Molecular Biology, Vol. 35, No. 6, 2001, pp. 818–825. Translated from Molekulyarnaya Biologiya, Vol. 35, No. 6, 2001, pp. 961–969.

UDC 557.323.5+576.315.42

Comparative Analysis of Methods Recognizing Potential Transcription Factor Binding Sites

M. A. Pozdnyakov[1], E. E. Vityaev[2], E. A. Ananko[1], E. V. Ignatieva[1], O. A. Podkolodnaya[1], N. L. Podkolodnyi[1], S. V. Lavryushev[1], and N. A. Kolchanov[1]

[1] *Institute of Cytology and Genetics, Siberian Division, Russian Academy of Sciences, Novosibirsk, 630090 Russia; E-mail: mike@bionet.nsc.ru*

[2] *Sobolev Institute of Mathematics, Siberian Division, Russian Academy of Sciences, Novosibirsk, 630090 Russia*

Received May 12, 2001

Abstract—A complex approach to recognition of transcription factor binding sites (TFBS) has been developed, based on four methods: (i) weight matrix, (ii) information content, (iii) multidimensional alignment, and (iv) pairwise alignment with the most similar representative of known sites. It has been shown that none of the methods considered is optimal for all kinds of sites, so in each case the appropriate way of recognition should be chosen. The approach proposed allows one to minimize the errors in TFBS recognition. The program available through the Internet (http://www.sgi.sscc.ru/mgs/programs/multalig/) has been created to search for the potential TFBS in nucleotide sequences set by the user.

Key words: transcription, recognition, binding sites, transcription factors, multiple alignment, promoter

INTRODUCTION

The development of methods recognizing transcription factor binding sites (TFBS) is important for computer annotation of genomic DNA. The widely applied approaches to TFBS recognition include consensus analysis [1–4], weight matrices [1], oligonucleotide matrices [5], estimation of physicochemical characteristics [6], information content [7, 8], neural networks [9], various statistical methods [10], etc. Despite the variety of approaches, one cannot regard the problem of precise TFBS recognition as completely solved nowadays.

The reason is the large variety of context, physicochemical, and conformational TFBS features; DNA–protein interactions between TFBS and transcription factors; the specificity of TFBS context in different kinds of regulatory regions (promoters, enhancers, silencers, locus-controlling regions, etc.); the extent of conservation of the context; and others.

At the same time, each of the approaches listed above takes into account certain peculiarities of the context or structural TFBS organization. A situation is typical when one method provides good results with one TFBS group and low accuracy of recognition with the other. Therefore, promising methods should be based on different modes of considering and revealing significant features of the TFBS structural organization and their context. To achieve this, it is necessary to develop methods taking into account the features of both TFBS and their neighboring DNA.

In the present work, we consider a complex approach to TFBS recognition based on methods known before and suggested by us: (i) weight matrix; (ii) information content; (iii) multiple alignment; and (iv) pairwise alignment with the most similar representative of known TFBS. We assess the mean accuracy of recognition to choose the optimal method for each kind of TFBS.

On the basis of experimental information on the structural–functional organization of TFBS accumulated in the TRRDatabase [11], we have developed methods recognizing 25 TFBS deposited in TRRD. For each TFBS set, four methods were created and integrated into the MMSite program available through the Internet (http://www.sgi.sscc.ru/mgs/programs/multalig/) allowing one to search for potential TFBS in unknown nucleotide sequences. The results of the analysis of eukaryotic gene promoter regions from TRRD have been demonstrated.

EXPERIMENTAL

Formation of the TFBS sets on the basis of the information from TRRD. The TFBS nucleotide sequences were extracted from the EMBL database on the basis of the information on site location deposited in TRRD [11] and intended for accumulation of experimental data on regions controlling eukaryotic gene transcription. These data were treated with the TRRD-Pars program created by us. The TFBS sets were formed out of 25 nucleotide sequences. The names of

0026-8933/01/3506-0818$25.00

the transcription factors under consideration are shown in the table.

As negative selections, we used the random sequences created as described below.

