Knowledge Discovery for Promoter Structure Analysis

Vityaev E.E. Sobolev Institute of Mathematics, Novosibirsk, 630090, Russia, vityaev@math.nsc.ru
Orlov Yu.L., Pozdnyakov M.A., Vishnevsky O.V., Kolchanov N.A. Institute of Cytology and Genetics SB RAS, Lavrentieva ave., 10, Novosibirsk, 630090, Russia, {orlov,mike,oleg,kol}@bionet.nsc.ru
Kovalerchuk B.K., Department CS, Central Washington University, Ellensburg, WA 98926-7520, borisk@cwu.edu

Abstract—This paper presents implementation of Data Mining and Knowledge Discovery techniques for searching for regularities in tables of context features of DNA sequences involved in regulation of transcription. The goal is to discover regularities that relate nucleotide sequences to the functional classes of these sequences. The search patterns for regularities have been constructed in the first-order logic augmented by probabilistic estimates. To this aim, the PC software system "Gene Discovery" has been designed. This system accepts molecular-genetical data retrieved from a database by using SQL queries. Nucleotide sequences of promoters of several functional systems were extracted from the TRRD database and analyzed. The data include nucleotide sequences of erythroid-specific gene promoters, endocrine system gene promoters, promoter regions of the genes controlling cell cycle, promoter of genes regulating lipid metabolism, and muscle-specific gene promoters. Several regularities that relate the nucleotide sequences in the regulatory DNA with each functional class have been found. In addition, sets of transcription factors binding sites and donor splice sites were analyzed. Recognition procedure based on regularities was developed.

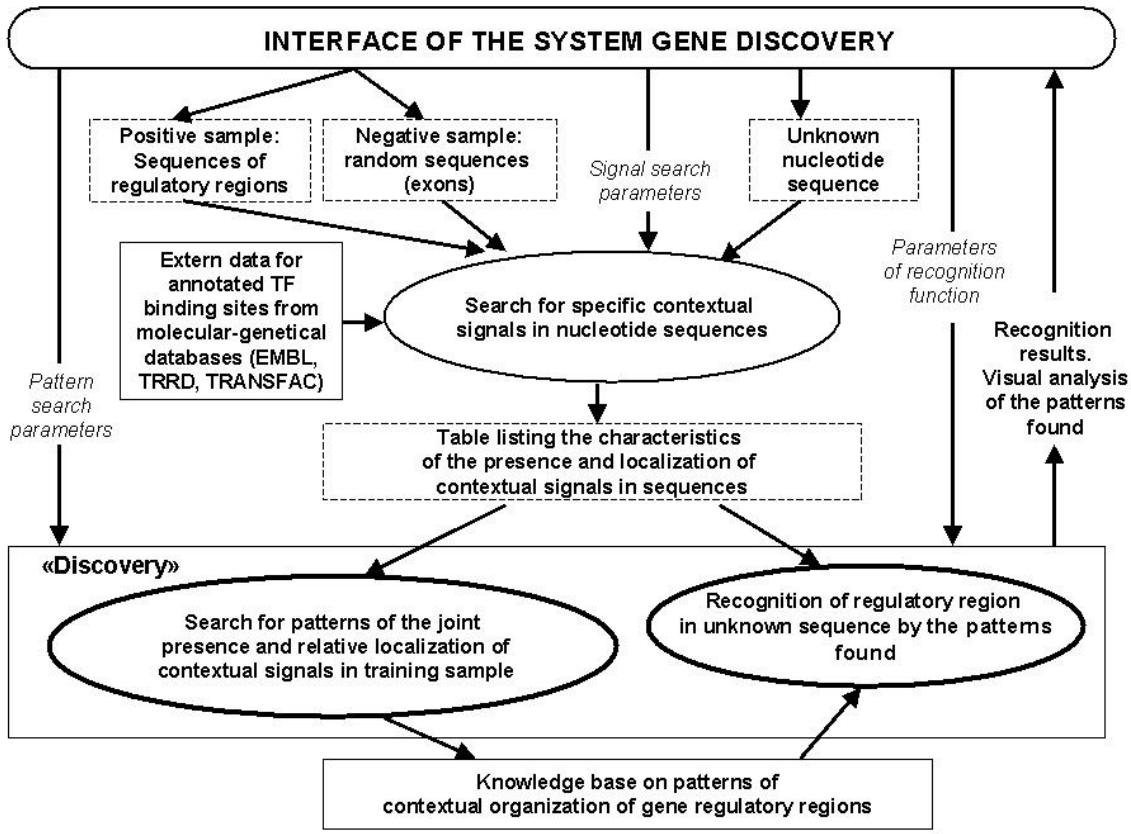
Keywords: — Machine Learning, Knowledge Discovery, Data Mining, bioinformatics, eukaryotic promoter recognition, transcription factors binding sites.

