doi 10.18699/bgrs2024-1.4-30

Bioinformatics of transcription initiation in eukaryotes: TATA binding protein (TBP) and TBP binding sites in gene promoters

Ponomarenko M.[1]*, Chadaeva I.[1], Zolotareva K.[1], Podkolodnyy N.[1, 2], Sharypova E.[1], Ponomarenko P.[1], Rasskazov D.[1], Bogomolov A.[1, 3], Savinkova L.[1], Kolchanov N.[1, 3]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia*

[3] *Novosibirsk State University, Novosibirsk, Russia*

* *pon@bionet.nsc.ru*

Key words: eukaryote, gene, promoter, TATA box; TATA-binding protein; SNP; *in silico* estimate; *in vivo* test; handmade annotation; target-assisted breeding

Motivation and Aim: One of the largest worldwide scientific projects now is “1000 Genomes”, which has already sequenced the individual genomes of many hundreds of thousands peoples. Their consensus sequence of the most frequent nucleotides to each position of genomic DNA is the reference human genome, difference of which from individual human genomes is single nucleotide polymorphism (SNP). The dbSNP database contains over one hundred of million experimentally proven SNPs [1]. The clinical search for biomedical SNP markers is the identification of SNP, the frequencies of which alleles in patients are statistically significantly different in comparison with healthy people that is obligatory criterion without alternative because of there is a chance to take such SNP-markers into account during the treatment of people. It is slowly, expensively and labor intensively applying to each of the all nine billion potential SNPs in humans. Preliminary computer-based analysis of unannotated SNPs can make this search for clinical SNP-markers much faster, cheaper, and more targeted due to rejecting neutral SNPs, which make up the absolute majority in humans according to both neutral evolution theory and Haldane's dilemma. The best studied is the SNP of protein-coding regions of genes, which disrupts the structure and function of proteins in the same way in all cells, tissues, and organs. That is why, these SNPs are easier to find, but their pathogenic effects cannot be avoid without genome editing, which is intensively studied in the case of human disease models using laboratory animals, as well as for plants and microorganisms, but is strictly prohibited in the case of humans. At the same time, the least studied is the regulatory gene region SNPs, which keep proteins in norm and change protein concentrations, which vary in cells, tissues, and organs for many reasons. That is why, they are more difficult to find, but its pathological effects can be corrected with medication and lifestyle changes. That is why, regulatory SNP-markers seem to be most promising for the current practice within genome-based personalized medicine. Among all regulatory SNPs, the best studied are those for the TATA-binding protein (TBP) binding sites, the canonical form of which is the TATA box. Most often TBP-sites can be found nearly 30 bp just in front of any transcription start sites of promoters as the obligatory regulatory element responsible for the primary initiation of transcription starting from just these sites, which is no longer required by subsequent re-initiation of transcription, but without which the phenomenon of transcription as such is impossible [2–4].

Methods and Algorithms: Therefore, in this review, we briefly recall how a three-step model of TBP binding to the promoters of human genes was proposed [5], which was then confirmed experimentally by independent authors [6]. Besides, we will describe how we have next created the Web-service SNP_TATA_Comparator [5] was, and how to use it in practice for genome-wide analysis of SNP within the human gene promoters. *Results*: As well, we present the genome-wide knowledge base Human_SNP_DATAdb [8], which among the all 11,384,651 SNPs within 90 bp proximal promoters of the all 17,079 protein-coding genes pinpoints to 602,547 SNPs altering significantly the TBP-promoter affinity and, thereby, the expression of the human genes under these promoters. *Conclusion*: Finally, Human_SNP_DATAdb has 57,612 annotations on how some of the above-said 602,547 candidate SNP-markers of the human gene expression changes can manifest in susceptibility to human diseases according to the PubMed database [9].
Funding: The study is supported by the Government Budget Project FWNR-2022-0020.

References

1. Day I.N. dbSNP in the detail and copy number complexities. *Hum Mutat*. 2010;31(1):2-4. doi 10.1002/humu.21149
2. Martianov I., Viville S., Davidson I. RNA polymerase II transcription in murine cells lacking the TATA binding protein. *Science*. 2002;298(5595):1036-1039. doi 10.1126/science.1076327
3. Auble D.T. The dynamic personality of TATA-binding protein. *Trends Biochem Sci*. 2009;34(2):49-52. doi 10.1016/j.tibs.2008.10.008
4. Rhee H.S., Pugh B.F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*. 2012;483(7389):295-301. doi 10.1038/nature10799
5. Ponomarenko P.M., Savinkova L.K., Drachkova I.A., Lysova M.V., Arshinova T.V., Ponomarenko M.P., Kolchanov N.A. A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. *Dokl Biochem Biophys*. 2008;419:88-92. doi 10.1134/s1607672908020117
6. Delgadillo R.F., Whittington J.E., Parkhurst L.K., Parkhurst L.J. The TATA-binding protein core domain in solution variably bends TATA sequences via a three-step binding mechanism. *Biochemistry*. 2009;48(8):1801-1809. doi 10.1021/bi8018724
7. Ponomarenko M., Rasskazov D., Arkova O., Ponomarenko P., Suslov V., Savinkova L., Kolchanov N. How to use SNP_TATA_Comparator to find a significant change in gene expression caused by the regulatory SNP of this gene's promoter via a change in affinity of the TATA-binding protein for this promoter. *Biomed Res Int*. 2015;2015:359835. doi 10.1155/2015/359835
8. Filonov S.V., Podkolodnyy N.L., Podkolodnaya O.A., Tverdokhleb N.N., Ponomarenko P.M., Rasskazov D.A., Bogomolov A.G., Ponomarenko M.P. Human_SNP_TATAdb: a database of SNPs that statistically significantly change the affinity of the TATA-binding protein to human gene promoters: genome-wide analysis and use cases. *Vavilov J Genet Breed*. 2023;27(7):728-736. doi 10.18699/VJGB-23-85
9. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford)*. 2011;2011:baq036. doi 10.1093/database/baq036