Computer System "Gene Discovery" for Promoter Structure Analysis

Nikolay A. KOLCHANOV, Mikhail A. POZDNYAKOV, Yury L. ORLOV,
Oleg V. VISHNEVSKY, Nikolay L. PODKOLODNY
Institute of Cytology and Genetics SB RAS, Lavrentieva ave., 10, Novosibirsk, 630090, Russia. E-mails: {kol,mike,orlov,oleg,pnl}@bionet.nsc.ru

Eugenii E. VITYAEV
Sobolev Institute of Mathematics SB RAS, Acad. Koptyug prospect, 4, Novosibirsk, 630090, Russia. E-mail: vityaev@math.nsc.ru

Boris KOVALERCHUK
Department of Computer Science, Central Washington University, Ellensburg, WA, 98926-7520, USA. E-mail: borisk@cwu.edu

Abstract. Data Mining and Knowledge Discovery techniques proved to be efficient tools for variety of complex tasks in biology including DNA research. This paper presents implementation of these techniques for searching regularities in tables of context features of DNA sequences involved in transcription regulation. The goal is to discover regularities that interrelate nucleotide sequences with the functional class of these sequences. The search for regularities is implemented in a software system "Gene Discovery" which is based on first-order probabilistic logic. The "Gene Discovery" system provides a general scenario of functional annotation of an arbitrary nucleotide sequence. This system accepts molecular-genetical data retrieved from the database using SQL queries. Sequences of non-homologous gene promoters extracted from the TRRD database have been analysed using this system. Several regularities have been detected. These regularities relate the context of regulatory DNA nucleotide sequence and its location relative to transcription start with the functional class. Our approach based on Data Mining methods selects a specific oligonucleotide pattern for description of the functional class of a gene. The method relies to original Knowledge Discovery approach that can be used for wide variety of complex bioinformatics problems.

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

1. Introduction

Techniques of the large-scale Data Mining, Knowledge Discovery, and other computational approaches of Machine Learning were intensively used in bioinformatics [1,2] usually for text database analysis. Data mining Systems based on first-order logic is a special class of Data Mining techniques with highest expressive abilities to represent complex patterns. These methods have been successfully applied for many problems in psychology, physics, medicine, finance, and other fields [3,4,5,6,7], and web site http://www.math.nsc.ru/LBRT/logic/vityaev/). As with any technique based on logic rules [8], this technique allows one to obtain human-readable forecasting rules that are interpretable in biological language. The discovery has two faces: (1) discovery of rules and (2) discovery of properties of promoter regions and record them as functional annotations

of genes. A biologist may evaluate both the correctness of the final discovery and rules themselves. In this paper we apply a machine learning method and system "Gene Discovery" [3,4] for functional annotation of regulatory regions [9,10]. The system discovers statistically significant first-order logic rules for this problem.

Analysis of gene regulatory regions is of great interest for understanding molecular mechanisms of transcription. Regulatory sequences constitute a small fraction of the roughly 95% of the mammalian genome that does not encode proteins, but they determine the level, location and chronology of gene expression [11]. Despite the importance of these non-coding sequences in gene regulation, our ability to identify and predict functions for this category of DNA is extremely limited.

The control of eukaryotic gene expression is primarily determined by relatively short sequences (signal/motif) in gene promoter region. These sequences vary in length, position, redundancy, orientation in DNA chain, and bases. Eukaryotic promoters are characterised by the absence of exact localization of context signals and the weakness of such signals [12]. Diversity of promoters is the main difficulty for developing of recognition programs [13].

The availability of consensus target sequences for many of the known transcription factors has been used to construct databases that can be searched to identify potential transcription factor binding sites (TFBS) in a DNA sequence [14,15]. Although useful data sets have been generated, the identification of such sites still presents a formidable challenge. We refer to a number of site prediction programs as first step to extract knowledge on promoter structure [16,17,18,19,20]. Despite the fact that some transcription factors bind to highly specific DNA sequences, most have a small invariant core sequence (about 4-6 bp) surrounded by a variable number of degenerate nucleotides.

We are solving this problem by using several methods:
(1) using specialised database, such as TRRD and its sections [14],
(2) combining of various statistical prediction programs [20], and
(3) estimating statistically defined degenerate oligonucleotides as potential TFBS [21].

TFBS or potential sites serve as input table characters from the Data Mining viewpoint. The computational detection of gene regulatory regions is a powerful complement to novel experimental approaches. If known structures provide a template, simple consensus searches, matrix approaches and also programs taking into account specific features, structural constraints and energy values are available. Such works are reviewed in [22]. If no genomic template structures are available then other approaches are implemented (neural networks [23,24,25]; language based approaches [26] and other non-consensus search methods, e.g. [27]).

The large number of false-positive returns is one of the major difficulties with the output from transcription factor database search. The short length and degenerate nature of TFBS account for most of these misleading predictions. Predictions can be further strengthened if sequence context in which a predicted site found is taken into account. Other approaches include clustered binding sites [28,29]. Recently several computational approaches have been suggested to address challenges of combinatorial regulation of transcription [30,31]. In particular, they concern computer selection of specific oligonucleotides [32] and mining associations between them [33,34].

Our approach based on Data Mining methods selects a specific oligonucleotide pattern for description of the functional class of a gene [9,10]. The method relies to original Knowledge Discovery approach that can be used for wide variety of other complex bioinformatics problems.

The basis for software development and use is a training sample of nucleotide sequences of promoter region. It is hard to describe all eukaryotic promoter sequences by a common pattern due to a huge variability of different transcription factor binding sites. To overcome this difficulty, the sets of promoters of genes performing the similar function

were extracted from the TRRD database [14]. 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 that describe a subgroup of the training set. Input data are converted into first-order logic form and search for such patterns is guided by probabilistic estimates [4,35].

2. Analysis of Gene Structure

This paper discovers the functional type of regulatory region based on a predicted set of protein binding sites and the context parameters of the nucleotide sequence. We consider primarily gene promoters for testing the methodology.

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. Promoters in eukaryotic organisms 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 transcription start and region approximately –60 to +40 bp relative to it [36]. 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 [37]. The presence and location of the transcription factor binding sites in 5' regulatory regions of genes are related to the tissue- and stage-specific features of gene expression in an organism. In any case these characteristics make the problem computationally difficult even for signal finding [33]. In spite of a large number of RNA Polymerase II promoter recognition methods the problem of recognition accuracy still remains [38,39,40].

Thus, our particular task is to develop new approach for promoter prediction based upon selected patterns of potential binding sites, that refer to the problem combinatorial transcription regulation [30,31].

The main goal of this research is to make gene functional annotation by using a set of integrated methods for the recognition of regulatory elements and transcription factors binding sites.

Analysis of a sequence has several stages:

(1) Making computer-assisted discovery of potential binding sites within the sequence of interest and marking their locations;

(2) Detecting the type of either regulatory or structural gene region (e.g., Promoter, 5' UTR, 3'UTR, coding sequence, enhancers) using predicted potential sites;

(3) Comparing predicted structural or functional regions with similar regions on related genes (using information accumulated in available databases);

(4) Providing functional annotation of the gene sequence.

Our approach is based on Data Mining methods for construction and description of specific oligonucleotide promoter patterns [9]. The suggested method uses significant patterns of promoter regions as the first step for functional annotation of genes. The system uses training samples of nucleotide sequences of promoter regions.

