Automatic Annotation of Genomic Regulatory Sequences by Searching for Composite Clusters

O.V. Kel-Margoulis, T.G. Ivanova, E. Wingender, and A.E. Kel

Pacific Symposium on Biocomputing 7:187-198 (2002)

AUTOMATIC ANNOTATION OF GENOMIC REGULATORY SEQUENCES BY SEARCHING FOR COMPOSITE CLUSTERS

O.V. KEL-MARGOULIS[1,2], T.G. IVANOVA[2], E.WINGENDER[3], A.E. KEL[1,2]

[1] *BIOBASE GmbH, Halchtersche Strasse 33, 38304 Wolfenbuettel, Germany;* [2]*Institute of Cytology & Genetics SB RAN, 10 Lavrentyev pr., 630090, Novosibirsk;* [3]*Research Group Bioinformatics, Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-38124 Braunschweig, Germany*

A new method was developed for revealing of composite clusters of cis-elements in promoters of eukaryotic genes that are functionally related or coexpressed. A software system "ClusterScan" have been created that enables: (i) to train system on representative samples of promoters to reveal cis-elements that tend to cluster; (ii) to train system on a number of samples of functionally related promoters to identify functionally coupled transcription factors; (iii) to provide tools for searching of this clusters in genomic sequences to identify and functionally characterize regulatory regions in genome. A number of training samples of different functional and structural groups of promoters were analysed. Search for composite clusters in human chromosomes 21 and 22 reveals a number of interesting examples. Finally, a decision tree system was constructed to classify promoters of several functionally related gene groups. The decision tree system enables to identify new promoters and computationally predict their possible function.

1. Introduction

Besides the fact that genomes of eukaryotic organisms contain rather limited number of genes (Ng) [1], the number of different intracellular molecular states (Ns) is enormously huge (Ns >>Ng). In multicellular organisms these are states of cellular ontogenesis in different tissues, organs and cell types, a number of developmental stages and cell cycle phases, the huge amount of influences of different external and internal signals. Every state is characterized and precisely organized by differential expression of specific sets of genes. Therefore it becomes obvious that most of the genes in genome are expressed in various cellular states (gene expression pattern), and it is a combination of active genes that is state specific (gene expression profile). For the majority of genes, transcription regulation plays the most important role in regulation of gene expression. 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. These multi-protein complexes are often

referred to as enhncesomes [1]. Functionally related genes involved in the same molecular-genetic, biochemical, or physiological process are often regulated coordinately by specific combinations of transcription factors. On the level of DNA, the blueprint of such common mechanisms of regulation may be seen as specific combinations of TF binding sites located in a close proximity to each other. We call such structures as "composite clusters". We are aiming to reveal a variety of such composite clusters in regulatory regions of eukaryotic genes. Different composite clusters could serve as good benchmarks for identification of new promoters and other regulatory regions in genomes and for functional characterization of the expression of the corresponding new genes as for understanding molecular mechanisms of their regulation. The goal is to find efficient means for automatic annotation of genomic regulatory sequences.

Last years, several computational approaches have appeared addressing the problem of combinatorial regulation of transcription. Specific TF binding site combinations were used for identification of muscle-specific promoters [2,3] for liver-enriched genes [4] and for yeast genes [5]. Recently, we have shown that search for specific combinations of two TF sites - composite elements - is a very effective tool for predicting gene expression patterns. We have demonstrated this approach for promoters of genes highly induced upon immune response [6]. Promoters of genes regulated during cell cycle could be recognized by combination of E2F binding sites with a dozen of oligonucleotide motifs [7]. A number of known examples of composite elements is collected in COMPEL database [8]. This data together with computationally predicted composite structure provide a key for annotation of regulatory regions in genomes.

Annotation of gene regulatory regions requires computational approaches that work with high sensitivity and specificity. One possible way to increase specificity is to develop methods that are trained on groups of co-regulated promoters rather than all promoters. Specific combinations of cis-elements for the vast variety of gene functional groups have to be determined to develop methods for automatic annotation of regulatory genomic sequences.

We have developed a method for revealing of composite clusters of cis-elements in promoters of eukaryotic genes that are functionally related or coexpressed. A software system "ClusterScan" have been created that enables: (i) to train system on representative samples of promoters to reveal cis-elements that tend to cluster; (ii) to train system on a number of samples of functionally related promoters to identify functionally coupled transcription factors; (iii) to provide tools for searching of this clusters in genomic sequences to identify and functionally characterize regulatory regions in genome. A number of training samples of different functional and structural groups of promoters were analysed. Search for composite clusters in human chromosomes 21 and 22 reveals a number of potential cell cycle regulated sequences. Finally, a decision tree system was

constructed to classify promoters of several functionally related gene groups. The decision tree system enables to predict an expression pattern of a new potential promoter by classifying it to the one of these promoter groups.

