Journal of Bioinformatics and Computational Biology
KULAKOVSKIY et al, J. Bioinform. Comput. Biol. (2013) DOI: 10.1142/S0219720013400040

FROM BINDING MOTIFS IN CHIP-SEQ DATA TO IMPROVED MODELS OF TRANSCRIPTION FACTOR BINDING SITES

IVAN KULAKOVSKIY

Laboratory of Bioinformatics and Systems Biology, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilov str. 32, Moscow, 119991, GSP-1, Russia[*]
Department of Computational Systems Biology, Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina str. 3, Moscow, 119991, Russia
ivan.kulakovskiy@gmail.com

VICTOR LEVITSKY

Laboratory of Molecular Genetics Systems, Institute of Cytology and Genetics of the Siberian Division of Russian Academy of Sciences, Lavrentiev Prospect 6, Novosibirsk, 630090, Russia
Faculty of Natural Sciences, Novosibirsk State University, Pirogova str. 2, Novosibirsk, 630090, Russia
levitsky@bionet.nsc.ru

DMITRY OSCHEPKOV

Laboratory of Molecular Genetics Systems, Institute of Cytology and Genetics of the Siberian Division of Russian Academy of Sciences, Lavrentiev Prospect 6, Novosibirsk, 630090, Russia
diman@bionet.nsc.ru

LEONID BRYZGALOV

Laboratory of Regulation of Gene Expression, Institute of Cytology and Genetics of the Siberian Division of Russian Academy of Sciences, Lavrentiev Prospect 6, Novosibirsk, 630090, Russia
leon_l@bionet.nsc.ru

ILYA VORONTSOV

Department of Computational Systems Biology, Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina str. 3, Moscow, 119991, Russia
Yandex Data Analysis School, Data Analysis Department, Moscow Institute of Physics and Technology, Leo Tolstoy Str. 16, Moscow, 119021, Russia
vorontsov.i.e@gmail.com

VSEVOLOD MAKEEV

Department of Computational Systems Biology, Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina str. 3, Moscow, 119991, Russia
State Research Institute of Genetics and Selection of Industrial Microorganisms, 1st Dorozhny proezd, 1 Moscow, 117545, Russia
Moscow Institute of Physics and Technology, Institutskii per. 9, Dolgoprudny, 141700, Moscow Region, Russia
vsevolod.makeev@gmail.com

[*] Present address.

Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) became a method of choice to locate DNA segments bound by different regulatory proteins. ChIP-Seq produces extremely valuable information to study transcriptional regulation. The wet-lab workflow is often supported by downstream computational analysis including construction of models of nucleotide sequences of transcription factor binding sites in DNA, which can be used to detect binding sites in ChIP-Seq data at a single base pair resolution. The most popular TFBS model is represented by positional weight matrix (PWM) with statistically independent positional weights of nucleotides in different columns; such PWMs are constructed from a gapless multiple local alignment of sequences containing experimentally identified TFBSs. Modern high-throughput techniques, including ChIP-Seq, provide enough data for careful training of advanced models containing more parameters than PWM. Yet, many suggested multiparametric models often provide only incremental improvement of TFBS recognition quality comparing to traditional PWMs trained on ChIP-Seq data. We present a novel computational tool, diChIPMunk, that constructs TFBS models as optimal dinucleotide PWMs, thus accounting for correlations between nucleotides neighboring in input sequences. diChIPMunk utilizes many advantages of ChIPMunk, its ancestor algorithm, accounting for ChIP-Seq base coverage profiles ("peak shape") and using the effective subsampling-based core procedure which allows processing of large datasets. We demonstrate that diPWMs constructed by diChIPMunk outperform traditional PWMs constructed by ChIPMunk from the same ChIP-Seq data. Software website: http://autosome.ru/dichipmunk/

Keywords: positional weight matrix, PWM, transcription factor binding sites, TFBS, TFBS model, gapless multiple local alignment, GMLA, ChIP-Seq, dinucleotide frequencies