It is difficult to describe all eukaryotic promoter sequences by a common pattern due to variety of different transcription factor binding sites. To reduce such variety we to studied co-regulated sequences. However, even these functional sets could not reveal a single oligonucleotide pattern common for all sequences. The algorithm has a **flexibility** to search for structural patterns that are typical for a whole set of sequences as well as for a subset of sequences. An oligonucletide patterns could contain different numbers of oligonucleotides. To construct patterns to be tested various logical constrains are imposed. For instance, the

algorithms uses: (1) the position of the oligonucleotides relative to the transcription start, (2) mutual location of the oligonucleotides in the pattern and (3) orientation of the oligonucleotides in DNA double helix. In spite of complexity of construction of patterns, the same pattern could be observed in the negative sample of nucleotide sequences. Therefore, we should take into consideration probabilistic nature of such patterns. To meet this challenge the structural patterns were constructed in the first-order probabilistic logic.

3. "Gene Discovery": Technology of Data Mining

Software system "Gene Discovery" (GD) has been developed in C++ for analysis of structural organization of eukaryotic promoters. It uses information of context signals (experimentally confirmed and computer-predicted). This system is a version of the system "Discovery" [3,4,6] specifically adapted for analysis of genetic sequences. Friendly graphical user interface helps a user to work with this software. "Gene Discovery" consists of three main modules: (1) the module for on-line representation of context signals from DNA sequence in standard table form; (2) the module "Discovery" for searching regularities; (3) the module for recognition of the class of the sequence using the regularities found. Figure 1 shows the scheme of the "Gene Discovery" system. Data Mining module "Discovery" is represented there as a search block “Search for patterns of the joint presence and relative localization of contextual signals…”

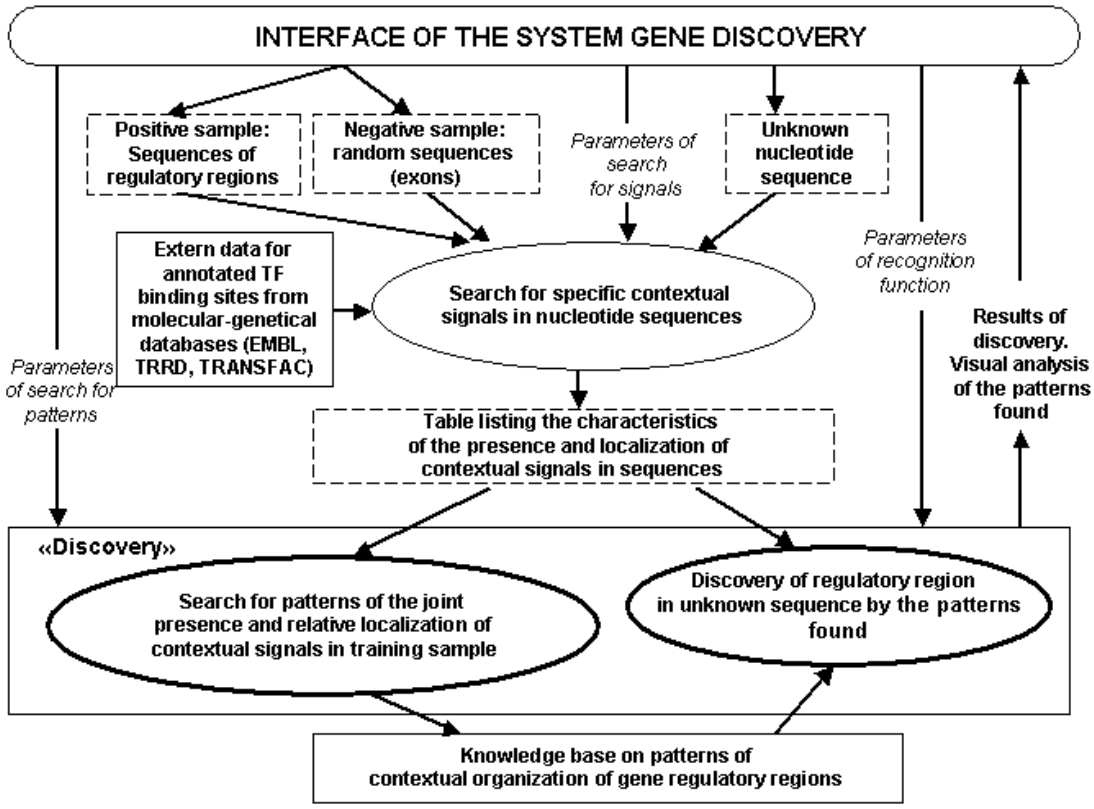

Figure 1. The scheme of the "Gene Discovery" system.

The discovery module is shown in Figure 1 as "Discovery of a regulatory region in unknown sequence by using patterns found". Other modules of the system serve for preparation and interpretation of the molecular-genetic data.

A machine learning method and system "Gene Discovery" [3,9], discover statistically significant first-order logic rules for functional annotation of regulatory regions. The technique allows one to obtain human-readable forecasting rules that are interpretable in biological language and also provides promoter recognition (functional annotation). A biologist may evaluate both the correctness of the recognition and that of the rules themselves.

Let us consider the particular example of oligonucleotide motif in 15-lettered alphabet is CWGNRGCN. It means C(A/T)G(A/T/G/C)(A/G)GC(A/T/G/C) in 4-lettered consensus record. This motif of length 8 bp is pre-selected as specific for the set of promoters by the program ARGO [21]. A complex rule uses several such motifs. Let us consider the example of the forecasting rule:

If CWGNRGCN<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& S_3…&S_k, where k>1. The program automatically defines the number of the signals in such a pattern [9,10].

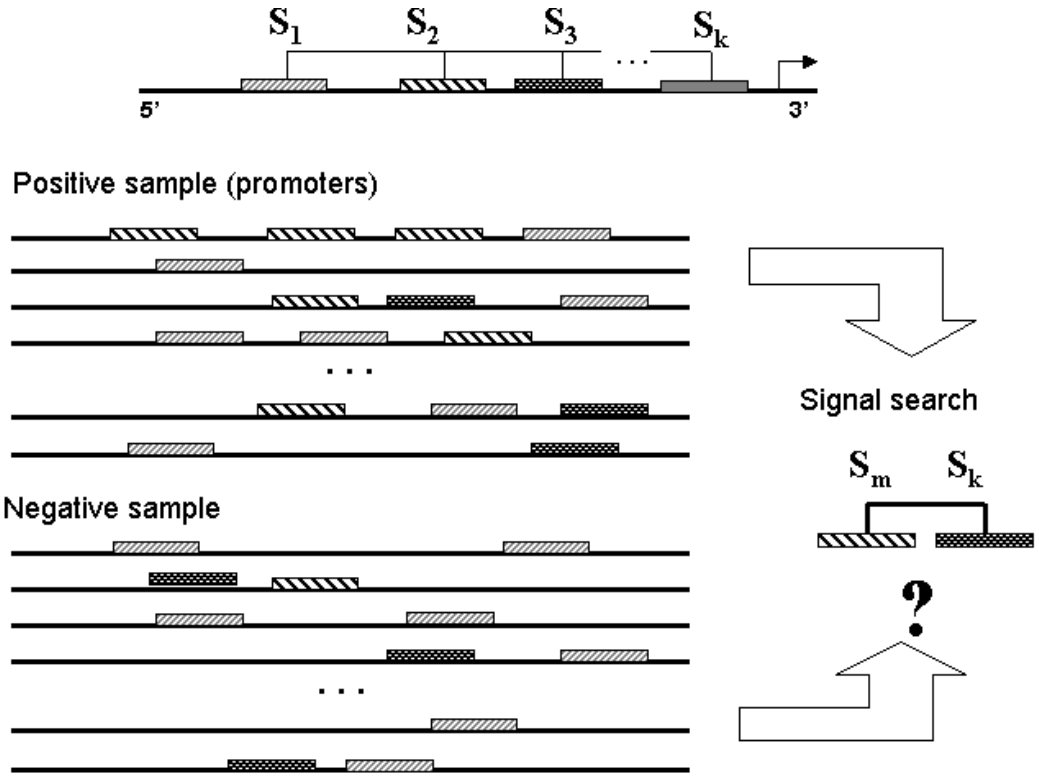

Figure 2. The problem of search of complex signals in promoters. An oligonucleotide pattern of S_1&S_2&S_3…&S_k located relative to the transcription start (denoted by asterisk) presented at the top of the picture. Distribution of oligonucleotides under analysis in sequences of positive and negative sample is given below.

