Application of a new method of pattern recognition in DNA sequence analysis: a study of *E.coli* promoters

Nickolai N.Alexandrov and Andrei A.Mironov
Genetics of Microorganisms Institute, 1-st Dorozhny pr. 1, Moscow 113545, USSR

Received October 18, 1989; Revised and Accepted March 12, 1990

ABSTRACT

An algorithm from the pattern recognition theory 'generalized portrait' was used to find a distinguishing vector (scoring matrix) for *E.coli* promoters. We have attempted to solve three closely linked problems: (i) the selection of significant features of the signal; (ii) subsequent multiple alignment and (iii) calculation of the vector coordinates. Promoters with known strength have been successfully ranked in the correct order using this vector. We demonstrate the use of this method in predicting the location of promoters. A revised consensus promoter sequence is also presented.

INTRODUCTION

Signal localization within DNA sequences is one of the most important objectives in computer analysis of biological sequences. Many approaches have been developed to solve the site recognition problem, but all have inherent limitations. Common to all of these approaches is the following optimistic proposition: if the primary structure of several DNA fragments responsible for the same function is known then one can deduce specific features of this signal and recognize such signals in other nucleotide sequences. Thus, previous attempts have been made to recognize promoters [1–4], ribosome-binding sites (RBS's) [5], intron-exon junctions [6], terminators [7] and many other functional sites.

Approaches to recognition

Historically the first recognition site algorithm was designed to search DNA sequences for similarity fragment homologous to the consensus sequence. This method is usually applied when we have little data for learning and provides, as a rule, unreliable predictions. The exclusions are simple signals such as restriction sites.

The increase in the volume of learning data permits us to employ a statistical method for the analysis of functional sites. It usually consists of the construction of a recognition matrix with 4n elements, where n is the number of considered nucleotides. Each matrix element is based on the frequency of a certain nucleotide at a certain position. The homogeneity of the data makes the use of simpler statistical methods inappropriate. Specifically, the occurrence of several identical sequences or of very similar mutant derivatives is undesirable. Morover these methods cannot solve the recognition matrix existence problem (see below).

Previous attempts to apply the pattern recognition theory and discrimination analysis seem quite promising. The specificity of such algorithms is that they employ not only the compilation of site sequences, but also a set of DNA sequences which are not signals (non-sites). The first results of a well-known algorithm, Perceptron, were obtained by Stormo [5] for RBS's. Iida [6] has used the discrimination theory for localizing splice junctions. We report here the results of an application of the algorithm 'generalized portrait', originally developed by Vapnik et.al. [8], for signal recognition, using *E.coli* promoters as an example.

RECOGNITION STRATEGY

We dissect the recognition problem into the following tasks:
1. Compilation of site;
2. Multiple alignment;
3. Choice of significant sign;
4. Calculation of the recognition matrix or discriminative vector.

Ideally, in compiling sites, all of them should have been characterised by the same experimental method. Similarly for the set of non-sites, all of them must be tested in the same experiments.

The next three problems are closely associated and the order of their solving is not strictly defined. In order to align the sites, for instance, it is necessary to know the significance or the 'weight' of nucleotides within the site. On the other hand, these weights are defined by the recognition matrix which depends on the alignment. The iterative procedure, presented in Fig.1 provides a possible way of solving this problem.

COMPILATION

E.coli promoters are well-investigated signals. Their most conserved features comprise two boxes of six nucleotides approximately 10 and 35 bases upstream from the transcription origin. The spacer between the boxes varies in the range 15 to 21 nucleotides. The distance between the '−10' box and transcription origin can range from 4 to 8, according to Hawley and McClure [9], and from 4 to 12 according to Harley and Reynolds [10].

