SOFTWARE FOR ANALYSIS OF GENE REGULATORY SEQUENCES BY KNOWLEDGE DISCOVERY METHODS

E.E. Vityaev[1, 3*], T.I. Shipilov[2], M.A. Pozdnyakov[1], O.V. Vishnevsky[1, 2], A.L. Proscura[1], Yu.L. Orlov[1], P. Arrigo[4]

[1] Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, prosp. Lavrentieva 10, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, ul. Pirogova 2, Novosibirsk, 630090, Russia; [3] Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, prosp. Koptyuga 4, Novosibirsk, 630090, Russia, e-mail: vityaev@math.nsc.ru; [4] ISMAC, via De Marini 6, 16149 Genova, Italy

[*] Corresponding author

Abstract: Application of knowledge discovery techniques to search for and analysis of regularities in the context signals in DNA sequences involved in transcription regulation is described. The software developed allows for interactive construction and visualization of complex signals for samples of DNA regulatory sequences and evaluation of their statistical significance. The complex signals in DNA comprise simple context signals (oligonucleotides), the signals with specified localizations in promoters, predicted and experimentally detected transcription factor binding sites, and other user-specified signals. The regularities are statistically assessed by first-order logic with probabilistic estimates. The computer tool developed was tested using samples of nucleotide sequences of eukaryotic gene promoters extracted from databases according to tissue specificity, type of gene regulation, or joint expression in a functional system. A set of conservative context signals relating the nucleotide sequences and gene functional class was found. The software is available by request to the corresponding author.

Key words: eukaryotic promoter; recognition; transcription factor binding sites; knowledge discovery; data mining

1. INTRODUCTION

Analysis of gene regulatory sequences is a challenge to the theory of data mining and machine learning. Promoters in eukaryotic genomes act as the molecular ‘switches’ that turn genes on and off (Zhang, 1998). They contain transcription factor binding sites (TFBS)—short stretches of DNA sufficiently conserved to provide a specific recognition by the corresponding protein. The presence and location of transcription factor binding sites in regulatory regions of genes correspond to the tissue- and stage-specific features of gene expression in an organism. The problem arises to both determine the TFBS in a DNA sequence and predict location of promoter, which determines regulation of gene expression, according to the localization pattern of particular TFBS (Qiu, 2003).

The problem of TFBS prediction is computationally difficult due to a tremendous diversity of the experimentally detected protein-binding nucleotide sequences compiled in databases (Zhang, 1998; Ohler and Niemann, 2001). Several machine learning approaches to TFBS prediction are developed, including statistics, Markov models, and neural network predictors (Ohler and Niemann, 2001; Vityaev et al., 2001; 2002; Thijs et al., 2002).

Positional weight matrix is a traditional model for predicting binding sites (Chen et al., 1995). Use of the predicted TFBS as signals in nucleotide sequence forms the background for forecast of gene promoters and further determination of the gene function connected with the manner how its transcription is regulated (Vityaev et al., 2001; Kolchanov et al., 2003).

Although many computational methods consider identification of individual transcription factor binding sites, very few of them focus on the analysis of mutual location and interactions between these sites (Cartharius et al., 2005). For example, several approaches are developed to search for clusters of sites in promoters (Zheng et al., 2003; Yu et al., 2004).

Nonetheless, such software is able to give only a general estimate on the degree of randomness in localization of a given number of sites; however, these programs take into account neither the overall diversity of specific features of site localizations nor other characteristics potentially important for regulation of gene transcription. In particular, the latter signals include polytracts, the regions capable of forming noncanonical DNA structures, and sites of preferred nucleosome formation (Orlov et al., this issue).

The main goal of this work is to predict the gene function by using a set of integrated methods for recognition of regulatory elements and transcription factor binding sites. The computer program developed searches for the regularities in locations of binding sites in a regulatory region. For this purpose, the set of context characteristics (specific features) and potential (predicted) TFBS should be first detected for the contrast training sample of promoter sequences. The features considered include computationally predicted binding sites in the sequence analyzed (predicted basing on the consensus, weight matrices, and homology to known sites stored with databases), specific oligonucleotides, low complexity regions, and several other characteristics.

