

Computer annotation of plant protein sequences based on sequence similarity and orthology

Malyugin E.^{1,2*}, Mustafin Z.^{2,3}, Pronozin A.^{2,3}, Genaev M.^{2,3}, Afonnikov D.^{2,3}

¹ Novosibirsk State University, Novosibirsk, Russia

² Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

³ Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

* evgeny.malyugin98@gmail.com

Key words: annotation, orthologs, gene age

Motivation and Aim: With the advent of new generation sequencing methods, the number of sequenced genome sequences is increasing tremendously. All this makes it possible to recognize the structure of the genome and predict the localization of protein-coding genes with high accuracy. However, for recognized genes, the function is usually unknown. Methods existing for this problem mainly use the search for homologous sequences with known functions described in terms of Gene Ontology to predict the function of genes. In this case, an important factor in sequence annotation is the age of the gene [1, 2], which is associated with the presence of its orthologs in a large number of taxa. For genes that have occurred recently during evolution (young or taxon-specific genes), orthologs are absent or poorly represented in other organisms. For them, function prediction is difficult and its accuracy is lower than for ancient genes, for which homologs are more widely represented. Thus, the development of a method for annotation of gene functions that would work with high accuracy regardless of the age of the genes is relevant.

Methods and Algorithms: We proposed an algorithm for gene function annotation based on similar sequence search and orthologous group analysis in the OrthoDB v. 10.0 [3]. The search for homologous sequences was performed in the database using the Usearch v 11 algorithm [4]. The method was tested on the sequences of protein-coding genes of a number of plant genomes, including *Arabidopsis thaliana*, rice, moss, algae, etc. Orthoscape software was used to estimate gene ages [5].

Results: The accuracy of the proposed algorithm when analyzing the *A. thaliana* genome was comparable with the eggNOG-mapper method [6]. Thus, the proportion of correct predictions (TP) among all positive predictions for our method was 86.6 %, for eggNOG-mapper method 85 %. It was shown that with certain parameters of the algorithm, the accuracy of function prediction in genes of different ages (including young genes) is close. Thus, the proposed method can reduce the differences in the accuracy of function prediction for genes of different ages, for better performance.

Acknowledgements: Development of algorithms was funded by the Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, agreement with the Ministry of Education and Science of the Russian Federation, No. 075-15-2019-1662.

References

1. Liebeskind B.J., McWhite C.D., Marcotte E.M. Towards consensus gene ages. *Genome Biol Evol.* 2016;8(6):1812-1823.

2. Mustafin Z.S., Zamyatin V.I., Konstantinov D.K., Doroshkov A.V., Lashin S.A., Afonnikov D.A. Phylostratigraphic analysis shows the earliest origination of the abiotic stress associated genes in *A. thaliana*. *Genes (Basel)*. 2019;10(12):963.
3. Kriventseva E.V., Kuznetsov D., Tegenfeldt F., Manni M., Dias R., Simão F.A., Zdobnov E.M. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 2019;47(D1):D807-D811.
4. Robert C.E. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460-2461.
5. Mustafin Z.S., Lashin S.A., Matushkin Yu.G., Gunbin K.V., Afonnikov D.A. Orthoscape: a Cytoscape plugin for grouping and visualization KEGG based gene networks by taxonomy and homology principles. *BMC Bioinformatics*. 2017;18(1):1-9.
6. Cantalapiedra C.P., Hernández-Plaza A., Letunic I., Bork P., Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol.* 2021;38(12):5825-5829.