

Application of kernel method to reveal subtypes of TF binding motifs. Causal analysis of gene expression data.

Alexander Kel^{1,3}, Yury Tikunov², Nico Voss¹, Jürgen Borlak⁴ and Edgar Wingender.^{1,5}

¹ BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany,
ake@biobase.de

[\[ake,nvo,ewi\]@biobase.de](mailto:[ake,nvo,ewi]@biobase.de)

<http://www.biobase.de>

² Institute of Cytology and Genetics SB RAN, Lavrentyev pr.,10, 630090, Novosibirsk, Russia

kel@bionet.nsc.ru

³ United Institute of Geology, Geophysics and Mineralogy SB RAN, Koptyug pr.,3, 630090, Novosibirsk, Russia

tikunov@uiggm.nsc.ru

⁴ Fraunhofer Institute (Fh-ITEM) of Toxicology and Experimental Medicine, Center for Drug Research and Medical Biotechnology, Nikolai-Fuchs-Str. 1, D-30625 Hannover, Germany.

borlak@item.fraunhofer.de

⁵ Dept. of Bioinformatics UKG, University of Goettingen, Goldschmidtstr. 1, D-37077 Goettingen, Germany

e.wingender@med.uni-goettingen.de

Abstract. Transcription factor binding sites often contain several subtypes of sequences that follow not just one but several different patterns. We developed a novel sensitive method based on kernel estimations that is able to reveal subtypes of TF binding sites. The developed method produces patterns in form of positional weight matrices for the individual subtypes and has been tested on simulated data and compared with several other methods of pattern discovery (Gibbs sampling, MEME, CONSENSUS, MULTIPROFILER and PROJECTION). The kernel method showed the best performance in terms of how close the revealed weight matrices are to the original ones. We applied the Kernel method to several TFs including nuclear receptors and ligand-activated transcription factors AhR. The revealed patterns were applied to analyze gene expression data. In promoters of differentially expressed genes we found specific combinations of different types of TF binding patterns that correlate with the level of up or down regulation.

1 Introduction

Sites in genomic DNA that serve as targets for binding of a certain transcription factor (TF) share common patterns that are often described by consensus sequences or position weight matrices (PWMs). Elucidation of the structure of TF binding sites is a

very important problem because it enables to understand the mechanism of gene regulation. Several methods have been developed in the recent years for identification of patterns shared by a set of functionally related sequences, e.g. CONSENSUS (Hertz and Stormo, 1999), Gibbs Sampler (Lawrence et. al., 1993), MEME (Bailey and Elkan, 1994), ANN-Spec (Workman and Stormo, 2000), PROJECTION (Buhler and Tompa, 2002), combinatorial approaches (Pevzner and Sze, 2000), MULTIPROFILER (Keich & Pevzner, 2002). A combination of these methods was used for identifying target sites of cooperatively binding factors (Thakurta and Stormo, 2001). The methods that are able to reveal patterns in the form of PWMs are of the most interest now. Position weight matrices currently are the state of the art in modeling the structure of TF binding sites. They are clearly superior to the consensus description. More complex models such as HMM showed quite good performance (Ellrott et al., 2003), but their application is limited now to a few TFs with the high number of known sites.

It is known now that often one set of TF binding sites may contain several subtypes of generally similar but different patterns. This happened in most cases because of the lack of knowledge. Often we don't know what particular isoform of the factor, which homo- or heterodimer, modified variant, or in a complex with which co-factor or other cooperating factor, it actually binds to *in vivo*. Methods for revealing such patterns by sub-clustering of the sets of sequences are urgently needed.

We have developed a novel method for discovery of subtypes of patterns based on the kernel estimation of a probability density function. A first variant of this method was used for the investigation of aligned sequences near the start of transcription of eukaryotic genes (Tikunov and Kel, 2000). Here we present the improvement of this method. It can now be applied to the analysis of unaligned nucleotide sequences. Using simulation of random sequences with implanted sites we have compared the developed kernel method with several other methods such as Gibbs sampling (<http://bayesweb.wadsworth.org/gibbs/>) (GIBBS), MEME (<http://meme.sdsc.edu/meme/>), CONSENSUS (<http://ural.wustl.edu/>), MULTIPROFILER (<http://www.cs.ucsd.edu/groups/bioinformatics/>) and PROJECTION (<http://www.cs.wustl.edu/~jbuhler/projection.html>). The Kernel methods showed the best performance in terms of how close the revealed patterns are to the original ones. The Kernel method was able to distinguish two very similar patterns (with more than 30% of the same nucleotides in the consensus) whereas most of the other programs had rather high level of identification errors. The program is available for on-line use at: www.biobase.de/cgi-bin/biobase/cbs2/bin/template.cgi?template=cbscall.html. Source code is available upon request.