1. Introduction

Analysis of promoter structure is of great interest for understanding molecular mechanisms of gene transcription. The presence and location of transcription factor binding sites in 5' regulatory regions of genes correspond to the tissue- and stage-specific features of gene expression in an organism. The control of eukaryotic gene expression is primarily determined by relatively short sequences (signal/motif) in the region surrounding a gene. These sequences vary in length, position, redundancy, orientation in DNA chain, and bases. Eukaryotic promoters are characterized by the absence of exact localization of context signals and the weakness of such signals [5]. Diversity of promoters is the main difficulty for developing of recognition programs [3]. During the last years, such techniques as the large-scale data mining, knowledge discovery, and other computational approaches of Machine Learning were intensively used in bioinformatics [5,12,15]. Recently several computational approaches have been suggested to address challenges of combinatorial regulation of transcription [7,9]. In particular, they concern computer selection of specific oligonucleotides [17] and mining associations between them [4].

Our approach based on Data Mining methods for specific oligonucleotide pattern selection and functional class of a gene description [30]. The program developed on the basis of the training sample of nucleotide sequences of promoter region.[1] It is hard to describe all eukaryotic promoter sequences by common patterns due to a huge variability of different transcription factors binding sites. To overcome this difficulty, the sets of genes promoters, which performing the similar function were extracted from the TRRD database [10, 24]. However, even such functional sets lack a single oligonucleotide pattern describing all sequences. Distinctive feature of the algorithm is the usage of specific feature patterns describing a subgroup of the training set. The search patterns for regularities are constructed in the first-order logic augmented by probabilistic estimates.

[1] The demo-version of the program is available upon the request addressed to the authors (vityaev@bionet.nsc.ru).



2. Systems and methods

2.1."Gene Discovery": Technology of Data Mining

The system "Gene Discovery" Fig. 1 is an adaptation of the system "Discovery" [12-14,19] to the tasks of molecular biology. "Gene Discovery" consists of three main modules: (1) the module for on-line representation of the context signals from DNA sequence in a standard table form; (2) the module "Discovery" for regularities searching; (3) the module for recognition of the sequence class by using the regularities found. The program is written in the C++ and it is supplied by a user-friendly interface.

System "Discovery" [19], retrieve statistically significant first-order logic rules for functional annotation of regulatory regions. This system base on first-order representations of data and hypotheses and have been successfully applied for many problems in psychology, physics, medicine, finance, and other fields [12-14,19] (see also www-site "Scientific Discovery" [25]).

2.2.Oligonucleotide Patterns as Forecasting rules

As with any technique based on logic rules [16], this technique allows one to obtain human-readable forecasting rules that are interpretable in biological language and also provides promoter recognition (functional annotation). An expert in biology may evaluate both the correctness of the recognition and that of the rules themselves.

An example of oligonucleotide motif in 15-lettered alphabet is CWGNRGCN.

Let us consider as example the forecasting rule:

If WGNRGCN<NGSYMTAM<MAGKSHCN
Then: Sequence class = promoter.

The symbol "<" here designates that positions of corresponding oligonucleotides are ordered relative to the transcription start.

This rule means: if motifs CWGNRGCN and NGSYMTAM and MAGKSHCN present in sequence under analysis, and their non-overlapping mutual location is fixed, then the sequence under

analysis contains promoter of the gene of an endocrine system.

In such a way, all the statistically significant oligonucleotide patterns are constructed in the form $S_1\&S_2\&\ldots\&S_k$, where k>1. The program automatically defines the number of the signals in such a pattern [30].

