ISSN 0006-2979, Biochemistry (Moscow), 2007, Vol. 72, No. 11, pp. 1187-1193. © Pleiades Publishing, Ltd., 2007.
Original Russian Text © T. I. Merkulova, D. Yu. Oshchepkov, E. V. Ignatieva, E. A. Ananko, V. G. Levitsky, G. V. Vasiliev, N. V. Klimova, V. M. Merkulov, N. A. Kolchanov, 2007, published in Biokhimiya, 2007, Vol. 72, No. 11, pp. 1459-1467.

REVIEW

Bioinformatical and Experimental Approaches to Investigation of Transcription Factor Binding Sites in Vertebrate Genes

T. I. Merkulova[1,2]*, D. Yu. Oshchepkov[1,2], E. V. Ignatieva[1,2], E. A. Ananko[1], V. G. Levitsky[1,2], G. V. Vasiliev[1], N. V. Klimova[1], V. M. Merkulov[1], and N. A. Kolchanov[1,2]

[1]Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, ul. Akademika Lavrentieva 10, 630090 Novosibirsk, Russia; fax: (3832) 331-278; E-mail: merkti@niboch.nsc.ru
[2]Novosibirsk State University, ul. Pirogova 2, 630090 Novosibirsk, Russia; fax: (3832) 339-101; E-mail: kol@bionet.nsc.ru

Received May 17, 2007
Revision received July 7, 2007

Abstract—The development of computer-assisted methods for transcription factor binding sites (TFBS) recognition is necessary for study the DNA regulatory transcription code. There are a great number of experimental methods that enable TFBS identification in genome sequences. The experimental data can be used to elaborate multiple computer approaches to recognition of TFBS, each of which has its own advantages and limitations. A short review of the characteristics of computer methods of TFBS prediction based on various principles is presented. Methods used for experimental monitoring of predicted sites are analyzed. Data concerning DNA regulatory potential and its realization at the chromatin level, obtained using these methods, are discussed along with approaches to recognition of target genes of certain transcription factors in the genome sequences.

DOI: 10.1134/S000629790711003X

Key words: genome annotation, recognition of transcription factor binding sites, computer and experimental methods, databases, training samples

The main elements of DNA regulatory transcription code are comprised of short (5-20 bp) sequences—regulatory elements specifically interacting with transcription factors (TFs). Various combinations of regulatory elements in promoter and enhancer gene regions and TF ensembles interacting with them define the level of gene expression depending on the stage of organism development, cell type, and external (environmental) effects [1-3]. The elaboration of methods for recognition of transcription factor binding sites (TFBSs) in nucleotide sequences is the first step on the way to developing methods for prediction of gene expression level in different situations based on the primary structure of its regulatory regions. TFBS recognition by computer methods is also necessary for revealing groups of coordinately regulated genes and reconstruction of the signal transduction pathways using experimental data on transcription changes obtained by methods of differential display, serial analysis of gene expression (SAGE), or analysis of gene expression using microarrays, as well as for solution of many other problems.

One of the most important problems is searching in the genome for target genes of a certain TF, which is especially important for studying mechanisms of the cell or organism response to various stimuli. Thus, in investigation of the response to hypoxia the revealing of the HIF1 target genes is necessary [4]. Studying the immune response, induced by interferon-α or -β, suggests a search for ISGF3 target genes [5]. The list of such examples could be continued.

The complexity of organization and high degeneracy of the regulatory transcription DNA code lead to serious difficulties in developing computer methods for TFBS recognition. An additional difficulty is connected with estimation of the reality of predicted sites including both the verification of a computer method and the evaluation of functionality of predicted sites, which is not the same. Theoretical approaches used for this purpose, such as

Abbreviations: ChIP) chromatin immunoprecipitation; EMSA) electromobility shift assay; GRE) glucocorticoid-responsive element; SAGE) serial analysis of gene expression; TF) transcription factor; TFBS) transcription factor binding site.

* To whom correspondence should be addressed.

comparative genomics [6-8] and evaluation of similarity of expression profiles of TF and TF-regulated genes [9], give us some insight into the functionality of predicted sites and only an indirect estimation of the accuracy of the recognition method. Direct experimental verification of predicted TFBSs is more efficient for evaluation of the method. However, there are only a few works concerning this problem [10-13], since they are rather labor-consuming and expensive because they require checking each predicted site. Also, it is only possible to carry them out under close collaboration of such scientists as specialists in bioinformatics elaborating methods for TFBS recognition and experimentalists capable of their testing.

