doi 10.18699/bgrs2024-1.2-27

PARADE: deep learning for prediction and rational design of untranslated regions with desired tissue-specific expression

Zinkevich A.[1]*, Aristova E.[1], Lee S.[2], Mittmann T.[2], Penzar D.[3, 4], Yousefi H.[2], Khoroshkin M.[2], Kulakovskiy I.V.[4]

[1] *Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia*

[2] *University of California, San Francisco, CA, USA*

[3] *Artificial Intelligence Research Institute (AIRI), Moscow, Russia*

[4] *Institute of Protein Research, RAS, Pushchino, Russia*

* *arsenii.z.work+bgrs@gmail.com*

Key words: synthetic biology; diffusion neural network; generative modelling; rational design; regulatory genomics, RNA-regulation; artificial intelligence

Motivation and Aim: Gene expression is a multi-level process that determines the morphological and functional characteristics of a cell. Different levels of gene expression are controlled by diverse regulatory mechanisms, from DNA epigenetic marks to post-translational modification of proteins. A significant regulatory layer is mediated by mRNA untranslated regions (5'UTR and 3'UTR), which significantly influence the mRNA translation and stability of the transcribed RNA molecules.

With the advent and rapid development of massively parallel reporter assays, it became possible to profile the regulatory effects of thousands of UTRs and identify and model the sequence-level regulatory grammar with the help of deep learning. Yet, existing studies fail to properly capture and predict the UTR-dependent pattern of cell type-specific mRNA translation and stability.

Methods and Algorithms: This study uses the results of MPRAs [1] performed with 20k+20k 5' and 3' UTR sequences in five distinct cell types. To analyze the data, we adapted the LegNet neural network [2] to the regression of the normalized reporter fluorescence from the UTR sequence in the particular cell type. The neural network was implemented in Python within the PyTorch framework.

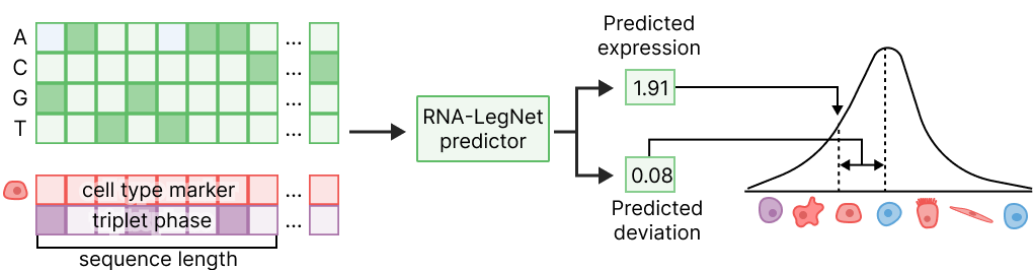

Fig. 1. Simplified double regression task. A neural network model is trained to simultaneously predict UTR sequence expression and expression difference from the mean

Results and Conclusions: Here we present the PARADE (Prediction and rational design of RNA UTRs) framework, a deep learning solution for estimating cell type-specific activity of 5'- and 3'UTRs, and performing rational design of UTR sequences constrained by desired expression patterns. To generate new sequences, PARADE employs 3 different methods: fine-grain filtering of random sequences, genetic algorithm, and diffusion-denoising neural network. The performance of the methods was evaluated and validated experimentally.

In the end, PARADE predicts the expression levels for the test set with r = 0.824 for 5'UTR and r = 0.793 for 3'UTR sequences, outperforming the previous specialized Optimus-5-prime model [3], and successfully generates sequences with desired directional tissue-specific activity.

Funding: A.Z. was supported by a personal fellowship from the Non-commercial Foundation for Support of Science and Education "INTELLECT". D.P. and I.V.K. were supported by RSF grant 20-74-10075.

References

1. Litterman A.J., Kageyama R., Le Tonqueze O. et al. A massively parallel 3' UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome Res.* 2019;29(6):896-906. doi 10.1101/gr.242552.118
2. Penzar D., Nogina D., Noskova E. et al. LegNet: a best-in-class deep learning model for short DNA regulatory regions. *Bioinformatics*. 2023;39(8):btad457. doi 10.1093/bioinformatics/btad457
3. Sample P.J., Wang B., Reid D.W., Presnyak V., McFadyen I.J., Morris D.R., Seelig G. Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol.* 2019;37(7):803-809. doi 10.1038/s41587-019-0164-5