3. ALGORITHM

3.1. Logical Probabilistic Rules

The critical issue in applying data-driven forecasting systems is generalization. "Discovery" software systems generalize data through "law-like" logical probabilistic rules. The "Law-like" rule definition satisfies all properties of scientific laws**:** (1) high level of generalization; (2) simplicity (Occam's razor); and, (3) refutability.

Formally, an IF-THEN "Law-like" rule is $C = (A_1\&\ldots\&A_k \Rightarrow A_0)$, where the IF-part, $A_1\& \ldots \&A_k$, consists of true/false logical statements A_1, …, A_k, and the THEN-part consists of a single logical statement A_0. Statements A_i may have negations. Rule C allows one to generate sub-rules $\text{SubFormular}(A_1\&\ldots\&A_k) \Rightarrow A_0$ with a truncated IF-part, e.g. $A_1\&A_2 \Rightarrow A_0$; $A_1\&A_2\&A_3 \Rightarrow A_0$ and so on. It is known that the sub-rule is logically stronger than the rule. Thus, if some rule C and its sub-rule C' classify correctly the same set of examples, then the sub-rule is preferred. In general, there are three reasons to prefer the sub-rule:

1) The sub-rule is more general (logically stronger and describes the same set of events);

2) The sub-rule is simpler than the rule, it consists of fewer statements in the IF-part;

3) Sub-rule is better testable (more refutable) than the rule, because the larger set of possible examples may falsify it (the IF-part of the sub-rule is less restrictive).

Thus, if a rule C covers the set of examples, then one should test that none of its sub-rules C' also covers the same set of examples. Otherwise, this sub-rule will be preferred. So, "law-like" rule can be defined as a rule without sub-rules covering the same set of examples. In other words, "law-like" true for some set of examples, but none of its sub-rules is true on it.

If examples contain noise, which is typical for life sciences, the probabilistic characteristics of the expressions are used instead of crisp (true/false) values. The conditional probability $P(C) = P(A_0|A_1\& \ldots\&A_k)$, assuming that $P(A_1\& \ldots\&A_k) > 0$ of the rule C is used in the "Discovery" system as such characteristic. Similarly, conditional probabilities $P(A_0|A_{i1}\&...\&A_{ih})$ are defined for sub-rules $C_i = (A_{i1}\&...\&A_{ih} \Rightarrow A_0)$, assuming that $P(A_{i1}\&...\&A_{ih}) > 0$.

The rule is a probabilistic "law-like" rule iff all of its sub-rules have a statistically significant lower conditional probability than the rule. Another definition of "law-like" rules can be given in terms of generalization. The rule is "law-like" iff it can't be generalized without producing a statistically significant reduction in its conditional probability.

The "Discovery" system searches all chains $C_1, C_2, \ldots, C_{m-1}, C_m$ of nested "law-like" subrules, where C_1 is a subrule of rule C_2 , $C_1 = \text{sub}(C_2)$, C_2 is a subrule of rule C_3, $C_2 = \text{sub}(C_3)$ and finally C_{m-1} is a subrule of rule C_m, $C_{m-1} = \text{sub}(C_m)$. Also $P(C_1) < P(C_2), \ldots, P(C_{m-1}) < P(C_m)$. The algorithm stops generating new rules when they become too complex (i.e., statistically insignificant on data, even if the rules are highly accurate).

There is theorem [21] that all rules with maximum values of conditional probability may be found at the end of such chains. Theoretical advantages of "law-like" rules are presented in [19,21].

3.2. Implementation for Oligonucleotide Patterns

Promoters of genes with common function could be characterized by groups of oligonucleotides motifs. Occurrence of such oligonucleotides reliably higher for promoter sequences. The task is to analyze mutual occurrence and location of these motifs in promoter sequences.

Let us name a group of oligonucleotide motifs forming a certain pattern of relative location in promoter sequences as a complex signal. The presence of such complex signal could be treated as the condition for A_0 to belong to the promoter class. Let us consider the simplest complex signal (S_1, S_2) formed by a pair of oligonucleotides and specified as follows:

$(S_1, S_2) = (\text{Pos}(S_1) < \text{Pos}(S_2))$

where S_1 and S_2 are oligonucleotides in the object-character table; $\text{Pos}(S_1)$, $\text{Pos}(S_2)$ - positions

of these oligonucleotides in the sequence relative to the transcription start. We can consider condition A_1 as (S_1, S_2), and test hypothesis $A_1 \Rightarrow A_0$ for all DNA sequences that contains S_1 and S_2.

