ExpertDiscovery system application for the hierarchical analysis of eukaryotic transcription regulatory regions based on DNA codes of transcription

***Khomicheva I.V.*[1,2*], *Vityaev E.E.*[2], *Ananko E.A.*[1], *Shipilov T.I.*[3], *Levitsky V.G.*[1,3]**

[1]Institute of Cytology and Genetics SB RAS, Lavrentyev aven., 10, 630090 Novosibirsk, Russia
E-mail: {khomicheva, ananko, levitsky}@bionet.nsc.ru
[2]Sobolev Institute of Mathematics, Koptyug aven. 4, 630090 Novosibirsk, Russia
E-mail: vityaev@math.nsc.ru
[3]Novosibirsk State University, Pirogova, 2, 630090, Novosibirsk, Russia
E-mail: tshipilov@gmail.com
[*]corresponding author

Abstract.
We developed Relational Data Mining approach which allows to overcome essential limitations of the Data Mining and Knowledge Discovery techniques. In the paper the approach was implemented to adapt the original 'Discovery' system to the computational biology needs. The objects under consideration, eukaryotic transcription regulatory regions, are characterized by the great variety of context physicochemical and conformational DNA features. The currently available tools aimed at the regulatory regions analysis are sensitive to specific DNA features; therefore they produce poor results on complex heterogeneous data. Development of a method integrating the results of different recognition programs is a challenging task. We have developed the 'ExpertDiscovery' system, which discovers the hierarchically complicating set of complex signals based on different elementary signals. It provides a powerful tool to construct a model of regulatory region generalizing the results of different programs. Besides, the system is an independent tool for analysis. In the paper we demonstrate that 'ExpertDiscovery' outperforms the position weight matrix in the case when the elementary signals introduced to the system are nucleotides at specific positions. The system is able to discover biologically significant, simple to complex models of potential transcription factor binding sites for regulatory regions of interferon-inducible genes.

Keywords: *hierarchical complex signals*, *transcription regulatory regions*, *discovery.*

1. Introduction

The 'ExpertDiscovery' system [36] discovers complex signals, recursively defined in the first order logic based on elementary signals revealed by other programs. The main advantage of the 'ExpertDiscovery' system is that it provides a powerful tool to construct models of regulatory regions integrating different DNA characteristics. Generally, the program appears to be an adaptation of the Relational Data Mining approach and the original system 'Discovery' [14] to knowledge acquisition tasks on DNA sequences.

Eukaryotic 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 [8]. Regulatory regions are characterized by complex modular hierarchical structure and as the first level of organization possess transcription factor (TF) binding sites (TFBSs). A pair of neighboring TFBSs organizes the so called composite element and in that case their joint action appears to be synergetic or antagonistic, different from if they act independently [11]. The up next level of organization consists of composite elements, promoters, silencers and enhancers. The block-like organization of 5'-regulatory regions means the existence of alternative promoters generally located at a considerable distance from each other. The block-hierarchical structure of regulatory regions provides flexible regulation on the level of transcription by switching separate elements.

Thus, each level of organization sets its own task to investigators. The first of them is TFBSs prediction, methodologically difficult by itself due to the high variety of DNA binding proteins and the tissue- and stage-specific mechanisms of regulations. The up next task is the discovery of TFBSs pattern, in other words, the task of a localization of a promoter belonging to a certain functional class according to its transcription regulation specificity [26]. The highest hierarchical level corresponds to the system of integral regulation of transcription defined by the superposition of different DNA codes [31].

The tasks of regulatory region analysis are complicated and challenging the Knowledge Discovery in Databases and Data Mining (KDD&DM) approaches. KDD&DM algorithms applicable to bioinformatics tasks are: decision trees, neural network, Hidden Markov Models (HMM), genetic algorithms, etc. [30]. The traditional approach to predicting TFBSs is the position weight matrix (PWM), indeed a very powerful tool, but still possessing obvious drawbacks and limitations [29]. Despite the evident importance of noncoding sequences in gene regulation, our ability to describe and properly localize them is extremely limited. These known approaches turn out to be rather restricted, confronting the lack of sufficient training data and degenerate nature of biological objects under analysis.

To overcome the above drawbacks, we propose that the Relational Data Mining approach be applied and the 'Discovery' system be adapted to the tasks of regulatory region analysis. For this purpose, we developed the special 'ExpertDiscovery' system. From the biological viewpoint, it allows to integrate the great variety of context, physicochemical and conformational DNA characteristics, which could be cached by different bioinformatics tools, and thus to overcome in part the shortage of training data. The Relational Data Mining approach allows extending the class of hypotheses to be tested.

The rest of the paper is organized as follows: Section 2 covers the state of the art of the complicated problem under consideration. Section 3 describes the 'Discovery' system based on the Relational Data Mining approach. Section 4 introduces the 'ExpertDiscovery' system adapted for bioinformatics tasks. Section 5 presents the experimental results, obvious examples of the 'ExpertDiscovery' application to real biological data. We cite two cases, in the first one the elementary signals to build the complex are nucleotides at certain positions, and in the other, they are potential TFBSs.

2. State of art.

Analysis of gene transcription regulatory regions is of great importance for understanding molecular mechanisms of transcription. The complex and challenging task

assumes the number of subgoals to be determined and different biological contexts to be considered.

First of all, it is the task of TFBSs prediction, methodologically difficult by itself due to the high variety of DNA binding proteins and the tissue- and stage-specific mechanisms of regulations resulting in the degenerate nature of TFBSs. These sequences vary in length, position, redundancy, orientation in the DNA chain, and bases. As a direct consequence, the problem of the large number of false-positives necessarily takes place manifesting itself in the poor predictive performance of the corresponding software.

