Computer System "Gene Discovery" for Functional Annotation of DNA Sequences

Vityaev E.E.[1], Orlov Yu.L.[2]*, Vishnevsky O.V.[2], Kovalerchuk B.K.[3], Belenok A.S.[2], Podkolodny N.L.[2], Kolchanov N.A.[2]

[1] Sobolev Institute of Mathematics SB RAS, Acad. Koptyug prospect, 4, Novosibirsk, 630090.

[2] Institute of Cytology and Genetics SB RAS, Acad. Lavrentiev ave., 10, Novosibirsk, 630090.

[3] Boris Kovalerchuk Computer Science Department, Central Washington University, Ellensburg, WA, 98926-7520, USA.

[*] Corresponding author. E-mail: orlov@bionet.nsc.ru

Summary

Methods of Data Mining and Knowledge Discovery were implemented for the search of regularities in tables of context features of DNA sequences involved in transcription regulation. The task was to retrieve regularities connecting nucleotide sequences with the functional class of those sequences. The search patterns were constructed in first-order logic with probability. For discovering regularities a computer program "Gene Discovery" was designed. The program accepts molecular-genetics data by SQL queries. Sequences of erythroid-specific promoters and promoters of genes of endocrine system from TRRD database (Kolchanov N.A. et al., 2000) were analysed by this system. Regularities connecting the nucleotide sequences in regulatory DNA and its location relative to the start of transcription and functional class were found. The recognition method of regulatory DNA promoter class founded on these regularities was developed.

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

BIOLOGICAL TASK

Analysis of the promoters structure is of great interest for understanding of molecular mechanisms of gene transcription. The core (basal) promoter is the main element of the gene regulatory region necessary for transcription initiation. Promoters in eukaryotic organisms (higher organisms, including mammals and human) act as the molecular "switches" that turn genes on and off. Each gene has at least one promoter upstream of the protein encoding part of the gene. It is considered as a minimal DNA sequence necessary for proper initiation of transcription in vitro. The sequence of a core promoter contains the point at which transcription starts and region approximately –60 to +40 b.p. relative to it (Zhang M.Q., 1998; Arnone M.I. & Davidson E.H., 1997). Promoter is the main regulatory region for gene expression. It contains

transcription factor binding sites - short stretches of DNA, sufficiently conserved to allow specific recognition by the corresponding protein (Nikolov D.B. & Burley S.K., 1997).

The presence and location of the transcription factor binding sites in 5' regulatory regions of genes corresponds to tissue- and stage-specific features of gene expression in organism. One gene can contain several promoters to define expression of different protein products or proteins with different levels of specific functional activity. Moreover these eukaryotic promoters characterised by the absence of exact localisation of context signals and the weakness of such signals (Goodrich J.A. et al., 1996). It is diversity of promoters which is the main difficulty for the developing of the recognition programs (Fickett J.W. & Hatzigeorgiou A.G., 1997). The problem of recognition accuracy remains even though a large number of computer methods for the recognition of RNA Polymerase II promoter have been created (Pedersen A.G. et al., 1999; Solovyev V. & Salamov A., 1997; Kondrakhin Y.V. et al., 1995). Proposed promoter recognition methods are based on weight matrices or some other measures of similarity to the sequences from databases. At the same time, the choice of sequence region used to estimate similarity were based on biological suggestions. These suggestions were not independently statistically tested.

Data mining in other fields tends to use larger databases and discover larger sets of rules. At the same time, experimental data distributed in literature and specialised molecular-biological databases around the world contains thousands of experimental results for DNA sequences involved in transcription regulation. Currently, about 300 molecular-biological databases are available on the Internet (Baxevanis A.D., 2001). Such efforts provide the opportunity for large-scale data mining and knowledge discovery for bioinformatics (Jakobsen I.B. et al., 2001).

