Analysis and Prediction of Regulatory Regions of Eukaryotic Genes by integrated UGENE and ExpertDiscovery Systems

Irina Khomicheva[1,2*], Evgenii Vityaev[2,3], Yurii Vaskin[3], and Timur Shipilov[3]

[1]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
khomicheva@bionet.nsc.ru
[2]Sobolev Institute of Mathematics, Novosibirsk, Russia
vityaev@math.nsc.ru
[3]Novosibirsk State University, Novosibirsk, Russia
vaskin90@gmail.com, tshipilov@gmail.com

Abstract. Appearance of advanced experimental technologies in modern biology resulted in exponential growth of experimental data, which need to be analyzed and mined. New methods of intelligent data analysis are challenged to solve the task of integration of primary raw experimental data, that are poorly consistent and structured, contain gaps, and each method taken separately can't reconstruct completely the biologic system or process. First, we developed the relational Data Mining method ExpertDiscovery, discovering the complex regularities of eukaryotic DNA regulatory regions organization. As the elementary signals to build the complex ones the system takes the different relevant DNA characteristics. Second, we integrated ExpertDiscovery system into UGENE project, uniting tens of state-of-art computing algorithms from the field of genetical and protein sequences analysis. The integration allowed expanding of the ExpertDiscovery elementary signals vocabulary by important information available through UGENE annotations of biological data.

Keywords: Integrated system, complex signal, relational Data Mining, hierarchical analysis, regulatory regions of genes, annotation, recognition, accuracy comparison

1. Introduction

At the heart of creating new generation medical preparations of prevention, prophylaxis of hereditary diseases, etc., there is a problem of eukaryote genes expression control. Genes expression is a complex multiphase process which first stage is the transcription. The transcription for eukaryotic organisms is carried out in cells nuclei. During a transcription there is a synthesis of certain quantity of genes products - RNA molecules. The intensity of each specific gene transcription is regulated depending on cellular conditions (phylum of cells and tissues, a stage of

organism development, a cellular cycle, inducers or repressors, influencing cells, etc.).

Possibility of flexible regulation of eukaryote genes transcription is provided by presence of extensive genes regulatory regions having complex block-hierarchical structure [1-2].

The first level of hierarchy includes various transcription factors binding sites (TFBS), the short regions of DNA serving as a place of alighting for regulatory proteins (transcription factors) [3]. TFBSs occurrence and locating in genes regulatory regions reflects tissue- and stage-specific features of their expression regulation. All methods of TFBS recognition have high enough over-prediction levels. The reason of this is a big diversity of DNA-protein interactions between the sites and the transcription factors, various tissue-, stage-specific mechanisms of transcription regulation, specificity of the context surrounding the TFBS in regulatory regions [4].

The next level of hierarchy corresponds to the arranged composition of the regulatory elements: composite elements, constituted of the pairs of neighboring TFBSs; core promoter, necessary to assemble the basal complex of transcription; cis-regulatory TFBS modules, etc.

The highest level of hierarchy of genes regulatory regions corresponds to the system of the transcription integrated regulation based on the superposition of different DNA codes (linear, conformational) [5].

The analysis and recognition of genes regulatory sequences represent an actual problem of biology and a challenge for working out new methods of KDD&DM (Knowledge Discovery in Databases and Data Mining). To solve the problem in general case it is necessary to consider various contextual, physical, chemical and conformational features of DNA, thus, modeling the regulatory regions recognition process by eukaryotic transcription machine. Constructing the integrated method of recognition, which would involve the signals of various types received as a result of application of other methods and thus, would create a model of regulatory region, is an actual problem, considered in the paper.

In the current work we have made the integration of mutually complementary tools, the ExpertDiscovery system [6-9], a powerful tool for hierarchical analysis of genes regulatory regions, and UGENE - integrated multifunctional tool for molecular biologists, [`http://ugene.unipro.ru`].

