

Databases and tools for *in silico* analysis of regulation of gene expression *

Alexander Kel^{1,2}, Olga Kel-Margoulis^{1,2}, Jürgen Borlak³, Dimitry Tchekmenev¹
and Edgar Wingender^{1,4}

¹ BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbuettel, Germany;

² Lab. of Molecular Genetic Systems, Institute of Cytology and Genetics, SB RAS, 630090, Novosibirsk, Russia;

³ Fraunhofer Institute (Fh-ITEM) of Toxicology and Experimental Medicine, Center for Drug Research and Medical Biotechnology and Center of Pharmacology and Toxicology, Medical School of Hannover, Nikolai-Fuchs-Str. 1, D-30625 Hannover, Germany

⁴ Dept. of Bioinformatics, UKG, University of Goettingen, Goldschmidtstr. 1, D-37077 Goettingen, Germany

Author details:

1. Alexander Kel, PhD, corresponding author

address: BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbuettel, Germany
e.mail: ake@biobase.de

2. Olga Kel-Margoulis, PhD

address: BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbuettel, Germany
e.mail: oke@biobase.de

3. Jürgen Borlak, PhD, Prof.

Fraunhofer Institute (Fh-ITEM) of Toxicology and Experimental Medicine, Center for Drug Research and Medical Biotechnology and Center of Pharmacology and Toxicology, Medical School of Hannover, Nikolai-Fuchs-Str. 1, D-30625 Hannover, Germany
e.mail: borlak@item.fraunhofer.de

4. Dimitry Tchekmenev

address: BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbuettel, Germany
e.mail: dte@biobase.de

5. Edgar Wingender, PhD, Prof.

address: Dept. of Bioinformatics, UKG, University of Goettingen, Goldschmidtstr. 1, D-37077 Goettingen
e.mail: e.wingender@med.uni-goettingen.de

* Handbook of Toxicogenomics: Strategies and Applications. Edited by Jürgen Borlak, Copyright © 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, ISBN: 3-527-30342-1, pp. 253-290.

1. Introduction.

Regulation of gene expression becomes the key problem of the era of "Functional Genomics". Now we know that genes in genomes of higher eukaryotic organisms are regulated mainly by means of multiple regulatory proteins - transcription factors (TF), acting through specific regulatory sequences (TF binding sites) that are usually located in the proximity of the genes when constituting a promoter, or at more remote locations when acting as part of an enhancer. Having available genomic sequences on one hand and the massive though phenomenological gene expression data on the other hand, the challenge is to understand regulatory mechanisms of every single gene in the genome by computer analysis of the gene regulatory sequences and by integrating this data with biological knowledge of the gene signal transduction, metabolic and physiological networks. Sophisticated computational regulatory sequence analysis tools that employ powerful statistical and machine learning algorithms driven by the rich databases that collect known biological facts enable us to make profound *in silico* predictions and formulate experimentally testable hypotheses. Such *in silico* driven experiments can greatly speed up the process of our understanding of gene regulatory mechanisms and the identification of new target genes. The understanding of how gene regulation mechanisms are encoded in the genomic regulatory sequences will give us a powerful means for deciphering causes of major human diseases.

2. Concepts of gene regulation.

In multicellular organisms the number of different possible intracellular molecular states is extremely large. These states correspond to different stages of cellular ontogenesis in different tissues, organs and cell types, to a number of developmental stages and cell cycle phases, to the huge number of cell responses, to different external and internal signals and influences. Every state is characterized and precisely organized by differential expressions of specific sets of genes. To generate this huge diversity of cellular molecular states, most of the genes in genome organize their expressions in multiple ways producing a complex pattern of gene expression through various cellular conditions. Gene expression is regulated on several stages, namely transcription (including initiation, elongation, and termination), splicing, translation, and protein degradation. In many cases, transcription initiation is the most important and tightly regulated level of gene expression. Multiplicity of gene expression on the transcription level is provided by means of combinatorial regulation. Combinatorial regulation of transcription is organized through binding of a multiplicity of transcription factors (TFs) to their target sites (cis-elements) in regulatory regions. Corresponding TFs interact with each other and with particular components of the basal transcription complex as well as with coactivators/corepressors, histone acetylases/deacetylases, therefore making up function-specific multiprotein complexes which are often referred to as enhanceosomes.

2.1. Transcription factors

Transcriptional regulation is achieved by a functionally defined large family of proteins: the transcription factors (McKnight & Yamamoto, 1992; Wingender, 1993). They interact with the DNA of promoters and enhancers in a more or less sequence-specific manner, recognizing defined sequence patterns and/or structural features. In contrast to prokaryotes, where the major control mechanism is to repress a generally active transcription machinery, eukaryotes have to match much more complex requirements to coordinate the execution of genetic programs. This is achieved by directed activation of those genes whose products are needed under certain cellular conditions, in general only a few percent of all genes of the genome. Once bound to the DNA, these factors may influence transcription through several mechanisms:

- (i) in most cases studied so far, they enhance the formation of the pre-initiation complex at the TATA-box / initiator element through interaction of a *trans*-activation domain with components of the basal transcription complex, either directly or through co-activators / mediators;
- (ii) some transcription factors cause alterations in the chromosomal architecture, rendering the chromatin more accessible to the RNA polymerase(s);
- (iii) some are auxiliary factors, adjusting an optimal DNA conformation for the activity of another transcription factor;
- (iv) some factors exert repressing influences, either directly by an active inhibiting domain, or by disturbing the required ensemble of transcription factors within a regulatory array (promoter, enhancer);
- (v) finally, there is a group of transcription factors that do not directly bind to DNA but rather assemble into higher order complexes through protein-protein interactions.

A definition of "transcription factors" has been proposed earlier (Wingender, 1997): **A transcription factor is a protein that regulates transcription after nuclear translocation by specific interaction with DNA or by stoichiometric interaction with a protein that can be assembled into a sequence-specific DNA-protein complex.**

Most transcription factors are modularly composed. They may comprise

- (a) a DNA-binding domain (DBD);
- (b) an oligomerization domain, since most factors bind to DNA as dimers, some also as higher order complexes; in most cases, this region forms a functional unit with the DBD;
- (c) a *trans*-activation (or *trans*-repressing) domain, which is frequently characterized by a significant overrepresentation of a certain type of amino acid residue (e. g., glutamine-rich, proline-rich, serine-/threonine-rich or acidic activation domains);
- (d) a modulating region which is often the target of modifying enzymes, mostly protein kinases;
- (e) a ligand-binding domain.

Most likely with the exception of the last domain type, all others may be redundantly present in a single polypeptide chain.

2.2. Modern concepts of the structure and function of the gene regulatory regions in genome.

Blueprints of gene regulation are encoded in the structure of gene regulatory sequences. It is generally accepted that gene regulatory regions of eukaryotic organisms have a modular structure. The fundamental principle of how molecular genetic systems are organized can be described as a hierarchy of modules (Ratner, 1990). It is realized in the modular structure of genomic DNA, both in its structural and regulatory parts (Ratner, 1992). In recent years much attention has been paid to the investigation of the modular structure of regulatory regions that control transcription of eukaryotic genes (Dyanan, 1989; Johnson & McKnight 1989; Werner, 1999; Struhl, 1999). This is a very important principle for understanding molecular mechanisms of functioning of these regions, their evolution and what is particularly important for deciphering complex mechanisms of differential gene expression in multicellular organisms.

2.2.1. Modular hierarchical structure of the gene regulatory regions.

Regulatory DNA is characterized by a modular hierarchical structure. An elementary module corresponds to a single transcription factor-binding site. Next, hierarchical levels are occupied by composite elements, promoters and enhancers, and finally by an integral system of gene transcription regulation.

Regulatory regions of every gene contain a number of binding sites for structurally and functionally different transcription factors, thus providing combinatorial regulation – one of the major principles of genome structure and functioning. Gene-specific regulation in a number of cellular situations is achieved through the formation of multi-component protein complexes on regulatory DNA. Both specific protein-DNA and protein-protein interactions contribute to gene-specific transcriptional regulation.

We consider several levels of structural hierarchy:

1) The minimal functional modules are binding sites for transcription factors. They are short DNA elements (5-20 bp) that can be specifically recognized by certain transcription factors (TF). There are a lot of different TF binding sites in accordance with the great variety of different transcription factors. Presently for human, more than 1200 TFs are collected in the TRANSFAC database. As an example, a collection of binding sites for AhR factors are shown in the Figure 1. The function of DNA binding sites is the specific binding of TFs and their tethering in a particular orientation relative to the other components of multi-protein complexes. Binding sites for the same TF can be found in regulatory regions of a number of different genes and therefore could be considered as standard modules constituting a regulatory region. But these sites may differ in one or several positions, which makes the problem of recognizing such sites by computer programs extremely difficult.

Fig. 1

2) Regulatory modules of the next, second, hierarchical level are composite regulatory elements. The term "composite element" was introduced while studying the glucocorticoid response element in the mouse proliferin promoter where glucocorticoid receptor binding site is adjacent to an AP-1 site (Diamond *et al.*, 1990). Further, this term was applied to quite different pairs of interacting sites and factors (Gutman and Wasylyk, 1990; Jackson *et al.*, 1993; Du *et al.*, 1993; Rooney *et al.*, 1995, and others). Based on the known examples, we define a composite element as a minimal functional unit within which both protein-DNA and protein-protein interactions contribute to a highly specific pattern of gene transcriptional regulation (Kel, O.V. *et al.*, 1995b, 1997). A specialized database TRANSCompel (Kel-Margoulis, O.V. *et al.*, 2000, 2002) collects information about known composite elements.

Binding of a transcription factor to the regulatory region is determined not only by the structure of the cis-element, but also by the possible protein-protein interactions; in other words, by its ability to specifically contact -- whether directly or indirectly -- the factors binding to other sites of the given regulatory region. Two or three closely situated binding sites for different TFs in combination form a composite element, which is a functional unit with new regulatory advantages. Structurally similar elements are present in quite different genes, which apparently implies that such regulatory modules are functionally significant.

Fig. 2

Composite elements are composed of the modules of the previous hierarchical level, the individual binding sites (Figure 2). Similar binding sites may constitute parts of functionally different CEs. For instance, AP-1 binding sites are parts of AP1/ETS composite elements, NFAT/AP1, NF-kappaB/AP1, AP1/Oct. CEs of AP1/ETS type provide gene activation in response to the variety of proliferative signals and constitute so-called Ras- and oncogene response units (Wasylyk *et al.*, 1993). NFAT/AP1 composite elements provide cross coupling of Ca²⁺-dependent and Ras/Raf/MEK signaling pathways (Rao *et al.*, 1997). NF-kappaB/AP1 CEs contribute to gene activation in response to hypoxia. In turn, TFs of the ETS family are shown to cooperate with CEs not only with AP-1 factors, but also with a variety of different TFs, for example, CBFalpha, SRF, Sp1, ETS, among others. Factors of the NF-kappaB family can be found with CEs with C/EBP, STAT, Sp1, IRF, HMG I, among others.

At this hierarchical level, potential variability is considerably higher compared to individual binding sites. Within the set of similar CEs, several parameters could vary: individual binding sites themselves, sequence and distance between individual sites, and in some cases mutual orientation of individual sites. For instance, CEs of AP1/ETS type are different in terms of mutual location of individual sites (Figure 2, compare N 1 and 2). Individual binding sites could be immediately adjacent (Figure 2, N 1), overlapping (Figure 2, N 4), or separated by several nucleotides (Figure 2, N 3). Thus, the composite regulatory elements, as modules of the second hierarchical level, display some new functions that cannot be provided by individual binding sites. New functions resulting from protein-protein interactions, along with DNA-protein interactions, are the following:

- stabilization of DNA-protein complex through direct or indirect interactions between the corresponding TFs (Chen *et al.*, 1998, and others);
- cross-coupling of intracellular signal transduction pathways and as a result, new functions in gene transcription regulation (to be considered in detail later).

3) The next level in the hierarchical organization of gene regulatory regions is formed by promoters, enhancers, and distal regulatory regions. The structural similarity between promoters and enhancers is that both are composed of the modules of the previous hierarchical levels, composite elements and individual binding sites for TFs. A crucial difference between promoters and enhancers is that basal promoter elements, TATA-box, Inr-element and some others are responsible for the formation of the basal transcription complex, the precise definition of start point and the direction of transcription. Enhancers and distal regulatory regions provide modulation of the rate of transcription initiation - in many cases-, tissue-specific or inducible regulation, as well as some other features. Within eukaryotic genes, enhancers could be located differently:

- immediately upstream from promoters, for example, in the human interleukin-2 gene, human apolipoprotein A1 gene, among others;
- in the far upstream regions: for example, the T-cell specific inducible enhancer in the human GM-CSF gene is located at -3500 bp, and in the human IL-3 gene - at -13,000 bp.
- in the introns: for instance, the cell cycle-dependent enhancer of the human *pcna p120* gene is located in the first intron. The liver-specific enhancer of the human apolipoprotein B gene is situated in the first intron, and a liver/small intestine specific enhancer in the second intron. A p53-inducible enhancer is located in the third intron of the human GADD45 gene.
- in 3' gene flanks, for example in the human and mouse IgH genes.

Promoters and enhancers are formed by several modules of the previous hierarchical levels: composite elements and individual binding sites. A great number of possible parameters is required to describe the structure of the promoters and enhancers: number and set of individual binding sites and composite elements, their variations, mutual location and orientation, as well as the distance between them. The consequence of the variability is that each promoter or enhancer is practically unique. Along with that, some of the promoters or enhancers are similar in terms of sets of individual binding sites.

