ISSN 0006-3509, Biophysics, 2011, Vol. 56, No. 1, pp. 114–116. © Pleiades Publishing, Inc., 2011.
Original Russian Text © I.V. Kulakovskiy, A.S. Kasianov, A.A. Belostotsky, I.A. Eliseeva, V.J. Makeev, 2011, published in Biofizika, 2011, Vol. 56, No. 1, pp. 136–139.

LETTERS TO THE EDITOR

Preferred Distances between Transcription Factor Binding Sites

I. V. Kulakovskiy[a,b], A. S. Kasianov[a], A. A. Belostotsky[b], I. A. Eliseeva[c], and V. J. Makeev[b]

[a] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia*
[b] *Research Institute for Genetics and Selection of Industrial Microorganisms, Moscow, 117545 Russia*
[c] *Institute for Protein Research, Russian Academy of Sciences, Pushchino, 142290 Russia*
E-mail: alexbel.monster@gmail.com, ivan.kulakovskiy@gmail.com

Received July 6, 2010

Abstract—Transcriptional regulation of gene expression in higher eukaryotes is driven by elaborate protein complexes of transcription factors. At the DNA level, these complexes interact with composite elements consisting of specific binding sites for different proteins. We use the hypoxia-response system to identify preferred localization distances between "hypoxia-induced factor-1 – cofactor" binding site pairs in promoter DNA regions of the human genome. Such characteristic co-localization distances agree with a supposed scale of regulatory regions while being significantly longer than the typical binding site length. We speculate that this phenomenon can provide a key to decipher the structure of DNA regulatory regions in higher eukaryotes.

Keywords: transcription factor, binding sites, co-localization, hypoxia, Homo sapiens
DOI: 10.1134/S0006350911010155

INTRODUCTION

Adaptation of eukaryotic cells to hypoxic conditions depends on expression of a huge number of genes responsible for angiogenesis, cell migration, energy metabolism, cell growth and apoptosis. The hypoxia-induced transcription factor HIF-lα forms a complex with the ARNT protein and other cofactors to activate the transcription of target genes. At the DNA level, the typical binding site for the HIF-lα:HIF-lβ(ARNT) dimer is known as HRE (hypoxia-responsive element). HRE is a comparably short DNA segment containing the highly conserved TACGTG consensus. The short length of the consensus sequence and its high similarity with the AHR motif corresponding to the xenobiotic response system (CACGTA consensus, XRE element [2]) implies that some additional mechanisms are required to specifically control gene expression for HRE and XRE systems. One such mechanism can be specific binding of a selected set of transcription factors (TFs). The regulatory DNA segments are expected to contain corresponding binding sites at specific distances from each other. TRANSCOMPEL [3] database illustrates the effect of close co-localization of binding sites for distances comparable to the lengths of binding motifs. Another research [4] shows the co-localization effect for consensus sequences, but it is not focused on a set of transcription factors for a specific regulatory system.

DATA AND METHODS

Building the set of DNA segments. To construct a set of DNA segments containing regulatory regions, we used the hg18 human genome annotation [http://genome.ucsc.edu/]. We extracted the regions around the 43854 annotated transcription starts (TSS). We searched for HREs within segments of 20000 bp around the TSS (i.e. 10000 bp upstream and downstream). We searched for cofactor binding sites within segments of 22000 base pairs (i.e. 11000 bp in both directions). The maximum distance between HRE and cofactor binding sites was set at 1000 bp.

Construction of the binding motif models. We used the set of transcription factors (TFs) from the system regulating HIF-1-induced expression of erythropoietin (EPO). This set of TFs consists of HNF4α, SMAD3, SMAD4, p300, and Spl [5]. To construct DNA binding motifs, we used data extracted from TRANSFAC database [3]. To make a correct model for HRE, we additionally used annotated data from [6] and additional ChIP-chip data [7]. Sequence sets for each TF were supplied to the ChIPMunk algorithm [8] to construct the optimal gapless local multiple alignment with the maximum motif length limited by 20 bp. The corresponding motif logos are presented in Fig. 1. For the motif model we used the positional weight matrix (PWM) normalized for background nucleotide frequencies estimated from the human genome. The PWM thresholds were taken as the mean

plus three standard deviations for the PWM score distribution over all words of fixed length [9].

Identification of positional preferences. To identify preferred distances between pairs of sites, we counted the number of sequences (i.e. the number of corresponding genes) having at least one occurrence of the "HRE – cofactor binding site" pair. Using such strategy, one can partially avoid systematic noise from satellite repeats or fragments of coding regions.

