DOI 10.18699/SBB-2022-328

# Integration of transcriptomic, genomic, and proteomic data in order to assess the impact of translation efficiency features on protein abundance for various prokaryotes

Korenskaia A.E.[1, 2]*, Matushkin Yu.G.[3], Lashin S.A.[1, 2, 3], Klimenko A.I.[1, 3]

[1] *Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Novosibirsk State University, Novosibirsk, Russia*

[3] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

* *korenskaia@bionet.nsc.ru*

**Key words:** translation efficiency, protein abundance prediction, codon usage

*Motivation and Aim*: Due to fact that the protein abundance controls many physiological processes, the calculation of protein abundance is of great importance for medical and biological studies. The protein abundance is determined by the regulation of the gene expression on the transcription level, the efficiency of the translation stage, and the post-translation level regulation. It is known that the level of transcript expression impacts protein abundance the most. The correlation coefficient between the expression at the transcriptomic and the proteomic level approaches to 0.5, which has been shown for *Escherichia coli* [3] and *Desulfovibrio vulgaris* [4]. At the same time, translation efficiency tends to reflect gene expression at the basal level, as a result, it could highly correlate with the protein abundance – more than 0.5 for several prokaryotes. Hence, the combination of the transcriptomic data and the translation efficiency features, calculated based on the sequence, are supposed to effectively predict the protein abundance. According to the studies [2], translation efficiency depends on the translation initiation efficiency, which is affected by the mRNA structure near the start codon, and translation elongation efficiency, which impact by codon bias on mRNA, and secondary structures on mRNA. It should be noted that for various prokaryotes these features impact on the protein abundance differently. The goal of this study is to assess the impact of those factors on a protein abundance, considering the gene expression at the protein and transcriptome levels, as well as to build a set of regression models to predict protein abundance in various prokaryotes.

*Methods and Algorithms*: The transcriptomic (RNA-seq) and proteomic (obtained by LC-MS/MS spectroscopy) data were collected from the studies, where these quantifications were carried out under the same conditions. The data gathered for prokaryotes representing various phylogenetic groups. For the provided genomes the translation efficiency features for each protein-coding gene have been calculated. The translation initiation efficiency features are calculated using the ostir [5] tool, the translation elongation efficiency features are obtained by the EloE tool [6] and the BCAWT tool [1]. A number of regression models were built, in which a set of proteins' expression levels acts as a dependent variable, while a set of transcripts' expression levels as well as a translation efficiency parameters act as independent variables for each of the organisms under study. To assess the impact of each parameter, a partial regression coefficient was calculated for each of these features. In addition, Pearson correlation coefficients between a parameter value and expression data were calculated. The

obtained coefficients were compared between the translation efficiency parameters of an organism, as well as between the prokaryotes belonging to various phylogenetic groups. *Results and Conclusion*: It has been demonstrated that both correlation coefficients and partial regression coefficients between translation efficiency features and protein abundance show diversity for various prokaryotes. For example, for *Leptospirillum ferriphilum* it has been shown that a feature, related to the number of secondary structures on mRNA impacts protein abundance the most. Notably, in different conditions correlation between expression at the level of mRNA and at the protein level is diverse (0.55 and 0.63), but adding the discussed parameter increases the correlation coefficient by 0.05. In conclusion, the usage of a combination of data on mRNA abundance and its secondary structure allows one to improve the protein abundance prediction power for certain prokaryotes. However, for some organisms other parameters are more significant, in particular, the characteristics of the codon usage. Thus, for effective protein prediction it is necessary to take into account different factors of translation efficiency for various prokaryotes.

*Acknowledgements*: The study is supported by the Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS (075-15-2019-1662).

*References*

1. Anwar A. BCAWT: Automated tool for codon usage bias analysis for molecular evolution. *J Open Source Softw.* 2019;4(42):1500. doi: 10.21105/joss.01500.
2. Chiaruttini C., Guillier M. On the role of mRNA secondary structure in bacterial translation. *Wiley Interdiscip Rev RNA.* 2019;11(3):1-21. doi: 10.1002/wrna.1579.
3. Guimaraes J.C., Rocha M., Arkin A.P. Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic Acids Res.* 2014;42(8):4791-4799. doi: 10.1093/nar/gku126.
4. Nie L., Wu G., Zhang W. Correlation between mRNA and protein abundance in Desulfovibrio vulgaris: A multiple regression to identify sources of variations. *Biochem Biophys Res Commun.* 2006;339(2):603-610. doi: 10.1016/j.bbrc.2005.11.055.
5. Roots C.T., Lukasiewicz A., Barrick J.E. OSTIR: open source translation initiation rate prediction. *J Open Source Softw*. 2021;6(64):3362. doi: 10.21105/joss.03362
6. Sokolov V., Zuraev B., Lashin S., Matushkin Y. Web application for au.tomatic prediction of gene translation elongation efficiency. *J Integr Bioinform*. 2015;12(1):256. doi: 10.2390/biecoll-jib-2015-256.