BIOINFORMATICS DISCOVERY NOTE Vol. 27 no. 19 2011, pages 2621–2624
doi:10.1093/bioinformatics/btr453

Sequence analysis

Advance Access publication August 18, 2011

A deeper look into transcription regulatory code by preferred pair distance templates for transcription factor binding sites

I. V. Kulakovskiy[1,2,*], A. A. Belostotsky[2], A. S. Kasianov[1], N. G. Esipova[1], Y. A. Medvedeva[2,3,†], I. A. Eliseeva[4] and V. J. Makeev[2,3]

[1]Laboratory of Bioinformatics and System Biology, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, [2]Laboratory of Bioinformatics, Research Institute for Genetics and Selection of Industrial Microorganisms, Moscow 117545, [3]Laboratory of Systems Biology and Computational Genetics, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow 119991 and [4]Group of Protein Biosynthesis Regulation, Institute of Protein Research, Russian Academy of Sciences, Pushchino 142290, Russia

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Modern experimental methods provide substantial information on protein–DNA recognition. Studying arrangements of transcription factor binding sites (TFBSs) of interacting transcription factors (TFs) advances understanding of the transcription regulatory code.

Results: We constructed binding motifs for TFs forming a complex with HIF-1α at the erythropoietin 3′-enhancer. Corresponding TFBSs were predicted in the segments around transcription start sites (TSSs) of all human genes. Using the genome-wide set of regulatory regions, we observed several strongly preferred distances between hypoxia-responsive element (HRE) and binding sites of a particular cofactor protein. The set of preferred distances was called as a preferred pair distance template (PPDT). PPDT dramatically depended on the TF and orientation of its binding sites relative to HRE. PPDT evaluated from the genome-wide set of regulatory sequences was used to detect significant PPDT-consistent binding site pairs in regulatory regions of hypoxia-responsive genes. We believe PPDT can help to reveal the layout of eukaryotic regulatory segments.

Contact: ivan.kulakovskiy@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 13, 2011; revised on June 29, 2011; accepted on July 31, 2011

1 INTRODUCTION

The regulatory code controlling gene expression in higher eukaryotes still remains elusive. It is challenging to understand how a 1D DNA text directs formation of the protein complex that controls gene expression in a particular cell type in specific conditions. Some insight is gained by using the well-known concept of 'composite elements' consisting of binding sites for different regulatory proteins separated by specific distances (Matys *et al*., 2006). Despite more than 15 years of study, information about the scale and specificity of possible distances between binding sites remains insufficient.

New technologies like ChIP-Seq have dramatically increased the quality of identification of TFBSs, both *in vitro* and *in vivo*. Recently, it was shown (Shelest *et al.*, 2010; Yokoyama *et al.*, 2009) that in some cases there are several preferred distances between binding sites of some TFs. Here, we illustrate that distance preferences themselves are extremely well exhibited and that this phenomenon appears to be much more common, at least in the case of *Homo sapiens* TFs involved in regulating responses for hypoxic conditions.

We studied distributions of distances between TFBSs 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.

2 METHODS

Data used in the study: for our analysis, we used the UCSC human genome annotation (hg18). The genome-wide set of regulatory sequences was constructed by taking segments of 3000 bp centered at transcription start sites (TSSs) for all annotated genes. For closely located TSSs, we merged together corresponding regulatory segments if they overlapped for >50% (resulting in a total of 36 271 segments). The hypoxia-regulated set consisted of 3000 bp long regulatory sequences for known hypoxia-dependent genes (Ortiz-Barahona *et al.*, 2010) (158 sequences total, centered at TSS and merged if overlapping for >50%). The masked genome set was created from the genome-wide set by substituting exons, repBase repeats and fuzzy tandem repeats (Boeva *et al*., 2006) for poly-N sequences. An additional set of sequences demonstrating a low level of transcription (the low-transcription set) was built using CAGE-tag depleted regions (a total of 70 955 non-overlapping 3000 bp regions) based on FANTOM4 data (Kawaji *et al.*, 2009) (see details in Section 1 in Supplementary Material).