The goals of this work were analysis of computer methods for TFBS prediction and the ways of their experimental monitoring, discussion of results obtained using experimentally verified computer methods in investigation of genome sequences, as well as discussion of methods for identification of target genes for certain TF.

COMPUTER APPROACHES

Sources of training samples for TFBS recognition techniques. The elaboration of methods for TFBS recognition is traditionally carried out on the basis of analysis of samples of TFBS sequences, for which the ability to interact with the TF and/or functionality were detected experimentally. Several specialized information resources contain eukaryotic TFBS sequences: TRANSFAC [2], TRED [14], TRRD [3], ooTFD [15], and MPromDb [16].

Usually such databases contain information concerning the extent of previous experimental study of these sites. These data are represented either as "quality" of the site (TRED, TRANSFAC) with indication of the experiment type (such as electromobility shift assay (EMSA), DNase I footprinting) or as digital code of the experiment (TRRD), which makes it possible to arrange sets of TFBS sequences meeting the user's requirements. No doubt the arrangement of good training samples of TFBSs requires proofs of the ability of each sequence from the sample to interact with TF. As a rule the following experiments are carried out to obtain such proofs: EMSA using the purified protein; DNase I footprinting with the purified protein; EMSA using nuclear extract and antibodies to transcription factor. When only data on the involvement of an element in gene expression regulation are used, the deterioration of training sample with binding sites of other TFs due to the existence of so-called tethering elements is possible. For example, well-known elements of glucocorticoid regulation (glucocorticoid-responsive elements) (GRE) directly interact with the glucocorticoid hormone receptors. However, there is a sufficiently large group of "joined" GRE, actually being sites for binding Fos/Jun [17], Stat5 [18], Smad3 [19], and some other TFs, with which the glucocorticoid receptor interacts via protein–protein interactions without contacting DNA.

Quite a number of information resources are known with already prepared TFBS matrix: TRANSFAC [2], JASPAR [20], and ARTSITE [21]. The TFBS matrix present in the above-mentioned databases can be divided into two types: designed on the basis of either natural (genomic) sequences of given TF binding sites, or artificial, selected *in vitro* by high affinity to a certain TF. The method for selection of sequences used as a basis for matrix deduction should be also considered in the development of recognition techniques.

In some situations, the TFBS sample may dissociate to sub-samples corresponding to structural variants of sites. For example, analysis of 160 glucocorticoid receptor binding sites from TRRD [3] has shown that only 54% of such sites are homologous to the palindrome GRE (AGAACAnnnTGTTCT), to which there binds a homodimer of receptor protein. The other 40% of sites are represented by hexanucleotide half-sites (TGTTCT), to which there binds the receptor monomer form. In this case, most such sites are involved in glucocorticoid regulation [22]. Functional sites of the estrogen receptor [23] and CTF1/NF1 [24] are also represented by two variants: a palindrome and a half-site. It is especially important to take into consideration the separation of sites into groups (site grouping) when sites for binding one and the same TF may be represented both by direct and inverted repeats of a conserved motif. Binding sites for androgen receptor [25] and SREBPs (sterol-responsive element binding proteins) [26] are examples of this. TFBS can also differ by the length of a spacer between half-sites. Thus, the RAR/RXR heterodimer binds to direct repeats AGGTCA with a spacer both of one and five nucleotides (direct repeats DR1 and DR5) [27], whereas PPAR/RXR interacts with sites that represent the different type direct repeats (DR1, DR0, and DR2) [27, 28]. The optimal length of a spacer dividing the inverted repeat TTTC...GAAA in the STAT1 binding sites is 3 bp, but sites with 2 bp spacers are also known [29, 30]. The absence of separation of samples of binding sites for such TFs to corresponding sub-samples may result in a significant (sometimes fatal) qualitative deterioration of many recognition techniques.

Computer technologies of TFBS recognition. Each TFBS is known to be able to interact specifically with regulatory DNA sequences differing by the context. The peculiarities of these interactions impose limitations on the DNA sequence exhibited in a partial conservativeness of the TFBS base sequence. Just this situation is the basis of most methods for TFBS prediction.

