Bull. Inst. Chem. Res., Kyoto Univ., Vol.71, No.3, 1993

A Method to Extract Functional Motifs for Transcriptional Regulation in Eukaryotic Sequences

Wataru FUJIBUCHI* and Minoru KANEHISA*

Received August 6, 1993

We present a heuristic method to define sequence motifs for transcription regulation from a set of DNA sequences upstream of the transcription initiation site. The method first identifies weakly conserved blocks within a given window relative to the initiation site by the search and merge of six-base patterns. Then most conserved portions of these blocks are extracted by calculating the information content after similar blocks are multiply aligned. The procedure was applied to primate promoters and the result was evaluated with the Transcription Factor Database (TFD).

KEY WORDS: Consensus Sequence / Information Content / Promoter / Transcription Factor

1. INTRODUCTION

The compilation of biosequence databases will bring useful information for molecular biologists if appropriately accompanied by theoretical studies. Sequence alignment is a well-established method to identify similar regions, including functional sites, of nucleic acid and protein sequences when they contain relatively long stretches of similarity. However, it is often the case that functional sites share too weak and localized similarity to be detectable by simple sequence alignment. By compiling many sequences and applying more sensitive methods, it becomes possible to identify sequence motifs that signal important biological functions.

For protein sequences such motifs are relatively well defined, organized in libraries like Prosite[1], and utilized for biological interpretation of amino acid sequence data. In contrast, nucleic acid sequence motifs still require major efforts to collect and properly organize relevant data and to develop computational methods to define and extract biologically important sequence patterns. We are interested here in extracting transcriptional control signals from a set of eukaryotic sequences. In the past, certain statistical methods[2,3] were applied to DNA sequences supposed to contain promoter signals and detected conserved sequence patterns. None of these methods, however, was able to inform the actual length of the regulatory site. The definition of a regulatory site may better be achieved, when all sequences in a data set are known to contain regulatory sites, by studying the information content[4,5] of the DNA sequence.

Our aim is to identify regulatory sequence patterns which have optimized lengths from a set of sequences which may or may not contain regulatory sites. We have developed a heuristic method which first identifies weakly conserved blocks by the search and merge of

*藤渕 航, 金久 實: Division of Molecular Biology and Information III, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611

short patterns and then extracts most conserved portions of these blocks by a combination of alignment and information content algorithms. This process may reflect biological situations, where transcriptional signal sites can be somewhat different in different sequences but they may contain a highly conserved consensus core and extended regions in one or both directions. Whether our method can reproduce biologically significant patterns is estimated by the use of the Transcription Factor Database (TFD)[6] as a kind of evaluation function.

2. MATERIALS AND METHODS

2.1. System and Dataset

A dataset of primate sequences upstream of the transcription initiation site was compiled from the EMBL database release 31.0 on a Sun SPARCstation 2 using the promget program coded in C on the UNIX operating system. First, sequences were selected by the keyword "prim_transcript" in the EMBL features table, and there were 322 sequences out of 16,172 primate entries. We examine up to 1000 bases upstream of the transcription initiation site, which are numbered in the 5' to 3' direction from -1000 to -1 where -1 is the base position immediately preceding the initiation site. They are divided into 10 regions each containing 100 bases. When aligned at the base position -1, each region may contain truncated sequences because the data set contained sequences with varying lengths. For simplicity, we require all sequences in one region be 100 bases long, removing truncated portions.

Next, to remove duplicate entries of similar sequences, all pairwise alignments of collected sequences were performed in each of the 100 base regions. For each group of sequences with above 80% similarity, only one representative was retained. As the result, there were 225 sequences in the first (-100 to -1) region and 94 sequences in the last (-1000 to -901) region of the dataset.

The actual analysis described below was done on a CRAY Y-MP2E/264 with programs coded in C and implemented on the UNIX operating system.

2.2. Search for blocks of conserved patterns

1) Putative functional fragment

In order to get sequence patterns which may bear biological significance, we search conserved patterns at specific locations relative to the transcription initiation site. For that purpose, we define a putative functional fragment (PFF) as:

an l-base pattern found in f% of all the sequences allowing up to s% substitutions.

A window search method[2] is used to count the frequency of each fragment in order to accomodate some variations of the location. As illustrated in Fig.1, the sequences of 100 bases long taken from one of the 10 regions of the dataset are aligned and the sequence patterns of 6 bases ($l = 6$) are searched in the window of 16 bases.