The developed method of pattern discovery makes a new advance in our capabilities in interpretation of gene expression data. Regulation of gene expression is accomplished through binding of multiple transcription factors to large regulatory regions of genes. 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 transcription factors rather than single transcription factors drive gene transcription and define its specificity.

We apply the developed Kernel method on the analysis of a set of gene expression data from toxicogenomics studies. We found two pattern subtypes of binding sites for AhR transcription factor, which plays a key role in regulation of genes during anti-toxic response of cells. In promoters of AhR activated or repressed genes we found specific combinations of different types of TF binding sites including two AhR pattern subtypes that correlate with the level of up- or down-regulation.

2 Data and Methods

We use two databases in the analysis: TRANSFAC® (BIOBASE GmbH, Wolfenbüttel, Germany) is a database that collects information about gene regulation in eukaryotes based on binding of transcription factors to their target sites; and TRANSCompel® which contains known composite regulatory elements in mammalian genes. We used TRANSFAC® Professional rel.6.4 and TRANSCompel® rel.6.4.

2.1 Kernel model for nucleotide sequences

We consider some set of nucleotide sequences s with length m . Let us denote a whole set of possible sequences as Ω . So the number of elements of set Ω is 4^m . The expected frequency of every possible sequence s is defined with appropriate probability p_s . Let us assign to every sequence some nonnegative weight w_s . We call this function $w(s)=w_s$ the weight kernel. We define for the weight kernel $w(s)$ the averaged weight sum $S_0(w)$ and volume $V_h(w)$, where

$$S_0(w) = \sum_{s \in \Omega} w_s \cdot p_s \quad (1a)$$

$$V_h(w) = \left(\sum_{s \in \Omega} w_s^{1+h} \right)^{1/(1+h)} \quad (1b)$$

$$\Phi_{0,h}(w) = \frac{S_0(w)}{V_h(w)} \quad (1c)$$

The ratio of averaged weight sum $S_0(w)$ to volume $V_h(w)$ is a functional of averaged density $\Phi_{0,h}(w)$ with respect to kernel $w(s)$.

Since $w_s \geq 0$ and $p_s \geq 0$ the Hölder's inequality is true

$$\sum_{s \in \Omega} w_s \cdot p_s \leq \left(\sum_{s \in \Omega} w_s^y \right)^{1/y} \cdot \left(\sum_{s \in \Omega} p_s^z \right)^{1/z} \quad (2)$$

where $y > 1$, $1/y + 1/z = 1$. If we put $y = h+1$, $z = (h+1)/h$ then we get from relation (2) to the inequality:

$$\frac{\sum_{s \in \Omega} w_s \cdot p_s}{\left(\sum_{s \in \Omega} w_s^{h+1} \right)^{1/(h+1)}} \leq \left(\sum_{s \in \Omega} p_s \frac{h+1}{h} \right)^{h/(h+1)} \quad (3)$$

The left part of inequality (3) is the averaged density functional $\Phi_{0,h}(w)$. The inequality (2) turns into equality when $w_s^y = (c \cdot p_s)^z$, where c is an arbitrary positive normalization factor. Hence

$$w_s^h = c \cdot p_s \quad (4)$$

It follows that the kernel is the densest when weight kernel function satisfies the equation (4).

If we have some sample of sequences $\Omega_n = \{s_1, s_2, \dots, s_n\}$ we can construct empirical averaged weight sum $S_n(w)$ and empirical averaged density functional $\Phi_{n,h}(w)$ with respect to kernel $w(s)$ as

$$S_n(w) = \frac{1}{n} \cdot \sum_{s \in \Omega_n} w_s \quad (5a)$$

$$\Phi_{n,h}(w) = \frac{S_n(w)}{V_h(w)} \quad (5b)$$

The mathematical expectation of empirical averaged weight sum $S_n(w)$ is defined with expression (1a) and the expectation of $\Phi_{n,h}(w)$ is defined with expression (1c). In accordance with the law of large numbers the value of functional $\Phi_{n,h}(w)$ converges to the true value in probability under $n \rightarrow \infty$. So we can estimate the functional $\Phi_{n,h}(w)$ with any accuracy under $n \rightarrow \infty$. If we have some set of kernels $w_\alpha(s): \alpha \in \Lambda$ the densest kernel defines the probabilities of sequences $p(s)$ in accordance with relation (4). The proposed functional of averaged density $\Phi_{n,h}(w)$ is remarkable because it allows us to reconstruct the probabilities $p(s)$ by putting more weights to the more frequent sequences s . Smoothing parameter h regulates the weights for sequences with different expected probabilities.

