Секция «Биоинформатика»

Улучшение предсказания 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

В работе [3] было проведено количественное определение активности всех потенциальных энхансеров в клеточных линиях HepG2, K562 и WTC11 при помощи лентивирусных векторов (lentivirus-based massively parallel reporter assays, lentiMPRA). Используя эти данные, можно обучить различные модели машинного обучения, для предсказания энхансерной активности. В оригинальной статье были использованы биохимические признаки — результаты различных полногеномных экспериментов (ChIP-seq, ATAC-seq, DNase-seq) и нуклеотидные последовательности энхансеров.

Для предсказания на основе биохимических признаков, в исходной статье была использована сравнительно простая модель – линейная регрессия. При этом сравнение данной модели проводится с куда более сложными моделями, основанными на нейронных сетях и их ансамблях. Для того, чтобы понять, какое качество достижимо только с помощью биохимических признаков, нами был использован градиентный бустинг, реализованный в библиотеке CatBoost, и полносвязная нейронная сеть. Оказалось, что эти модели работают лучше на 2 из 3 клеточных линиях – корреляция Пирсона составила 0.79 для линии K562 (вместо 0.72 с помощью регрессии), 0.77 для HepG2 (вместо 0.73) и 0.72 для WTC11 (вместо 0.71).

Следующим шагом работы является использование биохимических признаков совместно с последовательностью для построения нейросетевой модели на основе LegNet [2] и сравнение качество полученной модели с моделью, предсказывающую активность энхансера исключительно на последовательности.

После этого подобное же сравнение будет проведено между моделью, обученной только на последовательности и моделью, обученной на последовательности и признаках из модели Enformer [1].

Данная работа поможет ответить на вопрос – достаточно ли данных представленных в работе для выучивания всех регуляторных закономерностей в данных с нуля или же биохимические признаки и/или признаки, выученные нейронной сетью Enformer на внешних данных все еще необходимы для достижения оптимального качества.

Источники и литература

1) 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). https://doi.org/10.1038/s41592-021-01252-x
2) Dmitry Penzar, Daria Nogina et al., LegNet: a best-in-class deep learning model for short DNA regulatory regions, Bioinformatics, 2023; doi: 10.1093/bioinformatics/btad457
3) Vikram Agarwal, Fumitaka Inoue, Max Schubach, Beth K. Martin, Pyaree Mohan Dash, Zicong Zhang, Ajuni Sohota, William Stafford Noble, Galip Gürkan Yardimci, Martin Kircher, Jay Shendure, Nadav Ahituv, Massively parallel characterization

of transcriptional regulatory elements in three diverse human cell types bioRxiv 2023.03.05.531189; doi: 10.1101/2023.03.05.531189