

Essential and advantageous genes detection in prokaryotic genomes

Moshensky D.*, Alexeevski A.

A.N. Belozersky Institute of Physico-Chemical Biology MSU, Moscow, Russia

**moshenskydenis@gmail.com*

Key words: essential gene, genes annotation, comparative genomics, Ka/Ks, orthologs alignment

Among all proteins, there are those whose loss of functionality leads to a significant adaptation decrease. Such proteins are claimed as essential proteins. Essential proteins genes detection is an important challenge for the fight against pathogens and for biotechnologies development. In the first case, essential proteins could be the targets for new antibiotics. In the second, essential proteins should be remained while an industrial strains modification or should be replaced with proteins with the same functions, resistant to required conditions. Essential proteins detection is conducted with both experimental and computational methods. Experimental methods have a number of drawbacks: extreme difficulty, strong dependency on the quality of a genome annotation, weak reproducibility. Computational methods use comparative genomics approaches, genes expression analysis, proteins interactions and other approaches. Most of them produce accurate results only for well-studied organisms because these methods require qualitative annotation not only for analyzed genome but also for related organisms. Besides, generally, these methods are based on additional experimental data like RNA-seq data or transposon mutagenesis data. We have developed a program EAGLE (Essential and Advantageous Genes Location Explorer, <https://github.com/loven-doo/EAGLE>) for essential genes detection in bacterial genomes. This program predicts essential and advantageous (other non-essential functional genes) genes for an input genome sequence using only genome sequences for related bacteria that are organized in the database called EAGLEdb. EAGLEdb consists of basic taxons (taxons of same level or a set of same level taxons). EAGLE detects all ORFs in input genome and calculates 11 features for each ORF. Among these features are the difference between reference phylogenetic tree and phylogenetic tree of an ORF orthologs from the basic taxon genomes, standard deviation from the uniformity for conservative columns in ORF orthologs alignment, Ka/Ks ratio [1–3] the length of the ORF, representation in the basic taxon genomes, two features describing stop-codons distribution, four features based on distances between sequences in ORF orthologs alignment.

We have applied our program to the genomes from DEG database [4]. Functional genes in complete genome of an organism can be detected with up to 90 % accuracy using only relative genomes without any annotation.

Acknowledgements: This work is supported by RSF grant No. 21-14-00135.

References

1. Yang Z., Bielawski J.P. *Trends Ecol Evol.* 2000;15(12):496-503.
2. Yang Z., Nielsen R. *Mol Biol Evol.* 2000;17:32-43.
3. Wang D., Zhang Y., Zhang Z., Zhu J., Yu J. *Genomics Proteomics Bioinform.* 2010;8(1):77-80.
4. Zhang R. *Nucleic Acids Res.* 2004;32:D271-D272.