A distinctive feature of our approach is use of specific feature patterns describing a subgroup of the training set (Vityaev et al., 2001). The search patterns for regularities are constructed in the first-order logic augmented by probabilistic estimates. The program is written in C++ and supplied with a user-friendly interface.

We analyzed sets of promoter nucleotide sequences extracted from TRRD and EMBL databases according to tissue-specificity and type of gene regulation. A set of regularities relating the nucleotide sequences and gene functional class were found.

2. SYSTEMS AND METHODS

An interactive system ExpertDiscovery was developed; it allows the user (an expert biologist) to construct hierarchically the complex signals, visualize locations of these signals in DNA sequences, and determine statistical parameters of the signals in a contrast data sample analyzed. Complex signals are determined recursively basing on the primary signals. The following signals may represent the primary signal *S:*

1. A potential functional site predicted by the homology (or weight matrix) with annotated sequences in specialized molecular biological databases (Pozdniakov et al., 2001);
2. A context signal, i.e., a sequence of symbols in a 15 single letter–based code;
3. A site with conserved conformational or physicochemical features (i.e., double-helix angle twist or DNA melting temperature; Oshchepkov et al., 2004);
4. A secondary structure element (Z-DNA or RNA hairpin); and
5. A low complexity region (polytracks); (Orlov and Potapov, 2004).

The complex signal is described hierarchically as

- The primary signal itself;
- Orientation of the signal (direct, symmetrical, or inverted);
- Repetition of the signal N times $(2 \leq N_{min} \leq N \leq N_{max})$; the distance between neighbor copies of the signal falls into a user-specified range;
- Occurrence of the complex signal in a certain range relative to transcription start (or the beginning of a phased sequence); and
- An ordered pair of the complex signals S_1 and S_2 with the distance between them varying in a particular range.

The binding sites can be known and annotated in databases or predicted by certain external software. A universal format of signal layout representation is used to input information about sites and any other context signals into the system ExpertDiscovery. An example of formal TFBS presentation as a signal is given in the scheme below:

```
<Matrics_TFBS>
 Signal_Number 272
 <Signal 1>
name 103_AP2_AS00103
TAGAAAGCCCCGGT
   method_name weight_matrics
 </Signal 1>.
```

The system ExpertDiscovery allows the user to

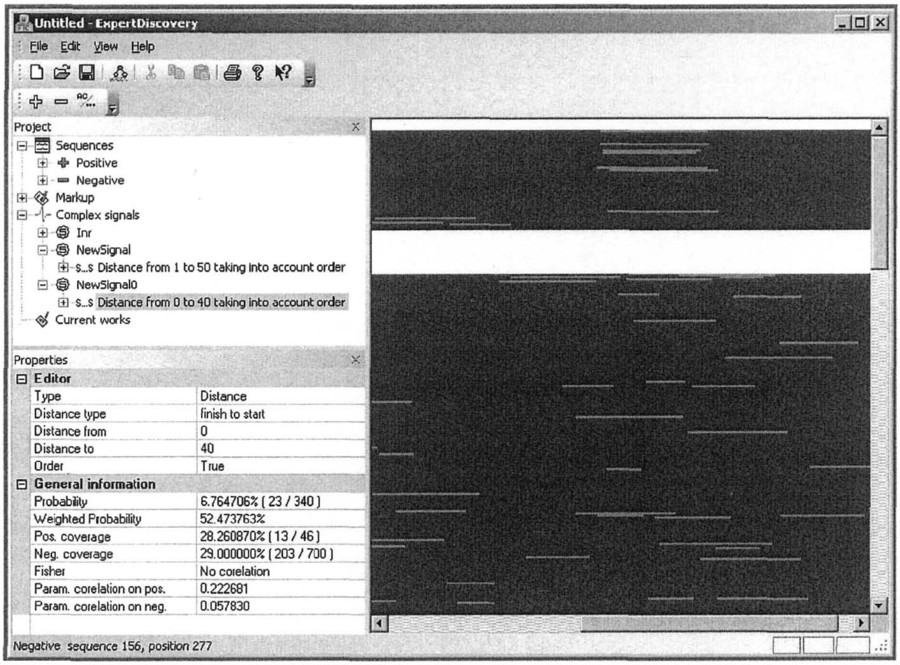
- Input data as a positive and negative samples of DNA sequences;
- Input layout of external signals for the samples in question;
- Specify complex signals in an interactive manner using signal editor and obtain visual pattern of signal localization in DNA sequences and statistical significance of the signal in question;
- Edit any complex signal altering its parameters and involving additional primary signal;
- Detect automatically the statistically significant complex signals specifying beforehand the predicate operations for the search to be performed (for example, all the pairs or all the triplets of primary context signals); and
- Save the project and the relevant data for further expert work.