We constructed the binding motifs in the form of positional weight matrices (PWMs) (Stormo, 2000) for the HIF-1α:ARNT dimer, HNF4α, SMAD3, SMAD4, p300 and Sp1 transcription factors (TFs) forming a complex at 3′ enhancer of erythropoietin gene (EPO) (Sánchez-Elsner *et al.*, 2004). To construct a positional weight matrix (PWM) for the HIF1α:ARNT dimer binding motif [known as the hypoxia-responsive element (HRE)], we used human-curated binding site data from Ortiz-Barahona *et al.* (2010) and from the SITE table of the TRANSFAC database (Matys *et al.*, 2006), as well as ChIP-chip data published in Xia and Kung (2009). PWM was created using the ChIPMunk tool (Kulakovskiy *et al.*, 2010). For the sake of consistency, we did not use TRANSFAC motifs for other TFs as well, but created original PWMs from binding sites stored in the TRANSFAC

*To whom correspondence should be addressed.
†Present address: Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia.

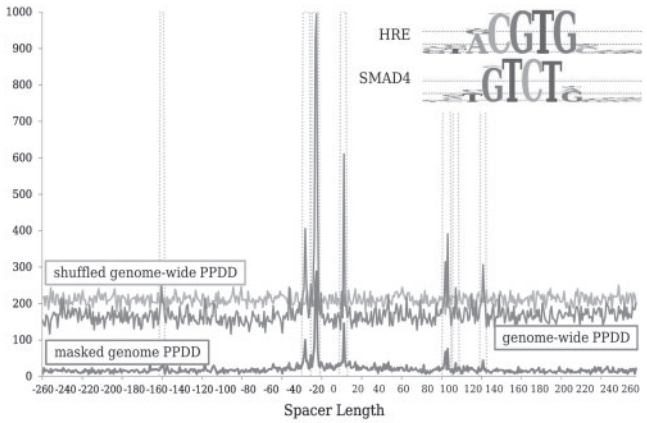

Fig. 1. The PPDD and PPDT peaks for the HRE–SMAD4 binding site pair. PPDDs from three different sequence sets are shown. The *Y*-axis displays the number of sequences (taking both strands of each segment independently) in the set having a pair of sites separated by the selected spacer (at the *X*-axis). HRE is located at zero X, see details in text). The shuffled genome-wide set was constructed by shuffling letters in the sequences from the genome-wide set so that only the base composition was preserved. Motif logos correspond to the PWMs and binding site orientation.

SITE table (see details in Section 2 in Supplementary Material). Sequence segments having PWM scores above a preselected threshold, were adopted as motif occurrences (Stormo, 2000). For a given threshold one can calculate the motif *P*-value, the probability that a random word has a PWM score no less than a given threshold. This *P*-value was calculated using the AhoPro tool (Boeva *et al.*, 2007). For each motif, we selected a PWM threshold in such a way that the corresponding motif *P*-value was equal to 10^{-3} (if not stated otherwise). The motif logos are presented in Supplementary Figure 1.

The 2400 bp long DNA segments centered at TSS were used to search for occurrences of the HRE motif, the principle DNA element controlling hypoxia response. The 600 bp windows centered at each putative HRE were used to search for occurrences of binding motifs for TFs operating as HIF-1α cofactors.

Preferred pair distance distributions: to evaluate preferred distances between the HRE and cofactor binding sites, we used a strategy similar to that described in Kulakovskiy *et al.* (2011). Essentially, for each spacer length from −300 to 300, we counted the number of sequences having a binding site of the selected TF located in a given orientation at a selected distance from the HRE. In other words for a selected orientation of the cofactor binding site relative to HRE, we counted the number of sequences having an 'HRE-cofactor binding site' pair separated by a spacer of the selected length. Negative/positive spacer values refer to upstream/downstream location of the cofactor binding site relative to the HRE. Both strands of each sequence were examined independently. Our strategy has two advantages. First, the genome sequence set provides a statistically representative set of possible distances between TFBSs. Second, when one counts the number of sequences containing at least one pair of TFBS, rather than the total number of pairs of TFBSs aggregated from all sequences, the result becomes relatively undistorted by contributions from homotypic TFBS clusters (Gotea *et al.*, 2010; Lifanov *et al.*, 2003) and repetitive DNA regions.

The corresponding preferred pair distance distribution (PPDD) for the 'HRE-SMAD4' pairs is given in Figure 1. It displays a somewhat noisy background with a set of markedly exhibited peaks at a number of selected distances. It is noteworthy that the masked genome set shows a very similar distribution of peaks in PPDD. Supplementary Figures 2–4 show genome-wide PPDDs for pairs formed by HRE and binding sites of HIF-1α cofactors for all cofactors considered in this study.