RESULTS AND DISCUSSION

For all HIF-1α cofactors studied, we have found that there are preferred co-localization distances for "HRE - cofactor binding site" pairs. An example of such positional preference is shown in Fig. 2. One can see that the preferred distances form relatively short continuous series (i.e. "peaks"). The hypothesis that the observed positional preferences correspond to the functional co-localization of binding sites is supported by various observations. These include the relatively small width of the peaks, their irregular positioning relative to each other, the absence of strict periodic behavior (showing no direct linkage to satellite repeats), and the dramatic differences in distance preferences depending on the orientation of the sites (which is probably directly related to the spatial orientation of binding of corresponding proteins). The table lists the characteristic positional preferences for HRE and binding motifs for HIF-1α cofactors.

Specific intersite distances appear to be connected either with direct interaction between corresponding proteins (for short distances ~10 bp) or indirect interaction through adapter proteins (for distances of a few dozen base pairs), or with specific chromatin structures (nucleosomes or chromatin loops) providing "remote" direct or indirect interaction. All these cases suggest the existence of an optimal structure for the protein complex defined by the localization of binding sites within the regulatory DNA segment.

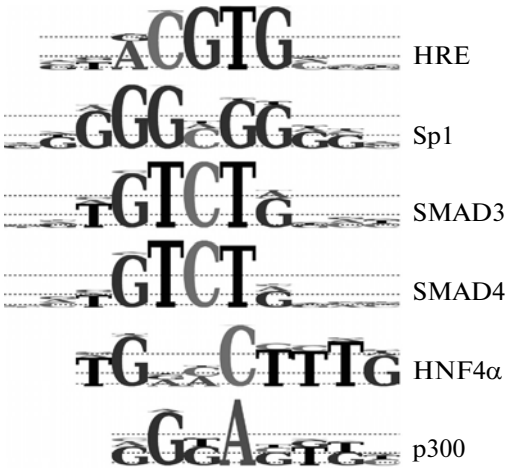

Fig. 1. Motif logos for HRE and cofactors. Discrete Information Content [8] used for letter scaling. The number of binding site-containing sequences extracted from the TRANSFAC and used for the motif construction: Sp1 – 1223, SMAD3/4 – 41/56, HNF4α – 190, p300 – 18 sequences. For HRE – 453 sequences including 363 sequences identified by ChIP-chip.

Localization of HREs relative to binding sites of known HIF-1α cofactors (including p53, BRCA1, JUN, STAT3) as well as that for XRE allows theorizing that similar patterns in distances between binding sites could be observed for other pairs of interacting TFs.

It is important to note that preferred distance templates are very sensitive to small changes in binging site motif models (as shown for SMAD3/4) which puts increased requirements on PWM quality.

We believe that usage of the preferred distance templates can reasonably improve in silico recognition of the regulatory regions using existing motif models. This can be achieved by reducing the false-positive rate by taking into account only properly positioned pairs of binding sites.

Positional preferences in co-localization of HRE versus cofactor binding sites depending on their orientation

Site pair	+	–
HRE-HRE	–10	–124, 10
HRE-SMAD3	–252, –120, –36, –31, 13, 131	–257, –139, –95, 113, 127
HRE-SMAD4	–41, –32,13, 104, 125, 135	–261, –126, –93, –74, 40
HRE-Sp1	–98, –91, –88, 27, 37, 43, 47, 52, 59, 66, 69	–194, –163, –134, –126, –62, –34, –38, –28, 11, 16
HRE-p300	–77, 67, –22, 58, 63, 68	–56, –24, 79
HRE-HNF4α	–117, –107, –49, 19, 28, 51, 61	–185, –85, –75, –17, 56, 118

Notes: (+) direct orientation of cofactor binding sites relative to HRE; (–) reverse orientation. The numbers correspond to the relative coordinates of peak 'summits' for peaks higher than the mean plus 3 SD. We counted the number of genes containing the selected pair at a given distance from the [–1000; 1000] range. It is notable that very similar SMAD3/4 motifs show huge differences in their binding preferences relative to HRE.