2 Method

2.1 Revealing of composite clusters of TF binding sites.

It is known that most of 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 possibility for protein-protein interactions between these factors. Therefore, it is expected that such sites for many different factors will make clusters in 5' regulatory regions that we call: "composite clusters" (CC) . Presence of such composite clusters in genomic sequence might be a good indication of regulatory regions of genes.

We have developed a method for identifying composite clusters of binding sites that are specific for promoter sequences. The method first analyses structure of promoter sequences from a training set of promoters trying to reveal clusters of transcription factor binding sites. For that, the whole library of weight matrices collected in TRANSFAC database [9] were considered. The method is based on genetic algorithm. It selects matrices and optimises cut-off values for every considered matrix in order to maximize the number of clusters in the training set of promoter sequences in contrast to a control set of non-promoter sequences.

Let's M is the set of all weight matrices from TRANSFAC. The following parameters are used for revealing composite clusters: K – a subset of weight matrices selected from the set M; $q^{(k)}_{cut-off}$ $(k \in K)$ - cut-off values of the matrix score (a site s considered to be present in a given position of the sequence if the score of the matrix k at this position exceeds the cut-off value $q^{(k)}(s) > q^{(k)}_{cut-off}$); max$d$ – the maximal distance between adjacent binding sites in a cluster. For example, when *mind* = 20bp, the algorithm considers only those clusters where distances between adjacent sites shorter then 20bp. The borders of the clusters are defined by sites that separated from the neighbour sites by the distance longer then max$d$. For a fixed set Γ of mentioned above parameters we can search for all clusters in every promoter sequence x. Then, we calculate the following function, that we call "cluster score":

$$CC\_score_{\lambda}(x) = \sum_{i=1,n} number\_of\_sites(i) \times density\_of\_sites(i) \qquad (1)$$

where n – is the number of found clusters in the sequence x; *number_of_sites(i)* – number of sites in the i-th cluster; *density_of_sites(i)= number_of_sites(i)/length_of_cluster(i)* – density of sites in the i-th cluster.

To reveal the best parameter set λ_{best} that exposes clusters in the promoters we apply a *genetic algorithm* (GA-1) that selects the subset K and optimises the values of the parameters *mind* and $q_{cut-off}^{(k)}$. We use the following fitness function:

$$J(\lambda) = \left(\frac{1}{|Y|}\sum_{y \in Y} CC\_score_{\lambda}(y) - \frac{1}{|Z|}\sum_{z \in Z} CC\_score_{\lambda}(z)\right) - R(|K|) \qquad (2)$$

In this fitness function we calculate difference between average values of the cluster score of the positive sample (Y) and negative sample (Z). The training set of promoters is the positive sample Y. As the negative sample Z we use a set of exon sequences. *R(m)* – is a function constructed similar to the Akaike Information Criteria [10] that decreases fitness of the models while the number of weight matrices increases. This criterion rescues the model to be over-fitted by getting the high number of free parameters.

2.2 Revealing of functionally coupled transcription factor sets

To analyse in more details the structure of found composite clusters we consider the following basic model of the composite promoter structure. Every eukaryotic promoter (or, more generally, a transcription regulatory region) contains numerous binding sites for different transcription factors that are organized in a number of functionally coupled subsets of factors – "functional sets" (FS). Every FS consists of a group of transcription factors that work together in one regulatory process by synergy or in antagonism through binding to their target sites that located in a close proximity to each other in regulatory regions of genes. Such FSs provide a framework for building up a specific complex of interacting TFs that supply a distinct regulatory function. Many examples of the simplest FSs consisting of two TFs with two adjacent binding sites are collected in the database of composite elements COMPEL [8]. Such FSs may provide gene induction in response to a complex condition, e.g. tissue-specific response to a certain extracellular signals. More complex FSs consisting of several interacting factors and may contain DNA signals of complex origin, such as TATA and GC boxes, Inr element and others. A family of functionally related promoters shares FSs that contain "obligatory" factors with target sites found practically in all promoters of the set and defining the "main" function of these promoters and "facultative" sites that may vary from

promoter to promoter and modulate the function in a specific manner. Such FSs being revealed as common for a promoter sample may be good benchmarks for promoter classification.