All data were taken from compilations [9] and [10]. We did not take promoters with undefined or multiple transcription origins, neither those from compilation [10] which have a distance between the −10 box and transcription origin of more than 8 or less then 4 nucleotides. Each promoter is represented by a sequence of approximately 60 bases with the first transcribed

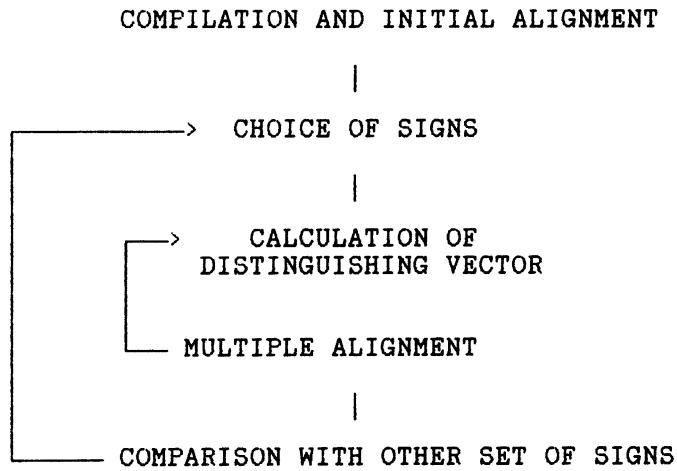

Figure 1. The general algorithm for signal recognition.

Table 1. Features used for promoter recognition. Besides the two boxes of 16 nucleotides, the spacer between them was considered.

Feature number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
-35 box												T	T	G	A	C	A

Feature number	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
-10 box								T	A	T	A	A	T			

nucleotide marked. Since almost all promoters were located in different experimental conditions it is incorrect to compare different sequences with each other. Hence, the aim of learning cannot be in finding such a recognition matrix for which any promoter has a greater weight than that for all non-promoters. Instead, finding the matrix which provides a maximum weight for the promoter *within* the corresponding sequence from compilation is a more correct aim. Regulated promoters were included in the compilation since it was shown that the regulators alter the promoter strength but do not change the transcription origin [11].

Finally, the compilation consisted of 166 promoters, each within a sequence of 50–60 nucleotides. All the other possible meanings of the set of features (including complementary strand) represented our non-promoter compilation. Thus each promoter has its own non-promoter compilation of approximately 200 non-sites. The total compilation was divided into two equal parts, the learning and the control compilations.

CHOICE OF FEATURES

Firstly the learning was made with the standard set of features consisting of two hexanucleotide boxes separated by 15–21 nucleotides. Then the boxes were expanded to 16 nucleotides (Table 1) and the optimal set of features was found. One more feature was the distance between these boxes, which has 7 values. The set of features may be formed by different types of sequence characteristics, such as the presence of secondary structure, G–C content and so on. The algorithm of learning does not depend on the set of features, only the sequence coding is changed.

Table 2. Dinucleotide features with the most significant x^2 values. The value before the colon is the expected number of dinucleotides where as the one after is the actual number.

Feature number	22				
		A	C	G	T
32	A	6: 8	8: 3	17:21	9:10
	C	4: 6	5: 4	12:12	7: 9
	G	6: 6	7:17	17: 8	9:10
	T	5: 4	7: 5	15:22	8: 6

x^2=24

Feature number	6				
		A	C	G	T
20	A	12:19	8: 6	11: 5	11:15
	C	9: 5	6:11	8:13	8: 3
	G	9: 6	6: 4	8:11	8:12
	T	11:13	8: 9	10: 9	10:10

x^2=23

Feature number	23				
		A	C	G	T
31	A	9:12	8: 5	10:11	13:13
	C	9: 4	8:18	10: 6	13:13
	G	6: 8	5: 6	7: 7	8: 7
	T	9:10	8: 2	10:14	13:15

x^2=22

Among other types of features only one was investigated—pairs of nucleotides at certain positions in the −10 and −35 boxes. This class of features characterizes the correlation of different positions in boxes. The statistical analysis of independent promoters was performed to reveal the significance of such dinucleotide features. The quantitative measure of correlation between features k and l was the following function:

$$x^2 = \sum_i \sum_j (n_{kl}^{ij} - n_s)^2 / n_s,$$

where n_{kl}^{ij} is the number of promoters with the k-th feature equaled i and the l-th one—j; n_s is expected number of such promoters: $n_s = n_k^i \cdot n_l^j / n_p$, where n_p is the total number of promoters. The degree of freedom equals 9. After all the mutant derivatives were excluded, the compilation consisted of 151 promoters. The features that have significant correlation ($x^2 > 20$) are shown in Table 2. These results did not permit us to conclude that dinucleotide features were correlated. We therefore decided to ignore them during learning.

Encoding the sequences

Each feature has several sets of values: the nucleotide at a certain position in the box can be A, C, G or T; the spacer between boxes alters at defined intervals of natural numbers; energy of the secondary structure has a continuous spectrum of real numbers. Thus, the sequences may be encoded into an ordered set of numbers or into a vector with coordinates corresponding to features.

It is more convenient to deal with binary features, i.e. features with two possible values—0 and 1. Any feature with k values can be changed by k binary features. In the class of features 'nucleotide at certain position' each nucleotide is encoded by four binary features and the feature 'spacer between boxes', by $l_{max} - l_{min} + 1$ binary ones, where l_{max} and l_{min} are the maximal and minimal distance between boxes.

As a result each sequence is encoded by a binary vector. For example, subsequence CTGT corresponds to vector (0100 0001 0010 0001). Nucleotides are ordered alphabetically so the first

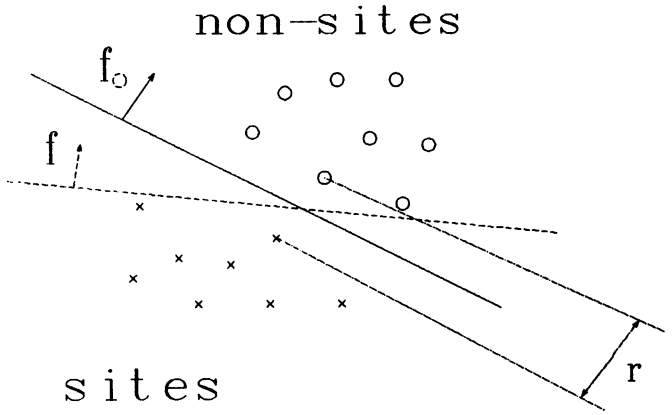

Figure 2. There are some different hyperplanes that discriminate the sets of sites and non-sites. The algorithm 'generalized portrait' provides the maximum distance, r, between the two sets.

number at every group of four features corresponds to A, second to C, third to G and the fourth to T.

Positing the problem

Now we can formulate the recognition problem more strictly. We say that two sets X and Y are distinguishable, if there exists vector f and the real number r such that for all $x \epsilon X$ and $y \epsilon Y$ $(x \cdot f) > r$, $(y \cdot f) < r$. To solve the recognition problem one must find distinguishing vector, f. Just now the problem is the same as formulated in [5]. Note that the terms recognition matrix and distinguishing vector are equal.

CALCULATION OF THE DISTINGUISHING VECTOR

Geometry interpretation

Many ideas of the recognition theory may be illustrated geometrically. By encoding the sequence into a vector we can represent it as a point in the feature space. If sites are shown as crosses and non-sites as circles then we will obtain a pattern as shown on Fig.2. The problem is to find a hyperplane which distinguishes the two sets.

The distinguishing vector is directed perpendicular to the hyperplane and defines its orientation in the feature space, and the discriminating number r defines its position. Any point z of the distinguishing hyperplane satisfies the equation: $(z \cdot f) = r$. As shown on the figure there are several distinguishing planes. PERCEPTRON [1,5] finds one of these planes and some of its disadvantages stem from the small value of distance r between the sets. We have used the algorithm 'generalized portrait' to find the distinguishing plane which provides the maximum value of r.

