From Composite Patterns to Pathways –Prediction of Key Regulators of Gene Expression.

Alexander Kel[1], Nico Voss[1], Tatyana Konovalova[2], Dmitry Tchekmenev[1], Philipp Wabnitz[4], Olga Kel-Margoulis[1] and Edgar Wingender[1,5]

[1] BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbuttel, Germany;
[2] Institute of Cytology and Genetics, 630090, pr. Lavrentyev-10, Novosibirsk, Russia.
[4]INGENIUM Pharmaceuticals AG, Fraunhoferstr. 13, D-82152 Martinsried, Germany.
[5]Dept. of Bioinformatics, UKG, University of Gottingen, Goldschmidtstr. 1, D-37077 Gottingen. E.mail: ake@biobase.de

Abstract: Motivation: Functionally related genes involved in the same genetic, biochemical, or physiological process are often regulated co-ordinately by specific combinations of transcription factors (TFs) that bind to specifically arranged binding sites on DNA (composite modules (CMs)). Different signal transduction pathways leading to the activation of these TFs converge at key molecules that master-regulate cellular processes under study. Results: We developed a novel computational approach to interpret gene expression data and to propose complexes of transcription factors and further "upstream" key signaling molecules that are able to master-regulate the observed gene expression. The approach utilizes data from two databases (TRANSFAC® and TRANSPATH®) and consists of two major components: 1) *CMFinder* analyzes 5'-upstream regions of co-expressed genes and applies a genetic algorithm to reveal CMs consisting of co-occurring single TF binding sites and composite elements; 2) *ArrayAnalyzer* is a fast network search engine that analyzes signal transduction networks upstream of the TFs revealed in the previous step and finds key molecules which can be responsible for the observed concerted gene activation. The approach was applied to a set of microarray data on differential gene expression in liver of growth hormone-deficient mice (*Sma1)*. The results obtained in our analysis allow to confirm the specific role of growth hormone (GH) in altering of gene expression in liver of the *Sma1* mutant mice. In addition we proposed a number of other highly relevant key molecules.

1 Introduction

New technologies generate mass data on gene expression patterns and profiles. However, the available capacities to interpret these data and to filter candidate targets for drug development are very limited.

Regulation of gene expression is accomplished through binding of TFs to their DNA binding sites and transmission of the regulatory signal to the basal transcription complex. Some of these TFs are specific for a particular tissue, a definite stage of development, or a given extracellular signal, but most transcription factors are involved in gene regulation under a rather wide spectrum of cellular conditions. It is clear by now that combinations of TFs rather than single factors drive gene transcription and define its specificity. Dynamic function-specific complexes of many different transcription factors, so called enhanceosomes (Merika and Thanos, 2001) are formed at gene promoters and enhancers driving gene expression in specific manners. At the level of DNA, the blueprints for assembling of such variable TF complexes on promoter regions may be seen as specific combinations of TF binding sites located in close proximity to each other. We call such structures "composite modules (CMs)". There may be several different types of CMs located in the regulatory region of one gene, that can be spaced (e.g. liver specific and muscle specific enhancers of one gene) or overlapping. CMs consisting of two/three closely located sites belong to the smallest lowest hierarchical level of CMs. CMs of a higher level may include more sites as well as CMs of a lower level such as CEs.

Specific TF binding site combinations were used for an identification of muscle-specific promoters (Frech et al., 1998; Wasserman and Fickett, 1998), promoters of liver-enriched genes (Tronche et al., 1997), of yeast genes (Brazma et al., 1997), of immune-specific genes (Boehlk et al., 2000; Fessele et al., 2001; Kel et al., 1999), and promoters of genes regulated during cell cycle (Kel et al., 2001). In the database TRANSCompel® (Kel-Margoulis et al., 2002) known CEs that are specific combinations of pairs or triplets of TF binding sites located in close proximity to each other are collected. Recently, a number of approaches identifying composite motifs were described: BioProspector (Liu et al., 2001), Co-Bind (GuhaThakurta and Stormo, 2001), MITRA (Eskin and Pevzner, 2002), dyad search (van Helden et al., 2000). These programs help to discover new regulatory sites for yet unknown transcription factors, but an "ab initio" motif finding method is limited by the length of sequences and may not be suitable for the analysis of long regulatory regions of genes of human and other higher eukaryotic organisms. A valuable source to identify TF binding sites is the TRANSFAC® database (http://www.biobase.de), (Matys et al., 2003). Novel methods have been developed that utilise this information (ClusterScan: Kel-Margoulis et al., 2002b, D'Souza et al., 2003; TOUCAN system - Aerts et al., 2003).