2. ExpertDiscovery System – Hierarchical Analysis of Biological Data.

The task of analysis and prediction of eukaryote genes regulatory regions is complex enough. For solution of this task we applied the Relational Data Mining approach Discovery [`http://www.math.nsc.ru/AP/ScientificDiscovery`], [10-13]. We transformed the Discovery system into the ExpertDiscovery system presenting the information extracted from DNA by complex signals (CS), more details are available in [8]. CS is defined recursively in the following way.

- *The elementary signal* is CS;

– CS *orientation* (straight line, symmetric, inverted) is a CS;
– CS ordered pair with a *distance* between them, varied in a certain interval, is a CS. A user specifies that the distance between signals can vary from min to max and the order of elements should be taken into account.
– CS accessory to some *interval* concerning the transcription start (or the beginning of the phased sequence) is a CS. It's specified that an input signal should be searched only in the range from min to max, where min and max are absolute values concerning the first symbol of the sequence.
– CS *repetition N* of times ($2 \le N_{min} \le N \le N_{max}$) is a CS. Thus the distance between the next copies of a signal belongs to the range set by the user. The user specifies N_{min}, N_{max} and distances range (from min to max) between the neighboring repetitions. The manually defined parameters of CS can be set up by user according to his *prior* knowledge about the problem or can be set to maximum values, if user has no such *prior* knowledge about the number of repetitions, distance and interval values.

CS can be presented as a hierarchical tree (Fig. 1). The example of such a tree in case when elementary signals are nucleotides, is presented in Fig.1. In this example the nucleotide G is located between the nucleotides A and T, and the distance between A and G, G and T is varied in the ranges set by the user. This complex signal repeats at least twice (N_{min}), and maximum N_{max} of times in DNA sequence.

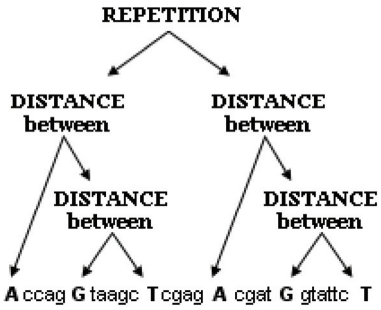


Fig. 1. Hierarchical tree of complex signal.

Elementary signals can be manually specified by the expert online, and also be loaded into the system in the form of DNA sequences annotations. They can be received by applying the known programs of signals recognition. A simple example of an elementary signal is a letter. A more complex example is some word.

In the system the following criteria of CS selection are used:

- *a threshold of positive data coverage* - the minimum percent of sampling YES sequences containing CS. Here the class YES contains the biological data under analysis, for example, the agents of regulatory regions of certain group of genes as the interferon-induced genes, or TATA-containing promoters.
- *a threshold of CS conditional probability* - the minimum value of conditional probability *P* which the signal should have;
- *a threshold of the CS statistical significance* by criterion - Fisher exact test of contingency tables.

According to the Discovery methodology ExpertDiscovery step by step complicates the current CS and finds all chains of nested signals. The complication is implemented by the semantic probabilistic inference [10-12] in such a way when the elementary signals in the CS (Fig. 1) are replaced by the predicates "repetition", "orientation", "interval" or "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.

Every separate CS describes biologically-sensible subgroups of analyzed sequences with significant characteristics. The set of CSs represent a hierarchical model of the objects being analyzed.

3. UGENE System – the Unified Bioinformatics Toolkit.

UGENE is a cross-platform multipurpose application for molecular biologists, uniting tens of important computing algorithms from the field of genetic and protein sequences analysis [`http://ugene.unipro.ru`]. The package gives a powerful and convenient visual interface, and also a possibility of performing of the high-efficiency distributed calculations.

UGENE provides access to a variety of bioinformatics algorithms: pattern search, local sequence alignment (Smith-Waterman), multiple sequence alignment (MUSCLE, KAlign), HMMER, restriction sites analysis, DNA assembly (Bowtie) and many others. 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 the tools' input and output. 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 and includes modules for visualization of such biological structures as annotated DNA/RNA or protein sequence, multiple sequence alignment, biological 3D structure and DNA assembly. UGENE also has capabilities to request key biological online databases such as NCBI Genbank, PDB and others.