THE COMPUTER PROGRAM "GENE DISCOVERY": PRINCIPAL SCHEME

The computer program "Gene Discovery" was developed for analysis of structural organization of eukaryotic promoters using information of experimentally proved and computer-predicted sites. System "Gene Discovery" is an adaptation of the system "Discovery" (Kovalerchuk B. & Vityaev E., 2000) to molecular biology tasks. "Gene Discovery" consists of 3 main modules: (1) the module for on-line representation of context signals from DNA sequence in standard table form; (2) the module "Discovery" for regularities search; (3) the module of recognition of the sequence class using the regularities found. The program is written in C++ and it has user-friendly interface. The Data Mining module "Discovery" is designated in the Figure 1 as “Search for patterns of the joint presence and relative localization of contextual signals…”.

Other modules of the system are supplements for preparation and interpretation of the molecular-genetic data.

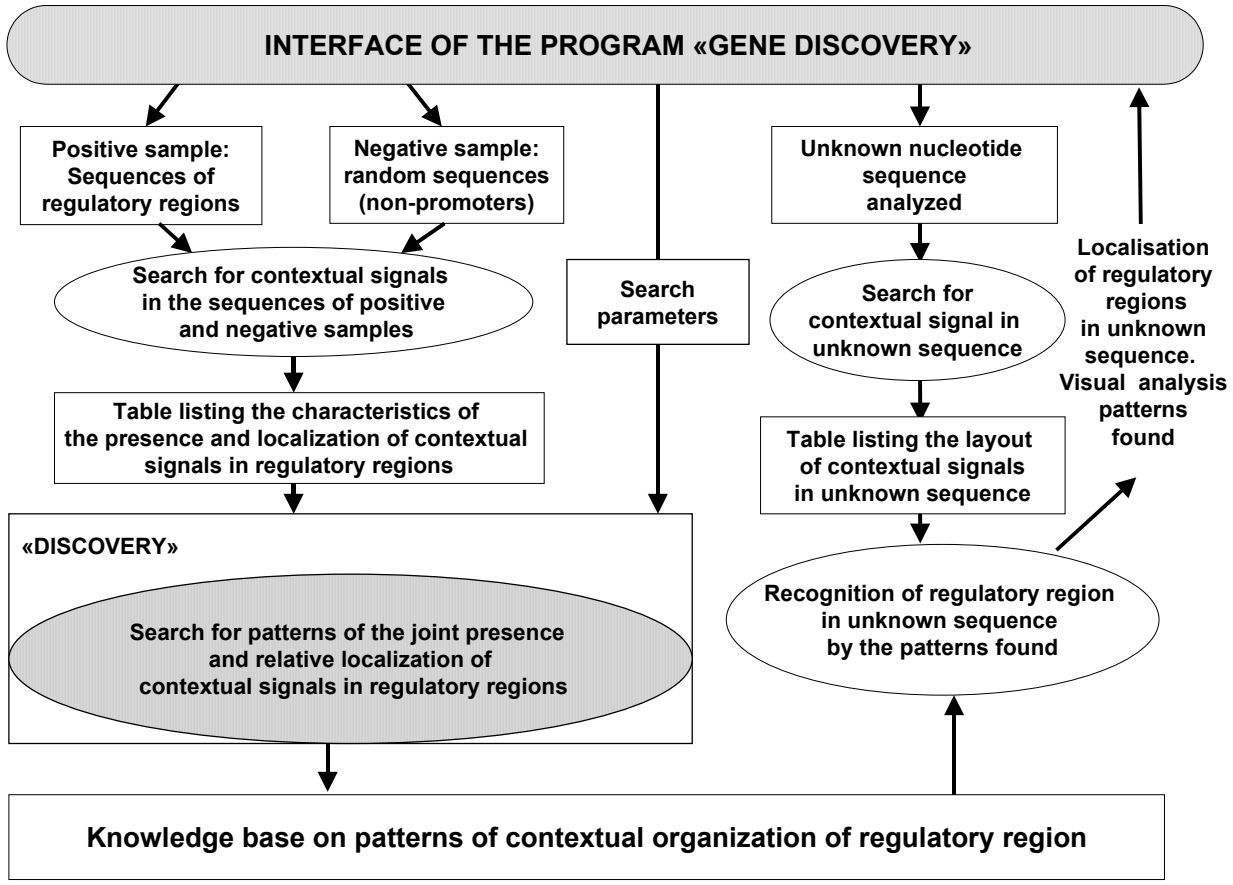

Figure 1. The scheme of the "Gene Discovery" system.

