Molecular Biology, Vol. 35, No. 6, 2001, pp. 810–817. Translated from Molekulyarnaya Biologiya, Vol. 35, No. 6, 2001, pp. 952–960.

UDC 577.214.625

Computer System "Gene Discovery" to Search for Patterns in Eukaryotic Regulatory Nucleotide Sequences

E. E. Vityaev[1], Yu. L. Orlov[2], O. V. Vishnevsky[2], A. S. Belenok[2], and N. A. Kolchanov[2]

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

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

Received May 31, 2001

Abstract—A method is proposed to automatically search for patterns in the mutual location of context signals in regulatory DNA sequences. The procedure is based on the methods of Data Mining and Knowledge Discovery software, implemented in a computer system Gene Discovery. This system was used to study erythroid-specific promoters and promoters of the endocrine-system genes from TRRD. We detected some trends in occurrence and localization of specific oligonucleotide groups.

Key words: recognition of eukaryotic promoters, Machine Learning, Knowledge Discovery, Data Mining, oligonucleotide motifs, transcription factor binding sites

INTRODUCTION

It is essential to study promoter structure to understand the mechanisms of transcription in eukaryotes. The basic element of transcription initiation is a core (basal) promoter, that is, a minimal DNA sequence required for correct initiation of gene transcription *in vitro* [2]. The core promoter includes the transcription start point within the region –60 to +40 bp around it [3, 4]. Each regulatory region contains binding sites specific for certain transcription factors [5]. Occurrence and distribution of the sites to bind transcription factors in 5'-terminal regulatory gene regions reflects tissue- and stage-specific patterns in regulation of their expression. A gene may have multiple promoters functioning to start translation resulting in various protein products or in products of various level of specific functional activity. Moreover, a common feature of eukaryotic promoters is the absence of exact location of the context signals essential for functioning, and these signals are rather weak [6].

Diversity of the gene promoter structure creates great problems in development of the software for promoter recognition. Numerous methods were proposed to recognize promoters of RNA polymerase II in eukaryotic genomes [7–10], however, as a whole, the problem how to increase the efficiency of promoter recognition remains unsolved.

In order to search for the structural organization patterns of eukaryotic promoters from databases, we developed the computer system "Gene Discovery" (basing on the system "Discovery" [11, 12]), which contains a module to search for the patterns in tables of molecular biological data showing significant correlation with the class of regulatory sequences, and the module to recognize promoters in a nucleotide sequence basing on the selected patterns. The software is written in C++ using VisualStudio 6.0 and designed for interactive work on PC. The system "Gene Discovery" was used to study the promoters of the endocrine-system genes from the database of eukaryotic transcription regulatory regions TRRD [1], and the promoters of erythroid-specific genes from database EpoDB [13]. We found a range of trends in occurrence and distribution of specific oligonucleotide motifs within these promoters.

EXPERIMENTAL

"Gene Discovery": Principal Scheme and Procedure of Data Analysis

Getting new knowledge with the help of database KDD&DM (Knowledge Discovery in Databases and Data Mining) is an extensively developed direction of research. In this frame, relational methods of data extraction from the databases (Data Mining) and software "Discovery" using the first-order logic language were developed [11, 12, 14]. This system was used to solve a wide range of problems, for example to construct diagnostic system of oncological diseases, analysis of psychophysical experiments, to predict time series, and many others [12, 15] (see also the Internet site "Scientific Discovery": http://www.math.nsc.ru/LBRT/logic/vityaev/, section "comparison"). Principles of the system "Discovery" allows its application in molecular biology, a rather hardly formalizable area.

