal bases

We examined various combinations of $f$ and $s$ for the pattern extraction process, and the lengths of extracted segments were optimized by various values of the information content threshold $I_{thres}$. For each set of parameters we made a discriminant function from training promoter data and calculated the prediction rate in test promoter data. The result is shown in Fig. 4.

Overall, the prediction rate was best for housekeeping gene promoters and worst for brain promoters (data not...
Table 1. List of promoter sequences used in this study.

| (a) Promoters for genes expressed in liver (36 genes) |  |
|------------------------------------------------------|--------------------------------------------------|
| HSFIBBR1[X05018] | Human fibrinogen beta gene 5’ region and exon 1 |
| HSFBRGG[X02415] | Human gene for fibrinogen gamma chain |
| HSAPOA1[X00098] | Human apolipoprotein A-I and C-III genes, complete cds. |
| HSAPOA2[X04898] | Human gene for apolipoprotein AI |
| HSAPOB1[M19808] | Human apolipoprotein B-100 (apoB) gene, exons 1 and 2. |
| HSAPOC2G[X05151] | Human apoC-II gene for preproapolipoprotein C-II |
| HSAPOA1[X00098] | Human apolipoprotein A-I and C-III genes, complete cds. |
| HSAPOE4[M10065] | Human apolipoprotein E (epsilon-4 allele) gene, complete cds. |
| HLSBE1[X13075] | Human albumin gene, exon 1 and 5’ flank. |
| HSLBG[M11158] | Human serum prealbumin gene. |
| HSAFGP1[M10949] | Human alpha-fetoprotein (AFP) gene, exon 1 with 5’ flanking region. |
| HSRFB1[X02775] | Human gene fragment for retinol binding protein (RBP) (exon 1-4) |
| HSGGLT1[X06492] | Human theta 1-globin gene |
| HSHB2[M34058] | Human beta-globin gene from a thalassemia patient, complete cds. |
| HSMRDI1A3[M57450] | Human multidrug-resistance (MDR1) gene, exon 1. |
| HSPH1G1[X01793] | Human haptoglobin (Hp1) gene exon 1 and 5’ flanking region |
| HSAAG5E[X03258] | Human argininosuccinate synthetase gene 5’ end |
| HSAARG1[KO2O48] | Human gene for lecithin-cholesterol acyltransferase (LCAT) |
| HSSOD1G1[X01780] | Human aldolase B (ALDOB) gene, exons 2 through 6. |
| HSLPH1[X35452] | Human hepatic lipase gene, exon 1. |
| HSRFCM11228 | Human protein C gene, complete cds. |
| HSFILG1[K02048] | Human factor IX gene, exon 1. |
| HSMH2A1[M15902] | Human MHC class III HLA factor B (Ba fragment) gene, 5’ end, clone cosmid Cos10, subclone pH1.3B. |
| HSMHCP42[M12792] | Human steroid 21-hydroxylase (P-450(C21)) B gene, complete cds, clone lambda-C21B-1. |
| HSOFG2AP[X05331] | Human insulxin/IGF-II intergenic region with adult IGF-II promoter |
| HSIFF1EF[X30308] | Human insulin-like growth factor IGFII gene leader exon (fetal promoter) |
| HSAANG1[X13520] | Human angiotensinogen gene 5’ region and exon 1 |
| HSSEGA1[M16789] | Human HER2 gene, promoter region and exon 1. |
| HSAATP[K20212] | Human alpha-1-antitrypsin gene (S variant), complete cds. |
| HSAATP[H20023] | Human S variable segment 5’ of antithrombin III gene (AT III) |
| HSAATP | Human renin-binding protein gene, 5’ flanking region and exon 1. |
| HSC5GN[M72430] | Human C5 gene, 5’ end. |
| HSC4BNDL[L11246] | Human C4b-binding protein beta-chain, exons 2-3b. |
| HSC4BNDL | Human genomic fragment for factor VIII (5’ terminus) |