Under $h = 0$ the functional $\Phi_{n,h}(w)$ is similar to Parzen-Rosenblatt kernel estimation of probability density which is often used in mathematical statistics for reconstruction of probability density (Rosenblatt, 1956; Parzen, 1962). A.I.Orlov (Orlov, 1991) adopted Parzen-Rosenblatt estimation for analysis of nonnumeric data. However, using of proposed functional $\Phi_{n,h}$ enables the search of the best probability function $p(s)$ suitable not for all space Ω but only for some of its local compact part.

For patterns that are represented by weight matrices $\|f_{ji}\|$ we can apply the described theory and construct an algorithm that allows us to reveal all the patterns in a set of sequences by searching for the clusters which are characterized with local maxima of $\Phi_{n,h}$. We assume that in each such cluster the probability distribution of sequences s is described with this matrix in accordance with independent distribution

of nucleotides in different positions $f_s = \prod_{j=1}^m f_{jl_j^s}$ where f_s is the frequency of sequence s ; f_{jl} is the frequency of letter l in position j (the elements of weight matrix $\|f_{jl}\|$); l_j^s is the letter of sequence s in position j . A consensus s_c that corresponds to a weight matrix $\|f_{jl}\|$ is a sequence that contains the most probable letters in every position. In accordance with equation (2) we define the weight kernel $w_s = (c \cdot f_s)^{1/h}$. Let us put the normalization factor c equal to $1/f_c$, where f_c is the frequency of the consensus sequence. So the weight of consensus sequence equals 1. Hence, the weight kernel

may be defined as $w_s = \prod_{j=1}^m w_{jl_j^s}$ and $w_{jl} = \left(\frac{f_{jl}}{f_{jl_j^c}} \right)^{1/h}$; here, l_j^c is the letter of consensus sequence in position j ; w_{jl} is the weight kernel coefficient. If we put $R_s = \ln(f_c/f_s)$

then $R_s = \sum_{j=1}^m \gamma_{jl_j^s}$ and $\gamma_{jl} = \ln(f_{jl_j^c} / f_{jl})$. R_s may be considered as a distance of

sequence s to consensus and γ_{jl} are the distance coefficients. The larger the distance the less weight it is assigned to this sequence in the model of the given consensus. Smoothing parameter h regulates the dependence of sequence weights from the distance.

Using the inputted above denotations one can derive an equation for the volume of kernel $w(s)$ based on a weight matrix $\|f_{jl}\|$

$$V_h(w) = \left(\sum_{s \in \Omega} w_s^{1+h} \right)^{1/(1+h)} = \left(\prod_{j=1}^m \sum_{l \in \mathbf{A}} e^{-\gamma_{jl} \cdot (1+h)/h} \right)^{1/(1+h)} \quad (6)$$

where l is a letter of alphabet \mathbf{A} . When functional $\Phi_{n,h}$ reaches its maximum its derivative equals zero. So from equation (5a), (5b), (6) by substituting the appropriate denotations we get

$$d(\ln \Phi_{n,h}) = \frac{1}{h} \cdot \sum_{j=1}^m \left(\frac{S_{n,jl}}{S_n} - \frac{e^{-\gamma_{jl} \cdot (h+1)/h}}{\sum_{\lambda \in \mathbf{A}} e^{-\gamma_{j\lambda} \cdot (h+1)/h}} \right) d\gamma_{jl} = 0 \quad (7)$$

where

$$S_{n,jl} = \frac{1}{n} \cdot \sum_{s \in \Omega_n(jl)} w_s = \frac{1}{n} \cdot \sum_{s \in \Omega_n(jl)} e^{-R_s/h} \quad (8)$$

$\Omega_n(jl)$ is a subsample of sample Ω_n those that the letter l is situated in position j . The $d(\Phi_{n,h})$ equals zero when every member of the sum from right part of equation (7) equals zero. We get from the equation (7) to the following equation system

$$f_{jl} = g \cdot \frac{S_{n,jl}}{e^{-\gamma_{jl}/h}}; \quad (9)$$

The value of g is calculated from the normalization requirement $\sum_{\lambda \in \mathbf{A}} f_{j\lambda} = 1$. The densest kernel corresponds to the probabilities $p(s)$ that in the best way describe the distribution of sequences near the local maxima.