4. Algorithm Background

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 [41] for $Y \Rightarrow X$ ("that allow us to deduce that Y is a probable cause of X 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 $\text{CondProb}(A_0|\text{SubFormular}(A_1\& \ldots\&A_k))$ of A_0 under any subcondition $\text{SubFormular}(A_1\&\ldots\&A_k)$ is strictly less than the conditional probability $\text{CondProb}(A_0|A_1\&\ldots\&A_k)$ of A_0 under full condition. In particular, for the expression $Y \Rightarrow X$, if we get $\text{SubFormular}(Y) = \varnothing$, then the conditional probability $\text{CondProb}(X|\text{SubFormular}(Y)) = \text{CondProb}(X|\varnothing) = \text{Probability}(X)$ must be strictly less then CondProb(X|Y) — the probability of X given Y. This gives us the following definition [4].

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

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

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

$\text{SubFormular}(A_1\& \ldots\&A_k)=A^s_1\& \ldots\&A^s_k$, $\{A^s_1, \ldots,A^s_k\} \subset$ (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 $\text{SubFormular}(A_1\& \ldots\&A_k) \Rightarrow A_0$ 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 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 $\text{Prob}(C) = \text{CondProb}(A_0|A_1\& \ldots\&A_k)$, assuming that $\text{Prob}(A_1\& \ldots\&A_k) > 0$. Similarly, conditional probabilities $\text{Prob}(A_0|A_{i1}\&...\&A_{ih})$ are defined for sub-rules $C_i = (A_{i1}\&...\&A_{ih} \Rightarrow A_0)$, assuming that $\text{Prob}(A_{i1}\&...\&A_{ih}) > 0$. 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 [7,35].

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 (Figure 3) $C_1, C_2, \ldots, C_{m-1}, C_m$ 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 $\text{Prob}(C_1) < \text{Prob}(C_2)$, … , $\text{Prob}(C_{m-1}) < \text{Prob}(C_m)$. There is a theorem [35] that all rules, which have a maximum value of conditional probability, can be found at the end of such chains.

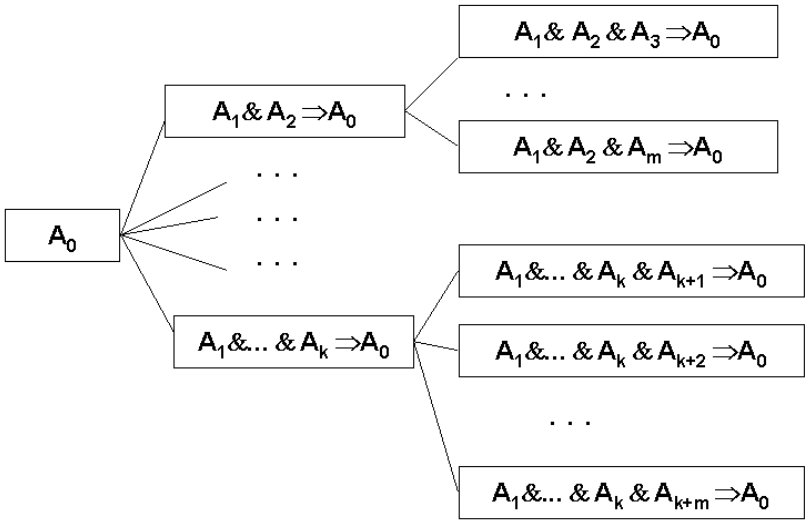

Figure 3. An example of the rule search for hypothesis A_0

5. Related Algorithms

Paper [42] presents probabilistic relational models (PRMs) that allow the properties of an object to depend probabilistically both on other properties of that object and on properties of related objects. This paper states that PRMs are (1) significantly more expressive than standard models and permit to use relational databases directly without "flattening" the relational data. "Flattening" removes richer structure of relational data. This lost structural information might be crucial in understanding the data. This algorithm is illustrated with a simple example from genetics.

Authors state that until now inductive logic programming (ILP) [43] has been the primary learning framework with capability to learn logical Horn rules for determining when some first-order predicate holds. "While ILP is an excellent solution in many settings, it may be inappropriate in others. The main limitation is **the deterministic nature** of the rules discovered. In many domains interesting correlations are far from being deterministic". The starting point if this paper is the structured representation of probabilistic models as an extension of the techniques of Bayesian network learning for first-order logic. Authors state that their probabilistic relational framework can be mapped into a variety of specific relational

systems, including the probabilistic logic programs of [44] and the probabilistic frame systems [45].

The goal of the paper [42] is, given a skeleton structure of the regularity (dependency), to define a probability distribution over all completions of the skeleton. The model consists of two components: the qualitative dependency structure, S., and the parameters associated with it, Θ_s. Parameters are learned using training data and Bayesian prior probabilities. Paper considers a conditional probability distribution (CPD) that specifies P(X.A | Pa(X.A)). This set of probabilities contains probabilities for each individual value of the attribute A on the data set X conditioned by values of the same attribute A on another entity Pa called a parent. An attribute for the parent is called a parent attribute. More generally, this paper investigates probabilistic relations y.b → x.a between attributes of entities y and x. Entities should satisfy some relation R(x,y), where y is called a parent of x. An assumed structure, S, determines the set of parents for each attribute, that is structure S is given by the relation R on the set of entities X. Each entity may have more than one parent. The learning task is to learn parameters of CPD for this structure. The specific parameter to be learned is a conditional probability P(X.A=v | Pa(x.A) =u), that is a probability that child's arbitrate A has value v conditioned that the same parent's attribute has value u. The maximum likelihood value for this parameter is a ratio of two counts, C_1/C_2. The first one, C_1 is a straight count of all (v,u) pairs (child, parent). The second one, C_2 is the count of the total number of all (v_t,u) pairs , where v_t can be any value of the attribute A in the data set X. Because this C_1/C_2 estimate may not be robust authors suggest using Bayesian prior probability distribution to smooth irregularities in training data. To learn structure S authors suggest to use a scoring function designed as a marginal likelihood function. This function uses the probability of a given value of parameter θ_s **c**onditioned by a structure S. Structures that provide highest values for θ_s are selected.

Authors note that the full problem of finding the best solution is NP-hard [46]. Thus, they use a heuristic search mechanism, which is similar to our heuristic search mechanism. This algorithm is greedy hill-climbing search. The algorithm maintains a current candidate structure and iteratively improves it. Local maxima are worked out by using random restart. Our approach is similar but we also test statistical significance of discovered structure S. This way we may refuse even high values if they are not statistically significant avoiding overfitting problem.