2) Block mapping

Once relatively conserved six-base patterns of PFFs are found in each 16-base window, the actual locations of the six-base patterns are then examined on the original sequences allowing again up to s% substitutions. If one PFF overlaps another with five bases sharing, i.e., if one PFF is shifted by one base from another, the two PFFs are combined. We call a

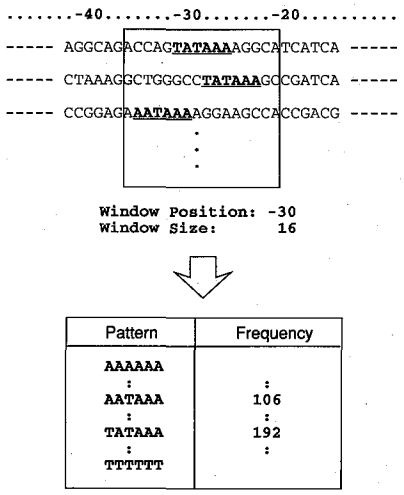


Fig.1. Selection of putative functional fragments (PFFs) by the window search method. The occurrences of 6 base patterns are searched in the window of 16 bases in order to accomodate variations of the location. In addition, the search is made allowing base substitutions in order to include similar patterns.

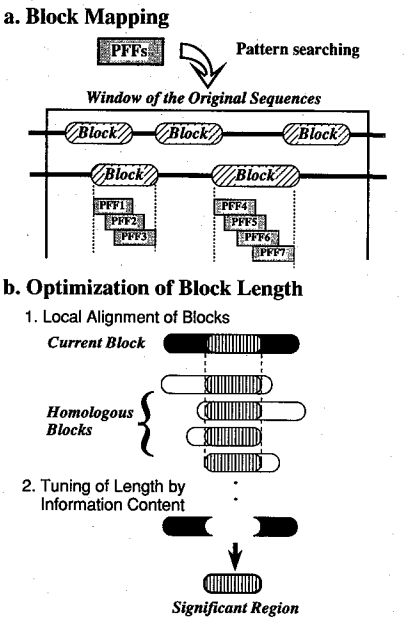


Fig.2. A schematic illustration of how the optimized block is obtained. (a) An initial block is obtained by combining adjacent PFFs in the original window. (b) Then the block is aligned with homologous blocks with the same or shorter lengths, and the significant region is extracted by the information content.

combined region of PFFs a block. Thus, many PFFs found in a given window in one sequence form a limited number of blocks (Fig.2). By this process which we call block mapping, a set of sequences is converted into a set of blocks in a given window.

2.3. Optimization of block length

Blocks containing the same PFFs have, in general, various lengths with a highly conserved core region and extended region(s). To extract only the significant region in each block, we have devised an iterative refinement procedure as follows.

1) Local alignment of blocks

Suppose one block is selected for refinement. First it is aligned with other blocks of equal or shorter lengths. The local homology score of two blocks is defined by :

$$S_{ij} = \max_{i,j,k} \{ \sum_{i,j}^{i+k,j+k} e_{ij} \} \quad , k(\geq 1)$$

$$e_{ij} = \{1\,(\text{base } i = \text{base } j),\ -1\,(otherwize)\}$$

where eij is a single base similarity score between position i of one block and position j of the other. Sij is the locally maximal homology score for the segment of length k when two blocks are aligned without any gap. If the base substitution of the k-base segment is equal or under the substitution parameter (s%) of PFF, the aligned block is retained. Thus, the block being considered is multiply aligned with a number of locally homologous blocks. This multiple alignment is converted into the base frequency matrix which has elements of four times the length of the block being considered.

2) Information content

In order to extract a significant portion of the block, the information content was calculated using the base frequency matrix. We adopt the definition of the information content following the work of Iijima and Kanehisa[5]:

$$I = \sum_{b=A}^{T} p_b * \log(p_b) - \sum_{b=A}^{T} f_b * \log(f_b)$$

$p_b = (n_b + 1) / (N + 4)$

N : the number of sequences

n_b: the frequency of base b in the column

f_b : the frequency of base b in the entire window

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

(a) The total number of bases at each column is no less than the half of the average number in all columns.

(b) The information content on both ends of the columns is more than the threshold I_{thres}.

The base frequency matrix is first inspected with (a) and, if the selected columns form a segment of more than 4 bases, it is then inspected with (b).