4) Finally, the highest level of hierarchy is represented by the integrity of all regulatory regions of a gene. On the basis of the known examples, the following functions of the integral system of gene regulatory regions can be considered:

- determining the local chromatin structure which influences the accessibility of particular DNA regions by TFs;
- overall control of the transcription at all stages including initiation, elongation and termination;
- contribution to the 3-dimensional DNA structure;
- providing a unique expression pattern for each gene;
- providing coordinated gene expression.

3. Databases on gene regulation

Despite the fact that gene regulation is one of the focuses of functional genomics, there are currently only a few databases that store data on gene regulation. The reason is that the molecular mechanisms of gene regulation appear to be very complex, which leads to a wide variety of different experimental techniques used for analysis of gene regulation. Laborious manual annotation of scientific literature is

needed to systematize all this data and store it in a computer readable form. Recently appearing mass data coming from micro array gene expression experiments is very useful for understanding gene regulation and can be more easily organized in a database. But this phenomenological data only partially fulfils the needs of the biological community working on functional genomics, since data regarding the details of molecular mechanisms of gene regulation are necessary to understand the causality of the observed gene expression.

Common nucleotide sequence databases and genomic databases such as GenBank and EMBL, Ensembl and RefSeq contain some pieces of information related to the regulation of gene expression. First of all, some types of gene regulatory sequences such as promoters, enhancers, LCR, S/MARs, translation: 5'UTR, 3'UTRs, translation enhancers, and TF binding sites are represented in EMBL and GeneBank. But this information is very sporadic, suffers from non-uniformity in the format of description and is sometimes even contradictory. This information is not the main focus of these databases and is based on practically uncontrolled annotation. A very important source of information on gene regulation is the protein databases PIR and Swiss-Prot. Structural as well as functional description of transcription factors can be found in these databases. These two protein databases also contain quite valuable information about tissue-specific expression of many genes.

Recently appearing genomic databases such as Ensembl (<http://www.ensembl.org/>) and UCSC genome browser (<http://genome.ucsc.edu>) put a great deal of emphasis on store structural information about genes. Users of these resources can obtain any kind of functional information, including information on gene regulation, through the links to the corresponding entries in the sequence databases described above.

There are three major groups of specialized databases that deal with gene regulation. The first group comprises databases that store information on regulatory sequences, including gene regulatory regions, promoters, information about transcription factors and their binding sites. The most recognized databases in this group are EPD and TRANSFAC. Another group of databases considers regulatory networks and signal transduction pathways. Among them, the most popular databases are CSNDB and TRANSPATH. The third group includes databases that store information about gene expression that is based primarily on micro array data.

Two ontology databases: GO (Gene Ontology)(??) and CYTOMER[®] (??) are becoming very important sources of information in the studies of gene regulation. Using GO genes can be classified into many different functional categories, thus providing a way to study correlation between gene expression and gene function. CYTOMER[®] is a database on physiological systems, developmental stages, anatomical structures and substructures, and their constituting cell-types for particular organisms (Chen et al., 1999; Fricke et al., 2001).

3.1. TRANSFAC database

The TRANSFAC[®] database on eukaryotic transcriptional regulation consists of data on transcription factors, their target genes and regulatory binding sites. At the core of the database is the interaction of transcription factors (FACTOR table) with their DNA-binding sites (SITE table), through which they regulate their target genes (GENE table). Apart from

genomic sites, 'artificial' sites, which are synthesized in the laboratory without any known connection to a gene, e.g., random oligonucleotides, and IUPAC consensus sequences are also stored in the SITE table. Sites must be experimentally proven to be included in the database. Experimental evidence for the interaction with a factor is given in the SITE entry in the form of the method that was used (gel shift, footprinting analysis,...) and the cell from which the factor was derived (factor source). Based on these, method and cell, a quality value is given to describe the 'confidence' with which an observed DNA-binding activity could be assigned to a specific factor. From a collection of binding sites for a factor, nucleotide weight matrices are derived (MATRIX table).

According to their DNA-binding domain, transcription factors are assigned to a certain class (CLASS). In addition to the more 'planar' CLASS table, a hierarchical factor classification system was proposed, as was also done some time ago (Wingender, 1997). Table 2 shows the number of entries in the different tables/flat files for the release. TRANSFAC[®] contains data from a wide variety of eukaryotic organisms, ranging from human to yeast.

4. Regulatory sequence analysis tools and approaches.

The achievements in developing rich databases that collect various information on gene regulation accelerated the development of computer programs for the analysis of gene regulatory sequences. The ultimate goal of the computational approaches is to reconstruct *in silico* the structural organization of the regulatory genomic regions by analyzing only the DNA sequence, thus providing computer methods for predicting the full spectrum of their regulatory functions. All hierarchical elements of these regions (single TF binding sites, composite elements, promoters and enhancers, long regulatory regions, S/MARs, LCRs and others) are to be categorized and modeled so as to be able to recognize them in genomic sequences. This allows us to make reasonable *in silico* predictions of the regulatory function of the sequences under study.

Over several years many computational tools for the analysis of regulatory sequences were developed (Bucher, 1999; Fickett & Wasserman, 2000)). Among them are rather simple tools as well as quite sophisticated systems that employ powerful statistical and machine learning algorithms. In all of these approaches two major strategies are used which we call "top-down" and "bottom-up" strategies. In the "top-down" strategy researchers analyze the structure of higher hierarchical elements such as promoters or full regulatory regions without making strong *a priori* hypothesis about their internal structure ("black-box" approach). In such strategies oligonucleotide frequencies as well as many other calculable parameters of DNA sequences are used in the analysis, without focusing on the biological and physical properties of such parameters. In the "bottom-up" strategy, researchers are making models of the complex units using the lower hierarchical elements as building blocks. The structure of promoters is analyzed on the basis of their elements: specific TF binding sites and other signals. The top-down strategy is useful during the initial steps of the analysis when there is limited knowledge about the internal structure of the regulatory units. The bottom-up strategy makes sense in the knowledge-rich areas and it is the next logical step of the analysis of regulatory genomic regions. Let us consider some of these approaches in more detail.

4.1. Motif analysis

The first methods that appeared for analysis of regulatory sequences were different methods of motif analysis that belong to the “top-down” strategy.

As described in the previous section, regulatory regions in genomes are not uniform in their internal structures. They contain DNA signals, most of which are short contiguous stretches of nucleotides that serve a particular type of regulatory function, such as target sites for binding of transcription factors, DNA bending sites of a particular type, heteroduplex formation signals, regions of Z and H forms of DNA, etc.. Signals of the same type can be grouped into an ideal pattern called motif. It means that the observed signals in sequences are different instances of the given motif.

There are several “languages” to describe motifs:

1) A motif is described by a *consensus* sequence of a length l , which contains the most frequent nucleotide in each position of the observed signals. The particular number (or percentage) of positions k that can mismatch with the consensus are typically given. Such (l,k) models for describing motifs are suitable for many different DNA signals, but they usually fail to describe properly the TF binding sites, where different nucleotide positions have totally different degrees of variability and could not be described by a single parameter k .

2) A motif is described by a single sequence that allows for several letters at each position in the motif. For example, the sequence AWCTTB describes a motif that allows letters A and T at the second position ($W = A/T$) and letters T, G or C at the last position ($B = \text{not A}$). Sometimes this description is complemented by the mismatch parameter k as in the first description, which might be different for the exact nucleotides and for the ambiguous ones.

3) A more biologically relevant representation of a motif is a probability matrix that assigns a different probability to each possible letter at each position in the motif (Schneider et al., 1986).

4.1.1. Motif finding algorithms (pattern and sequence driven)

The goal of motif finding, given a set of sequences, is to identify new motifs that are common for all or most of the sequences in the set. These new motifs are believed to correspond to some *a priori* unknown DNA signals that are important for the function of the given set of sequences.

There are two major strategies in motif discovery algorithms. The first is called the pattern-driven (PD) approach. It looks through all possible motif representations in a given solution space and finds the one that fits best to the set of sequences being analyzed. The second is called the sequence-driven approach (SD), which comprises algorithms that compare sequences to each other, trying to find local similarities between them to build motifs (Brazma et al, 1997). The PD algorithms are able to find the most optimal motifs, but they are slow and only practical for motifs of very limited size. The SD algorithms are fast, but do not guarantee finding the optimal motifs. Both types of algorithms are used for analysis of regulatory genomic sequences. Let us consider some of them a bit more in detail.

There have been many approaches developed for automatic motif discovery that apply SD and PD strategies as well as combinations of them. Among the best performing are Gibbs sampler (Lawrence et al., 1993), MEME (Bailey & Elkan, 1995), CONSENSUS (Hertz & Stormo, 1999),

PROJECTION (Buhler & Tompa, 2002), combinatorial approaches (Pevzner & Sze, 2000) and many others.

Pioneering work in this field was carried out by Gribskov et al. ((1990)). A scoring matrix called profile was introduced in this work. This method was applied first for finding motifs in protein sequences in the form of weight matrices. This approach was further developed and applied to the regulatory DNA sequences as well. Stormo and Hartzell used a greedy algorithm (1989) that was later improved by applying EM (expectation maximization) (Lawrence and Reilly, 1990).

Still later, an iterative Gibbs sampling algorithm (Lawrence et al., 1993) was introduced, which became the most frequently used algorithm for searching motifs. It can discover multiple motifs, but the number of occurrences of each motif in each sequence in the dataset must be specified in the input. There are many other algorithms developed for revealing multiple motifs using the PD approach, such as MOTIF (Smith et al., 1990), which searches for words consisting of three letters separated by a certain distance. Another algorithm was developed which discovers multiple motifs by clustering words of length k , which have at least r matches (Saqi and Sternberg, 1994; similar to Smith et al., 1990). Clustering is applied to the most frequently occurring patterns in order to combine related ones and obtain a reduced set of motifs. The exhaustive PD approaches were used for analyzing DNA motifs in promoter sequences. Van Helden et al. (1998) compared the frequency of conserved words in a given set of promoters to the frequencies in a reference set, thus revealing promoter-specific motifs. This approach was further developed by Kielbasa et al. (2001). In the recent work of Sinha (2002) the search for overrepresented motifs in a "positive" set of sequences (typically, promoters of co-regulated genes) versus a "negative" set of sequences is supported by a powerful formalism of computing p-values for the motifs. Kel et al. (1998) proposed another approach of finding overrepresented motifs using genetic algorithms (GA)).

4.1.2. *Heterogeneity. Search for multiple motifs.*

Ideally, every transcription factor is characterized by a specific DNA motif implicated in its target sites. In reality such a unique motif is difficult to find and sometimes does not exist. First of all, in vivo transcription factors bind to DNA in a complex with other factors, co-activators, and nucleosome components, and these complexes may be very different for different places in the genome. So, the local interaction environments impose specific constraints that certainly influence the DNA binding specificity of the transcription factor. So, we should talk about an in vivo TF motif that can be rather sparse or even split into several complex -specific motifs for the same transcription factor.

Moreover, in reality, sets of binding sites that can be retrieved from a database or collected from the literature typically comprise sites for a certain family of transcription factors rather than for a single TF. These factors makes it very important for a computational method to be able to reveal subsets of sites from a mixture of many local subtypes.

Recently we have developed a new powerful algorithm to search for multiple motifs in one set using Kernel functions (Tikunov & Kel, 2002). The method is able to reveal several motifs (in the form of weight matrices) in a set of unaligned sequences. Every weight matrix characterizes a pattern that can be found in a significant subset of sequences under analysis. The comparison to CONSENSUS

and Gibbs sampling shows that the Kernel method is clearly superior in identifying several motifs from noisy data (Kel et al., 2003).

The high sensitivity of the Kernel method results from the fact that the estimation of sequence distribution probabilities is mainly built on the basis of sequences located near the consensus, whereas all other methods estimate the background probabilities based on the complete set of sequences.

Heterogeneity of TF binding sites was taken into account in several other approaches, always yielding to the better recognition accuracy (Kel et al., 1995b; Shelest et al, 200??).

4.2. Recognition of TF binding sites

There are several thousands of different transcription factors functioning in human cells. More than 800 of them are well studied and characterized in databases such as SWISS-PROT and TRANSFAC. Even though TRANSFAC contains information about more than 5000 genomic sites (see above; about 1500 in human genes), it is far from being complete. Taking into account 33,000 genes in human genome and the fact that every gene might have up to 100 functioning sites in all regulatory regions including promoters, enhancers and far upstream regulatory regions, we could expect millions of sites in human genome. Knowledge about the number, the sequence and the position of all these sites in genome will bring us to a whole new level of understanding of how genes are regulated during development and functioning in the organism and how genes are deregulated in the cases of diseases.

Computer analysis of genome sequences provides the means for predicting binding sites for different transcription factors. Most transcription factors can be characterized by a specific DNA motif that is common for most of their binding sites. Given a known motif, there are many different methods for searching the motif in DNA sequences.

4.2.1. Search by consensus/pattern.

In an early stage of analysis of binding sites for a new transcription factor, only a few examples of known sites for this factor are available. These known examples are used as patterns in the search for new potential binding sites for the considered transcription factors. Since information on the variability of the positions in the pattern is extremely limited, the simple pattern search methods usually allow for a certain percentage of mismatches in any position of the pattern. While more examples of the known sites become available, **consensi** are created that describe the observed variability of nucleotides in different positions of the sites using the ambiguous letter code (e.g., the letter S marks a position where both nucleotides G or C are equally observed, W corresponds to either A or T, and so on.). Pattern search tools such as SIGNAL SCAN (Prestridge, 1991; Prestridge & Stormo, 1993) or PatSearch (Wingender et al., 1996), or SITE (Solovyev & Rogozin, 1986) can use consensi in the search and assign different mismatch penalties to the different letters of the ambiguous code. Pattern search methods are quite useful for finding new potential sites, but they are rather inaccurate and characterized by relatively high false negative as well as false positive rates.

