Knowledge Discovery for Gene Regulatory Regions Analysis

Nikolay A. KOLCHANOV, Mikhail A. POZDNYAKOV, Yury L. ORLOV,
Oleg V. VISHNEVSKY, Eugenii E. VITYAEV*, Boris Y. KOVALERCHUK#
Institute of Cytology and Genetics, Acad. Lavrentieva ave., 10, Novosibirsk, 630090, Russia.
E-mails: {kol,mike,orlov,oleg}@bionet.nsc.ru
** Sobolev Institute of Mathematics, Acad. Koptyug prospect, 4, Novosibirsk, 630090, Russia.*
E-mail: vityaev@math.nsc.ru
 # Computer Science Department, Central Washington University, Ellensburg, WA, 98926-7520, USA. E-mail: borisk@cwu.edu

Abstract. This paper presents implementation of Data Mining and Knowledge Discovery techniques for analysis of gene regulatory regions. The method is designed to search for regularities in tables of context features of DNA sequences involved in regulation of transcription. The goal is to discover regularities that relate nucleotide sequences to the functional classes of these sequences. The search patterns for regularities are constructed in the first-order logic augmented by probabilistic estimates. A software system "Gene Discovery" has been developed to implement the method. This system accepts molecular-genetical data retrieved from a database by using SQL queries. Nucleotide sequences of promoters of several functional systems were extracted from the TRRD database and analysed. The data include nucleotide sequences of erythroid-specific gene promoters, endocrine system gene promoters, promoter regions of the genes controlling cell cycle, promoter of genes regulating lipid metabolism, and muscle-specific gene promoters. Several regularities that relate the nucleotide sequences in the regulatory DNA with each functional class have been found. A was developed Discovered Using discovered regularities a a recognition procedure was devised.

Keywords: Transcription Factors Binding Sites, Oligonucleotide Patterns, Eukaryotic Promoter Recognition, Machine Learning, Knowledge Discovery, Data Mining, Bioinformatics.

1. Introduction

Published experimental data and specialized molecular-biological databases contain a tremendous number of experimental results for DNA sequences involved in transcription regulation and gene networks functioning [1,2]. Currently, about 300 different molecular-biological databases are available on the Internet. This provides an opportunity for large-scale data mining and knowledge discovery for bioinformatics [3,4]. Our approach was applied mainly to gene regulatory region analysis. The data source is Transcription Regulatory Regions Database (TRRD) [1]. It was designed for accumulation of experimental information on the structure-function organization of regulatory regions of eukaryotic genes.

The main goal is to make an annotation of an arbitrary nucleotide sequence. We use a collection of methods aimed at recognition of regulatory elements and their binding sites. Analysis of a sequence has several stages:

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

(2) detection the type of either regulatory or structural gene region (e.g., Promoter, 5' UTR, 3'UTR, coding sequence, splice sites, enhancer) on the basis of predicted potential sites;

(3) comparison of predicted structural or functional regions to regions accumulated in the databases and functional annotation of the gene sequence analysed.

The paper describes recognition of the functional type of regulatory region using a predicted set of transcription factor binding sites and context parameters of the 1ucleotide sequence. We consider gene promoters as regulatory regions. Analysis of the promoter structure is critical for understanding molecular mechanisms of transcription. Promoters in eukaryotic organisms act as the molecular "switches" that turn genes on and off. The presence and location of transcription factor binding sites in 5' regulatory regions of genes correspond to the tissue- and stage-specific features of gene expression in an organism. The control of eukaryotic gene expression is primarily determined by relatively short sequences (signal/motif) in the region surrounding a gene. These sequences vary in length, position, redundancy, orientation in DNA chain, and bases. Eukaryotic promoters are characterised by the absence of exact localisation of context signals and the weakness of such signals [5]. Diversity of promoters is the main difficulty for developing of recognition programs [6]. For the last several years, a variety of data mining, knowledge discovery, and machine learning techniques have been used on a large-scale in bioinformatics [4,7,8]. Recently several computational approaches have been suggested to address challenges of combinatorial regulation of transcription [9,10], e.,g., selection of specific oligonucleotides [11] and mining associations between them [12].

