Секция «Биоинформатика»

Создание модели с использованием нейросетей для предсказания связывания факторов транскрипции с олигонуклеотидами, полученными из экспериментов SMiLE-seq

Машкова София Дмитриевна
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: maskovasofa9@gmail.com

В настоящее время быстрыми темпами развиваются экспериментальные методы, которые позволяют с высокой точностью идентифицировать регуляторные элементы генома, в том числе участки связывания транскрипционных факторов. Одним из них является технология SMiLE-seq (селективное обогащение лигандов на микрогидродинамической основе с последующим секвенированием) [1], с помощью которой можно надежно детектировать и выделять взаимодействующие комплексы ДНК с транскрипционными факторами. Несмотря на накопленный массив данных, остается проблема в понимании тонких механизмов регуляции конкретных генов, то есть как определенная последовательность влияет на связывание и сборку белковых комплексов. Создание модели связывания факторов с последовательностями позволит пролить свет на понимание сложных регуляторных механизмов, а также связанных с ними болезней и проявлений других фенотипических признаков.

Наша задача - построить модель бинарной классификации последовательностей с использованием архитектур на основе сверточных нейронных сетей.

Структура данных

Положительный класс в данных представлен последовательностями длиной 30 нуклеотидов для 45 транскрипционных факторов. Отрицательный класс был получен перемешиванием исходных последовательностей с сохранением частот динуклеотидов. Полученные данные кодировались методом one-hot encoding.

Промежуточные результаты

Для референсных PWM-моделей мотивов транскрипционных факторов были получены значения AUROC (площадь под ROC-кривой), относительно которых будет оцениваться качество нейросетей. Для подсчета AUROC использовались выборки из 10000 последовательностей, которые подавались на вход программе для анализа мотивов ChiPMunk.

Первая модель нейронной сети имеет простую архитектуру с шестью сверточными слоями. С помощью нее удалось получить качество, сопоставимое с референсными моделями PWM. На следующих этапах мы планируем изменить метод обучения модели и ее архитектуру для повышения точности классификации.

Список используемой литературы

[1] Isakova A. et al. SMiLE-seq identifies binding motifs of single and dimeric transcription factors //Nature methods. - 2017. - Т. 14. - №. 3. - С. 316-322.