Multidimensional multiple alignment of nucleotide sequences. The first step of building methods of the TFBS recognition is the multiple alignment traditionally based on a pairwise alignment of all nucleotide sequences comprising a set, then building a similarity matrix, and on its basis, a similarity tree: step-by-step alignment of sequences, whose order is determined by the tree [12–15]. Such an approach comes from a suggestion that the considered group of nucleotide sequences has been originated from a common ancestor as a result of divergent evolution. The validity of this suggestion has been confirmed with the use of numerous extended nucleotide and amino acid sequences [12–15]. At the same time, the evolution of TFBS probably differs from that of extended

Characteristics of four methods of TFBS recognition

Factor	M1		M2		M3		M4		$E1_{min}$	M($E1_{min}$)	$E2_{min}$	M($E2_{min}$)
	*E*1	*E*2	*E*1	*E*2	*E*1	*E*2	*E*1	*E*2				
USF	5	1	0	24	0	11	0	15	0	M2, M3, M4	1	M1
CDP	33	27	33	27	0	42	0	4	0	M3, M4	4	M4
c-Fos/c-Jun	0	19	25	17	25	29	0	52	0	M1, M4	17	M2
c-Myc	25	6	0	32	0	3	0	21	0	M2, M3, M4	3	M3
CAN	0	13	0	18	0	8	0	8	0	M1, M2, M3, M4	8	M3, M4
CIIIB1	0	12	0	12	0	11	0	11	0	M1, M2, M3, M4	11	M3, M4
C/EBPδ	0	12	0	29	0	12	0	12	0	M1, M2, M3, M4	12	M1, M3, M4
E2F-1/DP-1	4	2	0	28	0	4	0	14	0	M2, M3, M4	2	M1
E2F	11	11	11	44	11	5	11	22	11	M1, M2, M3, M4	5	M3
EAR-2	0	9	0	9	0	9	0	9	0	M1, M2, M3, M4	9	M1, M2, M3, M4
EBP-1	0	8	0	7	0	7	0	7	0	M1, M2, M3, M4	7	M2, M3, M4
Elf1	0	2	0	19	0	2	0	19	0	M1, M2, M3, M4	2	M1, M3
GATA-3	27	11	0	68	0	1	0	54	0	M2, M3, M4	1	M3
HNF-3α	0	14	0	13	0	14	0	13	0	M1, M2, M3, M4	13	M2, M4
ISGF3	17	18	17	46	17	13	17	16	17	M1, M2, M3, M4	13	M3
LAP	4	5	2	16	2	2	0	23	0	M4	2	M3
NF-kB (p65)	0	4	0	2	0	2	0	2	0	M1, M2, M3, M4	2	M2, M3, M4
NF-Atp/c	31	15	15	73	8	4	8	64	8	M3, M4	4	M3
p53	17	48	67	44	17	1	17	34	17	M1, M3, M4	1	M3
Ptx1	0	1	0	1	0	1	0	1	0	M1, M2, M3, M4	1	M1, M2, M3, M4
STAT1	0	6	0	5	0	5	0	5	0	M1, M2, M3, M4	5	M2, M3, M4
TCF-1α	0	22	0	22	0	22	0	22	0	M1, M2, M3, M4	22	M1, M2, M3, M4
TTF-1	27	7	0	78	0	9	9	55	0	M2, M3	7	M1
XHSF1	0	28	0	13	0	0	0	0	0	M1, M2, M3, M4	0	M3, M4
Mean error	13	15	11	30	4	13	3	21	2		7	

Note: *E*1, first-type error in control (percentage).
*E*2, second-type error in control (percentage).
M1, weight matrix.
M2, information content.
M3, multidimensional alignment.
M4, alignment with the most similar representative.
$E1_{min}$, minimal first-type error for each TFBS type.
M($E1_{min}$), method providing the minimal first-type error.
$E2_{min}$, minimal second-type error for each TFBS kind.
M($E2_{min}$), method providing the minimal second-type error.

sequences: any eukaryotic genome contains an enormous amount of TFBS not connected by a common evolutionary origin and located in the 5'- and 3'-terminal gene regions.

Therefore, it seemed expedient to develop a method of TFBS analysis independent from the suggestions about their primary structure. This method based on the multidimensional alignment of nucleotide sequences is described below.

The multidimensional alignment is a generalization of a traditional two-dimensional alignment [12–15] that is performed in four steps.

(1) To align sequences of lengths L_1 and L_2, the two-dimensional matrix S is built of size $(L_1 + 1)(L_2 + 1)$. In the case of local alignment, elements of the first column and first row of the matrix are represented by zeros.

(2) To determine the internal elements according to (1), the maximum out of the three following values is required: (i) the value of the upper element minus penalty for deletion; (ii) the value of the left element minus penalty for deletion; and (iii) the value of the diagonal (upper left) element plus the value of nucleotide similarity:

$$S(i, j) = \max\begin{cases} S(i-1, j-1) + s(i, j) \\ S(i-1, j) - d \\ S(i, j-1) - d. \end{cases} \quad (1)$$

Here d is a penalty for deletion and $s(i, j)$ is the value of similarity between a nucleotide in position i of the first sequence and a nucleotide in position j of the second sequence.

(3) When the matrix S is completed, in the case of local alignment, the most weighted element $S(m_1, m_2)$ is chosen, where $0 \le m_1 \le L_1$ and $0 \le m_2 \le L_2$.