1. Introduction

Transcription regulation in higher eukaryotes is a complex process involving specific regulatory proteins, transcription factors (TFs), binding to specific DNA sites. High-throughput wet-lab methods based on chromatin immunoprecipitation reveal tens of thousands of DNA segments that are bound by particular protein in particular conditions *in vivo*. Special dry-lab workflow is required to process this data. Among other aims, including proper detection of genes, regulated by a particular TF, researchers are interested in precise locations of transcription factor binding sites (TFBSs) in DNA and common text patterns, often called as motifs, overrepresented in experimentally identified TF binding sequences. Both these tasks are tightly interweaved in the field of motif discovery, which has been developing for more than 20 years[1,2]. Modern motif discovery methods can identify a sequence pattern, specifically recognized by some TF in DNA segments, and use the pattern to precisely identify locations of binding sites in the experimental data obtained by various techniques, including high-throughput experiments[3]. The sequence patterns produced during this step can be used as TFBS models for computational TFBS prediction in genomic sequences of interest.

A positional weight matrix (PWM) is the most widely used TFBS model. PWM is typically produced directly on a basis of a gapless multiple local alignment (GMLA) of sequences. The set of sequences corresponds to the TF-bound DNA regions. Each GMLA position (the column of the alignment) produces positional weights for nucleotides occupying this position. The important assumption made here is that the nucleotide frequencies in different alignment columns are independent[2], i.e. there are no correlations between nucleotides in different alignment columns. Most of methods to detect DNA motifs in modern high-throughput data[4] are still based on simple mononucleotide PWMs.

Recent attempts to train more complex models using ChIP-Seq data failed to clearly outperform[5] traditional PWMs with independent positional weights.

Yet, assumption of non-independent positions is one of the weaker properties of traditional PWMs. A very straightforward PWM extension is a similar matrix of positional weights that takes into account correlations of nucleotides in neighboring alignment positions[6,7]. Such correlations possibly arise from interactions between nucleotides neighboring in DNA and contributing to formation of DNA secondary structure[8,9]. Dinucleotide PWMs were demonstrated to outperform classic mononucleotide PWMs if a training set of sequences was large enough[10]. Sequence-dependent conformational and physico-chemical properties based on dinucleotide statistics were also successfully applied for TFBS recognition[11]. Recently, a very successful analysis of protein-binding microarray data was done on the ground of TFBS models that accounted for frequencies of neighboring dinucleotides[12].

Thus, dinucleotide PWMs provide simple yet powerful extension of PWMs that allows accounting for neighboring nucleotides within TFBSs. In this report we present an approach applying dinucleotide models analysis of ChIP-Seq data.

2. Methods

Previously, we have developed an effective motif discovery tool, ChIPMunk[13], that performed well in several independent benchmark studies[14,15]. ChIPMunk can account for ChIP-Seq base coverage data, i.e. use the shape of ChIP-Seq tags pileup as prior positional information for probable locations of binding sites within extended ChIP-Seq peak sequences. In this paper we present a novel algorithm and a corresponding computation tool, diChIPMunk, able to construct an optimal dinucleotide PWM (diPWM that basically defines a Markov order 1 TFBS model) that accounts for dependencies of nucleotides in neighboring alignment positions. We demonstrate how ChIPMunk workflow can be modified to use diPWMs as a TFBS model. Using independent control datasets we perform careful comparison of dinucleotide PWMs versus traditional mononucleotide PWMs.