4.2.2. **Weight matrices. MatInspector, Match, and other programs.**

Weight matrices are used to describe highly degenerate TF binding sites. The basis of any weight matrix is the counts for the observed nucleotides in the corresponding position of known sites. To build a reliable weight matrix, one needs a collection of a few known binding sites for a certain transcription factor. The most up-to-date collection of weight matrices is presented in the TRANSFAC database www.biobase.de. Several weight matrix-based search programs were developed, e.g. ConInspector (Frech et al., 1993), MATRIX SEARCH (Chen et al., 1995) or MatInspector (Quandt et al., 1995), Match (Gössling et al., 2001) and TRANSPLOERER (www.biobase.de). Different programs use different formalisms for the calculation of weight matrices from the nucleotide counts based on thermodynamic considerations: (Berg & von Hippel, 1987; Berg & von Hippel, 1988), information theory (Stormo, 1998); applying pseudo-counts (Henikoff & Henikoff, 1996) and forbidden nucleotides (Tronche et al., 1997). The search algorithms are greatly speeded up by using hash tables (Goessling et al., 2001). These tools were applied intensively over the last several years for the analysis of regulatory regions of many different functional classes of genes. Among them are globin genes (Hardison R. et al., 1997), muscle and liver specific genes (Wasserman & Fickett, 1998), genes involved in the regulation of cell cycle (Kel et al., 2001) and many others.

4.2.3. **Match algorithm.**

The algorithm uses two score values: the matrix similarity score (weight) and the core similarity score (Goessling et al., 2001), which resembles the algorithm previously published in "Quandt et al., 1995". The matrix similarity score denotes the quality of a match between the sequence and the whole matrix, whereas the core similarity score is a weight for the quality of a match between the sequence and the matrix core (the five most conserved consecutive positions in a matrix). Both scores range from 0 to 1, where 1 denotes an exact match.

The main steps of the algorithm are:

1. For each matrix all possible core matches in sequence s are identified. Since the length of a core is defined as 5, all subsequences x of the length 5 within the sequence s are found.
2. For each of these subsequences, the start position in the sequence s and the core similarity score are stored in a table.
3. For each entry with a core similarity score higher than a certain cut-off, each occurrence of this subsequence is looked up in the sequence s and is extended at both ends so that it fits the matrix length. Then the matrix similarity score is calculated and those matches with a matrix similarity score higher than a certain cut-off are shown in the program output.

The score for the matrix similarity of a subsequence x of the sequence s of length L is calculated in the following way (Kel et al., 1999):

$$mat_sim = mat_sim(x) = \frac{Current - Min}{Max - Min} \quad (7)$$

$$\text{here, } Current = \sum_{i=1}^L I(i) f_{i,b_i} ; \text{ Min} = \sum_{i=1}^L I(i) f_i^{\min} ; \text{ Max} = \sum_{i=1}^L I(i) f_i^{\max} ;$$

where, $f_{i,B}$: frequency of nucleotide B to occur at the position i of the matrix

($B \in \{A, C, G, T\}$); f_i^{\min} and f_i^{\max} are the frequency of the nucleotide that is the rarest and the most frequent in the position i in the matrix;

The information vector $I(i) = \sum_{B \in \{A, T, G, C\}} f_{i,B} \ln(f_{i,B})$ describes the conservation of the positions i in a

matrix (Quandt et al., 1995). Multiplication of the frequencies with the information vector leads to a higher acceptance of mismatches in less conserved regions, whereas mismatches in highly conserved regions are reduced.

The core similarity score is calculated in the same way as the matrix similarity score, but only the 5 nucleotide core part of the matrix is taken into account.

We call the collection of known binding sites which is used for building a weight matrix a "training set" of positive examples. We need a "test set" of negative examples (or control set) for estimation of the rate of false positives (FP), in other words, to estimate how often the considered matrix will wrongly predict that a site is a binding site for the transcription factor when in reality it is not. We often use a set of second and third exons as a test set of negative examples. Alternatively, different kinds of computationally randomized sequences are used as such a set. Of course, we cannot guarantee that all predicted sites in such sequences are negative, but usually we believe that number of real sites in the sequences is minimal.

Ideally, in addition to the training set, a separate set of sites for the same transcription factor should be used for estimating the potential rate of false negatives (FN). We call it "test set" of positive examples. Estimations of FN rate that are done on the basis of the training set are usually more "optimistic" and far from reality compared with estimations using an independent test set. Unfortunately, often the number of known examples is very limited and all of them are used for building the matrix, so it is impossible to create a separate test set.

Different kinds of bootstrap methods are used to overcome this problem by reusing randomly selected subsets of sites from the training set. A special kind of bootstrap is the so called "Jack-knife" method, which is most often applied for the evaluation of weight matrices. For a set of N sites we sample N -times a new subset, which contains all but one sequence from the original set. A new weight matrix is built on the basis of this subset. This new matrix is applied to the site that was left out and was not used for the matrix construction. The results of N corresponding recognitions are averaged, so that the FN rate is estimated.

The application of the Jack-knife method produces more realistic results, although often it gives estimations that are a bit too "pessimistic", since the matrix is always changing during the test, thus adding some additional variability to the method.

Matrix methods have been quite successful in the recognition of binding sites of many transcription factors such as CREB, AP-1, E2F (Kel et al., 1999; Kel et al., 2001). Nevertheless it is still quite weak in the recognition of target sites for some other factors that are characterized by a very

vague motif or by a complex motif that contains several conserved modules in a variable distance from each other (for example, NF-1 and C/EBP sites, sites for some nuclear receptors).

4.2.4. *TRANSPLOERER.*

TRANSPLOERER (TRANSCRIPTION exPLOERER) is a software package for the analysis of transcription regulatory sequences. It includes a tool for the prediction of potential binding sites for transcription factors in any sequence that may be of interest. Currently, the TRANSPLOERER site prediction tool uses collections of position weight matrices (PWM). It is able to use several matrix sources: the largest and most up-to-date library of matrices derived from TRANSFAC® Professional database ("<http://www.biobase.de/pages/products/databases>"), but other matrix libraries as well as any user-developed matrix libraries can be invoked as well. This means that it provides an opportunity to search for a great variety of different transcription factor binding sites. A search can be made using all of the matrices or subsets of matrices from the libraries.

Fig. 3

TRANSPLOERER has an advanced user interface (see Figure 3) and comes with a great number of filtering options, which allow you to specify which kinds of sites you want to see in the program output. You can view the results in a table view or in an elaborate graphical representation, in which specific color schemes are helpful for visualizing different kinds of information about the sites found. TRANSPLOERER provides a variety of options, allowing you to adjust the program output to your personal preference.

As an additional feature, TRANSPLOERER allows you to specify your search by using profiles. The term "profile" is used for a specific subset of matrices from one or several libraries with optimized cut-offs for each matrix. TRANSPLOERER provides a tool for creating (editing, deleting) such matrix profiles. In addition, TRANSPLOERER itself offers a number of optimized tissue-specific profiles.

The cut-offs in TRANSPLOERER™ are thresholds that the scores of a match must exceed to be shown in the results output. The appropriate cut-off selection for TRANSPLOERER™ depends largely on the user's objectives. We have pre-calculated three different cut-offs for each matrix presented in the library:

- 1) to minimize the false positive (over prediction error) rate,
- 2) to minimize the false negative (under prediction error) rate,
- 3) to minimize the sum of both errors.

4.2.5. *Correlation between positions. Dinucleotide weight matrices.*

In the classical weight matrix method individual positions are considered as independent and their contribution to the binding energy as additive (Berg & von Hippel, 1987). Both these assumption are not quite correct for the binding sites of transcription factors. To deal with the non-additive effect of

different positions the information measure was suggested that assigns higher weights to the more conserved positions (Quandt et al., 1995).

To consider the correlation between neighboring site positions, Hidden Markov Model (HMM) approaches were applied (Ehret et al., 2001). HMMs calculate the probability of nucleotides in a certain position depending on the preceding nucleotide(s). Another approach is based on di- and trinucleotide matrices that consist of frequencies of all doublets or triplets occurring at a given position (Kondrakhin et al., 1994). Long distance correlation between nucleotides in non-neighboring positions of sites can also contribute to the recognition of sites (Kel et al., 1995a).

Dinucleotide weight matrices were applied for the recognition of a vast number of transcription factor binding sites (Kel et al., 1995b). In most cases the dinucleotide matrices showed sufficiently better recognition accuracy than mononucleotide matrices constructed under the same conditions. Dinucleotide matrices are able to take into account the correlation among neighboring site positions, thus providing better recognition. Further confirmation of the importance of correlation between neighboring positions in TF binding sites was given recently by experimental work with the use of oligonucleotide micro arrays (Bulyk et al, 2002).

4.2.6. Influence of site flanks. Local context. SITEVIDEO system.

A weight matrix captures position-specific preferences in a short region only, which often corresponds to the most conserved part of the site, whereas the weight matrix does not cover regularities in the flanking regions of the sites. Therefore, new methods were developed that take flanking regions of the sites into account. The program ConsInspector makes a pair-wise alignment of a sequence to the set of sites that includes the core of the sites as well as the flanking regions (Frech et al., 1993). Similarities found in the flanks are used for supporting the site recognition. Various approaches of artificial intelligence are used to develop new methods for the recognition of cis-elements trying to reveal additional features in the flanking regions of sites. An application of the perception method was developed for the recognition of GREs (glucocorticoid regulatory elements; Seledtsov et al., 1991). It creates a model that includes 30bp of each flank of the GRE sites.

A pattern recognition software SITEVIDEO (Kel et al., 1993) was used to build recognition programs for TF binding sites (Kel et al, 1995b). SITEVIDEO system allows an analysis of a set of sites including their flanking regions and the design of high accuracy recognition methods by using a number of standard multi-dimensional discriminating methods and modern approaches from the field of artificial intellect such as perceptron, neural networks, and others (Minsky & Papert, 1969; Bolch & Huang, 1974). SITEVIDEO considers quantitative characteristics such as: a) statistical properties: characteristics related to the frequencies of mono- or oligonucleotides and their uneven location in the functional sites; b) physical properties: charge, stacking energy, mass, volume, polarity, hydrophathy; c) chemical properties: presence of certain atom groups in the nucleotides and certain other distinctive features of their chemical structure. In addition, a new approach to visual monitoring of recurrent design of the recognition methods is applied (Kel A.E. et al., 1993). Recently this technique was used for building a new recognition method for binding sites of E2F transcription factors – the key regulators of cell cycle. An exhaustive search for contextual motifs was made in the flanking regions of these

sites. Evaluating found vs. represented motifs, a “score of context” was defined that is used in addition to the weight matrix search. As a result, the new method of E2F site recognition provides a level of accuracy that is 5 times higher than the conventional weight matrix method (Kel et al., 2001).

Thus, many reliable methods have been developed that can be applied for the recognition of TF binding sites. Unfortunately, the best approaches usually need to be trained on valuable training samples, and therefore up to now could only be applied to a limited number of transcription factors for which many binding sites are already known.

Searches for individual TF binding sites in DNA sequences, though being widely used (for review see Frech et al., 1997), could hardly be applied directly to the characterization of transcriptional regulation of individual or sets of genes due to three main reasons: (i) poor recognition ability of most programs currently being used (high rate of false positives and false negatives); (ii) very incomplete lists of TF binding sites being searched; (iii) individual sites that have been found correctly often incompletely define the specificity of transcriptional regulation of a gene because of the combinatorial nature of regulation.

4.3. Recognition of composite regulatory elements

Combinatorial regulation is the basic mechanism of gene expression control in eukaryotic organisms. The pattern of expression of eukaryotic genes is encoded in the structure of their transcription regulatory regions mainly by the combination of transcription factor (TF) binding sites. Over the last several years, several computational approaches have appeared addressing the problem of combinatorial regulation of transcription. Specific TF binding site combinations were used for the identification of muscle-specific promoters (Wasserman & Fickett, 1998; Frech et al., 1998), of liver-enriched genes (Tronche et al., 1997) and for the characterization of yeast genes (Brazma et al., 1997).

As described in the introduction, the first and the most important level of this combinatorial regulation is provided by composite regulatory elements (CE) – combinations of target sites for two different transcription factors, that interact with each other resulting in a particular expression pattern common for genes that contain this CE (Kel O et al., 1995). A number of known examples of composite elements have been collected in the TRANSCompel database (Kel-Margoulis et al., 2001).

4.3.1. Motif finding techniques for analysis of composite elements.

The “*ab initio*” motif finding techniques, which do not take into account any previous knowledge about possible known motifs, often have problems in revealing TF binding sites that are “too weak”. These are instances of sites that differ significantly from their consensus while still serving as targets for the transcription factors. As described before, such “weak” sites can function due to synergism with other sites in composite elements. Usually the binding of transcription factors to such weak sites is stabilized by protein-protein interactions of this TF with other TFs that are binding to the closely located sites. Since the traditional motif finding algorithms usually find one (or a few) high scoring

patterns, they often fail to find composite elements consisting of pairs of weak TF sites. One or both sites in such pairs may not be statistically significant on their own. An example of such a composite element is shown in the Figure 4.

Fig. 4

One can see that in this composite element the AP-1 site differs very much from the canonical AP-1 consensus (shown below). It is clear that such site cannot be found alone.