2.2 Algorithm for detection of multiple patterns in the unaligned sequence sets .

In the algorithm each weight matrix is calculated on the basis of subsequences (words) of length m picked up from the sample (one subsequence from each sequence). The present algorithm is initialized by a random choice of a starting subsequence of the length m from one random sequence of the sample. From all other sequences, one subsequence of the length m is picked up which is the closest to the starting subsequence. On the basis of all these subsequences the initial weight matrix is calculated by just counting of the letters in the appropriate positions. When the initial weight matrix $\|f_{j,l}\|$ is built the appropriate weight kernel coefficients and distance coefficients are calculated. The algorithm makes several recursive iterations calculating the next weight matrix coefficients from the equation (9) using current weight kernel and distance coefficients. The iterations are stopped when no further changes in the matrix elements f_{jl} are observed.

3 Results

First, we tested the Kernel method on a set of simulated data and compared its performance with other methods of pattern discovery. After that we applied it to analyze three sets of TF binding sites. And finally, we use the found patterns in the analysis of gene expression data.

3.1 Comparison of the kernel method with other motif search algorithms using simulated data.

To compare the developed algorithm with other known algorithms we have prepared several samples of simulated data using a setup similar (Workman and Stormo PSB 2000). We generated sets of random sequences in which we implanted sites using predefined weight matrices.

We generated sets of 200 random sequences of the length 24bp each ($n=200$). Sites of length 10 ($m=10$) were implanted in a randomly chosen position into half of the sequences (one site in a sequence). The other half of the sequences remains just random. The weight matrix X that was used for generation of the implanted sites contains in every position one nucleotide with the maximal weight ξ varying from 0.65 to 0.95. We call this nucleotide as “consensus” nucleotide. All other nucleotides have got weights $(1-\xi)/3$. In this way we can simulate more conserved or less conserved matrices. Six programs have been compared: the kernel method developed in this work

(Kernel), Gibbs sampling program (GIBBS), MEME, CONSENSUS, MULTIPROFILER and PROJECTION.

The default parameters of the programs were used to perform the test. Each program runs several times on different sets of generated sequences. After each run the matrix Y calculated by the program was compared with the original matrix X (distance between matrices is measured by $D = \sum_{jl} (p_{jl}^{initial} - p_{jl}^{calculated})^2$). In order to

align matrices we slide matrix Y along the matrix X by 3 positions left and right to find the best fit. (In the case of mismatching we set frequencies 0.25 to shifted part of the X matrix). In Figure 1 we present the results of the comparison of the first four programs. It is obvious that the lower the parameter ξ the more difficult for a program to reveal correctly the matrix.

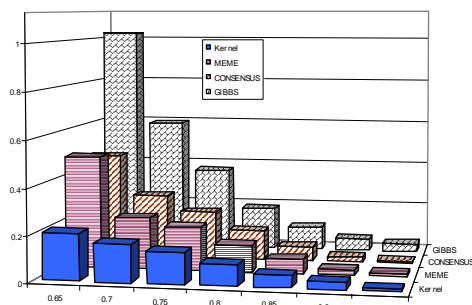


Fig. 1. Result of comparison of four different pattern discovery programs on the sets of simulated sequences with implanted TF binding sites for one matrix; y-axis: the averaged sum of squared differences (D) between revealed matrix and the original one (minimum 5 runs); x-axis: ξ values, that are the probabilities of “consensus nucleotide” in each position of the matrix. The smaller the value the better is the recognition ability of the program.

Programs MULTIPROFILER and PROJECTION approach the precision of the Kernel method, but only for the smallest values of $\xi = 0.65$ and 0.7 ($D = x$ and y , resp.).

3.2 Application of the kernel method to the sets of TF binding sites.

AP-1 and CREB. We take a mixture of binding sites for AP-1 and CREB transcription factors from database TRANSFAC. It is known that often transcription factors of these two different families bind to the same sites. The analyzed sample contained 155 sequences. Every sequence contained a TF binding site in the center and additional 10 nucleotide flanking both sides of the site.