One of the most widespread strategies for searching putative sites is based on revealing coincidences with a weight matrix obtained from a training sample of experimentally confirmed TFBS sequences and describing the frequency distribution probability of four nucleotides in

each position [31]. This approach is justified by methods of statistical physics [32], is the natural development of searching for consensus sequences [33], and uses a suggestion concerning the independence of each nucleotide interaction with protein. Hundreds of the weight matrix variants were used for TFBS recognition. Thus, the popular MatInspector [34] uses >700 weight matrices from the TRANSFAC database [2]. However, it was shown that the quality of TFBS recognition by the weight matrix technique may differ significantly in different situations [35]. Also, it is assumed that the use of weight matrices produces many overpredictions [36]. A serious drawback of the weight matrix technique is its inability to account for variations in the spacer length and half-site orientation within TFBS [37]. Since most TFs interact with DNA as homo- and/or heterodimers, the use of weight matrices for TFBS recognition often requires separation of training samples to corresponding sub-samples. Nonobservance of this requirement can be responsible for significant deterioration of the recognition quality.

Detailed analysis of experimental data on different affinity of TF to its binding sites on DNA made doubtful the idea of a nucleotide-independent interaction with protein [35, 38-40]. It was shown that revealing weak correlations between nucleotides simultaneously with the use of weight matrices can improve the quality of recognition [41]. In particular, to achieve this, weight matrices based on the dinucleotide frequency are constructed [42]. However, since the alphabet of dinucleotides exceeds fourfold that of nucleotides, the large volume of training samples is necessary to obtain reliable results. At the present time, samples of such a volume can not be collected for many TFs.

As an alternative approach to creation of TFBS recognition methods, the stock of conformational and physicochemical properties of the DNA region defined by base sequence of the latter is used [43]. These properties are mainly responsible for the existence of TFBS because local conformation of DNA, determined by the context, is one of the main factors of specificity of DNA–protein interactions [44, 45]. In particular, it was shown on the example of MetJ TFBS analysis that accounting for the DNA helix properties together with traditional methods, based on context analysis, improves the quality of recognition [46]. The DNA conformational parameters can be determined both for dinucleotide and for the trinucleotide codes [47]. Thus, the SITECON method of TFBS recognition is based on the analysis of physicochemical properties of DNA double helix in the region of TFBS [48]. The use of dinucleotides parameters in the SITECON training automatically results in accounting for nucleotide interdependence upon interaction with protein. An unquestionable merit of this approach is that its construction does not require large training samples. However, as in the case of weight matrices, this method does not account for deletions or insertions in the site core sequence, as well as alterations in the half-site orientations [49].

The method of hidden Markov chains [50] can be used to model the variability in the half-site orientation and in the length of a spacer between them, to use matrix fragments upon construction of the method for searching potential insertions, to generate sequences in accordance with the model, and to evaluate the similarity of any model with the chosen one [51]. These parameters of the method have allowed one, in particular, to design a realistic approach to recognition of sites for binding the subclass NR1-2 and NR4-6 of nuclear receptors [52], organized as direct or inverted repeats of the TGACCT motif with a spacer varying from 0 to 9 bp [37]. The approach based on the methods of discriminant analysis is also not very sensitive to deletions and insertions [53, 54]. In the case of nucleotide sequences, the recognition is based on the analysis of their statistical characteristics. Such characteristics include oligonucleotide frequencies.

Such approaches as Gibbs sampler [55], principle of maximal plausibility [56, 57], neural networks [58], etc. are used for TFBS recognition. Their detailed review is given in [59, 60]. In this case the quality of recognition is comparable with that by the weight matrix technique.

EXPERIMENTAL APPROACHES USED FOR MONITORING OF TFBS PREDICTED BY COMPUTER METHODS

Practically the whole arsenal of experimental approaches elaborated for testing of TFBSs predicted by different computer methods in regulatory regions of individual genes is used for experimental monitoring of such sites as well as for revealing functional elements in regulatory regions. Thus, in numerous works different tools from this arsenal have been used for experimental verification of different TFBSs found in DNA sequences of regulatory regions using consensus or various Internet-available programs (TESS [61], MatInspector [34], MATCH [62], Matrix Search [63], etc.). However, only some experimental approaches appear to be suitable for immediate monitoring of accuracy and efficiency of developed computer methods of TFBS recognition. The main requirements for such approaches include the possibility of rapid analysis of several tens of sites, lack of ambiguity of the result interpretation, as well as relative economy in time and money of the method.

The EMSA fully meets these requirements. This method allows quick estimation of the TF under study binding to several tens of double-stranded oligonucleotides corresponding to predicted sites. The source of TF can be either nuclear extract of cells expressing the given protein (and then the method should be complemented with cross-competition or supershift assays) or the purified recombinant protein. Thus, the EMSA with

nuclear extract together with cross-competition analysis were successfully applied by Tronche et al. [10] for verification of predicted HNF1 binding sites. The search for putative sites was carried out using a weight matrix in annotated genomic sequences of vertebrates. Fifty-four potential sites underwent experimental monitoring and 52 of these were "strong" HNF1 sites. This means that the false-positives of the computer method proposed by the authors was less than 5%.

