I.V. Kulakovskiy[1,2,*], A.A. Belostotsky[2], A.S. Kasianov[1], I.A. Eliseeva[4], V.J. Makeev[2,3]

(1) Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilov str. 32, Moscow 119991, Russia;

(2) Research Institute for Genetics and Selection of Industrial Microorganisms, 1st Dorozhny proezd 1, Moscow 117545, Russia;

(3) Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina str. 3, Moscow 119991, Russia;

(4) Institute for Protein Research, Russian Academy of Sciences, Institutskaya str. 4, Pushchino 142290, Russia.

E-mail: ivan.kulakovskiy@gmail.com

PREFERRED PAIR DISTANCE TEMPLATES REVEAL FUNCTIONAL TRANSCRIPTION FACTOR BINDING SITES

The regulatory code controlling gene expression in higher eukaryotes still remains unclear. It is a complex task to understand how a one-dimensional DNA text of multiple possibly overlapping “words” directs formation of protein complex that controls gene expression in specific tissues in specific conditions. Some insight is given by the well-known concept of “composite elements” consisting of binding sites for different regulatory proteins separated by specific distances [Matys2006]. Despite more than 15 years of study, information about possible intersite distance scale and specificity remains fragmentary.

The quality of identification of transcription factor binding sites (TFBS) both *in vitro* and *in vivo* has increased dramatically with the advent of new technologies like ChIP-Seq. Recently it was observed [Yokoyama2009], [Shelest2010] that there are several preferred distances between some pairs of TFBS (see esp. Fig. 7A in [Yokoyama2009] for pairs of NFY binding sites). Here we illustrate that this phenomenon appears to be much more common at least in the case of *Homo sapiens* TFs involved into regulation of response for hypoxia conditions. We believe this phenomenon can be used to distinguish functional binding sites from false positives, for both experimental identification of TFBS and their prediction *in silico*.

We studied distributions of distances between TFBS identified *in silico* for TFs involved in known protein-protein interactions. As a case study we took TFs participating in the regulation of the erithropoetin (EPO) gene expression in hypoxia response in human cells.

Materials and methods

The data used in the study

The genome-wide set of regulatory sequences (GW) was constructed by taking segments of 3000bp centered around transcription starts (TSS) for all annotated genes for UCSC hg18 human genome. Regulatory segments overlapping for more than 50% were merged together (a total of 36271 segments). Similar masked genome set (MG) was created from the hg18 genome assembly with masked exons, repBase repeats and fuzzy tandem repeats [Boeva2006]. The positive set (PS) contained the subset of all regulatory segments taking only those corresponding to known hypoxia-dependent genes (a total of 156 sequences) [Ortiz-Barahona2010].

We constructed the binding motifs for the HIF-1α:ARNT dimer, HNF4α, SMAD3, SMAD4, p300 and Sp1 proteins involved in the hypoxia-dependent regulation of the EPO expression [Sánchez-Elsner2004]. TRANSFAC [Matys2006] database was used as a source of binding site data. A position weight matrix (PWM) was adopted for a motif model. We did not use predefined TRANSFAC motifs but instead constructed PWMs with the help of our high-performance ChIPMunk tool [Kulakovskiy2010]. For the HIF-1α:ARNT dimer binding motif (known as the hypoxia responsive element, HRE) we also incorporated additional pregenomic and ChIP-chip data [Ortiz-Barahona2010, Xia2009]. The PWM thresholds were selected as the mean plus 3 SD for PWM score distribution over all possible words of a fixed length. The motif logos are presented in **Figure 1**.

A HRE motif, the principle DNA element controlling hypoxia response, was searched in the 2400bp long DNA segments centered at TSS. HIF-1α cofactor binding motifs were

searched in 600 bp windows centered at each putative HRE.

Preferred pair distance distributions

We used the strategy described in [Kulakovskiy2011] to evaluate preferred distances between the HRE and cofactor binding motifs. Basically we counted the number of sequences (i.e. the approximate number of genes) where the binding site of a selected TF was located in a given orientation at a selected distance from HRE. The corresponding positional pair distance distribution (PPDD) for the “HRE-reverse complementary Sp1” pairs are given in **Figure 2**. It displays a somewhat noisy background with a set of markedly exhibited peaks at a number of selected distances. It is noteworthy that the MG sequence set (produced by masking the genome from exons, repBase repeats and fuzzy tandem repeats) shows very similar distribution of the peaks in PPDD. It is noteworthy, that such distribution is very similar for motifs from different sources, e.g. for HRE and Sp1 taken from SwissRegulon database [Pachkov2007] (data not shown).

Our strategy has two advantages. Firstly, the genome sequence set provides the statistically representative set of possible distances between site pairs. Secondly, when the number of sequences containing a site pair is counted rather than the number of site pairs *per se* the final result becomes relatively undistorted by contributions from homotypic clusters [Lifanov2003] and repetitive DNA regions.