Among multiple components of UGENE project there are two unique components that are worth mentioning. The first one is the Workflow Designer, a visual tool for building complex analysis pipelines. The second one is the Query Designer, a tool which allows researchers to analyze a nucleotide sequences using different algorithms at the same time imposing constraints on the positional relationship of the results obtained from the algorithms. Both Workflow Designer and Query Designer have user-friendly interface and utilize all UGENE built-in tools and algorithms.

UGENE is a free software and is provided free of charge.

4. UGENE and ExpertDiscovery Systems Integration.

We have integrated the ExpertDiscovery system into UGENE as a built-in-plugin.

This integration allows ExpertDiscovery to "natively communicate" with other UGENE components and algorithms, extracting from DNA various signals.

UGENE possesses extensive possibilities on allocating fragments of sequences or marking of sequences. In this system they are called the sequences annotations. An annotation is defined by a name, coordinates on the sequence, and by additional qualifiers. Such markings are supplied to the ExpertDiscovery system for the further analysis.

Now as elementary signals for construction complex ones the ExpertDiscovery system, which has been built in the context of UGENE project can use:

(1) Nucleotides, contextual signals, any words in an extended IUPAC code;

(2) Elements accessible through the UGENE Query Designer (open reading frames, repeats, restrictions sites, primers and so on);

(3) Potential binding sites found by the traditional weight matrix method;

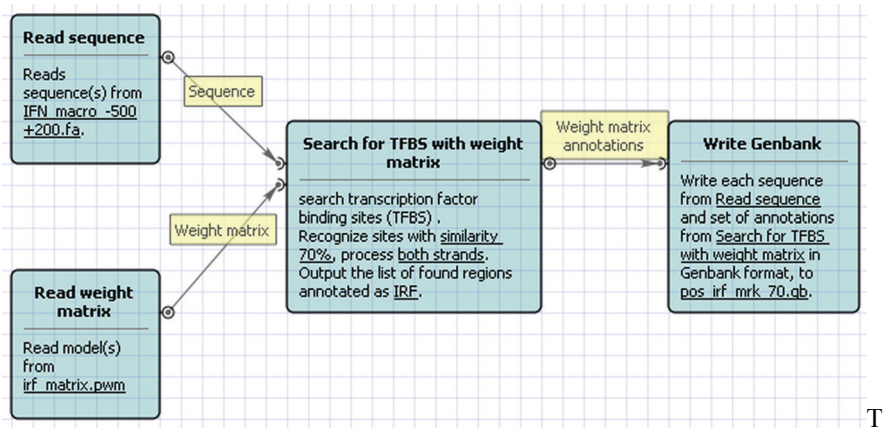
(4) Potential binding sites with conservative conformational or physical and chemical properties (Sitecon method) [14];

(5) Complex signals discovered by the ExpertDiscovery system on the previous stage of data analysis. For example, at first the Expert Discovery program is trained to distinguish TFBSs, potentially critical for functioning of a certain genes group. At the following stage it is possible to implement a hierarchical analysis of TFBSs relative positioning for this genes group;

(6) Signals provided by the UGENE users worldwide.

With UGENE Workflow Designer it is possible to easily create the scheme to mark the sequences being analyzed with potential TFBSs recognized by a weight matrices (Fig. 2). For this purpose the elements of sequences reading, of a matrix reading, a recognition element and an element of result writing are added to the scheme. In the given example in (Fig.2) the step-by-step execution of the scheme allows recognizing IRF TFBSs on the sequences of regulatory regions of interferon-induced genes. The recognition result consists of the sequences annotations specifying the sites locations. The sequences with annotations are written to a file in GenBank unified format [http://www.ncbi.nlm.nih.gov/collab/FT/#5], further this file is supplied to the Expert Discovery system in the form of a marking file. It is worth noticing, that the same way could be used to obtain the marking by SITECON method and by any UGENE tool.