(4) From the chosen element $S(m_1, m_2)$, by reconstruction of the course of alignment, the desired alignment is obtained [12–15].

By analogy with the two-dimensional alignment, the multidimensional alignment is also performed in four steps.

(1) To align N sequences of length $L_1, L_2, \ldots, L_N$, the multidimensional matrix S of size $(L_1 + 1)(L_2 + 1)\ldots(L_N + 1)$ is built. In the case of local alignment, all elements located at the matrix sides are represented by zeros.

(2) To fill the internal matrix positions, the maximum is determined out of $2^N - 1$ values of neighbor elements (since the aligned sequences may contain 0 to $N - 1$ deletions) calculated as described for two-dimensional alignment.

(3) In the case of local alignment, the most weighted element $S(m_1, m_2, \ldots, m_N)$ is chosen, where $0 \le m_1 \le L_1, 0 \le m_2 \le L_2, \ldots, 0 \le m_N \le L_N$, according to (2):

$$S(m_1, m_2, \ldots, m_N) = \max_{i1 \le L1,\, i2 \le L2,\, \ldots,\, iN \le LN} (S(i_1, i_2, \ldots, i_N)). \quad (2)$$

(4) From the chosen element $S(m_1, m_2, \ldots m_N)$, by reconstruction of the course of alignment, the desired alignment is obtained.

Our program of multidimensional multiple alignment MMSite asks for parameter n, the number of sequences subjected to simultaneous alignment ($1 < n \le N$, where N is the number of all sites). If $n < N$, then alignment is done step-by-step by multidimensional alignment of n out of N sequences. Thus, the MMSite program can perform both multidimensional multiple alignment mode and regular, two-dimensional mode.

The methods of TFBS recognition were built on the basis of the multidimensional alignment.

Methods of TFBS recognition. Four methods were used to recognize TFBS: (i) weight matrix; (ii) information content; (iii) recognition by multidimensional alignment; and (iv) recognition by comparison with the most similar representative of known TFBS. Methods (1) and (2) are well-known [16]. Methods (3) and (4) were first used by us. In method (3), a potential TFBS is aligned with a set of aligned real TFBS, whereas in method (4), a potential TFBS is aligned with each real TFBS and the best alignment is chosen.

The decision rule for all methods was determined by one and the same mode. As an example, let us consider a procedure for building the decision rule for multidimensional alignment.

Let us designate a set of real TFBS containing N sites as Q_N. After the excluding of one TFBS designated as R_{site}, a set Q_{N-1} remains. R_{site} was aligned with a set Q_{N-1}, and the R_{site} weight $W(R_{site})$ was determined by formula (2). Thus, according to the procedure of multidimensional multiple alignment, weight $W(R_{site})$ reflects the similarity between a nucleotide sequence R_{site} and a set of TFBS.

The weight distribution $W(R_{site})$ was obtained by sequential excluding each R_{site} out of a set Q_N. In addition, the weight distribution $W(R_{rnd})$ was obtained for random sequences. On this purpose, one R_{site} was sequentially chosen out of a set Q_N to serve as a basis for generating a random sequence R_{rnd}, which was aligned with a set Q_{N-1}. The weight of this alignment determined by formula (2) was considered as weight $W(R_{rnd})$. Thus, according to the procedure of multidimensional multiple alignment, weight $W(R_{rnd})$ reflects the similarity between a random sequence R_{rnd} and a set of TFBS.

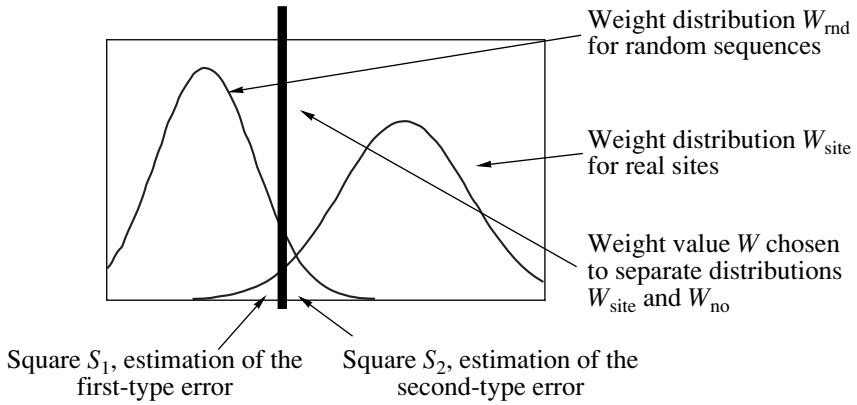


Fig. 1. Scheme of weight distributions for real sites (W_{site}) and random sequences (W_{rnd}). Weight distributions were obtained as described in the text.