| (b) Promoters for housekeeping genes (20 genes) |  |
|--------------------------------------------------|--------------------------------------------------|
| HSRNU1[V00501] | Human gene for small nuclear RNA U1. |
| HSUG2A[K03023] | Human small nuclear RNA U2 gene. |
| HSUG3EP1[M14061] | Human U3 small nuclear RNA gene. |
| HSUG4AC1[M15957] | Human U4C small nuclear RNA gene, complete cds. |
| HSMG14A[X21339] | Human non-histone chromosomal protein HMG-14 gene, complete cds. |
| HSMG17G[X13546] | Human HMG-17 gene for non-histone chromosomal protein HMG-17 |
| HSSNCEL0[M60858] | Human nucleolin gene, complete cds. |
| HSSNRP1[M21253] | Human small nuclear ribonucleoprotein (snRNP) E gene, exons 1-4, and 7 Alu repeats. |
| HSRPS1[M13934] | Human ribosomal protein S14 gene, complete cds. |
| HSRG1A[M32405] | Human homologue of rat insulinoma gene (rig), exons 1-4. |
| HSACTBP1[R00474] | Human beta-actin gene 5’-flanking region |
| HSAUG4CA[M15957] | Human HMG CoA reductase gene, exon 1, and promoter region. |
| HSRODG[X00948] | Human URO-D gene for uroporphyrinogen decarboxylase 5’-region |
| HSTPP1[M10036] | Human triosephosphate isomerase mRNA, complete cds. |
| HSPG11[K00159] | Human phosphoglycerate kinase (pgk) gene, exon 1. |
| HSG6PD[X14520] | Human gene for glucose 6-phosphate dehydrogenase G6PD exon 1 (5’ UTR ) |
| HSSOD1G1[X01780] | Human superoxide dismutase (SOD-1) gene exon 1 and 5’ flanking region |
| HSYU1G1[X04803] | Human ubiquitin gene (3 repeats) |
| HSDP1F1[M84327] | Human ADP-ribosylation factor 1 gene, exon 1. |
| HSUHLP[J03589] | Human ubiquitin-like protein (GdX) gene, complete cds. |
Figure 4. The prediction rates of liver and housekeeping gene promoters are plotted against the information content threshold $I_{\text{thres}}$ with various values of the other two parameters: the pattern probability $f$ and the substitution frequency $s$.

3.3. Number of extracted patterns

Table 2 shows the average number of extracted patterns for the six (liver) or four (housekeeping) training data sets with various parameter values. It was natural to observe that as more substitutions were allowed or as more probable 6-bp patterns were considered, the number of extracted patterns increased, while by increasing the information content threshold, the number of extracted patterns decreased.

The parameter sets with the highest prediction rates are indicated by asterisks. Note that the linear discriminant function is derived from the same number of variables as the number of extracted patterns shown in Table 2. The maximal prediction rate for housekeep-
The housekeeping group and HNF1 (hepatocyte nuclear factor-1) and LF-A2 (liver specific factor-A2) in the liver group. The pattern GACCTT in the liver group was also found to be embedded in the LF-B1 recognition sequence and the information content threshold for the length adjustment was performed here with the substitution frequency may be defined separately for each pattern, which corresponded to a peak in the bell-shaped curve of the prediction rate versus the information content threshold (Fig. 4). In contrast, the optimal prediction rate of liver promoters was only 62.9%, which required a large number of patterns (Table 2) and which was obtained with a relatively small number of (Table 2) and which was obtained from the more or less flat curves in the parameter optimization (Fig. 4). Thus, we consider that only the prediction of housekeeping gene promoters was successful. Liver promoters were more difficult to predict than housekeeping gene promoters, probably because the liver promoter group, as defined herein actually represented a collection of less specific promoters expressed in multiple tissues.

3.6. Further Improvements

The length adjustment was performed here with the discriminant function by changing the single parameter of the information content threshold. It is expected that better prediction may be achieved when each pattern is optimized by each threshold value. Similarly, the parameter for the substitution frequency may be defined separately for each pattern, which would again improve the prediction accuracy. Furthermore, the use of a non-linear discrimination method such as a neural network may also be introduced in order to incorporate a cooperativity of related patterns into the prediction. Although these improvements in methodology are important, we think that the critical aspect for further improvements is the database itself. In the current promoter databases such as EPD, the data is still insufficient to develop a practical prediction system for different expression specificity. We are looking into the possibility of using the EST data for identifying tissue-specific gene expression. Once such data are properly organized, the method presented here has the potential to make an extensive cat-
Table 3. Highly discriminative patterns matched with Transcription Factor Database (TFD) entries.

