Conserved Regions of DNA Forming Nucleosomes in Transcriptional Regulatory Modules Are Close in Space

A. P. Lifanov[a], V. J. Makeev[a,b], and N. G. Esipova[a]

[a] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, ul. Vavilova 32, Moscow, 119991 Russia*

[b] *Vavilov Institute of General Genetics, Russian Academy of Sciences, ul. Gubkina 3, Moscow, 119991 Russia*

e-mail: johnnie_me@list.ru

Distribution of genomic elements, i.e. segments of their increased or decreased density, may be related to DNA domains performing specific functions (e.g. CpG islands [1], clusters of transcription factor binding sites (TFBS) [2], or segments enriched or depleted in nucleosome binding sites [3]). Similarly, specific sequence features with regular and biased distribution can indicate yet unknown functional domains. If such elements are found in DNA segments conservative between several organisms at different evolutionary distances this increases the likelihood of their functional significance.

We aligned loci of genes involved in the early development of several *Drosophila* species: *D.melanogaster, D.pseudoobscura*, *D.erecta*, *D.littoralis*, and *D.willistoni* with the OWEN software tool [5]. Alignment parameters, the window size and degree of similarity, were selected in a way that all functional elements of the well-studied *even-skipped* gene locus for all studied *Drosophila* species were unambiguously aligned with the sequence of the reference species *D.melanogaster*. It turned out that conserved domains in the multiple alignment were well represented by the alignment of *D.melanogaster* with *D.pseudoobscura,* with all known TFBS overlapping segments with the conservation degree of no less than 90% (the coincidence of nine nucleotides (nt) in a 10-nt window). In this pairwise alignment the positions of conserved domains were determined for genes of the *pair-rule* group [6]: *even-skipped, hairy, odd-skipped, paired, runt, fushi tarazu, odd-paired, sloppy paired*, and *ten*. The positions of TFBS and known regulatory and coding segments in these gene loci were determined previously [2].

The identified conserved domains had the average length of 30 to 70 nt, the value, which is between the TFBS length (usually about 7-10 nt) and the length of the nucleosome repeat units (165-210 nt). Taken together all conserved domains occupied no more than a half the total enhancer length but overlapped the majority of TFBS.

Positions of the identified conserved regions were compared with nucleosome positions determined by Mavrich et al. [3].

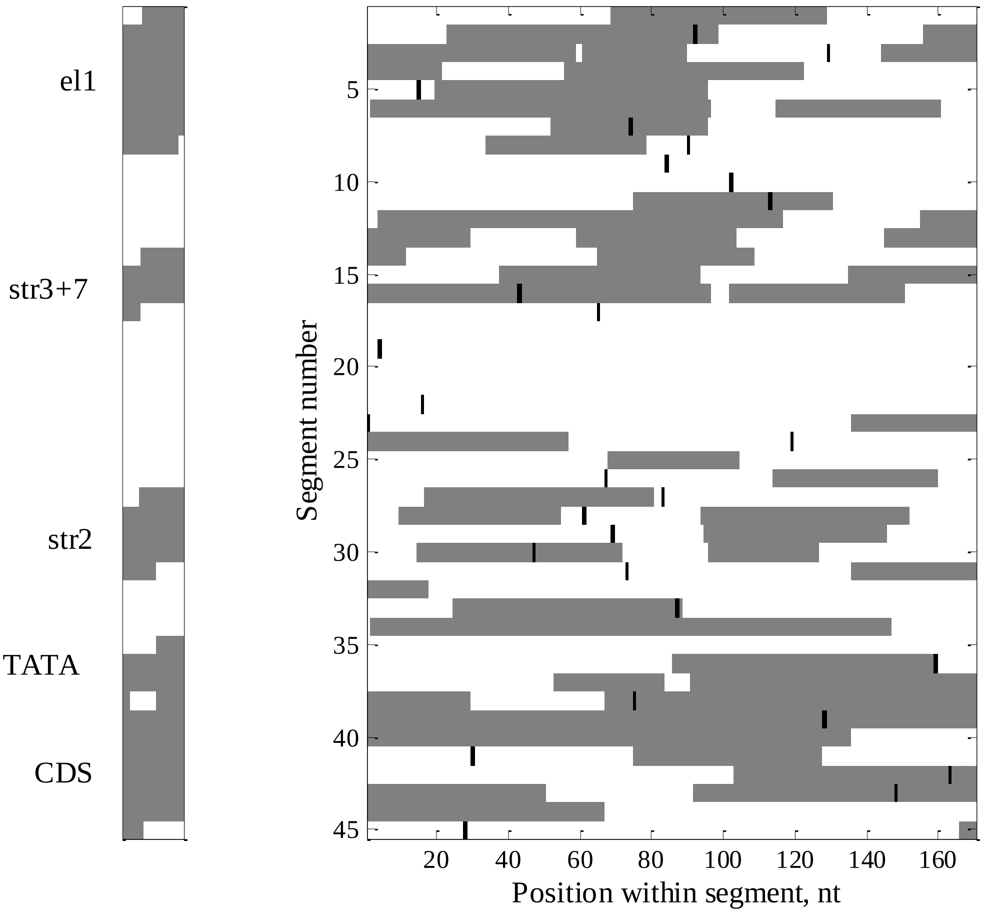

Fig. 1. Relative positions of conserved domains and nucleosomes in the proximal region of *D.melanogaster even-skipped* gene locus.

