DOI 10.18699/SBB-2022-050

enRest tool for transcription factor binding sites overrepresentation analysis in RNA-seq data

Tsukanov A.V.*, Levitsky V.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
** tsukanov@bionet.nsc.ru*

Contemporary transcriptomic technology such as RNA-sequencing provides reliable approach to identify differentially expressed genes (DEG) across different cell types, tissues or conditions of stimuli/treatments. This standard analysis reveals many co-expressed genes or even gene networks that coordinately explain the observed biological responses. However, specific mechanisms of DEG regulation still too far from a full clarification. Hence, identifying the transcription factors responsible for observed changes in gene expression patterns is an important step in understanding gene regulatory networks. To determine the overrepresentation of transcription factor binding sites (TFBS) in regulatory regions of DEG we developed the tool *enRest* that took results of the RNA-seq experiment and the list of motifs (positional frequencies matrices, PFMs) from a public library, e.g. HOCOMOCO, CISBP and JASPAR. The tool scans 5' upstream regions (–2000; +1) of transcription start sites of all protein-coding genes of genome, hereafter called promoters. The tool divides promoters into the foreground and background sets. The former consists of promoters of DEG (RNA-seq experiment, adjusted p-value < 0.05). The latter includes the same number of promoters randomly sampled from the rest genes of a genome. To estimate significance of the overrepresentation of potential TFBSs we applied the Monte Carlo approach. We repeated many sampling iterations of the background set to estimate the distributions of the number of sites per sequence in this set. Finally, we computed the enrichment p-value for each motif. The tool was tested on RNA-seq data of *H. sapiens* and *A. thaliana*. Our new tool can efficiently identify TFBSs that are statistically overrepresented in DEG.
Acknowledgements: The study was supported by Russian Science Foundation project No. 21-14-00240.