0026-8933/01/3506-0810$25.00

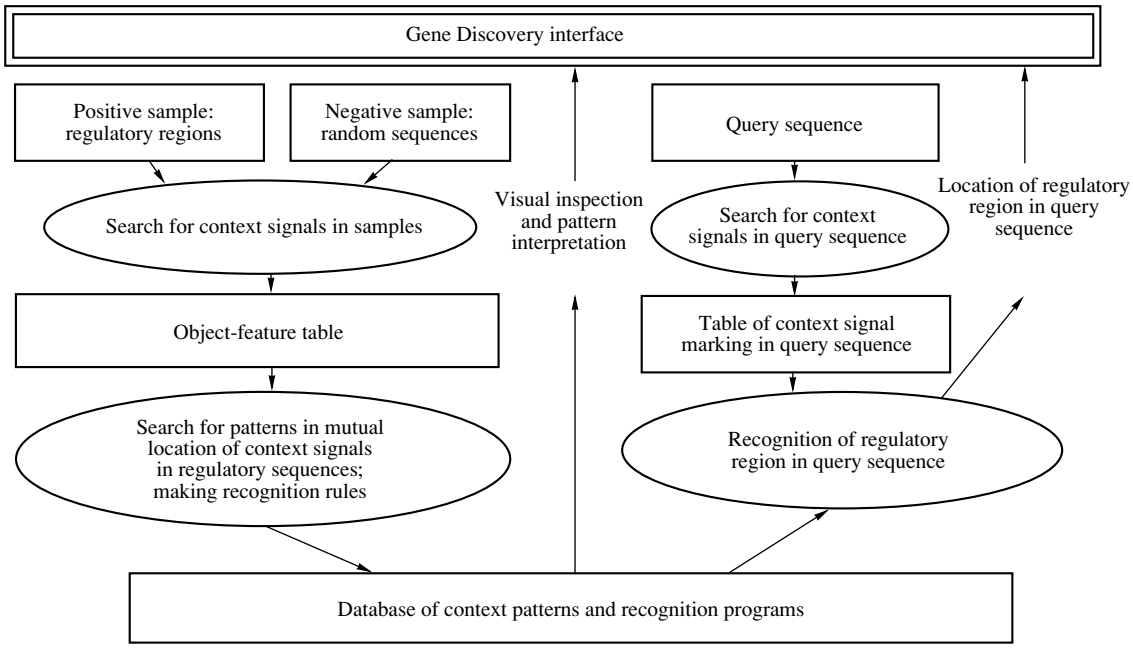

Fig. 1. Block scheme of the system "Gene Discovery." The data are shown as rectangles, the stages of information processing are shown with ovoids.

Computer-assisted system "Gene Discovery" is an adaptation of the system "Discovery" to analyze nucleotide sequences of the regulatory regions. The principal scheme of the system is shown in Fig. 1. The system is entered with a learning sample of nucleotide sequences from two alternative classes: class 1, promoters and class 2, the sequences lacking promoter functions (e.g., random sequences with the same nucleotide frequencies, sequences of exons, introns, etc.)

A block of programs searches for context signals in sequences of these two classes. A signal may be:

• context (a short oligonucleotide "word," a functional site, etc.),

• conformational (a DNA site showing certain conformational or physicochemical features, e.g., low-melting DNA fragments, highly bent DNA, etc.),

• structural (e.g., Z-DNA, a hairpin of the secondary RNA structure, etc.).

Only one type of the context signals is considered in this work, imperfect nucleotides.

Isolation of Individual Context Signals

Oligonucleotide signals specific for a given group of promoters were isolated using software ARGO [16], (see also Vishnevsky and Vityaev, this issue, and http://wwwmgs.bionet.nsc.ru/mgs/programs/argo/). We consider an oligonucleotide signal or motif a word of eight bases written in generalized 15-letter IUPAC alphabet: {A, T, G, C, R = G/A, Y = T/C, M = A/C, K = T/G, W = A/T, S = G/C, B = T/C/G, V = A/G/C, H = A/T/C, D = A/T/G, N = A/T/G/C}.

Program ARGO analyzes two contrasting set of signals, one of which has a specific function (a positive sample, the promoter set in our case) and the other lacks this function (a negative sample, random sequences in our case). The analysis of these two sets using software ARGO reveals oligonucleotide motifs according to the following criteria: (1) the motif has high frequency in the sequences of positive sample and low frequency in the sequences of negative sample; (2) this difference in frequency between the two samples is statistically significant.

This program was used to analyze promoter sequences of the endocrine-system genes compared with random sequences of the same nucleotide composition. The sample of 40 promoters was taken from ES-TRRD (http://wwwmgs.bionet.nsc.ru/mgs/papers/ignatieva/es-trrd/index.html). The length of promoter sequences was 120 bp, –100 to +20 of transcription start. As shown by homology analysis, no pair of promoters had more than 60% similarity. A negative "non-promoter" sample contained 1000 random sequences of the same length and with the same nucleotide composition as the sample of promoter sequences. Random sequences were generated using the program from http://wwwmgs.bionet.nsc.ru/mgs/dbases/nsamples/.

Analysis of endocrine-system gene promoters with software ARGO revealed 68 specific oligonucleotides in 15-letter code (below they are called motifs). Some examples are shown in Table 1 with information on

Table 1. Examples of oligonucleotide motifs specific for promoters of the endocrine-system genes (in the region between –100 and +20 of transcription start)

No.	Sequence in 15-letter code	Portion of the sequences from the learning set containing the motif*	
		Promoters	Non-promoters
1	KNCMAGDG	0.325000	0.031000
2	KRCCWGNR	0.350000	0.044000
3	ANANANCA	0.275000	0.032000
4	GKNCAGRG	0.225000	0.010000
5	CANAGCMN	0.250000	0.013000
6	RGSNRGRG	0.400000	0.041000
7	KGRSSAGR	0.275000	0.017000
8	KGRSCNGR	0.375000	0.043000
9	YRGRGNCA	0.250000	0.028000
10	CWGWGNCN	0.300000	0.042000
…	…	…	…
68	CTGNNCAN	0.250000	0.035000

* The difference for promoters and random sequences is statistically significant assuming binomial distribution of nucleotides. Missing data are marked with dots.