We describe a FS - μ characteristic for a group of functionally related promoters, by the following set of parameters: *P* a set of different TF weight matrices that compose the μ (including "obligatory" and "facultative" matrices). A certain cut-off value $q_{cut-off}^{(p)}$ and importance value $\varphi^{(p)}$ are assigned to every weight matrix *p* ($p \in P$) in μ. For every promoter sequence *x* we calculate the following function, that we call a "functional score":

$$FS\_score_{\mu}(x) = \sum_{p \in P} \varphi^{(p)} \times q_0^{(p)}(x), \qquad (3)$$

where $q_0^{(p)}(x)$ is the score of the best site found in the sequence *x* by the matrix *p* ($q_0^{(p)} = 0$, if no sites were found with score $q > q_{cut-off}^{(p)}$).

Optimisation of the parameters of the "functional score" is done by a modification of genetic algorithm (GA-2) similar to the one described in the previous section. In this case, the positive samples (Y) were sets of functionally related promoters. As the negative sample (Z) we use a full set of promoters (EPD database) where Y promoters are excluded.

2.3 Decision tree for classification of promoters

To classify promoters we build a decision tree (*T*) in the similar way as in [11]. The bottom nodes (*i*) of the tree (leafs) contain *L* different promoter classes. The internal nodes (*j*) of the tree represent different types of FSs - $\mu^{(j)}$. To classify a promoter sequence *x* the functional score $FS\_score_{\mu^{(j)}}(x)$ is calculated according to the equation (2) at every node as the sequence is passed to the tree. Cut-off values $FS\_score_{cut-off}$ are assigned to every internal node. If $FS\_score_{\mu^{(j)}}(x) > FS\_score_{cut-off}$ the sequence is passed to the left downstream node otherwise to the right downstream node. Finally, the sequence is classified to the one of the *L* promoter classes.

The decision tree was built by a variant of the *genetic algorithm* (GA-3), that optimizes the structure of the decision tree and cut-off values of the corresponding functions. The algorithm selects the components of $\mu^{(j)}$ at every node of the tree. The fitness function ρ is calculated on the basis of misclassification rate of decision tree *T* :

$$\mathrm{p}(T) = \prod_{i=1,L} \frac{N_{predict}^{(i)}}{N_{real}^{(i)}} \Big/ R(\sum m^{(j)}) \qquad (4)$$

Here, $N_{real}^{(i)}$ – is the number of promoters of the class (i) in the training set, $N_{predict}^{(i)}$ – is the number of correctly classified promoters of the class (i) , R – is the same function as in (2) calculated on the total number of weight matrices $m^{(j)}$ used in the decision tree.

3 Results

3.1 Composite clusters in promoter sequences of mammalian genes.

We extracted promoter sequences of mammalian genes from EPD database. The considered region was from –500 to +99 relative to the start of transcription. 349 promoters were extracted. We applied GA-1 method and revealed a set of matrices that exposes clusters in the promoters of this group. The following 25 matrices were selected by the algorithm: V$MSX1_01, V$PAX8_B, V$CDXA_01, V$GEN_INI3_B, V$P300_01, V$BARBIE_01, V$E2F_2Q6, V$E2F1_Q6, V$SP1_01, V$AP1_Q6, V$AP1_Q4, V$PAX4_04, V$NFKB_Q6, V$FOXD3_01, VUSF_Q6, VLDSPOLYA_B, V$OCT1_07, V$HNF3B_01, V$STAT_01, V$E2F_Q4, V$ETS1_B, V$OCT1_02, V$MYCMAX_02, V$SRF_C, V$VMAF_01. The maximal distance between adjacent sites *maxd* = 23bp. The average size of the clusters in promoter regions was 2.8 sites per cluster and in exon sequence 0.2 sites per cluster. It means that practically no clusters composed by these sites were observed in exon sequences, whereas in many promoters these sites make clusters of 3–5 sites.

3.2 Functionally coupled transcription factor sets.

Seven sets of promoters were obtained from different sources: promoters for cell-cycle related genes (43 promoters) and brain enriched genes (45 promoters) (collected in this work on the base of literature search), muscle-specific (25 promoters) and immune cell specific genes (24 promoters) [6], erythroid specific genes (10 promoters) (http:/www.bionet.nsc.ru), liver enriched genes (39 promoters) and housekeeping genes (26 promoters) (EPD rel.62). The promoter sequences of the length 600 bp (from –500 to +99 relative start of transcription) were extracted from EMBL database. We have selected these sets since they represent the most distinct functional classes of promoters

Applying the GA-2 method we have revealed functional sets of transcription factors specific for the promoter classes described above (see Table 1).

