Секция «Биоинформатика»

Количественное предсказание доступности хроматина в клеточных линиях по индивидуальному диплоидному геному методами глубокого обучения.

Латорцева Дарья Дмитриевна
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: latortsevad@gmail.com

Экспрессию генов в клетке регулируют некодирующие участки ДНК, которые могут находиться на большом расстоянии от контролируемого гена. Также экспрессия в различных клеточных линиях может сильно отличаться. Эти и другие факторы осложняют решение вопроса о регуляции генов, являющегося важной проблемой современной генетики. Методы глубокого обучения все чаще успешно используются для прогнозирования регуляторных участков по последовательности ДНК.

DNase-Seq позволяет идентифицировать области открытого хроматина на ДНК посредством секвенирования областей, чувствительных к расщеплению ДНКазой I. Полученные в ходе секвенирования риды картируются на геном, в результате чего получается трек покрытия ими всего генома. Доступная база данных проекта ENCODE, одной из целью которого является картирование гиперчувствительных участков ДНК на гаплоидный референсный геном, является одним из основных источников данных для обучения моделей по предсказанию открытости хроматина.

Интерес представляет не только определение того, открыт или закрыт данный участок ДНК, но также и количественное предсказание его доступности. Решение подобной задачи было предложено авторами моделей Basenji [1] и Enformer [2], также обучавших нейросети на данных ENCODE.

Нашей лаборатории были предоставлены данные о диплоидных наборах хромосом в различных клеточных линиях для нескольких индивидуумов. Была сформулирована гипотеза о том, что информация со второй хромосомы может увеличивать качество предсказания нейронной сети. В ходе работы планируется сравнить качества предсказаний нейронных сетей, обученных на гаплоидном и диплоидном наборах хромосом. Также особенность данных позволяет использовать не референсный геном, а индивидуальные геномы, что также может повлиять на качество.

Были выбраны готовые препроцессированные данные, на которых мы применили несколько архитектур нейронных сетей для решения задач бинарной классификации клеточных линий. На данный момент работа ведется на этапе препроцессинга данных, за основу выбран подход, описанный авторами Basenji.

Источники и литература

1 - Kelley, David R et al. “Sequential regulatory activity prediction across chromosomes with convolutional neural networks.” Genome research vol. 28,5 (2018): 739-750. doi:10.1101/gr.227819.117

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). https://doi.org/10.1038/s41592-021-01252-x

3 - The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.