Algorithm 'generalized portrait'

This algorithm, described by Vapnik et.al. [8], required minor modification for solving our problem. Let the discrimination number, r, be equal 1. It is needed only for normalization and does not change the recognition problem. The distinguishing conditions can then be written as:

$$(f \cdot x) \geq 1 \qquad (1)$$
$$(f \cdot y) \leq k, \text{ where } k < 1$$

The generalized portrait is the minimal module of all possible distinguishing vectors, f, which provides the maximum of

Table 3. The distinguishing vector obtained with algorithm 'generalized portrait'. Feature 33 corresponds to the box spacer.

Feature number	A	C	G	T	"Pribnow" and "Gilbert" boxes
1	0	0	0	0	
2	0	0	0	0	
3	0.497	-0.587	1.163	-0.074	
4	0	0	0	0	
5	0	0	0	0	
6	0	0	0	0	
7	-0.283	-1.139	0.489	0.933	
8	0	0	0	0	
9	0	0	0	0	
10	-1.690	-0.285	0.065	1.910	T
11	-1.976	-0.289	-0.469	2.734	T
12	-2.971	0.252	3.209	-0.489	G
13	2.126	-0.938	-1.351	0.163	A
14	0.818	0.826	-0.885	-0.759	C
15	0.672	-0.565	0.444	-0.550	A
16	-0.942	0.153	-0.163	0.953	
17	0	0	0	0	
18	0	0	0	0	
19	0.579	-0.315	-0.527	0.264	
20	0	0	0	0	
21	-0.445	0.380	-0.813	0.878	
22	0	0	0	0	
23	0	0	0	0	
24	-0.048	-0.143	-1.122	1.314	T
25	3.409	-1.320	-1.490	-0.598	A
26	0	0	0	0	T
27	1.651	-0.832	0.103	-0.923	A
28	0.605	0.218	-1.227	0.403	A
29	-1.476	-0.638	0.000	2.114	T
30	0	0	0	0	
31	0	0	0	0	
32	0	0	0	0	

L	15	16	17	18	19	20	21
33	-0.221	-0.214	1.714	1.400	-0.044	-1.955	-0.678

distance, r. Let us create set $Z = \{z_{ij} = x_i - y_j\}$, where i and j run independently all possible values. The finding of the generalized portrait equals to definition of the module minimum vector f_0, satisfying condition $(z_{ij} \cdot f_0) \geq 1 =$, for all $z_{ij} \epsilon Z$. It is possible to prove that f_0 is represented as the sum:

$$f_0 = \Sigma\Sigma z_{ij} \cdot a_{ij}$$
$$\text{and } a_{ij}(z_{ij} \cdot f_0 - 1) = 0;\ a_{ij} \geq 0.$$

Vectors z_{ij}, for which $z_{ij} \cdot f_0 = 1$ are called informative, and x_i and y_j that form z_{ij} are called boundary. Thus the generalized portrait may be represented as a linear combination of boundary vectors. It deduces the hyperplane building problem to maximization of quadratic form $Q(a) = \Sigma\Sigma a_{ij} - 1/2(f \cdot f)$. The distance between sets will be $r(f) \geq 1 = /\sqrt{(2Q(a_0))}$. If the volume Q(a) becomes greater then the a priory given Q_0 it will mean that the distance between two sets is less than the given r_0 value. This criterion permits us to make a conclusion on the undiscriminatability of sets.

The maximum point Q(a) can be found by the following procedures. Choose an initial group of vectors z_{ij}, build a quadratic form Q(a) on it, find its maximum point by the conjugated gradient method and find the corresponding vector f_{01}. Then find vectors x_m and y_m which provide extreme values of the scale product:

$$x_m \cdot f_{01} = \min(x_i \cdot f_{01})$$
$$y_m \cdot f_{01} = \max(y_j \cdot f_{01})$$

If discriminating condition (1) is not performed, create vector $z_m = x_m - y_m$, add it to the selected group and make new iteration. Continue until condition (1) will be performed or until discrimination will be impossible $(Q(a) > Q)$.

