In Silico Biology 11 (2011/2012) 97–108
DOI 10.3233/ISB-2012-0448
IOS Press

ExpertDiscovery and UGENE integrated system for intelligent analysis of regulatory regions of genes

Y.Y. Vaskin[a,*], I.V. Khomicheva[b], E.V. Ignatieva[c] and E.E. Vityaev[b]
[a]*Novosibirsk State University, Department of Information Technology, Novosibirsk, Russia*
[b]*Institute of Mathematics, SD RAS, Novosibirsk, Russia*
[c]*Institute of Cytology and Genetics, SD RAS, Novosibirsk, Russia*

Abstract. The task of automatic extraction of the hierarchical structure of eukaryotic gene regulatory regions is in the junction of the fields of biology, mathematics and information technologies. A solution of the problem involves understanding of sophisticated mechanisms of eukaryotic gene regulation and applying advanced data mining technologies. In the paper the integrated system, implementing a powerful relation mining of biological data method, is discussed. The system allows taking into account prior information about the gene regulatory regions that is known by the biologist, performing the analysis on each hierarchical level, searching for a solution from a simple hypothesis to a complex one. The integration of ExpertDiscovery system into UGENE toolkit provides a convenient environment for conducting complex research and automating the work of a biologist. For demonstration, the system has been applied for recognition of SF1, SREBP, HNF4 vertebrate binding sites and for the analysis the human gene regulatory regions that promote liver-specific transcription.

Keywords: Complex signal, hierarchical analysis, recognition, gene regulatory regions, bioinformatics

1. Introduction

Analysis of gene regulatory regions and searching for structural and functional patterns are actual problems of biology which are far from a final solution. In general it is due to the complex structure of gene regulatory regions and variety of mechanisms of transcription regulation. There is a need to analyze various data which concern physical, chemical, structural, information properties of gene regulatory regions and experimental data of their functions.

The fundamental property of genes is gene expression (i.e., the ability to produce biologically active products – proteins or RNA). Gene expression occurs in two major stages. At first, genes are transcribed into RNA and then translated to make proteins. Transcription is the first step leading to gene expression. During transcription, a DNA sequence is read by an RNA polymerase to produce an RNA molecule (a primary transcript) with essentially the same sequence as the gene. Given the complexity of multicellular eukaryotes, transcriptional regulation in these organisms needs to be very complex [30].

Intensity of each gene transcription is precisely regulated depending on cellular conditions (a type of cells and tissues, a stage of organism development, a cell cycle, inducers or repressors, influencing the cell) [4]. A great number of different regulatory proteins are involved in the process of transcription regulation, including transcription factors (TFs), coactivators, cosuppressors, mediators [15]. TFs play one of the most important roles in the process. They interact with exact DNA regions – transcription factor binding sites (TFBSs) – in a specific way. Besides interacting with DNA, TFs participate in protein-protein interactions with other regulatory proteins, forming multiprotein complexes which activate or suppress gene transcription.

Possibility of flexible regulation of eukaryotic gene transcription is provided by the presence of extensive gene regulatory regions that have complex block-hierarchical structure [12].

*Corresponding author: Y.Y. Vaskin Novosibirsk State University, Department of Information Technology, Pirogova St., 2, Novosibirsk, Russia. E-mail: vaskin90@gmail.com.

The first level of regulatory regions hierarchy includes various transcription factors binding sites, the short regions of DNA (10–20 nucleotides) recognized specifically by regulatory proteins (transcription factors) [17].

The next level is represented by composite elements which include neighboring TFBSs which acquire new regulation properties as a result of protein-protein interations with corresponding TFs. Composite elements of synergetic type provide a non-additively high level of transcription activation as a result of protein-protein interactions. Composite elements of antagonistic type include overlapping or very closely located TFBSs. In this case two transcription factors compete with each other for binding to DNA, so a stimulating effect of an activator is changed to suppressing effect of an inhibitor or otherwise, depending on cellular conditions [12].

Regulatory units (promoter regions, enhancers, silencers) form the next level in the system of gene regulatory regions hierarchical structure. Their functions are implemented since they contain TFBS and composite elements interacting with regulatory proteins [5]. The location of the regulatory units and their length vary considerably. Enhancers and silencers are units activating or suppressing transcription of a specific gene and they may be located very far from the transcription start site (up to 50000 bp). Enhancers and silencers may be situated in 5'- and 3'- flanking regions of genes or in introns. Promoter regions are regulatory units located right before the start of gene transcription. Their size usually varies from 200 to 1000 bp. [1].

The highest level of regulatory regions hierarchical structure corresponds to the system of integrated transcription regulation [13] which is carried out with the participation of complex system of regulatory proteins interacting with the whole set of regulatory units and elements of a specific gene. Composition of multiprotein complexes is determined by DNA-protein interactions based on superposition of different DNA codes (linear, conformational, etc.).

There is a wide range of structures of regulatory regions because each specific gene should be regulated in a particular way depending on a cellular situation. For instance, according to recent data, the human genome encodes about 1500 TFs [2]. One may expect that the human genome (like genomes of many other eukaryotic organisms) may contain just about the same amount of different types of TFBSs, providing functional properties of gene regulatory regions.

Regulatory regions of each gene include a unique combination of TFBSs of different types. According to TRRD [28], regulatory regions of a specific gene may contain more than 20 different TFBSs, which are experimentally proved. Therefore the whole system of gene integrated regulation may include dozens of regulatory units [14]. The fact that the length and the location of regulatory units may be very different emphasises their variety considerably.

Nowadays there are plenty of widespread computer methods providing analysis of regulatory regions, each of them dealings with certain hierarchical level. For TFBSs recognition such methods as PWM [35], SITECON [18], SiteGA [16] and other are used. But the task of TFBSs recognition is methodologically very difficult due to the high variety of genomic transcription factor binding sites. As a result, all TFBSs prediction methods developed so far have the well known shortcomings such as high over- or under-prediction rates (false positives or false negatives) [12].

The task of recognition of relationships between TFBS recognition corresponds to the other hierarchical level of regulatory regions structure [5]. However, since regulatory regions contain unique TFBSs combinations, the methods face bad representativeness of training sets which do not include enough particular cases of a general situation.