Preferred pair distance templates: Figure 1 displays preferred distances forming a comb of well-defined peaks. Additionally, the PPDD curve exhibits a 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. It is noteworthy that the principle peaks are usually much less sensitive to changes in motif model parameters (Supplementary Fig. 6 and 7) and to the sequence set used (Supplementary Fig. 9). Thus, a peak extraction procedure was needed to distinguish significant peaks from the variable background. We did this by identifying extreme points of the numerical derivative averaged over three-points (using PPDD with an 11bp-sliding window baseline correction). Peaks having both the derivative and the baseline values higher than their mean + SD were selected. Then we extracted the set of positions covered by significant PPDD peaks and called it 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 in a given orientation. Supplementary Figures 2 and 3 and Supplementary Table 1 show PPDDs and corresponding PPDTs for HRE–cofactor binding site pairs.

Assessing relevance of PPDT-consistent binding site pairs: to test whether PPDT is related to functional TFBS arrangements, we used the sequence set containing regulatory regions of hypoxia-dependent genes.

Suppose we have selected a cofactor TFBS and its orientation relative to HRE. For each sequence, we independently counted the total number, n, of 'HRE-cofactor TF binding site' pairs within 600 bp windows centered at HRE, and the number of such pairs having a spacer corresponding to one of the PPDT distances, k. We selected subsets from the genome-wide set and the hypoxia-regulated set containing sequences having n above zero (i.e. having at least one HRE–cofactor TFBS pair at any distance within 600 bp HRE-centered windows). Let N_g be the size of the genome-wide '$n > 0$'-subset of sequences and N_h be the size of the hypoxia-regulated '$n > 0$'-subset of sequences.

Using PPDT constructed from the genome-wide dataset, we evaluated the enrichment of pairs of HRE and its cofactor TFBS located at any of the PPDT-consistent distances in the hypoxia-regulated set. To this end, for a given threshold t for each set we counted the number of sequences having k/n ratio no less than the threshold (denoted as C_h and C_g for the hypoxia-regulated set and the genome-wide set respectively). Let us assume that in the genome-wide set, sequences with different k and n are found at random. The probability $P_{k/n}$ to pick a sequence at random having $k/n \geq t$ is estimated as $P_{k/n} = C_g/N_g$. Adopting the null hypothesis that the set of hypoxia-regulated sequences have a similar enrichment with PPDT-consistent distances as the genome-wide dataset, one can calculate the probability to draw N_h sequences having at least C_h sequences satisfying $k/n \geq t$ condition by chance. This value can be calculated with the help of the binomial distribution. The binomial distribution $B(l,s,p)$ is the discrete probability of having exactly s successes ('yes' answers) in a series l independent binary ('yes'/ 'no') trials, each of which can yield a success with the given probability P. With the help of this distribution, the k/n ratio P-value of having at least C_h sequences with $k/n \geq t$ from N_h writes as:

$$P\text{-value} = 1 - \Sigma_i B(N_h, i, P_{k/n}), \quad i = 0..C_h - 1$$

3 RESULTS

Hypoxia-regulated set is enriched with PPDT-consistent binding site pairs: Figure 2 shows k/n ratio P-values for the pairs of HRE–HRE and HRE–Sp1 binding sites estimated using genome-wide PPDTs. Supplementary Figure 5 display similar graphs for HRE and other cofactor binding sites in different orientations. Better (lower) P-values are exhibited for small k/n threshold levels. Specifically this means that regulatory regions of hypoxia-dependent genes often exhibit at least one PPDT-consistent pair of HRE–cofactor TFBS. The k/n ratio test demonstrated that positioning of the following TFBS pairs can be used to distinguish the hypoxia-regulated set from

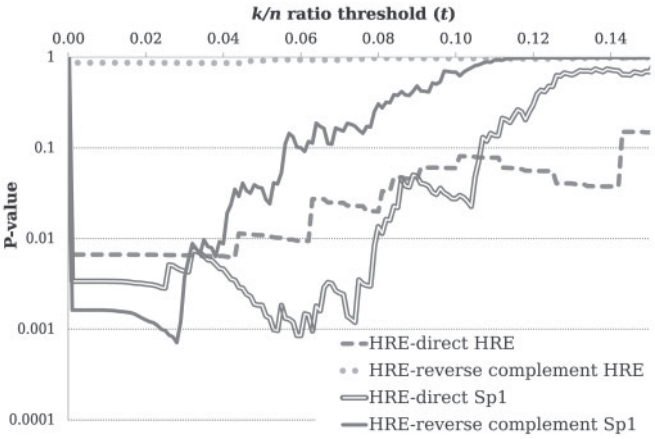

Fig. 2. k/n ratio test used to distinguish the hypoxia-regulated regions from the genome-wide set (see details in Section 3). The X-axis corresponds to the k/n ratio threshold. The Y-axis shows the corresponding P-value.