The weight distributions W_{site} and W_{rnd} are schematically shown in Fig. 1. The prearranged border between these distributions cuts a part of W_{site} on the left (the square is designated as S_1) and a part of W_{rnd} on the right (S_2). These squares are estimations of the first- and second-type errors, respectively. In the methods developed by us, the border W_0 has been determined so that the correlation coefficient CC(W) characterizing the separation of two weight distributions (3) was maximal $W_0 = \arg\max(\mathrm{CC}(W))$.

$$\mathrm{CC}(W) = \frac{(1-S_1(W))(1-S_2(W)) - S_1(W)S_2(W)}{[(1-S_1(W)+S_2(W))(1-S_2(W)+S_1(W))]^{1/2}}. \quad (3)$$

Here, $S_1(W)$ and $S_2(W)$ are estimations of the first- and second-type errors, respectively, which depend on where the border separating the real and random TFBS is running (Fig. 1).

Let us consider now a nucleotide sequence X, which is under question whether it is of the TFBS type represented by a set Q. To build the weight distribution W_X for the sequence X, each sequence was excluded in turn from a set Q_N, and sets Q_{N-1} were aligned with the sequence X. As before, the alignment weight W_X was calculated according to (2). The mean value $W_{x\,mean}$ was determined for the weight distribution W_x. The following decision rule is used to recognize TFBS: if $W_{x\,mean} > W_0$, then X is a potential site; otherwise, X is not a site.

Such methods of the TFBS recognition as the weight matrix and information content were built by an analogous way. The matrix P of relative frequencies of nucleotides was built on the basis of a set Q_{N-1}. Then the recognized sequence X was aligned with a set Q_{N-1}. By using this alignment and the matrix P, the X weight was found according to formulas (4) and (5) for the weight matrix and information content, respectively:

$$W = -\sum_{i=1}^{L} \ln(P(a_i)), \quad (4)$$

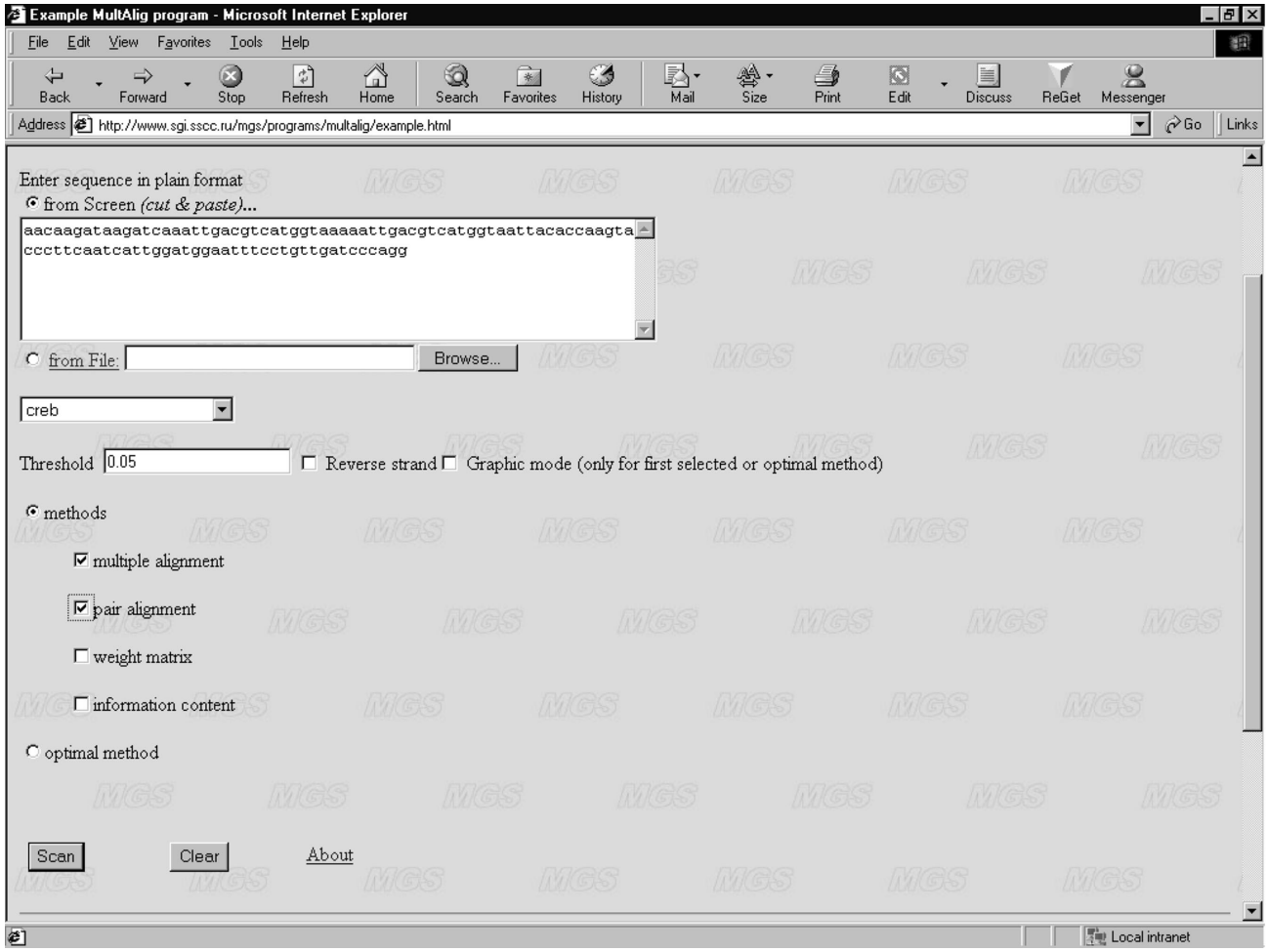
$$W = -\sum_{i=1}^{L} P(a_i)\ln(P(a_i)). \quad (5)$$