Upon solving the site recognition problem we build the non-site set Y_i for each element $x_i \epsilon X$. This set contain all vectors formed from i-th sequence, apart from x_i. If long sequence

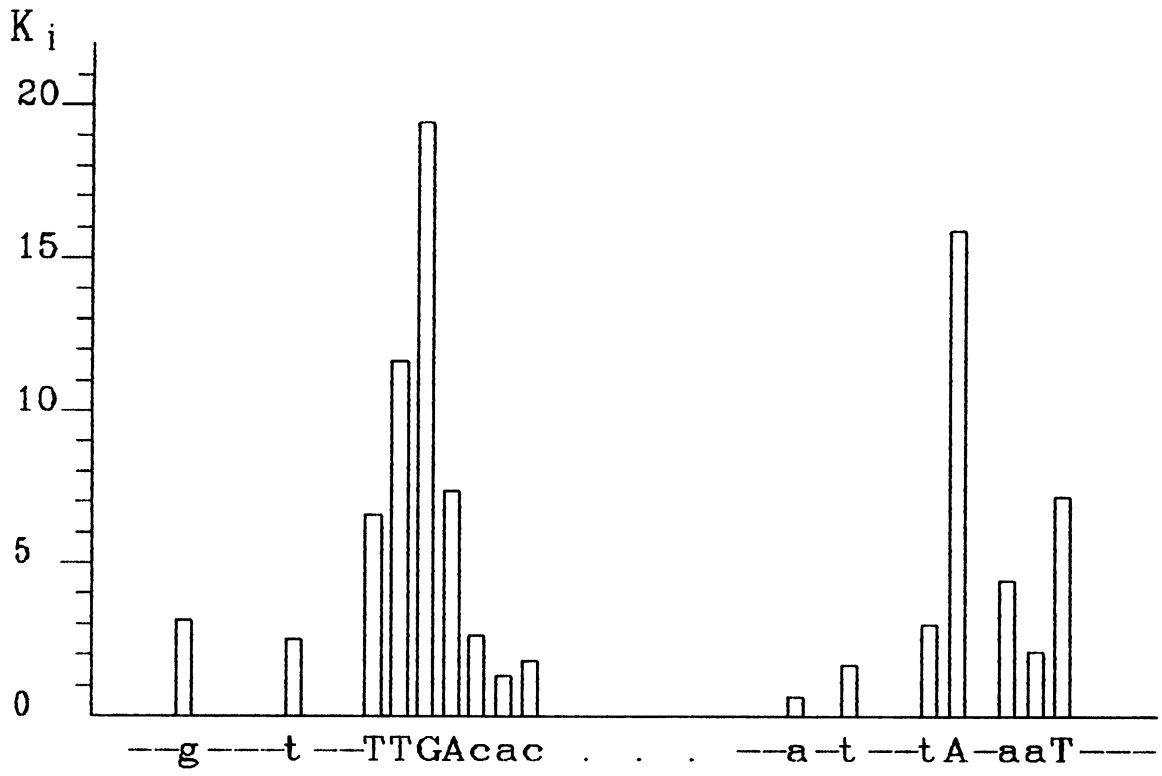

Figure 3. The significant features of *E.coli* promoter signal as derived from this study. A dashed line indicates a single nucleotide.

contains two or more promoters we do not include other promoters and regions surrounding them (± 100 bp) because of possible crypt signals. Thus we make the discriminating value r different for each sequence. The latter being more correct because the experimental methods used for identifying the promoters were not identical for all sequences.

Strength of sites

Set Z is significantly expanded if we use information about the relative efficiency (strength) of sites.If site x_i is stronger then site x_j, we include all elements of set Y_j and x_j in set Y_i. The identity of experimental conditions is very important here.