Recently, a couple of new approaches appeared for revealing such composite motifs in the sets of sequences: BioProspector (Liu et al., 2001), Co-Bind (GuhaThakurta & Stormo, 2001), MITRA (Eskin & Pevzner, 2002). The first two are based on an extension of the Gibbs sampling techniques for finding significant motifs consisting of two modules that have some flexible distance between them. The algorithm maximizes the joint likelihood of co-occurrence of two motifs. The MITRA approach is the pattern-driven approach based on enumeration of l-mers. The algorithm uses a mismatch tree data structure to split the space of all possible patterns into disjoint subspaces that start with a given prefix. This way it avoids an explosion of the search space for the long composite motifs consisting of two parts.

All of these approaches prove their efficiency for some examples of composite motifs in yeast and bacterial genomes.

4.3.2. Matching algorithms for searching composite elements.

Several tools were designed for searching known composite elements in nucleotide sequences directly. With the program "FastM" (Frech et al., 1998) the user can specify a pair of weight matrices and set a preferable distance range between corresponding sites in the composite element. Another tool: "Catch" (www.biobase.de) searches the composite elements using the pattern-matching approach. Composite elements from the TRANSCompel database are used as patterns for the Catch program. Several parameters are available to restrict the search, such as maximal mismatches in the cores of site1 and site2 comprising the composite elements, maximal variation of the distance between the two sites (in percent), cut-off value for a general score for the CE. The CE score reflects how well the match coincides with the known example of the composite element in TRANSCompel. Score function takes into account the number of mismatches in both sites and the distance between them. All found matches are directly linked to the TRANSCompel entries containing the corresponding composite elements.

We have applied the Catch program for 5'-regions of genes expressed in activated T-cells (T-genes). For this analysis we have chosen a set of CEs that are situated in the regulatory regions of genes expressed in activated T-, B- and myeloid cells and collected in the TRANSCompel database. The set included the following types of CEs: AP-1/NF- κ B, AP-1/Oct, AP-1/NFAT, AP-1/Ets, NF- κ B/HMG, NF- κ B/IRF, C/EBP-a / AML, C/EBP-a / PU.1, Ets/AML. The frequency of the potential CEs in the 5'-regions of T-genes is 3 times higher than in the random sequences with the same nucleotide composition (Table 3). (Kel-Margoulis et al., 2000)

Table 3

It is clear that the list of known composite elements is far from complete. To reveal new types of Ces, some statistical estimations were made for finding pairs of TF binding sites that are close to each other in a sequence and can participate in a composite element. Application of χ^2 statistics allows pairs of sites to be revealed that often can be found at a short distance from each other (Kel et al., 1995a). Using this tool, unknown pairs were revealed such as CREB/SP-1 and GATA/NF- κ B, which can be new potential composite regulatory element types.

4.3.3. Composite score.

A method called “composite score” was developed for revealing NFAT/AP-1 composite elements (Kel et al., 1999). It includes two matrices for two corresponding transcription factors. The range of allowed distances between matrix matches and their mutual orientation are taken into account, as well as the coordinate variation of the matrix scores for these two factors. A low score of one matrix is compensated by a high score of another matrix, thus providing an optimal binding energy for the protein-DNA complex on the composite element.

The set of 13 NFAT/AP-1 CEs was extracted from the COMPEL database release 2.1. For each CE we apply the Match program and compute two scores: q_{NFAT} and q_{AP-1} for the two corresponding binding sites constituting the composite element. From these scores two parameters: $\pi_{NFAT} = \log(1 - q_{NFAT})$ and $\pi_{AP-1} = \log(1 - q_{AP-1})$ are calculated, estimating the binding energy of these two factors to their binding sites. To model the synergistic binding of two factors to DNA, we combine two parameters: π_{NFATp} and π_{AP-1} , and design a method for recognition of composite elements. For combining these two parameters into one recognition function, we use the SITEVIDEO software (Kel et al., 1993), which provides a means for obtaining the best discrimination between a training set of CEs and control data (random sequences).

It was shown that identifying composite elements with this method is a very effective tool for predicting gene expression patterns. It has been demonstrated for promoters of genes highly induced upon immune response. It was shown that NFAT/AP-1 composite elements are found in high concentration in the promoters of genes that are induced upon immune cell activation. (Figure 5).

Fig. 5

Clusters of these composite elements provide a good landmark for identifying promoters of immune-specific genes. A number of genes potentially regulated through this mechanism were revealed by genome search and suggested for experimental verification (Kel et al., 1999).

4.4. Analysis of promoters.

Computer-assisted prediction of eukaryotic promoters is one of the most straightforward approaches to the analysis of transcription regulation using sequence data (Bucher, 1990).

4.4.1. Types of promoters. Core promoter recognition. (TATA-rich; TATA-less; composite core promoters; “null promoters”)

A promoter can be defined as a structural part of a gene that defines its transcription start point and mediates and controls the initiation of transcription. Promoter sequences may include a TATA box, the initiator region (Inr), upstream activating elements, and downstream elements. However, not all of these elements are always required. This means that the DNA pattern that defines a so-called core promoter (around -50 to +10) can vary significantly in different cases and can be divided into four classes. (reviewed in [1, 2, 3]??). Briefly, these classes comprise (i) core promoters which consist of a TATA-box only, which directs transcriptional initiation at a position about 30 bp downstream; (ii) core promoters which do not contain any TATA box (and, therefore, are referred to as TATA-less). In these promoters, the exact position of the transcriptional start point is controlled mainly by an initiator element Inr (Smale, 1994; and Smale, 1997); (iii) composite promoters consisting of both a TATA box and an initiator element; and (iv) so-called Null-promoters, which have neither a TATA box nor an initiator and rely exclusively on upstream and downstream promoter elements (PDE) (Smale, 1994; Novina and Roy, 1996).

4.4.2. Brief survey of promoter recognition programs

There are basically two different approaches applied so far for promoter recognition. The first approach is a purely *sequence-based* one making use of different statistics of nucleotides and oligonucleotides in the promoter sequences without considering known promoter signals and cis-elements. Specific features of the distribution of nucleotides and oligonucleotides of different lengths (di-, tri- and longer oligonucleotides) are revealed in the training set of known promoters and these features are then used to build recognition procedures. In most of the applications that use the sequence-based approach the authors did not pay too much attention to the biological meaning of the found high/low frequent oligonucleotides and their distribution in promoter sequences. We call this approach "top-down" (see above).

In contrast to this, in the *signal-based* approach (bottom-up), authors build their tools on the basis of features and signals (such as TATA, CAAT boxes, consensi of TF binding sites, and physical and chemical properties of DNA) that are known to have high biological significance for promoter structure.

Both approaches have their merits. Using known signals makes much more sense than the simple oligonucleotide counting since it can capture crucial features of promoter structure that may be dissolved by the total oligonucleotide approach. On the other hand, oligonucleotide analysis provides a general sketch of all possible signals that might still be unknown and therefore could be missed by procedures that are based on the known signals only.

In Table 4 we survey a number of promoter recognition programs currently available.

Testing of the promoter prediction tools has demonstrated that nearly all of the available tools have a rather low level of recognition accuracy (specificity vs. sensitivity) (Fickett & Hatzigeorgiou,

1997). The accuracy of some more recently developed tools, however, is much higher, first of all because of larger training sets, better signal databases and the implication of new powerful techniques of machine learning and pattern recognition. But the recognition accuracy is still relatively low, so that these programs can hardly be used for direct annotation of a genome without additional information. It became obvious that prediction of a “general promoter structure“ is not just a difficult task but also a misleading one, and that each promoter should be described on the basis of its specific composition of regulatory elements. Addressing more specifically the biological features of promoters, the most promising trend is to identify specific promoter structures that are common for a group of functionally related promoters (e.g. tissue-specific promoters). Identification of composite regulatory elements is now very important for the recognition of promoters and the prediction of their structure.

4.5. Functional classification of promoters and prediction of gene regulation

4.5.1. Functional classification based on combinations of binding sites

Over the last several years, several computational approaches have appeared addressing the problem of combinatorial regulation of transcription. Specific TF binding site combinations were used for the identification of muscle-specific promoters (Wasserman & Fickett, 1998; Frech et al., 1998) for liver-enriched genes (Tronche et al., 1997) and for yeast genes (Brazma et al., 1997) and for immune specific genes (Kel et al., 1999) (see above). Promoters of genes regulated during cell cycle could be recognized by the combination of E2F binding sites with a dozen additional oligonucleotide motifs (Kel et al., 2001). It becomes clear that TF site combinations provide a key to functional classification of promoters and to the annotation of regulatory regions in genomes.

4.5.2. Decision tree

For automatic annotation of genomic regulatory regions, methods of classification of promoters to different functional groups are necessary.

We have developed a decision tree classifier of the 7 functional classes of promoters using combinations of TF binding sites (Kel-Margoulis et al., 2002). The bottom nodes of the tree contain 7 different promoter classes. The training set of 212 promoters described above was used for optimizing the decision tree structure with the help of genetic algorithms. The topology of the decision tree obtained in the analysis is shown in Figure 6.

Fig. 6.

The following set of TF binding sites appeared to be the most effective for classification of the mentioned sets of promoters: E2F, Oct-1, NF-AT, MyoD, SRF and ER.

The percentages of the promoters correctly classified by the tree are shown below each bottom node. One can see that cell cycle-related and erythroid-specific promoters are classified best (65 – 70% of correct classifications). In contrast, promoters of housekeeping genes and brain-enriched genes are the most difficult to classify (34% and 20% of correct classifications correspondingly). It is

known that these two classes contain genes with very heterogeneous function and expression. More effort should be made in the initial grouping of promoters into functionally unified classes.

4.5.3. Clusters of sites - Composite clusters

It is known that most TF target sites are located in 5'-regions of genes. We assume that binding sites for transcription factors that bind together to a regulatory region of a gene tend to be co-localized in a relatively short region inside the 5'-regulatory region in order to provide the possibility for protein-protein interactions between these factors. Therefore, it is expected that such sites for many different factors are clustered in 5'-regulatory regions that we call "composite clusters" (or "hetero-clusters"). The presence of such composite clusters in genomic sequences might be a good indication of regulatory regions of genes. The structure of composite clusters can tell us a lot about regulatory mechanisms.

Clusters of binding sites for the same transcription factors (homo-clusters) are believed to help increase the probability of binding these factors to their target regions in genome. Statistical estimations made by Karlin (Karlin & Macken, 1991) are used for revealing statistically significant homo-clusters of TF binding sites in nucleotide sequences (Kel et al., 1999). There is an option implemented in the "Match" tool for revealing such clusters. By applying this tool, clusters of E2F binding sites were found in the sequences of human chromosomes 21 and 22 (Kel-Margoulis et al., 2001b).

A new statistical technique recently appeared that is suitable for the identification of statistically significant homo-clusters as well as hetero-clusters (Wagner, 1999). The technique was applied on two transcription factors, Mcm1 and Stel2, involved in cell cycle and mating control of the yeast *Saccharomyces cerevisiae*. Clusters of binding sites for these factors were revealed in the yeast genome (Wagner, 1999).

A tool called "Cister" was developed for revealing hetero-clusters (Frith et al, 2001). It uses a Hidden Markov Model (HMM) based method for searching for clusters of cis-elements (TFs). The found clusters are considered to be landmarks for detecting promoters and other regulatory regions in DNA sequences. According to the estimation given by the authors, the program achieves a sensitivity of promoter predictions of 67%, while making one prediction per 33 kb of non-repetitive human genomic sequences. A web interface is available at <http://sullivan.bu.edu/~mfrith/cister.shtml>. The user can search for site clusters in a sequence and adjust search parameters.

To reveal site clusters it is absolutely necessary to set correct cut-off parameters for searching TF sites by the weight matrices. It is known that some sites that are part of composite elements often differ significantly from the consensus. In such cases the lack of binding energy for a transcription factor is compensated by protein-protein interactions with other factors. To find such cryptic sites as parts of hetero-clusters (or "composite clusters"), a program called ClusterScan was developed (Kel et al., 2001b). It applies a genetic algorithm that is able to find optimal parameters for searching composite clusters of TF sites in regulatory genomic sequences. The composition of the identified clusters can tell a lot about regulatory mechanisms and provide a means for functional classification of

regulatory regions and for automatic annotation of regulatory regions in genomes (Kel-Margoulis et al., 2001b)

4.6. Phylogenetic Footprinting

Phylogenetic Footprinting is a new approach to revealing potential transcription factor binding sites in promoter sequences. The idea is based on the assumption that functional sites in promoters should evolve much slower than other regions that do not carry any conservative function. Therefore, potential transcription factor (TF) binding sites that are found in the evolutionarily conserved regions of promoters have a higher chance of being considered “real” sites.

The global comparison of human and mouse genomes now becomes real and there are several groups doing this systematically. One of the available resources of the global comparison of human and mouse genomes is Berkeley Genome Pipeline (<http://pipeline.lbl.gov/>). They use a system called VISTA (Mayor et al., 2000) for handling the global alignment and for revealing so-called conservative non-coding sequences (CNS) that are good landmarks on genome for finding functionally important promoters, enhancers or silencers (Duret & Bucher, 1997).

The most difficult step of the Phylogenetic Footprinting is the alignment of promoter sequences between different organisms. The conventional alignment methods often cannot align promoters due to the high level of sequence variability. We have developed a new alignment method that takes into account the similarity in distribution of potential binding sites (Cheremushkin & Kel, 2002) <http://compel.bionet.nsc.ru/FunSite/footprint/>.