Here, L is the TFBS length, a_i is a nucleotide in position i of the sequence X, and $P(a_i)$ is an element of the matrix P in position i for the nucleotide a_i.

The method of TFBS recognition by alignment with the most similar representative is as follows. The sequence X is aligned in turn with each site of a set Q, and the alignment weight is estimated according to (2). The highest weight serves a characteristic of the similarity between X and the set Q. To obtain the decision rule, the distribution of these characteristics is built for sets of random and recognized sequences by the procedure described above (see building the decision rule for multidimensional alignment).

Method of search for potential TFBS. On the basis of four methods of recognition, the MMSite program has been developed to search for potential TFBS in extended nucleotide sequences. The program interface is shown in Fig. 2a. The user sets a nucleotide sequence and the name of transcription factor, whose binding sites are to be recognized, the second-type error in the field "Threshold," the direction of the introduced sequence by turning on or off "Reverse strand," and the mode of result presentation, text or graphic, by turning on or off "Graphic mode." In Fig. 2a, the search is set by two methods: the multidimensional alignment (designated as multiple alignment) and the pairwise alignment with the most similar representative (pair alignment).

(a)



(b)

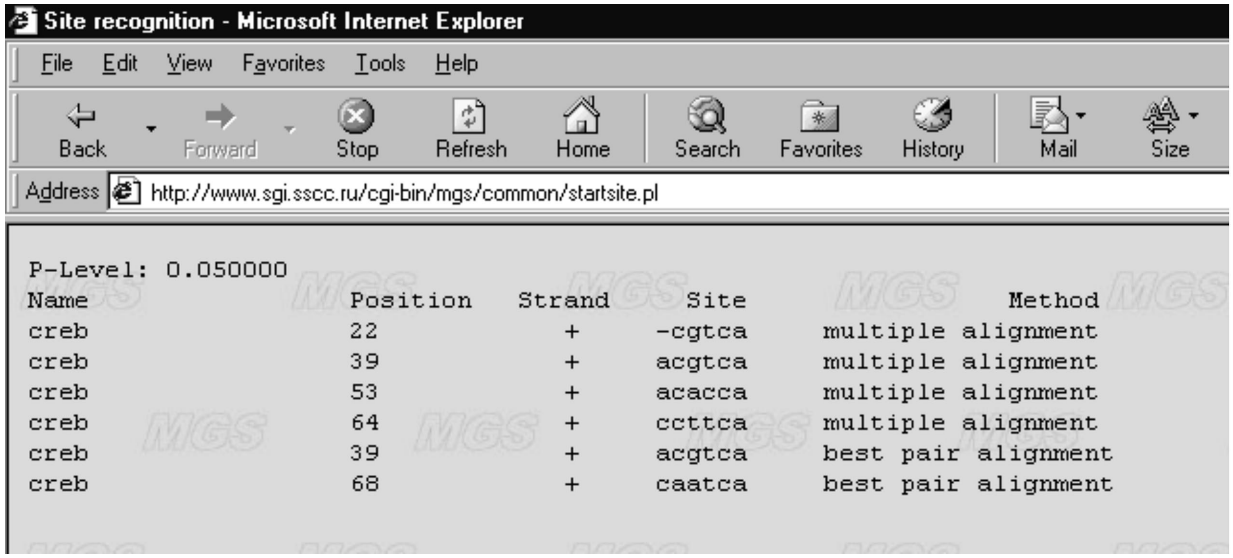


Fig. 2. The Web site with the MMSite program: (a) interface and (b) an example of the search results. The user can set a nucleotide sequence, indicate the name of the transcription factor whose binding sites are required for a search in the given sequence, and chose any combination of four recognition methods or "Optimal method." In the latter case, the program determines the most precise method for the given TFBS. In "Threshold," the user indicates the second-type error for the given TFBS.