But the presence of only two oligonucleotides (S_i,S_j) may be insufficient. So, we should consider all oligonucleotide triples in DNA sequences such as (S_1,S_2,S_3) = ($Pos(S_1) < Pos(S_2) < Pos(S_3)$). Formally this triple could be treated as two pairs (S_1, S_2) and (S_2, S_3). The hypothesis for testing now is $A_1 \& A_2 \Rightarrow A_0$. Thus, using first-order logic we construct more and more complex conditions including the presence of these oligonucleotides in direct or inverted DNA strains, overlapping of the oligonucleotides and so on.

3.3. Data preparation. Signal construction.

The computer system "Gene Discovery" adapts the methods described above to the analysis of nucleotide sequences of regulatory regions.

The learning sample of nucleotide sequences of two alternative classes was used as input. The teaching sample consists of promoters sequences specific for functional system (class 1) and some random sequences (class 2). It could be computer-generated random sequences with the same nucleotide frequencies or real sequences of neighboring regions not corresponding to this regulatory function such as exons.

There is the program block for searching the context signals in the sequences of these two classes (Figure 1). The signal could be:

(1) context (user-defined short nucleotide word (oligonucleotide) or functional site, presented in the specialized molecular-biology database TRRD);

(2) site with conformational or physico-chemical peculiarities (such as angles twist, roll, rise, DNA melting temperature, etc.);

(3) structural element (Z-DNA, RNA hairpin).

All these signals may be recognized using knowledge about DNA properties and the consensus scheme based on experimental data stored in specialized databases. Here we consider two tasks: (i) promoter analysis and recognition using specific degenerate oligonucleotides as signals; (ii) donor splice sites recognition using separate nucleotide bases.

Promoter sequences were extracted from TRRD [10] and divided into several groups according to the transcription regulation specificity (promoters of endocrine gene system, lipid system, heat shock response system, interferon-regulated, glucocorticoid-regulated, cell-cycle system, etc). Here we present analysis of promoter sequences of endocrine gene system. The sample contained 40 sequences of 120 bp length (from –100 bp to +20 bp relative to transcription start). The homology level between any sequence pair did not exceed 60%.

The program ARGO was used to select the specific oligonucleotides of length 8 bp. [1]. The term "degenerate oligonucleotides" is used to denote 15-lettered IUPAC coding for nucleotides.

The selected context signals (degenerate oligonucleotides) in these nucleotide sequences were located and presented in the data table "object-attribute" using input module of "Gene Discovery". In this table DNA sequences are called objects, and attributes show presence of the context signals and their locations relative to the experimentally defined transcription start. This table contains several thousand strings. It contains sequences of the context signals S_i and their positions $Pos(S_i)$ in the promoter region. For example in the first promoter S_1= TGACCAAT, $Pos(S_1)$=-67, S_2=RCCAATND $Pos(S_2)$ =-65, etc. The testing hypothesis A_0 was: "Does the sequence belong to class 1 (promoters)?"

The program developed can use as input any sequence set in FASTA format. A functional sample could be extracted from TRANSFAC [22,24], TRRD [24], EpoDB [23]. Analogously, other functional sets of promoters extracted from the TRRD database were analyzed, including erythroid-specific gene promoters, promoter regions for the cell cycle controlling genes, promoters of genes controlling lipid metabolism, and promoters of genes expressed in muscle.

Hypotheses about complex signals also could have weaker or stronger demands for mutual oligonucleotide location. The great number of regularities for joint appearance of the context signals in the promoter regions was found by "Gene Discovery" system. The number of regularities depends on the user-defined parameters of search.

3.4. Visual presentation and interpretation of complex signals

The regularities found could be analyzed by a molecular biology expert as unique complex signals, which are significant for proper promoter functioning.