Our approach based on data mining methods selects specific oligonucleotide pattern for description of the functional class of a gene [13]. The program uses a training sample of nucleotide sequences of a promoter region. The challenge is describing all eukaryotic promoter sequences by a common pattern. It is difficult because of a huge variability of different transcription factor binding sites. To overcome this difficulty, the sets of promoters of genes that perform a similar function were extracted from the TRRD database [1]. However, even such functional sets lack a single oligonucleotide pattern describing all sequences. A distinctive feature of the algorithm is the use of specific feature patterns. These patterns describe a subgroup of the training set. As we already mentioned the search patterns for regularities are constructed in the first-order logic augmented by probabilistic estimates.

2. Knowledge Discovery System

"Gene Discovery" software was developed for analysis of structural organisation of eukaryotic promoters. The system uses information of both experimentally proved and computer-predicted sites. "Gene Discovery" [14] is an adaptation of a generic relational data mining system called "Discovery" [15,16]. "Gene Discovery" consists of three main modules for: (1) on-line representation of the context signals from DNA sequence in a standard table form; (2) searching for regularities (“Discover” module); (3) recognition of a sequence class by using regularities found. The program is written in the C++. The system is equipped with user-friendly graphical interface.

The method and “Discovery” system reveal statistically significant first-order logic rules for the functional annotation of regulatory gene regions. Learning systems based on the first-order representations have been successfully applied to solving many problems in psychology, physics, medicine, finance, and others [16] (see also "Scientific Discovery" web-site: http://www.math.nsc.ru/LBRT/logic/vityaev/, section "applications"). Since this technique is based on the logic rules, it allows deducing human-readable forecasting rules. Such rules could be interpreted in a biological language and, could support promoter recognition (functional annotation) [17]. A biologist may evaluate both the correctness of the recognition and that of the rules themselves.

Let us consider an example of oligonucleotide motif in 15-lettered IUPAC alphabet: CWGNRGCN=C(A/T)G(A/T/G/C)(A/G)GC(A/T/G/C).

Such motifs could correspond to transcription factors binding sites from the database [18]. The example of the complex forecasting rule including this motif as component part is given below:

IF	CWGNRGCN<NGSYMTAM<MAGKSHCN
THEN	Sequence class = promoter.

The symbol "<" here designates that positions of corresponding oligonucleotides are ordered relative to the transcription start.

This rule means: if motifs CWGNRGCN and NGSYMTAM and MAGKSHCN present in sequence under analysis, and their non-overlapping mutual location is fixed, then the sequence under analysis contains a promoter of the gene of an endocrine system.

In such a way, all the statistically significant oligonucleotide patterns are constructed in the form S_1& S_2& S_3…&S_k, where k>1. The program automatically defines the number of the signals in such pattern [13,14]. "Gene Discovery" implements the methods described above to the analysis of nucleotide sequences of regulatory regions (for detail see [13]).

The learning sample of nucleotide sequences of two alternative classes is used as system input. The learning sample consists of sequences of promoters specific to the functional system ("positive" set) and some random sequences ("negative" set). The latter is a computer-generated random set of sequences with the same nucleotide frequencies or a set of real sequences within the neighbouring regions, which do not perform the particular regulatory function. Consequently, we consider degenerate oligonucleotides as context signals specific for promoters.

Let us describe an analysis of the endocrine system gene promoters to illustrate the issue. A sample of 40 sequences was extracted from the database TRRD (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/). The sequences were of 120 bp in length (from –100 bp to +20 bp relative to the transcription start). The level of homology between any pair did not exceed 60%.

The program permits to analyse any input sequence set in the FASTA format. A functional sample could be extracted from the EPD, TRANSFAC (http://www.gene-regulation.de/), or TRRD databases.

The program ARGO (http://wwwmgs.bionet.nsc.ru/mgs/programs/argo/) was used to select the specific oligonucleotides of 8 bp in length [19]. The term "degenerate oligonucleotides" is used to denote 15-lettered IUPAC coding for nucleotides.

Similarly, other functional sets of promoters extracted from the TRRD database were analysed, including erythroid-specific gene promoters, promoter regions for the cell cycle controlling genes, promoters of genes controlling lipid metabolism, and promoters of genes expressed in muscle.

