УДК 575.112

Использование сверточных нейросетей для детекции и сегментации VJ фрагментов

Беляков М.Д.[1] (бакалавр)

Научный руководитель - кандидат биологических наук, Шугай М.А.[3,4]**, аспирант, Власова Е.К.**[2,4]

[1] *СПБГУ*

[2] *Университет ИТМО*

[3] *Институт Биоорганической химии РАН*

[4] *Российский Национальный Исследовательский Медицинский Университет*

belyakov06mat@mail.ru

Данный проект направлен на исследование возможности детекции и сегментации VJ-фрагментов в сырые последовательностях ДНК. Наш метод основан на использовании архитектуры сверточного автоэнкодера, который применяется на эмбеддингах последовательности нуклеотидов (с возможным включением N).

Введение.

Проблема поиска VJ-фрагментов в последовательностях имеет множество решений. IgBLAST[1] способен с высокой точностью аннотировать последовательности. Он считает большое количество метрик, но скорость обработки последовательностей остается крайне низкой. Из-за алгоритма выравнивания и необходимости поиска по базе данных обработка сэмплов может занять несколько десятков часов. При таком подходе становится актуальным наличие тула, который мог бы отсеять большое количество лишних последовательностей, оставив лишь те, которые с большей вероятностью имеют в себе VJ-фрагменты. Одним из таких тулов является Vidjil[2], который работает на основе k-меров. Он работает быстрее IgBLAST, но все равно достаточно долго. Также его разметка в большинстве своем неудобная для чтения и анализа.

Основная часть.

Во время работы над проектом нами был разработан метод, который решает проблему поиска VJ-фрагментов с помощью нейронных сетей. Разработанная нами сверточная нейронная сеть состоит из энкодера и декодера и, по своей сути, является адаптацией модели для анализа изображений - UNet. Для тренировки модели мы брали данные, обработанные IgBlast. В выходном AIRR-formatted файле находились все необходимые нам данные для обучения - координаты начала и концов индексов в последовательностях. Сырые последовательности ДНК мы преобразовали в эмбеддинги. Нуклеотиды кодировались числами от 0 до 5 (включая N - 4, и специальный нуклеотид для паддинга P - 5). Полученные эмбеддинги мы использовали для тренировки нейронной сети. Наша сеть состоит из трех блоков: энкодера, bottleneck, и декодера. Энкодер состоит из 4 блоков, каждый из которых содержит по 2 сверточных слоя с слоями BatchNorm и Dropout. После энкодера находится блок bottleneck, который создает латентное пространство для извлечения признаков. Далее идет блок декодера, который схож с блоком энкодера, но информация с соответствующих блоков энкодера пробрасывается в декодер по принципу Residual Connections. Таким образом информация энкодера у нас сохраняется и учитывается

при декодировании. В конце расположена финальная свертка и разделение на 2 головы. Каждая голова состоит из двух финальных сверточных блоков, которые преобразуют последовательность в выходную, равную по длине входной. Каждая голова учится отдельно определять V или J фрагменты. На выходе 2 последовательности - для V и J. Длина каждой последовательности равна исходной длине и состоит из чисел от 0 до 1, и чем выше число, тем вероятнее данный нуклеотид принадлежит к V или J фрагменту. Полученная модель имеет значительно более высокий recall, чем Vidjil, при этом скорость обработки более, чем в 2 раза выше, чем у последнего.

Выводы.

Таким образом наша модель демонстрирует возможность использования нейросетевых моделей в детекции и сегментаций последовательностей ДНК для нахождения V и J фрагментов. Более того, полученные результаты демонстрируют возможность расширения нашей модели для детекции более специфических регионов, таких как CDR и FWR фрагменты.

Литература

1. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013 Jul;doi: 10.1093/nar/gkt382.

2. Duez M, Giraud M, Herbert R, Rocher T, Salson M, Thonier F. Vidjil: A Web Platform for Analysis of High-Throughput Repertoire Sequencing. PLoS One. 2016 Nov; doi: 10.1371/journal.pone.0166126