"DISCOVERY": TECHNOLOGY OF DATA MINING

A machine learning method and system “Discovery” (Vityaev E.E. & Moskvitin A.A., 1993), discover statistically significant first-order logic rules for functional annotation of regulatory regions. Learning systems based on first-order representations have been successfully applied to many problems in psychology, physics, medicine, finance, and other fields (Kovalerchuk B. et al., 1992, 1996, 1997; Kovalerchuk B. & Vityaev E. 2000, 2001) (see also www-site "Scientific Discovery": http://www.math.nsc.ru/LBRT/l/vityaev/, section "comparison"). As with any technique based on logic rules, this technique allows one to obtain human-readable forecasting rules that are interpretable in biological language and also provides a promoter recognition (functional annotation) (Mitchell T., 1997). A specialist in biology can evaluate the correctness of the recognition as well as the rules itself. The critical issue in applying data-driven forecasting systems is generalisation. “Discovery” software systems generalise data through “law-like” logical probabilistic rules.

The concept of probabilistic causality (P.Suppes, 1970) for $Y \Rightarrow X$ ("that allow us to deduce that X is a probable cause of Y when the probability of X given Y is different from the probability of X") may be generalized to expressions $A_1\& \ldots \& A_k \Rightarrow A_0$ as follows: expressions $A_1, \ldots, A_k$ allows us to deduce that they are a probable cause of A_0 if the conditional probability $CondProb(A_0|SubFormular(A_1\&\ldots\&A_k))$ of A_0 under any subcondition $SubFormular(A_1\&\ldots\&A_k)$ is strictly less than the conditional probability $CondProb(A_0|A_1\&\ldots\&A_k)$ of A_0 under full condition. In particular, for the expression $Y \Rightarrow X$, if we get $SubFormular(Y) = \varnothing$, than the conditional probability $CondProb(X|SubFormular(Y)) = CondProb(X|\varnothing) = Probability(X)$ must be strictly less than $CondProb(X|Y)$ — the probability of X given Y. This gives us the following definition (Kovalerchuk B. & Vityaev E., 2000):

Definition: $A_1, \ldots, A_k$ is a probabilistic "Law-like" rule (probable cause) of A_0 if:

$CondProb(A_0|SubFormular(A_1\& \ldots\&A_k)) < CondProb(A_0|A_1\& \ldots\&A_k)$

for any $SubFormular(A_1\& \ldots\&A_k)$, where

$SubFormular(A_1\& \ldots\&A_k)=A^s_1\& \ldots\&A^s_k$, $\{A^s_1, \ldots, A^s_k\} \subset$ (not equal)$\{ A_1\& \ldots\&A_k\}$.

The "Law-like" rule definition satisfies all properties of scientific laws. Conceptually, **law-like rules** came from the philosophy of science. These rules attempt to capture mathematically the essential features 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 $A_1, \ldots, A_k$, 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 $SubFormular(A_1\& \ldots\&A_k) \Rightarrow A_0$ 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 a sub-rule is logically stronger than the rule used to construct the sub-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, because 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 (or perhaps some of its sub-rules) will be preferred, because this sub-rule is simpler, more general and more refutable. In the **deterministic case**, a **"law-like" rule** can be defined (for some set of examples) as a rule

without sub-rules covering this set of examples. In other words, "law-like" rule is the rule, which is true for some set of examples, but none of its sub-rules is true for this set.

