DOI 10.18699/SBB-2022-014

Analysis of multiple transcription factors binding in the model plant genomes

Kononov V.A.[1], Dergilev A.I.[2], Orlova N.G.[3, 4], Orlov Y.L.[1, 5]*

[1] *Agrarian and Technological Institute, Peoples' Friendship University of Russia, Moscow, Russia*

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

[3] *Financial University under the Government of the Russian Federation, Moscow, Russia*

[4] *Moscow State Technical University of Civil Aviation, Moscow, Russia*

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

** orlov-yul@rudn.ru*

Key words: plant biology, genomics, transcription factor binding, ChIP-seq, databases, *Arabidopsis thaliana*

Motivation and Aim: The development of high-throughput genomic sequencing coupled with chromatin immunoprecipitation technologies allows studying the binding sites of the protein transcription factors (TF) in the genome scale. The growth of data volume on the experimentally determined binding sites raises qualitatively new problems for the analysis of gene expression regulation, prediction of transcription factors target genes, and regulatory gene networks reconstruction. Genome regulation remains an insufficiently studied though plants have complex molecular regulatory mechanisms of gene expression and response to environmental stresses. It is important to develop new software tools for the analysis of the TF binding sites location and their clustering in the plant genomes, visualization, and the following statistical estimates.

Methods and Algorithms: We used experimental ChIP-seq data to model TF binding as the alternative to sequence-based prediction [2, 4]. This study presents application of the analysis of multiple TF binding profiles in evolutionarily distant model plant organisms. The construction and analysis of non-random ChIP-seq binding clusters of the different TFs in mammalian embryonic stem cells were discussed earlier using similar bioinformatics approaches and computer scripts [3]. We used ChIP-seq data in several plant plants, including *Arabidopsis thaliana, Physcomitrella patens,* and *Chlamydomonas reinhardtii* to model TF binding clusters in their genomes and estimate the distribution of the TF in the clusters found [2]. The data on TF binding were collected from GEO NCBI, PlantTFDB (http:planttfdb.cbi.pku.edu.cn, http://ucsc.gao-lab.org/), PlantRegMap (http:plantregmap.cbi.pku.edu.cn).

Results: The clusters of TF binding sites may indicate the gene regulatory regions, enhancers and gene transcription regulatory hubs. Further, it can be used for analysis of the gene promoters as well as a background for transcription networks reconstruction. The heatmaps of TF binding in Arabidopsis allow delineating two major groups of factors: (GATA, ARF, BBR-BPC, HSF, LBD, FAR1, C2H2, GRAS, MYB, TCP, E2F/DP, HD-ZIP, CH3, NAC, AP2, bZIP, WRKY) and (B3, G2-like, ZF-HD, MADS, DOF, YABBY), or, in other words, relating to the GATA and related G2-like transcription factors. We discuss the statistical estimates of the TF binding sites clusters in the model plant genomes. The distributions of the number of different TFs per binding cluster follow same power law distribution for all the genomes studied. the evolutionarily

more ancient factors GATA and MYB are presented in the clusters of sites in all studied plant species. The binding clusters in *Arabidopsis thaliana* genome will be discussed in detail.

Conclusion: We show that regulatory gene networks structure in plants corresponds to similar transcription factor binding in plant genomes. The statistical estimates of the transcription factor binding clusters in genomes could be extended to other evolutionary not related genomes. Integration of experimental genomic information, Big Data, in general, represents an important problem in bioinformatics, requiring the integration of existing software tools and approaches [1, 5]. The presented statistical analysis of clusters of binding sites in plant genomes helps in solving the problem, developing approaches to the study of the evolutionary origin of enhancers and gene networks

Acknowledgements: This publication has been supported by the RUDN University Scientific Projects Grant System, project No. R.3-2022-ins).

References

1. Chen M., Harrison A., Shanahan H., Orlov Y. Biological big bytes: Integrative analysis of large biological datasets. *J Integr Bioinform.* 2017;14(3):20170052. doi: 10.1515/jib-2017-0052.
2. Dergilev A.I., Orlova N.G., Dobrovolskay O.B., Orlov Y.L. Statistical estimates of multiple transcription factors binding in the model plant genomes based on ChIP-seq data. *J Integr Bioinform*. 2021;19(1):20200036. doi: 10.1515/jib-2020-0036.
3. Dergilev A.I., Spitsina A.M., Chadaeva I.V., Svichkarev A.V., Naumenko F.M., Kulakova E.V., Vityaev E.E., Chen M., Orlov Y.L. Computer analysis of colocalization of the TFs' binding sites in the genome according to the ChIP-seq data. *Russ J Genet Appl Res*. 2017;7(5):513-522. doi: 10.1134/S2079059717050057.
4. Kononov V.A., Orlova N.G., Orlov Y.L. Computer analysis of plant stress response: gene networks and transcriptional regulation. In: Proceedings of the International Conference "Scientific research of the SCO countries: synergy and integration". Part 3. Reports in English. March 31, 2022. Beijing: PRC. Scientific publishing house Infinity, 2022;143-152.
5. Orlov Y.L., Galieva A.G., Orlova N.G., Ivanova E.N., Mozyleva Y.A., Anashkina A.A. Reconstruction of gene network associated with Parkinson disease for gene targets search. *Biomeditsinskaya Khimiya*. 2021;67(3):222-230. doi: 10.18097/PBMC20216703222.