We developed a new method for detecting *de novo* composite modules using information from TRANSFAC® about positional weight matrices (PWMs) that characterize DNA binding signature of many different transcription factor families. This method allows us to interpret gene expression data and to propose complexes of transcription factors and further "upstream" key signaling molecules that are able to master-regulate the observed gene expression.

2 Data and Methods

2.1 Databases and PWM based site recognition method.

Two databases for prediction of gene expression were used (BIOBASE GmbH, Wolfenbuttel, Germany, www.biobase.de) . TRANSFAC® (Matys et al., 2003) is a database that collects information about gene regulation in eukaryotes based on annotation of experimentally proven binding of transcription factors to their target sites. TRANSPATH® (Krull et al., 2003) is a database that comprises data about molecules participating in signal transduction and the reactions they undergo, thus spanning a complex network of interconnected signalling components. TRANSPATH® focuses on signalling cascades that aim at transcription factors and thus alter the gene expression profile of a given cell. We used TRANSFAC® Professional rel.8.1 and TRANSPATH® Professional rel.5.1.

For predicting potential TF binding sites in nucleotide sequences we used the most widely used method which is the application of positional weight matrices (PWMs) (Quandt et al., 1995). TRANSFAC® provides the largest collection of weight matrices for eukaryotic transcription factors (www.biobase.de; www.gene-regulation.com). We use Match program (Kel et al., 2003) to search for potential TF binding sites using PWMs. This program requires two parameters for each matrix: *core cut-off* and *matrix cut-off* (for the 5 nucleotide most conserved core of the matrix and for the whole matrix respectively). These two cut-off can vary from 0.0 to 1.0 and they define how specific and how sensitive is the method. The higher the cut-offs the more specific and less sensitive is the search. We uses a set of 522 matrices that correspond to different TFs of vertebrate organisms. The cut-offs were set to guarantee for each matrix at least 90% of sensitivity for recognition of the corresponding TF binding sites.

2.2 A genetic algorithm to determine composite regulatory modules (CM).

Composite module (CM) is a particular combination of PWMs for different TFs, which is associated with a specific functional type of gene regulatory regions. In this paper we use the following model of CMs. We define a CM as a relatively small set of individual PWMs and pairs of PWMs, with given cut-offs for matrices and relative distances between site matches in the pairs. CM analysis was performed using the CMFinder tool, an advanced version of a promoter analysis tool ClusterScan based on a genetic algorithm as recently described (Kel-Margoulis et al., 2002b). CMs are characterized by the following parameters: *K,* the number of individual PWMs in the module, *R*, the number of pairs of PWMs, matrix cut-off values $q_{cut-off}^{(k)}$, relative impact values $\phi^{(k)}$ maximal number of best matches $\kappa^{(k)}$ that were assigned to every weight matrix *k* (*k=1,K*), as well as matrix cut-off values $q_{cut-off}^{(r)}$, relative impact values $\phi^{(r)}$ and maximal distances $d_{\max}^{(r)}$ that were assigned to every matrix pair *r* (*r=1,R*) in the CM.

Matrices are selected to be included into the CM by the program from a given library of PWMs. A composite module score (CM score) is calculated for a DNA sequences *X* (promoter of a gene) according to the following equation:

$$F_{CM}(X) = \sum_{k=1,K} \phi^{(k)} \times \sum_{i=1}^{\kappa^{(k)}} q_i^{(k)}(X) + \sum_{r=1,R} \phi^{(r)} \times (q_1^{(r)}(X) + q_2^{(r)}(X))$$

where $q_i^{(k)}(X) < q_{cut-off}^{(k)}$ and $q_{1,2}^{(r)}(X) < q_{cut-off}^{(r)}$; and distance between site matches of two matrices of a pair *(r)* : $d^{(r)} < d_{\max}^{(r)}$.