Table 2. Examples of oligonucleotide motifs specific for promoters of the erythroid-specific genes (in the region between –100 and +20 of transcription start)

No.	Sequence in 15-letter code	Portion of the sequences from the learning set containing the motif	
		Promoters	Non-promoters
1	TGACCAAT	0.341463	0.000000
2	RCCAATND	0.585366	0.022000
3	RRYCAATR	0.512195	0.009000
4	YTGRYCAA	0.390244	0.008000
5	CCAATRDN	0.487805	0.017000
6	CMMTGRSH	0.585366	0.035000
7	MYTKRCCW	0.439024	0.022000
8	CMHTSACY	0.390244	0.012000
9	WWTSAYTT	0.243902	0.018000
10	TNNYYTCA	0.390244	0.038000
…	…	…	…
160	NYRCWTYT	0.317073	0.035000

specific oligonucleotide motifs in 15-letter code (column 2), and on the parts of the positive and negative sets containing the motif at least once (columns 3 and 4). For example, motif KNCMAGDG is found in 32% sequences of the promoter sample and only in 3% sequences of the random sample.

Since the motifs are described using the 15-letter code, they are variable, i.e., almost every position of a motif may be occupied by either of more than one nucleotide. For example, motif KNCMAGDG (no. 1 in Table 1) written in 15-letter code really is like (T/G)(A/T/G/C)C(A/C)AG(A/T/G)G. Each of the shown motifs is found in at least once per 20–40% promoters. At the same time, the frequency of the motif in promoters is significantly higher than in random sequences (Table 1). However, we should note that despite significantly higher frequency in promoters none of the revealed motifs was present in all promoter sequences. Therefore, we consider the revealed oligonucleotide motifs as quasi-invariant signals, short conserved fragments common for subgroups of the studied set of promoter sequences.

A similar procedure was used to analyze the sample containing 41 promoter sequence of the erythroid-specific genes taken from EpoDB [13] (http://www.cbil.upenn.edu/EpoDB/release/version_2.2/epodb.html) (Table 2).

To summarize, program ARGO allowed us to find short oligonucleotide motifs common for different groups of specific promoters. The next stage of our study was to search for localization of the revealed motifs within all sequences of a specific promoter group. The results of this search for endocrine-system gene promoters are presented in object-feature form in Table 3. Each row of this table corresponds to a single sequence and includes description of an oligonucleotide signal found within this sequence. If the sequence contains more than one oligonucleotide signals, it is represented with more than one row. For example, the first row in Table 3 corresponds to sequence no. 1 of the positive sample (promoter of the human adrenodoxin gene, EMBL ID M23665), which contains the context signal 6: RGSNRGRG. This signal is located at position –65 of transcription start point in direct chain with orientation from the 5' end to the 3' end. The second row of the Table 3 shows that the same signal is located at position –60 overlapping with the copy of position –65. The last column of Table 3 indicates that this sequence is of promoter class.

Search of the Promoters for Complex Context Signals

The promoters of each of the two studied types have common groups of quasi-invariant oligonucleotide motifs, and frequency of these motifs in promoter sequences is significantly higher than in random sequences. It appeared interesting to study the problem of similar occurrence and relative location of these motifs within promoter sequences. Below we use the name “complex signal” for a group of quasi-

invariant oligonucleotide motifs with certain features of relative location within promoter sequences.

If we take a simple complex system (S_1, S_2) formed by a pair of oligonucleotides, it may be presented as follows:

$$(S_1, S_2) = (\mathrm{Pos}(S_1) < \mathrm{Pos}(S_2) \\ \& (\mathrm{Sign}(S_1) = z_1) \& (\mathrm{Sign}(S_2) = z_2), \quad (1)$$

where S_1 and S_2 are oligonucleotides in the object-feature table; $\mathrm{Pos}(S_1)$ and $\mathrm{Pos}(S_2)$ are positions of these oligonucleotides within a sequence, $\mathrm{Sign}(S_1)$ and $\mathrm{Sign}(S_2)$ are oligonucleotide signs showing that the oligonucleotide is located in either direct (+) or complementary (–) chain: $z_1, z_2 \in \{+, -\}$.