An example of the results presented in text mode is shown in Fig. 2b. The first line indicates the significance level ("Threshold" value). Then follows a brief description of the columns produced by the program. The first column shows the transcription factor name; the second, positions of potential TFBS in the inquired nucleotide sequence; the third, the direction of search; the fourth, the sequence of potential TFBS with possible deletion symbols "–" designating the alignment of this region; and the fifth, the method name.

RESULTS AND DISCUSSION

Recognition of TFBS Described in TRRD

The table shows the first- and second-type errors for the methods created by us to recognize binding sites for 25 transcription factors, which were obtained for the control TFBS sets and the sets of random sequences that were not present in training sets. The first-type error $E1$ characterizes a part of unrecognized sites from a control set of real TFBS according to (6):

$$E1 = \frac{n^{-}_{\text{site}}}{N_{\text{site}}}. \quad (6)$$

Here n^{-}_{site} is the number of real TFBS recognized as non-sites from a control set and N_{site} is the number of sites in a control set.

The second-type error $E2$ characterizes a part of random sequences recognized as TFBS from a control set of random sequences according to (7):

$$E2 = \frac{n^{+}_{\text{rnd}}}{N_{\text{rnd}}}. \quad (7)$$

Here n^{+}_{rnd} is the number of random sequences recognized as TFBS from a control set and N_{rnd} is the number of random sequences in a control set.

The table shows the first- and second-type errors ($E1$ and $E2$) obtained for each TFBS recognized by four methods under consideration.

Obviously, in most cases, different methods yield different errors of TFBS recognition. For example, in the case of TFBS USF, errors $E1$ and $E2$ are respectively equal to 5 and 1% for the weight matrix method, 0 and 24% for the information content method, 0 and 11% for the multidimensional alignment, and 0 and 15% for the alignment with the most similar representative.

Methods with the least first- and second-type errors, $M(E1_{\min})$ and $M(E2_{\min})$, as well as their errors, $E1_{\min}$ and $E2_{\min}$, are shown for each TFBS in the last four columns of the table. For example, in the case of TFBS CDP, the least first-type error is provided by methods M3 and M4 (multidimensional alignment and alignment with the most similar representative), while the least second-type error is provided by method M4 (alignment with the most similar representative).

Obviously, no method ensuring the least errors, $E1_{\min}$ and $E2_{\min}$, for all considered TFBS occurs among these methods. Thus, method M3 (multidimensional alignment) provides the minimal error $E1$ for USF, CDP, c-Myc, CAN, CIIIB1, and C/EBP, while in the case of c-Fos/c-Jun, the least first-type error is provided by the weight matrix method.

The bottom line of the table shows mean errors $E1$ and $E2$ for each method. The methods may be ranged according to the increase of the mean first-type error as follows: (i) alignment with the most similar representative, 3%; (ii) multidimensional alignment, 4%; (iii) information content, 11%; and (iv) weight matrix, 13%.

According to the increase of the mean second-type error, the methods may be ranged as follows: (i) multidimensional alignment, 13%; (ii) weight matrix, 15%; (iii) alignment with the most similar representative, 21%; and (iv) information content, 30%.

One the basis of the results obtained, we can recommend the following strategy of TFBS recognition in unknown nucleotide sequences.

(1) To search for a certain TFBS, the method should be chosen providing the minimal first- or second-type error ($E1_{\min}$ or $E2_{\min}$) depending on a specific task. Thus, for long genomic sequences, methods providing the minimal second-type error $E2_{\min}$ are preferable to avoid prediction of the large amount of false TFBS. On the other side, to obtain full information about potential TFBS of a certain type in short sequences, it is expedient to apply a method with the minimal first-type error $E1_{\min}$ to avoid possible missing of real TFBS. As follows from the table, to search for potential TFBS E2F-1/DP-1 with the minimal first-type error, one should use multiple alignment or alignment with the most similar representative (first-type errors for these methods are equal to 0%), whereas for a search with the minimal second-type error, one should apply weight matrix method ($E2_{\min}$ is equal to 2%).

(2) When it is necessary to recognize simultaneously TFBS of different types in one nucleotide sequence with the minimal first-type error, it is recommended to use a combination of the methods indicated in column $M(E1_{\min})$ of the table. Such an approach provides the mean first-type error $E1_{\min}$ of 2%, which is substantially lower than $E1_{\min}$ for each method.