The CMFinder program takes as an input two sets of sequences (the set to be analyzed, we call it set ***Y,*** and a control set ***N***) and a library of PWMs for transcription factors. In the case of analysis of microarray data, set ***Y*** consists of promoters of up- or down-regulated genes or genes that are belong to one expression cluster, and the set ***N*** consists of background sequences (e.g. promoters of genes that did not change their expression). For defined parameters *K* and *R* and over a number of iterations, the program optimizes the CM by searching for a "best" combination of matrices, their cut-offs and their relative impacts that discriminates sets ***Y*** and ***N***. The output of the program is the best discriminative CM with the optimized parameters. We define the goal function of the genetic algorithm as a weighted sum of false negative, false positive errors and the statistical significance (t-test) over several random iterations of bootstrap procedure by splitting the initial ***Y*** and ***N*** sets into the training and testing subsets. In addition, we tested the normal distribution of the *F* function in the ***Y*** and ***N*** sets. Calculating the *goal function* allows to assess the usability of the obtained solutions for classification of individual sequences.

2.3 *ArrayAnalyzer*™ – a fast search engine that analyses signal transduction network.

To understand the mechanisms of gene expression, microarray data should be analyzed in the context of complex regulatory networks of a cell. Through such networks, few regulators can control expression of large sets of genes. We have developed a tool *ArrayAnalyzer*™ integrated in the TRANSPATH® signal transduction database that allows fast search for key upstream regulators in the networks of signal trunsduction that is composed of known experimentally verified signal trunsduction reactions annotated in TRANSPATH®. *ArrayAnalyzer*™ starts its upstream search from a list of so called target molecules which are considered as targets of a cascade of signal transduction. These target molecules include products of those genes that are differentially expressed in the considered gene expression experiment. A list of TFs that are proposed on the previous step of analysis are used as input for *ArrayAnalyzer*™.

The program searches for so called “key nodes” (or key molecules) in the signal transduction network upstream from the target molecules in such a way that a signal from these key molecules can reach the maximal number of target molecules through a minimal number of steps. By this way the program can find molecules providing coordinated signal transduction and, in the case of transcription factors, it can suggest a “master-regulator” of differentially expressed genes.

The algorithm of *ArrayAnalyzer*™ uses compressed shortest-distance matrices (***D***), which are calculated by, e.g., Floyd's algorithm. The compression is performed by a delta-like algorithm. The identification of key nodes is done by assigning a score to every node (molecule or gene) in TRANSPATH, depending on a given maximal distance *dmax* (“search radius”). The score *S* of each node represents the number *N* of genes from a certain given array experiment, which can be reached from this node (true positives), in relation to the number of all other nodes *M* in the whole network that can be reached as well (false positives), according to the chosen distance matrix ***D***, radius *dmax* and a value of “penalty on false positives” (P).

ArrayAnalyzer™ performs a search of the potential key nodes in the TRANSPATH network taking the TFs that are linked to the PWMs of the best discriminating CMs revealed on the previous step (see para. 2.2) as a starting set of target molecules. The found potential key nodes are ranked according to the calculated score *S*.

3 Results

3.1 Testing of the *CMFinder* on simulated data.

We generated a number of sets of random sequences. The nucleotide composition of these random sequences was adjusted to that of known genomic promoters. We implanted into the random sequences several binding sites of certain TFs by placing them in random positions. We used the *CMFinder* program to compare the random sequence set with and without implanted sites. We varied the minimal scores of the sites and their frequencies to find limitations of the method.

Table 1. Results of testing *CMFinder* on simulated data.

# of matrices	Implanted sites[1)]	% of sequences with implanted sites[2)]		
		50	70	100
2	**AhR, AP-1**	+/- (**AhR**, GATA)	+[3)]	+
4	**AhR, OCT, C/EBP, AP-1**	+/- (**AhR, C/EBP, OCT,** HNF3)	+	+
6	**AhR, OCT, C/EBP, AP-1, NF1, HNF1**	-/+ (**AhR, C/EBP**, HNF4, HNF3A, ROR)	+/- (**AhR, C/EBP**, **AP-1**, GATA, ROR, **HNF1**, COUP)	+/- (**AhR, OCT, C/EBP**, ROR, **NF1**, **HNF1**)

[1] here the names of TFs are given whose matrices were used for implantation into the random sequences. We did three tests by implanting 2,4 and 6 matrices in the sets of 100 random sequences of the length 600bp each;
[2] the corresponding percent of sequences in the set was used for implantation, other sequences in the set were left unchanged;
[3] + means that all implanted matrices were revealed correctly by the algorithm. +/- means that not all matrices were revealed correctly, the "correct" factors are shown in bold.