This method has been used effectively for the alignment of human/mouse CNS revealed by the Berkeley Genome Pipeline. New potential binding sites for various transcription factors were revealed. Binding sites for the transcription factors that belong to the same family having overlapping locations on the alignment are considered to be the positive match at the phylogenetic footprint. The list of 17117 CNS of the total length of alignment 2418267 bp was analyzed. We applied a set of 240 weight matrices from TRANSFAC rel. 5.3 with the cut-offs optimized to minimize the sum of false positive and false negative errors. We found 58106 conserved TF binding sites.

5. Analysis of gene expression data

Functionally related genes involved in the same molecular genetic, biochemical, or physiological process are often regulated coordinately. This coordination is maintained by transcription factors which, after being activated, become able to switch on or off a number of target genes. This is done through binding these TFs to their binding sites in the regulatory regions of the coordinately expressed genes. Therefore, it is tempting to reveal TF binding sites which regularly appear in the regulatory regions of co-expressed genes. Our knowledge about the functioning of the corresponding transcription factors can then be used for generating working hypotheses about the molecular mechanisms of coordinate gene regulation.

5.1. Analysis of the promoters of co-regulated genes

Mass data on gene expression coming from the micro array experiments provide valuable information for deducing gene regulatory mechanisms. Groups of co-expressed genes can be revealed from this data using various clustering techniques . A number of techniques are used for the analysis of regulatory regions of the co-expressed genes in the clusters. First of all are the motif-finding algorithms that search for oligonucleotide motifs which are overrepresented in the promoters of the co-expressed genes. The found motifs can correspond to the binding sites for transcription factors. Approaches that have been described above are: Gibbs sampling, MEME, Consensus, ClustalW, and AlignACE and all are applied for this tasks. Another approach is to search for potential TF binding sites using a collection of known weight matrices. Finally, the most promising techniques searching for specific combinations of TF binding sites that correlate with gene expression patterns have to be applied as well.

Recently, an analysis of pairs of TF binding sites in correlation with gene expression data was done for yeast genes (Pilpel et al., 2001). The authors searched for so called "synergistic" pairs of TF sites when the expression coherence score (EC) of genes containing both sites in their promoters was significantly greater than that of genes containing either motif alone. The EC score for a set of K genes is defined as p/P , where $P = 0.5 * K * (K - 1)$ is the number of all gene pairs and p is the number of gene pairs that have very similar expression patterns (similar expression values in all measured conditions). A number of synergistic pairs were revealed that regulate expression of yeast genes in cell cycle and sporulation (Pilpel et al, 2001). A general motif synergy map was proposed that shows a network of functional interactions between different pairs of TFs.

ClusterScan (Kel et al., 2001b) provides the means for analyzing more complex combinations of TF binding sites. There are several new methods that appeared recently for revealing composite modules in the sets of co-regulated promoters (Frith et al., 2002; Hannenhalli & Levy, 2002; Sharan et al, 2003). They are based on different statistics. Let's consider in a bit more detail the most effective method that is based on genetic algorithm (Kel et al, 2003 submitted).

5.1.1. Genetic algorithm to determine composite regulatory modules (CM)

We define a composite module CM as a set of TF weight matrices with given matrix cut-offs and other parameters, which is associated with a specific functional type of gene regulatory regions. We have developed a new computational method to determine CMs in a set of promoter (or other regulatory) sequences of co-expressed genes. This method is based on a *genetic algorithm* (a prototype of this method is realized in the tool package ClusterScan (Kel et al., 2001b). The CMs are characterized by the following parameters: K – number of PWMs in the module (typically 6 to 12). The program selects these K matrices from a library of all considered matrices. We use different profiles including the profile vertebrate_minFN62.prf, which includes 410 matrices for different transcription factors of vertebrate organisms (TRANSFAC® rel. 6.4). A certain cut-off value $q_{cut-off}^{(k)}$, relative importance value $\phi^{(k)}$ and maximum number of best matches $\kappa^{(k)}$ are assigned to every weight matrix k ($k=1, K$) in the CM. Some matrices are organized in pairs. A parameter R is defined that puts a limit on the

distance between matches of these matrix pairs (at least one pair of matches should be found fitting this limit). When all of these parameters are defined, we can calculate a “composite module score” (CM score) for any sequence X using the following equation:

$$F_{CM}(X) = \sum_{k=1, K} \phi^{(k)} \times \sum_{i=1}^{\kappa^{(k)}} q_i^{(k)}(X) \quad (2)$$

, where $q_i^{(k)}(X)$ are the $\kappa^{(k)}$ best scoring sites found in the sequence X by the matrix (k). An implementation of the *genetic algorithm* is used to determine the parameters of CMs that are specific for a particular set of promoters. The general description of *genetic algorithms* is available elsewhere (see, ??).

We define the goal function G as a weighted sum of false negative and false positive errors and the value of T-test that are calculated over several random iterations of bootstrap procedures of splitting of the initial set into a training and testing subset. In addition, we test the normal-likeness of the distributions of the F function over the set of positive and negative sequences. This algorithm for calculating the goal function gives us evidence for the usability of the obtained solutions for classification of individual sequences.

The program realization of the method is called CompositeModuleFinder (cmf). It takes as input two sets of sequences (the set which is analyzed and a background set) and a set of weight matrices for transcription factors. For defined parameters K and R, over a number of iterations, the program optimizes the set of matrices selected, their quantity, their cut-offs, the relative importance and the maximum number of best matches. The user can vary parameters K and R and compare results of the program. The output of the program is a profile ready to run by Match™ or TRANSPLOER® (see Methods section).

5.1.2. Analysis of promoters of Ah-receptor regulated genes.

The aryl hydrocarbon receptor (AhR) is a ligand-activated nuclear transcription factor that mediates responses to a variety of toxins. Among them are halogenated aromatic toxins such as 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), polynuclear aromatic hydrocarbons, combustion products, and numerous phytochemicals such as flavonoids and indole-3-carbinol (I3C). Experimental data obtained by RT-PCR on AhR-responsive genes were investigated with the CompositeModuleFinder to reveal transcription factor binding sites and their specific combinations related to AhR-responsiveness (Kel et al., 2003, submitted)

A number of matrices in various combinations were revealed in the 5' regions of these genes. Among the most discriminative matrices were those for the following transcription factors: HNF1, AhR, GR, OCT, C/EBP and some others. Among the most prominent matrix pairs selected by the algorithm are: HNF-1/Sp-1 and AP-1/NF-1 for the max. distance R=100bp; E2F/NF-1, AhR/Myb, HNF3/NFY, HNF6/NF-kappaB, and Sp-1/Myb for the max. distance R=50bp and HNF-1/GR for the maximum distance R=40bp. It was interesting to observe that not all matrices found individually were also found in pairs and that not only discriminating individual TFs appear in pairs. This gives us an additional idea of the composite structure of the promoters under study. In Table 5 we present a combination of

individual matrices and matrix pairs that give one of the best discriminations of the set of promoters of AhR regulated genes (-2000 +2000) from other promoters. We call this set of matrices AhR-associated composite regulatory modules CM_{AhR} of the promoters under study.

Fig. 7.

As can be seen in Figure 7, the distribution of the composite module score F_{CM} is clearly different in these two sets of promoters (T-test value is 12,12, p value = $1.7 \cdot 10^{-28}$). The AhR-responsive promoters have clearly higher values of F_{CM} score than the background promoters.

Acknowledgments

The authors are indebted to Volker Matys (BIOBASE GmbH) and Yuri Tikunov (Institut Cytology and Genetics, Novosibirsk), Susanne Reymann (Fraunhofer Institute (Fh-ITEM)) for their great contribution to the preparation of the manuscript; Aida G. Romashchenko and Vadim A. Ratner (Institut Cytology and Genetics, Novosibirsk) for fruitful discussion of the results presented in the manuscript. Parts of this work were supported by the Siberian Branch of Russian Academy of Sciences, by a grant of the European Commission (BIO4-95-0226), by grant of Volkswagen-Stiftung (I/75941) and by grant BioProfile Braunschweig/Göttingen/Hannover (0313092).

References

1. Bailey T.L. & Elkan C. (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, **21**, 51.
2. Bajic V.B., Seah S.H., Chong A., Zhang G., Koh J.L.Y. and Brusica V. (2002). Dragon Promoter Finder: recognition of vertebrate RNA Polymerase II promoters. *Bioinformatics*, **18**, 198-199.
3. Bajic VB, Seah SH. (2003) Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes. *Nucleic Acids Res.* **31**,3560-3563.
4. Berg, O.G. & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical Theory and Application to Operators and Promoters. *J. Mol. Biol.* **193**, 723-750.
5. Berg, O. G., von Hippel, P. H. (1988). Selection of DNA binding sites by regulatory proteins. II. The Binding Specificity of cyclic AMP Receptor Protein to Recognition Sites. *J. Mol. Biol.* **200**, 709-723.
6. Bucher P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol.* **212**, 563-278.
7. Bucher P. (1999) Regulatory elements and expression profiles. *Curr Opin Struct Biol.* **9**,400-407.
8. Bolch B.W. & Huang C.J. (1974) Multivariate statistical methods for business and economics, Prentice-Hall,Inc., Unglued Cliffs, N.J..
9. Brazma A., Jonassen I., Eidhammer I., Gilbert D. (1997) Approaches to the automatic discovery of patterns in biosequences. Technical Report (accepted for publication in the Journal of Computational Biology), Department of Informatics, University of Bergen, TR-113, 1995, Bergen, Norway, 42 pp. (available <http://industry.ebi.ac.uk/~brazma/Papers/survey.ps>.)
10. Brazma, A., Vilo, J. & Ukkonen, E. (1997b) Finding Transcription Factor Binding Site Combinations in the Yeast Genome. In *Proceedings of the German Conference on*

- Bioinformatics* GCB'97, Kloster Irsee, Bavaria, Sept. 22-24, 1997 (H.W.Mewes and D.Frishman eds.), 57-60.
11. Buhler J, Tompa M. (2002) Finding motifs using random projections. *J Comput Biol* **9**, 225-242.
 12. Bulyk M.L., Johnson P.L., Church G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* **30**, 1255-1261.
 13. Chen Q.K., Hertz G.Z., Stormo G.D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput.Applic. Biosci.*, **11**, 563-566.
 14. Chen L., Glover J.N., Hogan P.G., Rao A., and Harrison S.C. (1998) Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. *Nature* **392**, 42-48.
 15. Chen,X., Dress,A., Karas,H., Reuter,I. and Wingender,E. (1999) A database framework for mapping expression patterns. In Proceedings of the German Conference on Bioinformatics GCB'99. Hannover, Germany, pp. 174–178
 16. Cheremushkin E., Kel A. (2002) PromoterFootprint: A new method for alignment of regulatory genomic sequences. Phylogenetic footprinting of TF binding sites. In Liliana Florea, Brian Walenz, Sridhar Hannenhalli (eds) *Currents in Computational Molecular Biology 2002. RECOMB 2002*, Washington D.C., pp. 40-41.
 17. Davuluri R.V., Grosse I., Zhang M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genetics*, **29**, 412-417.
 18. Diamond M.I., Miner J.N., Yoshinaga S.K., and Yamamoto K.R. (1990) Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science*. **249**, 1266-1272.
 19. Du W., Thanos D., and Maniatis T. (1993) Mechanisms of transcriptional synergism between distinct virus- inducible enhancer elements. *Cell* **74**, 887-898.
 20. Duret, L. & Bucher, P. (1997). Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7**, 399-406.
 21. Dynan W.S. (1989) Modularity in promoters and enhancers. *Cell* **58**, 1-4.
 22. Ehret G.B., Reichenbach P., Schindler U., Horvath C.M., Fritz S., Nabholz M., Bucher P. (2001) *J Biol Chem* **276**, 6675-88.
 23. Fickett J.W. & Hatzigeorgiou A.G. (1997). Eukaryotic promoter recognition. *Genome Res.*, **7**, 861-878.
 24. Fickett J.W., Wasserman W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr Opin Biotechnol.* **11**, 19-24.
 25. Frech K., Herrmann G. and Werner T. (1993) Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res.*, **21**, 3117-3118.
 26. Frech K, Quandt K, Werner T. (1997) Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem Sci.* **22**,103-104.
 27. Frech, K., Quandt, K., Werner, T. (1998) Muscle actin genes: A first step towards computational classification of tissue specific promoters. *In Silico Biology* **1**, 0005, <http://www.bioinfo.de/isb/1998/01/0005/>.
 28. Fricke,E., Land,S., Rotert,S., Karas,D. and Wingender,E. (2001) Cytomer: A database on gene expression sources. Proceedings of the German Conference on Bioinformatics GCB'01. Braunschweig, Germany, pp. 149–151.
 29. Frith M.C., Hansen U., Weng Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA, *Bioinformatics* **17**, 878-889.
 30. Frith M.C., Spouge J.L., Hansen U., Weng Z. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.* **30**, 3214-24.
 31. Goessling E., Kel-Margoulis O.V., Kel A.E. and Wingender E. (2001) MATCHTM - a tool for searching transcription factor binding sites in DNA sequences. Application for the analysis of