If examples contain noise, which is typical in life sciences, the probabilistic characteristics of the expressions are used instead of crisp (true/false) values. The conditional probability of the rule is used in the "Discovery" system as this characteristic. The conditional probability of rule C is defined as $Prob(C) = CondProb(A_0|A_1\& \ldots\&A_k)$, assuming that $Prob(A_1\& \ldots\&A_k) > 0$. Similarly, conditional probabilities $Prob(A_0|A_{i1}\&...\&A_{ih})$ are defined for sub-rules $C_i = (A_{i1}\&...\&A_{ih} \Rightarrow A_0)$, assuming that $Prob(A_{i1}\&...\&A_{ih}) > 0$. The conditional probability Prob(C) is used for estimating the forecasting power of the rule to predict A_0. In addition, the conditional probability is a major tool for defining **non-deterministic** (**probabilistic**) "**law-like" rules** (Kovalerchuk B. et al., 2001; Vityaev E.E., 1992).

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.** "Law-like" rules defined in this way hold all three properties listed above (properties of scientific laws), i.e., these rules are (1) general from a logical perspective, (2) simple, and (3) refutable.

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 , $C_1 = sub(C_2)$, C_2 is a subrule of rule C_3, $C_2 = sub(C_3)$ and finally C_{m-1} is a subrule of rule C_m, $C_{m-1} = sub(C_m)$. Also $Prob(C_1) < Prob(C_2), \ldots, Prob(C_{m-1}) < Prob(C_m)$. There is a theorem (Vityaev, 1992) that all rules, which have a maximum value of conditional probability, can be found at the end of such chains.

The goal of the "Discovery" system is to find all strongest predictive rules (ST rules).

Definition 1. Rule $C = (A_1\& \ldots\&A_k \Rightarrow A)$, where $k > 1$ and $Prob(A_1\& \ldots\&A_k) > 0$ is called a strongest predictive rule for atomic formula A on data D iff:

1) $Prob(C) = Prob(A|A_1\& \ldots\&A_k) > Prob(A)$, where $A_1\& \ldots\&A_k$ are conditions generated using data D;

2) Rule C has maximum of conditional probability Prob(C) among rules C* satisfying condition 1 and generated by the same data as C, and

3) For any rule C* satisfying conditions 1, 2, we have $C \Rightarrow C^*$.

Condition 1 means that the rule C has a reasonable If-part (premise), i.e., the conditional probability Prob of rule C is greater then the probability of the atomic formula A itself. If this condition is not satisfied then there is no reason to add the premise for forecasting A. If atom A has a high probability itself then it can be predicted without a premise.

Condition 2 brings the "strongest" rule, i.e., the rule with maximum conditional probability among rules satisfying condition 1 above for the same data.

Condition 3 means that a ST rule is the most "general" among rules satisfying conditions 1 and 2, i.e., a BST rule covers the widest set of cases from D for which it can be applied.

There is a **theorem** (Vityaev, 1992) that **all ST rules can be found at the end of chains** C_1 , C_2 , …, C_{m-1}, C_m. So the goal of the "Discovery" system is achieved.

The algorithm stops generating new rules when they become too complex (i.e., statistically insignificant for the data) even if the rules are highly accurate on training data. The Fisher statistical criterion (exact Fisher test for contingency tables) is used in this algorithm for testing statistical significance (Kendall M.G. & Stuart A., 1977).

The obvious other stop criterion is limitation of the number of conditions A_k (the number of data fields in the analysed table).