2.1. *ChIPMunk core workflow: from mono- to dinucleoties*

Original ChIPMunk algorithm uses a subsampling approach and a greedy model optimization procedure. A random starting "seed" PWM and a corresponding GMLA are optimized on a random subset of the initial sequence set. The obtained PWM is then optimized on the total sequence set. The greedy optimization here is similar to that originally presented in CONSENSUS[16] with the best PWM hits from each sequence used to reconstruct the GMLA and build a PWM for the next iteration. Greedy optimization procedure converges rather quickly, and the two-step optimization procedure based on subsamples allows further speed up the convergence. Iterative random subsampling also allows the algorithm to avoid partially locally optimal alignments, since any PWM obtained from the total set of TFBSs is then submitted to the main optimization procedure as a starting seed. There it is firstly driven out of local optimum by optimization on a random TFBS subset and then is optimized on the total set of TFBSs to compete with PWMs obtained in other runs. More details on the core procedure can be found in[17] and in Supplementary materials at the ChIPMunk website (http://autosome.ru/ChIPMunk/).

The diChIPMunk algorithm uses the very same core procedure with the only major difference of diPWM as the TFBS model. The important points here are: (1) how to handle the sequences in a way able to account for neighboring dinucleotides (2) how to construct a diPWM from a GMLA and (3) how to select the optimal diPWM and the optimal GMLA in a “dinucleotide” sense.

2.1.1. *Handling sequences with dinucleotide alphabet*

We convert all DNA sequences from mono- to dinucleotide alphabet. Each dinucleotide is constructed from two single neighboring nucleotide letters. Conversely, each nucleotide of a DNA sequence, except for the fist and the last nucleotides, is included in two neighboring (overlapping) dinucleotides. This conversion from the mononucleotide to dinucleotide alphabet allows seamlessly incorporating dependencies (i.e. correlations) between neighboring nucleotides into the model. The dinucleotide alphabet contains 16 letters with each letter being a dinucleotide (AA, AC, AG, .., TT). E.g., ACGT nucleotide sequence is written as A-C-G-T in nucleotides and as AC-CG-GT in dinucleotides.

It should be noted, that sequences over ACGT-letter mononucleotide alphabet form only a subset of all sequence over AA-..-TT-alphabet, since dinucleotide alphabet allows only sequences like XY-YZ, where the second nucleotide of each preceding dinucleotide is the same as the first nucleotide of each subsequent dinucleotide.

2.1.2. *Constructing a diPWM from a GMLA*

If the sequences are written in the dinucleotide alphabet, the procedure is the same as for PWM in mononucleotide alphabet[13]:

$$S_{\alpha,j} = log\left(\frac{x_{\alpha,j} + cq_\alpha}{(N+c)q_\alpha}\right) \tag{1}$$

where $S_{\alpha,j}$ is the score of letter $\alpha \in \{\text{AA,AC,AG, .., TT}\}$ in position j, $x_{i,j}$ is the number of occurrences of dinucleotide letter α in the j-th column of the GMLA, q_α is the background frequency of dinucleotide α, N is the total number of sequences in the GMLA, and c is the pseudocount parameter set as $\sqrt{N}$. The score of a word written in dinucleotide alphabet is then estimated in the following way:

$$score(\text{word}) = \sum_{j=1}^{l} S_{\text{word}[j],j} \tag{2}$$

where l is the word length (equal to the number of columns of the diPWM) and $S_{\text{word}[j],j}$ is the diPWM element for the j-th letter in the word.

2.1.3. *Defining the dinucleotide optimality of a GMLA*

Any obtained GMLA is evaluated whether its column-specific dinucleotide composition deviates from the given background, e.g. if some dinucleotide is highly prevalent rendering the overall distribution far from the uniform. While solving a similar problem ChIPMunk uses a criterion based on Discrete Information Content (DIC) and its extension with Kullback term (KDIC). If the sequences are written in dinucleotide alphabet one can use a similar measure to estimate GMLA quality in the dinucleotide sense. We call this measure as Kullback Dinucleotide Discrete Information Content, KDIDIC and define it as follows:

$$\begin{aligned}\mathrm{DIDIC} &= \sum_{j=1}^{l} \sum_{\alpha\in\{\mathrm{AA},..\mathrm{TT}\}} \left(\log\left(x_{\alpha,j}!\right) - \log N!\right) \\ \mathrm{KDIDIC} &= \mathrm{DIDIC} - \sum_{j=1}^{l} \sum_{\alpha\in\{\mathrm{AA},..\mathrm{TT}\}} x_{\alpha,j}\log\left(q_{\alpha}\right)\end{aligned} \qquad (3)$$

