DISTANCE PREFERENCES IN DISTRIBUTION OF BINDING MOTIFS AND HIERARCHICAL LEVELS IN ORGANIZATION OF TRANSCRIPTION REGULATORY INFORMATION

Makeev V.J.1*, Lifanov A.P.2, Nazina A.G.3, Papatsenko D.A.3

1 Scientific Center “Genetika”, Moscow, Russia; 2 Engelhardt Institute of Molecular Biology, Moscow, Russia; 3 Department of Biology, New York University, New York, USA

* Corresponding author: e-mail: makeev@imb.ac.ru

Keywords: transcription, regulation, factor, binding site, enhancer, periodicity, composite element

Resume

Motivation: Initiation of eukaryotic transcription is a complex process that involves huge protein complexes, which interact with regulatory DNA at many DNA-protein binding sites. A huge amount of regulatory information is contained in the mutual positioning of binding sites. Investigation in detail of interrelation of site location can help to understand the spatial conformation of DNA helix within these regions as well as facilitate developing of new computerized methods of locating of regulatory modules.

Results: We explored distance preferences in distribution of binding motifs for five transcription factors (Bicoid, Kruppel, Hunchback, Knirps and Caudal) in a large set of Drosophila Cis-Regulatory Modules (CRM). We established that the vast majority of high-affinity binding sites are positioned in the CRMs at distances close to 10, 20, 30 etc. base pairs, i.e. approximately on the same side of the DNA helix. We also assess site overlapping and positioning of site pairs at specific distances. We discuss hierarchical levels in organization of transcription regulatory information and its role in detection of transcription regulatory regions in genome.

Availability: http://homepages.nyu.edu/~dap5/

Introduction

Initiation of tissue-specific or spatio-specific transcription in multicellular organisms requires binding of multiple transcription factor molecules to transcription regulatory regions, such as promoters and enhancers. Multiplicity of binding motifs and binding sites for the same motif in the regulatory regions, are often characterized as regulatory clusters (Wagner, 1999). However, the number of sites and their affinity is not sufficient to describe regulatory information encoded by the binding motifs: specific arrangement of these binding motifs within the regulatory regions appears to be necessary for achieving proper biological function. The arrangement of binding motifs carries signature of 3D protein complexes, involved in the initiation of specific transcription. The precisely arranged pair of distinct binding motifs, involved into formation of specific DNA-protein-protein complexes are called composite elements (CE). The database of composite elements, TRANSCompel (Kel-Margoulis et al., 2002) combines 256 (Release 6.0) elements from different organisms.

In the current work we explored presence of preferential site distances in cis-regulatory modules (CRM) of Drosophila developmental genes. The term CRM stands for extended transcription regulatory units (~1Kb range), often located far from transcription start site and responsible for spatio-temporal expression of their cognate developmental genes (Berman et al., 2002). Database of known functional Drosophila CRMs is available (see NYU web site: http://homepages.nyu.edu/~dap5/PCL/appendix2.htm) along with the list of matrices for a number of transcription factors and known transcriptional interactions, thus providing an excellent original dataset for computational study.
Results

Measuring distances between binding sites requires careful mapping of binding motifs in a relevant set of sequence data. We selected Position Weight Matrices (PWMs), generated from good alignments and limited our choice by five best known binding motifs for transcriptional regulators Bicoid, Caudal, Hunchback, Kruppel and Knirps, having relatively large number of instances in our CRM database. At the next step, we selected from our database of experimental CRMs only sequences containing significant homotypic clusters for these five binding motifs (Lifanov et al., 2003).

The total size of analyzed CRMs after the described prescreening procedure comprised more than 68 Kb of sequence data and contained 58 separate homotypic clusters in 33 non-overlapping contigs (see: http://homepages.nyu.edu/~dap5/PCL/appendix2.htm) along with the contigs themselves (Papatsenko et al., 2002; Lifanov et al., 2003).

We measured the distances between the centers of binding site cores, which were established on the basis of informational content of motif. We considered motifs at the same or at different DNA strands, corresponding to the tandem of motif arrangement respectively.