The task of analysis of gene regulatory units and the whole system of integrated transcription regulation is considerably much more complex than the tasks of analysis of TFBSs and their combinations. The reason of it is the existence of a huge variety of regulatory regions structures which is caused by presence of different elementary signals in regulatory regions (TFBSs, conformational, physical, chemical features) and also variability of regulatory regions lengths and locations. From an informatics point of view, the task of eukaryotic gene regulatory regions analysis implies hierarchical analysis of genetic information. To solve the problem it is required to apply the state-of-art computer technologies of intelligent data analysis (Data Mining and Knowledge Discovery). Nowadays none of the known methods can completely solve this problem. In the most cases in order to achieve a biologically significant result scientists have to manually analyze huge amount of information, which may be contradictory in some cases.

In general, automatic methods of gene regulatory regions analysis and recognition must consider various contextual, physical, chemical and conformational features of DNA. So, constructing an integrated method of recognition which would involve signals of different

types acquired as results of applying of other recognition methods is quite an actual problem.

In the current work a new method of intelligent regulatory regions analysis is presented. Our method is based on the integration of mutually complimentary tools, ExpertDiscovery system, which is a powerful tool for hierarchical gene regulatory regions analysis, and UGENE suite, which is a multiplatform bioinformatics toolkit, containing different algorithms for analysis of genetic information [7–11,20]. ExpertDiscovery system was integrated into UGENE toolkit as a plugin. UGENE plugins are unified by common interface and logic of operating. Thus, results obtained from the modules can be combined and provide the potential of ExpertDiscovery system in complex rules extraction. That is we can perform recognition of signals at the lower hierarchical levels by UGENE and pass the results to ExpertDiscovery which is able to do more complex hierarchical analysis. Unlike the well-known methods providing regulatory regions analysis that focus on extraction of single signals, ExpertDiscovery uses integrative attitude for analysis of genetic information. ExpertDiscovery system is able to perform analysis at any level: from nucleotide context to any characteristics of extensive genomic sequences, and gives a possibility to integrate special methods at each level.

The applicability of ExpertDiscovery system for analysis of complex structure of gene regulatory regions was demonstrated by two examples. First, ExpertDiscovery was used to recognize single TFBSs (HNF4, SREBP, SF-1) and recognition accuracy was checked against PWM [35] method. Second, the system had been applied for revealing complex signals present in the regulatory regions of human genes expressed in a tissue-specific manner. For this purpose potential signals found by SITECON [18] (or other site recognition methods) may be used. It was demonstrated that the system can automatically build complex signals (CSs), and in fact some CSs were found to coincide with known composite elements, which functional significance were confirmed by experiments described in scientific papers.

2. Applying the relation approach for analysis of regulatory regions. Expert Discovery system

ExpertDiscovery system applies an original knowledge discovery approach (Relational Data Mining) [20,25]. The approach was used in Discovery system which has been successfully applied for solution some particular problems in the fields of psychophysics, cancer diagnostics and securities rates prediction. The heart of the system is semantic probabilistic inference [25].

The idea of new knowledge discovery is to sequentially increase accuracy of hypotheses so that on each step the hypotheses have the higher probability and definition level. Also the level of significance of the results is tested by statistical criterions.

The *semantic probabilistic inference* is such a sequence of rules C1, C2, … , Cn, where A^i_j, G – atoms. that:

1. $C_i = (A^i_1 \& \ldots \& A^i_{k_i} \Rightarrow G)$, $i = 1,\ldots,n$;
2. C_i – *sub-rule* of a rule C_{i+1}, e.g. $\{A^i_1,\ldots, A^i_{k_i}\} \subset \{A^{i+1}_1,\ldots, A^{i+1}_{k_{i+1}}\}$;
3. $\text{Prob}(C_i) < \text{Prob}(C_{i+1})$, $i = 1,2,\ldots,n\text{-}1$, where *conditional probability* of a rule C_i is defined as: $\text{Prob}(C_i) = \text{Prob}(G/A^i_1\&\ldots\&A^i_{k_i}) = \text{Prob}(G\&A^i_1\&\ldots\&A^i_{k_i})/\text{Prob}(A^i_1\&\ldots\&A^i_{k_i})$;
4. C_i – *Probability law*, e.g. for each rule $C' = (A^i_1\&\ldots\&A^i_{ki} \Rightarrow G)$ of a rule C_i, $\{A_1,\ldots, A_j\} \subset \{A^i_1,\ldots, A^i_{k_i}\}$ an inequality holds: $\text{Prob}(C') < \text{Prob}(C_i)$;
5. C_n – *The strongest probability law*, e.g. rule C_n is not a sub-rule of any other probability law.

Discovery system implements semantic probabilistic inference with knowledge discovery as a set of probability laws, the strongest probability laws and maximally specific laws.

ExpertDiscovery is an adaptation of the Discovery system which is configured to knowledge discovery in sets of nucleotide sequences, according to semantic probabilistic inference, as complex signals with specified parameters.

The following definitions are required for description and formalization of ExpertDiscovery method.

As we know, genetic information is represented as letter sequences in the 4- or 15-letter (IUPAC nomenclature) alphabet.

Signal is a set of rules defining features of DNA sequences. Elementary signal is an indivisible signal which is characterized by a name and locations in a sequence, where it is.

Experts' hypotheses are represented as Complex Signals (CSs) which are defined recursively based on elementary signals and operation applied to them (Fig. 1):

1. The elementary signal is CS.
2. The result of the (see further) "repetition" or "interval" operation applied to CS is CS.

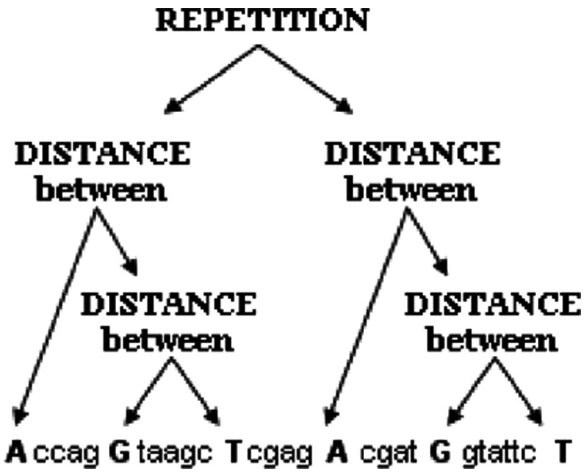

Fig. 1. Hierarchical tree of a complex signal. In that case elementary signals are letters designating DNA nucleotides, and operations are "distance" (between letters and CS) and "repetition" (of two CSs). It follows from the CS definition that each letter and each sub-tree is a CS.