Exclusion of ambiguous experimental data

The possibility that incorrect data are included in the learning set always exists. The probability of mistakes is especially large within the set of sites because of the large spectrum of methods involved and their complexity. Our algorithm can exclude ambiguous data. We choose the sequences most difficult for distinguishing and then excluded them when discrimination became impossible. If the total number of excluded sequences will be greater than 50 we decided that the separation is impossible.

Optimal set of features

Unessential features can be excluded by our program. We ordered features according to sum: $K_i=\Sigma(f_{it})^2$, where f_{it} corresponds to the t-th gradation of the i-th feature. In our case t changes from 1 to 4 or from 'A' to 'T'. We excluded features with the lowermost values of K_i until the discrimination became impossible.

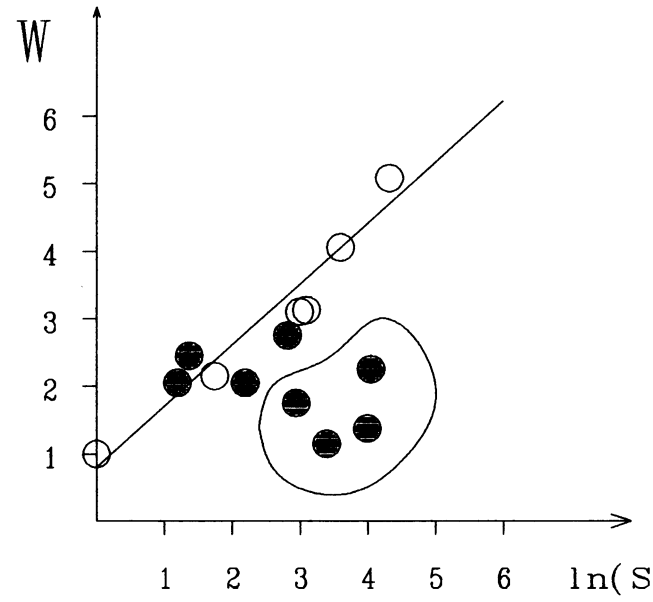

Figure 4. Correlation between promoter 'weight' $w=(f\cdot x)$ and log(s), where s is the promoter strength measured in *bla* units [12,13]. White circles correspond to promoters in learning set. T5 phage promoters (encircled) are separated from the rest.

Multiple alignment

We used an iterative algorithm to choose the box positions. Initial alignments were formed by the statistical vector found in [2]. The following iterations were obtained using 'generalized