The result of this simulation (see Table 1) shows that the *CMFinder* program was able to determine correctly all TF matrices used for site implantation. However, the more matrices were used for implantation the more difficult it became for the program to reveal the sites correctly. Importantly, matrices were found to behave differently in the test. For instance, the matrices for C/EBP and AhR were easily identified whereas recognition of the matrix for OCT factors was more difficult, and the AP-1 matrix was most difficult to identify.

3.2 Testing of *ArrayAnalyzer*™ on the simulated sets of transcription factors.

To test the ability of *ArrayAnalyzer*™ to identify key molecules for different sets of TFs we ran an extensive random simulation. During simulations we were selecting randomly a certain set of TFs (2 to 10) from all factors that are currently present in TRANSPATH® (590 TF family nodes in total). For each selected set we performed an upstream *ArrayAnalyzer*™ search in order to find all possible key molecules that can reach all target molecules in the set. Each time, we select from all found key molecules the one that has the closest distance to the set of target molecules. The distance is calculated by adding up the number of steps from the key molecule to each target molecule. We assume that the closest key molecule is the most probable one in transducing the signal to the target molecules. The results of these simulations show that the total chances to find a key molecule in this random simulations is about 6%. It is possible to find it for pairs or even triples of factors, but rather unlikely for higher number of factors in the set (0.4% for four targets). The most frequent families of molecules that were found as key molecules in these simulations were: Jak, PKA, PKC, Rad23 and AKT.

3.3 Analysis of microarray data on a mouse model of the complex disease: growth hormone (GH) deficiency. Blind validation of the method.

Microarray analysis of 4200 murine genes expressed in the liver was performed to compare growth hormone-deficient Sma1 and wild type mice (Wabnitz et al., 2004, in preparation). This study aimed at elucidating molecular mechanisms and suggested potential new players responsible for the differences in growth control, body composition, lipid- and steroid metabolism between Sma1 and wild type mice. Importantly, the search for potential composite modules and following pathway analysis were done without any knowledge that mice under study are growth hormone deficient.

A data set contains 82 genes whose relative expression values differ 2 times and more between Sma1 and wild type mouse samples. To retrieve the promoter sequences of the genes we use Ensembl and DBTSS databases (Suzuki et al., 2002). The beginning of the annotated first exon was considered as a tentative transcription start site (TSS). However, for some genes several possible TSS positions were annotated due to possible alternative transcription starts. In these cases several possible 5'-regions were considered. In total we extracted 40 5' regions for up-regulated genes and 43 for down-regulated genes. For the analysis we selected the 600bp regions around TSS (-500 +100). A set of 788 promoters for vertebrate genes from EPD database (Praz et al., 2002) was used as a background set.

One of the best discriminative composite modules includes matrices for C/EBP, GR, SMAD3, Stat and Fox factors (Fig. 1). Combination of these matrices allows to differentiate promoters of Sme1 specific genes from other promoters.

a)

0.0983*V$TCF11MAFG_01(0.821)
0.0471*V$FOXO4_01(0.961)
0.0301*V$IPF1_Q4(0.852)
0.0410*V$AR_01(0.851)
0.0766*V$GR_Q6(0.971)
0.0482*V$STAT1_02(0.995)
0.0508*V$CEBPB_01(0.98)
0.0281*V$STAT5A_02(0.826)
0.1040*V$CETS1P54_02(0.949)-50-V$TCF4_Q5(0.908)
0.0751*V$TCF1P_Q6(0.726)-50-V$STAT6_01(0.861)
0.0728*V$SF1_Q6(0.684)-50-V$SMAD3_Q6(0.833)
0.0419*V$ELK1_02(0.862)-50-V$GRE_C(0.842)

b)

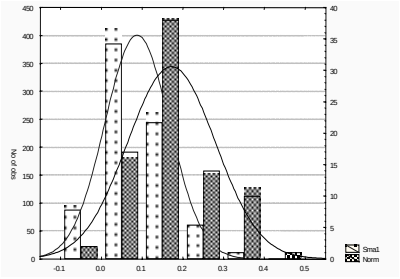

Fig. 1 Composite module found in promoters of differentially expressed genes in liver of growth hormone-deficient mice (*Sma1)*. a) A list of position weight matrices and matrix pairs consisting the CM. It includes the impact values (first value in each row), the weight matrix identifier (ID), corresponding cut-offs (in parenthesis after the matrix ID) and the maximal distances between sites in the matrix pairs (shown as "-50-"). b) Distributions of composite module scores. The plot shows discrimination between differentially expressed promoters (black bars) and non changed promoters (white bars). The CM score is given on the abscissa. The significance of discrimination by T-test p=5.6E-18.