- human chromosomes. In: *Proceedings of the German Conference on Bioinformatics (GCB2001)*, October 7-10, 2001, Braunschweig, pp.158-160.
32. Ghosh,D. (1991) *TIBS* **16**, 445-447.
 33. Gribskov M., Luthy R., and Eisenberg D. (1990) Profile analysis. *Methods in Enzymology*, **183**, 146-159.
 34. GuhaThakurta D., Stormo G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics* **17**, 608-621.
 35. Gutman A. and Wasylyk B. (1990) The collagenase gene promoter contains a TPA and oncogene-responsive unit encompassing the PEA3 and AP-1 binding sites. *EMBO J.* **9**. 2241-2246.
 36. Hannenhalli S, Levy S. (2002) Predicting transcription factor synergism. *Nucleic Acids Res.* **30**, 4278-4284.
 37. Hardison R.C., Oeltjen J., Miller W. (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* **7**, 959-66
 38. van Helden, J., Andre, B., and Collado-Vides, J., (1998) Extracting regulatory sites from the upstream regions of yeast genes by computational analysis of oligonucleotide frequencies. *J.Mol.Biol.* **281**, 827-842.
 39. Hertz G.Z., Stormo G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563-77.
 40. Jackson D.A., Rowader K.E., Stevens K., Jiang C., Milos P., and Zaret K.S. (1993) Modulation of liver-specific transcription by interactions between hepatocyte nuclear factor 3 and nuclear factor 1 binding DNA in close apposition. *Mol. Cell. Biol.* **13**, 2401-2410.
 41. Johnson P.F. and McKnight S.L. (1989) Eukaryotic transcriptional regulatory proteins. *Annu. Rev. Biochem.* **58**, 799-839.
 42. Karlin, S. & Macken, C. (1991). Assessment of inhomogeneities in an E.coli physical map. *Nucleic Acids Res.* **19**, 4241-4246.
 43. Kel A., Kel-Margoulis O., Babenko V., Wingender E. (1999) Recognition of NFATp/AP-1 Composite Elements within Genes Induced upon the Activation of Immune Cells *J. Mol. Biol.*, **288**, 353-376.
 44. Kel A.E., Kel-Margoulis O.V., Farnham p.J., Bartley S.M., Wingender E. and Zhang M.Q. (2001) Computer-assisted identification of cell-cycle related genes: New targets for E2F transcription factors. *J.Mol.Biol.* **309**, 99-120.
 45. Kel A., Kel-Margoulis O., Ivanova T., Wingender E., (2001b) ClusterScan: A Tool for Automatic Annotation of Genomic Regulatory Sequences by Searching for Composite Clusters. In: *Proceedings of the German Conference on Bioinformatics GCB 2001*, Braunschweig, Germany, October 7-10, 2001, p.96-101.
 46. Kel A.E., Kolchanov N.A., Kapitonov V.V., Ponomarenko M.P., Likhachev E.A., Lim H.A , Milanesi L. (1993b) Computer analysis and recognition of functional sites on the base of oligonucleotide patterns distributions. In: *Cantor, C.R. and Lim H.A. 8eds) Proc. of the Second International Conference on Electrophoresis, Supercomputing and the Human Genome, June 1992, St.Petersburg, Florida, USA.*, 521-544.
 47. Kel A., Kondrakhin Yu., Kolpakov Ph., Ponomarenko M., Wingender E., Kolchanov N., (1995) Computer analysis of the structure of transcription factor binding sites, *SAMS*, **18-19**: 819-822.
 48. Kel A.E., Kondrakhin Yu.V., Kolpakov Ph.A., Kel O.V., Romashenko A.G., Wingender E., Milanesi L., Kolchanov N.A. (1995b) Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences. In: *Proc. of the 3-d Intern. Conf. on Intelligent Systems for Molecular Biology, AAAIPress, California, 1995*, pp.197-205.
 49. Kel A.E., Ponomarenko M.P., Likhachev E.A., Orlov Yu.L., Ischenko I.V., Milanesi L., Kolchanov N.A. (1993) SITEVIDEO: A computer System for Functional site Analysis and Recognition. Investigation of the human splice sites. *Comput. Appl. Biosci.*, **9**, 617-627.

50. Kel A., Ptitsyn A., Babenko V., Meier-Ewert S., Lehrach H. (1998) A genetic algorithm for designing gene family-specific oligonucleotide sets used for hybridization: the G protein-coupled receptor protein superfamily *Bioinformatics* **14**, 259-270.
51. Kel, A., Tikunov Y., Voss N. and Wingender E. (2003) Recognition of multiple patterns in unaligned sets of sequences. Comparison of kernel clustering method with other methods. GCB, in press.
52. Kel O.V., Romaschenko A.G., Kel A.E., Wingender E. and Kolchanov N.A. (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res.* **23**, 4097-4103.
53. Kel,O.V., Kel,A.E., Romaschenko,A.G., Wingender,E., and Kolchanov,N.A. (1997) Composite regulatory elements: classification and description in the COMPEL database. *Mol. Biol. (Mosk)*, **31**, 498-512.
54. Kel-Margoulis,O.V., Romaschenko,A.G., Kolchanov,N.A., Wingender,E. and Kel,A.E. (2000) TRANSCompel: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res.* **28**, 311-315.
55. Kel-Margoulis O.V., Romaschenko A.G., Deineko I.V., Kolchanov N.A., Wingender E., Kel A.E., Database on composite regulatory elements in eukaryotic genes (COMPEL). (2000b) In *Proceedings of the second international conference on bioinformatics of gene regulation and structure BGRS 2000*, August 7-11, 2000, Novosibirsk, v.1, p.45-48.
56. Kel-Margoulis, O. (2001) Automatic annotation of the regulatory regions of cell cycle related genes on human chromosomes. In: "*Genome Sequencing & Biology*". Cold Spring Harbor Symposia, 2001
57. Kel-Margoulis O.V, Kel A.E., Reuter I., Deineko I.V., Wingender E. (2002a) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res* **30**, 332-334.
58. Kel-Margoulis O.V., Ivanova T.G., Wingender E., Kel A.E. (2002b) Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pac Symp Biocomput* 2002, 187-198
59. Kielbasa S.M., Korbel J.O., Beule D., Schuchhardt J. and Herzog H. (2001) Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics* **17**, 1019-1026.
60. Knudsen S., (1999) Promoter 2.0: for the recognition of PolII promoter sequences *Bioinformatics* **15**, 356-361.
61. Kolchanov N.A., Rogozin I.B., Kel A.E., Ponomarenko M.P., Lihachov J., and Milanesi L. (1991) In: *Kanehisa,M. (ed), Prossidings of Genome informatics workshop II . Japan*, pp 104-107.
62. Kondrakhin Yu.V., Shamin V.V., and Kolchanov N.A. (1994) Construction of a generalized consensus matrix for recognition of vertebrate pre-mRNA 3'-terminal processing sites. *Comput. Applic. Biosci.* **10**, 597-603.
63. Kondrakhin Yu.V., Kel A.E., Kolchanov N.A., Romashchenko A.G., Milanesi L. (1995) Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Applic. Biosci.* **11**, 477-488.
64. Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F., Wootton J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208-14.
65. Lawrence C. E. and Reilly A. A.. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *PROTEINS: Structure Function and Genetics* **7**, 41-51.
66. Liu X., Brutlag D.L., Liu J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*.127-138.
67. Mayor C., Brudno M., Schwartz J. R., Poliakov A., Rubin E. M., Frazer K. A., Pachter L. S. and Dubchak I. (2000) VISTA: Visualizing Global DNA Sequence Alignments of Arbitrary Length. *Bioinformatics* **16**, 1046.

68. Minsky M. & Papert S. (1969) *Perceptrons*, M.I.T. Press, Cambridge.
69. McKnight S. L., Yamamoto K. R. *Transcriptional Regulation*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1992.
70. Novina C.D., Roy A.L. (1996) Core promoters and transcriptional control. *Trends Genet* **12**, 351-355.
71. Ohler U., Niemann H., Liao G., and Rubin G.M. (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics* **17**, 199-206.
72. Pevzner P.A. & Sze S (2000). Combinatorial approaches to finding subtle signals in DNA sequences. *In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp 269-278.
73. Pilpel Y., Sudarsanam P., Church G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet.* **29**, 153-159.
74. Prestridge, D. S. & Stormo, G. (1993). SIGNAL SCAN 3.0: new database and program features. *Comput. Appl. Biosci* **9**, 113-115.
75. Prestridge D. S. (1995). Predicting pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* **249**, 923-932.
76. Ptitsyn A.A., Rogozin I.B., Grigorovich D.A., Strelets V.B., Kel A.E., Milanese L. and Kolchanov N.A. (1996) AutoGene: A computer system for nucleotide sequence analysis. *Mol. Biol. (Mosk.)*, **30**, 258-264.
77. Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995) *Nucleic Acids Res.*, **23**, 4878-4884.
78. Rao A., Luo C., and Hogan P.G. (1997) Transcription factors of the NFAT family: regulation and function. *Annu. Rev. Immunol.* **15**, 707-747.
79. Ratner V.A. (1990) Towards the Unified Theory of Molecular Evolution (TME). *Theor. Popul. Biol.*, **38**, 233-261.
80. Ratner V.A. (1992) *Genetica (Mosk)*, **28**, 5-24.
81. Reese, M.G., and Eeckman, F.H. (1995). Novel Neural Network Prediction Systems for Human Promoters and Splice Sites', *Proceedings of the Workshop on Gene-Finding and Gene Structure Prediction*, Pennsylvania, Philadelphia, edited by D. Searls, J. Fickett, G. Stormo and M. Noordewier.
82. Reuter, I. (2000), Dissertation, <<http://www.biblio.tu-bs.de/ediss/data/20000317a/20000317a.html>>
83. Rooney J.W., Hoey T., and Glimcher L.H. (1995) Coordinate and cooperative roles for NF-AT and AP-1 in the regulation of the murine IL-4 gene. *Immunity* **2**, 473-483.
84. Saqi M. A. and M. J. Sternberg. (1994) Identification of sequence motifs from a set of proteins with related function. *Protein Engineering*, **7**, 165-171.
85. Scherf, M., Klingenhoff, A., Werner, T. (2000) Highly Specific Localization of Promoter Regions in Large Genomic Sequences by PromoterInspector: A Novel Context Analysis Approach. *J. Mol. Biol.* **297**, 599-606.
86. Schneider T. D., Stormo G. D., Gold L., and Ehrenfeucht A.. (1986) Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, **188**, 415-431.
87. Seledtsov, I. A., Solovyev, V. V., Merkulova, T. I. (1991). New elements of glucocorticoid-receptor binding sites of hormone-regulated genes. *Biochim. Biophys. Acta* **1089**, 367-376.
88. Sharan R, Ovcharenko I, Ben-Hur A, Karp R.M. (2003) CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, Suppl 1, I283-I291.
89. Sinha S. (2002) Discriminative Motifs. In *RECOMB2002 Proceedings of the Sixth Annual International Conference on Computational Biology*, pages 291-298. Washington, DC, USA, April 18-21, 2002.

90. Smale S.T., (1994) Core promoter architecture for eukaryotic protein-coding genes. *In: Conaway, R.C. Conaway, J.W. (eds) Transcription: Mechanisms and regulation*. Raven Press, New York, 63-81
91. Smale, S.T. (1997) Transcription initiation from TATA-less promoters within eukaryotic protein-encoding genes. *Bioch. Biophys. Acta* **1351**, 73-88.
92. Smith H. O., Annau T. M., and Chandrasegaran S. (1990) Finding sequence motifs in groups of functionally related proteins. *Proc. Natl. Acad. Sci. USA* **87**, 826-830,.
93. Solovyev V.V. et al. (1997) *In: Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (Gaasterland, T., Karp P., Karpus K., Ouzounis C., Sander C. & Valencia A., eds)*, pp.294-302.
94. Solovyev V.V., Rogozin I.B. (1986) The program package of the context analysis of DNA, RNA and protein sequences 1. Search for homology and functional sites. Institute Cytology and Genetics of the USSR Academy of Science, Novosibirsk, (Russ), 1-70.
95. Stormo G.D. and Hartzell G.W. (1989) A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA* **86**, 1183-1187.
96. Stormo G.D. (1998) Information content and free energy in DNA--protein interactions. *J Theor Biol* **195**, 135-137.
97. Struhl K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98**, 1-4.
98. Tikunov Y., Kel A. (2000) Kernel method for estimation of functional site local consensi. Classification of transcription initiation sites in eukaryotic genes *In: Proceedings of the German Conference on Bioinformatics (GCB00)*, October 5-7, 2000, Heidelberg, p. 83-88
99. Tikunov Y., Kel A. (2002) Kernel method for identification of local patterns in unaligned sets of functional sites . *In The Third International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002)* (accepted).
100. Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M. & Pontoglio, M. (1997). Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.* **266**, 231-245.
101. Wagner A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **15**, 776-784.
102. Wasserman W.W. & Fickett J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J.Mol.Biol.* **278**, 167-181.
103. Wasylyk B., Hahn S.L., and Giovane A. (1993) The Ets family of transcription factors. *Eur. J. Biochem.* **211**, 7-18.
104. Werner, T. (1999). Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome* **10**, 168-175.
105. Wingender E. *Gene Regulation in Eukaryotes*. VCH, Weinheim, 1993.
106. Wingender, E. (1997) Classification scheme of eukaryotic transcription factors. *Mol. Biol. (Mosk)* **31**, 483-497.
107. Wingender E., Karas H. and Knueppel, R. (1996) TRANSFAC Database as a Bridge between Sequence Data Libraries and Biological Function. *Pacific Symposium on Biocomputing '97 (PSB'97)*, R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein (eds.). *World Scientific*, pp. 477-485.
108. Zhang M.Q. (1998) Identification of Human Gene Core Promoters in Silico, *Genome Res.* **8**, 319-326.
109. Zhang M.Q. (1998b) Statistical features of human exons and their flanking regions, *Hum.Mol.Genet* **7**, 919-932.

