DOI 10.18699/SBB-2019-08

# Statistical analysis, clusterization and visualization of genome distribution of transcription factor binding sites

Dergilev A.I.[1]*, Tsukanov A.V.[1, 2], Luzin A.N.[1], Babenko R.O.[1], Orlov Y.L.[1, 2]
[1] *Novosibirsk State University, Novosibirsk, Russia*
[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
** e-mail: arturd1993@yandex.ru*

**Key words:** genome, transcription factor, medical informatics, ChIP-seq

*Motivation and Aim*: The analysis of gene transcription regulation based on the data of modern technologies of high-performance sequencing is an actual task of bioinformatics [1]. It requires the development of new computer tools including supercomputer applications. We consider the problems of processing of genome ChIP-seq profiles for detections of transcription factors binding site in a genome, determining the peaks of such profiles and search the binding sites in the nucleotide sequences of the peaks.
*Methods and Algorithms*: The computer programs have been developed to analyze the location of the binding sites in the genome relative to gene regions, to calculate clusters of such sites and visualize their positions in the genome. Clusters of binding sites of transcription factors in the human genome have been calculated using the Cistrome database.
*Results*: We have calculated matrices of the joint occurrence of pairs of binding sites of different transcription factors in the genome for various types of tissues and cells. A computational experiment on the computer generation of random clusters in the genome was carried out, as well as an assessment of the occurrence of large clusters for experimentally obtained binding sites of transcription factors in the human genome. The patterns of occurrence of binding sites of pluripotency factors in embryonic stem cells were described. The developed software is available on request to the authors.
*Conclusion*: Problem of analysis of genome distribution of transcription factor binding sites in human genome is of importance for personalized medicine and genomics studies.
*Acknowledgements*: The work was supported by the RFBR and ICG SB RAS budget project (0324-0019-0040).

*References*

1. Spitsina A.M., Bragin A.O., Dergilev A.I., Chadaeva I.V., Tverdokhleb N.N., Galiyeva E.R., Tabikhanova L.E., Orlov Y.L. Computer tools for analysis of transcriptomics data: program complex ExpGene. *Program Systems: Theory Applications*. 2017;8:2(33):45–68. DOI 10.25209/2079-3316-2017-8-2-45-68. (in Russian)