The traditional approach to prediction of TFBSs is the position weight matrix, indeed a very powerful tool, but still possessing the obvious drawbacks and limitations. [29]. PWM and consensus-based methods involve explicit assumption concerning the independent contribution of each nucleotide position to the binding affinity, producing the cumulative effect to the binding strength. In PWM, elements simply correspond to the probabilities of observing each nucleotide at each position. Numerous works [4, 20, 3, 32] indicate that the nucleotides of TFBSs cannot be treated independently. This assumption is invalid and contradicts the processes underlying the biological model. Predictions can be further strengthened by taking into account the sequence context in which a predicted site is found.

TFBSs clusters controlling eukaryotic gene expression form cis-regulatory modules. The transcription rate of a gene is guided by the artful interplay between DNA-binding TFs. The complexity of regulation is determined by the variety of possible states of regulatory regions when one or other subset of elements is switched on (a particular TF binds to the corresponding site) and by the diversity of combinations of interacting TFs [5]. The diversity of cis-regulatory regions is again the main challenge to recognition software.

Cis-regulatory module discovery methods include statistical approaches, in particular using dynamic programming to find the most significant clusters of putative binding sites of user-defined TFs [10], HMMs [27, 2, 39, 7], HMMs involving Bayesian statistics for estimation of some parameters [37].

The principal concern is identification of targets for two TFs binding in coordination, a so called composite element, as the simplest model of cooperative binding. A pair of neighboring TFBSs forms a composite element and in that case their joint action appears to be synergetic or antagonistic, different from the situation when they act independently [11]. FastM is the method developed to search for task-specific models of transcription regulatory units, masks of binding sites in conserved order, separated by spacer [12]. The program allowed defining the model of the experimentally verified NFκB/IRF1 regulatory module and correctly predicting it in the positive set of HLA-A, HLA-B, β-2-microglobulin and β-IFN gene promoter sequences.

Methods to discover cis-regulatory modules possess considerable drawbacks arising from the incomplete set of known TFs critical for biological system function; rather limited training data on known binding sites; and small number of special cases of literature-derived regulatory modules for predominant tissues, insufficient to grasp a more general pattern.

While control of transcription rates is implemented at the levels of the initiation of transcription by RNA polymerase II, activation or repression of transcription by TFs binding to regulatory elements, three-dimensional structure of chromatin, the highest hierarchical level corresponds to the system of integral regulation of transcription defined by the superposition of different DNA codes [31].

This suggests that it is important to consider a sufficient body of information to recognize the regulatory regions as the transcriptional machinery does, taking into account the variety of context, conformational and physicochemical DNA features [25].

An example of such approach to regulatory region identification, namely, promoters discovery, integrating different DNA encodings is the program PromoterExplorer [38]. As input it takes a high-dimensional vector of various DNA features as the local distribution of pentamers, positional CpG island features and the digitized DNA sequence.

3. Relational Data Mining approach and 'Discovery' system

Molecular data are so complicated that it is insufficient now to apply or develop a particular method for verification of a particular kind of hypotheses. If we do not "guess" the kind of a hypothesis corresponding to the regularity hidden in the data, we will fail to obtain the result. The achievement of a biological result demands tools for analysis of a task when the regularity is not known a priori but should be detected. We have to perform flexible study of regularities hidden in data, and this study should constantly outstep the limitations imposed by both the data and the class of hypotheses checked. Therefore, in spite of the advantages of KDD&DM methods, their potential in data analysis is limited, and they do not allow arbitrary alteration of the data under study or classes of hypotheses verified.

Any KDD&DM method explicitly or implicitly assumes:

(1) some quantities for input data;
(2) language (ontology) of a particular KDD&DM method to manipulate and interpret data and results;
(3) class of hypotheses to be tested on data (knowledge space of the KDD&DM approach) in terms of that language;
(4) the ontology of the particular KDD&DM method should be interpretable in the ontology of the subject domain.

The knowledge extracted by a KDD&DM method is the set of confirmed hypotheses that are interpretable in the ontology of the KDD&DM method and in the ontology of the subject domain. Thus, the main characteristics of a particular KDD&DM method are: the ontology of the method and its knowledge space.

To overcome the limitations of KDD&DM methods determined by their ontology and knowledge space we developed a Relational Data Mining approach [14] based on knowledge extraction from data using various classes of hypotheses in the first-order logic. For that purpose a special data mining method 'Discovery' was developed. This approach overcomes limitations of particular KDD&DM methods in the following ways:

a) extends the data type notion, using the first-order logic and the measurement theory for representation of various data types in the many-sorted empirical systems;
b) uses any background knowledge presented in the first-order logic;
c) extracts various hypotheses classes, formulated in the first-order logic.

The 'Discovery' system allows study of a priori unknown regularities of genetic sequence structures. The advantages of the 'Discovery' system [14] over other DM&KDD methods are:

- It can process data of any type measured with reference to any scales, in particular:
  - relations: preference, partial order, grids, etc.;
  - scales: names, order, loglinear, etc.;

- structures: trees, networks, graphs, etc.;
- combinations of all these values, which are of particular importance for heterogeneous databases.

