

Использование персонализированных геномов для предсказания эффектов однонуклеотидных мутаций в регуляторных областях

Научный руководитель – Пензар Дмитрий

Смирнова Диана Алексеевна

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова, Факультет
биоинженерии и биоинформатики, Москва, Россия

E-mail: didi2003n@gmail.com

Целью работы является обучение моделей глубокого обучения на персонализированных данных об экспрессии генов и предсказание при помощи обученных моделей влияния однонуклеотидных вариантов в регуляторных районах генов.

Важной проблемой биоинформатики является прогнозирование влияния индивидуальных вариаций генома на экспрессию генов. Из успешных моделей в этой области Enformer демонстрируют высокую точность, он использует мультизадачный подход, предсказывая результаты множества экспериментов одновременно (CAGE, ChIP Histone, ChIP TF, DNase/ATAC seq), что требует значительных вычислительных ресурсов, однако в контексте предсказания одной задачи это не всегда оправданно. Модели, предсказывающие результаты одного эксперимента, не уступают Enformer в контексте одной задачи, это говорит о целесообразности использования более простых моделей.

В то же время известно, что все существующие модели, обученные на полногеномных экспериментах, плохо справляются с задачей предсказания эффектов однонуклеотидных мутаций в регуляторных областях. Для задачи была выбрана CNN модель LegNet, обучающаяся на предсказание каждого эксперимента индивидуально [4] и являющаяся победителем конкурса DREAM-2022 по прогнозированию экспрессии генов.

Модель была обучена на данных lentiMPRAs, которые позволяют тестировать последовательности в их естественном хроматиновом контексте. Данные включали измерения активности более чем 680 000 последовательностей в трех типах клеток (HepG2, K562, WTC11).

Затем были взяты 67 библиотек PRO-cap LCL, обработаны, выделены пики и сопоставлены с восстановленными персонализированными последовательностями. Модель LegNet была дообучена на этих данных для улучшения ее возможности предсказывать эффекты однонуклеотидных вариантов на экспрессию.

Модель была протестирована на данных MPRA. Также мы сравнили качество предсказаний с другой моделью, обученной на тех же данных - CLIPNET [2]. Дообученная модель превзошла CLIPNET на тестовых данных PRO-cap, достигнув корреляции Пирсона 0,6. Также модель была протестирована на базах данных ADAstra и UDACHA [1], содержащих информацию об аллель-специфичных событиях.

Источники и литература

- 1) Abramov S., Boytsov A., Bykova D., Penzar D.D., Yevshin I., Kolmykov S.K., Fridman M.V., Favorov A.V., Vorontsov I.E., Baulin E., Kolpakov F., Makeev V.J., Kulakovskiy I.V. Landscape of allele-specific transcription factor binding in the human genome // Nat. Commun. – 2021. – Vol. 12, № 1. – P. 2751.
- 2) Adam Y. He, Charles G. Danko, Dissection of core promoter syntax through single nucleotide resolution modeling of transcription initiation, bioRxiv 2024.03.13.583868; doi: <https://doi.org/10.1101/2024.03.13.583868>

- 3) Agarwal, V., Inoue, F., Schubach, M. et al. Massively parallel characterization of transcriptional regulatory elements. Nature (2025). <https://doi.org/10.1038/s41586-024-08430-9>

Иллюстрации



Рис. : Схема обработки данных и получения датасета для дообучения

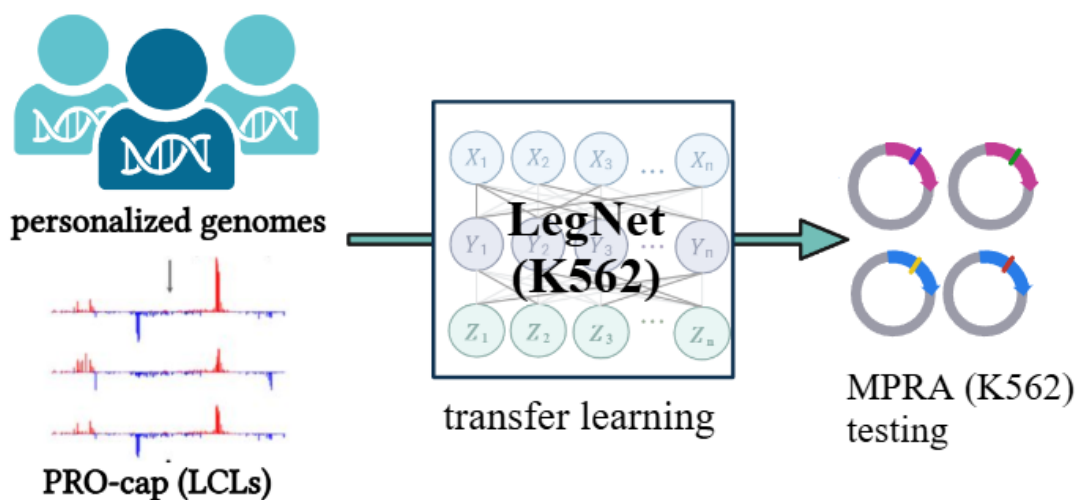


Рис. : Схема дообучения и тестирования модели