In Fig. 3 the ExpertDiscovery window in the UGENE system is presented. Functions of the document management, loading the marking, marking the signals etc. are accessible from the top toolbar. The window is divided into three areas. The left top area contains the system elements: sequences (in this case promoter regions of endocrine system genes were analyzed), the markings folders, the complex signals. The left bottom area displays the properties of the chosen element. The complex signal corresponding to the cis-regulatory TATA-box element is found on 27 sequences from 46 (Pos. coverage), and on 65 from 700 background sequences (Neg. coverage). On the right you can see the sequences area. A signal currently selected is displayed in the form of annotations on these sequences.



T

Fig. 2. Scheme built in the UGENE Workflow Designer for marking the sequences being analyzed by potential TFBSs recognized by a method of weight matrices.

The system is implemented in the interactive mode with the feedback possibility. Being in the dialogue with the system one can visualize a complex signal, i.e. to look through the hierarchical tree of the complex signal (Fig. 1,3) and to observe how the complex signal is projected to the data. The system allows editing the complex signals, manipulating the predicate's degrees of freedom (for example, the number of "repetitions", and the range of "interval").

5. Discovery of the Transcription Factor Binding Sites in the Unaligned DNA Sequences.

Regulatory regions of genes contain transcription factors binding sites (TFBS). The TFBSs computer annotation is important for understanding the gene expression regulation.

In the course of research the ExpertDiscovery system effectiveness in experiments on aligned TFBS sequences recognition has been shown [6-9].

Comparison of the system with the position weight matrix method (PWM) has shown that on the investigated examples the ExpertDiscovery system catches the nucleotide regularity context and has accuracy, comparable, or over performing the PWM method. Considerable improvement can be reached in case of the adequate training data size containing the TFBSs representative sampling.

Within TFBSs there are the conservative regions (cores) divided by variable regions (spacers). The number of such core regions can be from one to several. The presence of core regions is bound by that the transcription factors have a modular structure, and can contain several domains entering into them, performing specific functions [15-16]. Overcoming the PWM limitation concerning the simplifying assumption of independent contribution of each nucleotide position to the binding