- To retrieve this information, it applies the results of the measurement theory, which shows how all information stored in measured values can be presented by a set of interpreted relations and operations and,
- It can use any background knowledge represented in the first-order logic;
- It can detect and verify any kind of hypothesis formulated in the first-order logic;
- It detects hypotheses in the form of 'law-like' rules that meet all major requirements imposed on natural laws: maximum falsifiability, simplicity, and least numbers of parameters.
- The exhaustive search of specified hypotheses may be immense. For efficient search, the 'Discovery' system uses directed hypothesis search. This search is based on the following theoretical results defining systems of hypothesis ordering:
  a) arrangement by the strength of scales used in the rules. The weaker is the scale, the simpler is the rule. Rules for more complicated scales are refinements (by the information they contain) of simpler rules;
  b) arrangement of rules by logical strength. The search should start from logically strongest rules;
  c) it can be proven that not all rules should be sought but only 'law-like' ones [34].
  d) search of rules, taking into account conditions a—c, is performed by a special semantic probabilistic inference, using only 'law-like' rules.
- The resulting knowledge base can be applied to prediction by using the aforementioned semantic probabilistic inference. It has been proven [35] that predictions by a semantic probabilistic inference yields the highest estimate of the conditional probability of the prediction and avoid the problem of statistical ambiguity.

As for bioinformatics problems, the 'Discovery' system should be adjusted with regard to the specific features of input data and hypotheses. For this purpose, we did the following:

(1) developed input formats and the system of visualization of symbol sequences with indication of signals revealed by other methods;
(2) in order to investigate structural regularities of regulatory gene regions: developed a system for on-line specification of biologically important relations and operations on a symbol genetic sequences including signals revealed by other methods. The 'ExpertDiscovery' system, described below, allows specifying various intervals, distances between signals, and signal repeats.
(3) developed a system for on-line specification of a biologically important class of hypotheses for subsequent check on data;
(4) adapted the method of hypothesis discovery developed for the 'Discovery' system to this type of problems;

As a result an 'ExpertDiscovery' system was developed that is described in the next section.

4. 'ExpertDiscovery' system

The interactive system 'ExpertDiscovery' allows the expert-biologist to investigate the regularities of the molecular-genetics structure by discovering the knowledge in the form of the complex signals. The complex signals are built from the elementary signals generally obtained by other methods. As the elementary signals the system takes:

(1) nucleotides, the context signals, any words in the 15 IUPAC code;
(1) putative functional sites, discovered by the well established approaches, such as PWM, HMM, BLAST [21] and others;
(2) degenerate oligonucleotide motifs [33];
(3) sites with conservative conformational or physical-chemical features (such as double-helix angle twist, DNA melting temperature) [24];
(4) secondary structure element (Z-DNA, RNA hairpin);
(5) low complexity region (polytracks) [23]
(6) the nucleosome formation sites [17]

The complex signal is defined hierarchically and by the induction:

- the elementary signal (e.g. nucleotide, oligonucleotide, see above) is the complex signal;
- the "repetition" N times ($2 \le N_{\min} \le N \le N_{\max}$) of the complex signal is the complex signal. The distance between the neighbor complex signals varies in the user specified range;
- "orientation" (forward, reverse) of the complex signal is the complex signal;
- location of the complex signal (relative to the transcription start) restricted to the certain "interval", defined by user, is the complex signal;
- a pair of ordered complex signals located on some "distance" from each other is the complex signal. Distance varies in the user specified range;
- the logic "conjunction" of the complex signals is the complex signal.

Generally, you are free to organize the most suitable to the data domain list of predicates (such as "repetition", "orientation" and so on) that would participate in the complex signal notion. The main advantage of 'ExpertDiscovery' system is that it provides a powerful tool to formulate the verifiable hypothesis, to choose the language of hidden regularities.

In 'ExpertDiscovery' system the 'law-like' rules appear to be the complex signals characterized by the set of parameters: conditional probability value, significance level (according to Fisher criterion applied to contingency tables), positive/negative coverage (the number of sequences that satisfy the complex signal).

According to the 'Discovery' methodology 'ExpertDiscovery' step by step complicates the current complex signals and finds all chains of nested signals. The complication is realized by the semantic probabilistic inference in such a way when the elementary signals in the complex signal (Fig. 1) are replaced by the predicates "repetition", "orientation", "interval", "distance" from the user-specified list of predicates. The current signal becomes complicated, if the new complicated signal possesses the higher conditional probability value and the lower significance level.

Essentially the complex signal can be expressed by the hierarchical tree (Fig. 1).

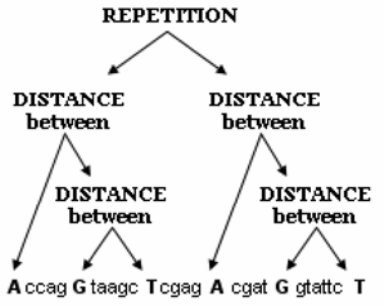


Fig. 1. Complex signal hierarchical tree. Consider the case when the elementary signals to build the complex are nucleotides (the bold nucleotides on figure). Let us follow the left branch of the tree. The nucleotide G is located between the nucleotides A and T, distances between the nucleotides A and G, G and T vary in the user-specified borders. This complex signal is repeated along the DNA length.

The principal merit of 'ExpertDiscovery' system distinguishing it from other systems [38, 9] is the ability of hierarchical analysis of the regulatory regions by means of the hierarchically complicating complex signals. Analyzing the nucleotide context of regulatory regions sequences as the prior information the system takes the markup of specific nucleotides and their positions, thus each complex signal defines the specific nucleotide pattern. The totality of discovered complex signals allows recognizing the short DNA stretches (such as TFBSs with flanks). At the up next level of hierarchy taking as the prior knowledge the conformational or physical-chemical DNA features, overrepresented degenerate oligonucleotide motifs, three-dimensional structure of chromatin and all the signals obtained on the previous level, 'ExpertDiscovery' reveals the complex signals that correspond to the system of integral regulation of transcription defined by the superposition of different DNA codes [31].