First we compared the distribution of binding motifs in a hole set of regulated CRMs using the geometric distribution evaluated from the same set as the reference hypothesis. The comparison was performed for different levels of site quality, i.e. different PWM cutoff value. To insure site independence we excluded close distances (less than 5 bases), which often correspond to the overlapping sites. The described distance distribution test clearly shows that the binding sites in Drosophila CRMs are distributed in a non-random fashion and large fraction of them has spacing not exceeding 50–60bp (see (Makeev et al., 2003)). This distance range is comparable with the size of composite elements (CE) reported earlier.

To test whether the binding motifs distributed in the CRMs periodically we calculated all distances between the binding sites. In a random Bernoulli sequence the all-distance occurrence has uniform distribution, independent from the distance itself. Distance distribution between particular different binding sites may be also interesting, especially if they involved in synergistic or antagonistic interactions. Fourier spectra were used to assess periodical positioning.

First, we built distributions and the corresponding Fourier spectra for motifs of the same types. The most striking result, entirely confirming the hypothesis of ‘helical phasing’ was obtained for Bicoid. The vast majority of the high-affinity Bicoid sites are positioned at distances close to 10, 20, 30 etc. base pairs (see the Figure). The periodicity in the distribution of Bicoid (Bcd) sites drops very fast with decreasing site affinity (PWM score), supporting biological role of this specific spacing, rather than short-range correlations in DNA sequence itself.

Similarly, but not identical periodic signal was revealed in the distribution of Hunchback (Hb) (see the Fig.). In this case however, the period was equal not to 10 but to 11 base pairs. We believe that this difference in periodicity is the result of slightly different DNA conformation of the two binding motifs (compare CCTAATCCC, – consensus for Bicoid and TTTTTTTT, – consensus for Hunchback). Surprisingly, but distribution of another binding motif, Kruppel (Kr), showed no presence of periodic signals, corresponding to the ‘helical phasing’. This can be a reflection of different biological role of those factors. Whereas Bicoid and Hunchback are transcriptional activators mostly, involved in cooperative polymerization and protein-protein interactions with other transcription factors, Kruppel is a repressor protein. Caudal motif alone also displayed ‘helical phasing’ and it is also an activator protein. Knirps motif had low number of occurrences in our dataset and was not considered alone.

The analysis of distance preferences between specific combinations of binding motifs demonstrated that Bcd-Hb hunchback motif pair demonstrated lower periodic signal than the sites for both proteins taken independently. In contrast the motif pair, Bcd-Kr shows a prominent periodic signal the phase of which contrasts the helical pitch (17 bp) (Makeev et al., 2003). Thus, non-
overlapping Bcd and Kr sites have a tendency to be distributed on opposite surfaces of DNA. This result may suggest that the non-overlapping Bcd and Kr sites may belong to distinct composite elements, performing different functions.

Finally, we extracted periodic signals from combination of all five binding motifs and detected presence of the same ‘helical phase’ signal, though having smaller amplitude and, likely, contributed by distances Bcd-Bcd and Hb-Hb. It is important, however, that the distribution of non-overlapping binding motifs in regulatory regions cannot be described by the simple ‘helical phasing’ formula. The more detailed description of these finding can be found in (Makeev et al., 2003).

Discussion

Binding motifs are distributed in Drosophila CRMs in a non-random fashion, where a large fraction of sites exhibit distances not exceeding 50–70 base pairs. We believe that the composite elements represented by small groups (2–5) of specifically arranged (spaced) binding motifs comprise the intermediate organizational level. Size of promoter or cis-regulatory module (05–1Kb) allows fitting several such composite elements, perhaps acting independently in their response to verity of transcriptional signals. For example, repression of the same promoter (CRM) by two transcription factors might be achieved through independent action of two corresponding composite elements. In this respect, our finding of the opposite to the helical phase (x17 bp) in distribution of Bcd-Kr pair represents special interest.
Acknowledgements

This study was partially supported by RFBR grant 04-07-90270-а, and by Program in Molecular and Cellular Biology of Russian Academy of Sciences, coordinator V.G. Tumanyan, V. Makeev and A. Lifanov are also supported by CRDF RBO-1268-MO-02 and Howard Hughes Institute Grant 55000309.

References