The region is divided into consecutive 170-nt segments (from top to bottom), with gray bands showing the conserved domains and vertical black lines indicating the centers of nucleosome binding regions determined experimentally. The bar on the left shows the arrangement of functional elements in the locus: enhancers (el1, str3+7, str2), the proximal promoter (TATA), and the coding segment (CDS).

A locus map is constructed for each of the aforementioned genes. This map consists of nonzero values at positions within the conservative domains and zeros at all other positions. The centers of nucleosome binding segments are also located on the map.

To evaluate similarity between the distribution patterns of conserved domains and nucleosome positions, the map is divided into consecutive 170-nt segments (close to the minimum length of nucleosome repeat unit [7]), which are placed one above another. Such a segmental map for the proximal region of *D.melanogaster even-skipped* gene locus is shown in the Fig. 1. Comparing the "segmentation patterns" of this and other gene loci, some common characteristic features could be distinguished.

We observe that distributions of conserved domains and nucleosomes are interrelated and have a common quasi-periodic pattern with a period (interval between elements) agreeing with the minimum length of the nucleosome repeat unit: the boundaries of domains from adjacent segments are approximately aligned in a vertical direction (see the Figure). This allows one to conclude that DNA in the regulatory elements of the locus is likely to be packed into nucleosomes. A similar distribution of genomic elements agreeing with the nucleosome length has also been revealed in other regions of the genome, namely, exons and introns in the coding regions of collagen genes [9].

Furthermore, a 170-nt segment usually incorporated two conserved domains and two intermediate nonconserved inserts, with the characteristic distance between the neighboring domains of about 84 nt, i.e., equivalent to the pitch of nucleosomal DNA superhelix [6, 7]. Such a similarity in the length can indicate that the identified conserved domains are located in the neighboring coils of the nucleosome DNA superhelix and thus could be termed as "co-phased blocks".

The characteristic length of a triplet consisting of two consecutive co-phased blocks separated by a nonconserved insert is equivalent to about 1.5 coils of nucleosomal DNA superhelix, which is only slightly smaller than the total length of such DNA (1.65-1.8 coils [7,8]). Thus, the DNA of such a triplet almost completely overlaps with the histone core of the nucleosome, with its central nonconserved segment lying at the nucleosome symmetry axis. The co-phased blocks are located on the opposite side of the histone core that binds two DNA complementary strands and, hence, are spatially converged.

According to one of generally accepted three-dimensional models of the nucleosome, this structure also includes histone H1, which stabilizes flanking nucleosome DNA segments [10]. In this context, the deficiency of conservation in the central part of the triplet and the simultaneously under-representation of TFBS in this segment could be attributed to its protection by histone H1.

The complete version of this study is published in [11].

This study was supported by the Russian Foundation for Basic Research, project nos. 14-04-90034-Bel_a and 12-0401776-a.

1. F. Antequera (2003) *Cell. Mol. Life Sci.* **60**, 1647.

2. A. P. Lifanov, V. J. Makeev, A. G. Nazina, and D. A. Papatsenko (2003) *Genome Res.* **13**, 579.

3. T. N. Mavrich, C. Jiang, I. P. Ioshikhes, et al. (2008) *Nature* **453**, 358.

4. J. Touchman (2010) *Nature Education Knowledge* **3** (10), 13.

5. A. Y. Ogurtsov, M. A. Roytberg, S. A. Shabalina, and A. S. Kondrashov (2002) *Bioinformatics* **18**, 1703.

6. L. Wolpert and C. Tickle (2002) *Principles of Development* (Oxford Univ. Press, Oxford).

7. G. Felsenfeld and M. Groudine (2003) *Nature* **421** (6921), 448.

8. L. Marino-Ramirez, M. G. Kann, B. A. Shoemaker, and D. Landsman (2005) *Expert Rev. Proteomics* **2**, 719.

9. A. P. Lifanov, P. K. Vlasov, V. J. Makeev, and N. G. Esipova (2008) *Biophysics (Moscow)* **53** (3), 245.

10. N. Happel and D. Doenecke (2009) *Gene* **431** (1-2), 1.

11. A. P. Lifanov, V. J. Makeev, and N. G. Esipova (2015) *Biophysics (Moscow)* **60** (1), 5 (in english).