According to (1), complex signal (S_1, S_2) is formed by two signals S_1 and S_2, signal S_1 located upstream of signal S_2 ($\mathrm{Pos}(S_1) < \mathrm{Pos}(S_2)$) and each of the two signals shows certain orientation ($\mathrm{Sign}(S_1) = z_1$) & ($\mathrm{Sign}(S_2) = z_2$).

The hypothesis about relation of the complex signal (S_1, S_2) with the promoter class is written as the following logical expression:

$$\forall R \exists S_1, S_2((S_1, S_2) \Rightarrow (\mathrm{Class}(R) = 1)), \quad (2)$$

where R is a nucleotide sequence; $\mathrm{Class}(R)$ is the number of the class containing this sequence (1, promoters or 2, random sequences).

Hypothesis (2) states that any sequence R belongs to promoter class (class 1) if it contains signals S_1 and S_2, of location $\mathrm{Pos}(S_1) < \mathrm{Pos}(S_2)$ and oriented similarly with z_1 and z_2.

The block for trend search "Gene Discovery" tries within the object-feature table all possible variations of hypothesis (2) for complex signals (S_i, S_j), where $i, j = 1, \ldots N$, N is a number of individual signals (oligonucleotide motifs). Fisher's test was used to estimate significance of signal relation to certain object class (class 1, promoter or class 2, non-promoter) for each variation. Significance level by Fisher's test may be written as $P(N_1, N_2, N_3, N_4)$:

	Class 1	Class 2
Condition (1) fit	N_1	N_2
Condition (1) not fit	N_3	N_4
Number of object in class	$N_1 + N_3$	$N_2 + N_4$

where N_1, N_2 is number of promoters (class 1) and non-promoters (class 2), respectively, containing signals S_1, S_2, which fit condition (1); N_3, N_4, number of promoters and non-promoters, respectively, not fitting condition (1).

Beside Fisher's test, we estimated conditional probability $PC(N_1, N_2) = N_1/(N_1 + N_2)$ of sequence attribution to promoter class if this sequence contains complex signal (S_1, S_2).

Table 3. Object-feature table describing oligonucleotide signals revealed in promoters of endocrine-system genes

No. of the sequence in the set*	No. of the context signal	Sequence of the context signal	Position of the site start relative to the transcription start point	Orientation**	Class***
1	6	RGSNRGRG	–65	1	1
1	6	RGSNRGRG	–60	1	1
1	8	KGRSCNGR	–100	–1	1
1	66	YTSCWGNW	+13	–1	1
1	67	TCMAGNMN	+13	1	1
…	…	…	…	…	…
40	67	TCMAGNMN	–65	–1	1
40	68	CTGNNCAN	–79	1	1
40	68	CTGNNCAN	–61	1	1
41	1	KNCMAGDG	–51	1	2
41	10	CWGWGNCN	–13	–1	2
…	…	…	…	…	…
1040	56	HNNKGCTG	–64	1	2
1040	56	HNNKGCTG	–12	1	2
1040	64	NCWGGGNC	–8	1	2

* 1–40, promoters; 41–1040, random sequences of the same nucleotide composition. Only part of the whole data (16,300 rows) is shown. Missing rows are marked with dots.

** 1, direct chain; –1, complementary chain.

*** 1, promoters; 2, non-promoters.

Running the analysis of the object-feature table, the system starts from the simplest complex signals (S_1, S_2) and performs directed step-by-step search of more complex signals, gradually increasing the complexity by adding new individual signals. Therefore, in general case we consider a complex signal as (S_1, S_2, …, S_m) at $m > 1$, fitting the condition:

$$\begin{aligned}&((\mathrm{Pos}(S_1) < \mathrm{Pos}(S_2)) \,\&\, (\mathrm{Pos}(S_2)\\ &< \mathrm{Pos}(S_3)) \,\&\, \ldots \,\&\, (\mathrm{Pos}(S_{m-1}) < \mathrm{Pos}(S_m))\\ &\quad \&\, (\mathrm{Sign}(S_1) = z_1) \,\&\, (\mathrm{Sign}(S_2)\\ &\quad = z_2) \,\&\, \ldots \,\&\, (\mathrm{Sign}(S_m) = z_m).\end{aligned} \tag{3}$$

The tested hypothesis about relation of the signal (S_1, S_2, …, S_m) with promoter class written using formal logic language looks as follows:

$$\begin{aligned}\forall R \exists S_1, S_2, \ldots, S_m((S_1, S_2, \ldots, S_m)\\ \Rightarrow \mathrm{Class}(R) = 1),\end{aligned} \tag{4}$$

where R is a nucleotide sequence; Class(R) is number of the class containing this sequence (1, promoters; 2, random sequences); m = 1, 2, …, number of signals (oligonucleotides) considered within the tested hypothesis.