| (a) Patterns extracted from liver promoters | (b) Patterns extracted from housekeeping promoters |
|-------------------------------------------|--------------------------------------------------|
| **Extracted pattern [Prediction Rate]** | **Matching [TFD sequence: matchingscore]** |
| AGAGGT[60.8] AP-1{gagaggag: 5/6} | AP-1{tgacttcct: 5/6} CREB{tgacgttct: 5/6} |
| PR{tgcttctc: 5/6} | TFII-I{agcttctc: 5/6} |
| CTGCC[60.1] (Sp1){cccccgcgc: 5/6} | (Sp1){gggggcgg: 5/6} |
| C/EBP{tcctcctc: 5/6} | H-2RIBP{tgtcctct: 5/6} |
| Sp1{gggggggg: 5/6} | Spl{cccccgcgc: 5/6} |
| TGTCCT[59.1] CREB{tgacgttct: 5/6} | GR{tgcttct: 6/6} |
| GR{tgcttct: 6/6} | Lb{ccgcgtggt: 5/6} |
| (b) Patterns extracted from housekeeping promoters | |
| GAGCCG[72.4] (Sp1){cccccgcgc: 5/6} | (Sp1){gggggcgg: 5/6} |
| CAGCCG[65.1] (Sp1){cccccgcgc: 5/6} | Sp1{cccccgcgc: 5/6} |
| GAGCG[63.8] EIIaE-A{aagggcgc: 5/6} | MFI-1{gcgcgcgc: 5/6} |
| GAGGCC[62.4] (Sp1){cccccgcgc: 5/6} | Sp1{gggggcgg: 5/6} |
| T-antigen{gagggc: 5/6} | |
| CGCGG[67.9] EIIaE-A{aagggcgc: 5/6} | |
| ACGGCC[61.9] (Sp1){cccccgcgc: 5/6} | |
| GCGGAG[61.4] (Sp1){cccccgcgc: 5/6} | |
| GCAGCCG[60.5] (Sp1){cccccgcgc: 5/6} | |
| GACGGCC[60.0] ATF{tgacgttct: 5/7} | |
| PEBP2{gcgcgcgc: 5/7} | |
| CAGGCCTC[60.0] ATF{tgacgttct: 5/7} | |
| PEBP2{gcgcgcgc: 5/7} | |

A logical expression of gene expression signals and to provide automatic analysis of genomic sequences for functional identification.

Acknowledgments: We thank Dr. Kenta Nakai for helpful discussions and comments on the manuscript. This work was supported in part by a Grant-in-Aid for scientific research on the priority area “Genome Science” 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. Bucher, P. and Trifonov, E. N. 1986, Compilation and analysis of eukaryotic POL II promoter sequences, *Nucl. Acids Res.*, 14, 10009–10026.
2. Waterman, M. S., Galas, D. J., and Arratia, R. 1984. Pattern recognition in several sequences: consensus and alignment, *Bull. Math. Biol.*, 46, 515-527.
3. Galas, D. J., Eggert, M., and Waterman, M. S. 1985, Rigorous pattern-recognition methods for DNA sequences, Analysis of promoter sequences from *Escherichia coli*, *Bull. Math. Biol.*, 18, 117-128.
4. Ulyanov, A. V. and Stormo, G. D. 1995, Multi-alphabet consensus algorithm for identification of low specificity protein-DNA interactions, *Nucl. Acids Res.*, 18, 1434-1440.
5. Sticklee, E. E., Emmrich, C., Grob, U., and Nielsen, P. J. 1990, Statistical analysis of nucleotide sequences, *Nucl. Acids Res.*, 18, 6641–6647.
6. Pesole, G., Prunella, N., Liuni, S., Attimonelli, M., and Saccone, C. 1992, WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences, *Nucl. Acids Res.*, 20, 2881–2885.
7. Stormo, G. D. and Hartzell III, G. W. 1989, Identify-
9. Prediction of Promoter Specificity

8. Hertz, G. Z., Hartzell, G. W., and Stormo, G. D. 1990, Identification of consensus patterns in unaligned DNA sequences known to be functionally related, *CABIOS*, 6, 81-92.

9. Iijima, T. and Kanehisa, M. 1991, Analysis of DNA functional sites by information content, *Bull. Inst. Chem. Res.*, 69, 226-233.

10. Lawrence, C. E. and Reilly, A. A. 1990, An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences, *Proteins*, 7, 41-51.

11. Cardon, L. R. and Stormo, G. D. 1992, Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments, *J. Mol. Biol.*, 223, 159-170.

12. Ghosh, D. 1990, A relational database of transcription factors, *Nucl. Acids Res.*, 18, 1749-1756.

13. Prestridge, D. S. 1995, Predicting Pol II promoter sequences using transcription factor binding sites, *J. Mol. Biol.*, 249, 923-932.

14. Kondrakhin, Y. V., Kel, A. E., Kolchanov, N. A., Romashchenko, A. G., and Milanesi, L. 1995, Eukaryotic promoter recognition by binding sites for transcription factors, *CABIOS*, 11, 477-488.

15. Fujibuchi, W. and Kanehisa, M. 1993, A method to extract functional motifs for transcriptional regulation in eukaryotic sequences, *Bull. Inst. Chem. Res.*, 71, 317-326.

16. Fujibuchi, W. and Kanehisa, M. 1993, Construction of a functional word dictionary for primate promoter sequences, *Proceedings Genome Informatics Workshop IV*, Universal Academy Press, pp. 275-282.

17. Arratia, R. and Gordon, L. 1989, Tutorial on large deviations for the binomial distribution, *Bull. Math. Biol.*, 51, 125-131.

18. Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. 1986, Information content of binding sites on nucleotide sequences, *J. Mol. Biol.*, 188, 415-431.