The examples of such complex context signals for endocrine system gene promoters are presented in the table 1. Let us consider signal CWGNRGCN < NGSYMTAM < MAGKSHCN. The symbol "<" here designates that positions of corresponding oligonucleotides are ordered relative to the transcription start.

Let us consider selected rules for simultaneous presence of oligonucleotides in promoter as large complex signals. Following additional conditions were used to interpret these complex signals:

(1) oligonucleotides in the complex signal are not overlapped on the promoter sequence;

(2) the observed number N of promoters possessing the complex signal is greater than the expected number N^*, $N>N^*$.

For example, for pattern CWGNRGCN < NGSYMTAM < MAGKSHCN the expected by random number N* was equal 0.47 (i.e. less than 1). But it was present in 6 promoters, that is approximately 13 times greater than expected.

An example of the location of the complex signal is presented in Figure 2. The promoter sequences are aligned relative to the transcription start (position +1 bp), indicated by arrows.

The EMBL identifiers of promoters studied are given in parentheses. The eight-bp oligonucleotide motifs composing the complex signal are shown as dark rectangles; positions of the first nucleotides are indicated relative to the transcription start. Red rectangles mark the positions of TATA-boxes, indicated in the TRRD database; positions of its first and last nucleotides are italicized. It is interesting that only one oligonucleotide in the complex signal corresponds to the annotated context signals in the site sequence.

Table 1. The examples of the complex signals in the endocrine system gene promoters

№	Complex signal (regularity)[1]	Cond. probability of such signal[2]	Fisher statistical criterion[3]	N[4]
1	CWGNRGCN<NGSYMTAM<<CAGGRNCH	0.875	0.00054	4
2	KGRSSAGR<CYCYNSCY<<CWGSNYCH	1.0	0.00012	4

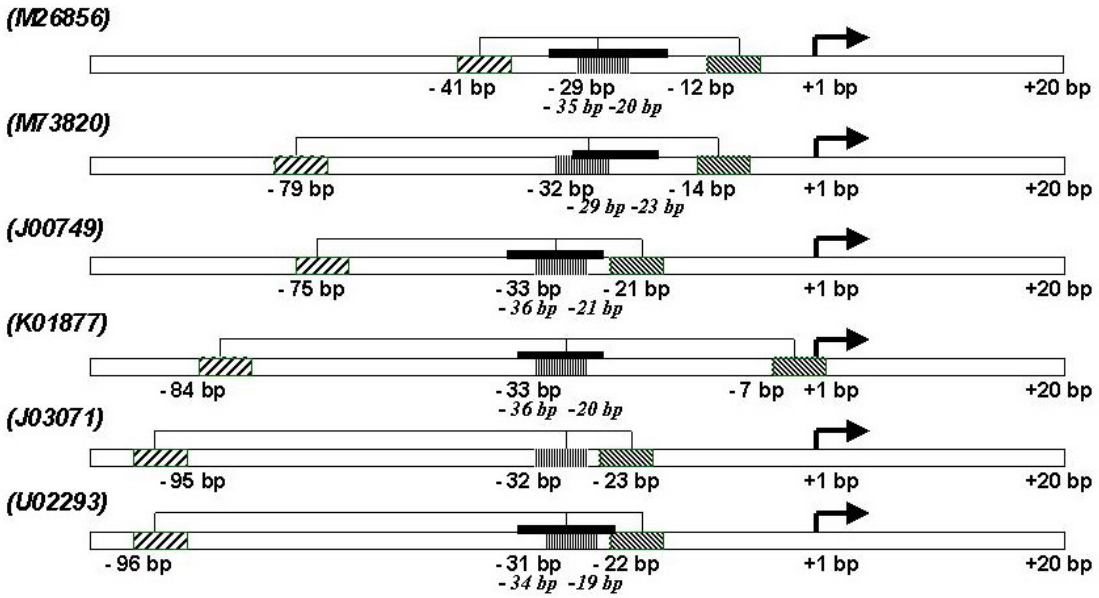


Fig. 2. Schematic localisation of the complex signal CWGNRGCN<NGSYMTAM<MAGKSHCN in promoters of endocrine system genes. The promoter sequences are aligned relative to the transcription start (position +1 bp), indicated by arrows. The EMBL identifiers of the promoters studied are given in parentheses to the left. The eight-bp oligonucleotide motifs composing the complex signal are shown as hatched rectangles; positions of the first nucleotides are indicated relative to the transcription start. The black rectangles mark experimentally determined positions of the TATA-box in the TRRD database. Positions of its first and last nucleotides are italicised.