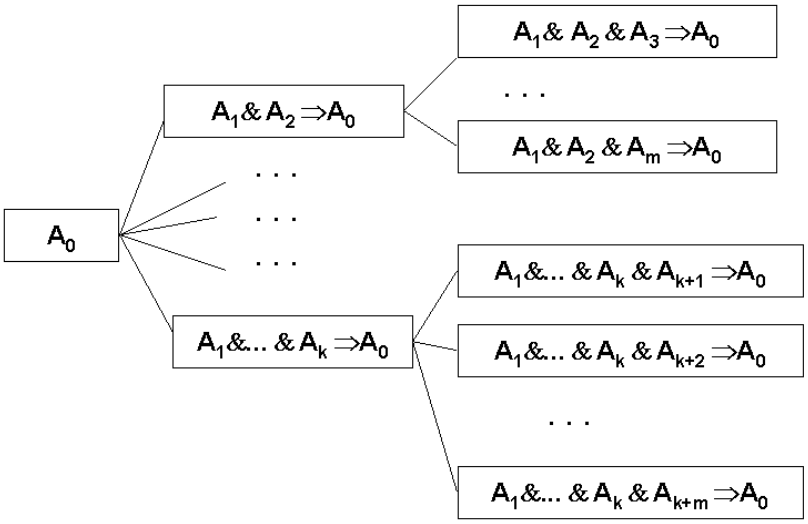

Figure 2. An example of the rule search for hypothesis A_0.

Theoretical advantages of this generalisation are presented in theorem (Kovalerchuk B. et al., 1996; Vityaev E.E., 1992; Vityaev E.E., Moskvitin A.A., 1993). This approach has some similarity with the hint approach (Abu-Mostafa, 1990). We use mathematical formalisms of first-order logic rules described in (Russel S. & Norvig P., 1995; Halpern J.Y., 1990; Krantz D.H. et al. 1971, 1989, 1990). Note that a class of general propositional and first-order logic rules covered by the methods is wider than a class of decision trees (Mitchell T., 1997).

DATA PREPARATION

The computer system "Gene Discovery" adapts the methods described above to the analysis of nucleotide sequences of regulatory regions. The principal scheme is presented in Figure 1.

The teaching sample of nucleotide sequences of two alternative classes is used as input to the system. The teaching sample consists of the sequences of promoters specific to the 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 neighbouring regions not corresponding to this regulatory function such as exons.

There is the program block to search for the context signals in the sequences of these two classes (Figure 1). The signal could be: context (user-defined short nucleotide word (oligonucleotide) or functional site, presented in the specialised molecular-biology databases TRRD and Transfac); conformation (DNA region is characterised by peculiarities of physico-chemical properties, for example easily melting DNA region, curved DNA etc.); structural (Z-DNA, RNA hairpin).

All these signals may be recognised using knowledge about DNA properties and the consensus scheme based on experimental data stored in specialised databases.

Here we will consider degenerate oligonucleotides as context signals specific to promoters.

The sequences of endocrine system genes promoters were analysed. The sample of 40 sequences was extracted from the database ES-TRRD (http://wwwmgs.bionet.nsc.ru/). The length of the sequences was 120 b.p. (from –100 b.p. to +20 b.p. relative to transcription start). The level of homology between any pair did not exceed 60%.

The program ARGO was used to select the specific oligonucleotides of length 8 b.p. (BabenkoV.N. et al., 1999), (see also http://wwwmgs.bionet.nsc.ru/mgs/programs/argo/). The term "degenerate oligonucleotides" is used to denote 15-lettered IUPAC coding for nucleotides. This is the standard way to present the similar nucleotide strings as one signal. Thus, the preliminary step for data preparation was fulfilled. Similar signals could be obtained from databases by homology to protein binding sites.

The selected context signals (degenerate oligonucleotides) in these nucleotide sequences were located and presented in the data table by using a module of "Gene Discovery". So the data were presented in the table "object-attribute". In this table objects are DNA sequences, attributes are presence of the context signals and their location relative to the experimentally defined transcription start. By the same way other samples of promoters were analysed, particularly the promoters of erythroid genes. So we construct table for data mining of several thousand strings. It contains sequences of the context signals S_i and their positions Position(S_i) in the promoter region. For example for the first promoter in the sample under analysis S_1=TGACCAAT, Position(S_1)=-67, S_2=RCCAATND, Position(S_2)=-65, etc.