3. The result of the "distance" operation applied to a pair of CSs is CS.

To CS the following operations can be applied:

Distance between signals. The input CSs are s1 and s2. It is specified that the distance between them varies from *min* to *max* and the order is taken into account. The result CS is found in a position if in the position s1 is found, and on a distance from min to max from it s2 is found. In the case if order is unimportant, s2 may be found before s1. Min and max parameters are specified by expert.

Repetition of a signal. A result case is a repetition of an input signal s from N_{min} to N_{max} times and the distance between neighbor repetitions varies from min and max. N_{min}, N_{max}, min and max are specified by expert.

Belonging of a signal to an interval. The input CS must be in the interval from min to max, where min and max are absolute distance values from the first symbol of a sequence. The operation makes sense only for an aligned set. Min and max parameters are specified by expert. Also the distance between a pair of CSs can be determined by any of the following ways:

- from the end of the first signal to the start of the second.
- from the start of the first signal to the start of the second
- from the middle of the first signal to the start of the second

A way of the distance determination is a parameter of the corresponding operation and specified by expert.

By specifying the parameters of the operations an expert specifies a set of operations (SetO) which can be used to create CS as hypotheses and also a set (SetCS) of all the CS which are needed to be tested by expert or to be extracted automatically.

User specifies the set of operations (SetO) which will be applied to CS thus defining the expert hypotheses as CS and sequentially increasing their accuracy. Also it is needed to specify the parameters of CS selection.

The elementary signals are taken as an initial population of signals at the first step of the algorithm. Next steps increase the accuracy of the signals in the population. The following procedure is performed to improve the quality of the CS:

1. Choose one of the elementary signals T of current CS;
2. Choose an operation O from the set of the operations (SetO) and T is substituted by O which is applied for some other elementary signals;
3. The resulting CS is tested on the *selection criterion* (see further):
   a. If it is satisfied then the CS is written to the resulting set (ResCS).
   b. Otherwise the CS is tested on the *branching criterion* (see further). If it is satisfied the signal is transferred to the next population.
   c. If all of the previous criterions are failed then the CS is discarded.

Then the next signal in the population is taken. When there are no more signals in current population the algorithm goes to the next population. The cycle continues while the population is not empty. The result is the set of CS (ResCS). Note, that each resulting CS is more significant and probable than each of its sub-signal.

To test the CS, two sets are required – positive (YES set) and negative (NO set). YES set contains sequences that have some signals in advance. NO set sequences do not have the signals in advance or the sequences can be generated randomly and they are needed to test the statistical parameters of the CS.

ExpertDiscovery uses the following *selection criterions* of the CS:

- A *condition probability* threshold – the minimal value of the condition probability that a signal

must have. Also it is checked that the signal is more probable than the previous sub-signal;

- A *statistical significance by Fisher criterion* threshold – it is needed to check statements 3 and 4 of semantic probabilistic inference;
- If *minimization significance level by Fisher criterion* is set then it is checked that the signal is more significant than the previous sub-signal;
- A statistical significance by Ul criterion threshold [26];
- A positive set coverage threshold;
- Uniqueness check. On different steps the signals with a similar structure can be found. It is possible to choose between saving all the signals or unique signals only.

For branching criterions testing:

- A *condition probability* threshold – it is also checked that a resulting signal after branching is more probable than the initial;
- A statistical significance by Fisher criterion threshold;
- If *minimization significance level by Fisher criterion* is set, then it is checked that a resulting signal after branching is more probable than the initial;
- *Minimal complexity* (amount of contained operations) of the CS;
- Maximal complexity of the CS;
- Correlation of the "distance" operation arguments condition of the CS.

Selection and branching criterions use the following terms:

1. *Condition probability* P of belonging of the CS to YES set.

   $P = a_{11}/(a_{10} + a_{11})$,

   where:

   a_{11} – full amount of inclusions of the CS to YES set,

   a_{10} – full amount of inclusions of the CS to NO set.
2. *Statistical significance by Fisher criterion* (Fisher exact test of contingency tables [6]). Four values are used for calculation of the significance level (f):

   t_{00}-amount of negative sequences having the signal inclusions;

   t_{01} – full amount of inclusions of the CS to YES set

   t_{10} – full amount of inclusions of the CS to NO set

   t_{11} – amount of positive sequences not having the signal inclusions

   $f = (t_{00} + t_{01})!\,(t_{10} + t_{11})!\,(t_{00} + t_{10})!\,(t_{01} + t_{10})!\,/\,((t_{00} + t_{01} + t_{10} + t_{11})!\; t_{00}!\; t_{01}!\; t_{10}!\; t_{11}!)$
3. Statistical significance by Ul criterion [26]
4. *Positive set coverage* in percent (for positive set sequences having the signal);
5. *Negative set coverage* in percent (for negative set sequences having the signal);

For the "distance" operation *correlation level between arguments* is evaluated.

3. UGENE System

The goal of UGENE project is integration of various algorithms from the field of genetics and protein sequences analysis in a unified context [3, 24]. The range of the algorithms includes: patterns search, local alignment (Smith-Waterman), HMMER, restriction sites analysis, DNA assembly (Bowtie, UGENE genome aligner), phylogenetic analysis, multiple alignment (MUSCLE, KAilgne), etc. Two original and powerful designers are implemented – for workflow schemes (Workflow Designer [3]) and for complex queries (Query Designer). A key advantage of UGENE is that the most of the algorithms are integrated into the source package and modified to use internal UGENE data model. This allows one to avoid manual data conversion between input and output tools. Some of the algorithms are optimized for multicore environment and have GPU implementations. UGENE supports reading and writing for more than 20 biological data formats. UGENE has capabilities to request key biological online databases such as NCBI, Genbank, PDB and others.

Huge attention is paid to visualization of the algorithms results and development of an effective user interface. There are a lot of modules for visualization of such biological structures as annotated DNA/RNA or protein sequence, multiple sequence alignment, protein 3D structure and DNA assembly.