We have used the supershift assays for monitoring of SF1 binding sites. These sites were found in not yet studied genes of the mammalian steroidogenesis system by the SiteGA method based on detection of significant correlations between dinucleotide frequencies within local site regions [54] and by the SITECON technique revealing positions with a high level of conservativeness in the DNA double helix conformation and physicochemical properties [48]. In the first case, 15 of 18 predicted sites were confirmed, i.e. the overprediction was 17% [12]. In the second case, the ability to interact with SF1 was confirmed for all 18 predicted sites, which is indicative of a very low level of the overprediction [13].

Main advantage of the EMSA is the uniqueness of the interpretation of the result. The method registers only the fact of the presence or absence of TF binding to the predicted site, which makes possible the precise estimation of overprediction error and choosing appropriate parameters for the recognizing method that would allow establishing an optimal ratio of false-positives and false-negatives.

Another experimental approach to monitoring of TFBS predicted by computer methods is chromatin immunoprecipitation (ChIP). The method is based on fixation of DNA–protein interactions *in vivo* by formaldehyde cross-links and use of specific antibodies for isolation of complexes containing the protein under study [59]. The advantage of the method is a more pronounced, compared to the EMSA, approximation to estimation of the site functionality. However, estimation of the overprediction error (existing in any computer approach) by the method of chromatin immunoprecipitation is not as precise as that in the case of the EMSA. First, the TF under study can be cross-linked both to DNA and any other chromatin component, including other DNA-bound TF. Second, the use of ChIP makes impossible detection of site position within the 100-500 bp DNA fragment. Also, the method is much more labor consuming and expensive compared with EMSA. It requires selection of cell lines expressing both the factor under study and products of genes containing predicted sites, and it is often associated with the necessity of immunoprecipitation with antibodies to all members of TF family capable of interaction with the same sites.

Nevertheless, ChIP made possible successful monitoring of binding sites of several TFs predicted by different methods. Thus, the use of antibodies to six members of the E2F family confirmed the existence of the protein binding sites predicted using weight matrix in 10 genes involved in the cell proliferation control [11]. The same method confirmed Myc binding sites revealed by phylogenetic footprinting in the genes encoding enzymes of glycolysis [64].

DNA REGULATORY POTENTIAL AND ITS REALIZATION

The lack of ambiguity of interpretation of the result of the computer method control by EMSA makes possible the use of such verified methods for estimation of the number of TFBS under investigation in the genome. Thus, for HNF1 it is a single site per 10,000 bp [10], for SF1 it is 1.5 sites per 10,000 bp [12] or three sites per 10,000 [3]. This means that in the human genome there are hundreds of thousands of binding sites for these TFs. It is important to note that since the experimental checks have shown that only a few sites are false-positives (<5% HNF1 [10], from 0 to 17% SF1 [12, 13]), these estimations indicate the real presence of these TFBSs in the genomic sequences. Thus, the results point to a very high regulatory potential of eukaryotic DNA. The question of the realization of this potential, i.e. which TFBSs existing in the genome are involved in transcription regulation, is still open.

In this connection data obtained using a broad-scale variant of the chromatin immunoprecipitation assay ChIP-chip, theoretically allowing detection of the whole aggregate of these TFBSs, are of great interest [59]. These data show that tens of thousands of sites in the genome interact *in vivo* with the TF under study. Thus, ChIP-chip gives the estimate of 12,000 sites for Sp1, 25,000 for c-Myc [65], and 65,000 for p53 [66]. For CREB sites the estimate is 19,000 [67], which correlates with the estimation of the number of CREB molecules in a cell (40,000) [68].

Thus, there is a discrepancy of over one order of magnitude between regulatory potential of genomic DNA (hundreds of thousands of sites) and its realization (tens of thousands of sites). This discrepancy can be partially explained by the fact that in a concrete type of cells in a certain functional condition (which are used in the ChIP-chip experiment) only some sites are available for binding to TF in the chromatin context. In other cell types and under different situations, alterations in the chromatin state may result in the availability of another group of binding sites of the same TF. An additional factor responsible for such discrepancy is the existence of families of factors interacting with he same sites, whereas in the present-day works using ChIP-chip, usually binding of only a single member of the family is detected. However, even with accounting for these reasons, the discrepancy between the number of sites in genomic DNA and the

number of sites occupied by TFs *in vivo* is very high, which suggests that some fraction of TFBSs does not participate in transcription regulation. The existence of such "silent" sites is supported, in particular, by the discovery of sites for HNF1 and SF1 binding in the genes expressed in organs and tissues where neither these factors nor their homologs able to bind the same sites are expressed [10, 12]. This puts forward the problems of searching for ways to discriminate between "silent" sites and functioning ones, as well as of elucidating the question concerning the need for "silent" TFBSs.