Figure captions

Fig. 1. A collection of binding sites for AhR transcription factors. The sequence of each site, accession number and site ID in TRANSFAC database, gene name and position of the site relative to the start of transcription are shown. The core part of the site is shown in bold.

Fig. 2. Composite regulatory elements as second level of hierarchical structure of gene regulatory regions

Fig. 3. The output window of the TRANSPLOERER. Three different windows represent three scales to see potential TF binding sites (central part). Sites for factors of different families are represented in different colours as defined by the user (right, middle). Numerous filter options allow to focus in the results of particular interest (right, top). Database links and details of single pieces of information can be enlarged selectively (right, bottom).

Fig. 4. An example of NF-AT/AP-1 composite element in the promoter of mouse interleukin-2 gene. The AP-1 site differs from the canonical AP-1 consensus.

Fig. 5. Frequencies of NFAT/AP-1 composite elements ($q_{CE} > 10.0$) in the functional parts of immune-cell specific genes, muscle-specific genes and random sequences.

Fig. 6. A decision tree for classification of promoters into 7 functional classes. To classify a new promoter, the sequence (x) is passed down the tree beginning at the top. If the functional score: $F(x) > F_{cut-off}$ the sequence is passed down to the left, otherwise to the right. The functions $F(x)$ and cut-offs were optimised by genetic algorithm.

Fig. 7. Discrimination between (-2000 +2000) promoters (red bars) and background promoters from chromosome 21 (PR:-2000 +2000) (blue bars) by the composite module CM_{AhR} . The composite module score F_{CM} is given on the abscissa.

Site sequence	acc	site ID	gene	from	to
cacgtg gcgtg tcttgt	R02649	MOUSE\$CYTOP_01	CYP1A1 (cytochrome P450 1A1)	-1227	-1146
cagctag gcgtg acagca	R02650	MOUSE\$CYTOP_02	CYP1A1 (cytochrome P450 1A1)	-1076	-1048
ggagtt gcgtg agaaga	R02651	MOUSE\$CYTOP_03	CYP1A1 (cytochrome P450 1A1)	-1066	-977
ccgaat gcgtg cgatcg	R02652	MOUSE\$CYTOP_04	CYP1A1 (cytochrome P450 1A1)	-933	-869
tgtctc gcgtg gatcct	R02653	MOUSE\$CYTOP_05	CYP1A1 (cytochrome P450 1A1)	-893	-641
aagctc gcgtg agaagc	R02654	MOUSE\$CYTOP_06	CYP1A1 (cytochrome P450 1A1)	-509	-448
gtcgagg gcgtg cgttcc	R13150	MOUSE\$CYP1B1_01	Cyp1B1 (cytochrome P-450 1B1)	-870	-841
cgctgg gcgtg cagatg	R13159	HS\$CYP1A1_01	CYP1A1 (cytochrome P450 1A1)	-401	-391
tagctt gcgtg cgccgg	R13161	HS\$CYP1A1_03	CYP1A1 (cytochrome P450 1A1)	-900	-890
ggcgtt gcgtg agaagg	R13162	HS\$CYP1A1_04	CYP1A1 (cytochrome P450 1A1)	-988	-978
cccctc gcgtg actgcg	R13163	HS\$CYP1A1_05	CYP1A1 (cytochrome P450 1A1)	-1061	-1051
cgagtt gcgtg agaaga	R00270	RAT\$CYTOP_04	CYP1A1 (cytochrome P450, 1a1)	-1029	-1005
ctgctc gcgtg agaagc	R13271	RABBIT\$CYP1A1_01	CYP1A1 (cytochrome P450 1A1)	-1012	-984
cggctc gcgtg ctgggg	R13226	MOUSE\$AhRR_01	AhRR (Aryl hydrocarbon receptor repressor)	-59	-55
gacttag gcgtg ttcctc	R13227	MOUSE\$AhRR_02	AhRR (Aryl hydrocarbon receptor repressor)	-393	-387
ttaaagg gcgtg agccgt	R13228	MOUSE\$AhRR_03	AhRR (Aryl hydrocarbon receptor repressor)	-1302	-1296
cggccg gcgtg cgccgg	R13260	HS\$CATHD_02	CATH-D (cathepsin D)	-130	-126
tgcctt gcgtg tttgtg	R13262	MOUSE\$POLK_01	Polk (polymerase (DNA directed), kappa)		
agagtt gcgtg ccccctt	R13263	MOUSE\$POLK_02	Polk (polymerase (DNA directed), kappa)		
gaatgt gcgtg acaagg	R13264	RAT\$UGT1A1_01	Ugt1 (UDP-glucuronosyltransferase 1 family, member 1)	-134	-129
ttatgt gcgtg gtgata	R13237	Mouse\$IL2_15	IL-2 (interleukin-2)	-860	-831
gcatgt gcgtg cacatg	R13238	Mouse\$IL2_16	IL-2 (interleukin-2)	-823	-794
aagttc gcgtg acgaag	R13240	MOUSE\$ALDH3A1_01	Aldh3a1 (aldehyde dehydrogenase 3a1)	-98	-74
tggggg gcgtg ggcacac	R13248	HS\$CFOS_28	c-fos	-1163	-1159
gaactc gcgtg cagcag	R13268	HS\$UGT1A6_01	UGT1A6 (UDP glycosyltransferase 1 family, polypeptide A6)	-1502	-1498
attacag gcgtg agccac	R13274	HS\$PS2_02	PS2	-530	-508

Fig. 1.

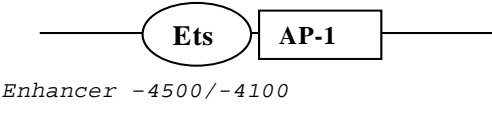
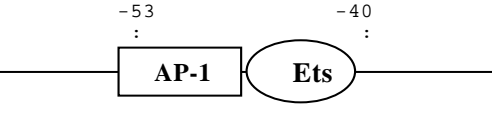
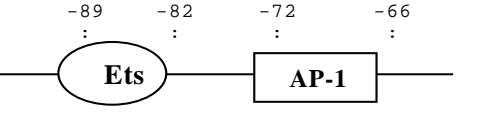
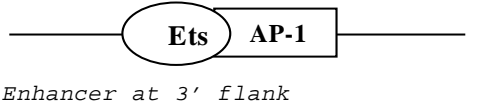
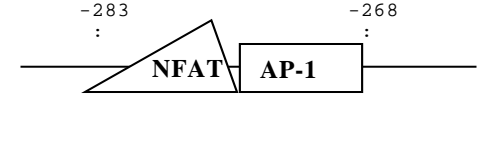
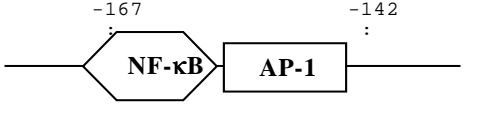
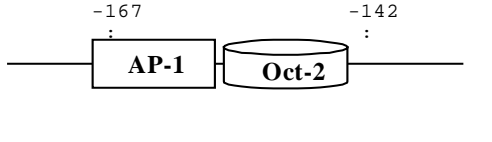
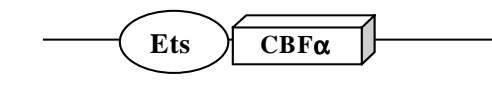
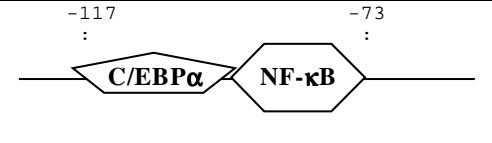
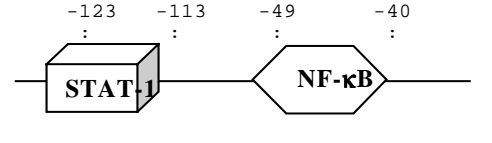
N	Gene	Schema and positions of a CE	TRANSCompel accession number
1.	Scavenger receptor, <i>Homo sapiens</i>		C00080
2.	GM-CSF, <i>Mus musculus</i>		C00081
3.	Collagenase, <i>Homo sapiens</i>		C00083
4.	IgH, <i>Mus musculus</i>		C00133
5.	Interleukin 2, <i>Homo sapiens</i>		C00109
6.	Interleukin 2, <i>Homo sapiens</i>		C00165
7.	Интерлейкин 2, <i>Mus musculus</i>		C00158
8.	IgH, <i>Homo sapiens</i>		C00173
9.	Serum amiloid A1, <i>Rattus norvegicus</i>		C00101
10.	IRF-1, <i>Mus musculus</i>		C00192

Fig. 2.

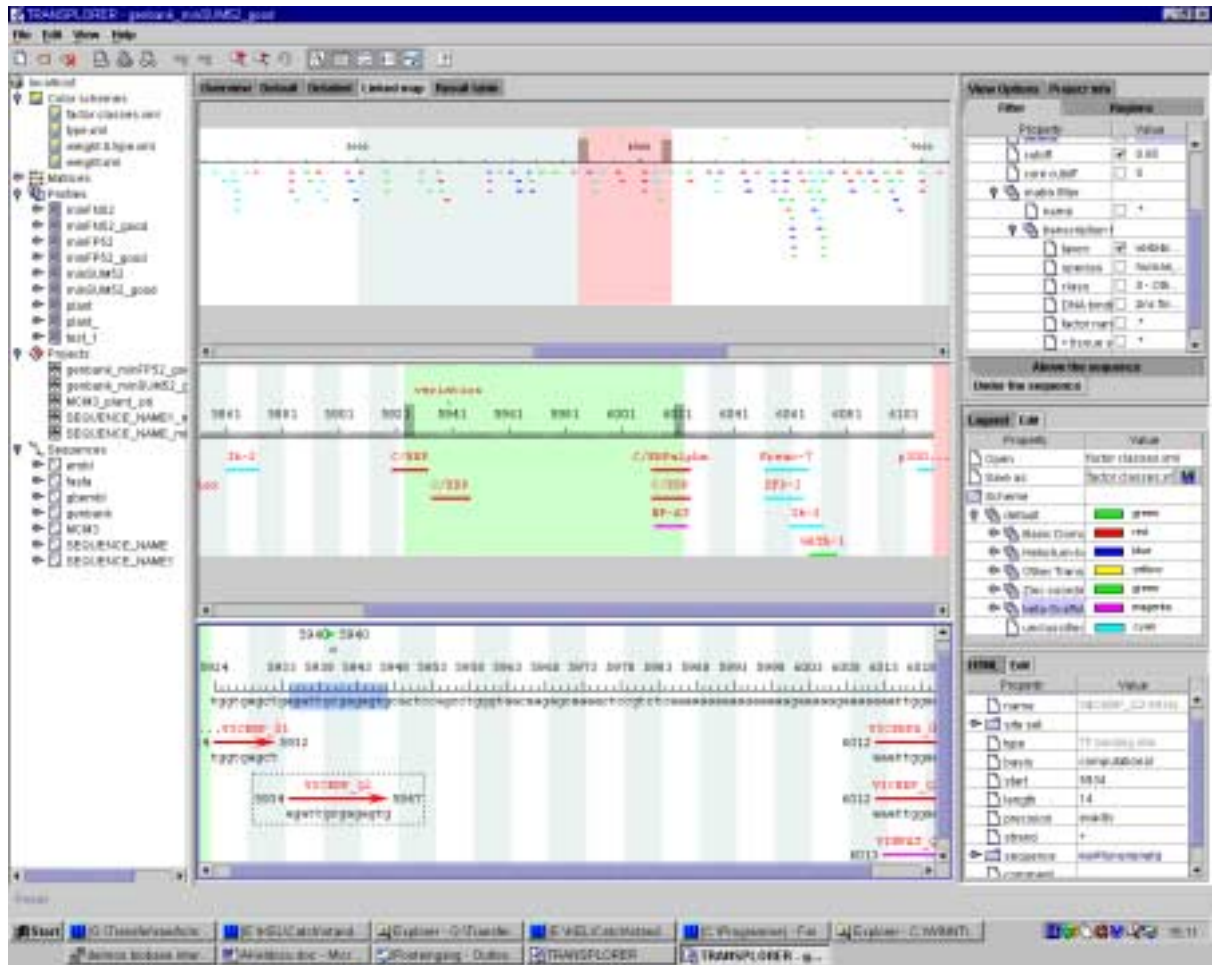


Fig. 3.

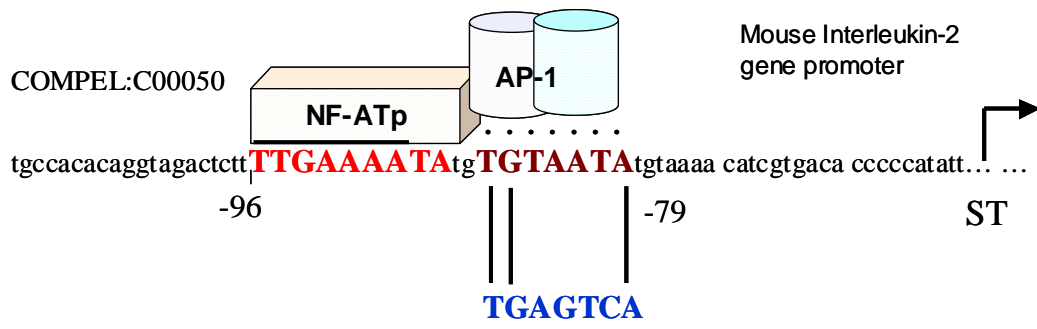


Fig. 4.