3) Iterative procedure

In practice, all blocks found in a given window are sorted and grouped according to the length. Starting with the longest one, each block among the same length group is in turn taken for multiple alignment with similar blocks of the same and shorter lengths by the procedure 1). The resulting base frequency matrix is used to optimize the block length by the procedure 2). The shortened block now belongs to another group of a shorter length and it is later utilized when blocks of that length are examined. To avoid order dependency, all base frequency matrices and information contents are calculated among the same length group before performing the shortening of blocks. The procedure is illustrated in Fig. 2.

3. RESULT

3.1. Determination of parameters

There are three parameters in our method which need to be estimated first: the fraction of sequences containing a given pattern f, the pattern substitution rate s, and the threshold information content I_{thres}. Since our purpose is to extract biologically meaningful patterns, we use the Transcription Factor Database (TFD)[6] as a reference of experimentally deter-

mined functional sites.

Before the critical evaluation, we first counted the number of unique mapped blocks to roughly estimate the range of parameters f and s to be used. We examined various values of f and s in the region of -100 to -1, which contained TATA box position around -30 to -25. When we searched PFF patterns of 6 bases (l = 6) in the window of 16 bases, we allowed 0, 1, 2 substitutions (s = 0, 1/6, and 1/3) with the values of f ranging from 10% to 50%. When s = 0 no blocks were extracted other than TATA related ones, while when s = 1/3 virtually all the 16 base windows were extracted as single blocks (data not shown). Thus, we chose 20%, 25% and 30% substitutions as the parameter s. The observed frequency of unique blocks, which exclude identical blocks of the same lengths, with s = 1/6 is shown in

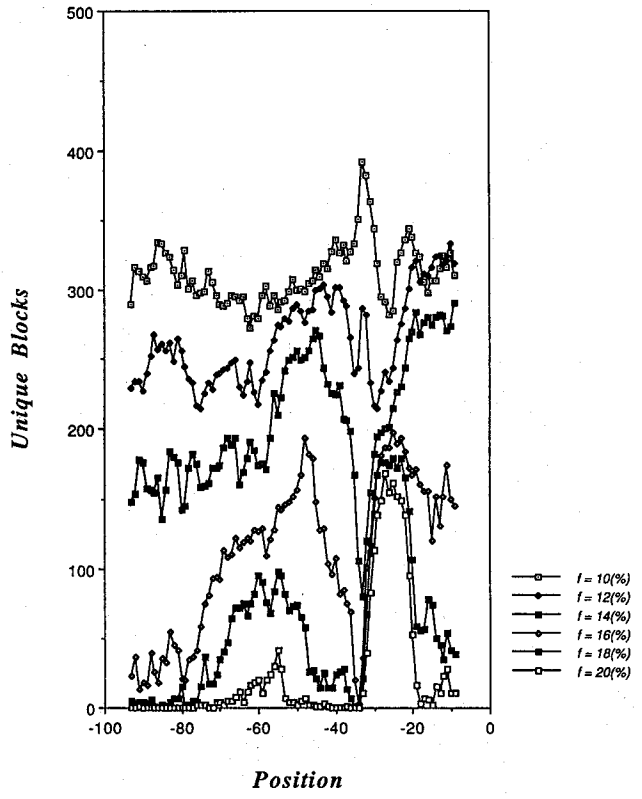


Fig.3. The number of unique blocks observed in the -100 to -1 region of primate promoter sequences with various values of the frequency parameter f. According to the method described in 2.2, blocks are obtained by the search of 6 base PFFs and subsequent block mapping in the window of 16 bases. The substitution parameter used was: s = 1/6. As the frequency parameter f takes smaller values, the TATA box-like peak around -30 to -25 appears clearly.

Fig.3 with various values of f. When f = 10% the TATA peak was not so distinct, while when f = 20% or higher the TATA peak was clear but most other peaks were suppressed. Therefore, we chose 12%, 14%, 16%, and 18% for the parameter f, which may well detect signals other than TATA box.

Using combinations of s and f chosen above, blocks were extracted and their lengths were optimized by the information content. As it happened, since blocks were constructed from PFFs there was a chance of observing very infrequent block patterns. In order to further exclude less frequent patterns, we performed an additional selection as follows. A block is removed if the observed frequency of itself and similar blocks up to s% substitutions combined is below f% of the number of sequences.

With the combinations of parameters s and f, we varied the last parameter I_{thres} ranging from 0.1 to 0.6, and the selected blocks were compared with TFD consensus sequences.