UGENE is developed in Qt4 (C++ framework) that allows it to work on different platforms: Win, *nix and Mac. UGENE is a free open source software and it is provided free of charge under the terms of GPLv2.

4. UGENE and Expert Discvoery integrated system

The integrated system is a quite powerful tool for hierarchical regulatory regions analysis. Generating markups with different methods, we allow the system to perform recognition on the high levels of the hierarchy which is impossible with other programs. So, we can extract and investigate a model of a complex regulatory region. All the functionality is accessible in the context of one program.

UGENE uses the system of plugins: each independent module is a plugin which can be switched on or off by user, the plugins can interact with each other.

ExpertDiscovery system algorithms fit well the UGENE concept. That's why it was decided to integrate them into UGENE as a plugin which would repeat and extend possibilities of ExpertDiscovery.

ExpertDiscovery plugin in UGENE has the following advantages:

1. Crossplatforming
2. The unite system
   a. Many algorithms within the bounds of one project, apparently, give more possibilities than many different individual narrow applications. Such an approach simplifies user's work: what is needed is to launch UGENE which gives the access to the wide range of the algorithms instead of launching different unrelated programs.
   b. UGENE plugins have unified interface and work logic. Also, user who is already familiar with UGENE could cope with a new module faster. Thus, ExpertDiscovery uses reliable interface and visualization solutions (sequence view, annotation view, task manager, etc.) of UGENE.
   c. Extension and combination of results possibilities appear (Table 1). For example, ExpertDiscovery markups can be UGENE algorithms' results (SITECON, Weight Matrix, Query Designer, etc.)
   d. Data formats. ExpertDiscovery can read sequences in any format which is supported by UGENE (FASTA, FASTAQ, Genbank, GFF, EMBL, etc.).

As for regulatory regions hierarchical analysis – using UGENE we can create sequences markups with elementary signals which are loaded to ExpertDiscovery for the further analysis and building more complex models of the regulatory regions. For example, it is easy to create markups with UGENE Workflow Designer building up a corresponding scheme (Fig. 2).

It is worth to say that as a markup for ExpertDiscovery any file with annotations in the GenBank format can be loaded. For instance, a markup can be generated with a scheme that uses SITECON method instead of weight matrix method or launching elements search with UGENE Query Designer or any other UGENE tools.

Logically, ExpertDiscovery consists of two parts: the part that extracts CS (ED Signal Extractor) and the part

Table 1
The elementray signals of ExpertDiscovery. UGENE extends ExpertDiscovery CS library. In the integrated system the elementary signals can be any signals from the table

Elementary signals	Description
Nucleotides, context signals, any words in the extended IUPAC code [33]	As an addition to the markups, user can load any context signals extracted by specific programs, for instance, "mask" searching programs [34].
Potential TFBS recognized by traditional methods (weight matrix, statistical methods).	UGENE has weight matrix from JASPAR (511 matrix) and UniPROBE (275 matrix) databases. Besides, it is possible to recognize TFBS with other matrix built from user's set of sequences or provided by user.
Potential TFBS recognized by SITECON method which is based on the analysis of conservative conformational or physical and chemical features of DNA [18]	It is possible to recognize 44 eukaryote binding sites for which UGENE has SITECON models of conservative conformational or physical and chemical features. Besides, if user has a training set of other TFBS he can build his own SITECON model for this type of TFBS in UGENE environment.
Markups generated by UGENE Workflow Designer.	The tool for simple creation of computational workflows allows generating markups using different methods.
Patterns found by UGENE Query Designer	UGENE Query Designer allows extracting signals with different functional significance: open reading frames, repeats, restriction sites, potential TFBSs (found by Weight Matrix or SITECON methods) and patterns.
Any sequence annotations loaded in the GenBank format [22].	Open databases contain annotations, including experimentaly verified data about location of TFBSs, in the widespread format for writing sequences and its annotations.
CS found by ExpertDiscovery system	Any CS can be added to the markup and can be used as the elementary signal for extracting more complex CS.

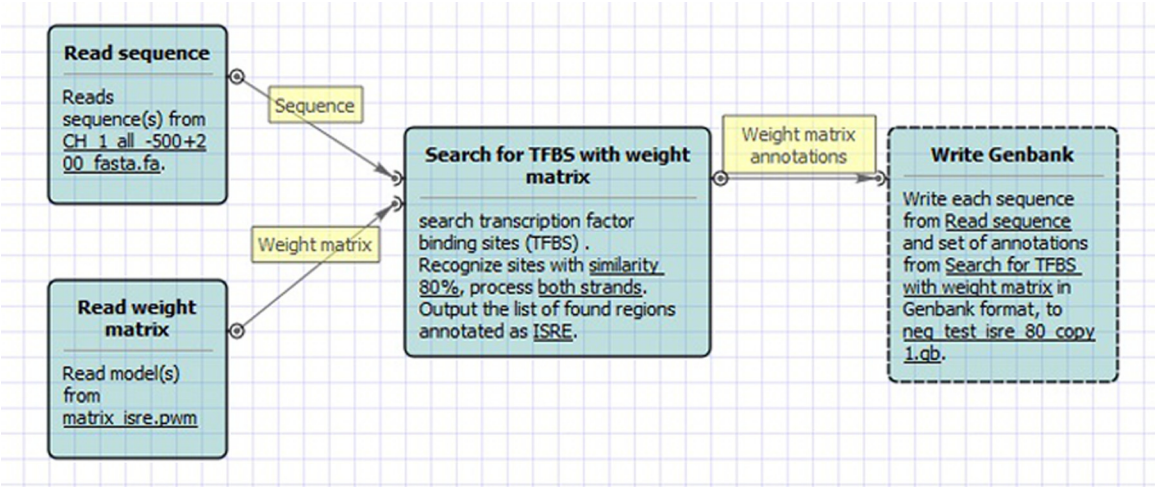

Fig. 2. UGENE Workflow Designer scheme for creating markups with weight matrix method. To create the markup the scheme contains the "Read sequence" element, the "Read weight matrix" element, the element for recognition and the "Write sequence"(and its annotation) element which writes the result into a file in genbank format. The resulting file is the file with annotations of the sequences of IRF binding site. Then the file can be loaded to ExpertDiscovery system as a markup with IRF site as the elementary signal.