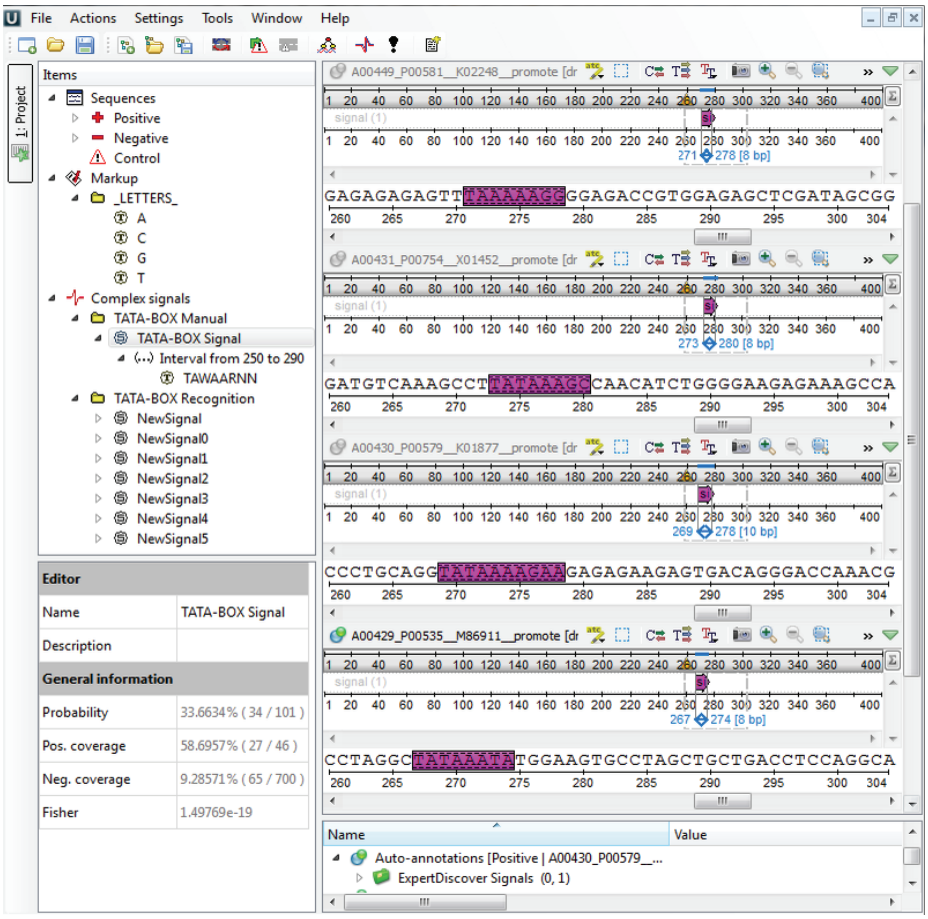


Fig. 3. The screen shot of the ExpertDiscovery system integrated into the UGENE.

affinity ExpertDiscovery system finds dependences between the nucleotides which located quite far from each other in a general case.

In the case when the sequences alignment information is not provided a priori to the ExpertDiscovery system, the system discovers complex signals (1) statistically significant, (2) hierarchically complicating, (3) containing gaps [9].

We have made an experiment on comparing the TFBSs recognition accuracy by the ExpertDiscovery system and PWM. As a modeling object an aligned CEBP TFBSs sampling and a sampling which was not exposed to preliminary procedure of alignment were used. The length of sequences in the sampling was 50 nucleotides, the sample volume - 96 sites. The data were extracted from Transcription Regulatory Regions Database (TRRD, [17]).

In Fig. 4 the results of the "jackknife" procedure [18] applied to compare the recognition accuracy of CEBP TFBSs by two methods are provided.

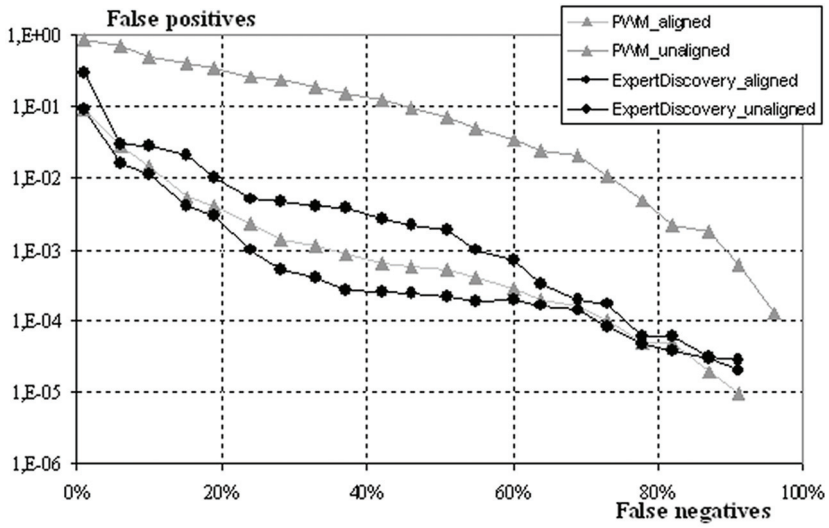


Fig. 4. False positives versus false negatives for the ExpertDiscovery system and PWM in two possibilities: aligned and unaligned sequences under analysis.

The comparison has shown, that in the case when the *a priori* alignment of TFBSs is known ExpertDiscovery outperforms position weight matrix (PWM) at any threshold cutoff. When the *a priori* alignment is not known ExpertDiscovery finds the common building patterns of TFBSs and performs much better than PWM at any threshold cutoff.

6. Hierarchical Analysis and Prediction of Regulatory Regions of Eukaryotic Genes.

To approach the prediction of extensive transcription regulatory regions it is important to consider the variety of biological signals, sufficient to recognize the entire biological situation (tissue, stage specifity, etc.).

