

Integration of bioinformatic tools for assessing promoter activity and gene translation elongation efficiency to predict microbial gene expression levels

Klimenko A.I.^{1, 2*}, Fairushina K.S.³, Ponomarenko M.P.¹, Kolchanov N.A.^{1, 2}, Lashin S.A.^{1, 2, 3}

¹ Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

² Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

³ Novosibirsk State University, Novosibirsk, Russia

* klimenko@bionet.nsc.ru

Key words: translation efficiency; protein abundance prediction; codon usage

Motivation and Aim: Prediction of gene expression in prokaryotes is a fundamental bioinformatics problem that is significant for further advances in both basic research and biotechnology. The gene expression values, explicated as the amount of active protein, depend on various factors, primarily transcriptional and translational regulation. Therefore, it is important to evaluate and integrate these upstream data to accurately predict gene expression. Contextual and structural characteristics determining promoter activity have a serious impact on the transcriptional activity of the corresponding genes, while in some microorganisms it has been shown that translation elongation is the second most important factor determining the amount of protein in the cell after the mRNA level. In this regard integration of bioinformatic tools for assessing promoter activity and gene translation elongation efficiency is relevant for solving the problem of predicting microbial gene expression levels.

Methods and Algorithms: The solution to this problem requires protein-level expression profiles covering a significant portion of the microorganism's genes. Within the scope of this study, we used the data gathered for prokaryotes representing various phylogenetic groups from Pax-DB database, which contains protein expression data with sufficiently high gene coverage [1]. For the provided genomes the translation efficiency features for each protein-coding gene have been calculated using the EloE tool [2]. The ACTIVITY [3] system implements an approach to predicting the activity of a site based on its primary nucleotide sequence, in particular the activity of promoters, on the basis of the PROPERTY database containing information on the physical and chemical properties of DNA. Regression models of gene expression at the protein level incorporating bioinformatic characteristics of gene translation elongation efficiency and corresponding promoter activity were built using Python tools.

Results and Conclusion: In this work, we have been integrating software tools aimed at predicting the efficiency of various stages of gene expression, namely translation elongation and promoter activity of the corresponding genes. The designed framework allows us to expand the set of considered predictors of expression level, including both context-dependent conformational and physico-chemical properties of associated sequences and various derived indices evaluating the properties associated with the efficiency of translation elongation. The integration of these approaches with

experimental expression data obtained at different levels appears promising for future gene expression prediction studies.

Funding: The study is supported by the Budget Project FWNR-2022-0020.

References

1. Huang Q., Szklarczyk D., Wang M., Simonovic M., Mering C. von PaxDb 5.0: Curated Protein Quantification Data Suggests Adaptive Proteome Changes in Yeasts. *Mol Cell Proteomics*. 2023;22(10):100640. doi 10.1016/j.mcpro.2023.100640
2. Sokolov V., Zuraev B., Lashin S., Matushkin Y. Web application for automatic prediction of gene translation elongation efficiency. *J Integr Bioinform*. 2015;12(1):256. doi 10.2390/biecoll-jib-2015-256
3. Ponomarenko M.P., Ponomarenko J.V., Frolov A.S., Podkolodny N.L., Savinkova L.K., Kolchanov N.A., Overton G.C. Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins. *Bioinformatics*. 1999;15(7-8):687-703. doi 10.1093/bioinformatics/15.7.687