At the next step, we have analysed the signal transduction pathways upstream of all found TFs using *ArrayAnalyzer*™. The goal of the pathway analysis was to reveal key molecules that could provide coordinate regulation of several transcription factors and thus can be considered as potential drug target molecules.

Growth hormone (GH) was identified by *ArrayAnalyzer*™ as one of the most weighted key molecules (among first 10 molecules with the maximal score *S*). Pathways upstream of transcription factors GR, STAT1, STAT5, C/EBP and others were crossing at this molecule (see Fig 2 as output of *ArrayAnalyzer*™). According to the data collected in TRANSPATH®, growth hormone can act through its cognate receptor (GHR) or prolactin receptor (PRLR). Downstream of these receptors, Jak kinases could be activated on one hand, and Src and PI3K kinases on the other hand.

Receptor Tyrosine Kinases (RTK) is another family of molecules that was found by the *ArrayAnalyzer*™ as a key molecule. Nck adaptor protein is recruited by Receptor Tyrosine Kinases (RTK) (for example Insulin receptor, IGF-I receptor and others) as well as by immune receptors (T-cell receptors, B-cell receptors) and serves to provide signal flow further downstream. PAK1 is a protein serine/threonine kinase downstream of Rac1. PAK1 is known to be activated in response to several growth factors, and is involved in cytoskeletal reorganization in activated immune cells.

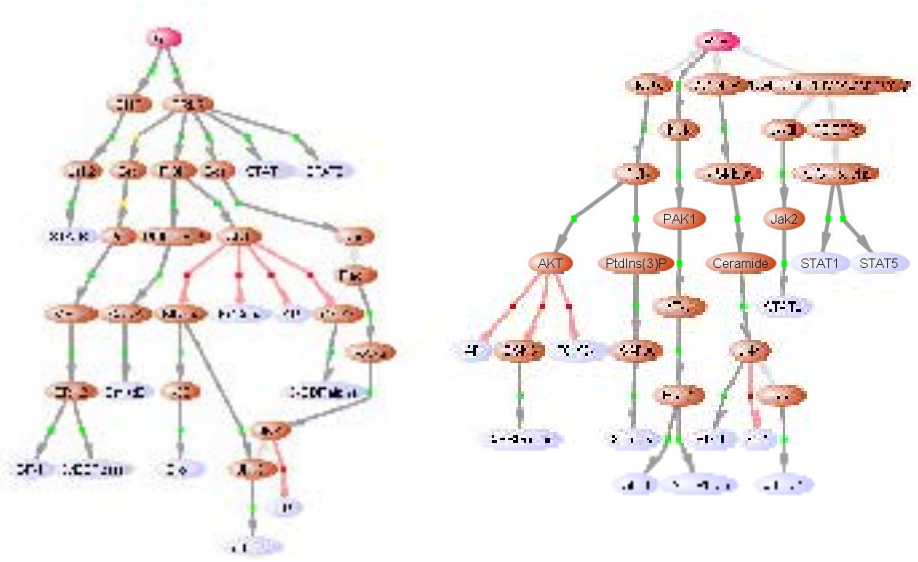

Fig. 2. Results of the *ArrayAnalyzer*™ search upstream from TFs resulting in identifying: growth hormone (GH) and receptor tyrosine kinases (RTK) as potential key molecules involved in differential expression of the genes in liver of growth hormone-deficient mice (*Sma1)*.

Any of these two cascades or both can be responsible for coordinated activation/repression of the found TFs: GR, STAT, C/EBP, SMAD, Ets and therefore for the regulation of genes that are differentially expressed in the liver of Sma1 versus wild type mice.

4 Conclusion

In this paper we describe a novel computational approach to interpret gene expression data. In this approach we analyse promoters of differentially regulated genes and are able to propose complexes of transcription factors that provide the observed concerted regulation of these genes. Further analysis of the signalling networks in the cell leading to activation of these transcription factors allow us to propose "key signaling molecules" that are able to master-regulate the observed gene expression. The approach was applied to a set of microarray data on differential gene expression in liver of growth hormone-deficient mice (*Sma1)*. The results obtained in our analysis allow to confirm the specific role of growth hormone (GH) in altering of gene expression in liver of the *Sma1* mutant mice.