FROM SITE RECOGNITION TO IDENTIFICATION OF TARGET GENES

The existence of any TFBS in the gene nucleotide sequence is a necessary but not sufficient condition to have this gene transcription under immediate control of this TF. As a rule, transcription regulation requires a series of additional conditions that may be different both for different TFs and for different genes. The elucidation of such conditions and their subsequent accounting in elaboration of computer methods for recognition of the different TF target genes is needed for successful work of these methods.

One similar condition is the cooperation of TF bound to its site on DNA with other TFs interacting with closely located sites. In some cases, this fact can be used successfully for recognition of certain TF target genes. Thus, during analysis of promoter regions of 62 genes of the immune system, regulated by NF-κB, the AP1, IK, IRF, and STAT, binding sites were revealed as the most frequent neighbors of NF-κB sites [69]. Using these data for selection of the NF-κB-regulated genes, products of which are supposedly involved in immune response, as well as in the course of estimation of the NF-κB site conservativeness in human and murine ortholog genes, the authors revealed 28 new genes of the immune system that are potential NF-κB targets [69]. In searching for NFAT potential target genes, expressed in activated T cells, the ability of NFAT to cooperate with AP1 as well as the inclination of the NFAT/AP1 composite elements to clustering were used [70]. In addition, approaches based both on simultaneous recognition of binding sites of several TF, interacting for carrying out a certain function, as well as on different combinations of such TFBSs, appeared to be a very profitable way for revealing functionally combined groups of genes. Thus, a model for recognition of regulatory regions of genes expressed only or mostly in liver was designed on the basis of simultaneous application of weight matrices for binding sites for HNF1, HNF3, HNF4, and C/EBP factors expressed mainly in this organ [71]. Similarly, recognition of binding sites for ISG3, STAT1, IRF1, and NF-κB, playing the key role in interferon production, as well as recognition of their different combinations was used as a basis of a method for searching for interferon-regulated genes in genomic sequences [72].

However, the approach to recognition of concrete TF target genes based on the simultaneous search for its partner binding sites, which together with the given TF provide for a certain functional response, is associated with a number of limitations. For example, this approach is difficult to use in searching for the TF target genes involved in control of a great variety of physiological processes and, as a result, having a great number of TF "partners", each of which is involved in regulation of a small number of genes. At the same time, the problems on revealing all target genes of such TFs emerge quite often. Thus, in regard to clinical problems, it is extremely important to reveal the great number of target genes of the PPAR TF family, synthetic ligands of which are widely used for the treatment of obesity and diabetes [73], as well as target genes of glucocorticoid receptor, whose natural and synthetic ligands are used as immunological depressants and anti-inflammatory drugs [74], whereas side-effects of such therapy remain largely unknown.

In such cases either clustering of concrete TF binding sites in the regulatory region [70] or certain regularities in the TFBS localization with respect to the transcription start site can be considered as additional conditions. In particular, based on studying such regularities in the arrangement of SF1 binding sites in its known target genes, we have suggested the following criterion for identification of potential target genes of this factor as the existence of at least one SF1 site in the region of (−300; +1) for the existence of at least one more SF1 site in the region of (−2100; −300) or (+1; +2100) [12]. Using this criterion, new SF1 target genes were revealed in the human genome among genes encoding cytokine receptors, growth factors involved in the subsequent steps of signal transduction, and among genes of the male reproductive system, which according to published data corresponds to physiological functions of these genes [12].

There are now many computer approaches to TFBS recognition; this is a necessary component of genome computer annotation and a basic instrument for analysis of gene expression data obtained with microarrays and required for solving many other problems of molecular biology, biochemistry, physiology, and molecular medicine. The main shortcoming of most of these approaches is a high level of overpredictions. However, the experimental verification of sites predicted by various methods is required to distinguish between overprediction error of a method and revealing "silent" TFBSs not participating in expression regulation.