3. Interpretation of the Rules

A great number of regularities for joint appearance of the context signals in the promoter regions was found as a result of "Gene Discovery" search. This number depends on user-defined search parameters, such as level of conditional probability (greater than 0.5) and level of conditional probability for Fisher criterion (greater than 0.95). Discovered regularities could be analysed further as unique complex signals (patterns of oligonucleotide motifs) significant for proper promoter functioning.

Let us consider selected signal CWGNRGCN<NGSYMTAM<MAGKSHCN. Here the symbol "<" means that positions of corresponding oligonucleotides are ordered relative to the transcription start. An example of the location of this complex signal is presented in Figure 1.

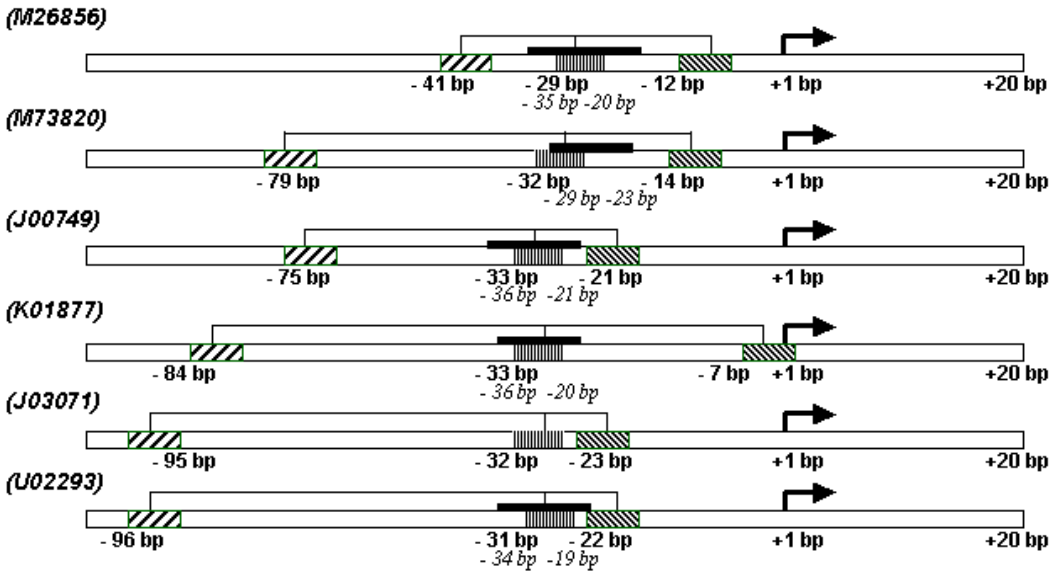

Figure 1. Schematic localisation of the complex signal CWGNRGCN<NGSYMTAM<MAGKSHCN in promoters of genes of endocrine system (E.E. Vityaev *et al,* 2002.).

The promoter sequences are aligned relative to the transcription start (position +1 bp), indicated by arrows. The EMBL identifiers of promoters studied are given in parentheses on the left. The eight-bp oligonucleotide motifs composing the complex signal are shown as shaded rectangles; positions of the first nucleotides are indicated relative to the transcription start. Black rectangles mark positions of the TATA-boxes, indicated in the TRRD database; positions of its first and last nucleotides are italicised. Interestingly, only a single oligonucleotide in the complex signal corresponds to the real annotated site, whereas the others could correspond to potential transcription factor binding sites or to the double-stranded DNA regions with specific physicochemical properties.

Thus, "Gene Discovery" permits to find complex signals in promoter regions. All these signals may be recognized using knowledge about DNA properties and the consensus scheme based on experimental data stored in specialized databases. Any other sample of nucleotide sequences could be analysed similarly. Functional meaning of a signal could be treated in terms of transcription factors binding sites or conformational properties of DNA [10,20].

Recognition rule based on discovered complex signals. This recognition rule for oligonucleotides signals is described in [18]. The score of object based on the score of all signals applied to this object . This score means the probability of appearance of signals on the random sequence. Using negative random samples we can estimate the level of the score, which guarantees some levels of first and second kind errors. If for some control sequence a score is greater then these levels then we predict that this sequence is from some functional class. This approach was extended for the complex signals.