Acknowledgments

Different parts of this study were supported by grants of the German Federal Ministry of Education and Research (Bioinformatics Competence Center "Intergenomics", grant no. 031U210B and "BioChance", grant no. 0312432), and by INTAS (Ref. Nr. 03-51-5218). TRANSFAC and TRANSPATH are registered trademarks of BIOBASE GmbH.

References

1. Aerts, S., Van Loo, P., Thijs, G., Moreau, Y., De Moor, B. (2003). Computational detection of cis -regulatory modules. Bioinformatics, Suppl 2:II5-II14.
2. Boehlk, S., Fessele, S., Mojaat, A., Miyamoto, N. G., Werner, T., Nelson, E. L., Schlondorff, D., and Nelson, P. J. (2000). ATF and Jun transcription factors, acting through an Ets/CRE promoter module, mediate lipopolysaccharide inducibility of the chemokine RANTES in monocytic Mono Mac 6 cells. Eur. J. Immunol. 30:1102-1112.
3. Brazma, A., Vilo, J., and Ukkonen, E. (1997). Finding Transcription Factor Binding Site Combinations in Yeast Genome, p. 57-59. In D. Frishman and H.W. Mewes (ed.), Computer Science and Biology. Proc. German Conference on Bioinformatics GCB '97, Martinsried, Germany.
4. D'Souza, U.M., Kel, A., Sluyter, F. (2003). From transcriptional regulation to aggressive behavior. Behav. Genet., 33:549-62.
5. Eskin, E., and Pevzner, P. A. (2002). Finding composite regulatory patterns in DNA sequences. Bioinformatics, 18:Suppl. 1: S354-S363.
6. Fessele, S., Boehlk, S., Mojaat, A., Miyamoto, N. G., Werner, T., Nelson, E. L., Schlondorff, D., and Nelson, P. J. (2001). Molecular and in silico characterization of a promoter module and C/EBP element that mediate LPS-induced RANTES/CCL5 expression in monocytic cells. FASEB J., 15: 577-579.

7. Frech, K., Quandt, K., and Werner, T. (1998). Muscle actin genes: a first step towards computational classification of tissue specific promoters. In Silico Biol. 1: 29-38.
8. Guha Thakurta, D. and Stormo, G. D. (2001). Identifying target sites for cooperatively binding factors. Bioinformatics, 17: 608-621.
9. Kel, A.E., Goessling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., Wingender, E. (2003) MATCH(TM): a tool for searching transcription factor binding sites in DNA sequences Nucleic Acids Res. **31**, 3576-3579.
10. Kel, A., Kel-Margoulis, O., Babenko, V., and 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.
11. 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.
12. Kel-Margoulis, O., Kel, A.E., Reuter, I., Deineko, I.V., and Wingender, E. (2002). TRANSCompel: a database on composite regulatory elements in eukaryotic genes. Nucleic Acids Res., 30: 332-334.
13. Kel-Margoulis, O.V., Ivanova, T.G., Wingender, E., and Kel, A.E. (2002b). Automatic annotation of genomic regulatory sequences by searching for composite clusters. Pac. Symp. Biocomput. 7, 187-198.
14. Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A., and Wingender, E. (2003). TRANSPATH®: an integrated database on signal transduction and a tool for array analysis. Nucl. Acids. Res. 31, 97-100.
15. Liu, X., Brutlag, D. L., and Liu, J. S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput. 127-138.
16. Lonze, B.E. & Ginty, D.D. (2002). Function and regulation of CREB family transcription factors in the nervous system. Neuron. 35: 605-623.
17. Merika, M., and Thanos, D. 2001. Enhanceosomes. Curr Opin Genet Dev., 11: 205-208.
18. Matys, V., Fricke, E., Geffers, R., Goessling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Muench, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender,E. (2003). TRANSFAC®: transcriptional regulation, from patterns to profiles. Nucl. Acids. Res. 31, 374-378.
19. Praz, V., Perier, R., Bonnard, C. and Bucher, P. (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucleic Acids Res. **30**, 322-324.
20. Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. **23**, 4878-4884.
21. Suzuki, Y., Yamashita, R., Nakai, K. and Sugano, S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res. 30:328-31.
22. Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M., and 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.
23. van Helden, J., Rios, A. F., and Collado-Vides, J. 2000. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. 28:1808-1818.
24. Wasserman, W. W., and Fickett, J. W. 1998. Identification of regulatory regions which confer muscle-specific gene expression. J. Mol. Biol. 278:167-181.