Analysis of available data on experimental monitoring of some predicted TFBS using the EMSA has shown that their representation in the genome is very high, up to several hundreds of thousands per genome [10, 12, 13],

and only a small number of them are pseudopredicted (<5% for HNF1 [10], from 0 to 18% for SF1 [12, 13]). In this case, some TFBSs are found in genes not regulated by the protein under study. A factor decreasing the excessive binding of TF is the inaccessibility of extended DNA regions for binding due to the DNA compact package in the cell nucleus in the form of chromatin. There are several levels of chromatin structural organization that can influence the binding site accessibility to different TFs [75]. On the other hand, a hypothesis explaining the abundance of TFBSs in the genome might be considering the existence of "traps" for transcription factors, providing for local increase in their concentration in a necessary region.

The observed abundance in the TFBS content in genomic DNA as well as the complex nature of the transcription regulation mechanisms indicate that the use of additional criteria is necessary for searching for target genes of a certain transcription factor. These can be the regularities in the arrangement relative the transcription start, clustering the single type TFBS, and certain regularities in mutual arrangement.

This study was financially supported by the Russian Academy of Sciences Program No. 2 "Molecular and Cell Biology" (project No. 10.4), the Integration project of the Siberian Branch of the Russian Academy of Sciences No. 115, and the Russian Foundation for Basic Research (grant 05-07-98012).

REFERENCES

1. Latchman, D. S. (2004) in *Eukaryotic Transcription Factors*, Elsevier Academic Press, N. Y., pp. 299-330.
2. Matys, V., Kel-Margoulis, O. V., 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, A. E., and Wingender, E. (2006) *Nucleic Acids Res.*, **34**, D108-D110.
3. Kolchanov, N. A., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Stepanenko, I. L., Merkulova, T. I., Pozdnyakov, M. A., Podkolodny, N. L., Naumochkin, A. N., and Romashchenko, A. G. (2002) *Nucleic Acids Res.*, **30**, 312-317.
4. Wenger, R. H., Stiehl, D. P., and Camenisch, G. (2005) *Sci. STKE*, **306**, re1. 2.
5. Platanias, L. C. (2005) *Nat. Rev. Immunol.*, **5**, 375-386.
6. Ho Su, S. J., Mortimer, J. R., Arenillas, D. J., Brumm, J., Walsh, C. J., Kennedy, B. P., and Wasserman, W. W. (2005) *Nucleic Acids Res.*, **33**, 3154-3164.
7. Li, X., Zhong, S., and Wong, W. H. (2005) *Proc. Natl. Acad. Sci. USA*, **102**, 16945-16950.
8. Chang, L. W., Nagarajan, R., Magee, J. A., Milbrandt, J., and Stormo, G. D. (2006) *Genome Res.*, **16**, 405-413.
9. Holloway, D. T., Kon, M., and DeLisi, C. (2005) *Genome Infor.*, **16**, 83-94.
10. Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M., and Pontoglio, M. (1997) *J. Mol. Biol.*, **266**, 231-245.
11. Kel, A. E., Kel-Margoulis, O. V., Farnham, P. J., Bartley, S. M., Wingender, E., and Zhang, M. Q. (2001) *J. Mol. Biol.*, **309**, 99-120.