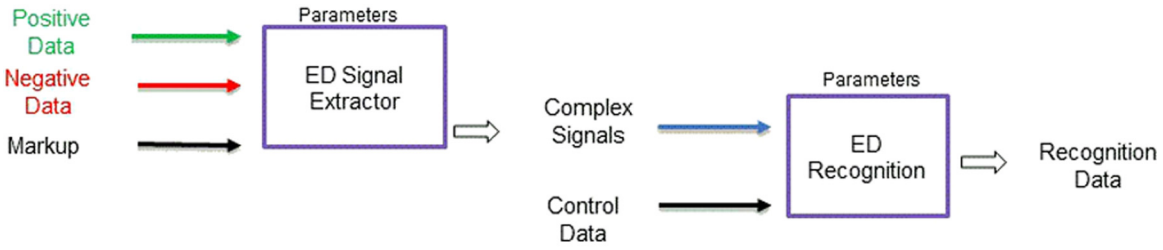

Fig. 3. The main iteration of work of ExpertDiscovery user. First, sequence sets and its markups are loaded into the system. Then the system builds CSs that can be recognized on a control set after.

that recognizes CS on sequences (ED Recognition) (Fig. 3).

Expert loads a positive set of sequences (Positive) containing a regulatory object of user's interest, and a negative set (Negative) which doesn't contain the object. Learning of the system will be based on these two sets. Also it is necessary to set the parameters (Parameters) and load markups of the sequences with elementary signals which will be used to extract complex signals. The output data of the algorithm are CSs (Complex Signals). Then, user can recognize any CS on sequences of the control set (Control). User sets the recognition bound and, as the result, obtains recognition data (Recognition Data) as an HTML report or a recognition profile.

ExpertDiscovery is integrated into UGENE as a plugin and launching from the Tools menu in the main UGENE window (Fig. 4).

4.1. Loading data

To build a model of a regulatory region the system requires the training set: positive and negative set of nucleotide sequences. Sequences of the positive set contain a region of expert's interest. It can be a set of sequences which contains binding sites of a specific type or set of a specific group of genes.

Usually, sequences which don't contain the investigated signal (or a set of signals) are used as a negative set. However, sometimes it is difficult to provide such a set before the stage of computer analysis. That's why the negative set can include so-called "random sequences" generated automatically saving the frequency of occurrence of symbols relatively to the positive set.

Loading of a set of nucleotide sequences dialog is launched by the "New ExpertDiscovery Document"

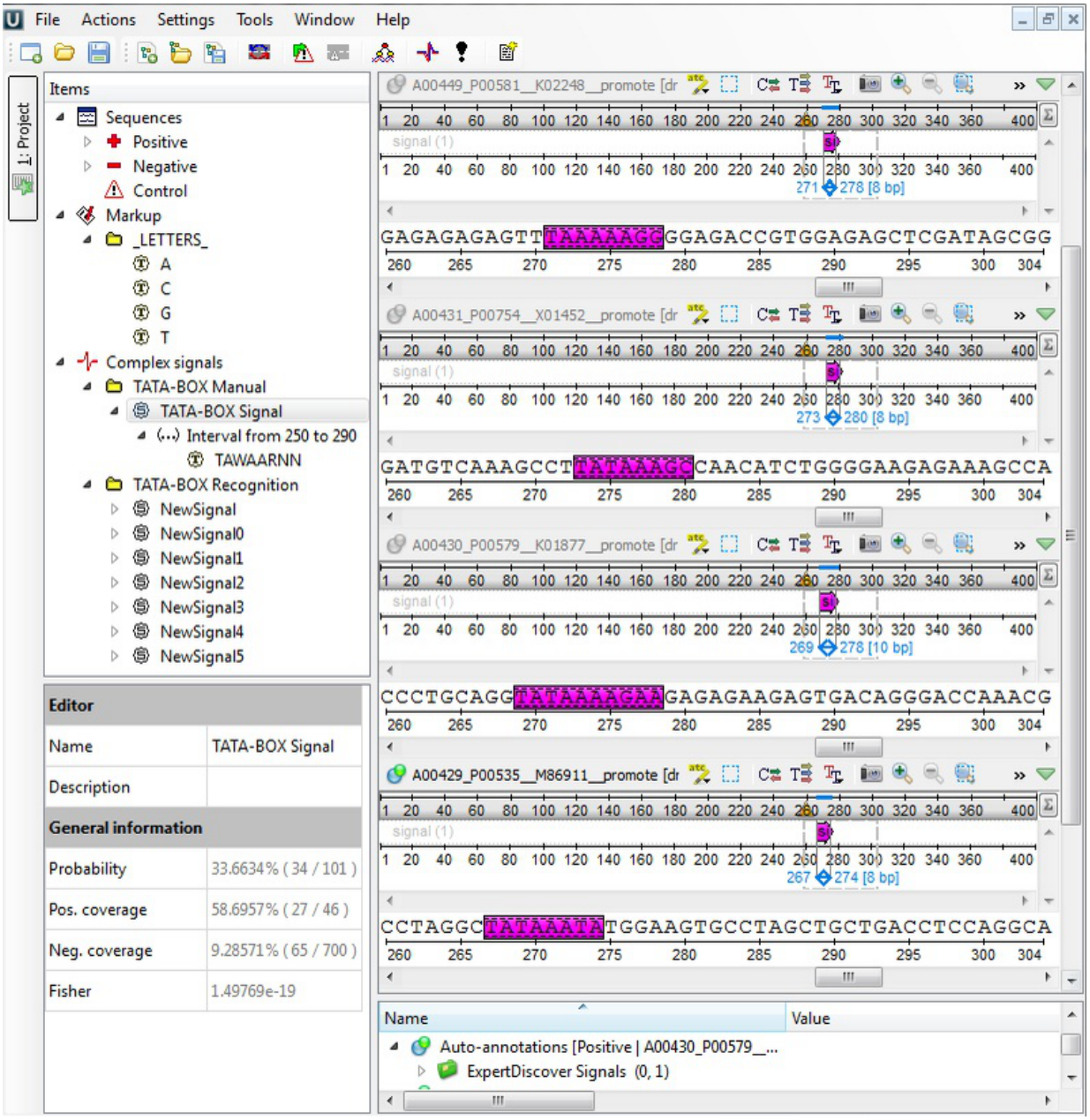

Fig. 4. The main ExpertDiscovery window. Functional of document management, loading markups, extracting signals, etc., can be found as buttons on the toolbar. The window is divided into three areas. The upper left area contains the hierarchical list of the system elements: sequences (positive, negative and control), markups, signals. The lower left area shows properties of a chosen element. On the right – the area of sequence view, a chosen signal is shown as an annotation of sequences.

button on the toolbar. Any sequence files in a format supported by UGENE may be chosen. Then it is needed to load a markup of the sequences with elementary signals which will be the basis for CS. User can chose nucleotides markup or load any markups file generated before. Usually, markups are characterized by locations on a sequence which have the signal and names of a family they are included in.