IMPLEMENTATION OF THE METHOD FOR DNA SEQUENCES

The testing hypothesis A_0 was: "Does the sequence belong to class 1 (promoters)?".

Let us name a group of oligonucleotide motifs displaying 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) = (Position(S_1)<Position(S_2))

where S_1 and S_2 are oligonucleotides in the object-character table; Position(S_1) and Position(S_2) are positions of these oligonucleotides in a sequence relative to the transcription start. So, 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 not be a satisfactory condition. So, we should consider all oligonucleotide triples in DNA sequences such as (S_1,S_2,S_3) = (Position(S_1)<Position(S_2) <Position(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.

The great number of regularities for joint appearance of the context signals in the promoter regions was found as a result of the "Gene Discovery" search. The number of regularities depends on the user-defined parameters of this search. If we define a low level of conditional probability (less than 0.5) then the number of resulting rules will be too large (up to several thousand). It is a complicated task for an expert to interpret such number of rules. Also we may demand high level of conditional probability, for example, greater than 0.95. So the number of rules would be small, but very significant from a biological point of view.

INTERPRETATION OF THE RULES

The regularities found could be analysed by a molecular biology expert as unique complex signals which are significant for proper promoter functioning. Let us consider selected rules for simultaneous presence of oligonucleotides in promoter as large complex signals. These additional conditions were used to interpret these complex signals:

(1) the 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*.

The examples of such complex context signals for endocrine system genes 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.

The expected by random number N* for this signal equals to 0.47 (i.e. less than 1). But it is present in 6 promoters, that is approximately 13 times greater than expected (see Table 1).

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

№	Complex signal (regularity)[1]	Conditional probability of such signal[2]	Fisher statistical criterion[3]	Number of promoters possessing the signal[4]	Number of promoters expected by random[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	0.4 (<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)

Notes: Data in the table 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. The gaps between the neighbouring oligonucleotides positions are not fixed.

2 – Conditional probability $PC(N_1,N_2)$ 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 – Number of promoters possessing the signal.

5 – The expected number of promoters in random conditions possessing the complex signal. It is estimated by product of frequencies of the oligonucleotides taking into account all the variants of their mutual location.

An example of the location of the complex signal is presented in Figure 3.

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 shaded rectangles; positions of the first nucleotides are indicated relative to the transcription start. Black rectangles mark the positions of TATA-boxes, indicated in the TRRD database; positions of its first and last nucleotides are italicised. It is interesting that only one oligonucleotide in the complex signal corresponds to the annotated site. Other oligonucleotides could correspond to potential transcription factor binding sites or to regions with specific physico-chemical properties of double-stranded DNA.