12. Klimova, N. V., Levitsky, V. G., Ignatieva, E. V., Vasiliev, G. V., Kobzev, V. F., Busygina, T. V., Merkulova, T. I., and Kolchanov, N. A. (2006) *Mol. Biol. (Moscow)*, **40**, 512-523.
13. Ignatieva, E. V., Klimova, N. V., Oshchepkov, D. Yu., Vasiliev, G. V., Merkulova, T. I., and Kolchanov, N. A. (2007) *Doklady Akad. Nauk*, **415**, 1-8.
14. Jiang, C., Xuan, Z., Zhao, F., and Zhang, M. Q. (2007) *Nucleic Acids Res.*, **35**, D137-D140.
15. Ghosh, D. (2000) *Nucleic Acids Res.*, **28**, 308-310.
16. Sun, H., Palaniswamy, S. K., Pohar, T. T., Jin, V. X., Huang, T. H., and Davuluri, R. V. (2006) *Nucleic Acids Res.*, **34**, D98-D103.
17. Jonat, C., Rahmsdorf, H. J., Park, K. K., Cato, A. C., Gebel, S., Ponta, H., and Herrlich, P. (1990) *Cell*, **62**, 1189-1204.
18. Stoecklin, E., Wissler, M., Moriggl, R., and Groner, B. (1997) *Mol. Cell. Biol.*, **17**, 6708-6716.
19. Song, C. Z., Tian, X., and Gelehter, T. D. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 11776-11781.
20. Vlieghe, D., Sandelin, A., de Bleser, P. J., Vleminckx, K., Wasserman, W. W., van Roy, F., and Lenhard, B. (2006) *Nucleic Acids Res.*, **34**, D95-97.
21. Khlebodarova, T., Podkolodnaya, O., Oshchepkov, D., Miginsky, D., Ananko, E., and Ignatieva, E. (2006) in *Bioinformatics of Genome Regulation and Structure II* (Kolchanov, N., and Hofestaedt, R., eds.) Springer Science+Business Media, Inc., N. Y. pp. 55-65.
22. Merkulov, V. M., and Merkulova, T. I. (2006) *Ekol. Genet. (Moscow)*, **4**, 20-31.
23. O'Lone, R., Frith, M. C., Karlsson, E. K., and Hansen, U. (2004) *Mol. Endocrinol.*, **18**, 1859-1875.
24. Roulet, E., Bucher, P., Schneider, R., Wingender, E., Dusserre, Y., Werner, T., and Mermod, N. (2000) *J. Mol. Biol.*, **297**, 833-848.
25. Schoenmakers, E., Alen, P., Verrijdt, G., Peeters, B., Verhoeven, G., Rombauts, W., and Claessens, F. (1999) *Biochem. J.*, **341**, 515-521.
26. Kim, J. B., Spotts, G. D., Halvorsen, Y. D., Shih, H. M., Ellenberger, T., Towle, H. C., and Spiegelman, B. M. (1995) *Mol. Cell. Biol.*, **15**, 2582-2588.
27. Khorasanizadeh, S., and Rastinejad, F. (2001) *Trends Biochem. Sci.*, **26**, 384-390.
28. Okuno, M., Arimoto, E., Ikenobu, Y., Nishihara, T., and Imagawa, M. (2001) *Biochem. J.*, **353**, 193-198.
29. O'Brien, C. A., and Manolagas, S. C. (1997) *J. Biol. Chem.*, **272**, 15003-15010.
30. Chen, H., Lee, J. M., Zong, Y., Borowitz, M., Ng, M. H., Ambinder, R. F., and Hayward, S. D. (2001) *J. Virol.*, **75**, 2929-2937.
31. Stormo, G. D., Schneider, T. D., and Gold, L. (1986) *Nucleic Acids Res.*, **14**, 6661-6679.
32. Berg, O. G., and von Hippel, P. H. (1987) *J. Mol. Biol.*, **193**, 723-750.
33. Mulligan, M. E., Hawley, D. K., and Entriken, R. (1984) *Nucleic Acids Res.*, **12**, 789-800.
34. Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005) *Bioinformatics*, **21**, 2933-2942.
35. Roulet, E., Fisch, I., Junier, T., and Mermod, N. (1998) *In Silico Biol.*, **1**, 21-28.