By the "Load Markup" button on the toolbar the loading of markups of sequences dialog is launched. If the "Append to Current Markup" flag is not checked then the old markup will be deleted.

4.2. CS editing

To manually create CS one can use the popup menu of the "Complex signals" item in the "Items" project window. Also, grouping folders are provided for convenience.

Under definition of CS, it is represented as a hierarchical tree in which the operations are nodes and markups items or words are leafs.

When CS is created and selected, its structure can be changed and parameters can be viewed in the parameters area. The available types of nodes are the "distance" operation (binary), the "repetition" operation, the "interval" operation, the markup items and words. CS is full determined when all its leafs have terminal symbols – words or markup items.

4.3. Creating of CS automatically

Using the training set (positive and negative set, markups) the system can construct a structure of a regulatory region as CS. The extracting wizard is launched by the "Extract signals" button on the toolbar.

In the first dialog window extraction parameters (see below) are set. Next windows are for setting operations which will be nodes of CS and choosing a folder for CS storing.

To see CS location in a sequence it is needed to pick sequences for representation with the popup menu of the sequence. Then, one can choose any CS and it will be shown as annotations on each displayed sequence. Moreover, it is possible to observe a few signals at once on the sequence; for that, a user checks signals for group representation with the popup menu. The same operation is used to choose signals for recognition.

4.4. CS recognition on a sequence

After the CSs are automatically extracted they can be recognized on any sequence. Such a set of sequences can be loaded as the control set.

For recognition some set of CSs is chosen, each of the signals is applied to a sequence. Then, to a symbol of the sequence, where CS is occurred, –log(1-P) score is added, where P is a value of conditional probability of the signal. Score of the sequence is a total score of all its symbols. The sequence is considered to be recognized when it has the selected CS, and its total score is higher than the recognition bound. Expert can choose the recognition bound using the training set. Choosing of the recognition bound is performed in the corresponding dialog by clicking the button "Set recognition bound" on the toolbar. In the dialog errors of the first and the second type are shown for choosing the value.

Also, for convenience, an HTML recognition report can be generated. The report includes statistical parameters and a recognition result for each sequence.

5. Application

The comparison of accuracies of the ExpertDiscovey system and the PWM method was performed using DNA sequences of binding sites of different transcription factors. Three training data sets (DNA sequences of SF1, SREBP, HNF4 vertebrate binding sites with flanks) were extracted from the TRRD database [28]. To reach the highest PWM recognition accuracy the optimal sequence length for PWM was clarified [11]. Then the positive training sets containing DNA sequences of SF1, SREBP, HNF4 binding sites of the same length were prepared. Negative training set consisted of 20 000 randomly generated DNA sequences with the same nucleotide frequencies as in the positive set. The false positive (FP) and the false negative (FN) rates were computed according to bootstrap procedure by jackknife resampling tests [29] relying on the negative and control sequence samples, correspondingly. A comparison of the two methods performance at identifying HNF4 binding sites (Fig. 5) demonstrates that in this case the accuracy of the ExpertDiscovery system is greater than the PWM method: at any given false negative rate ExpertDiscovery system has a lower false positive value. The similar results were obtained for SF1 and SREBP binding sites (not shown). The Table 2 confirms the better performance of the ExpertDiscovery versus PWM model for all three models (HNF4, SREBP, SF1), presenting lower false positive values for ExpertDiscovery provided the same FN rates (50%).

To investigate the complex structure of gene regulatory regions, a set of 303 human promoter regions responsible for liver specific transcription was analyzed. The set of liver specific genes was formed using data about tissue-specific transcription from TiGER [27] and the DNA sequences corresponding to promoters, -500 /-1 (relatively to TSS) were extracted from the UCSC Genome Browser Database [31]. We used SITECON algorithm that is implemented UGENE Workflow Designer and to get the markup for building more complex rules (finding composite elements). Built-in SITECON models of

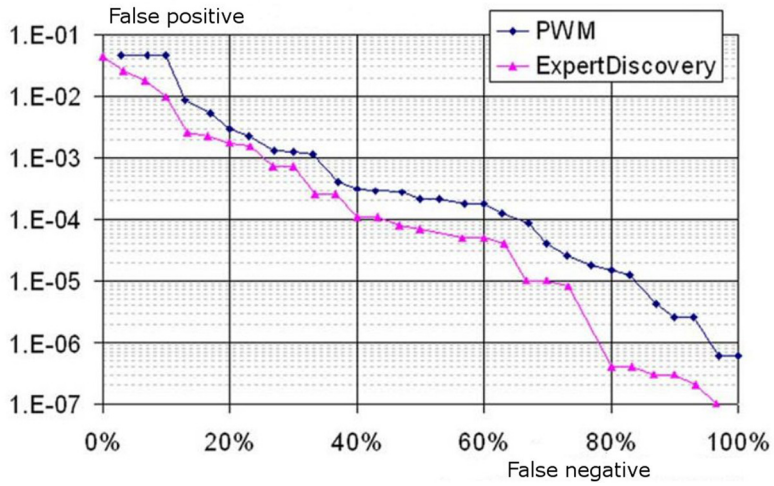

Fig. 5. False positives (FP) versus false negatives (FN) for ExpertDiscovery and PWM model of HNF-4 binding site. At any given false negative rate, better performance is indicated by a lower false positive value.