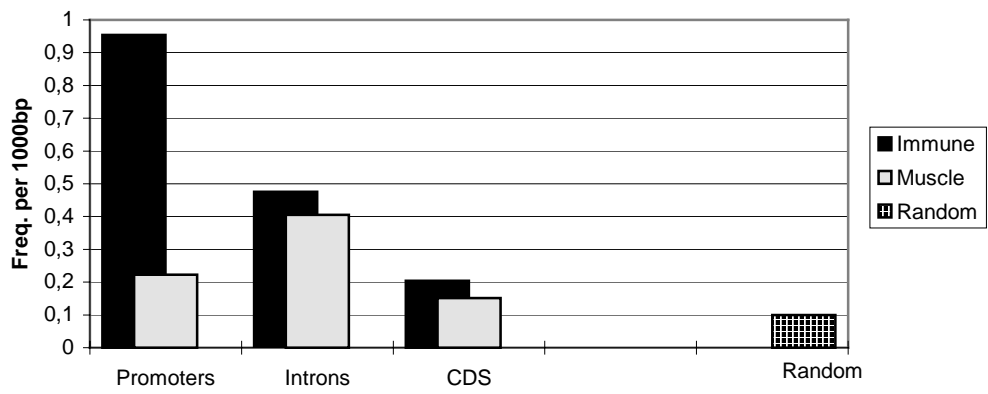


Fig. 5.

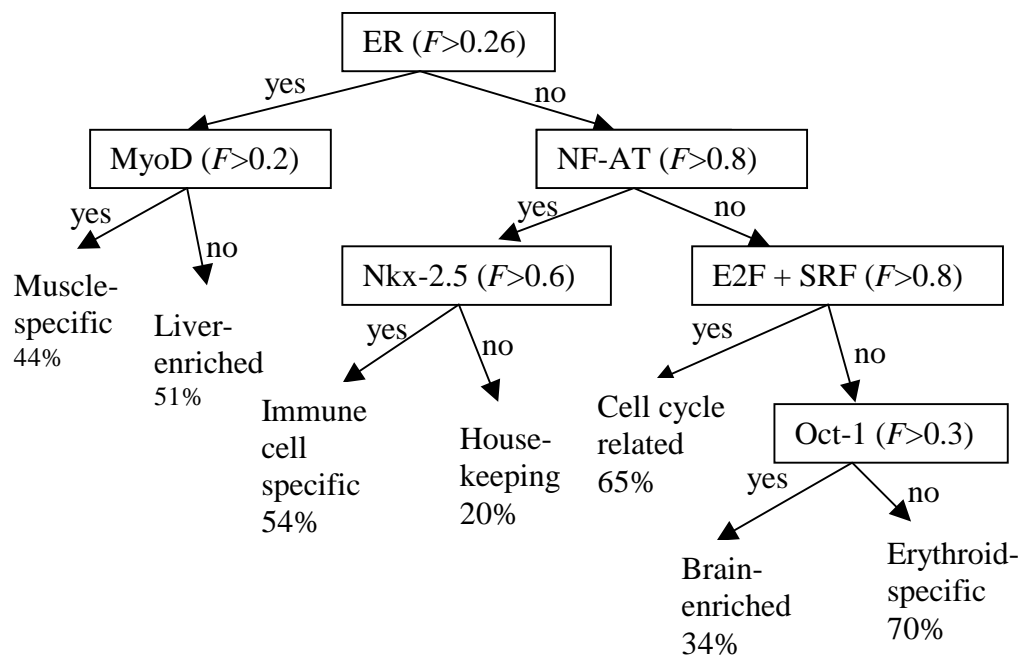


Fig. 6.

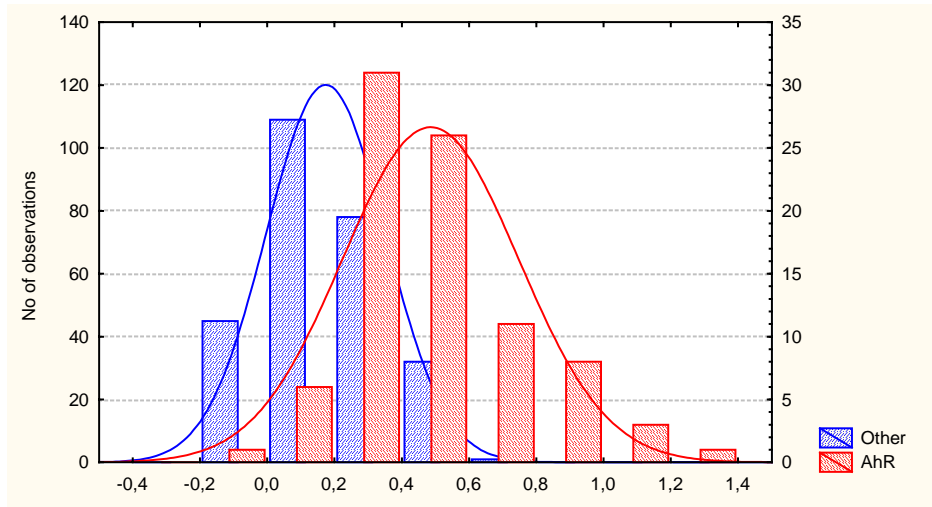


Fig. 7.

Table 1 Databases that have information on different aspects of gene regulation.

N	Database	Information on gene regulation	URL
1.	EMBL Nucleotide sequence database	For some gene there is information on location of transcription start site, TATA box, CAAT box and some TF binding sites.	http://www.ebi.ac.uk/embl.html
2.	GeneBank		http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html
3.	SWISS-PROT	Structure of transcription factors, domain structure, short functional description. For many proteins there is information on tissue-specific expression.	http://www.expasy.ch
4.	PIR: Protein Information Resource		http://www-nbrf.georgetown.edu/pir
6.	EPD - Eukaryotic promoter database	Location of starts of transcription; some information on gene expression; functional classification of gene products.	http://www.epd.isb-sib.ch
7.	DBTSS	Transcription start sites; genomic sequences (human and mouse); predicted TF binding sites.	http://dbtss.hgc.jp/index.html
8.	TRANSFAC	Transcription factors (TF), TF classification, their binding sites, weight matrices, structure of gene regulatory regions, gene expression..	http://www.biobase.de ; http://www.gene-regulation.com
9.	TRRD	Structure of transcription regulatory regions of genes; binding sites for TFs, gene expression.	http://www.bionet.nsc.ru/trrd/
10.	COMPEL, TRANSCompel	Composite regulatory elements; TF protein-protein interaction	http://compel.bionet.nsc.ru/ http://www.biobase.de
11.	TFD	Sites of TF binding; consenci	http://www.ifti.org/
12.	RegulonDB	Transcription regulation of E.coli gene. TFs and their binding sites, consenci, weight matrices.	http://www.cifn.unam.mx/Computational_Biology/regulondb
13.	PRODORIC	Information on prokaryotic gene expression; TFs; genomic TF binding sites; Regulatory networks	http://prodoric.tu-bs.de
14.	SCPD - The Promoter Database of <i>Saccharomyces cerevisiae</i>	Genes, promoters, TFs, sites, consenci, weight matrices	http://cgsigma.cshl.org/jian/
15.	Muscle-Specific Regulation of Transcription	Description of regulatory regions of muscle-specific genes; sites.	http://agave.humgen.upenn.edu/MTIR/HomePage.html
16.	EpoDB	Database of genes that relate to vertebrate red blood cells.	http://agave.humgen.upenn.edu/epodb/

17.	GeNet	Information on functional organization of regulatory genes networks acting at embryogenesis.	http://www.csa.ru/Inst/gorb_dep/inbios/genet/genet.htm
18.	PlantCARE	Transcription regulation in plants; regulatory cis-elements; TFs.	http://sphinx.rug.ac.be:8080/PlantCARE/
19.	PLACE		http://www.dna.affrc.go.jp/htdocs/PLACE/
20.	TRANSPATH	Information of molecules and reactions involved in signal transduction in the cell.	http://www.biobase.de
21.	GeneNet	Object-oriented databases that include information on a number of gene regulatory networks.	http://www.mgs.bionet.nsc.ru/systems/MGL/GeneNet/
22.	CSNDB	Cell Signaling Networks Database; Information on signal transduction reactions and signaling molecules.	http://geo.nihs.go.jp/csndb/
23.	SPAD: Signaling Pathway Database	Integrated database for genetic information and signal transduction systems.	http://www.grt.kyushu-u.ac.jp/spad/
24.	KEGG: the Kyoto Encyclopedia of Genes and Genomes	Information of some signaling molecules; graphical representation of several signal transduction networks.	http://www.genome.ad.jp/kegg/

Table 2. Total number of entries in each table of TRANSAC database.

Table	Release 7.2
<u>Factor</u>	5241
<u>Site</u>	12976
Gene	4571
<u>Matrix</u>	636
<u>Cell</u>	1592
<u>Class</u>	50
Method	87
Reference	9897
Journals	385

Table 3. Frequency of potential CEs that provide various aspects of lymphoid- and myeloid-restricted transcriptional regulation within different sequences.

Sequences under study	Frequency of potential CEs on 1000 bp
<i>T</i> -genes	1.636
5'regions of <i>T</i> -genes	2.894
Random [A]=[T]=[C]=[G]=0.25.	0.832
Random (nucleotide composition as in 5'-regions of <i>T</i> -genes) [A]=0.2865; [T]=0.2615; [G]=0.2196; [C]=0.2323	0.938

Table 4. Promoter recognition programs.

Program name	Description	Ref/URL
Autogen Promoter	The first program for recognition of eukaryotic promoters. Based on distinctive features of dinucleotide distribution in a sample of promoters from EPD database.	(Kel et al., 1993b)
PromFind	The algorithm operates by scoring the sequence using a differential hexamer measure. The author suggested to search for a peak score (independent of an absolute threshold) and has shown that it is more accurate than a search based upon a fixed scoring threshold.	(Hutchinson, 1996) (http://iubio.bio.indiana.edu/soft/molbio/mswin/mswin-or-dos/profin11.exe).
NNPP	Applies several neural networks for analysis of nucleotide composition of the core promoter region that includes TSS (transcription start site) and TATA box.	(Reese & Eeckman, 1995) http://www.fruitfly.org/seq_tools/promoter.html
Promoter 2.0	The neural network uses as input a window of DNA sequence, as well as the output of other neural networks. Genetic algorithms are used for optimisation of the weights in the neural networks to discriminate maximally between promoters and non-promoters	(Knudsen, 1999) http://www.cbs.dtu.dk/services/promoter
McPromoter V3.0	A neural network is used to combine features of nucleotide context, features describing CpG islands and some selected physico-chemical parameters of DNA.	(Ohler et al., 2001) http://promoter.informatik.uni-erlangen.de/
CorePromoter	Realizes the stepwise strategy based on initial localization of a functional promoter into 1- to 2-kb (extended promoter) region and further localization of a transcriptional start site into 50- to 100-bp (core promoter) region covering the interval -60 to +40. The method uses a position-dependent 5-tuple measure that is analyzed with the help of a quadratic discriminant analysis technique (QDA)	(Zhang, 1998) http://argon.cshl.org/genefinder/CPROMOTER/
PromoterInspector	The program is purely based on libraries of IUPAC words extracted from training sequences by an unsupervised learning approach.	(Scherf et al., 2000) http://genomatix.gsf.de/cgi-bin/promoterinspector/promoterinspector.pl
FirstEF	It uses a set of discriminant functions that can recognize structural and compositional features such as CpG islands, promoter regions and first splice-donor sites. The core of this algorithm is a decision tree that processes results of these discriminant functions.	(Davuluri et al., 2001) http://www.cshl.org/mzhanglab/
Dragon Promoter Finder	Identifies TSS positions using five independent promoter recognition models with Artificial Neuron Networks.	(Bajic et al., 2002) http://sdmc.lit.org.sg/promoter/ and www.biobase.de
Dragon Gene Start Finder	Currently the best performing program. Assesses the gene start in mammalian genomes and predicts a region which should overlap with the first exon of the gene or be in its proximity	(Bajic & Seah, 2003) http://sdmc.lit.org.sg/promoter/ and www.biobase.de

<u>PromoterScan</u>	TF binding sites were searched using a pattern matching algorithm based on the TFD database. The combined individual density ratios of all binding sites were then used to build a scoring profile. This profile, in combination with a weight matrix for TATA box, is used by the program to predict TSS locations.	(Prestridge,1995) http://bimas.dcrn.nih.gov/molbio/proscan/
TSSG and TSSW	The algorithm predicts potential transcription start positions by linear discriminant function combining characteristics that describe functional motifs and oligonucleotide composition of promoters.	(Solovyev et al., 1997) http://www.softberry.com
FunSiteP	Method is based on the localization of binding sites of transcription factors. Based on the fact that distribution of TF sites is uneven, the authors have constructed a weight matrix of binding site localization. FunSiteP recognizes promoters in nucleotide sequences and tentatively identifies the functional class the promoters must belong to (according Bucher's specifications)	(Kondrakhin et al. 1995) http://compel.bionet.nsc.ru/FunSite/fsp.html

Table 5 Composite module CM_{AhR} constructed for the (-2000 +2000) set of AhR-regulated promoters.

Factor or pair of factors (distance)	ϕ (see equation (4))
E2F	0,105086
OCT1	0,084289
GR	0,077050
YY1	-0,169821
IRF/SRY(50)	0,213636
HNF3/SRY(50)	0,164787
AP1/NF1(100)	0,149481
SP1/MYB(50)	0,138358
GR/HNF1(40)	0,137571
AHR/MYB(50)	0,124308
HNF1/SP1(100)	0,110810
E2F/ROR(100)	-0,115593
AHR/CREB(100)	-0,086788