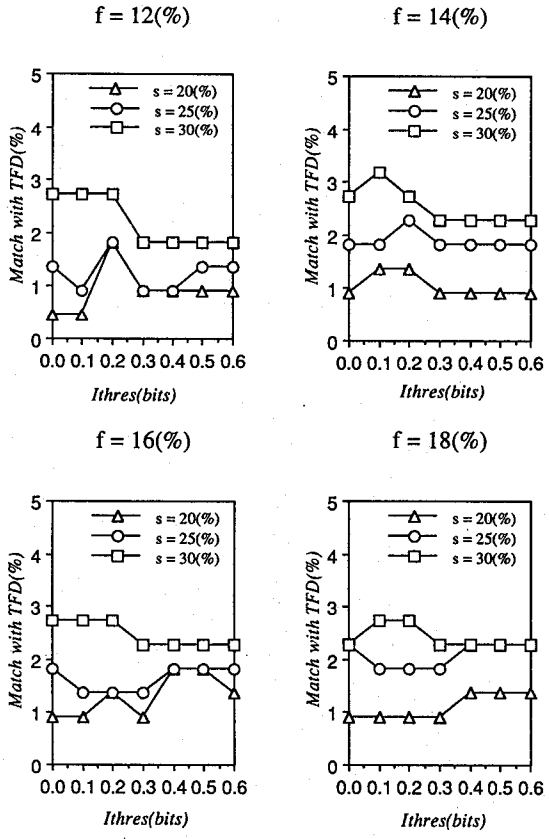


Fig.4. The match rate of optimized blocks with TFD consensus patterns is plotted against the threshold of information content at various values of the frequency parameter f and the substitution parameter s.

As shown in Fig.4 the number of perfect matches with TFD took the highest value when: f = 14%, s = 30% and I_{thres} = 0.1. This is the parameter set used in the present study.

3.2. Characterization of blocks by TFD

With the parameter set determined above, we extracted blocks from each of the ten regions that cover base positions -1000 to -1 and inspected the number of matches with TFD entries. A block is considered to match a TFD entry when the TFD pattern is included in the block pattern. Thus, each block is compared with all TFD entries with the same or shorter lengths and the result is shown in Table 1. Note that a block may correspond to multiple TFD entries. For each of the ten regions Table 1 shows the number of sequences examined, the number of extracted blocks, and the number and fraction of matched blocks, as well as the numbers of blocks which correspond to specific transcription factors.

Table 1. Blocks characterized by TFD.

Region	**-1000**	**-900**	**-800**	**-700**	**-600**	**-500**	**-400**	**-300**	**-200**	**-100**
Sequences	94	103	122	129	138	155	179	197	215	225
Extracted Blocks	196	197	86	75	126	71	60	135	209	308
〈Factors〉					**〈Corresponded Blocks〉**					
TFIID	0	0	1	0	0	0	0	0	0	93
(TFIID/TBF)	0	0	1	0	0	0	0	0	0	90
B-factor	0	0	0	0	0	0	0	0	0	68
TCF-1	7	11	3	5	2	4	1	1	1	59
LVc	5	4	7	2	10	3	3	7	16	16
T-Ag	5	3	2	1	11	2	10	30	24	16
GCF	0	0	0	0	0	0	0	0	12	16
Sp1	0	1	0	0	0	0	0	1	16	15
AP-2	0	0	0	0	0	0	0	0	8	13
GAL4	6	7	0	1	5	2	0	1	1	11
BGP1	0	0	0	0	0	0	0	1	8	8
(Sp1)	0	0	0	0	0	0	0	1	8	8
LSF	0	0	0	0	0	0	0	1	8	8
CTCF	5	3	1	1	4	1	2	11	12	7
GATA-1	0	0	0	0	0	0	0	0	0	4
malT	0	0	0	0	4	0	0	1	1	3
PuF	0	0	0	0	0	0	0	1	1	3
(SRF)	0	0	0	0	0	0	0	0	0	2
SRF	0	0	0	0	0	0	0	0	0	2
(TFIID/TBP)	0	0	0	0	0	0	0	0	0	2
zeste	0	1	1	0	0	0	0	0	0	2
CF1	0	0	0	0	0	0	0	0	0	1
sigma-70	0	0	0	0	0	0	0	0	0	1
MEF-2	0	0	0	0	0	0	0	0	0	1
LBP-1	4	6	0	2	2	2	12	1	3	0
H4TF-2	0	0	0	0	0	0	0	1	0	0
PEA3	3	6	0	0	1	0	1	0	0	0
GR	0	3	1	0	0	1	0	0	0	0
PU.1	2	2	0	1	1	0	0	0	0	0
GCN4	0	1	1	0	0	0	0	0	0	0
AP-1	0	1	0	0	0	0	0	0	0	0
mulutiple	4	1	0	0	0	0	0	0	0	0
c-Myb	2	0	0	0	0	0	0	0	0	0
Myb	2	0	0	0	0	0	0	0	0	0
v-Myb	2	0	0	0	0	0	0	0	0	0
Coresponded	38	45	16	12	35	15	27	54	84	198
Percentage	19.39	22.84	18.60	16.00	27.78	21.13	45.00	40.00	40.19	64.29