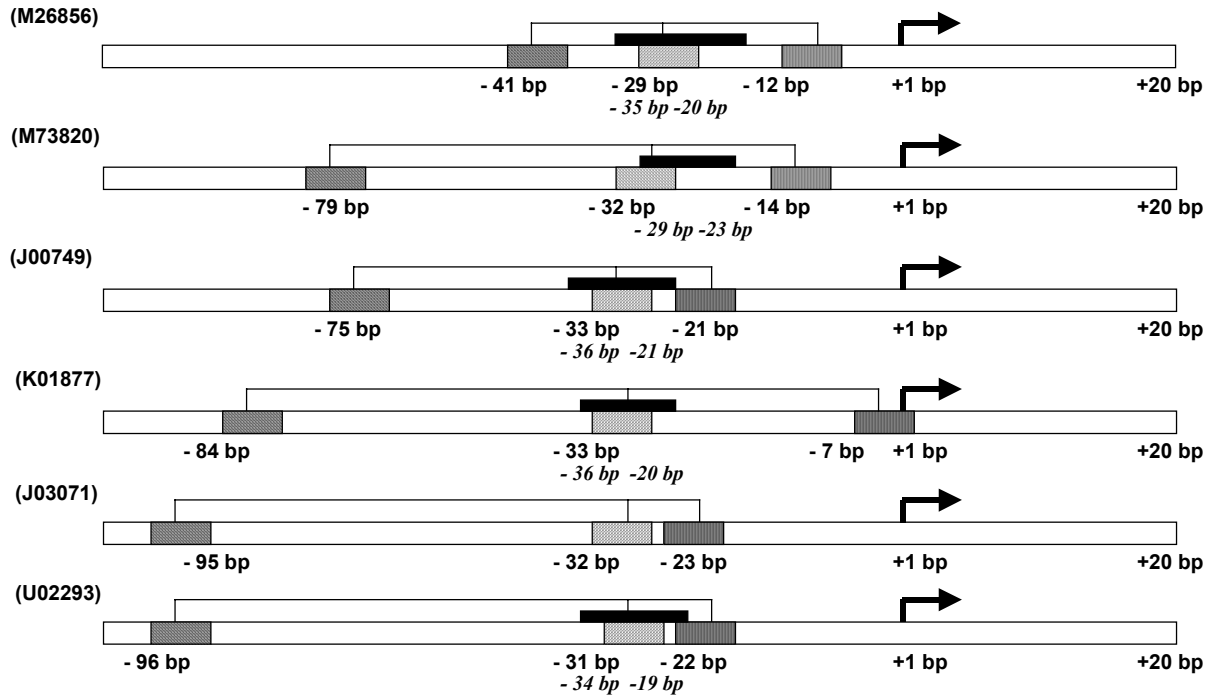

Figure 3. Schematic localization of the complex signal CWGNRGCN<NGSYMTAM<MAGKSHCN in promoters of endocrine system genes. The promoter sequences are aligned relative to the transcription start (position +1 bp), indicated by arrows. The EMBL identifiers of the promoters studied are given in parentheses to the left. The eight-bp oligonucleotide motifs composing the complex signal are shown as textured rectangles; positions of the first nucleotides are indicated relative to the transcription start. The black rectangles marks experimentally defined positions of the TATA-box indicated in the TRRD database. Positions of its first and last nucleotides are italicized.

Thus, the developed system "Gene Discovery" help us to find out complex signals in promoter regions. In a similar way any samples of nucleotide sequences could be analysed.

The functional meaning of the signal is proved by experts in biology. Its could be treated in terms of the transcription factors binding sites or the conformational properties of DNA (Kondrakhin Yu.V. et al., 1995; Klingenhoff A. et al., 1999).

Promoter recognition on the basis of regularities found by the system is a further topic for discussion. It should be noted that the system does not over-learn on the training samples and shows the level of false positive rate on control samples corresponding to the conditional probability of selected rule.

The authors are grateful to V.Levitsky, C.Gabel and D.Andor for scientific discussions and support of this work. The research was supported by RFBR and Siberian Division of RAS (Integration grant #65). Yu.Orlov was supported by INTAS (YSF 00-178).

References

1. Abu-Mostafa. Learning from hints in neural networks. *J Complexity* 1990, 6: 192-198.
2. Arnone M.I., Davidson E.H. The hardwiring of development: organization and function of genomic regulatory systems. *Development* 1997, 124(10): 1851-1864.