New individual signals are added to the complex signal when conditional probability $PC(N_1, N_2)$ of sequence attribution to promoter class strictly increases and Fisher's test $P(N_1, N_2, N_3, N_4)$ shows significance at < 0.05.

This search results if formation of complete group of the complex signals $Q = \{(S_1^1, \ldots, S_{m1}^1), \ldots, (S_1^i, \ldots, S_{mi}^i), \ldots, (S_1^n, \ldots, S_{mn}^n)\}$ and relative promoter-attributed patterns characterizing interaction of a complex signal with sequence class. Therefore, promoter j is analyzed for subset of complex signals $Q_j \subset Q$, which belongs to it.

RESULTS AND DISCUSSION

We used "Gene Discovery" to analyze tables of data related to promoters testing hypotheses (2)–(4). The program applied to promoters of erythroid-specific genes and endocrine-system genes detected numerous complex signals, the number of which (from dozens to dozen thousand) varied depending on critical value selected for the Fisher's test.

At the last step we selected complex signals to fit the following additional conditions: (1) the individual signals forming a complex signals do not overlap in the sequences of studied promoters; (2) the recorded number of complex-signal-containing promoters N is higher than N^* expected for random distribution: $N > N^*$.

The expected number N^* was estimated multiplying the frequencies of single nucleotides in promoters by total number of promoters, considering number of variations in mutual oligonucleotide location within promoter sequence. For example, expected number of promoters N^* with complex signal (S_1, S_2, S_3|Pos(S_1) < Pos(S_2) < Pos(S_3)), is calculated as:

$$N^* = P(S_1)P(S_2)P(S_3)M/6,$$

where N^* is the expected number of promoters; $P(S_1)$, $P(S_2)$, $P(S_3)$ frequencies of promoters containing oligonucleotides S_1, S_2, and S_3, respectively; M, total number of promoters in the analyzed set; 6 = 3! The number of possible variations of relative location of three oligonucleotides within a promoter.

Examples of the complex signals fitting these conditions and specific for endocrine-system gene promoters and for erythroid-specific promoters are shown in Tables 4 and 5. For example, signal CWGNRGCN < NGSYMTAM < MAGKSHCN in endocrine-system gene promoters has $N^* = 0.47$, i.e. less than 1, while this signal is found in 6 promoters, being about 13 times more frequent than in case of random distribution (see Table 4). Signal DNMYTTSA < DNYAADGG < RCAGMMDY in erythroid-specific gene promoters has $N^* = 0.54$, and was found in 8 promoters showing frequency about 14 times higher that expected for random distribution (see Table 5).

It appeared interesting to study the patterns of complex signal location within the promoters. As an example, Fig. 2 shown location of the complex signal CWGNRGCN < NGSYMTAM < MAGKSHCN (Table 4) in promoters of the endocrine-system genes. This signals was detected in six promoters in the region from –100 to +20 bp counting from the transcription start point: 9, 12, 22, 25, 32, 37 (ID EMBL: M26856, M73820, U02293, J00749, J03071, K01877, respectively). This complex signal is located in the region from –95 to –7 bp counting from the transcription start point: (we mark position of the first nucleotide of the motif). Position of the TATA box indicated in TRRD is marked with dashed rectangles One may note that the second oligonucleotide motif coincides with the TATA-box region. Moreover, Fig. 2 shows that in promoters 12, 25, and 37 the distance between second and third oligonucleotides is 42–51 and 12–26 bp, respectively. These oligonucleotides are also closely located in promoters 22 and 32 (63–65 and 9 bp). In 9, these distances are 12 and 17 bp, respectively.

Figure 3 shows an example of complex signal DNMYTTSA < DNYAADGG < RCAGMMDY location in 8 sequences of erythroid-specific gene promoters (numbers in the set 8–11, 14, 16, 17, 25, and 39). In this case one may also see common distances between individual signals. In promoters 8 and 10 the distance between first and second, second and third oligonucleotides of the complex signal is 16 and