These properties of 'ExpertDiscovery' system provide the powerful tool to solve the complicated task - constructing the model of regulatory region generalizing the results of different programs.

Fig. 2 presents the screen shot of the system interface; at the right there is the schematic sketch of the nucleotide sequences under analysis, namely the promoter regions of the endocrine system genes and after the blank the randomly generated sequences, constituting the promoter background. At the left there are the complex signals discovered by the 'ExpertDiscovery' system and parameters of some selected signal. Here as the elementary signals we used the degenerate oligonucleotide motifs discovered by ARGO system in the 15 IUPAC code [33]. The selected complex signal denotes the group of three ordered motifs located at the distance, varying from 20 to 30 nucleotides. The location of the complex signal on the promoter regions is reflected by the dark rectangles.

The system is realized in the interactive mode with the feedback possibility, being in the dialogue with the system one can visualize the complex signal, i.e. to look through the hierarchical tree of the complex signal (Fig. 1,2) and to observe how the complex signal is projected to the data. The system allows editing the complex signals, to manipulate the predicate's degrees of freedom (for example, the number of 'repetitions", and the range of "interval").

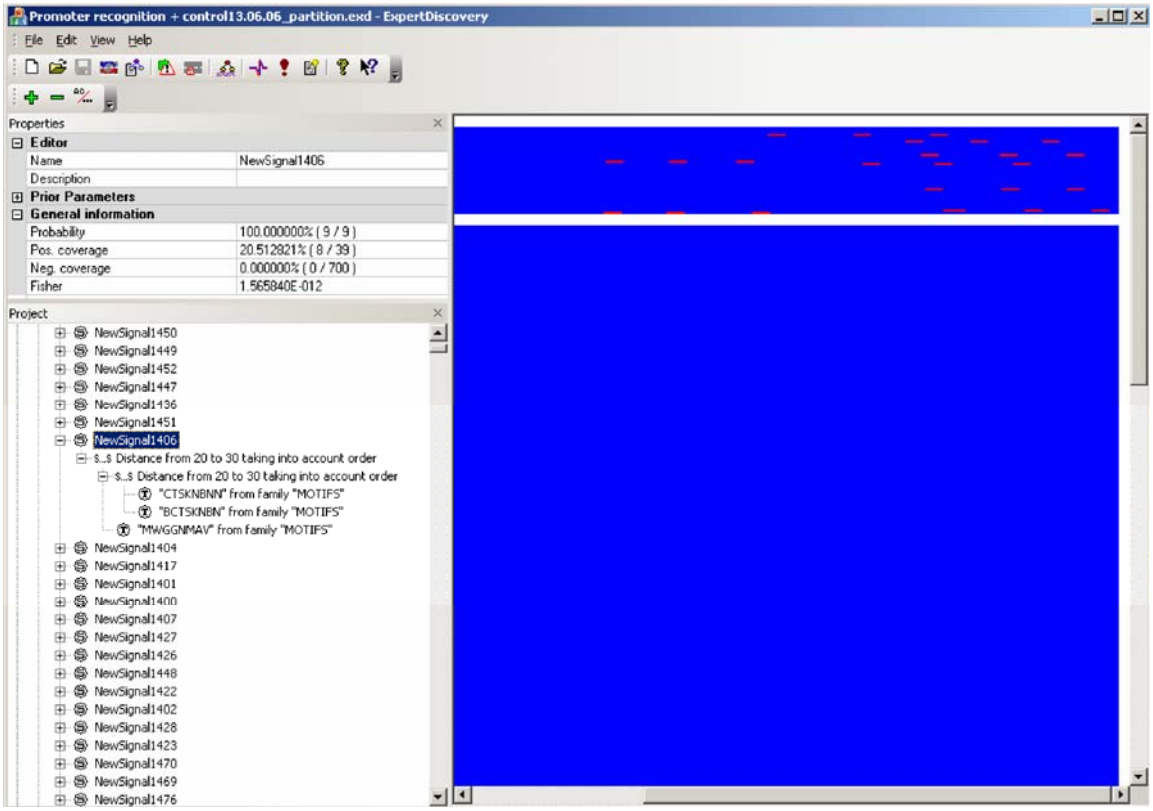


Fig. 2. The screen shot of the 'ExpertDiscovery' system.

5. Experiments.

5.1 Biological data analysis in the case when the elementary signals are nucleotides.

In the case when the elementary signals are nucleotides the system can discover regularities in the nucleotide context.

We analyzed the DNA targets of three protein families: steroidogenic factor-1 (SF1), sterol regulatory element binding protein (SREBP), and early growth response factor 1 (EGR1). The training data sets (sequences of TFBSs with flanks) were retrieved from the TRRD database [13].

We performed the accuracy comparison of the system and the PWM according to the bootstrap procedure [6]. Totally, the data sets contained 53 sequences (SF1 BSs), 38 (SREBP), 22 (EGR1). First of all, we tried the PWM on different sequences lengths as it was described in [16] to reach the highest PWM recognition accuracy. When the optimal sequence lengths for PWM were found to be 13 nucleotides (SF1 BSs), 18 (SREBP) and 10 (EGR1), we prepared positive training sets containing sequences of BSs of the same length. The negative training set consisted of randomly generated sequences with the same frequencies as in the positive set.