Distinctive feature of the algorithm is the use of a subset of sequences carrying a complex signal. Thus, prediction is applicable only for sequences with homology to a oligonucleotide pattern. Lengths of the gaps in pattern are not fixed. The sequences themselves could have very weak pair homology.

"Gene Discovery" is an integrative part of Internet-based system GeneExpress 2.1 (http://wwwmgs.bionet.nsc.ru/mgs/gnw/) designed for accumulation of experimental data, data navigation, data analysis, and analysis of dependencies in the field of gene expression regulation.

Acknowledgements. The authors are grateful to A.S.Belenok, E.V.Ignatieva, G.V.Orlova for help in research. The work was supported by the RFBR grant (00-04-49229, 00-07-90337, 02-07-90355, 02-07-90355, 01-07-90376, 00-04-49255) and the grant by Siberian Division of RAS (Integration grant N65).

References

[1] N.A. Kolchanov *et al.*, Transcription Regulatory Regions Databases (TRRD): its status in 2002. *Nucleic Acids Research* **30** (2002) 312-317.
[2] E.A. Ananko *et al.*, GeneNet: a database on structure and functional organization of gene networks. *Nucleic Acids Research* **30**(1) (2002) 398-401.
[3] I.B. Jakobsen *et al.*, TreeGeneBrowser: phylogenetic data mining of gene sequences from public databases. *Bioinformatics* **17** (2001) 535-540.
[4] R.D. King *et al.*, The utility of different representations of protein sequence for predicting functional class. *Bioinformatics* **17** (2001) 445-454.
[5] J.A. Goodrich *et al.*, Contacts in context: promoter specificity and macromolecular interactions in transcription. *Cell* **84** (6) (1996) 825-830.
[6] J.W. Fickett and A.G. Hatzigeorgiou, Eukaryotic promoter recognition. *Genome Res.* **7** (1997) 861-878.
[7] Y.-J. Hu, Biological Sequence Data Mining. In: De Raedt L., Siebes A. (eds.), PKDD'2001, LNAI 2168, Springer-Verlag Berlin Heidelberg, 2001, pp.228-240.
[8] E. Kretschmann *et al.*, Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT *Bioinformatics*. **17** (2001) 920-926.
[9] A.E. Kel *et al.,* Computer-assisted identification of cell cycle-related genes–new targets for E2F transcription factors. *J.Mol.Biol.* **309** (2001) 99-120.
[10] A. Klingenhoff *et al.*, Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15** (1999) 180-186.
[11] G. Thijs *et al.*, A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17** (2001) 1113-1122.
[12] J.-T. Horng *et al.*, Mining putative regulatory elements in promoter regions of *Saccharomyces cerevisiae. In Silico Biology* (2002) 0025, http://www.bioinfo.de/isb/2002/02/0025/.
[13] E.E. Vityaev *et al.*, Computer system "Gene Discovery" for regularities retrieving in eukaryotic regulatory sequences organization. *Mol. Biologia (Mosk).* **35** (2001) 952-960 (in Russian).
[14] E.E. Vityaev *et al.*, Computer system "Gene Discovery" for promoter structure analysis. *In Silico Biology* 2002, http://www.bioinfo.de/isb/2002/02/0024/
[15] B. Kovalerchuk *et al.*, Consistent and Complete Data and "Expert" Mining in Medicine, In: Medical Data Mining and Knowledge Discovery, Springer, 2001, pp.238-280.
[16] B. Kovalerchuk and E. Vityaev, Data Mining in finance: Advances in Relational and Hybrid Methods. (Kluwer international series in engineering and computer science; SECS 547), Kluwer Academic Publishers, 2000, 308 p.
[17] T. Mitchell, Machine Learning. New York: McGraw Hill. 1997.
[18] O.V. Vishnevsky and E.E. Vityaev Analysis and recognition of promoters of the erythroid-specific genes on the basis of degenerated oligonucleotide motifs. *Mol. Biologia (Mosk).* **35** (2001) 979-986 (in Russian).
[19] V.N. Babenko *et al.*, Investigating extended regulatory regions of genomic DNA sequences. *Bioinformatics* **15** (1999) 644-653.
[20] Y.V. Kondrakhin *et al.*, Eukaryotic promoter recognition by binding sites for transcription factors. *Com put Appl Biosci*. **11** (1995) 477-488.