One can see, that matrices for a number of class-specific factors (such as E2F, NF-AT, MyoD, ..) were taken by the method as "obligatory" (high importance values were assigned). These matrices were included only in one class-specific functional set. Other matrices for some of the ubiquitous factors (such as SP-1, SRF, AP-1…) have been included in many FSs. These factors appeared to play an important role in many types of promoters.

In Fig. 1 we show two distributions of the functional score for cell cycle promoters versus exon sequences. One can see that high values of the score are the characteristic feature of the most cell cycle related promoters.

Table 1. Functional sets of transcription factors specific for different promoter classes. Values of the matrix relative importance are shown in brackets in the front of each TF name.

Promoter class	TF factors selected	J(l) – score
Cell-cycle related	**E2F** (1.00), **TATA (**0.95), **CREB** (0.88), **Sp-1 (**0.81**)**	7.2
Brain enriched	BRLF1 (0.192), ATF (0.038), **CREB (0.450), Sp-1 (0.592), HFH2 (1.00)**	3.8
Muscle-specific	Tal-1 (0.50), **YY-1 (1.0),** Oct-1 (0.40), **MyoD (0.80)**, **SRF (1.0), PAX5 (0.80)**	5.2
Immune cell specific	COMP1 (0.024), STAF (0.017), **NF-kB (1.30), NF-AT (0.957),** Brn-2 (0.059)	6.6
Erythroid specific	**n-myc (0.31) ,** GR (0.08), **AP-4 (1.00),** RREB-1 (0.08), v-Maf (.08)	2.0
Liver enriched	**RORalpha1 (1.00),** Sp-1 (0.03), **SREBP-1 (1.00), HNF-1 (0.54),** ER (0.07), GATA-1 (0.03)	2.6
Housekeeping	Egr-2 (0.15), **AhR/Arnt (0.72), ZID (0.94), Elk-1 (0.79),** NRF-2 (0.54), **CREB (.62)**	7.2

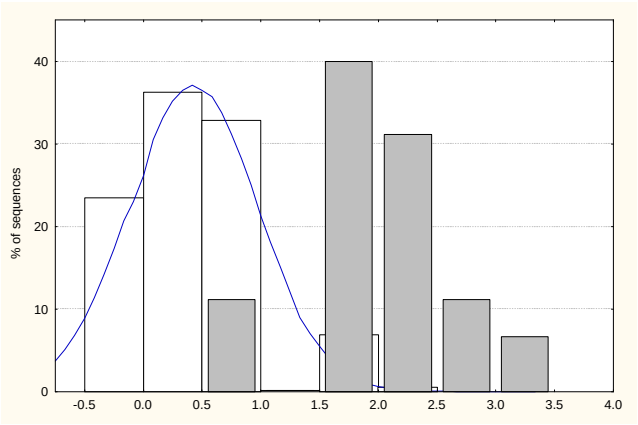

Fig. 1 Histograms of the functional score values in the cell-cycle related promoters (black) and exon sequences (white). The score is given along the x – axis. The functional score is calculated on the bases of set of factors specific for cell cycle promoters that are shown in the Table 1 (first column). In the histogram we show the percentage of the sequences in each set that exhibit the given value of "cell cycle functional score".

3.3 Decision tree classifier of promoters.

A decision tree classifier of the 7 classes of promoters was build by using of the weight matrices found in the previous step. The bottom nodes of the tree contain 7 different promoter classes. The training set of 212 promoters described above was used for optimising the decision tree structure with the help of GA-3. The topology of the one of the decision tree obtained in the analysis is shown in Figure 2.

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.

Percentage of the correct classification obtained by the tree is 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 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 efforts should be paid for initial grouping of promoters into functionally unified classes.

