

Development of a computational pipeline to search and analyze genes of multidomain protein families in plant genomes

Bocharnikova M.E., Afonnikov D.A.*

Kurchatov Genome Center of the ICG SB RAS, Novosibirsk, Russia,

*Federal Research Center Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia

E-mail: m.bocharnikova@g.nsu.ru

Modern sequencing technologies have led to a huge flow of data on plant genome sequences. This data requires a high-performance computer annotation to determine protein functions. In this paper, we propose the OrthoDom computational pipeline, which is aimed at searching for orthologous proteins taking into account their domain composition. The pipeline is implemented based on the Snakemake system. Protein sequences of genomes, sets of reference proteins for which functions and domains are known, as well as HMM profiles for the domains of these proteins are fed to the input of the pipeline. The pipeline outputs a list of orthologous sequences for these reference proteins, as well as information about their domain composition and phylogenetic trees for the identified orthologous families. The pipeline was used to identify orthologs in the genomes of cereals.

The work was funded by the Kurchatov Genome Center of the ICG SB RAS (project no. 075-15-2019-1662).

The main publications of authors on the subject of the abstract:

Turnaev I. I., Bocharnikova M. E., Afonnikov D. A. *Human phospholipases A2: functional and evolutionary analysis* // *Vavilov Journal of Genetics and Breeding*. - 2023. - T. 26. - №. 8. - C. 787-797.

Take-home message:

We propose the OrthoDom computational pipeline, which is aimed at searching for orthologous proteins taking into account their domain composition. We tested the systems accuracy and implemented it for plant protein analysis.