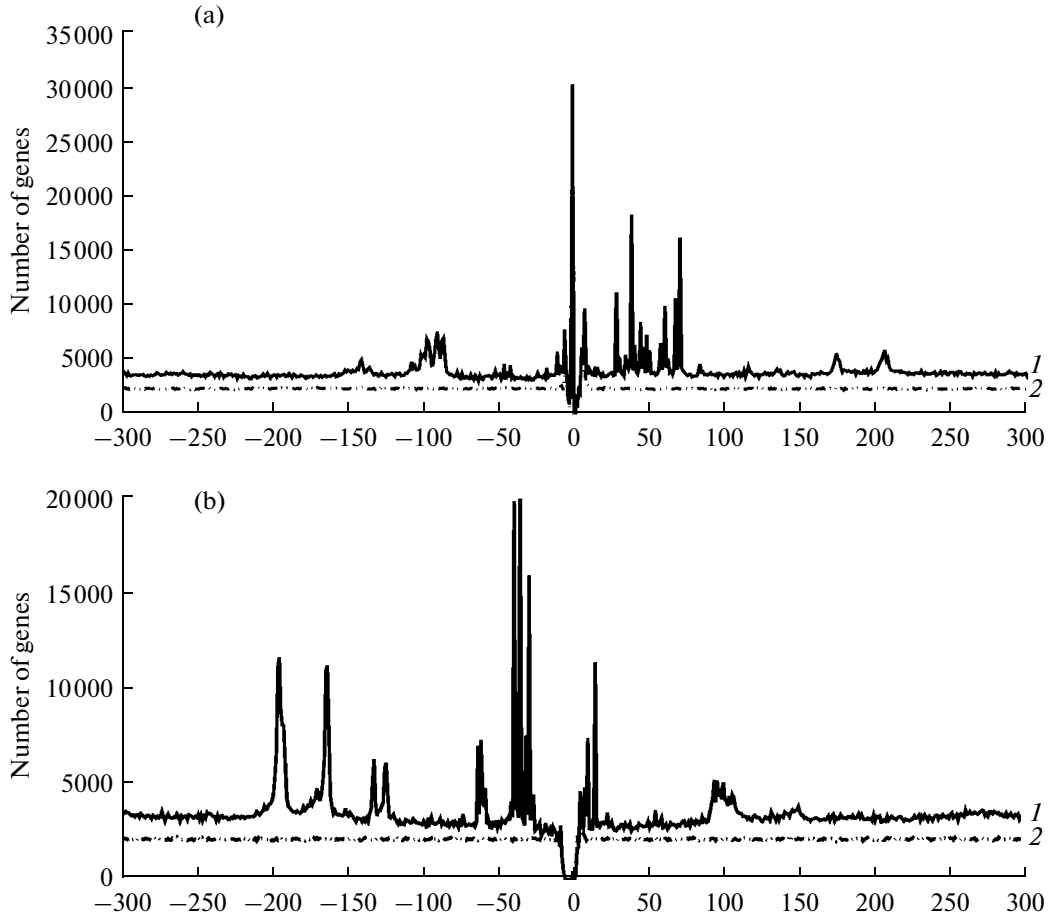

Fig. 2. Distance distribution between HRE and Sp1 binding sites. Y axis shows the number of genes for which at least one pair of binding sites is found at the selected distance (shown at X-axis, [−300;300] range) within the 10-kb upstream promoter region (*1*, solid). The control curve (*2*, dashed) corresponds to the sequence set derived by shuffling letters in each sequence of the initial set. Orientation of HRE in reference to Sp1: (a) direct; (b) reverse. The region around zero on X-axis corresponds to the overlapping of motifs. Here we present position preferences for the direct HRE orientation relative to TSS. For the reverse HRE orientation the graphs are symmetrically mirrored over the Y axis.

ACKNOWLEDGMENTS

This research is supported by the Russian Foundation for Basic Research (10-04-92663) and "Molecular and Cell Biology" program of the Presidium of Russian Academy of Sciences.

REFERENCES

1. R. H. Wenger, D. P. Stiehl, and G. Camenisch, Sci. STKE, **306**, re12 (2005).
2. H. I. Swanson, W. K. Chan, and C. A. Bradfield, J. Biol. Chem. **270**, 26292 (1995).
3. V. Matys, O. V. Kel-Margoulis, E. Fricke, et al., Nucleic Acids Res. **34**, D108 (2006).
4. K. D. Yokoyama, U. Ohler, and G. A. Wray, Nucleic Acids Res. **37**, e92 (2009).
5. T. Sanchez-Elsner, J. R. Ramirez, F. Sanz-Rodriguez, et al., J. Mol. Biol. **336**, 9 (2004).
6. A. Ortiz-Barahona, D. Villar, N. Pescador, et al., Nucleic Acids Res. **38**, 2332 (2010).
7. X. Xia, M. E. Lemieux, W. Li, et al., Proc. Natl. Acad. Sci. USA **106**, 4260 (2009).
8. I. V. Kulakovskiy and V. J. Makeev, Biophysics **54**, 667 (2009).
9. I. V. Kulakovskiy, A. V. Favorov, and V. J. Makeev, Bioinformatics **25**, 2318 (2009).