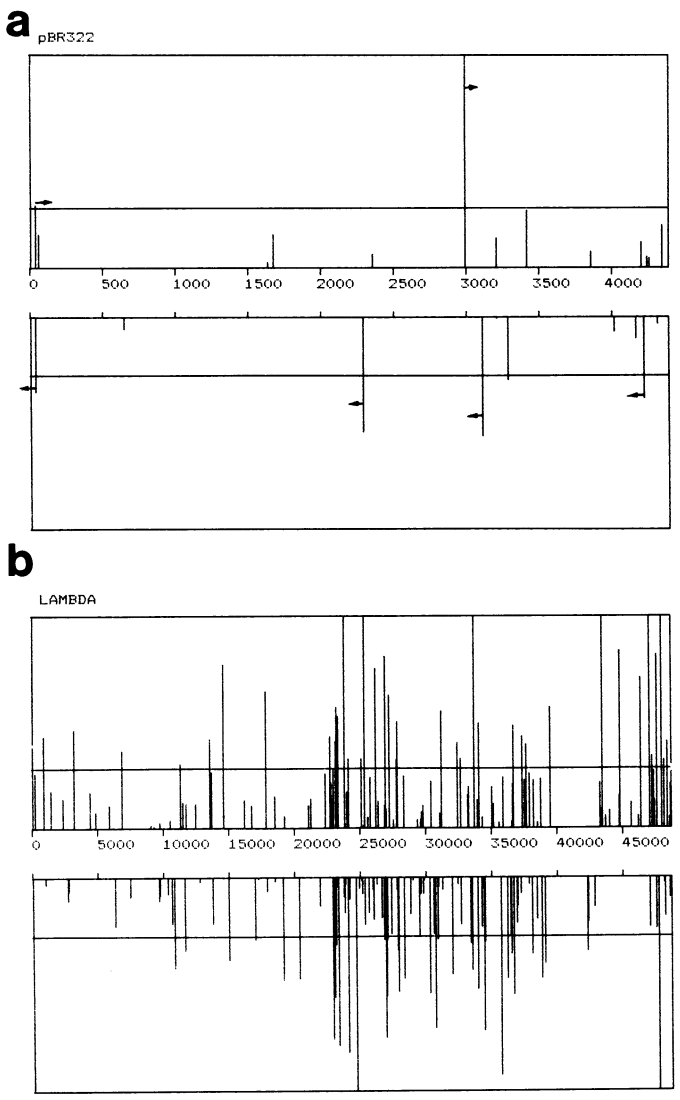

Figure 5. Distribution of the promoter-like sites along DNA sequences. a) plasmid pBR322, EcoR1-digested; known promoters are marked with arrows. b) phage lambda. The two graphs for each sequence represent each strand.

portrait'. New alignments were built until the positions of −10 and −35 boxes became fixed. The algorithm is similar to the one described by Bains (alignment by consensus) [12]. This method seems sufficient if the distinguishing vector is approximately known. The general scheme of the algorithm is shown on Fig.1.

RESULTS AND DISCUSSION

Distinguishing vector

The distinguishing vector was first calculated using only canonical features of *E.coli* promoters. They comprise two hexanucleotide sequences separated by 15−21 bases. The number of binary features equals $N=4\cdot6+4\cdot6+7=55$. The result of learning revealed that the discrimination with such a set of features was impossible. To solve the problem we included new features by increasing the box sizes. The size of each box became 16 nucleotides. The number of binary features—$N=4\cdot16+4\cdot16+7=135$. This set of features permitted us to calculate the distinguishing vector which precisely discriminated sets of promoters and nonpromoters in the 166 sequences tested.

The control of distinguishing vector was achieved by formation of two random compilations of promoter sequences. One was used for learning and the other for control. The control of learning should be performed for sequences which belong to defined classes: promoters or nonpromoters. It is wrong to test the distinguishing vector for example with the pBR322 sequence because of stringent transcriptional analysis deficiency. In all sequences from the compilation we know only the transcription initiation point. We suppose that the distance between the −10 and −35 boxes can have seven values and between the −10 box and the first transcribed nucleotide—five. So 35 different positions of the boxes correspond to the known initiation point. We suppose the promoter to be found correctly if the scalar multiplication ($f\cdot x$) reaches maximum at one of the 35 possible positions of −10 and −35 boxes. The number of all positions of boxes equals 200. The probability of the correct promoter localization with the random vector equals 0.2.

To perform complete discrimination, it was necessary to exclude the following promoter sequences from the compilation (as named in ref. 9): tyrT/6, ompR, micF, Tn501merR, Pori-r; the proposed positions of the −10 & −35 boxes of these sequences are not within acceptable distances from the transcription initiation point. More precise determination of the transcription start point may be required for these promoters.

According to the recognition theory the probability of the right classification of control objects after learning increases with the growth of the learning set and decreases with the growth of the number of features. By eliminating unimportant features we did indeed decrease the number of errors in the control set.

So, upon learning with the set of 83 sequences, the distinguishing vector was obtained. The vector gives three errors in the learning set of sequences. Testing with the control set of the other 83 sequences revealed 12 errors. For example, the statistical vector from reference [2] gives 19 errors in the first set and 15 in the second. Table 3 presents the vector, obtained from the learning set of 83 promoter sequences and pBR322. From the initial set of features consisting of 32 nucleotides, only 16 were chosen as the most important (Fig.3). The number of binary features becomes 71, versus 127 for the statistical vector.