The positive training sets were randomly sampled 7 times into the new subsets, each containing 90% of the whole data sets. The PWM and 'ExpertDiscovery' methods trained on the basis of these subsets were applied to the rest of sequences (control subsets, 10% of the whole data set). For each of the control TFBSs we estimated the false positive (FP) rate

relying on the sets of randomly generated sequences of sufficient size (each set of 1 000 000 sequences). Further we ranged the joint set of the control TFBSs according to the corresponding FP rates. Table 1 presents the FP rates with the false negative (FN) rate being equalled to 50%.

Table 1. Recognition accuracy of the 'ExpertDiscovery' system and PWM for TFBSs SF1, SREBP, and EGR1 estimated for the control sequences. False positive rates at the stringent threshold are defined by the false negative rate equal to 50%.

TFBS	'ExpertDiscovery'	PWM
SF1	5.01E-05	6.87E-05
SREBP	1.97E-04	8.32E-04
EGR1	8.09E-04	4.06E-03

The comparison shows that the 'ExpertDiscovery' system can discover regularities governing the nucleotide sequence content, it favorably competes with the optimized PWM in the TFBSs prediction task and even outperforms it (Table 1).

PWM and consensus-based methods involve an explicit assumption concerning the independent contribution of each nucleotide position to the binding affinity, producing the cumulative effect to the binding strength. A number of works [4, 20, 3, 32] indicate that nucleotides of TFBSs cannot be treated independently. This assumption is invalid and contradicts the processes underlying the biological model. Unlike PWM and consensus methods, the 'ExpertDiscovery' system reveals the mutual interdependences between the nucleotides rather distant from each other in the general case. Complex signals discovered by the system cover the groups of sequences intelligent from the biological viewpoint.

For details of 'ExpertDiscovery' application to the TFBSs analysis task let us consider the set of SF1 binding sites. The nucleotide content is depicted by the matrix of absolute nucleotide frequences (Table 2). The most conserved nucleotide positions adjust the consensus sequence of sites (capitalized, table 2), invariant core sequence. Analysis of the complex signals generated by 'ExpertDiscovery' brings to light valuable knowledge, which the system adds in comparison with that extracted from PWM.

Table 2. Matrix of the absolute nucleotide frequencies of the SF1 binding sites. The last row contains the most frequent nucleotide in the position and consensus sequence (capital).

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	7	8	3	**47**	**51**	0	0	3	3	36	9	15	17
T	10	25	1	0	0	0	0	34	10	2	10	13	9
G	27	5	6	6	1	**53**	**53**	1	2	7	18	14	21
C	9	15	**43**	0	1	0	0	15	38	8	16	11	6
	g	t	C	A	A	G	G	t	c	a	g	a	g

The complex signal automatically generated by 'ExpertDiscovery' and corresponding to the consensus sequence of SF1 binding sites is given on the Fig. 3. The hierarchical tree defines the five adjacent nucleotides, overrepresented in the data set. This regularity is fulfilled for 39 objects (74%) of the positive training set and 45 (0,3%) of the negative one (see Fig. 4, complex signal parameters).

According to the 'ExpertDiscovery' methodology the system complicates the current signal if the new complicated one possesses a higher conditional probability value and a lower significance level (section 4). Fig. 5 presents the complex signal that complicates the consensus complex signal (Fig. 3). The signal defines the two nucleotide positions flanking the consensus sequence probably informative for DNA binding mechanism.

The signal corresponding to the reinforced consensus turns to be one of the most significant complex signals, whose conditional probability value equals 93%. This regularity is fulfilled for 14 objects from 53 SF1 TFBSs, and just for the one object from the negative set (Fig. 6).

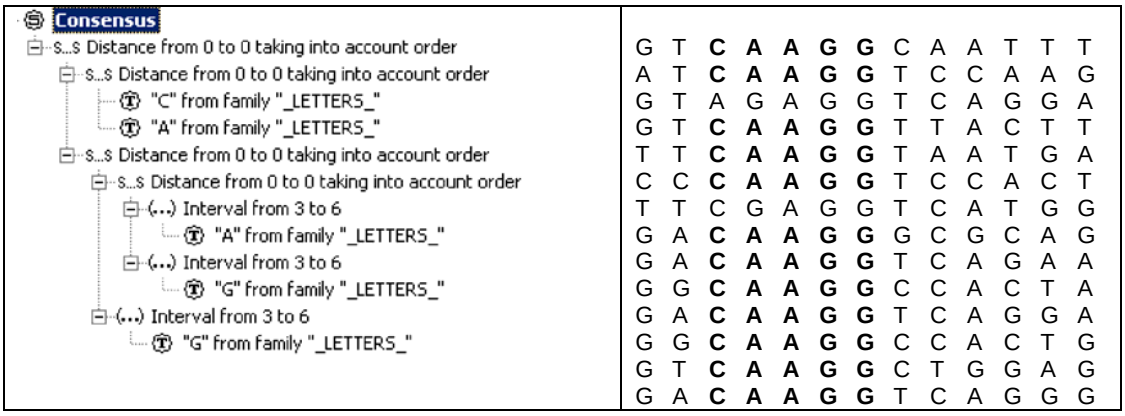


Fig. 3. The hierarchical tree of the complex signal corresponding to the consensus sequence of the SF1 TFBSs. At the right of the figure you can see which nucleotides participate in the complex signal (bold, capital). The notation 'Distance from 0 to 0 taking into account order' means the two adjacent (complex) signals.