where $\alpha\in\{\mathrm{AA},..\mathrm{TT}\}$ is the dinucleotide letter; q_α is the background frequency of α; j is the position (the column number) in GMLA; $x_{\alpha,j}$ is the frequency of α in j-th GMLA column; l is the width (the length) of the alignment and N is the total number of TFBS sequences. Thus the search for the optimal alignment is defined as the search for the alignment with the maximal KDIDIC value.

KDIDIC does not explicitly account for dependencies of neighboring dinucleotide columns, but we use only dinucleotide sequences which are unambiguously mapped to corresponding mononucleotide sequences written in 4-letter ACGT-letter alphabet. Thus neighboring positions (e.g. j and $j+1$) of the alignment are not independent since the neighboring dinucleotide letters overlap in the initial DNA nucleotide sequence (as described in 2.1.1).

KDIDIC maximum defines some optimal alignment for the set of all sequences written in 16 letter dinucleotide alphabet. DNA sequences written in mononucleotide alphabet correspond to the subset of all dinucleotide sequences. Thus KDIDIC maximum cannot select the “truly optimal” alignment. Still Eq. (3) gives an estimate how dinucleotide frequencies of a given alignment deviate from the given background dinucleotide distribution q_α. To clearly demonstrate that KDIDIC-optimal dinucleotide TFBS models allow good TFBS recognition we have performed a benchmark study which is described in a separate section below.

KDIDIC of terminal alignment columns also allows estimating the optimal alignment length. Default diChIPMunk motif length estimation procedure scans a given lengths range for the longest alignment with the terminal columns passing KDIDIC threshold. This criterion is based on the observation that flanking columns of the correct TFBS model should have their dinucleotide distribution different from a given background. This can be achieved by defining a threshold value for KDIDIC of a single column. Longer alignments can be discarded if they have terminal columns with dinucleotide distribution not passing a given KDIDIC threshold. The threshold was set equal to KDIDIC of a column missing any two dinucleotide letters and having the uniform distribution of frequencies of all 14 remaining dinucleotides.

2.2. *Assessing diPWM TFBS recognition performance*

We have extracted three independent ChIP-Seq datasets from ENCODE[18]. ChIP-Seq peaks for AP2A, HNF4A TFs were provided with the base coverage profiles while only peak segments were provided for REST transcription factor. DNA segments for AP2A and HNF4A were truncated by discarding flanking segments with base coverage profile values lower than 10% of the maximum height of the peaks. For all datasets we discarded too short peaks (<25bps) that possibly corresponded to PCR artifacts. The subsets of top 1000 peaks was taken for each case and 500 even (odd) ranked peaks were used for model training (control).

Three TFBS models were assessed for each TF: the longest existing PWM from TRANSFAC[19], a ChIPMunk PWM and diChIPMunk diPWM. ChIPMunk and diChIPMunk were searching for optimal alignments in a lengths range from 10 to 25bps using either peak shapes (for AP2A and HNF4A) or local nucleotide/dinucleotide composition as a background model (for REST).

To evaluate the model performance we plotted a set of receiver operator characteristic (ROC) curves using the same strategy as assessed in construction of HOCOMOCO (http://autosome.ru/HOCOMOCO/) TFBS model collection[20]. The control datasets were used to estimate the true positive (TP) rate independently from the learning set. TP rate was computed as the number of sequences from the independent control set with PWM hit scores no less than the threshold. Thus for a full range of TP rates we collected the corresponding set of threshold values. For each PWM or diPWM threshold we computed *P*-value, the fraction of all DNA segments that are recognized as binding sites by the model. Basically, *P*-value represents the probability to obtain a score of no less than the threshold in a selected position of the random DNA sequence. Thus *P*-value is used to estimate the False Positive (FP) rate as the probability to find at least one PWM hit scoring no less than the threshold in a random double-strand DNA segment of a fixed length L:

$$\mathrm{FP} = 1 - \left(1 - P\text{-value}\right)^{2(L-l+1)}. \tag{4}$$

Here *L* is selected as the median sequence length (as estimated based on the independent control set), *l* is the PWM/diPWM length and the model hits are assumed to be independent with their total number complying compound Poisson distribution.