As can be seen 64% of our blocks corresponded to TFD entries in the -100 region covering base positions from -100 to -1. The result shows the match rate decreases as the dataset region goes upstream. Although the match rate in the -900 region seems somewhat higher than the -1000 and -800 regions, the difference could well be within statistical variations, especially because the number of sequences is roughly half that of the -100 region.

In the -100 region TATA box- and GC box-related patterns (eg., ATATAAAAAG and GGCGGGG) found in factors such as TFIID, GCF and Sp1 represent a large fraction of blocks extracted. Thirteen blocks with AP-2 like patterns (eg., GGGCAGGC) were all matched with exact lengths. An interesting observation was the existence of TFIID-like blocks in the -800 region. Investigating the data in more detail, we found that TATA box-like blocks were located on the reverse strand. Since, in addition, we found Sp1-like blocks appeared in the -900 region, this may suggest that there are genes transcribed in reverse direction in that region.

3.3. Block length

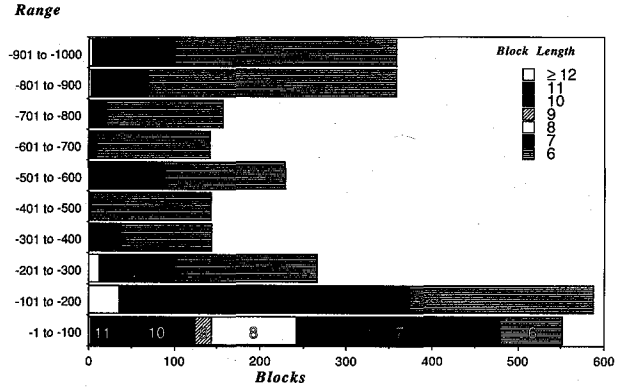


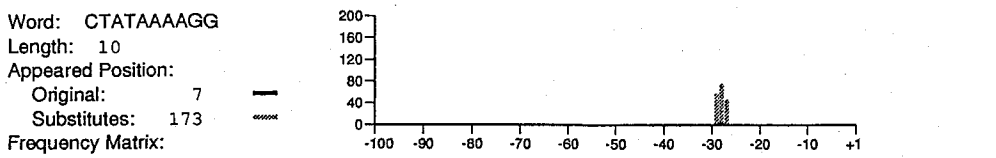
Fig.5. The length distribution of optimized blocks in each 100 base region of primate promoter sequences.

The distribution of block lengths are shown in Fig.5. In general 6 and 7 base blocks were most frequent in all regions, while significant number of longer blocks were found in the -100 region. We did not find 5 base or shorter blocks in any region. Again, it is not certain whether the apparent increase of blocks found in the -1000 and -900 regions is simply a result of statistical variations or contains any biological significance.

3.4. Motif dictionary

Toward making a dictionary of functional words for interpreting DNA sequence data, we have compiled our collection of blocks in a form shown in Fig.6. This is an entry of our motif dictionary which contains the block pattern, the observed frequencies of the block and similar blocks both in numerical and graphical representations, the base frequency matrix, and the corresponding TFD entries.