3	CWGNRGCN<NGSYMTAM< <MAGKSHCN	1.0	0.00009	6
4	CWGNRGCN<NGSYMTAM< <CMDGGNCH	0.846	0.00099	5
5	CNKSAGNT<NCARGRNC< <HNNKGCTG	1.0	0.01426	4
6	RNWGGCCN<DGRGNRGG< <TCMAGNMN	0.875	0.00118	4
7	RGSNRGRG<NNGSTWTA< <CNCNRKGC	1.0	0.02852	5
	...	...	...	..

Notes: Data is not full, gaps denoted as dots.

1 – Complex signals presented as oligonucleotides in 15-lettered coding IUPAC. Sign "<" denotes relation between the positions of corresponding oligonucleotides relative to the transcription start.

2 – Conditional probability $P_{Cond}(N)$ was calculated as quotient of the number of promoters possessing the signal to the total number of promoters.

3 – Probability to obtain in random conditions more observations of the signal than present. It is calculated by the exact Fisher criterion for contingency tables.

4 – *N*, Number of promoters possessing the signal.

Regularities obtained for splice sites contained sub-sequences of bases which being taken into account. These regularities permit to discriminate splice sites from random sequences.

3.5. Recognition procedure

Recognition rule based on complex signals that was found. For oligonucleotides signals the recognition procedure was described in [18].

In this paper the score of some object positions found as the score of all signals applied to this position. This score means the probability of appearance of such signals on the random sequence. Using negative random samples we can estimate the level of the score, which guarantee some levels of first and second kind errors. If on some control sequence score is greater then this levels then we predict that the sequence is from some functional class.

This approach may be extended to the complex signals. On the first step of recognition procedure we find all complex signals applied to the control sequence. As a result we have a sequence of complex signals $0<N<...<N_{total}$, where N_{total} – total number of signals. The order of the signals means the order of the complex signals on this sequence. Then similar score *P(S)* may be calculated for each position of the sequence. Recognition procedure is similar to that of simple signals. If score in some place of the sequence is greater then estimated levels of the first and second kind errors then this sequence is from some functional class.

For donor splice sites first and second types errors for tested data were 4,4% and 4,0% respectively.

4. DISCUSSION

Thus, the system "Gene Discovery" helps us to find complex signals in promoter regions. In a similar way any samples of phased nucleotide sequences could be analyzed. The functional meaning of the signal could be treated in terms of the transcription factors binding sites or the conformational properties of DNA [9,11].

Published experimental data and specialized molecular-biological databases contain a large number of experimental results for DNA sequences involved in transcription regulation. It provides an opportunity for large-scale data mining and knowledge discovery for bioinformatics [6,8]. Our approach was applied mainly to gene regulatory region analysis. Now we analyze context gene structure for all levels of gene hierarchy: promoter, regulatory regions, transcription factor binding sites and its modules, 5'UTR, splice sties.

5. ACKNOWLEDGEMENTS

The authors are grateful to A.S. Belenok, N.L. Podkolodny and G.V. Orlova for help in preparation of the manuscript and to I.B. Rogozin for providing splice site data. The work was partially supported by INTAS (YSF 00-178), RFBR (00-04-49229, 00-04-49255, 00-07-90337, 01-07-90376, 02-07-90355) and Integration project of SB RAS (N65).

6. REFERENCES

[1] Babenko, V.N., Kosarev, P.S., Vishnevsky, O.V., Levitsky, V.G., Basin, V.V. and Frolov, A.S. Investigating extended regulatory regions of genomic DNA sequences. *Bioinformatics*, 1999, **15**, 644-653.

[2] Baxevanis, A.D. The Molecular Biology Database Collection: an updated compilation of biological database resources. Nucleic Acids Research, 2001, 29, 1-10.

