DOI 10.18699/MM-HPC-BBB-2018-63

# ARGO_CEL: GPU based approach for potential composite elements discovery in large DNA datasets

O. Vishnevsky[1, 2]*, A. Bocharnikov[1], N. Kolchanov[1, 2]
[1] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
** e-mail: oleg@bionet.nsc.ru*

**Key words:** oligonucleotide motifs, composite elements, transcription regulation, ChIP-Seq

*Motivation and Aim*: Composite elements play an important role in the regulation of transcription. Existing methods for the revealing of potential composite elements (PCE) are usually based on assessment of the significance of the mutual presence of the predicted transcription factor binding sites (TFBS). In this case, the recognition of potential TFBSs is performed using weight matrices or other methods trained on samples of binding sites of known transcription factors. Thus, such methods essentially depend on the completeness of training samples and information on existing TFs.
*Methods and Algorithms*: We have proposed a method for *de novo* discovery of PCE, which does not require preliminary information about the localization of potential TFBS. Based on the proposed approach, the Internet-accessible resource http://argo.bionet.nsc.ru/cgi-bin/ARGO_CEL/Argo_CEL.cgi was created, which allows the user to obtain sets of mutually present groups of significant motives in the analyzed sample of nucleotide sequences, and annotate them. Such groups of motifs can correspond both to binding sites of known transcription factors, and to certain physico-chemical features of the nucleotide context of regulatory regions of genes. The method developed is based on *de novo* discovery of significant degenerate oligonucleotide motifs of a fixed length, written in a 15-letter IUPAC [1]. Discovered motifs are clustered into groups, corresponding to individual regulatory elements. After that, the significance of the mutual presence of the obtained groups of motives is evaluated and the PCE are identified. The ARGO_CEL system allows annotation of the constructed groups of motives on the base of Transfac TFBS.
*Results*: The developed approach was used to analyze the results of the FoxA ChIP-Seq dataset [2]. Several groups of significant motifs were obtained using the ARGO_CEL system. Some of them correspond to known FoxA binding sites, and some to other TFBS. It was shown that some of these groups reliably co-occur in the control set and can correspond to PCEs.
*Conclusion*: We have developed a *de novo* method for discovery of potential composite elements that does not require preliminary information on the TFBS. Using the proposed approach, context signals are identified in the ChIP-Seq dataset, which can correspond to potential composite elements.
*Acknowledgements*: The work was supported in part by ICG SB RAS budget project (0324-2016-0008) and integration project of SB RAS (0324-2018-0021).

## *References*

1. Vishnevsky O.V., Bocharnikov A.V., Kolchanov N.A. (2017) ARGO_CUDA: Exhaustive GPU based approach for motif discovery in large DNA datasets. Journal of Bioinformatics and Computation Biology, Dec 10:1740012.
2. Wederell E.D. et al. (2008) Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. Nucleic Acids Research. 36:4549-4564.