Figure 1 exemplifies the interface of ExpertDiscovery.

Promoter sequences were extracted from TRRD database (Kolchanov et al., 2002) and divided into several groups according to their tissue specificities (endocrine system genes, cholesterol homeostasis, heat shock response system, interferon-regulated, glucocorticoid-regulated, and cell cycle genes). The negative sets, containing the sequences that are not promoters, were constructed using genomic data (exons and noncoding nonregulatory regions). In addition, randomly generated sequences with the same nucleotide frequencies as in the positive set sequences were used as a negative set.

3. RESULTS AND DISCUSSION

The oligonucleotides specific of promoter regions detected by the program ARGO (Vityaev et al., 2001) and the TFBS predicted by the weight matrices constructed using TRRD database were the primary signals. The predicted sites were involved because the number of experimentally confirmed sites is insufficient for a large-scale statistical analysis. Expert Discovery succeeded in finding numerous regularities for joint presence of context signals in promoter regions.



Sequences
Positive
Negative
Markup
Complex signals
Inr
NewSignal
S...S Distance from 1 to 50 taking into account order
NewSignal0
S...S Distance from 0 to 40 taking into account order
Current works

Editor	
Type	Distance
Distance type	finish to start
Distance from	0
Distance to	40
Order	True
General information	
Probability	6.764706% (23 / 340)
Weighted Probability	52.473763%
Pos. coverage	28.260870% (13 / 46)
Neg. coverage	29.000000% (203 / 700)
Fisher	No corelation
Param. corelation on pos.	0.222681
Param. corelation on neg.	0.057830

Negative sequence 156, position 277

Figure -1. Interface of the system ExpertDiscovery. The right panel represents visualization of TATA box location in promoters of genes of the endocrine system.

The resulting regularities are stored in ExpertDiscovery as complex signals in an IF–THEN form:

$$(A_1\&\ldots\&A_k) \Rightarrow A_0,$$

where the IF part, $A_1\&\ A_2\&\ldots\&A_k$, consists of true/false logical statements $A_1, \ldots, A_k$ concerning the presence of context features (potential TFBS) in a sequence and the THEN part consists of a single logical statement A_0 concerning promoter class and gene function.

Oligonucleotides in a 15 single letter–based code may, for example, serve as context signals A_k. The record of a complex signal is simultaneously a human-readable forecasting rule, which in text format is accepted also by computer program.

Consider an example of a complex signal (YCTNNYTS, DRVSCAG, WTAWWWR) found by the program in promoters of lipid metabolism genes (Figure 2).

As a rule, complex signals characterize a number of sequences (a subset of sequences from a set); in this case, nine sequences (Figure 2). The probability to get a signal due to a random cause is indicated as well as the sequences wherein the signal is located.

```
Regularity 203
IF YCTNNYTS =  1   (Fisher criterion 0.004058)
AND NDRVSCAG =  1   (Fisher criterion 0.005992)
AND WTAWWWRN =  1   (Fisher criterion 0.020397)
THEN Class =  1   (with frequency 9 / (0 + 9) = 1.000000)
Regularity apply to objects: 2(+) 5(+) 16(+) 17(+) 24(+) 26(+)
31(+) 37(+) 50(+)
```

Figure -2. Signal {YCTNNYTS, NDRVSCAG, WTAWWWRN}. Output of the program.