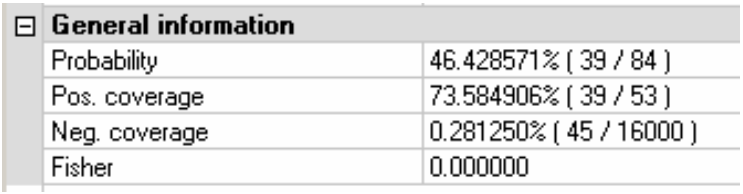


Fig. 4. Parameters of the complex signal given in Fig. 2. Here "Probability" is the conditional probability, "Pos./Neg. coverage" is the number of sequences that satisfy the complex signal, "Fisher" is the significance level (according to the exact Fisher criterion).

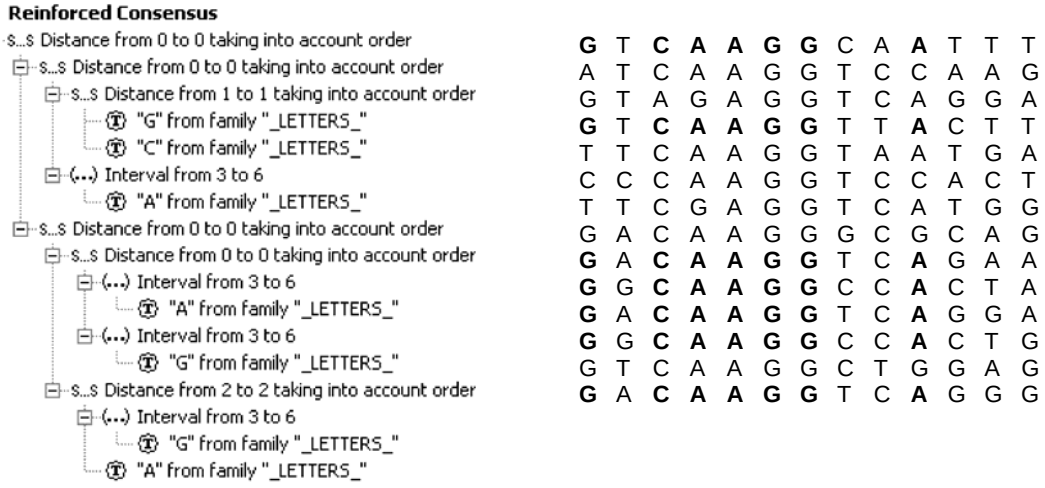


Fig. 5. The hierarchical tree of one of the most significant complex signals, corresponding to the reinforced consensus of the SF1 TFBSs. At the right of the figure you can see which nucleotides participate in the complex signal (bold, capital). The notation 'Distance from 1 to 1 taking into account order' means the two adjacent (complex) signals or (complex) signals having 1 nucleotide between them.

General information	
Probability	93.333333% (14 / 15)
Pos. coverage	26.415094% (14 / 53)
Neg. coverage	0.006250% (1 / 16000)
Fisher	0.000000

Fig. 6. Parameters of the complex signal given in Fig. 4. Here "Probability" is the conditional probability, "Pos./Neg. coverage" is the number of sequences that satisfy the complex signal, "Fisher" is the significance level (according to the exact Fisher criterion).

The decided superiority of the 'ExpertDiscovery' over PWM can become apparent only in the case of sufficient training data containing the most representative set of TFBSs.

5.2. Biological data analysis in the case when the elementary signals are potential transcription factor binding sites.

In the case when the elementary signals are results obtained from different programs, for instance, potential TFBSs located by the optimized PWM, the system discovers regularities over them.

Regulatory regions of genes were extracted from the TRRD database [13] and divided into several groups according to biological features (interferon-inducible, expressed in macrophages, cell cycle genes). The positive set comprised 101 regulatory regions of interferon-inducible genes, the aligned examples being 700 bp long (from –500 bp to +200 bp relative to transcription start site). The negative training set was generated relying on biological data (regulatory regions of genes from contrasting functional groups or of all available genes from human chromosome 1), or it consisted of randomly generated sequences with the same nucleotide frequencies as in the positive set.

At the initial step of our research we trained the optimized PWM on the data sets for more then 30 different TFs presumably sufficient for function of interferon-inducible genes. The thresholds for site types predefined for further analysis are listed in the table. The corresponding TP and FP rates were estimated according to the standard jackknife [6] procedure (see above) in such manner when during iterations exactly one positive object was left for control, thus the number of bootstrap iterations was equal to the number of sites in the TFBSs sample. The TP rates are presented in the table with the FP rate of the order of 1.E–4).

Table 3. True positive and false positive rates for the recognition of transcription factor binding sites, corresponding to the thresholds for further analysis by the 'ExpertDiscovery' system.

TFBS	Power of training set	TP rate	FP rate
Sp1	220	0.24	9.76E-04
Oct family	23	0.42	7.14E-04
CEBP family	88	0.62	7.39E-04
Stat1	32	0.55	1.58E-04
ISGF3	27	0.71	3.23E-04
NFκB family	44	0.84	6.84E-04
IRF family	30	0.77	1.96E-04
AP1 family	20	0.62	8.20E-04

The site types vary in their correlations of TP and FP rates, the best with regard to recognition accuracy appear to be the binding sites of the IRF and ISGF3 TFs, the worst is that of Sp1. We decided on using not the most stringent but concordant thresholds, as the methodology of the 'ExpertDiscovery' system allows to implement additional control and to eliminate noisy data.