In [47] a data-mining approach is applied to spacer oligonucleotide typing of Mycobacterium tuberculosis. This is achieved by a DNA analysis technique (genotyping), called *sp*acer *oligo*nucleotide *typing* (*spoligotyping*). The paper discovers intelligible knowledge rules from spoligotyping. This processing was achieved by applying the C4.5 induction algorithm. Finally, a Prototype Selection (PS) procedure was applied to eliminate noisy data. Algorithm C4.5 constructs rules but that rules are actually is prepositional logic rules, which are less expressive than first-order logic rules employed in our works [7,9].

6. Complex Signals as Oligonucleotide Patterns

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 $Prob(C_1) < Prob(C_2)$, … , $Prob(C_{m-1}) < Prob(C_m)$. 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.

Promoters of co-regulated genes could be characterised by groups of oligonucleotide motifs. We use term "motifs" to underline degenerate consensus of such oligonucleotides. The problem is to study mutual presence and location of these motifs.

Below a **complex signal** is a group of oligonucleotide motifs that display a certain pattern of relative location in promoter sequences. The presence of such complex signal could be treated as the condition for A_0 to belong to the promoter class. Specifically we consider a group of two oligonucleotide motifs (S_1, S_2) as a complex signal specified as follows:

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

where S_1 and S_2 are oligonucleotides in the object-attribute 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.

Complex signal term (S_1, S_2) could incorporate complementary oligonucleotides

$$(S_1, S_2) = (\text{Position}(S_1) < \text{Position}(S_2)) \,\&\, (\text{Sign}(S_1) = z_1) \,\&\, (\text{Sign}(S_2) = z_2)),$$

where S_1 and S_2 are oligonucleotides in the object-attribute table; Position(S_1) and Position(S_2) are positions of these oligonucleotides in a sequence relative to the transcription start. Sign(S_1) and Sign(S_2) mean strand in double-helix DNA, where the signals were located in; $z_1, z_2 \in \{+,-\}$, $z_1, z_2 \in \{+,-\}$ sign (+) means direct strain, i.e. from 5'- to 3'-end, (-) means inverted strain.

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 $(S_1,S_2,S_3)=(\text{Position}(S_1)<\text{Position}(S_2)<\text{Position}(S_3))$. 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.

More complex forecasting rules are testing by addition of new signals to condition (S_1,… S_{i-1},S_i), i=1,2,... . System "Gene Discovery" sorts out all variants of longer forecasting rule (S_1,… S_{i-1},S_i, S_j) to strengthen prediction, i,j=1,...N, N -number of motifs in the object-attribute table.

The Fisher statistical criterion (exact Fisher test for contingency tables) is used in this algorithm for testing statistical significance [48]. Statistical significance $P(N_1,N_2,N_3,N_4)$ used by the program to select better rules.

	Class 1	Class 2
Condition of the rule is satisfied	N_1	N_2
Condition of the rule is not satisfied	N_3	N_4
Number of object in the class	N_1+N_3	N_2+N_4

Here N_1, N_2 - number of promoters (Class 1) and non-promoters (Class 2) correspondingly, that contain the complex signal tested; N_3, N_4 - number of promoters and non-promoters correspondingly, not containing the complex signal.

Conditional probability $PC(N_1,N_2)= N_1/(N_1+N_2)$ to predict promoters correctly by the rule also is estimated by the program. We add a new motif S_j to the forecasting rule if conditional probability $PC(N_1,N_2)$ increases and significance value $P(N_1,N_2,N_3,N_4)$ is not exceed the threshold level 0.05.

Theoretical advantages of this generalisation are presented in [3,35]. This approach has some similarity with the hint approach [49]. We use mathematical formalisms of first-

order logic rules described in [50,51,52]. Note that a class of general propositional and first-order logic rules covered by such methods is wider than a class of decision trees [8].

7. Data Preparation and Signal Pre-Selection

The teaching sample of nucleotide sequences of two alternative classes is used as input to "Gene Discovery" 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:

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

(2) Site with conformational or physical-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 show the approach for 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 [14] and divided into several groups according to 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 the 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 [21] (see also http://wwwmgs.bionet.nsc.ru/mgs/programs/argo/). 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 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. The testing hypothesis A_0 was: "Does the sequence belong to class 1 (promoters)?"

The program can use any sequence set in FASTA format as input. A functional sample could be extracted from TRRD [14], TRANSFAC [15], EpoDB [53].

Similarly, other functional sets of promoters extracted from the TRRD database were analysed, 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.

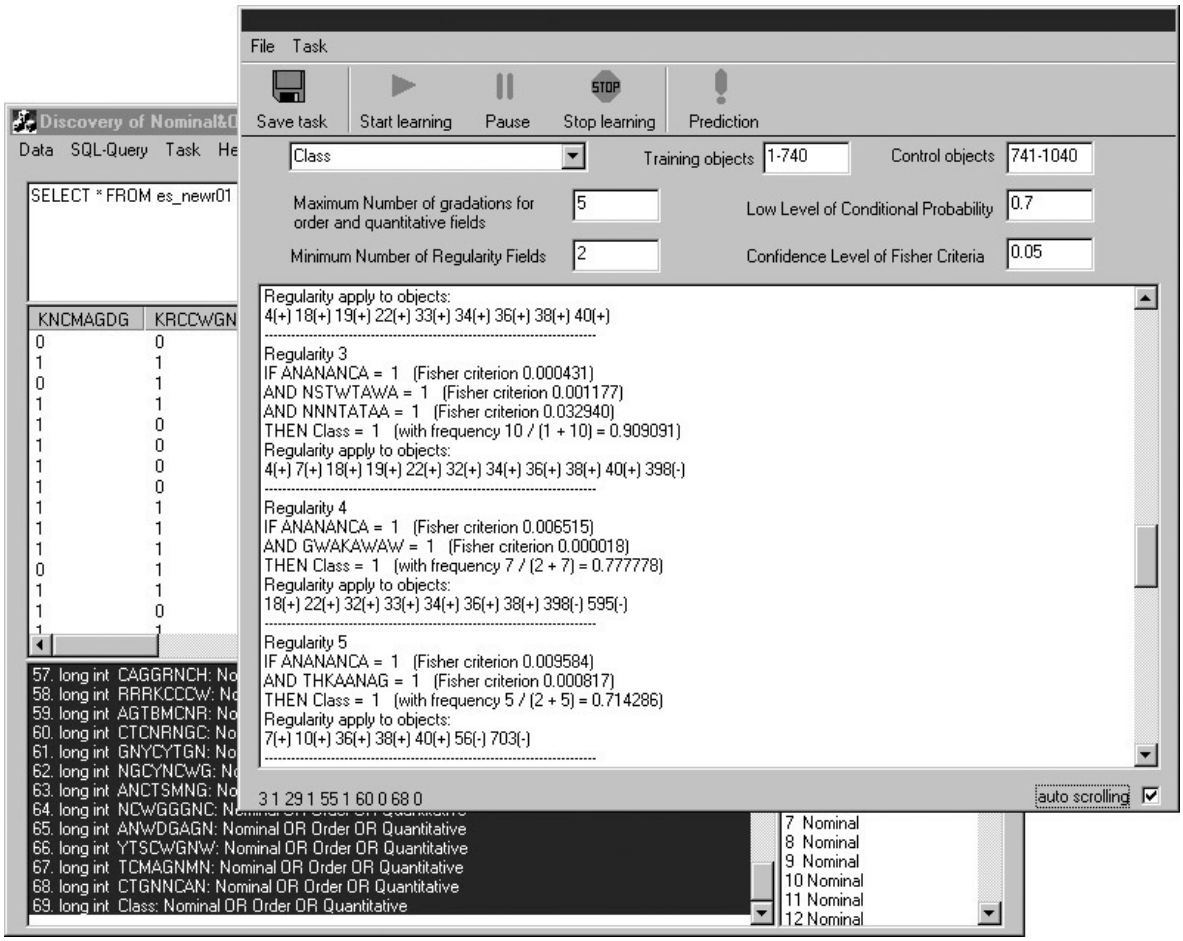

Figure 4. User interface of "Gene Discovery" software. An example of search of regularities for the sample of endocrine gene promoters is shown. Regularities have form of IF-THEN hypotheses. Condition "IF ANANANCA = 1 AND GWAKAWAW = 1" means that oligonucleotides ANANANCA and GWAKAWAW must be present in sequence under analysis. Conclusion "THEN Class = 1" means that the sequence belongs to the class of endocrine gene promoters. Such regularities in plain text format could be used for further recognition for a test sample.