the genome-wide set: HRE-p300 (direct orientation), HRE–HRE (direct orientation), HRE–Sp1 (both orientations), HRE–SMAD3 (reverse orientation). Preferred locations of other TFBS pairs were not exhibited in the hypoxia-regulated gene set, which may indicate that the binding of these TFs in such an orientation is not characteristic for regulation of hypoxia response or that in this case TF binding pattern is so complex that it cannot be easily recovered with the help of HRE-centered PPDTs.

Intersite distance preferences are stable and prevalent: PPDDs are obviously sensitive to motif modifications or changes of PWM score threshold. Still, general patterns of distance preferences are quite stable and seem not to be a specific feature of the particular model for the HRE or another TFBS. Supplementary Figure 6 shows the HRE–HRE and HRE–Sp1 PPDDs for different motif thresholds. Supplementary Figure 7 shows the comparison for PPDDs obtained for TRANSFAC PWMs and our PWMs constructed with ChIPMunk. Supplementary Figure 8 shows genome-wide PPDDs centered at Sp1 binding sites instead of HRE used as the anchoring element elsewhere in this article.

Mystery of low-transcription regions: in fact, HIF-1α can regulate hundreds of genes (Mole *et al.*, 2009). Different studies report dramatically different sets of HIF1a targets (see e.g. Venn diagram in Fig. 2 in Ortiz-Barahona *et al.*, 2010). Since we could not reliably guess which genes were not regulated by HIF-1α and thus could be taken as the negative control set, we tried to compare our findings with DNA sequence segments performing low transcription activity or no transcription activity at all. To this end, we selected a set of DNA segments located far from any DNA region, for which transcription activity was demonstrated in TSS-calling experiments (see the description of the set in the Section 2 and Section 1 in Supplementary Material). Surprisingly, in the low-transcription regions, PPDD peaks were exhibited even more explicitly than in the genome-wide set (Supplementary Fig. 9 and 10). Meanwhile, the relative number of binding site pairs in this case was much lower than in the genome-wide set (Supplementary Fig. 10). We failed to provide an exhaustive explanation of a clearer PPDD for the low transcribed dataset. Yet, it is noteworthy that preparation of the genome-wide dataset included the merging of regions centered at closely located alternative TSSs. It is possible that the genome-wide PPDD includes distances measured between TFBSs belonging to different homotypic clusters which regulate transcription from different TSSs and this interference smears the PPDD.

Moreover, we counted the number of CAGE-tags located in three subsets of low-transcription regions set: the first having no HRE–HRE or HRE–Sp1 binding site pairs, the second having pairs at any distance not consistent with PPDT within 600 bp HRE-centered windows and the third having pairs of TFBS at PPDT-consistent distances according to PPDT evaluated from the genome-wide set. Supplementary Table 2 displays the median of the number of CAGE tags per sequence. The sequences with binding site pairs located at PPDT-consistent distances tend to be relatively enriched with CAGE-tags. Thus, we suggest that the PPDT-consistent binding site pairs found in low-transcription regions are somehow linked to nearby transcriptionally active regions.

4 DISCUSSION

In our previous work (Kulakovskiy *et al.*, 2011), we reported the observation that a large genome-wide set of long segments exhibits some peaks in the distribution of distances between TFBSs of interacting TFs. In this study, we introduce the concept of PPDT, a set of selected preferred distances between TFBSs, which, as we demonstrate, are expected to be found in the regulatory regions of genes, regulated by the corresponding TFs.

Preferred distances between TFBS for different TF pairs form substantially different sets, but in all cases a general pattern of a peak comb over a background of more or less random distances is observed. It is tempting to believe that binding sites found at PPDT inconsistent distances are likely to form complexes with TFs other than HIF-1α or simply are false positives of PWM scanning. Yet, PPDT found in the low-transcribed regions allows the conclusion that there might be some yet unknown factor controlling possible TFBS positioning in DNA.

