doi 10.18699/bgrs2024-1.2-08

Enformer deep learning model for predicting mutation effects on gene transcriptional activity

Belokopytova P.[1, 2]*, Kuratov Y.[3], Fishman V. [1, 2, 3]

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

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

[3] *AIRI, Moscow, Russia*

* *belokopytova@bionet.nsc.ru*

Key words: deep learning; gene expression; enhancer–promoter interactions

Motivation and Aim: Investigation of mechanisms underlying the regulation of gene expression is one of the most relevant areas in modern biology. Genomic variants in the coding genomic regions can lead to the synthesis of defective proteins, which can be the cause of various diseases. However, there may also be mutations in non-coding genomic regions resulting in changes in transcriptional activity leading to pathology phenotype. Thus, the task of finding a connection between genomic variants and transcription activity of genes becomes especially actual. Currently, there are no convenient and accessible tools for evaluating the effects of mutations in non-coding genomic regions. The majority of current methods focus on the analysis of exomic variants, but due to the decreasing cost of sequencing, there is a gradual shift towards genome-wide analysis. This is making the study of single nucleotide variants and small structural rearrangements in non-coding regions of the genome increasingly accessible. This in turn requires the development of methods for their interpretation. In recent years, deep learning algorithms have been actively used in genomics, including the prediction of gene transcriptional activity.

Methods and Algorithms: We chose the state-of-the-art deep learning model Enformer [1], which was developed to predict gene transcriptional activity by DNA sequence. This model is based on a relatively new neural network architecture – Transformer. This neural network architecture makes it possible to integrate information about genomic interactions over a long distance (up to 200 kb bp). The attention layers of the neural network and *in silico* mutagenesis make it possible to determine the contribution of each nucleotide of the DNA sequence to the prediction. It has been shown that sequences with the greatest predictive weight are regulatory or insulatory nucleotide sequences. The ability to vary different genetic variants makes it possible to use the Enformer model to study quantitative trait loci (eQTL) without having to study hundreds or thousands of individual gene expression profiles. It is therefore promising to use the Enformer model as the basis for a tool that predicts changes in the transcriptional activity of genes associated with different genomic variants.

Results: We aimed to check how well the Enformer model deals with predicting changes of gene transcriptional activity on real experimental examples. In recent papers [2–4] has been shown that the Enformer model often incorrectly predicts the direction of transcriptional changes for individual genomes. It is worth noting that the level of gene transcription varies very slightly among healthy people and such changes of transcriptional activity are difficult to detect. There weren't any investigations of the Enformer model ability to predict a critical change in gene expression in the case of

single nucleotide or short genomic mutations and chromosomal rearrangements leading to pathology. We analyzed the literature and created a dataset that includes mutations in promoters and enhancers of genes leading to critical gene expression changes (at least two times).

Conclusion: We have shown that in the most investigated cases the Enformer model predicts right direction and level of transcriptional activity changes. This fact allows creating a tool for interpreting non-coding genomic variants that will expand the available information about genetic anomalies obtained by different genomic analyses. This may help medical genetics to put forward new hypotheses regarding the molecular mechanisms associated with the pathological phenotype.

Funding: Ministry of Education of Russia project № 075-15-2024-539.

References

1. Avsec Ž., Agarwal V., Visentin D. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*. 2021;18(10):1196-1203. doi 10.1038/s41592-021-01252-x
2. Huang C. et al. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat Genet*. 2023;55(12):2056-2059
3. Sasse A. et al. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat Genet*. 2023;5(12):2060-2064
4. Tang Z., Toneyan S., Koo P.K. Current approaches to genomic deep learning struggle to fully capture human genetic variation. *Nat Genet*. 2023;55(12):2021-2022