We have applied our program and have revealed two different patterns. The analysis showed that the overwhelming majority of sequences contain a pattern of length 7 that corresponds to the consensus: “TGAGTCA”. The second pattern has the length 8 and corresponds to the consensus: “TGACGTCA”, which differs from the first one by insertion of letter “C” in the fourth position. It is important to mention that

exactly these two patterns correspond to the known consensi of AP-1 site (the first pattern) and CREB site (second pattern). Classification of the investigated sites shows that some of them contain both of these motifs located at different locations but close to each other.

We have applied two other programs: CONSENSUS and Gibbs sampling to the same set of sites. Using default parameters of these two programs we were not able to reveal two different patterns. Only one pattern was revealed that presents a mixture of the original two: “T(g/a)(c/a)GTCA”. It is noteworthy to pay attention that the kernel method by its nature is able to reveal correctly two matrices that are very similar to each other where most of other methods have much higher failure rate.

AhR. We have analyzed a relatively small set of 24 binding sites for transcription factor AhR that plays a very important role in the antitoxic cellular response. We identified two pattern subtypes: “TTGCGTGA”(matrix V\$AHR_N1, see Fig.2) and “CTCGCGTG” (V\$ANR_N2) that differ mainly at their 5’ end and may correspond to two different groups of binding complexes.

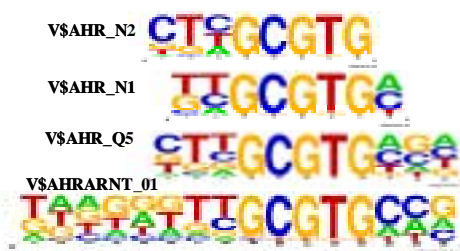


Fig. 2. Sequence logos of the weight matrices for AhR binding sites. Last two matrices are taken from TRANSFAC, first two are generated in this work.

3.3 Analysis of alteration of gene expression in human and rat hepatocytes by toxins.

We studied genes whose expression is regulated by a ligand-activated transcription factor AhR (aryl hydrocarbon receptor) that mediates responses to a variety of toxins. Expression of a number of genes was measured by RT-PCR in human and rat hepatocytes after treatment with Aroclor 1254 (artificial ligand of Ah – receptors). 111 (72 human and 39 rat) of differentially expressed genes were identified. 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 TSS (transcription start site). For the analysis we selected the regions around TSS: -1000/+100.

To analyze the structure of these promoters we applied a novel method *CMFinder* (Kel et al, 2004 submitted). This method applies a genetic algorithm to identify so called composite modules (CMs) – specific combinations of TF patterns (in the form of weight matrices) that correlate with the level of up or down regulation of the genes.

The program takes as input a full set of known TF matrices and computes an optimal combination of them that fits best to the observed expression changes of the genes.

We have included the new pattern subtypes for AhR binding sites that were found on the previous step of analysis. The result of the *CMFinder* program is shown in Figure 3. One can see that both new AhR patterns as well as some other previously constructed patterns were selected by the program. The composite module contained matrices for AhR, PPAR, HNF-6, STAT, ROR and ETS.

I	CC	CM	acc	Name	#
4.12	1.000	0.933	M00139	<u>V\$AHR_01</u>	3
1.00	0.655	0.639	M00528	V\$PPARG_03	3
1.00	1.000	0.900	A00002	<u>V\$AHR_N2</u>	4
0.48	1.000	0.905	M00235	<u>V\$AHRARNT_01</u>	2
2.08	0.861	0.904	M00340	V\$ETS2_B	2
4.46	1.000	0.927	M00639	V\$HNF6_Q6	2
-0.30	1.000	0.924	A00001	<u>V\$AHR_N1</u>	3
-0.74	1.000	0.956	MT00026	<u>V\$AHR_Q5</u>	4
-0.48	0.939	0.908	M00492	V\$STAT1_02	4
0.74	1.000	0.921	M00156	V\$RORA1_01	1

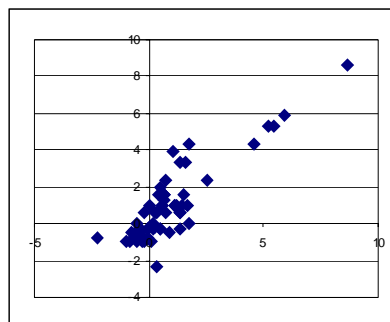


Fig. 3. Result of applying *CMFinder* program to the AhR gene expression data. A) The optimal composite module found by the program; two underlined matrices correspond to the two pattern subtypes of AhR sites; column I corresponds to the impact of the matrix into the correlation of the composite module score with the level of up or down regulation (positive impacting “up”, negative – “down” regulation); CC, CM – core and matrix cut-offs; Acc and Name – TRANSFAC accession number and the name of the matrix. B) Plot of the correlation between composite module score (x-axis) and the log of gene expression change (y-axis).