Functional Word Dictionary



A	85	6	178	0	180	178	172	159	74	37	0.47	0.03	0.99	0.00	1.00	0.99	0.96	0.88	0.41	0.21
T	8	158	2	179	0	1	1	7	1	9	0.04	0.88	0.01	0.99	0.00	0.01	0.01	0.04	0.01	0.05
G	35	4	0	0	0	1	4	14	99	100	0.19	0.02	0.00	0.00	0.00	0.01	0.02	0.08	0.55	0.56
C	52	12	0	1	0	0	3	0	6	34	0.29	0.07	0.00	0.01	0.00	0.00	0.02	0.00	0.03	0.19

TFD record:	IMAMM	(TFIID/TBP)	**TATAAA**
TFD record:	IDROS	B-factor	**TATAAAA**
TFD record:	CMAMM	TCF-1	**MAMAG**
TFD record:	IEUK	TFIID	**TATAAA**
TFD record:	IYEAST	TFIID	**TATAAA**
TFD record:	IMAMM	TFIID	**TATAAAA**
TFD record:	IMAMM	TFIID	**TATAAAA**
TFD record:	CMAMM	TFIID	**TATAWAW**

Fig.6. An entry of the DNA functional word dictionary. This word is extracted from the -100 to -1 region of primate promoter sequences. As shown here the dictionary contains the word pattern, the length of the word, the number and the position of original and similar words, and the consensus matrix showing possible variations in similar words. Actual usage of the word is given, if it exists, as corresponding TFD entries.

4. DISCUSSION

The method we presented in this paper is an approach toward automatic detection of functional signals dispersed in DNA sequences. Roughly speaking, it consists of two steps: (i) finding weakly conserved patterns in a given window (PFF search) and combining them as blocks (block mapping); and (ii) extracting most conserved portions of blocks by local alignment and calculation of the information content. It is known that the first step itself involves difficult statistical problems of estimating the frequency of observing specific DNA patterns. Thus far, Markov chain models[3,7] have been presented to estimate sequence probabilities. It may be possible to apply such stochastic models to the refinement of the first step of our procedure. In this paper we took a practical approach and adjusted the parameter values by experimental data stored in TFD. However, since the parameters were adjusted in the -100 region which is exceptionally rich in biologically significant patterns, the validity of the parameter values in others regions is not clear. It will be necessary to use other parameter values and examine, for example, if we can still observe more signals in the -1000 and -900 regions as indicated in Fig.5.

The second step of block length optimization by the information content contains our idea of distinguishing a highly conserved core and extended regions of a regulatory site. This may reflect the way transcription factors interact with DNA in the three-dimensional structures, which may require a common spatial contact and minor variations dependent on local DNA conformations. It is assumed that the information content of blocks reflects those conformational properties.

The result of comparison with TFD gives us new biological insights into the DNA signals. For example, TATA box is often described as 'TATAAA' or 'TATAAAA' in the textbook. However our result shows that TATA box-like blocks are formed mainly beginning with 'GTATA', 'CTATA' or 'ATATA', and blocks beginning with 'TTATA' exist much less frequently. In fact, the information content is high at the base position preceding 'TATA'. We think it better to call VTATA box rather than TATA box.

In this paper we concentrated on the application of our method to transcription signals, but it is also applicable to analyzing other regulatory signals in DNA and RNA sequences, such as translation initiation signals, splicing and other RNA processing signals. As the genome sequencing projects proceed, there will be more pressing needs for understanding specific signals that play important roles in gene regulation and expression. The approach presented in this paper is an important step towards efficient extraction and use of biological knowledge from rapidly expanding sequence data.

ACKNOWLEDGEMENT

We thank Mr. Atsushi Ogiwara for instructions of programming. This work was supported by the grant-in-aid for scientific research on the priority area 'Genome Informatics' from the Ministry of Education, Science and Culture, Japan. The computation time was provided by the Supercomputer Laboratory, the Institute for Chemical Research, Kyoto University.

REFERENCES

(1) A. Bairoch, *Nucl. Acids Res. Suppl.*, **20,** 2013(1991).
(2) D.J. Galas, M. Eggert, and M.S. Waterman, *J. Mol. Biol.*, **186,** 117(1985).
(3) G. Pesole, N. Prunella, S. Liuni, M. Attimonelli, and C. Saccone, *Nucl. Acids Res.*, 20, 2871 (1992).
(4) G.D. Stormo, and G.W. Hartzell, III, *Proc. Natl. Acad. Sci. USA*, 86, 1183(1989).
(5) T. Iijima, and M. Kanehisa, *Bull. Inst.Chem.Res., Kyoto Univ.*, **69,** 226(1991).
(6) D. Ghosh, *Nucl. Acids Res.*, **18,** 1749(1990).
(7) J. Kleffe, and M. Borodovsky, *CABIOS*, **8,** 433(1992).