(3) In order to simultaneously recognize TFBS of different types in one nucleotide sequence with the minimal second-type error, one should apply a combination of the methods shown in the table column

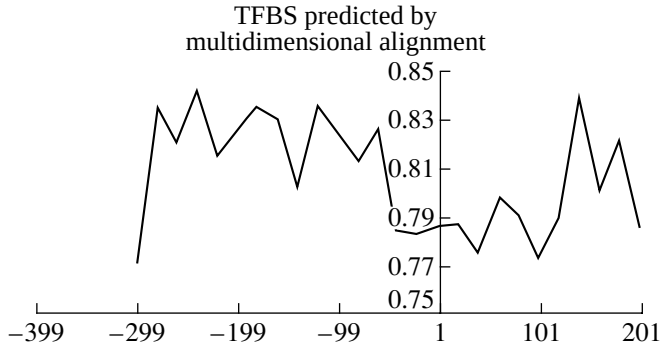


Fig. 3. Distribution of TFBS predicted by multidimensional alignment for eukaryotic gene promoter regions with positions –300 to +200 from the transcription initiation point. The X axis indicates the promoter positions from the transcription initiation point (position 1). The Y axis shows the frequency of TFBS calculated per promoter region. The set of promoter regions was created on the basis of the information from TRRD and included 516 sequences.

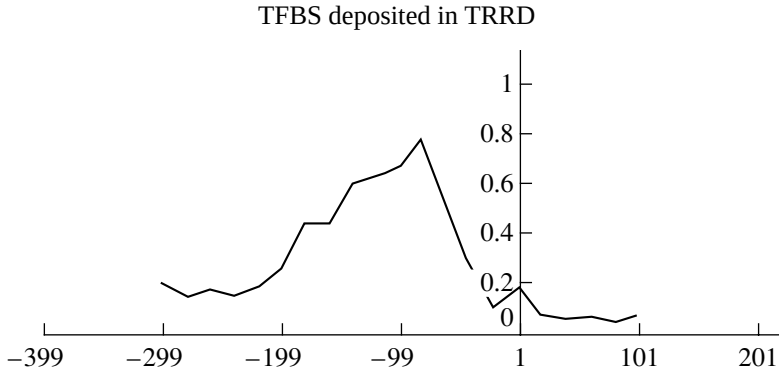


Fig. 4. Distribution of real TFBS in eukaryotic gene promoter regions with positions –300 to +200 from the transcription initiation point (on the basis of the information from TRRD). Axis designations as in Fig. 3. The set of promoter regions (516 sequences) was created on the basis of the information from TRRD.

M($E2_{min}$). Such an approach provides the mean second-type error $E2_{min}$ of 7%, which is substantially lower than $E2_{min}$ for each method taking separately.

Distribution of Potential TFBS along Promoters

Using the developed methods of TFBS recognition, we studied gene promoters (–300 to +200 from the initiation transcription point) deposited in the TRRD database. In total, 516 promoters were analyzed, in which 25 types of TFBS were recognized by multidimensional alignment. The data obtained are shown in Fig. 3. The distribution of real TFBS in promoter regions is represented in Fig. 4.

As follows from Figs. 3 and 4, the distribution of potential TFBS predicted by multidimensional alignment and that of real TFBS deposited in TRRD are specifically similar to each other in the region –300 to –1 and substantially differ in the region +1 to +200 from the initiation transcription point. The similarity of these distributions upstream the initiation transcription point testifies to the accuracy of the methods developed by us with regard to the prediction of annotated TFBS, as well as to the fact that a substantial part of real TFBS in gene promoters is already experimentally revealed.

At the same time, the number of predicted potential TFBS downstream the initiation transcription point is substantially greater than that of the experimentally revealed TFBS deposited in TRRD. This may suggest that the regulatory gene regions downstream the initiation transcription points are insufficiently studied. Therefore, according to the results of computer analysis, the study of these regions is of great interest since they may contain many new TFBS.

CONCLUSIONS

One of the current problems is the fast accumulation of new experimental data on the TFBS primary structure and their location in genomic DNA. Therefore, a “technological line” is required, which should

include the annotation of literature data containing information about TFBS, the deposition of the data in computer databases, the formation of TFBS sets and their analysis to build methods of TFBS recognition, and the accumulation of the methods in a computer database available through the Internet. Some important elements of this approach are described in the present work.

The main source of the data on TFBS and on the regulatory regions controlling transcription of eukaryotic genes is the database TRRD, continually updated [11]. It contains experimental information about the initiation transcription points, disposition of regulatory regions (promoters, enhancers, silencers, etc.) and TFBS according to the initiation transcription points, as well as references on nucleotide sequences from EMBL and GenBank. These data are processed by the TRRD-Pars program allowing one to extract the regulatory TFBS sequences from EMBL. Thus, the possibility of continual renovation of the TFBS data occurs to built methods of their recognition.