[3] Fickett, J.W. and Hatzigeorgiou, A.G. Eukaryotic promoter recognition. Genome Res., 1997, 7, 861-878.
[4] Horng, J.-T., Huang, H.-D., Huang, C.-C. and Kao, C.-Y. Mining Putative Regulatory Elements in Gene Promoter Regions. *GCB'2001*, October 7-10, Braunschweig, 2001, 90-95.
[5] Hu Y.-J. Biological Sequence Data Mining. In De Raedt, L. and Siebes A. (eds.): PKDD 2001, LNAI 2168, Springer-Verlag Berlin Heidelberg, 2001, pp. 228-240.
[6] Jakobsen, I.B., Saleeba, J.A., Poidinger, M. and Littlejohn, T.G. TreeGeneBrowser: phylogenetic data mining of gene sequences from public databases. Bioinformatics, 2001, 17, 535-540.
[7] Kel, A., Kel-Margoulis, O., Ivanova, T., Wingender, E. ClusterScan: A Tool for Automatic Annotation of Genomic Regulatory Sequences by Searching for Composite Clusters. In: GCB'2001, Oct. 7-10, 2001, 96-101.
[8] King, R.D., Karwath, A., Clare, A. and Dehaspe, L. The utility of different representations of protein sequence for predicting functional class. Bioinformatics, 2001, 17, 445-454.
[9] Klingenhoff, A., Frech, K., Quandt, K. and Werner, T. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. Bioinformatics, 1999, 15(3), 180-186.
[10] 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. Transcription regulatory regions database (TRRD): its status in 2000. Nucleic Acids Research, 2000, 28(1): 298-301.
[11] Kondrakhin, Y.V, Kel, A.E., Kolchanov, N.A., Romaschenko, A.G. and Milanesi, L. Eukaryotic promoter recognition by binding sites for transcription factors. Comput. Appl. Biosci., 1995, 11, 477-488.
[12] Kovalerchuk, B. and Vityaev, E. Data Mining in finance: Advances in Relational and Hybrid Methods, Kluwer Academic Publishers, 2000, 308 p.
[13] Kovalerchuk, B., Vityaev, E., Ruiz, J. Consistent Knowledge Discovery in Medical Diagnosis. IEEE Eng. in Medcine and Biol. Mag., 2000, July/August, 26-37.
[14] Kovalerchuk, B., Vityaev, E., Ruiz, J.F. Consistent and Complete Data and "Expert" Mining in Medicine. In: Medical Data Mining and Knowledge Discovery, Springer, 2001, 238-280.
[15] Kretschmann, E., Fleischmann, W. and Apweiler, R. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT Bioinformatics, 2001, 17, 920-926.
[16] Mitchell T. Machine Learning. NY: McGraw Hill. 1997.
[17] Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P. and Moreau Y. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics, 2001, 17, 1113-1122.
[18] Vishnevsky, O.V., Vityaev, E.E. Analysis and recognition of promoters of the erythroid-specific genes on the basis of degenerated oligonucleotide motifs. Mol. Biol. (Mosk). 2001, 35(6), 979-986 (Russian).
[19] Vityaev, E.E. and Moskvitin, A.A. Introduction to discovery theory. "Discovery" software system. Comp. Syst., Novosibirsk 1993, 148, 117-163 (Russian).
[20] Vityaev, E.E, Orlov, Yu.L., Vishnevsky, O.V., Belenok, A.S. and Kolchanov, N.A. Computer system "Gene Discovery" for regularities retrieving in eukaryotic regulatory sequences organisation. Mol. Biol. (Mosk). 2001, 35(6), 952-960 (Russian).
[21] Vityaev, E.E. Semantic approach to knowledge base development: Semantic probabilistic inference. *Com. Syst.,* Novosibirsk 1992, **146**, 19-49 (in Russian).
[22] Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, Pruss M, Schacherer F, Thiele S, Urbach S. (2001) The TRANSFAC system on gene expression regulation. Nucleic Acids Res. Jan 1 29(1), 281-3.
[23] EpoDB: http://www.cbil.upenn.edu/EpoDB/
[24] Transcription Regulatory Regions Database http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/ TRANSFAC database: http://www.gene-regulation.de/
[25] Scientific Discovery web-site http://www.math.nsc.ru/LBRT/logic/vityaev