An example of testing hypothesis of oligonucleotide pattern without fixed location of the oligonucleotides in the pattern is given in Figure 4.

8. Analysis of Complex Signals Found

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.

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) 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*.

Expected number N* was estimated as product of oligonucleotides frequencies in promoter multiplied by total number of promoters and divided by the number of variants of mutual location. For example, expected number of promoters N* possessing the complex signal (S_1,S_2,S_3|Pos(S_1)<Pos(S_2)<Pos(S_3)) is equal

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

where N* - expected number of promoters N* possessing the oligonucleotides S_1,S_2,S_3 ; $P(S_1)$, $P(S_2)$, $P(S_3)$ – frequencies of oligonulceotides S_1,S_2 and S_3, correspondingly; M - total number of promoters in the sample analysed; 6=3! - number of possible variants of mutual location of three oligonucleotides in the sequence.

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.

The expected by random number N* for this signal equals to 0.47 (i.e. less than 1). But the signal is present in 6 promoters; this 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 5. 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 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 physical-chemical properties of double-stranded DNA.

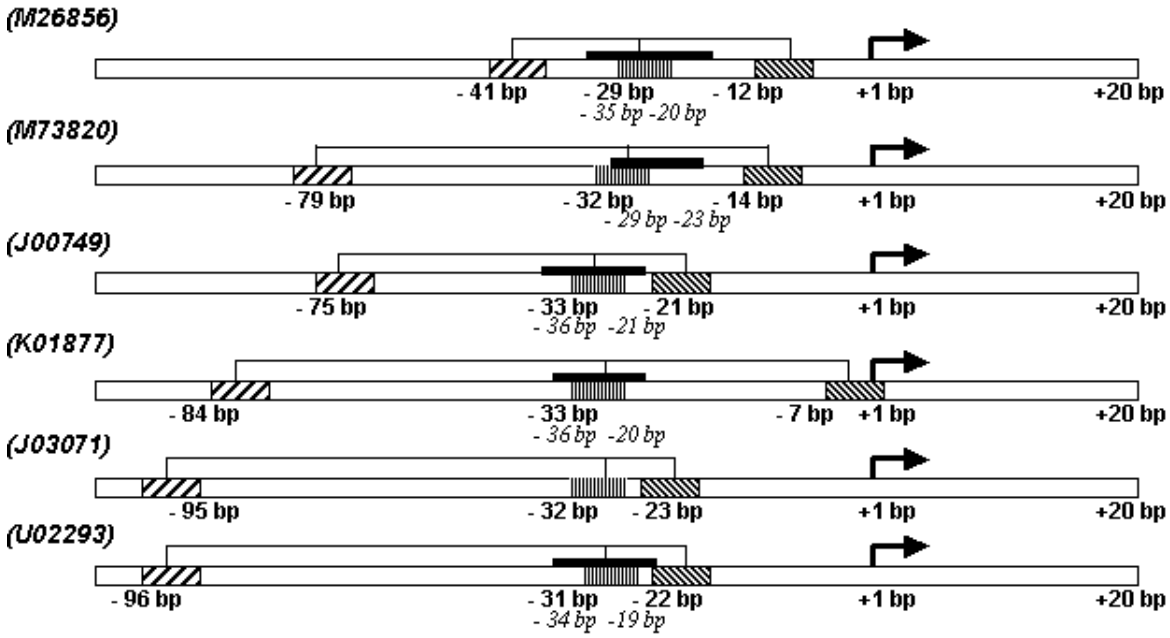

Figure 5. Schematic localization of the complex signal CWGNRGCN<NGSYMTAM<MAGKSHCN in promoters of endocrine system genes. The promoter sequences are aligned relative to the transcription start (position +1 bp), indicated by arrows. The EMBL identifiers of the promoters studied are given in parentheses to the left. The eight-bp oligonucleotide motifs composing the complex signal are shown as dashed rectangles; positions of the first nucleotides are indicated relative to the transcription start. The black rectangles mark experimentally defined positions of the TATA-box indicated in the TRRD database. Positions of its first and last nucleotides are italicised.

The signal presented in Figure 5 found in 6 promoters (EMBL ID: M26856, M73820, U02293, J00749, J03071, K01877, correspondingly). This complex signal is located in the region from -95 bp up to +7 bp relative to the transcription start. In Figure 5 position of each oligonucleotide is indicated as position of it's first nucleotide .The TATA-box location from the TRRD database is indicated by black rectangles. One can see coincidence of the second oligonucleotide motif with TATA-box region. It is shown similar distances between first and second and between second and third oligonucleotides.

Figure 6 gives an example of localization of the complex signal DNMYTTSA<DNYAADGG<RCAGMMDY in promoter sequences of erythroid-specific genes. In this case also one can see characteristic distance between oligonucleotides in th complex signal.

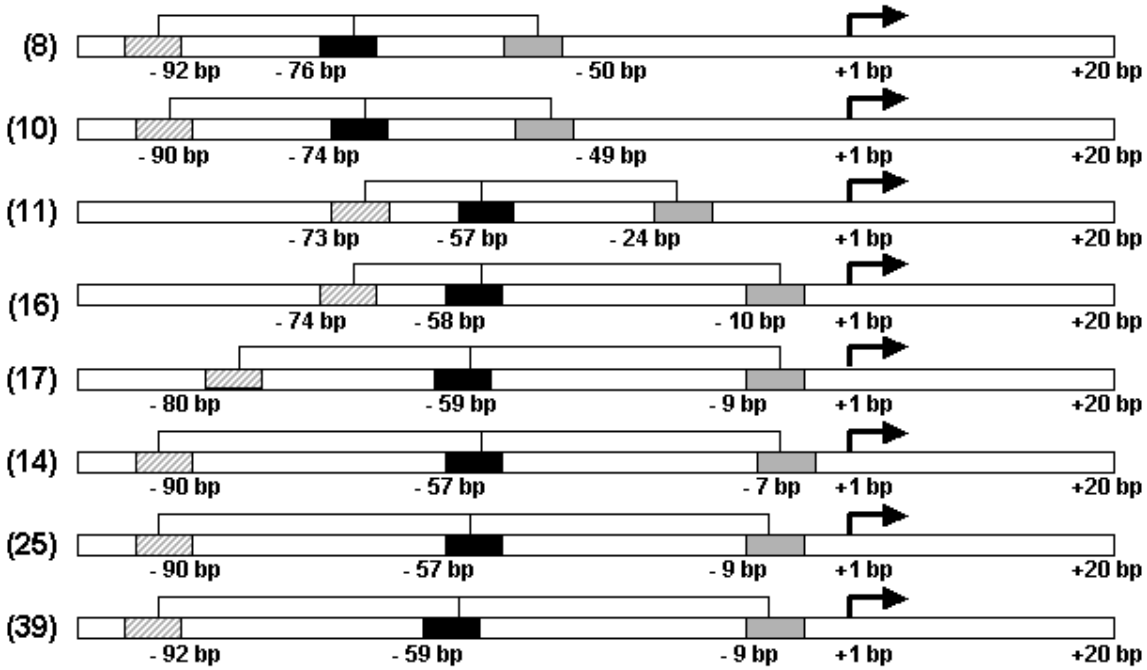

Figure 6. An example of localization of the complex signal DNMYTTSA<DNYAADGG<RCAGMMDY in eight promoter sequences of erythroid-specific genes. The promoter sequences are aligned relative to the transcription start (position +1 bp), indicated with arrows. Identifiers of the promoters studied are given in parentheses to the left. The eight-bp oligonucleotide motifs composing the complex signal are shown as black rectangles; positions of the first nucleotides are indicated relative to the transcription start.

