

Улучшение предсказания активности энхансерных последовательностей при помощи алгоритмов машинного обучения

Научный руководитель – Пензар Дмитрий Дмитриевич

Зубова Ева Александровна

Студент (специалист)

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

E-mail: ewzubow@mail.ru

Данная работа посвящена изучению регуляторных последовательностей человека. С мутациями в них ассоциировано множество клинически значимых заболеваний, таких как диабет, шизофрения и болезнь Паркинсона.

В работе [1] было проведено исследование активности энхансерных последовательностей в клеточных линиях HepG2, K562 и WTC11 при помощи лентивирусных векторов (lentivirus-based massively parallel reporter assays, lentiMPRA). Эксперимент заключался в интеграции потенциальных энхансерных последовательностей в геном, после чего измерялась сила их влияния на экспрессию репортерного гена. Используя эти данные, можно обучить различные модели машинного обучения для предсказания энхансерной активности. С этой целью в оригинальной статье были использованы биохимические признаки — результаты различных полногеномных экспериментов (ChIP-seq, ATAC-seq, DNase-seq) и нуклеотидные последовательности энхансеров.

Наилучшее качество в статье показала модель LegNet [3], обученная исключительно на нуклеотидной последовательности, ей уступила модель Enformer [2] предобученная на предсказание биохимических признаков по последовательности и дообученная на предсказание активности. В то же время, на данных малого размера, LegNet уступала предобученной модели Enformer. Для того чтобы выяснить, вносят ли дополнительную информацию биохимические признаки, выучиваемые Enformer, мы модифицировали архитектуру LegNet за счет добавления признаков, предсказанными Enformer. Видоизмененную архитектуру мы назвали Legformer.

Полученная сеть демонстрирует лучшее качество по сравнению с LegNet (рис. 1), что говорит о том, что для достижения оптимального качества в предсказание регуляторных закономерностей все еще нужны внешние данные (в данном случае выученные моделью Enformer). При валидации на независимых данных об *in silico* мутагенезе из статьи Kircher et al. [4], наша модель показала лучшее качество на 3 из 4 энхансерах при предсказании эффектов мутаций, по сравнению с моделями LegNet и Enformer.

Таким образом, разработанный нами гибридный подход позволяет более точно предсказывать эффекты регуляторных вариантов на экспрессию, компенсируя недостатки обоих ранее использовавшихся методов.

Источники и литература

- 1) Agarwal, V., Inoue, F., Schubach, M. et al. Massively parallel characterization of transcriptional regulatory elements. *Nature* (2025).
- 2) Avsec, Ž., Agarwal, V., Visentin, D. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 18, 1196–1203 (2021).
- 3) Dmitry Penzar, Daria Nogina et al., LegNet: a best-in-class deep learning model for short DNA regulatory regions, *Bioinformatics* (2023).

- 4) Kircher, M. et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat. Commun. 10, 3583 (2019).

Иллюстрации

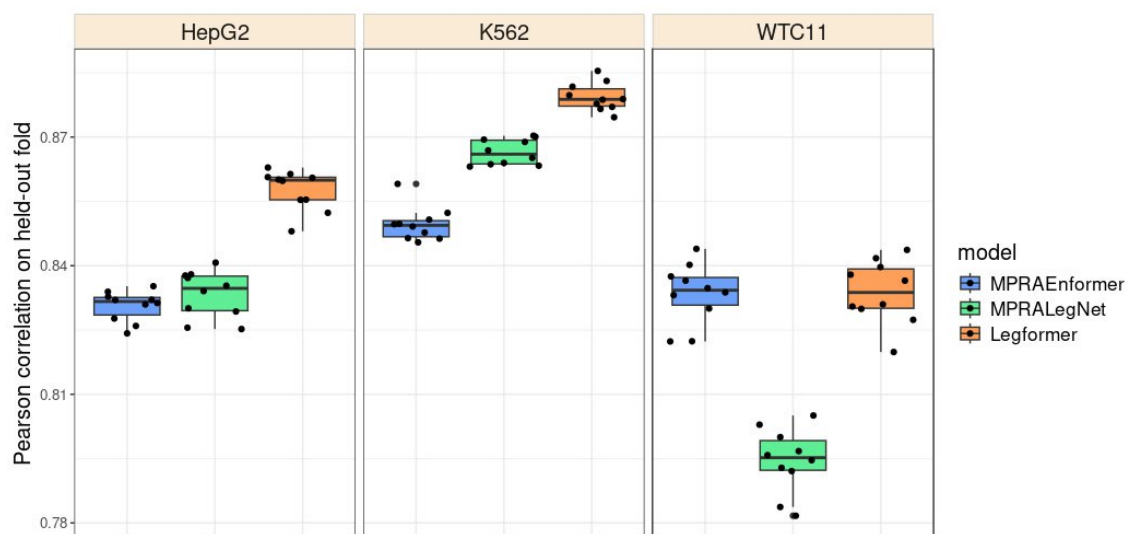


Рис. : Корреляция Пирсона, полученная для моделей Enformer, LegNet и Legformer (LegNet с добавлением признаков из модели Enformer) на тестовом фолде для трех клеточных линий.