Previous studies [13,14] have provided data on actual promoter strengths. We used only qualitative relations between promoters, such as: *lac* promoter is stronger than the *bla* promoter. Nevertheless the quantitative relations are in good agreement with experimental data. We used six sequences for learning, the predicted strength is shown on Fig.4. There is a logarithmic dependence between weight and promoter strength. The strength of phage T5 promoters have bad agreement with the predicted value probably due to the fact that some important features for T5 promoters were not included in our set [15]. The discrimination of the total compilation with strength arrangement is impossible even when an extended set of 32 features is employed.

Promoter localization within DNA sequences

For localizing *E.coli* promoters the program LOCSUN was written for IBM PC compatible computers. The program is included in the DNASUN package [16]. Examples of the use of this program are presented in Fig.5. The same analysis was made earlier using a statistical vector [17]. In spite of similarity between the distinguishing vectors the patterns of promoter

localization differ significantly. This difference is seen clearly when using random sequences (data not shown). It is surprising but the true sites as a rule give similar peaks for different vectors.

CONCLUSION

There are certain advantages for employing the algorithm 'generalized portrait', rather than the statistical methods or the simple algorithm PERCEPTRON. In particular, the former permits us to give an answer on the possibility to discriminate sets of sites and non-sites; to chose the optimal orientation of the discriminating hyperplane and the optimal set of site features.

ACKNOWLEDGEMENTS

We thank Dr K.Smith for carefull reading of manuscript and helpfull discusions.

REFERENCES

1. Alexandrov N.N, Mironov A.A. (1987) Molecular. biology (USSR) *20*, 242–249.
2. Mulligan M.E., Hawley D.K, McClure W.R. (1984) Nucl. AcidsRes. *12*, 789–800.
3. Staden R. (1984) Nucl. Acids Res. *12*, 505–519.
4. O'Neill M.C. (1989) J. Biol. Chem. *264*, 5522–5530.
5. Stormo G.D., Schneider T.D., Gold L., Ehrenfeucht A. (1982) Nucl. Acids Res. *10*, 2997–3011.
6. Iida Y. (1987) CABIOS *3*, 93–98.
7. Brendel V., Trifonov E.N. (1984) Nucl. Acids Res. *12*, 4411–4427.
8. Vapnik V.N., Glazkova T.G., Kotcheev V.A., Mikhalsky A.I., Chervonenskis A.Ya. In: Algorithms and programs for regression analysis (1984) M.: Nauka, 816p.
9. Hawley D.K., McClure W.R. Nucl. Acids Res. (1983), *11*, 2237–2255.
10. Harley C.B., Reynolds R.P. Nucl. Acids Res. (1987), *15*, 2343–2361.
11. Reznikoff W.S., Magnat L.E., Munson L.M., Johnson R.C., Mandecki W. In: Promoter: Structure and Function. (1982) N.Y.: Praeger Publishers, 80–95.
12. Bains W. Nucl. Acids Res. (1986), *14*, 159–177.
13. Mashko S.V., Gorovits R.L., Trukhan M.E., Demchuk E.Ya., Lebedeva M.I., Lapidus A.L., Podkovyrov S.M., Kozlov Yu.I, Debabov V.G. (1986) Dokl. AN USSR *291*, 1510–1513.
14. Deuschle U., Kammerer W., Gentz R., Bujard H. (1986) EMBO J. *5*, 2987–2994.
15. Gentz R., Bujard H. (1985) J.Bacteriol. *164*, 70–77.
16. Mironov A.A. In: Application of data bases and microcomputers in molecular biology. (1988) Jena, 66.
17. Mulligan E.M., McClure R.W. (1986) Nucleic Acids Res. *14*, 109–126.