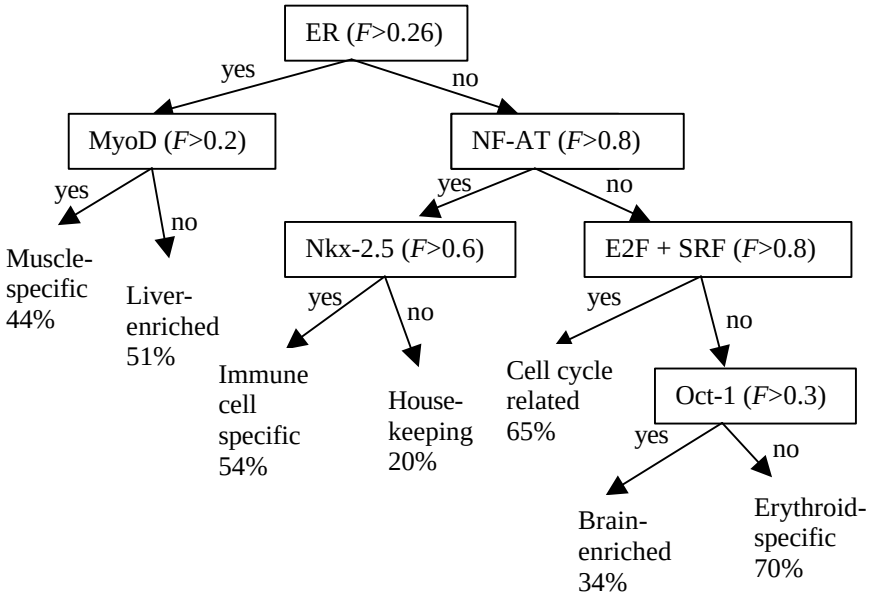

Fig. 2. A decision tree for classification of promoters into 7 functional classes. To classify a new promoter, the sequence (*x*) is passed down the tree beginning at the top. If the functional score: *F(x)* > *Fcut-off* the sequence is passed down to the left, otherwise to the right. The functions *F(x)* and cut-offs were optimised by GA-3.

We have applied the developed promoter classifier for identification of new potential cell cycle regulation for a number of known genes retrieved from EMBL. In our previous work [6] we developed a new method for context-specific identification of binding sites for E2F transcription factors – the main regulators of cell cycle progression. This method was applied to reveal new E2F target genes. We scanned EMBL release 6.0, divisions: hum, rod, vrt, and mam. 4611 promoters have been have been retrieved and analysed. As a result, 313 promoters were identified as new potential E2F targets [7].

In the present work, the promoter classifier was applied to the selected promoters to find the most probable target genes. After passing through the decision tree 103 promoters were classified as potential cell cycle regulated promoters.

Some of these promoters were inspected experimentally in our previous work using *in vivo* formaldehyde cross-linking technique to confirm the identity of potential E2F target genes that have been suggested computationally [7]. Using antibodies against various members of the E2F family, the specific E2F binding to several promoters under study in asynchronously growing HeLa cells have been confirmed. The following promoters bearing predicted E2F binding sites were experimentally confirmed to be cell-cycle dependent: *c-fos* and *junB*; the gene encoding TGF-β which acts as an antiproliferative agent to a majority of cell types; ARF locus encoding protein that binds to and stabilizes p53 and thus functions in tumor suppression; *mcm4* and *mcm5* involved in the initiation of DNA replication; von Hippel-Lindau (*VHL*) tumor suppressor gene; and *e2f-1*.

3.4 Search for composite clusters in human chromosomes and identification of new potential cell cycle related genes.

We have applied the ClusterScan system for scanning chromosome sequences of human genome in order to reveal composite clusters benchmarking new potential regulatory regions. As an example, we search for clusters that are specific for cell cycle related genes. We have previously shown that promoters of cell cycle genes are characterized by high frequency of the E2F binding sites [7]. The majority of promoters of cell cycle genes are GC-rich and TATA-less. In some of these promoters E2F binding sites are located just at the transcriptional start site. These data suggest that basal promoters of the cell cycle regulated genes may be characterized by a specific arrangement of known and yet unknown DNA elements resulting in specific composite clusters. In the training set of 29 cell cycle-dependent genes (no orthologs) we have revealed specific basal DNA elements by using Gibbs sampling program [12]. Within region [-45; -16] three motifs were revealed: TATA-like, GC box, and “CCT/ATT” motif. At the start site, [-15;+15], an E2F-like motif, an Inr-like pattern and the motif “CCC/A” were revealed. Downstream of the start site, within [+16; +45], a GAGA-like box was found. For all these motifs positional weight matrices were constructed. All the revealed motifs together with the E2F weight matrix were used for searching composite clusters in the chromosomal sequences.