With the set of TP and FP rates at hand we can plot a ROC curve and estimate the area-under-curve (AUC) as the single measure for the models quality. MACRO-APE software (http://autosome.ru/macroape/) was used to estimate the P-values for PWMs; dinucleotide version of MACRO-APE is available at diChIPMunk website (http://autosome.ru/dichipmunk/).

3. Results and Discussion

We have applied the strategy described in Methods to analyze three ChIP-Seq datasets and to estimate TFBS model quality using ROC curves and corresponding AUC values. ROC curves for AP2A models are shown in Figure 1. AUC values for all TFs are presented in Table 1. Motif LOGOs are shown in left panels of Figure 2. ChIPMunk models for all TFs showed better AUC values comparing to the TRANSFAC models; the difference between TRANSFAC and ChIPMunk was comparable to that of ChIPMunk and diChIPMunk. The notable exception was HNF4A with the negligible difference between model quality of diPWM and PWM. This agrees well with the published results for HNF4A[10]; the same effect was observed for other TFs in a wider study focused on dinucleotide models[21].

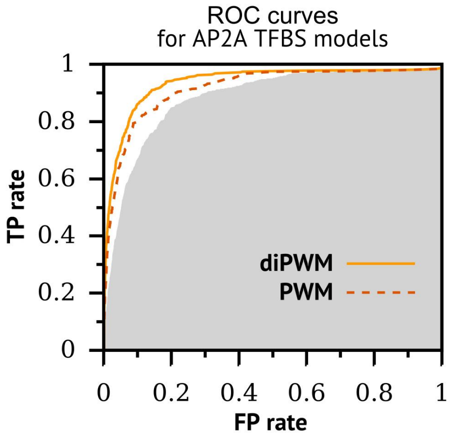

Fig. 1. Receiver Operating Characteristic (ROC) curves for TRANSFAC (shown as solid area), diChIPMunk (diPWM, solid line) and ChIPMunk (PWM, dashed line) TFBS models for TFBSs of AP2A transcription factor. TP and FP rates were estimated as described in Section 2.2. Larger area under curve corresponds to better model quality.

Table 1. Area-under-curve (AUC) values for ROC curves for TRANSFAC, ChIPMunk (PWM) and diChIPMunk (diPWM) models. Higher values correspond to better model quality (an ideal classifier would obtain the AUC value of 1.0).

		Transcription factor		
		AP2A	REST	HNF4A
TFBS models	diPWM	0.936	0.975	0.962
	PWM	0.915	0.957	0.957
	TRANSFAC	0.874	0.940	0.782

To confirm that diPWM TFBS predictions are more reliable comparing to those of a traditional PWM we have applied both models to predict binding sites in each control

ChIP-Seq dataset. We demonstrate that diPWM predictions are located closer to the peak summits in the cases, when the predictions of diPWM and traditional PWM are different. The position of the peak summit indicates the most probable location of actual TF binding sites at DNA[22].

We examined the peaks where PWM and diPWM had their best hits located farter apart than 5bps. We plotted the number of such peaks with the predicted best-scoring binding site within the given distance from the peak summit (<25bp to the peak summit, <50bp to the peak summit etc.; the base coverage profile was missing for REST data so the centers of peak segments were used as the putative peak summit locations). The results are presented at the right panels of Figure 2, clearly showing that diPWMs select binding sites located closer to the ChIP-Seq peak summits. It is notable, that for HNF4A usage of diPWM again gives no advantage over the traditional PWM.