Preferred pair distance templates

Figure 2 displays preferred distances forming a comb of well-defined 'peaks'. Additionally, PPDD curve exhibit some general trend decreasing from center to edge. The significance of this trend depends on the motif lengths, PWMs and PWM thresholds, and the nucleotide composition of sequence segments in the set. In contrast, the principle peaks usually withstand changes of these parameters. Thus, some 'peak extraction' procedure is needed to distinguish significant peaks from the variable background. We did this by identifying extremal points of a numerical derivative averaged over 3-points (using PPDD also averaged over 3-points). Peaks having the derivative values higher than its mean+SD were selected. We did not take into account the peak heights because we did not use any detrending. Thus, we extracted the set of positions covered by significant PPDD peaks and call it as the Preferred Pair Distance Template, PPDT. PPDT refers to the set of valid intersite distances (i.e. preferred spacers) for a selected pair of TFs.

Statistical significance of binding site pairs

To check whether PPDT is related to functional TFBS arrangements we used the positive sequence set containing regulatory regions for hypoxia-dependent genes. For each DNA strand we independently counted the total number of “HRE-cofactor binding site” pairs and the number of such pairs having a spacer corresponding to one of the PPDT distances. Assuming the distances between TF pairs to be independent random events we counted the P-value as the probability to observe no lesser than the given number of pairs with PPDT distances using 600bp windows centered at any of HREs. The P-values were calculated using the binomial distribution. To obtain the overall sequence-related P-value we multiplied P-values for two mutual “HRE-cofactor binding site” orientations at both DNA strands. As a baseline we used the PS set with the randomly generated spacers and a random subset of regulatory regions. The corresponding graph is shown in **Figure 3**. **Table 1** shows the PPDT listing for the PWMs of cofactors regulating the hypoxia-controlled EPO expression.

Discussion

Preferred distances between TFBS seem to be related either to the direct interaction between

TFs (for short distances around 10bp) or with the indirect interaction via adapter proteins (for medium distances around tens of base pairs) or with particular chromatin structures (nucleosomes or chromatin loops) mediating direct or indirect interaction of distant TFs. In all these cases formation of the protein complex is facilitated by particular positioning of TFBS within the DNA segment.

PPDDs constructed for different TF pairs exhibit different characteristic sets of preferred distances, but in all cases a general pattern of a preferred peak comb over a background of more or less random distances is observed. We believe, that it is very likely that binding sites found at “wrong distances” either form complexes with TFs other than HIF-1α or simply are false positives of PWM scanning. What is important is that this approach provides additional information allowing one to distinguish functional TFBS pairs from irrelevant ones. Currently we explore the potential of PPDD/T for recognition of DNA segments binding regulatory TF complexes, and thus for reconstruction of regulatory genetic networks by means of sequence-analysis. However, the difficulties of this approach should not be underestimated, because of complex arrangements of binding sites, often overlapping each other.

Acknowledgments and funding

We thank Alexander Kel for sharing the access to the TRANSFAC release 2010.1 and Dmitrijs Lvovs for reading and commenting on the manuscript. This study was supported by the Presidium of the Russian Academy of Science program in Cellular and Molecular Biology.

Table 1. *PPDT spacers list for the HIF-1 cofactors in two possible orientation relative to HRE.* -0/+0 refers to the cofactor binding site directly to the left/to the right of HRE.

p300 direct	[-289:-286], [-246:-243], [-81:-76], [-72:-65], [-63:-55], [-17:-12], [-2:-0], [+55:+63], [+126:+127], [+286:+289]
p300 rc	[-51:-43], [-18:-14], [+65:+71], [+122:+123]
HNF4α direct	[-111:-105], [-101:-93], [-56:-52], [-43:-35], [+7:+11], [+15:+23], [+39:+41], [+48:+54], [+73:+78], [+93:+97]
HNF4α rc	[-178:-173], [-69:-63], [-11:-5]
SMAD3 direct	[-287:-284], [-246:-233], [-113:-105], [-77:-75], [-67:-63], [-33:-29], [-27:-17], [-6:-0], [+0:+10], [+70:+71], [+80:+82], [+95:+97], [+109:+114], [+117:+123], [+136:+138], [+243:+246], [+287:+287]
SMAD3 rc	[-248:-242], [-176:-175], [-140:-138], [-130:-125], [-114:-109], [-87:-80], [-61:-58], [+0:+8], [+25:+33], [+104:+108], [+114:+120], [+152:+153], [+197:+198], [+248:+255], [+285:+287]
SMAD4 direct	[-287:-286], [-53:-51], [-33:-27], [-25:-14], [+0:+6], [+89:+97], [+113:+115], [+122:+128]
SMAD4 rc	[-253:-246], [-118:-113], [-87:-79], [-62:-58], [+26:+34], [+117:+118]
Sp1 direct	[-287:-286], [-132:-131], [-92:-85], [-83:-73], [+14:+20], [+24:+40], [+44:+52], [+54:+62], [+160:+164], [+285:+287]
Sp1 rc	[-187:-177], [-155:-148], [-123:-121], [-115:-110], [-53:-45], [-30:-26], [-24:-21], [-19:-14], [+0:+2], [+4:+9]

References

[Boeva2006] Bioinformatics. 2006 Mar 15;22(6):676-84. Epub 2006 Jan 10. Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. Boeva V, Regnier M, Papatsenko D, Makeev V.