Analysis of the human chromosome 21 resulted in 20 composite clusters. Of them, 7 clusters are located within annotated repeat families – SINE, LINE and LTR; 1 clusters within CpG islands; 2 within intron sequences of two genes; 4 clusters are found just 5’ to the annotated mRNA start of genes with unknown

function (see an example of such gene found in the chromosome 22, Fig.3); and 6 clusters do not coincide with any annotation.

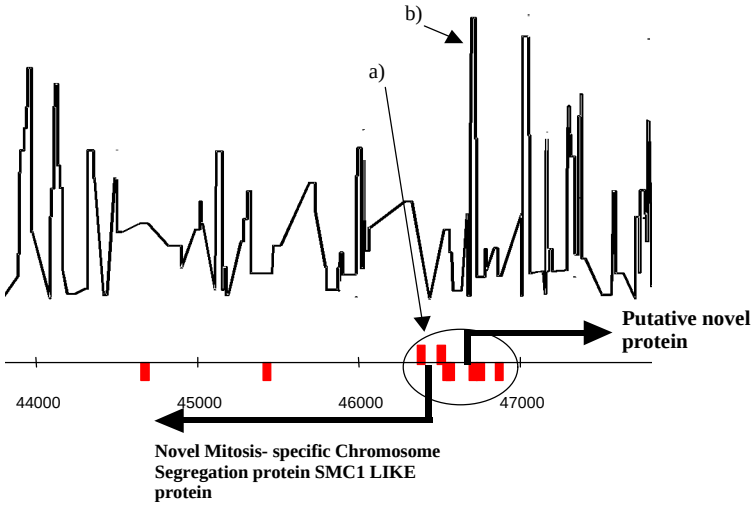

Fig.3 Prediction of a novel cell cycle regulated promoter in a fragment of the chromosome 22. Human DNA sequence from clone RP1-102D24 on chromosome 22 (AC: AL021391) is considered. a) a cluster of E2F sites at the potential starts of transcription of two genes. b) a pick of the composite cluster score (CC_score) comprising basal elements revealed in the training set of cell cycle related genes.

In summary, the computer method presented here allows us to search for clusters of potential cis-regulatory elements and to reveal promoters that belong to several definite functional categories. Experimental verification of some of these promoters confirms the computational predictions. With the advent of the large-scale sequencing projects, it becomes increasingly essential to develop computational methods enabling to analyse transcription regulatory regions of new genes and predict theire regulatory functions.

Acknowledgments

The authors are indebted to Vadim Ratner and Michael Zhang for fruitful discussion of the results. Parts of this work was supported by Siberian Branch of Russian Academy of Sciences, by grant of Volkswagen-Stiftung (I/75941)

References

1. Merika M. and Thanos D. Enhanceosomes. *Curr. Opin. Genet. Dev.* **11,** 205-208 (2001)
2. Wasserman, W. W., Fickett, J. W. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol*. **278** , 167-181 (1998)
3. Frech, K., Quandt, K., Werner, T. 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/ (1998)
4. Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M. & Pontoglio, M. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.* **266**, 231-245 (1997)
5. Brazma, A., Vilo, J. & Ukkonen, E. 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.), (1997) 57-60
6. Kel, A., Kel-Margoulis, O., Babenko, V., Wingender, E. " Recognition of NFATp/AP-1 Composite Elements within Genes Induced upon the Activation of Immune Cells" *J. Mol. Biol.* **288** , 353-376 (1999)
7. Kel A.E, Kel-Margoulis O.V., Farnham P.J., Bartley S.M., Wingender E., and Zhang M.Q. Computer-assisted identification of cell cycle-related genes - new targets for E2F transcription factors. *J. Mol. Biol.* **309** , 99 – 120 (2001)
8. Kel-Margoulis,O.V., Romaschenko,A.G., Kolchanov,N.A., Wingender,E. and Kel,A.E. TRANSCompel: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res*. **28,** 311-315 (2000)
9. Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt T., Pruss, M., Reuter, I., Schacherer, F. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28,** 316-319 (2000)
10. Akaike, H. *IEEE Trans. Autom. Control* **19**, 761−723 (1974)
11. Salzberg, S. Locating protein coding regions in human DNA using a decision tree algorithm *J. Comput. Biol*. **2** , 473-485 (1995)
12. Kel-Margoulis O., Kel A. and Wingender E. Automatic annotation of the regulatory regions of cell cycle related genes on human chromosomes. // Proceedings of the conference, Genome sequencing and biology. Cold Spring Harbor Laboratory, May 9-13, 2001. P.139