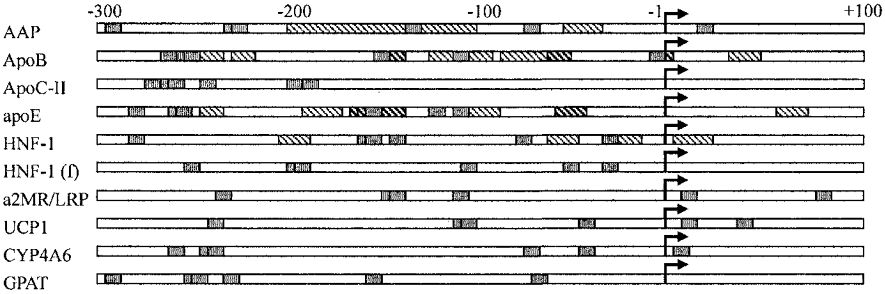


Figure -3. An example of location pattern of potential binding sites (grey rectangles) in gene promoter sequences phased relative to the transcription start (arrow). Genes: human AAP, human ApoB, human apoC-II, human apoE, rat HNF-1, clawed frog HNF-1, human alpha2MR/LRP, rat UCP1, rabbit CYP4A6, and mouse GPAT.

Oligonucleotides may be used for localizing the binding sites in promoters; however, layout of potential binding sites obtained using weight matrices is a more accurate signal. Such layout of the potential TFBS in promoters was made involving the weight matrices constructed using the ArtSite database (Khlebodarova et al., 2005, this issue). An example of a location pattern of potential binding sites in promoters of endocrine system genes is shown in Figure 3. The promoter sequences are aligned relative to the transcription start (position +1 bp), indicated by arrows. Gene names are given to the left.

Potential transcription factor binding sites composing the complex signal are shown as gray rectangles; hatched rectangles indicate those annotated in TRRD databases.The complex signal shown in Figure 3 comprises six potential binding sites, some of which are repeated. The formal signal representation is (239_USF)&(239_USF)&(240_USF)&(323_SPZ1)&(240_USF)&(239_USF).

Here, 239_USF is the weight matrix for USF1 (upstream stimulatory factor 1), determined according SELEX experiment (Khlebodarova et al., 2005, this issue); the protein binds to DNA as a homodimer; 240_USF is a weight matrix for USF1 (matrix for the second half-site); and 323_SPZ1 is the weight matrix for the transcription factor SPZ1 (spermatogenic Zip 1).

Thus, the system ExpertDiscovery helps construction and detection of such complex signals and regularities in promoter regions that cannot be formulated in terms of one of the used approaches (for example, only a group of nucleotides or a pair of binding sites). Functional meaning of the signal could be treated in terms of the transcription factor binding sites or DNA conformational properties. Our study demonstrates that not only pairs of context signals, but also triplets, quadruplets, and larger sets of ordered context signals, which may correspond to groups of jointly functioning transcription factors, are statistically significant.

The regularities found could be analyzed by a molecular biology expert as unique complex signals that are essential for proper promoter functioning. The research suggested that functional promoter modules could be detected by formal models independently of the degree of homology between sequences (Klingenhoff et al., 1999).

This data mining approach is applicable to analysis of the context gene structure at all levels of gene hierarchy: promoter, regulatory regions, and transcription factor binding sites (Liu and Wong, 2003). The algorithm is flexible enough to search for structural patterns that are typical of a whole set of sequences as well as a subset of sequences.

Study of the regulatory regions with the help of contextual signals and predicted transcription factor binding sites may be supplemented with methods of comparative genomics (Dieterich et al., 2005).

ACKNOWLEDGMENTS

The authors are grateful to E. Bedareva, O.A. Podkolodnaya, and N.A. Kolchanov for scientific discussion. This work was supported in part by the Russian Foundation for Basic Research (grants Nos. 05-04-08084, 05-04-49111, 05-07-90185, 05-07-98012, 03-04-48506, 03-07-90181, and 03-04-48469), Siberian Branch of the Russian Academy of Sciences (integration project No. 119), Presidium RAS Program on molecular physicochemical biology (grant No. 10.4), Program for Development of Science Potential of Higher School (grant No. 8319, 308-05), Novosibirsk State University (grant No. 3H-308-05), and NATO (grant No. LST.CLG 979815).