Секция «2.2 Биоинформатика»

Предсказание эффектов регуляторных однонуклеотидных вариантов

Научный руководитель – Пензар Дмитрий Дмитриевич

Фоменко Елизавета Антоновна

Студент (специалист)

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

E-mail: lizzzafomenko@gmail.com

Значительная доля клинически значимых некодирующих однонуклеотидных вариантов (ОНВ), ассоциированных с различными заболеваниями, находится в регуляторных элементах генома - промоторах и энхансерах[1]. Предсказание влияния таких вариантов на транскрипционную активность генов является одной из ключевых задач вычислительной регуляторной геномики. Существующие подходы детекции и функциональной аннотации некодирующих вариантов являются неточными и слабо интерпретируемыми, в связи с чем разработка новых инструментов, которые позволят перейти к персонализированной медицине, является критически важной. Для решения этой задачи в данной работе используется архитектура нейронной сети Arinette, предсказывающая доступность участка хроматина и его активность по нуклеотидной последовательности. Качество модели оценивалось на наборе регуляторных ОНВ lentiMPRA (массовый параллельный анализ репортеров на основе лентивирусов) CAGI7, а также с помощью бенчмарка PromoterAI[2].

Первой задачей работы было достижение качества популярной модели ChromBPNet на задаче предсказания эффектов регуляторных вариантов. Arinette была обучена на данных DNase-Seq клеточной линии HepG2. В ходе работы были подобраны оптимальные гиперпараметры. Наилучшие результаты достигаются при использовании в качестве негативной выборки (районов генома, где активность хроматина отсутствует) случайных последовательностей генома, сбалансированных по GC-составу относительно положительных примеров, в соотношении 1:1, а также двухкомпонентной функции потерь, включающей: MSE (Mean Squared Error, средняя квадратичная ошибка), отвечающая за точность предсказания суммарного числа прочтений в окне, и мультиномиальный лосс, контролирующий форму их распределения. На данных CAGI7 коэффициент корреляции Пирсона составил 0.430 и 0.436 для ChromBPNet и Arinette соответственно, при этом размер нашей модели в полтора раза меньше конкурента.

Одной из тенденций в области вычислительной регуляторной геномики является построение multitrack-моделей[2][3], обучаемых одновременно на данных нескольких клеточных линий. Такой подход позволяет увеличить объем обучающей выборки и учитывать клеточную специфичность регуляторных механизмов, однако потенциально может снижать точность предсказаний для отдельных клеточных линий. В работе реализована multitrack-версия Arinette, обученная на данных DNase-Seq трех клеточных линий - HepG2, HEK293 и K562. Использование нескольких треков позволило повысить качество модели на бенчмарке PromoterAI без существенной потери качества на данных CAGI7 (Рис. 1).

Еще одной современной тенденцией является использование данных с однонуклеотидным разрешением. В работе исследовано, приводит ли обучение на таких данных к улучшению точности предсказания эффектов регуляторных ОНВ. Модель обучалась на данных DNase-Seq трех образцов (клеточной линии HepG2, клеток печени и гепатоцитов) в двух вариантах разрешения: в оригинальном (1bp) и сглаженном (20bp). Оказалось, что сглаживание профиля практически не снижает точность предсказания по сравнению с

1bp данными, при этом существенно упрощая их хранение и использование в виду их меньшего объема.

Таким образом, нами разработана новая архитектура нейронной сети Arinette с малым количеством параметров, выступающая на уровне современных моделей-конкурентов, а в ряде случаев превосходящая их.

Источники и литература

1) Farh, KH., Marson, A., Zhu, J. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015). https://doi.org/10.1038/nature13835
2) Kishore Jaganathan et al. Predicting expression-altering promoter mutations with deep learning. Science 389, eads7373(2025). DOI:10.1126/science.ads7373
3) Avsec, Ž., Latysheva, N., Cheng, J. et al. Advancing regulatory variant effect prediction with AlphaGenome. Nature 649, 1206–1218 (2026). https://doi.org/10.1038/s41586-025-10014-0

Иллюстрации

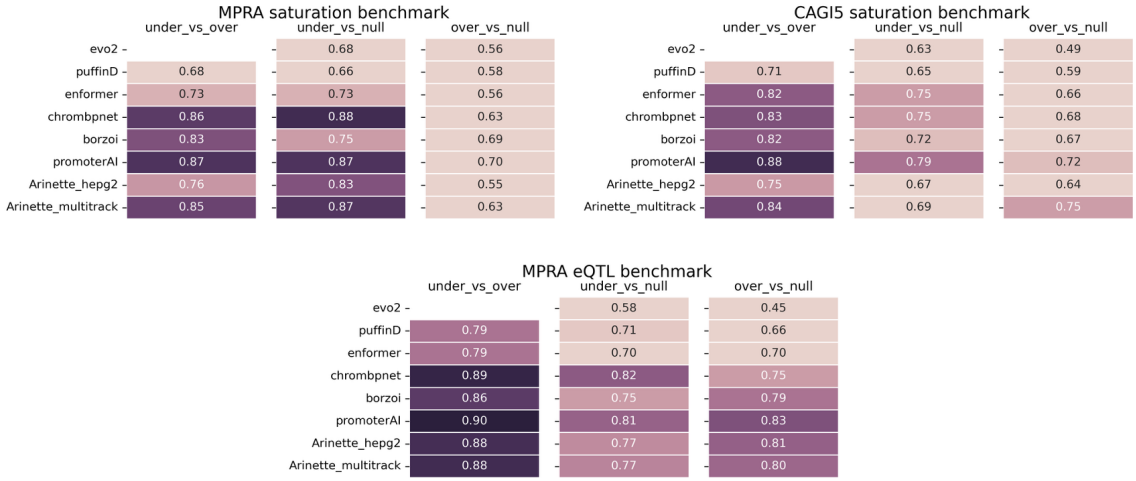

Рис. : Рис. 1 Сравнение качества предсказаний Arinette с другими популярными моделями машинного обучения на PromoterAI бенчмарке