Similar technique was implemented for donor splice sites of genes of primates. It contained 2343 sites, each contained positions from -11 to +10 relatively exon-intron junction. Separate nucleotide bases were used as context signals in the site sequence. 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.

Table 2 contains the examples of signals found. Complex signals presented as nucleotides. Sign "<" denotes relation between the positions of corresponding nucleotides relative to the exon-intron junction.

Table 2. The examples of the complex signals for donor splice sites

№	Complex signal (regularity)	Length of complex signal	Significance	Number of sites possessing the signal*
1	a<t	2	7.221685e-003	6011
2	a<g	2	4.549541e-002	7469
3	t<c<c<c<a	5	2.242927e-002	2467
4	c<a<c<a<t<t	6	1.886203e-002	770
5	c<c<a<c<a<a	6	2.004277e-002	726
6	t<c<c<a<c<a	6	1.602915e-002	902
7	g<c<c<a<c<a	6	1.644068e-002	880
8	g<c<a<c<a<g	6	2.211978e-002	696
9	a<c<a<c<a<t<t	7	2.358411e-002	304
-	...		...	...
1918	c<g<c<a<c<a<a	7	2.196624e-002	331

Note: * A signal (especially short one) could be presented in splice site sequence more than once.

Figure 7 presents location of signal g<c<a<c<a<g (№8 in Table 2) for a splice site.

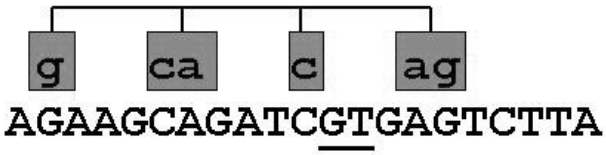

Figure 7. Example of regularity g<c<a<c<a<g applied for a splice site.

9. Discovery Technique

Discovery rule base on complex signals found. For oligonucleotides signals this recognition rule described in the paper [54].

In this paper the score of some object positions estimated based on the score of all signals applied to this position. This score means the probability of appearance these signals on the random sequence. Using negative random samples we can estimate the level of the score, which guarantee some levels of the first and second kind errors. If in some control sequence the score is greater then these levels then we predict that this 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 some control sequence. As a result we have a complex signals sequence $0 < N < ... < N_{total}$, where N_{total} – total number of signals. The order of the signals means the order of the applications of the complex signals to this sequence. Then similar score $P(S)$ may be calculated for each position of the sequence. Recognition procedure is similar: if the score in some place of the sequence is greater then estimated levels then this sequence is from some functional class.

Let us estimate probability $P(S)$ to obtain sequence $S=X_1X_2...X_n$ as product of probabilities X_i. Here X_i could correspond to one, two, three or even all possible nucleotides, $i=1,2,...,n$.

$$P(S)=\prod_{i=1}^{n} P(X_i)$$

If the sequence S is more degenerate (for example, nucleotide of any of four types could be presented in position i, and $P(X_i)=1$), then probability $P(S)$ closer to 1. If sequence S is less degenerate (for example, only one type of nucleotides could be presented in position i, $P(X_i)\approx 0.25$), then probability $P(S)$ closer to 0 (i.e. close to $1/4^n$). In last case we have more similarity to sequences from training set and have a class of more reliable predictions. To overcome the problem of computer presentation of small values we use logarithmic scale. Instead of product $P(X_i)$ we use the sum of logarithms $log\mathrm{P}(\mathrm{X_i})$ for estimation $P(S)$. But probabilities $P(X_i)$ could be context-dependent. In this case we could use all nucleotide sequences from the class "No" to estimate probability $P(S)$ and define some threshold for recognition function.

Prediction function base on some consensus sequence S. This sequence defines relatively short functionally important region in nucleotide sequences from the training set and is character only for some subset of the training set.

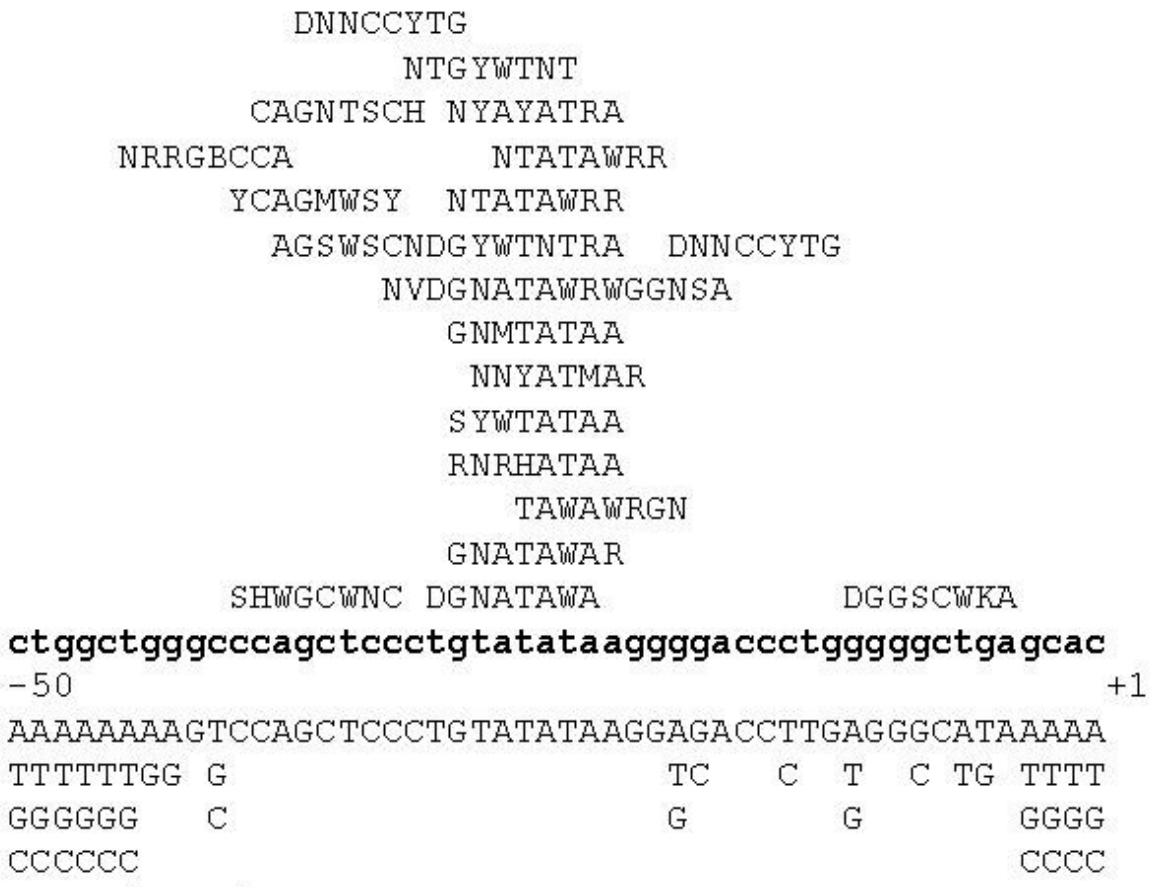

Figure 8. Location of signals (8-bp oligonucleotides) in the sequence under analysis (in lower case). Positions are indicated relative to the transcription start. Consensus sequence is given below.