Table 2
False positive rates for PWM and ExpertDiscovery provided the same FN rates (50%)

Transcription factor binding site	The number of DNA sequences in the control sequence sample	The optimal sequence length for PWM	FP rate	
			ExpertDiscovery	PWM
SF1 (steroidogenic factor-1)	53	13	5.01E-05	6.87E-05
SREBP (sterol regulatory element binding protein)	38	18	1.97E-04	8.32E-04
HNF4 (Hepatocyte nuclear factor 4)	30	13	7.00E-05	2.14E-04

liver-specific sites were used in the workflow scheme: CEBP, GATA1, GATA2, GATA3, TATA-Box, HNF1, HNF3, HNF4, COUP. As a result 80 complex signals were revealed. A number of biologically significant CSs were found among them. For example, two pairs of CSs (GATA-HNF and GATA-GATA) correspond to well known composite elements (Fig. 6), described in the literature [23] and in TRANSCompel database [32].

6. Conclusion

UGENE with ExpertDiscovery plugin is a convenient and effective tool for expert work automation. The integrated system allows combining recognition results obtained by different tools. The combination gives possibilities to extract complex patterns. Analysis is carried out within the environment of one suite which speeds up productivity due to effective and fast interaction between the modules and lack of data conversions.

An original approach to data integration of different methods and knowledge of experts was developed and it showed itself in practice. The results were published in a row of articles [7–12].

Acknowledgments

The work was founded by Russian Federation for Basic Research grant № 11-07-00560-a, SB RAS integrated projects (projects №№ 3, 39, 87, 130, 136), RAS presidium (programs №№ 6.8, 28 (subprogram 2), 30.29), Ministry of Education and Science of RF (Contracts №№ 857, 07.514.11.4003, 07.514.11.4023),

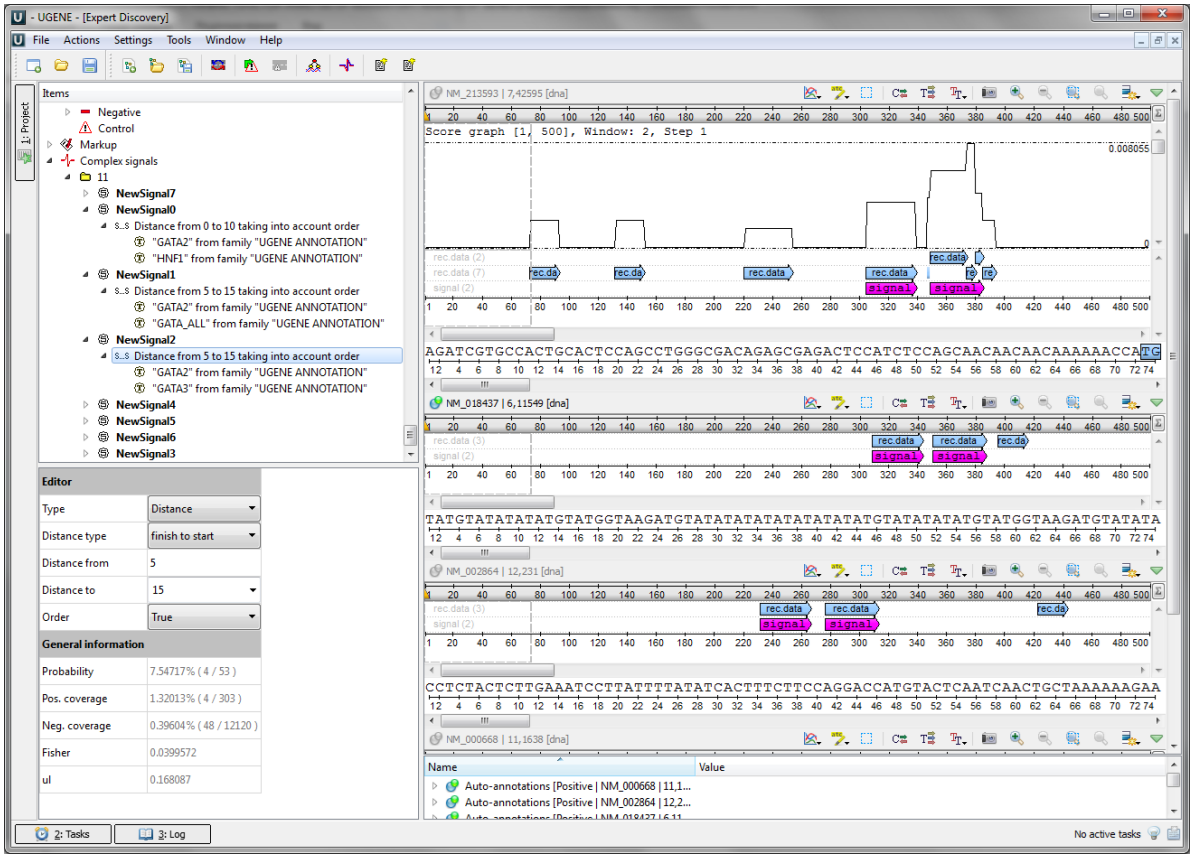

Fig. 6. Complex signals revealed by ExpertDiscovery and UGENE integrated system that correspond to well known composite elements GATA-GATA and GATA-HNF

the RF president's grant council, and support of the leading science schools (project NSH- 5278.2012.4, SS-276.2012.1)

References

[1] Caley M., Smale S.T. *Transcriptional Regulation in Eukaryotes*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2000, 640 p.

[2] Fulton D.L., Sundararajan S., Badis G., Hughes T.R., Wasserman W.W., Roach J.C., Sladek R. *TFCat: the curated catalog of mouse and human transcription factors*. Genome Biol. 2009;10(3):R29.

[3] Fursov, M. Y.; Oshchepkov, D. Y; Novikova, O. S. *UGENE: interactive computational schemes for genome analysis*. 2009, Proceedings of the Fifth Moscow International Congress on Biotechnology 3: 14–15. ISBN 5-7237–0372-2.

[4] Kel A.E., Kolchanov N.A., Kel O.V., Romashenko A.G., Ananko E.A., Ignateva E.V., Merkulova T.I., Podkolodnaya O.A., Stepenko I.L., Kochetov A.V., Kolpakov F.A., Podkolodniy N.L., Naumochkin A.A. 1997. *TRRD – eukaryotic gene regulatory regions database*. Mol. Biol., t. 31, p. 626–636

[5] Kel O.V., Romaschenko A.G., Kel A.E., Wingender E., Kolchanov N.A. *A compilation of composite regulatory elements affecting gene transcription in vertebrates*. Nucleic Acids Res. 1995 Oct 25;23(20):4097–103.