Table 4. Examples of complex signals in promoters of endocrine-system genes

No.	Complex signal[1]	Conditional probability[2]	Fisher's test[3]	Number of promoters containing the signal[4]	Number of promoters expected from random distribution[5]
1	CWGNRGCN<NGSYMTAM<CAGGRNCH	0.875	0.00054	4	0.24 (<1)
2	KGRSSAGR<CYCYNSCY<CWGSNYCH	1.0	0.00012	4	0.28 (<1)
3	CWGNRGCN<NGSYMTAM<MAGKSHCN	1.0	0.00009	6	0.47 (<1)
4	CWGNRGCN<NGSYMTAM<CMDGGNCH	0.846	0.00099	5	0.43 (<1)
5	CNKSAGNT<NCARGRNC<HNNKGCTG	1.0	0.01426	4	0.37 (<1)
6	RNWGGCCN<DGRGNRGG<TCMAGNMN	0.875	0.00118	4	0.40 (<1)
7	RGSNRGRG<NNGSTWTA<CNCNRKGC	1.0	0.02852	5	0.53 (<1)
8	NNGSTWTA<NMAGDGMC<CNCNRKGC	0.875	0.04755	5	0.53 (<1)
9	RGSNRGRG<NNGSTWTA<CMDGGNCH	1.0	0.03964	5	0.55 (<1)
10	RGSNRGRG<KGGNSAGD<ANCTSMNG	1.0	0.03964	4	0.45 (<1)
…	…	…	…	…	…
45	RGSNRGRG<NGSYMTAM<CNCNRKGC	1.0	0.03964	5	0.58 (<1)

Note: Only part of the whole dataset of large volume is shown. Missing rows are marked with dots.

[1] Complex signal is formed by oligonucleotides in 15-letter code linearly positioned within the sequence as written. "<" means that position number of the first oligonucleotide relative to the transcription start point is lower than that of the second. The distance between individual oligos is not fixed.

[2] Conditional probability PC(N_1, N_2) is calculated as a ratio of number of promoters containing the signals N_1 and total number of sequences containing the signal $N_1/(N_1 + N_2)$.

[3] Fisher's test is used to estimate probability to randomly obtain the signal in more promoters than observed, it is calculated for tables $P(N_1, N_2, N_3, N_4)$.

[4] Number of promoters in the learning set containing the signal.

[5] Number of promoters containing the signal expected from random distribution. It is calculated assuming independence of oligonucleotides as total number of promoters multiplied by oligonucleotide frequency in promoters considering variations in their relative positioning.

Table 5. Examples of complex signals in promoters of erythroid-specific genes*

No.	Complex signal	Conditional probability	Fisher's test	Number of promoters containing the signal	Number of promoters expected from random distribution
1	DNMYTTSA<DNYAADGG<RCAGMMDY	1.0	0.00201	8	0.54
2	DNMYTTSA<DNYAADGG<WKCWSANA	1.0	0.00201	8	0.58
3	ATNDNYTC<RAANNMAW<BYNNCACA	1.0	0.00020	9	0.66
4	DNMYTTSA<RAANNMAW<BYNNCACA	1.0	0.00099	10	0.74
5	ATNDNYTC<WGNRNCWG<NTGYWTNT	1.0	0.01426	11	0.82
6	ANYYTTGN<ATNDNYTC<CWGNYNCW	1.0	0.00118	8	0.6
7	ATNDNYTC<DSDGVWSA<TGANRCWK	1.0	0.02852	10	0.76
8	ATNDNYTC<NVDGNATA<NABHTGCT	1.0	0.04755	10	0.76
9	DNMYTTSA<DNYAADGG<RAANNMAW	1.0	0.03964	8	0.61
10	ATNDNYTC<GDSCCWGN<BYNNCACA	1.0	0.03964	9	0.69
…	…	…	…	…	…
357	TNNYYTCA<RAANNMAW<BYNNCACA	1.0	0.03964	9	0.7

* Structure as in Table 4.

25–26 bp, respectively. In promoters 17, 14, 25 and 39 these distances are 21–33 and 50 bp, respectively (see Fig. 3)

In conclusion, computer-assisted system "Gene Discovery" developed by us allows detection of individually essential motifs (varying quasi-invariant oli-