The library of optimized PWMs was used to map the locations of binding sites in the sequences of positive/negative training sets in both DNA strands. If more than one match to the matrix had been found, the selection would have been possible based on prior knowledge about biological model of binding. For example, in the case of the self-complementary binding sites (NFkB, STAT1 [1, 28]), when they were recognized in both strands, we preferred one with the highest score.

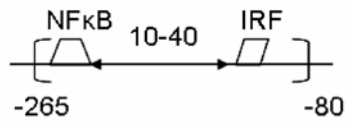
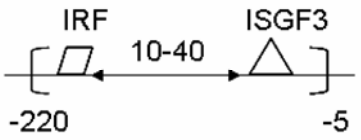
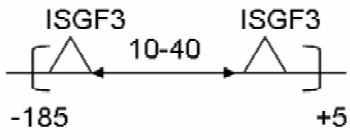
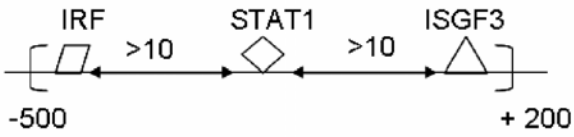
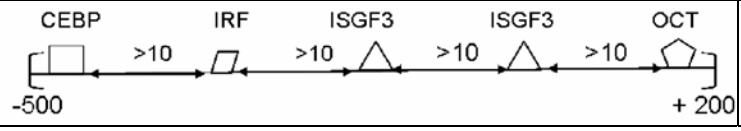
Analysis of the putative site distribution over the training data revealed the set of transcription factors verified in literature, which were critical for interferon-inducible genes function. The relative frequency of their binding site locations in the positive set was significantly different from those calculated for the negative training set. The set of 8 important TFs is presented in Table 3.

The putative binding sites were introduced to 'ExpertDiscovery' as the input markup of elementary signals for further analysis. Complex signals automatically generated by the system can be interpreted as the hierarchically nested models of signals from simple to complex. The simplest model consists of two sites located at the predefined distance from each other. If the distance ranges from 10 to 40 nucleotides, the adjacent sites constitute a composite element or, in general case, a statistically and biologically significant pair of

binding sites (Table 4, first three rows). 'ExpertDiscovery' selects the most reliable simple models and hierarchically complicates them if the new complicated model possesses the higher conditional probability value and the lower significance level according to Fisher criteria (Table 4, last three rows). Our research concludes that not only combinations of two and three ordered binding sites are statistically significant, but the pattern of four or even more binding sites is reliable, probably corresponding to the cis-regulatory module that regulates expression in a combinatorial manner.

In total, the system discovered more then 200 regularities. The number of regularities depends on the user-specified parameters of search: level of conditional probability (greater than 50%) and confidence level for Fisher criterion (less than 0.05).

Table 4. Hierarchically complicated complex signals corresponding to the cis-regulatory modules of interferon-induced genes, revealed by the 'ExpertDiscovery' system.

Simple to complex models of TFBSs	Biological relevance
	synergism, increase of induction [15]
	induction prolongation [18]
	stabilization of the DNA/protein complex, strengthening of the stimulative effect [19]
	cumulative effect [22]
	cumulative effect
	cumulative effect

Acknowledgments

Siberian Branch of the Russian Academy of Sciences (integration project no. 115). The work was in part by supported the President of the Russian Federation, Scientific Schools grant 4413.2006.1.

References

[1] P.A. Baeuerle and D Baltimore, NF-kappa B: ten years after, Cell ;87(1) 1996, 13-20.

[2] T.L. Bailey and W.S. Noble, Searching for statistically significant regulatory modules, Bioinformatics 19 Suppl. 2 (2003). 16-25.

[3] Y. Barash, G. Elidan, F. Friedman, T. Kaplan, Modeling dependencies in protein-DNA binding sites, RECOMB (2003), 28–37.

[4] P.V. Benos, M.L. Bulyk, G.D. Stormo, Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res 30 (2002), 4442-4451.

[5] E.H. Davidson, Genomic Regulatory Systems: Development and Evolution, San Diego, Academic Press (2001).

[6] B. Efron and G. Gong, A leisurely look at the bootstrap the jackknife and resampling. American Statistician 37 (1983), 36-48.

[7] M. Gupta and J.S. Liu, De novo cis-regulatory module elicitation for eukaryotic genomes, Proc. Nat. Acad, Sci USA 102 (2005), 7079-7084.

[8] R.C. Hardison, Conserved non-coding sequences are reliable guides to regulatory elements, Trends Genet. 16 (2000), 369-372.

[9] H. Huang, J. Horng, Yi. Sun1, A. Tsou, S. Huang, Identifying transcriptional regulatory sites in the human genome using an integrated system, Nucleic Acid Res. 32 (2004), 1948-1956.

[10] O. Johansson, W. Alkema, W.W. Wasserman, J. Lagergren, Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm, Bioinformatics 19 Suppl 1 (2003), 169-176.

[11] O.V. Kel-Margoulis, A.E. Kel, I. Reuter, I.V. Deineko, E. Wingender, Transcompel: a database on composite regulatory elements in eukaryotic genes, Nucl. Acids Res. 30 (2002), 332-334.

[12] A. Klingenhoff, K. Frech, K. Quandt, T. Werner, Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity, Bioinformatics 15(3) (1999), 180-186.

[13] N.A. Kolchanov, E.V. Ignatieva, E.A. Ananko, O.A. Podkolodnaya, I.L. Stepanenko, T.I. Merkulova, M.A. Pozdnyakov, N.L. Podkolodny, A.N. Naumochkin, A.G. Romashchenko, Transcription Regulatory Regions Database, (TRRD): its status in 2002, Nucleic Acid Res. 30 (2002), 312-317.