Obviously, PPDD/Ts provide additional information on the specificity of TF binding, especially when the total number of putative TFBSs in the regulatory regions is small. However, the difficulties of using PPDD/Ts for identification of functional binding sites should not be underestimated. Binding site arrangements in regulatory DNA segments are complex, with sites often overlapping each other. Most of the sequences in the hypoxia-regulated dataset contain homotypic clusters of HRE. So when one calculates all pairwise distances between binding sites of two transcription factors, the number of observed intersite distances becomes large. Paradoxically, the worst display of preferred distances is in the hypoxia-regulated dataset (Supplementary Fig. 11). This happens because this dataset is rather small, containing only 158 sequences, 25 of which do not contain any HRE elements. Many sequences contain only one or two pairs of binding sites of the specified type. In the hypoxia-regulated dataset, each distance is rarely found more often than five times. Thus, most numbers of preferred distance occurrences are statistically insignificant, and the PPDD/T pattern is unclear. This makes it difficult to conclude whether the hypoxia-regulated set really has some major characteristic differences in PPDDs when compared with the genome-wide set.

Another paradox is that PPDTs for HRE–Sp1 and HRE–HRE homotypic site pairs estimated on the genome-wide gene set can be efficiently used to distinguish the hypoxia-regulated set (Fig. 2). This

is also true for some other HRE–TFBS pairs (Supplementary Fig. 5). This property looks promising for identifying hypoxia-dependent genes, because the procedure excludes overfitting. Indeed, the positive gene set (the set of hypoxia-regulated genes) in this case works as a test dataset rather than as a training dataset.

A large number of HRE–HRE homotypic pairs tend to have very small spacers (see corresponding HRE–HRE PPDD/Ts in Supplementary Figs 2–4). To explain this result, it is enough to suggest that HIF-1α:ARNT dimers form higher order complexes bound to nearby HREs corresponding to the HRE–HRE PPDD/T). This suggestion is supported by the sequence motif found *de novo* using the data from (Ortiz-Barahona *et al.*, 2010), see Supplementary Figure 12.

ACKNOWLEDGEMENTS

We thank Biobase GmbH and personally Alexander Kel for granting us the access to the TRANSFAC release 2010.1. We thank Dmitrijs Lvovs, Valentina Boeva for reading and commenting on the article. We thank Dmitry Oshchepkov, Victor Levitskii, Mikhail Roytberg and Mikhail Gelfand for fruitful discussions during the preparation of the revised version of the article. We especially thank Adam Arents for his many important suggestions on the article.

Funding: Presidium of the Russian Academy of Sciences program in Cellular and Molecular Biology; Russian Ministry of Science and Education State Contract (07.514.11.4005); Russian Ministry of Science and Education State Contract (07.514.11.4006); Russian Foundation for Basic Research grant (10-04-92663).

Conflict of Interest: none declared.

REFERENCES

Boeva,V. *et al.* (2006) Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics*, **22**, 676–684.

Boeva,V. *et al.* (2007) Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules. *Algorithms Mol. Biol.*, **2**, 13.

Gotea,V. *et al.* (2010) Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.*, **20**, 565–577.

Kawaji,H. *et al.* (2009) The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biol.*, **10**, R40.

Kulakovskiy,I.V. *et al.* (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.

Kulakovskiy,I.V. *et al.* (2011) Preferred distances between transcription factor binding sites. *Biophysics*, **56**, 114–116.

Lifanov,A.P. *et al.* (2003) Homotypic regulatory clusters in Drosophila. *Genome Res.*, **13**, 579–588.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Mole,D.R. *et al.* (2009) Genome-wide association of hypoxia-inducible factor (HIF)-1alpha and HIF-2alpha DNA binding with expression profiling of hypoxia-inducible transcripts. *J. Biol. Chem.*, **284**, 16767–16775.

Ortiz-Barahona,A. *et al.* (2010) 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. *Nucleic Acids Res.*, **38**, 2332–2345.

Sánchez-Elsner,T. *et al.* (2004) A cross-talk between hypoxia and TGF-beta orchestrates erythropoietin gene regulation through SP1 and Smads. *J. Mol. Biol.*, **336**, 9–24.

Shelest,V. *et al.* (2010) DistanceScan: a tool for promoter modeling. *Bioinformatics*, **26**, 1460–1462.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Xia,X. and Kung,A.L. (2009) Preferential binding of HIF-1 to transcriptionally active loci determines cell-type specific response to hypoxia. *Genome Biol.*, **10**, R113.

Yokoyama,K.D. *et al.* (2009) Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. *Nucleic Acids Res.*, **37**, e92.