Recognition procedure based on regularities (complex signals) is similar to procedure described above. We define recognition function for the sequence under analysis based on regularities found in training sample.

Weight of objects (sequences) could be defined in several ways for object under analysis:

(1) N, total number of different signals found in the object;

(2) $\sum log\ P(S)$, sum of logarithms probabilities to obtain signals found in the object;

(3) N_r, number of signals found in the object and participating in regularities (complex signals, oligonucleotide patterns) that were correct for this object;

(4) $\sum log\ P(S_r)$, sum of logarithms probabilities to obtain signals found in the object. Only signals participating in regularities are used.

Four more weights can be achieved when to consider signals found in testing object and in most close object in the training set (nearest neighbour method). Among defined here four weights first two do not use regularities found in training set and other two weights do. So, for each task there is a possibility to estimate contribution of regularity using in comparison with simple taking into account number and probabilities of found signals.

Based on these ways of sequence weight estimation we have developed the program for donor splice sites prediction in user-defined nucleotide sequences.

Given weights for all objects in training and control sets first and second type errors can be estimated for training and test sets respectively. For donor splice sites first and second types errors for test data were 4,4% and 4,0% respectively.

10. Discussion

The developed system "Gene Discovery" helps us to find complex signals in promoter regions. In a similar way any samples of phased nucleotide sequences could be analysed. The functional meaning of the signal could be treated in terms of the transcription factors binding sites or the conformational properties of DNA [31,40].

It should be noted that the system does not over-trained on the training samples and shows the false positive rate on test samples similar to the conditional probability of a selected rule.

Automatic rule generation for functional annotation of genes could use also other approaches of Data Mining [2]. We plan to combine other table data of gene features connected not only to context signals from regulatory regions for prediction of gene function class.

Analysis shows high number of complex signals for promoters of endocrine gene system and promoter sequences of erythroid-specific genes. Functional meaning of complex signals is confirmed by similar location of oligonucleotides motifs relative to the transcription start and similar distances between these mot ifs. In addition promoters analysed have no significant sequence likelihood.

Oligonucleotide motifs could correspond to transcription factors binding sites (TFBS). It was shown that promoters region rich in potential TFBS relative to random sequences. Oligonucleotide motifs could also correspond to promoter regions with specific conformational or physical-chemical properties, such as high flexibility, low melting temperature etc. It was shown also specific patterns of potential TFBS with different maximums of sites location for different promoter parts. Thus, complex signals could affect preferable location TFBS in promoters. Importance of context features location in promoters was shown by M.Zhang [36]. Promoter dissection by parts with similar dinucleotide frequencies was revealed in [55]. It was shown that such promoter parts have similar conformational and physical-chemical properties. We suggest that complex signals in promoter could have both context and conformational origin reflecting specific promoter features.

In addition, special type of regulatory elements studied, so-called composite elements [56; http://compel.bionet.nsc.ru/]. Composite element formed by pair of transcription factor binding sites that acquire new regulatory properties due to protein-protein interaction. Such interaction provides more expressed effect on transcription. Analysis of regularities found by "Gene Discovery" system suggests new approach for computer discovery of composite elements.

Published experimental data and specialized molecular-biological databases contain a large number of experimental results for DNA sequences involved in transcription regulation. Currently, more than 300 molecular-biological databases are available on the Internet and this number continues to grow [57]. This provides great opportunity for large-scale data mining and knowledge discovery for bioinformatics [1,58].

Our approach was applied mainly to gene regulatory region analysis. We suggest a new way for promoter prediction for specific co-regulated gene groups. Now we analyse context gene structure for all levels of gene hierarchy: promoter, regulatory regions, transcription factor binding sites and its modules, 5'UTR, splice sties. The program "Gene Discovery" is developed as a component of Internet navigation system GeneExpress for resources and databases for gene transcription regulation [59; http://wwwmgs/bionet.nsc.ru/gnw/].

Acknowledgements

The authors are grateful to A.S. Belenok for help in programming, to G.V. Orlova for help in preparation of the manuscript, and to I.B. Rogozin for providing splice site data. Work was supported in part by the Russian Foundation for Basic Research (No. 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90355, 02-07-90359, 00-04-49229, 00-04-49255), Russian Ministry of Industry, Sciences and Technologies (№ 43.073.1.1.1501), Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65). Y.O. was supported by INTAS (YSF 00-178).

References

[1] R.D. King *et al*., The utility of different representations of protein sequence for predicting functional class, *Bioinformatics* **17** (2001) 445-454.
[2] E. Kretschmann *et al*., Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT, *Bioinformatics* **17** (2001) 920-926.
[3] E.E. Vityaev and A.A. Moskvitin, Introduction to discovery theory: Discovery software system, *Computational Systems (*Novosibirsk) **148** (1993) 117-163 (in Russian).
[4] B. Kovalerchuk and E. Vityaev, 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.
[5] B. Kovalerchuk *et al*., Design of consistent system for radiologists to support breast cancer diagnosis. In: *Proc Joint Conf Information Sciences*, Durham, NC, **2** (1997) 118-121.
[6] B. Kovalerchuk *et al*., Consistent Knowledge Discovery in Medical Diagnosis, *IEEE Engineering in Medicine and Biology Magazine* (Special issue: "Medical Data Mining", July/August) 2000, pp. 26-37.
[7] B. Kovalerchuk *et al*., Consistent and Complete Data and "Expert" Mining in Medicine. In: Medical Data Mining and Knowledge Discovery (Book chapter). Springer. 2001, pp. 238-280.
[8] T. Mitchell, Machine Learning. New York: McGraw Hill, 1997.
[9] E.E. Vityaev *et al*., Computer system "Gene Discovery" for regularities retrieving in eukaryotic regulatory sequences organisation, *Mol.Biologia (Mosk).* **35** (2001) 952-960 (in Russian).
[10] E.E. Vityaev *et al*., Computer system "Gene Discovery" for promoter structure analysis, *In Silico Biol.* **2** (2002) 0024 <http://www.bioinfo.de/isb/2002/02/0024/>
[11] R.C. Hardison, Conserved non-coding sequences are reliable guides to regulatory elements, *Trends Genet.* **16** (2000) 369-372.
[12] J.A. Goodrich *et al*., Contacts in context: promoter specificity and macromolecular interactions in transcription, *Cell* **84**, (1996) 825-830.
[13] J.W. Fickett and A.G. Hatzigeorgiou, Eukaryotic promoter recognition, *Genome Res.* **7** (1997) 861-878.
[14] N.A. Kolchanov *et al*., Transcription Regulatory Regions Databases (TRRD): its status in 2002, *Nucleic Acids Res*. **30** (2002) 312-317.
[15] E. Wingender *et al*., The TRANSFAC system on gene expression regulation. Nucleic Acids Res. **29** (2001) 281-3.
[16] K. Quandt *et al*., MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data, *Nucleic Acids Res.* **23** (1995) 4878-4884.
[17] T. Werner, Computer-assisted analysis of transcription control regions. Matinspector and other programs, *Methods Mol.Biol.* 132 (2000) 337-349.