From the point of view of computation efficiency, the performance of diChIPMunk remained quite acceptable taking several hours to construct diPWMs from 500 ChIP-Seq peaks in a multi-threaded mode (Core i7 CPU). Since a dinucleotide model has more parameters to estimate, the default number of starting random seeds and subsampling runs for diChIPMunk were doubled as compared to ChIPMunk. diChIPMunk is free software implemented in Java and thus can be used on any platform. The source code and Java classes are freely available online (http://autosome.ru/dichipmunk/).

While more and more high-throughput data becomes available today, the improved TFBS models finally have a chance to be widely used giving a reasonable improvement over traditional PWMs. The dinucleotide PWMs produce better predictions for actual TF binding to DNA, which is important for DNA analysis *in silico* including identification of target genes regulated by particular TFs when the direct experiment is missing. Definitely, correlation pattern in TF binding sites is likely to be longer than for neighboring nucleotides; from physical considerations one can expect correlations extending up to the half of DNA helix pitch, which is about 5bp. Yet, from the practical point of view, dinucleotide models provide a reasonable tradeoff between the model complexity and the recognition accuracy, catching correlations brought about by stacking interaction, the important contributor into DNA stabilizing energy[8].

Acknowledgments

This work was supported by a Dynasty Foundation Fellowship [to I.K.]; Russian Foundation for Basic Research [12-04-32082-mol_a to I.K.] and [12-04-01736-a to D.O.]; Presidium of the Russian Academy of Sciences program in Cellular and Molecular Biology; Presidium of the Russian Academy of Sciences Fundamental Research Subprogram "Gene pools dynamics and conservation"; Russian Federation Government Contract 8088 under program "Scientific and pedagogical personnel of innovative Russia".