On the basis of the nucleotide sequence sets formed, a number of the methods of TFBS recognition has been built: (i) weight matrix, (ii) information content, (iii) multidimensional alignment, and (iv) alignment with the most similar representative. The results obtained by us indicate that the use of a combination of methods to recognize each type of TFBS allows one to decrease the first- and second-type errors.

This approach is suggested to be developed in the following directions.

(1) Extension of the number of TFBS groups involved in building the recognition methods (taking into account the fact that about 700 different TFBS are deposited in TRRD at present).

(2) Extension of a set of recognition methods, including such approaches as hidden Markov models [17], discriminant analysis [18], neural networks [9], methods based on the context-dependent conformational and physicochemical characteristics of DNA [6], methods of knowledge discovery and regularity search (Vityaev *et al.*, this issue), and others. It is suggested that the use of a large amount of methods allows one to minimize the recognition errors.

(3) Building the TFBS recognition methods not only on the basis of the TFBS central regions (as in the present work), but also on the basis of their flanking regions.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project nos. 00-04-49229, 00-04-49255, 00-07-90337, 01-07-90376, and 99-07-90203), by the Human Genome program, and by the Integration project of the Siberian Division of Russian Academy of Sciences (no. 65).

REFERENCES

1. Schneider, T. and Stephens, R., *Nucleic Acids Res.*, 1990, vol. 18, pp. 6097–6100.
2. Ulyanov, A. and Stormo, G., *Nucleic Acids Res.*, 1995, vol. 23, pp. 1434–1440.
3. Kel, A.E., Kondrakhin, Y.V., Kolpakov, Ph.A., Kel, O.V., Romashenko, A.G., Wingender, E., Milanesi, L., and Kolchanov, N.A., *Proc. Third Int. Conf. on Intelligent Systems Mol. Biol.*, 1995, pp. 197–205.
4. Kondrakhin, Y.V., Kel, A.E., Kolchanov, N.A., Romashenko, A.G., and Milanesi, L., *CABIOS*, 1995, vol. 9, pp. 1–13.
5. Ponomarenko, M.P., Ponomarenko, J.V., Frolov, A.S., Podkolodnaya, O.A., Vorobiev, D.G., Kolchanov, N.A., and Overton, C., *Bioinformatics*, 1999, vol. 15, pp. 631–643.
6. Ponomarenko, J., Ponomarenko, M.P., Frolov, A.S., Vorobiev, D.G., Overton, C., and Kolchanov, N.A., *Bioinformatics*, 1999, vol. 15, pp. 654–668.
7. Schneider, T., Stormo, G.D., and Gold, L., *J. Mol. Biol.*, 1986, vol. 188, pp. 415–431.
8. Papp, P. and Chattoraj, D., *J. Mol. Biol.*, 1993, vol. 233, pp. 219–230.
9. Horton, P. and Kanehisa, M., *Nucleic Acids Res.*, 1992, vol. 20, pp. 4331–4338.
10. Sewell, R. and Durbin, R., *J. Comput. Biol.*, 1995, vol. 2, pp. 25–31.
11. Kolchanov, N., Podkolodnaya, O., Ananko, E., Ignatieva, E., Stepanenko, I., Kel-Margoulis, O., Kel, A., Merkulova, T., Goryachkovskaya, T., Busygina, T., Kolpakov, F., Podkolodny, N., Naumochkin, A., Korostishevskaya, I., Romashchenko, A., and Overton, G., *Nucleic Acids Res.*, 2000, vol. 28, pp. 298–301.
12. Apostolico, A. and Giancarlo, R., *J. Comput. Biol.*, 1998, vol. 5, pp. 173–196.
13. Fernandez-Baca, D., Seppalainen, T., and Slutzki, G., *Proc. 11th Annu. Symp. on Combinatorial Pattern Matching, Lecture Notes Computer Sci.*, Berlin: Springer, 2000, no. 1848, pp. 69–83.
14. Subbiah, S. and Harrison, S.C., *J. Mol. Biol.*, 1989, vol. 209, pp. 539–548.
15. Taylor, W.R., *Comput. Appl. Biosci.*, 1987, vol. 3, pp. 81–87.
16. Gelfand, M.S., *J. Comput. Biol.*, 1995, vol. 2, pp. 87–115.
17. Durbin, R., Eddy, S.R., Krogh, A., and Mitchson, G., *Biological Sequence Analysis*, Cambridge: Cambridge Univ. Press, 1998.
18. Zhang, M.Q., *Genome Res.*, 1998, vol. 8, pp. 319–326.