[6] Kendel M., Stuart A. *Statistical inferences and relationships*. M.: Science, 1973, p. 899.

[7] Khomicheva I., Demin A., Vityaev E. (2007b) *Transcription Factor Binding Site Discovery by the Probabilistic Rules*. PKDD Proceedings: Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, Andrzej Skowron (Eds.), Knowledge Discovery in Databases: PKDD 2007. 11th European Conference on Principles and Practice of Knowledge Discovery in Databases. Warsaw, Poland, September 17–21, 2007. Proceedings. Lecture Notes in Artificial Intelligence 4702, Springer 2007, ISBN: 978-3-540-74975-2. 104–109.

[8] Khomicheva I.V., Vityaev E.E., Ananko E.A., Levitsky V.G., Shipilov T.I. (2007a) *Hierarchical analysis of the eukaryotic transcription regulatory regions based on the DNA codes of transcription*. Proceedings of the 3-rd Moscow conference on computional molecular biology. Moscow, Russia, July 27–31, p. 142–144.

[9] Khomicheva I.V., Vityaev E.E., Ananko E.A., Shipilov T.I., Levitsky V.G., «ExpertDiscovery system application for the hierarchical analysis of the eukaryotic transcription regulatory regions based on the DNA codes of transcription». Intelligent Data Analysis, Vol. 12, № 5, 481–494, 2008a.

[10] Khomicheva I.V., Vityaev E.E., Shipilov T.I. *Discovery of the transcription factor binding sites in the aligned and unaligned DNA sequences*. Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2008, 22–28 June, Novosibirsk, Russia), ICG, Novosibirsk, 2008b, p. 116.

[11] Khomicheva I.V., Vityaev E.E., Shipilov T.I., Levitsky V.G., *Transcription factor binding sites recognition by the ExpertDiscovery system based on the recursive complex signals* // Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS2006, 16–22 July, Novosibirsk, Russia), ICG, Novosibirsk, 2006, v.1, pp.77–80

[12] Kolchanov N.A., Merkulova TI, Ignatieva EV, Ananko EA, Oshchepkov DY, Levitsky VG, Vasiliev GV, Klimova NV, Merkulov VM, Charles Hodgman T. *Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes*. Brief Bioinform. 2007 Jul;8(4):266–74.

[13] Kolchanov N.A., Podkolodnaya O.A., Ananko E.A., Ignateva E.V., Stepanenko I.L., Hlebodarova T.M., Merkulova T.I., Merkulov V.M., Mishenko E.L., Ibragimova S.S., Smirnova O.G., Podkolodniy N.L., Romashenko A.G., Oshepkov D. Y., Miginskiy D.S. *DNA regulatory regions: description in databases*. System biology, 2008, Editors: N.A.Kolchanov, S.S.Goncharov, V.A.Lokhoshvay, V.A.Ivanisenko, Novosibirsk, Publishment of SB RAS, 667 p.

[14] Kolchanov N.A., Podkolodnaya O.A., Ananko E.A., Ignatieva E.V., Stepanenko I.L., Kel-Margoulis O.V., Kel A.E., Merkulova T.I., Goryachkovskaya T.N., Busygina T.V., Kolpakov F.A., Podkolodny N.L., Naumochkin A.N., Korostishevskaya I.M., Romashchenko A.G., Overton G.C. *Transcription Regulatory Regions Database (TRRD): its status in 2000*. Nucleic Acids Res., 2000, 28, 1, 298–301.

[15] Lemon B., Tjian R. Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. 2000. V.14, №20. p. 2551–2569.

[16] Levitsky V.G., Ignatieva E.V., Ananko E.A., Turnaev I.I., Merkulova T.I., Kolchanov N.A., Hodgman T.C. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. BMC Bioinformatics. 2007 Dec 19;8(1):481.

[17] Nikolov D.B., Burley, S.K. (1997) RNA *polymerase II transcription initiation: A structural view*. Proc. Natl. Acad. Sci. USA, 94, 15–22.

[18] Oshchepkov D.Y., Vityaev E.E., Grigorovich D.A., Ignatieva E.V., Khlebodarova T.M. (2004) 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), 208–212.

[19] S. Sinha, M. Tompa (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation, Nucleic Acids Research, 30(24): 5549–5560.

[20] *Scientific Discovery Web Site*, http://www.math.nsc.ru/AP/ScientificDiscovery.

[21] Stormo G.D. *DNA binding sites: representation and discovery*. Bioinformatics 2000, 16:16–23.

[22] *The DDBJ/EMBL/GenBank Feature Table*: Definition Version 8.3 Apr 2010 http://www.ncbi.nlm.nih.gov/collab/FT/#5.1

[23] Trainor C.D., Omichinski J.G., Vandergon T.L., Gronenborn A.M., Clore G.M. A palindromic regulatory site within vertebran GATA-1 promoters requires both zinc fingers of the GATA-1 DNA-binfing domain for high-affinity interaction. Mol. Cell. Biol., 1996, vol. 16, no 5 2238–2247.

[24] *Unipro UGENE: an open-source bioinformatics toolkit*; http://ugene.unipro.ru

[25] Vityaev E. E. *Data Mining. Computer cognition*. Models of cognitive processes: Monogr.. // NSU, Novosibirsk, 2006.

[26] Yule U. *On the Association of Attributes in Statistics* // Philosophical Transactions of the Royal Society of London, Ser. A, Vol. 194, 1900, p. 257–319.

[27] Liu, X., Yu, X., Zack, D.J., Zhu, H., Qian, J. (2008) TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinformatics. 9:271.

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

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

[30] Caley M., Smale S.T. Transcriptional Regulation in Eukaryotes. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2000, 640 p.

[31] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun, 12(6).996–1006

[32] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Research 34(Database issue):D108–110.

[33] Cornish-Bowden A. (1985) Enzyme kinetics in Comprehensive Biotechnology (ed. M. Moo-Young), vol. 1, pp. 521–538, Pergamon, Oxford

[34] Grillo G, Licciulli F, Liuni S, Sbisà E, Pesole G, PatSearch: A program for the detection of patterns and structural motifs in nucleotide sequences. Nucleic Acids Res. 2003 Jul 1;31(13): 3608–12.

[35] Gershenzon NI, Stormo GD, Ioshikhes IP. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. Nucleic Acids Res. 2005 Apr 22;33(7):2290–301.