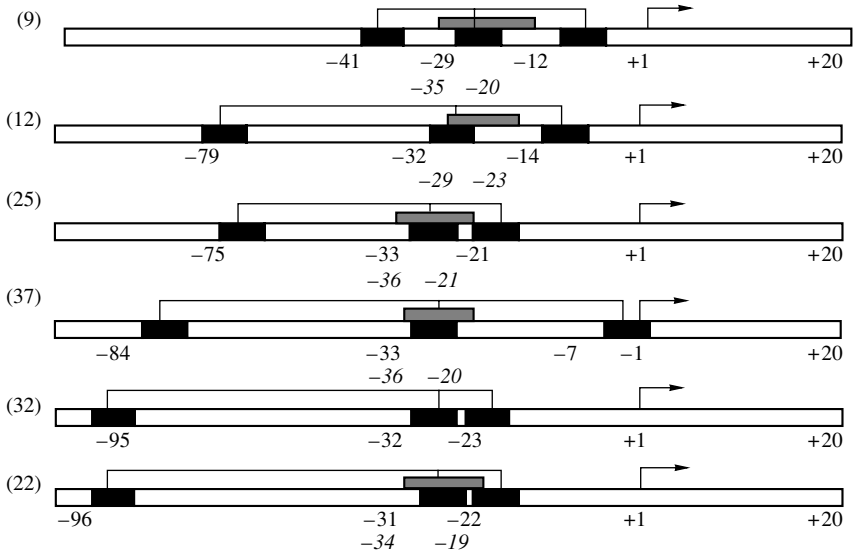

Fig. 2. Location of the complex signal CWGNRGCN < NGSYMTAM < MAGKSHCN in promoters of the endocrine-system genes. Promoter sequences are phased relative to the transcription start (position +1), shown with arrow. The number of promoter is shown on the left in parentheses. The 8-bp oligonucleotide motifs forming the complex signal are shown as black rectangles, position of the first nucleotide relative to the transcription start point is marked. Gray rectangles show position of the TATA box indexed in TRRD. Positions of the first and the last nucleotide of the TATA box are shown in italics.

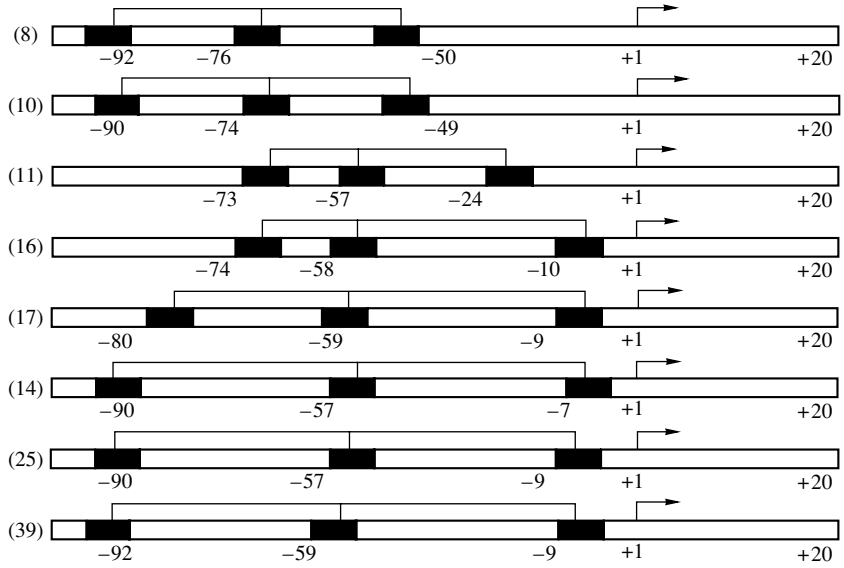

Fig. 3. Location of the complex signal DNMYTTSA < DNYAADGG < RCAGMMDY in eight sequences of erythroid-specific gene promoters. Designations as in Fig. 2.

gonucleotides) and of complex signals. Our analysis has shown that high content of these signals is common for promoters of endocrine-system genes and of erythroid-specific genes. Similar location of these signals within subgroups of specific promoters suggests their functional importance (see Figs. 2 and 3). Moreover, as mentioned above, complex signals may have similar distance between individual motifs. At the same time, analyzed promoter sequences no strong homology.

Individual motifs may correspond with transcription factor binding sites. In early works on promoter recognition and analysis it was already shown that, compared with random sequences, they are enriched with potential sites of transcription factor binding [17]. Individual motifs may also correspond with the

DNA sites responsible for specific conformational of physical and chemical properties: increased DNA curving, low-melting, etc. essential for promoter functioning.

Some interesting points concerning complex signals should be mentioned. First, some works [10, 18] have shown specific patterns of potential transcription factor-binding sites: different sites were mainly located in different promoter regions. Therefore, the observed complex signals may reflect location of different sites in certain promoter regions. This consideration [10] allowed better recognition of promoters. Importance of positioning for studying context features was also proved by Zang [3]. Levitsky and coworkers [19] (see also Levitsky and Katokhin, this issue) revealed promoter separation in local regions of common dinucleotide composition. These regions were shown to possess certain conformational or physicochemical properties. We suggest that complex signals may be either of context or context-conformation nature in relation with context specificity in certain promoter regions or with local patterns of DNA conformation essential for specific functions of the gene promoters.

