DOI 10.18699/SBB-2022-084

Textual and structural analysis of RNA polymerase II core promoter of evolutionary diverse organisms

Melikhova A.V.[1], Anashkina A.A.[2]*, Il'icheva I.A.[2]

[1] *I.M. Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow, Russia*

[2] *Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia*

* *anastasia.a.anashkina@gmail.com*

Key words: RNA polymerase II promotor, sequence specificity DNA structure

Motivation and Aim: The general structure and action of all eukaryotic and archaeal RNA polymerases machinery have an astonishing similarity despite the diversity of core promoter sequences in different species. The goal of our work is to find common characteristics of the DNA region that define it as a promoter for the RNA polymerase II (Pol II). We have analyzed the profiles of textual and a large number of physical and structural characteristics, namely mono- and dinucleotides distributions, frequencies of occurrence of tetranucleotides TATA and AAAA, stacking energy, roll, stiffness of the duplex structure to roll alteration, slide, stiffness of the duplex structure to slide alteration, mobility to bend towards major groove, ultrasound cleavage intensities, and DNase I cleavage intensities. The indexes of these characteristics in each position of core promoter sequence were averaged over representative sets of promoters for each of the evolutionary diverse species from animals, plants and unicellular fungi and a unicellular protozoan (*H. sapiens*, *M. mulatta*, *M. musculus*, *R. norvegicus*, *G. gallus*, *C. familiaris*, *D. melanogaster*, *A. mellifera*, *D. rerio*, *C. elegans*, *A. thaliana*, *Z. mays*, *S. cerevisiae*, *S. pombe*, *P. falciparum*).

Methods and Algorithms: The sets of nucleotide sequences from the core promoter regions of different organisms, aligned at the TSS, were retrieved from the EPD New section of the Eukaryotic Promoter Database (EPD) (http://epd.vital-it.ch). For our analysis we used numerical parameterization at the level of ten different base-pair steps, collected at the database DiProDB (http://diprodb.fli-leibniz.de), as well as the relative intensities of cleavage of the central phosphodiester bond in 16 dinucleotides and 256 tetranucleotides [1], and the new data on the propensities of hexanucleotides to the DNase I cleavage [2]. Methods are described in detail in the article [3].We have written programs in Python 3.10 to analyze sequences of RNA polymerase II promotors. Logos were made at http://weblogo.threeplusone.com.

Results: The core promoters (80 b.p. at positions from –50 to +30) of all analyzed organisms, representing the kingdoms of animals, plants, fungi, and protozoa have a common structural organization. Regardless of the AT content in genomic DNA, they have two singular regions: a hexanucleotide with coordinates –2 -- +4 (INR), surrounding the transcription start site (TSS), and an octanucleotide, separated from TSS at a distance of about 28 b.p., located upstream. In the TSS position (–1, +1), the occurrence of PyPu/PyPu is exceptionally high, with a noticeable predominance of the d(CA/TG) dinucleotide. The minor groove of TATA-box is broadened and conformational motion is decreased in that region. Special characteristics of conformational behavior are revealed in mammalian, best of all in *H. sapience*, at the

region, which connects the end of TATA-box and the transcription start site (TSS). The intensities of conformational motions in the complementary strands are periodically changed in that region in opposite phases.

Conclusion: Comparison of profiles of textual, structural, structural-dynamic and physico-chemical characteristics of the nucleotide sequences of promoters of fifteen organisms representing animals, plants, unicellular fungi, as well as unicellular protozoan (*P. falciparum*), allows us to conclude that there exist a general scheme of the structural organization of the core promoter regions for all organisms, which does not depend on the level of their evolutionary development and the AT content in their genomes. Se-quences in singular regions (TATA box and TSS) are not random. Their selection was based on the conformational characteristics of dinucleotides and their nearest neighbors. The number of sequence variants, that represent the TBP binding region (TATA-box) depends on the percentage of AT content in the genome. Conformational characteristics of the octanucleotide in this position can affect the kinetics of DNA-protein complex formation in the promoter of a particular gene. Obtained results may be useful in genetic engineering for artificial modulation of the promoter strength.

Acknowledgements: This research was funded by Russian Science Foundation, grant No. 21-14-00346.

References

1. Grokhovsky S.L. et al. Sequence-specific ultrasonic cleavage of DNA. *Biophys J.* 2011;100(1):117-25.
2. Lazarovici A. et al. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc Natl Acad Sci USA*. 2013;110:6376-6381.
3. Il'icheva I.A. et al. Structural features of DNA that determine RNA polymerase II core promoter. *BMC Genom*. 2016;17(1):973.