[Kulakovskiy2010] Bioinformatics. 2010 Oct 15;26(20):2622-3. Epub 2010 Aug 24. Deep and wide digging for binding motifs in ChIP-Seq data. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ.

[Kulakovskiy2011] [In Russian] Biofizika/Biophysics. 2011; 56(1):136-9. Preferred distances between transcription factor binding sites. Kulakovskiy IV, Kasianov AS, Belostotsky AA, Eliseeva IA, Makeev VJ.

[Lifanov2003] Genome Res. 2003 Apr;13(4):579-88. Homotypic regulatory clusters in Drosophila. Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA.

[Matys2006] Nucleic Acids Res. 2006 Jan 1;34(Database issue):D108-10. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. 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.

[Ortiz-Barahona2010] Nucleic Acids Res. 2010 Apr;38(7):2332-45. Epub 2010 Jan 8. Genome-wide identification of hypoxia-inducible factor binding sites and target genes by a probabilistic model integrating transcription-profiling data and in silico binding site prediction. Ortiz-Barahona A, Villar D, Pescador N, Amigo J, del Peso L.

[Pachkov2007] Nucleic Acids Res. 2007 Jan;35(Database issue):D127-31. Epub 2006 Nov 27. SwissRegulon: a database of genome-wide annotations of regulatory sites. Pachkov M, Erb I, Molina N, van Nimwegen E.

[Sánchez-Elsner2004] J Mol Biol. 2004 Feb 6;336(1):9-24. A cross-talk between hypoxia and TGF-beta orchestrates erythropoietin gene regulation through SP1 and Smads. Sánchez-Elsner T, Ramírez JR, Sanz-Rodriguez F, Varela E, Bernabéu C, Botella LM.

[Shelest2010] Bioinformatics. 2010 Jun 1;26(11):1460-2. Epub 2010 Mar 25. DistanceScan: a tool for promoter modeling. Shelest V, Albrecht D, Shelest E.

[Xia2009] Genome Biol. 2009;10(10):R113. Epub 2009 Oct 14. Preferential binding of HIF-1 to transcriptionally active loci determines cell-type specific response to hypoxia. Xia X, Kung AL.

[Yokoyama2009] Nucleic Acids Res. 2009 Jul;37(13):e92. Epub 2009 May 29. Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. Yokoyama KD, Ohler U, Wray GA.

Figure 1. The motif logos for the PWMs used in the study (direct orientation).

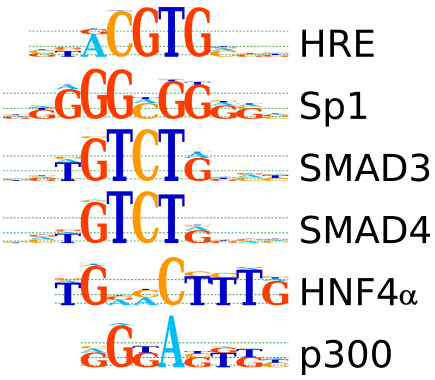

Figure 3. The statistical significance of the HRE-Sp1 spacers distribution. The sequences are independently sorted by corresponding P-value. See details in text.

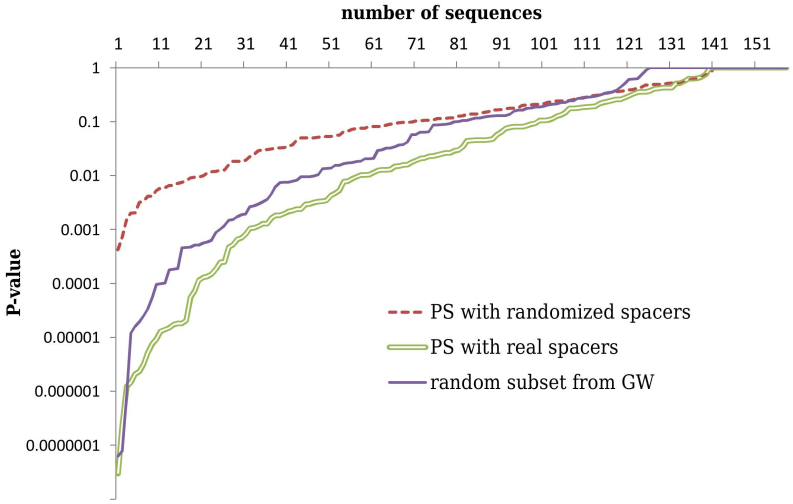

Figure 2. The PPDD and PPDT peaks for the HRE - “reverse complement Sp1” binding site pair. PPDDs from three different sequence sets are shown; the curves are normalized for their maximum. X axis displays the length of the spacer. See details in text.