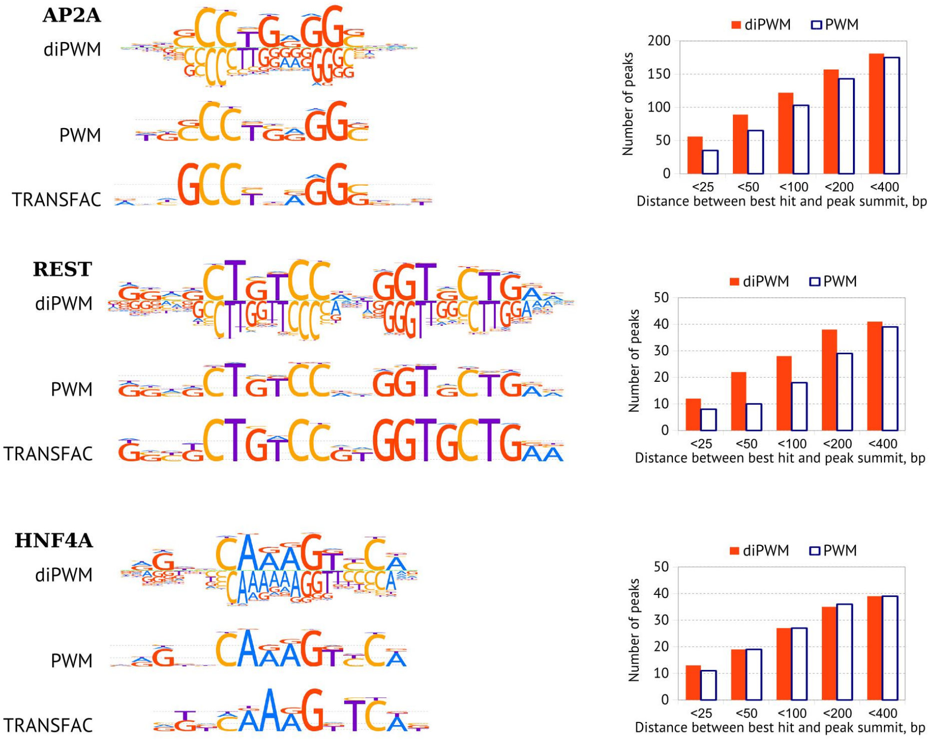

Fig. 2. Left panel: LOGO representation of ChIPMunk (PWM), diChIPMunk (diPWM) and TRANSFAC TFBS models; letters are scaled according to column KDIC values (KDIDIC for diPWMs). Right panel: cumulative distributions of distances between best hits of diPWM/PWM and the locations of peak summits (for each peak the best hit is defined as the position of binding site prediction with the highest diPWM/PWM score). Only ChIP-Seq peaks with differently located best hits of PWMs and diPWMs (5bps or farther) are considered.

References

1. Stormo GD, DNA binding sites: representation and discovery, *Bioinformatics* **16**(1):16-23, 2000.
2. Wasserman WW, Sandelin A, Applied bioinformatics for the identification of regulatory elements, *Nat Rev Genet* **5**(4):276-87, 2004.
3. Zambelli F, Pesole G, Pavesi G, Motif discovery and transcription factor binding sites before and after the next-generation sequencing era, *Brief. Bioinformatics* bbs016, 2012.
4. Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J, A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs, Supplementary table 1, *Nat Protoc.* **7**(8):1551-68, 2012.
5. Bi Y, Kim H, Gupta R, Davuluri RV, Tree-based position weight matrix approach to model transcription factor binding site profiles, *PLoS One* **6**(9):e24210, 2011.
6. Benos PV, Bulyk ML, Stormo GD, Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* **30**:4442-4451, 2002.

7. Gershenzon NI, Stormo GD, Ioshikhes IP, Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites, *Nucleic Acids Res.* **33**(7):2290-301, 2005.
8. Saenger W, *Principles of Nucleic Acid Structure*, Springer-Verlag, New York, 1984.
9. SantaLucia J Jr, Hicks D, The thermodynamics of DNA structural motifs, *Annu Rev Biophys Biomol Struct.* **33**:415-40, 2004.
10. Levitsky VG, Ignatieva EV, Ananko EA, Turnaev II, Merkulova TI, Kolchanov NA, Hodgman TC, Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions, *BMC Bioinformatics* **8**:481, 2007.
11. Oshchepkov DY, Vityaev EE, Grigorovich DA, Ignatieva EV, Khlebodarova TM, SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition, *Nucleic Acids Res.* **32**(Web Server issue):W208-12, 2004.
12. Zhao Y, Ruan S, Pandey M, Stormo GD, Improved models for transcription factor binding site identification using nonindependent interactions, *Genetics* **191**(3):781-90, 2012.
13. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ, Deep and wide digging for binding motifs in ChIP-Seq data, *Bioinformatics* **26**(20):2622-3, 2010.
14. Ma X, Kulkarni A, Zhang Z, Xuan Z, Serfling R, Zhang MQ, A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information, *Nucleic Acids Res.* **40**(7):e50, 2012.
15. Kuttippurathu L, Hsing M, Liu Y, Schmidt B, Maskell DL, Lee K, He A, Pu WT, Kong SW, CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments, *Bioinformatics* **27**(5):715-7, 2011.
16. Hertz GZ, Stormo GD, Identifying DNA and protein patterns with statistically significant alignments of multiple sequences, *Bioinformatics*, **15**(7-8):563-77, 1999.
17. Kulakovskiy IV, Makeev VJ, Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources, *Biophysics* **54**, 667-674, 2010.
18. ENCODE Project Consortium, A user's guide to the encyclopedia of DNA elements (ENCODE), *PLoS biol.* **9**: e1001046.
19. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E., TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res.* **34**(Database issue):D108-10, 2006.
20. Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ, to appear in *Nucleic Acids Res*.
21. Siddharthan R, Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix, *PLoS One* **5**(3):e9722, 2010.
22. Valouev, A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A, Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data, *Nat. Methods* **5**, 829–834, 2008.