In current research we performed the analysis of large-scaled promoter sequences (-500 to +200 bp) of lipid-metabolism genes extracted from TRRD [17], using as the elementary biological signals the statistically overrepresented motifs (gapped words) of fixed length. These motifs found, for example, by YMF or other state-of-art programs, correspond to the potential transcription factor binding sites that are critical for function of co-expressed genomic sequences [19]. For different characteristics of motifs (length, degenerate symbols and spacers number) the accuracy comparison of ExpertDiscovery performance was made according to bootstrap procedure [18]. The best were prefect hexamers. More then 80% of control data were predicted with false negative rate lower then 10^{-7}.

Acknowledgments

This work was funded in part by Russian Ministry of Science and Education (State Contract No. 14.740.12.0819), Russian Science Foundation grant #11-07-00560-a and Integration projects of the Siberian Division of the Russian Academy of science grants ##47, 111, 119.

References

1. Dynan, W.S.: Modularity in promoters and enhancers. Cell. 58(1). 1-4 (1989)
2. Arnone, M.I., Davidson, E.H.: The hardwiring of development: organization and function of genomic regulatory systems. Development. 124(10), 1851-1864 (1997)
3. Nikolov, D.B., Burley, S.K.: RNA polymerase II transcription initiation: A structural view. Proc. Natl. Acad. Sci. USA, 94, 15-22 (1997)
4. Stormo, G.D.: DNA binding sites: representation and discovery. Bioinformatics. 16, 16-23 (2000)
5. Trifonov, E.N.: Genetic level of DNA sequences is determined by superposition of many codes. Mol. Biol. (Mosk) 31, 759-767 (1997)
6. Khomicheva, I.V., Vityaev, E.E., Ananko, E.A., Levitsky, V.G., Shipilov, T.I.: Hierarchical analysis of the eukaryotic transcription regulatory regions based on the DNA codes of transcription. In: Proceedings of the 3-rd Moscow conference on computional molecular biology. pp.142-144. Moscow, Russia (2007)
7. Khomicheva, I., Demin, A., Vityaev E.: Transcription factor binding site discovery by the probabilistic rules. In: PKDD Proceedings: Kok, J.N., Koronacki, J., Mantaras, R.L, Matwin S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. pp.104-109. Springer, Warsaw (2007)
8. 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 (2008)
9. Khomicheva, I.V., Vityaev, E.E., Shipilov, T.I.: Discovery of the transcription factor binding sites in the aligned and unaligned DNA sequences. In: Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure, p. 116. Novosibirsk, Russia (2008)
10. Vityaev, E.: The logic of prediction. In: Mathematical Logic in Asia. Proceedings of the 9th Asian Logic Conference: Goncharov, S.S., Downey, R., Ono, H. (eds.), pp. 263-276. Singapore (2006)
11. Vityaev, E., Kovalerchuk, B.: Empirical Theories Discovery based on the Measurement Theory. Mind and Machine, vol.14(4), 551-573 (2004)

12. Vityaev, E.E., Kovalerchuk, B.Y.: Relational methodology for Data Mining and Knowledge Discovery. Intelligent Data Analysis. vol.12(2), pp. 189-210, IOS Press (2008)
13. Kovalerchuk, B., Vityaev, E.: Data Mining in Finance: Advances in Relational and Hybrid methods. Kluwer Academic Publishers, 2000, p.308. (2000)
14. 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 (2004)
15. Barash, Y., Elidan, G., Friedman, F.,. Kaplan, T.: Modeling dependencies in protein-DNA binding sites. RECOMB, 28–37 (2003)
16. Benos, P.V., Bulyk, M.L., Stormo, G.D.: Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res. 30, 4442-4451 (2002)
17. 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.: Transcription Regulatory Regions Database, (TRRD): its status in 2002. Nucleic Acid Res. 30, 312-317 (2002)
18. Efron, B., Gong, G.: A leisurely look at the bootstrap the jackknife and resampling. American Statistician. 37, 36-48 (1983)
19. Sinha, S., Tompa, M.: Discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Research, 30(24), 5549-5560 (2002)