36. Tompa, M., Li, N., Bailey, T. L., Church, G. M., de Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005) *Nat. Biotechnol.*, **23**, 137-144.
37. Sandelin, A., and Wasserman, W. W. (2005) *Mol. Endocrinol.*, **19**, 595-606.
38. Man, T. K., and Stormo, G. D. (2001) *Nucleic Acids Res.*, **29**, 2471-2478.
39. Bulyk, M. L., Johnson, P. L., and Church, G. M. (2002) *Nucleic Acids Res.*, **30**, 1255-1261.
40. Barash, Y., Elidan, G., Kaplan, T., and Friedman, N. (2003) *RECOMB*, 28-37.
41. Zhang, M., and Marr, T. (1993) *Comput. Appl. Biosci.*, **9**, 499-509.
42. Gershenzon, N. I., Stormo, G. D., and Ioshikhes, I. P. (2005) *Nucleic Acids Res.*, **33**, 2290-2301.
43. Gorin, A. A., Zhurkin, V. B., and Olson, W. K. (1995) *J. Mol. Biol.*, **247**, 34-48.
44. Starr, D. B., Hoopes, B. C., and Hawley, D. K. (1995) *J. Mol. Biol.*, **250**, 434-446.
45. Meierhans, D., Sieber, M., and Allemann, R. K. (1997) *Nucleic Acids Res.*, **25**, 4537-4544.
46. Liu, R., Blackwell, T. W., and States, D. J. (2001) *Bioinformatics*, **17**, 622-633.
47. Arauzo-Bravo, M. J., and Sarai, A. (2005) *Genome Inform.*, **16**, 12-21.
48. Oshchepkov, D. Y., Vityaev, E. E., Grigorovich, D. A., Ignatieva, E. V., and Khlebodarova, T. M. (2004) *Nucleic Acids Res.*, **32**, W208-W212.
49. Oshchepkov, D. Yu., Turnaev, I. I., Pozdnyakov, M. A., Milanesi, L., Vityaev, E. E., and Kolchanov, N. A. (2004) in *Bioinformatics of Genome Regulation and Structure* (Kolchanov, N., and Hofestaedt, R., eds.) Kluwer Academic Publishers, Boston-Dordrecht-London, pp. 93-102.
50. Durbin, R., Eddy, S., and Krogh, A. G. M. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge.
51. Eddy, S. R. (1998) *Bioinformatics*, **14**, 755-763.
52. Laudet, V., Auwerx, J., Gustafsson, J., and Wahli, W. (1999) *Cell*, **97**, 161-163.
53. Gunewardena, S., Jeavons, P., and Zhang, Z. (2006) *J. Comput. Biol.*, **13**, 929-945.
54. Levitsky, V. G., Ignatieva, E. V., Ananko, E. A., Merkulova, T. I., Kolchanov, N. A., and Hodzhman, T. S. (2006) *Biofizika*, **51**, 633-639.
55. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993) *Science*, **262**, 208-214.
56. Bailey, T. L., and Elkan, C. (1995) *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 1-29.
57. Grundy, W. N., Bailey, T. L., Elkan, C. P., and Baker, M. E. (1997) *Comput. Appl. Biosci.*, **13**, 397-406.
58. O'Neill, M. C. (1991) *Nucleic Acids Res.*, **19**, 313-318.
59. Elnitski, L., Jin, V. X., Farnham, J., and Jones, J. (2006) *Genome Res.*, **16**, 1455-1464.
60. Gelfand, S. (1995) *J. Comput. Biol.*, **2**, 87-115.
61. Stoeckert, C. J., Jr., Salas, F., Brunk, B., and Overton, G. C. (1999) *Nucleic Acids Res.*, **27**, 200-203.
62. Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003) *Nucleic Acids Res.*, **31**, 3576-3579.
63. Chen, Q. K., Hertz, G. Z., and Stormo, G. D. (1995) *Comput. Appl. Biosci.*, **11**, 563-566.
64. Kim, J. W., Zeller, K. I., Wang, Y., Jegga, A. G., Aronow, B. J., O'Donnell, K. A., and Dang, C. V. (2004) *Mol. Cell. Biol.*, **24**, 5923-5936.
65. Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. (2004) *Cell*, **116**, 499-509.
66. Wei, C. L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., Liu, J., Zhao, X. D., Chew, J. L., Lee, Y. L., Kuznetsov, V. A., Sung, W. K., Miller, L. D., Lim, B., Liu, E. T., Yu, Q., Ng, H. H., and Ruan, Y. (2006) *Cell*, **124**, 207-219.
67. Euskirchen, G., Royce, T. E., Bertone, P., Martone, R., Rinn, J. L., Nelson, F. K., Sayward, F., Luscombe, N. M., Miller, P., Gerstein, M., Weissman, S., and Snyder, M. (2004) *Mol. Cell. Biol.*, **24**, 3804-3814.
68. Hagiwara, M., Brindle, P., Harootunian, A., Armstrong, R., Rivier, J., Vale, W., Tsien, R., and Montminy, M. R. (1993) *Mol. Cell. Biol.*, **13**, 4852-4859.
69. Liu, R., McEachin, R. C., and States, D. J. (2003) *Genome Res.*, **13**, 654-661.
70. Kel, A., Kel-Margoulis, O., Babenko, V., and Wingender, E. (1998) *J. Mol. Biol.*, **288**, 353-376.
71. Krivan, W., and Wasserman, W. W. (2001) *Genome Res.*, **11**, 1559-1556.
72. Ananko, E. A., Kondrakhin, Y. V., Merkulova, T. I., and Kolchanov, N. A. (2007) *BMC Bioinformatics*, **8**, 56.
73. Lemay, D. G., and Hwang, D. H. (2006) *J. Lipid Res.*, **47**, 1583-1587.
74. Le Phuc, P., Friedman, J. R., Schug, J., Brestelli, J. E., Parker, J. B., Bochkis, I. M., and Kaestner, K. H. (2005) *PLoS Genet.*, **1** (2):e16 (0159-170).
75. Beato, M., and Eisfeld, K. (1997) *Nucleic Acids Res.*, **25**, 3559-3563.