It is interesting that five different AhR matrices were included by the algorithm in the CM. This suggests that sites for AhR in the promoters play an important role in the up or down regulation of these genes. V\$AHR_01 got the maximal impact value. It seems to be specific for several cytochrome P450 genes and influences their expression in response to AhR. V\$AHRARNT_01 was constructed on the basis of data from SELEX experiments whereas V\$AHR_Q5 was constructed on the basis of genomic sites. These two matrices as well as two matrices constructed by the Kernel method (V\$AHR_N2 and V\$AHR_N1) have very different impact values: Two of them (V\$AHRARNT_01, V\$AHR_N2) have a positive, the other two (V\$AHR_Q5, V\$AHR_N1) a negative impact value. Comparison of the structure of these two matrices shows that they are very similar in the core, but differ in some nucleotides at the flanks. For example, in position 10 of matrix V\$AHR_Q5, the most prominent nucleotide is G, whereas in the matrix V\$AHRARNT_01 nucleotide “G” in the corresponding position is absolutely “forbidden”. Matrices V\$AHR_N1 and V\$AHR_N2 are also different but on the other flank of the site. You can see in the Table 2 that the matrix V\$AHR_N1 have got in the position 2 the consensus nucleotide “T” whereas in the matrix V\$AHR_N2 the corresponding position is mainly occupied by nucleotide “C”. This could influence binding of some other factors such as repressors in the vicinity of AhR sites.

In general, the Kernel method described in the paper is applicable to sets of unaligned regulatory sequences of any length. In a separate study we demonstrated its ability to reveal multiple patterns in the sets of tissue specific promoters.

Acknowledgments

This work was mainly funded by BIOBASE GmbH (Wolfenbüttel, Germany). Parts of this work were supported by Siberian Branch of Russian Academy of Sciences, by a grant of Volkswagen-Stiftung to A.K. (I/75941) and by a grant of the Lower Saxony Ministry of Culture and Science to J.B.

References

1. Bailey T.L., Elkan C. (1994) Proc Int Conf Intell Syst Mol Biol. 2, 28-36.
2. Buhler, J. and Tompa, M. (2002) J. Comput. Biol., 9, 225-242.
3. Ellrott K., Yang C., Sladek F.M., Jiang T. (2002) Bioinformatics. Suppl 2,S100-S109.
4. Hertz G.Z. and Stormo G.D., (1999) Bioinformatics, 15, 563-577.
5. Keich U., Pevzner P.A. (2002) Subtle motifs: defining the limits of motif finding algorithms. Bioinformatics,18,1382-1390.
6. Lawrence, C.E., Altschul, S.F., Bogouski, M.S., Liu, J.S., Neuwald, A.F., and Wooten, J.C., (1993) Science, 262, 208-214.
7. Pevzner, P.A. and Sze, S. (2000). In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, 269-278.
8. Thakurta, D.G. and Stormo, G.D. (2001) Bioinformatics, 17, 608-621.
9. Tikunov, Yu. and Kel, A.E. (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, 83-88.
10. Tikunov, Y., Kel, A. (2004) Functional of averaged density: Application for estimation of probability density function. Submitted.
11. Workman, C.T. and Stormo, G.D. (2000) Pac. Symp. Biocomput., 5, 464-475.
12. Parzen E. (1962) On estimation of probability density function and mode, Annals of Mathematical Statistics, 33, 1065-1076.
11. Rosenblatt M. (1956) Remarks on some nonparametric estimates of a density function, Annals of Mathematical Statistics, 27, 832-837.
12. Orlov, A.I. (1991) Classification of nonnumeric objects on the basis of nonparametric density estimations. In: Problems of computer data analysis and modeling. Byelorussian State University, pp. 141-148.
13. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Priß, M., Schacherer, F., Thiele, S. and Urbach, S. (2001) The TRANSFAC system on gene expression regulation. Nucleic Acids Res. 29, 281-283.