Second, recently attention was drawn to analysis of a special type of regulatory elements controlling transcription, CE [20]. They are formed by paired transcription factor binding sites (overlapping, adjacent or located at a fixed distance) which show new regulatory properties because of protein-protein interactions between the respective transcription factors. Each of the sites within a CE is able to function independently, but their interaction provides considerably stronger activating or repressing effect on gene transcription. To date, more than 150 CE have been revealed experimentally [20] (see also http://compel.bionet.nsc.ru/). The study of the patterns of joint occurrence and relative location of the sites using the system "Gene Discovery" opens a way to create computer-assisted procedures to search for potential CE. We suppose that detection and analysis of complex signals should allow considerable increase in efficiency of specific promoter group recognition in future.

Recognition of promoters basing on the signals detected using our system "Gene Discovery" will be described elsewhere.

ACKNOWLEDGMENTS

The authors are grateful to V.G. Levitsky, E.V. Ignatieva, M.A. Pozdnyakov, and O.A. Podkolodnaya for their help and valuable comments. This work was supported by the Russian Foundation for Basic Research (projects nos. 00-04-49229, 00-07-90337, 99-07-90203, 01-07-90376, 00-04-49255, 01-04-06243mac), program Human Genome, Siberian Branch of the Russian Academy of Sciences), and INTAS (grant to Yu. Orlov YSF 00-178).

REFERENCES

1. Kolchanov, N.A., Podkolodnaya, O.A., Ananko, E.A., Ignatieva, E.V., Stepanenko, I.L., Kel-Margoulis, O.V., Kel, A.E., Merkulova, T.I., Goryachkovskaya, T.N., Busygina, T.V., Kolpakov, F.A., Podkolodny, N.L., Naumochkin, A.N., Korostishevskaya, I.M., Romashchenko, A.G., and Overton, G.C., *Nucl. Acids Res.*, 2000, vol. 28, pp. 298–301.
2. Singer, M. and Berg, P., *Genes and Genomes*, Moscow: Mir, 1998.
3. Zhang, M.Q., *Genome Res.*, 1998, vol. 8, pp. 319–326.
4. Arnone, M.I. and Davidson, E.H., *Development*, 1997, vol. 124, pp. 1851–1864.
5. Nikolov, D.B. and Burley, S.K., *Proc. Natl. Acad. Sci. USA*, 1997, vol. 94, pp. 15–22.
6. Goodrich, J.A., Cutler, G., and Tjian, R., *Cell*, 1996, vol. 84, pp. 825–830.
7. Fickett, J.W. and Hatzigeorgiou, A.G., *Genome. Res.*, 1997, vol. 7, pp. 861–878.
8. Pedersen, A.G., Baldi, P., Chauvin, Y., and Brunak, S., *Comput. Chem.*, 1999, vol. 23, pp. 191–207.
9. Solovyev, V. and Salamov, A., in *Proc. Fifth Int. Conf. on Intelligent Systems for Molecular Biology (ISMB-97)*, 1997, pp. 294–302.
10. Kondrakhin, Y.V, Kel, A.E., Kolchanov, N.A., Romaschenko, A.G., and Milanesi, L., *Comput. Appl. Biosci.*, 1995, vol. 11, pp. 477–488.
11. Vityaev, E.E. and Moskvitin, A.A., in *Vychislit. Sistemy*, 1993, issue 148, pp. 117–163.
12. Kovalerchuk, B. and Vityaev, E., *Data Mining in Finance: Advances in Relational and Hybrid Methods*, Kluwer Academic Publishers, 2000.
13. Stoeckert, C.J., Jr., Salas, F., Brunk, B., and Overton, G.C., *Nucleic Acids Res.*, 1999, vol. 27, pp. 200–203.
14. Krantz, D.H., Luce, R.D., Suppes, P., and Tversky, A., *Foundations of Measurement*, New York, London: Academic Press, vols. 1, 2, 3, 1971, 1989/1990.
15. Kovalerchuk, B., Vityaev, E., and Ruiz, J., *IEEE Engineering in Medicine and Biology Magazine. Special issue: Medical Data Mining*, 2000, pp. 26–37.
16. Babenko,V.N., Kosarev, P.S., Vishnevsky, O.V., Levitsky, V.G., Basin, V.V., and Frolov, A.S., *Bioinformatics*, 1999, vol. 15, pp. 644–653.
17. Prestridge, D.S., *CABIOS*, 1991, vol. 7, pp. 203–206.
18. Klingenhoff, A., Frech, K., Quandt, K., and Werner, T., *Bioinformatics*, 1999, vol. 15, pp. 180–186.
19. Levitsky, V.G., Katokhin, A.V., and Kolchanov, N.A., *Vichislitelnie Technologii*, 2000, vol. 5, pp. 41–47.
20. Kel-Margoulis, O.V., Romaschenko, A.G., Kolchanov, N.A., Wingender, E., and Kel, A.E., *Nucl. Acids Res.*, 2000, vol. 28, pp. 311–315.