3. Babenko V.N., Kosarev P.S., Vishnevsky O.V., Levitsky V.G., Basin V.V., Frolov A.S. Investigating extended regulatory regions of genomic DNA sequences. *Bioinformatics* 1999, 15: 644-653.
4. Baxevanis A.D. The Molecular Biology Database Collection: an updated compilation of biological database resources. *Nucleic Acids Research* 2001, 29: 1-10.
5. Fickett J.W., Hatzigeorgiou A.G. Eukaryotic promoter recognition. *Genome Res.* 1997, 7: 861-878.
6. Goodrich J.A., Cutler G., Tjian R. Contacts in context: promoter specificity and macromolecular interactions in transcription. *Cell* 1996, 84(6): 825-830
7. Halpern J.Y. An analysis of first-order logic of probability. *Artificial Intelligence* 1990, 46: 311-350.
8. Jakobsen I.B., Saleeba J.A., Poidinger M., Littlejohn T.G. TreeGeneBrowser: phylogenetic data mining of gene sequences from public databases. *Bioinformatics* 2001, 17: 535-540.
9. Kendall M.G., Stuart A. The advanced theory of statistics, 4th ed., Charles Griffin & Co LTD, London. 1977.
10. Klingenhoff A, Frech K, Quandt K, Werner T. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* 1999, 15(3): 180-186.
11. 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., Overton G.C. Transcription regulatory regions database (TRRD): its status in 2000. *Nucleic Acids Research* 2000, 28(1): 298-301.
12. Kondrakhin Y.V, Kel A.E., Kolchanov N.A., Romaschenko A.G., Milanesi L. Eukaryotic promoter recognition by binding sites for transcription factors. *Comput Appl Biosci*, 1995, 11: 477-488.
13. Kovalerchuk B. and Talianski V. Comparison of empirical and computed fuzzy values of conjunction. *Fuzzy Sets and Systems* 1992, 46: 49-53.
14. Kovalerchuk B., Triantaphyllou E., Ruiz J. Monotonicity and logical analysis of data: A mechanism for evaluation of mammographic and clinical data. In: Kilcoyne RF, Lear JL, Rowberg AH (Eds): Computer Applications to assist Radiology, Carlsbad, CA: Symposia Foundation, 1996, 191-196.
15. Kovalerchuk B., Vityaev E., and Ruiz J.F: Design of consistent system for radiologists to support breast cancer diagnosis. In *Proc Joint Conf Information Sciences*, Durham, NC, 1997, 2: 118-121.
16. Kovalerchuk B., Vityaev E. Data Mining in finance: Advances in Relational and Hybrid Methods. (Kluwer international series in engineering and computer science; SECS 547), Kluwer Academic Publishers, 2000, 308 p.
17. Kovalerchuk B., Vityaev E., Ruiz J. Consistent Knowledge Discovery in Maedical Diagnosis. *IEEE Engineering in Medicine and Biology Magazine*. Special issue: "Medical Data Mining", July/August 2000, 26-37.
18. Kovalerchuk, B., Vityaev E., Ruiz J.F. Consistent and Complete Data and "Expert" Mining in Medicine, In: Medical Data Mining and Knowledge Discovery (Book chapter), Springer, 2001, 238-280.
19. Krantz D.H., Luce R.D., Suppes P., Tversky A. Foundations of measurement. Vol. 1,2,3 - NY, London: Acad. press, 1971. 577 p., 1989, 493 p., 1990. 356 p.
20. Mitchell T. Machine Learning. New York: McGraw Hill, 1997.
21. Nikolov D.B., Burley S.K. RNA Polymerase II transcription initiation: A structural view. *Proc. Natl. Acad. Sci. USA*, 1997, 94: 15-22.
22. Pedersen A.G., Baldi P., Chauvin Y., Brunak S. The biology of eukaryotic promoter prediction - a review. *Comput. Chem*., 1999, 23: 191-207.
23. Russel S. and Norvig P. Artificial Intelligence. A Modern Approach. Englewood Cliffs, NJ: Prentice Hall, 1995.
24. Solovyev V., Salamov A. The gene-finder computer tools for analysis of human and model organisms genome sequences. In: Proceedings, Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB-97), 1997, 294-302.
25. Suppes P. A probabilistic Theory of Causality, North-Holland, Amsterdam, 1970.
26. Vityaev E.E., Moskvitin A.A. Introduction to discovery theory: Discovery software system. *Computational Systems* 1993, 148: 117-163, Novosibirsk (in Russian).
27. Vityaev E.E. Semantic approach to knowledge base development: Semantic probabilistic inference. *Computer Systems* 1992, 146: 19-49, Novosibirsk (in Russian).
28. Zhang M.Q. Identification of human gene core-promoters in silico. *Genome Res.* 1998, 8: 319-326.