[14] 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), p.308.

[15] J.F. Leblanc, L. Cohen, M. Rodrigues, J. Hiscott, Synergism between distinct enhanson domains in viral induction of TI the human beta interferon gene, Mol.Cell.Biol. 10(8) (1990), 3987-3993.

[16] V. Levitsky, E. Ignatieva, G. Vasiliev, N. Limova, T. Busygina, T.Merkulova, N. Kolchanov, The SiteGA tool for recognition and context analysis of transcription factor binding sites: significant dinucleotide features besides the canonical consensus exemplified by SF-1 binding site, In: Bioinformatics of Genome Regulation and Structure II. (Eds. N.Kolchanov, R. Hofestaedt ,L.Milanesi), Springer Science+Business Media, Inc. (2006), pp. 31-41.

[17] V.G. Levitsky, A.V. Katokhin, O.A. Podkolodnaya, D.P. Furman, N.A. Kolchanov, NPRD: Nucleosome Positioning Region Database, Nucl. Acids. Res. 33 (2005), 67-70.
[18] D.J. Lew, T. Decker, I. Strehlow, J.E. Darnell, Overlapping elements in the guanilate-binding protein gene promoter TI mediate transcriptional induction by alpha and gamma interferons, Mol.Cell.Biol. 11(1) (1991), 182-191.
[19] X. Li, S. Leung, C. Burns, G.R. Stark, Cooperative binding of Stat1-2 heterodimers and ISGF3 to tandem DNA elements, Biochimie 80 (1998), 703-710.
[20] T.K. Man, G.D. Stormo, Non-independence of Mnt repressoroperator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay, Nucleic Acids Res 29 (2001), 2471-2478.
[21] S. McGinnis and T.L. Madden, BLAST: at the core of a powerful and diverse set of sequence analysis tools, Nucleic Acids Res. 32 (2004), W20-5.
[22] J. Mirkovitch, T. Decker, J.E. Darnell, Jr, Interferon induction of gene transcription analyzed by in vivo footprinting, Mol.Cell.Biol. 12(1) (1992), 1-9.
[23] Y.L. Orlov and V.N. Potapov, Complexity: an internet resource for analysis of DNA sequence complexity, Nucleic Acids Res. 32 (Web Server issue) (2004), 628-633.
[24] D.Y. Oshchepkov, E.E. Vityaev, D.A. Grigorovich, E.V. Ignatieva, T.M. Khlebodarova, SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition, Nucleic Acids Res. 32(Web Server issue) (2004), 208-212.
[25] A.G. Pedersen, P. Baldi, Y. Chauvin, S. Brunak, The biology of eukaryotic promoter prediction – a review, Comput. Chem. 15 (1999), 191-207.
[26] P. Qiu, Recent advances in computational promoter analysis in understanding the transcriptional regulatory network, Biochem Biophys Res Commun. 309(3) (2003), 495-501.
[27] N. Rajewsky, M. Vergassola, U. Gaul, E.D. Siggia, Computational detection of genomic cis-regulatory modules applied to body-patterning in early Drosophila embryo, BMC Bioinoformatics (2002), 3-30.
[28] H.M. Seidel, L.H. Milocco, P. Lamb, J. E. Darnell, Jr, R. B. Stein, J. Rosen, Spacing of palindromic half sites as a determinant of selective STAT (signal transducers and activators of transcription) DNA binding and transcriptional activity, Proc Natl Acad Sci U S A 92 (1995), 3041-3045.
[29] G.D. Stormo, DNA binding sites: representation and discovery. Bioinformatics 16 (2000), 16-23.
[30] A.C. Tan and D. Gilbert, An empirical comparison of supervised machine learning techniques in bioinformatics, Proceedings of First Asia Pacific Bioinformatics Conference (APBC) (2003).
[31] E.N. Trifonov, Genetic level of DNA sequences is determined by superposition of many codes, Mol. Biol. (Mosk) 31 (1997), 759-767.
[32] I.A. Udalova, R. Mott, D. Field, D. Kwiatkowski, Quantitative prediction of NF-kB DNA–protein interactions, Proc. Natl. Acad. Sci. USA 99 (2002), 8167–8172.
[33] O.V. Vishnevsky and N.A. Kolchanov, ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoter, Nucleic Acids Res. 33 (2005), 417-422

[34] E. Vityaev and B. Kovalerchuk, Empirical Theories Discovery based on the Measurement Theory, Mind and Machine 14(4) (2004), 551-573.
[35] E. Vityaev, The logic of prediction, In: Mathematical Logic in Asia. Proceedings of the 9th Asian Logic Conference (August 16-19, 2005, Novosibirsk, Russia), World Scientific, Singapore, (2006), pp.263-276.
[36] E.E. Vityaev and T.I. Shipilov, Software for the analysis of gene regulatory sequences by knowledge discovery methods, Bioinformatics of Genome Regulation and Structure II. (Eds. N.Kolchanov and R. Hofestaedt) Springer Science+Business Media, Inc. (2006), 491-498.
[37] Jing Wu and Jun Xie, Computation-based discovery of cis-regulatory modules by hidden markov model, Journal of Computational Biology 14 (6) (2007), in press.
[38] X. Xie, S. Wu, K.-M. Lam, H. Yan, PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm, Bioinformatics 22 (2006), 2722-2728
[39] Q. Zhou and W.H. Wong, CisModule: De novo discovery of cisregulatory modules by hierarchical mixture modeling, Proc. Nat. Acad. Sci USA 10 (2004), 12114-12119.