[18] T. Werner, Models for prediction and recognition of eukaryotic promoters, *Mamm. Genome* **10** (1999) 168-175.
[19] D.S. Prestridge, Computer software for eukaryotic promoter analysis. *Methods Mol. Biol*. **130** (2000) 265-295.
[20] M.A. Poznyakov *et al*., Comparative Analysis of Methods Recognizing Potential Transcription Factor Binding Sites, *Mol.Biologia (Mosk)*. **35** (2001) 961-978 (in Russian).
[21] V.N. Babenko *et al*., Investigating extended regulatory regions of genomic DNA sequences, *Bioinformatics* **15** (1999) 644-653.
[22] T. Dandekar and K. Sharma, Regulatory RNA, Springer Verlag, Heidelberg. 1998.
[23] B. Demeler and G. Zhou, Neural network optimization for E. coli promoter prediction. *Nucl. Acids Res.* **19**, (1991) 1593-1599.
[24] A.G. Pedersen and J. Engelbrecht, Investigations of Escherichia coli promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional startpoint. *Intelligent Systems Mol. Biol.* **3**, (1995) 292-299.
[25] H. Ogura *et al*., A study of learning splice sites of DNA sequence by neural networks, *Comput. Biol. Med.* **27** (1997) 67-75.
[26] E.N. Trifonov, Interfering contexts of regulatory sequence elements. *Comp. Appl. Biosci.* **12** (1996) 423-429.
[27] S. Tiwari *et al*., Prediction of probable genes by Fourier analysis of genomic sequences *Comp. Appl. Biosci.* **13** (1997) 263-270.
[28] J. van Helden *et al*., Extracting regulatory sites from the upstream reion of yeast genes by computational analysis of oligonucleotide frequencies, *J.Mol.Biol.* **281** (1998) 827-842.
[29] A. Wagner, Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes, *Bioinformatics* **15** (1999) 776-784.
[30] A. Kel *et al*., ClusterScan: A Tool for Automatic Annotation of Genomic Regulatory Sequences by Searching for Composite Clusters. In: E. Wingender, R. Hofestaedt and I. Liebich (eds.): *Proceedings of the German Conference on Bioinformatics GCB'2001*, October 7-10, Braunschweig, 2001, pp. 96-101.
[31] A. Klingenhoff *et al*., Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity, *Bioinformatics* **15** (1999) 180-186.
[32] G. Thijs *et al.*, A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling, *Bioinformatics* **17** (2001) 1113-1122.
[33] Y.-J. Hu, Biological Sequence Data Mining. In: L. De Raedt and A. Siebes (eds.): *PKDD 2001*, LNAI 2168, ISBN 3-540-42534-9 Springer-Verlag Berlin Heidelberg. 2001, pp. 228-240.
[34] J.-T. Horng *et al*., Mining Putative Regulatory Elements in Gene Promoter Regions. In: E. Wingender, R. Hofestaedt and I. Liebich (eds.): *Proceedings of the German Conference on Bioinformatics GCB'2001*, October 7-10, Braunschweig, 2001, pp. 90-95.
[35] E.E. Vityaev, Semantic approach to knowledge base development: Semantic probabilistic inference, *Computer Systems,* Novosibirsk **146** (1992) 19-49 (in Russian).
[36] M.Q. Zhang, Identification of human gene core-promoters in silico, *Genome Res.* **8** (1998) 319-326.
[37] D.B. Nikolov and S.K. Burley, RNA Polymerase II transcription initiation: A structural view, *Proc. Natl. Acad. Sci. USA* **94** (1997) 15-22.
[38] A.G. Pedersen *et al*., The biology of eukaryotic promoter prediction - a review, *Comput. Chem.* **23** (1999) 191-207.
[39] V. Solovyev. and A. Salamov, 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, pp.294-302.
[40] Y.V. Kondrakhin *et al*., Eukaryotic promoter recognition by binding sites for transcription factors, *Comput Appl Biosci*. **11** (1995) 477-488.

[41] P. Suppes, A probabilistic Theory of Causality, North-Holland, Amsterdam, 1970.
[42] N. Friedman *et al*., Learning Probabilistic Relational Models. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), pp.1300-1307 (http://robotics.stanford.edu/~koller/papers/ijcai99lprm.ps).
[43] N. Lavrak and S. Dzeroski, Inductive Logic Programming: Techniques and Applications. Ellis Hodwood, 1994.
[44] L. Ngo and P. Haddawy, Answering queries from context-sensitive probabilistic knowledge bases. Theoretical Computer Science, 1996.
[45] D. Koller and A. Pfeffer., Probabilistic frame-based systems. In: *Proc. AAAI*, 1998.
[46] D.M. Chickering, Learning Bayesian networks in NP-complete. In: D. Fisher and H-J. Lehz, eds., Learning from Data: Artificial intelligence and Statistics, Springer Verlag, 1996.
[47] M. Sebban, I. Mokrousov, N. Rastogi and C. Sola, A data-mining approach to spacer oligonucleotide typing of Mycobacterium tuberculosis, *Bioinformatics* **18** (2002) 235-243.
[48] M.G. Kendall and A. Stuart, The advanced theory of statistics, 4th ed., Charles Griffin & Co LTD, London. 1977.
[49] Y.S. Abu-Mostafa, Learning from hints in neural networks, *J Complexity* **6** (1990) 192-198.
[50] S. Russel and P. Norvig, Artificial Intelligence. A Modern Approach. Englewood Cliffs, NJ: Prentice Hall, 1995.
[51] J.Y. Halpern, An analysis of first-order logic of probability, *Artificial Intelligence* **46** (1990) 311-350.
[52] D.H. Krantz *et al*., Foundations of measurement, Vol. 1,2,3 - NY, London: Acad. press, 1971, 577 p.; 1989, 493 p.; 1990, 356 p.
[53] C.J. Stoeckert *et al*., EpoDB: a prototype database for the analysis of genes expressed during vertebrate erythropoiesis, *Nucl. Acids Res.* **27** (1999) 200-203.
[54] O.V. Vishnevsky and E.E. Vityaev, Analysis and recognition of promoters of the erythroid-specific genes on the basis of degenerated oligonucleotide motifs, *Mol. Biologia (Mosk)*. **35** (2001) 979-986 (in Russian).
[55] V.G. Levitsky and A.V. Katokhin, Computer analysis and recognition of *Drosophila melanogaster* gene promoters *Mol. Biologia (Mosk)*. **35** (2001) 970-978 (in Russian).
[56] O.V. Kel-Margoulis *et al*., COMPEL: a database on composite regulatory elements providing combinatorial transcription regulation, *Nucleic Acids Res.* **28** (2000) 311.
[57] A.D. Baxevanis, The Molecular Biology Database Collection: 2002 update. *Nucleic Acids Res.*, **30** (2002) 1-12.
[58] I.B. Jakobsen *et al*., TreeGeneBrowser: phylogenetic data mining of gene sequences from public databases, *Bioinformatics* **17** (2001) 535-540.
[59] N.A. Kolchanov *et al*., GeneExpress: a computer system for description, analysis, and recognition of regulatory sequences of the eukaryotic genome. *ISMB-98* **6** (1998) 95-104.