Секция «Биоинформатика»

Предсказание областей связывания транскрипционных факторов с помощью моделей машинного обучения

Научный руководитель – Пензар Дмитрий Дмитриевич

Грызунов Никита Сергеевич

Студент (специалист)

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

E-mail: NikGR1@yandex.ru

Взаимодействие транскрипционных факторов с хроматином регулирует процесс транскрипции, и как следствие является одним из определяющих факторов жизнедеятельности клетки.

В настоящее время существует множество экспериментальных подходов для выявления сайтов связывания ТФ, например, ChIP-seq и genomic HT-SELEX. Эти методы учитывают множество различных факторов: доступность хроматина, конкуренция между различными ТФ, смещения в данных, характерные для разных подходов и т.д.

В случае данных ChIP-Seq связывание определяется не только мотивом самого фактора, но и мотивами кофакторов. В случае же GHT-SELEX влияние кофакторов нивелировано, но возникают паразитные сигналы из-за добавленных фланкирующих последовательностей. Поэтому данная работа посвящена изучению переносимости между данными ChIP-seq и genomic HT-SELEX.

Экспериментальные данные двух экспериментов были предоставлены консорциумом Codebook, после чего были получены пики (длиной 301 нт) из экспериментальных профилей (положительный класс), а также сгенерированы некоторые последовательности, которые не были получены из экспериментов и выступали в роли отрицательного класса.

Использовалось три типа негативных примеров:

1) foreigns — пики, принадлежащие другим ТФ,

2) random — последовательности, полученные случайным образом из генома человека hg38,

3) shades — последовательности, полученные при отступе от пиков положительного класса.

Соотношение количества последовательностей положительного и отрицательного классов — 1:100. При этом осуществлялся контроль, чтобы последовательности двух классов не пересекались, а также имели одинаковый GC-состав.

После этого было проведено обучение конволюционной модели LegNet на данных ChIP-seq и тестирование на данных HT-SELEX, и наоборот. Кроме того, было проведено обучение с инициализацией и «заморозкой» первого блока модели позиционными матрицами весов (соответствующего типа экспериментальных данных для обучения) из GRECOBIT/Codebook Motif Explorer, что приводило к улучшению переносимости моделей в случае нехватки последовательностей положительного класса.

Обученные модели показали переносимость классификации между двумя экспериментальными подходами, сравнимую с другими методами, в частности, ArChIPelago [unpublished], основанном на случайных лесах. Кроме того, обученные модели имели, в большинстве случаев, лучшее качество классификации, чем PWM. При сравнении моделей LegNet и ArChIPelago оказалось, что в случае переноса классификации из ChIP-seq в genomic HT-SELEX лучшее качество классификации имеет модель ArChIPelago, а при переносе из genomic HT-SELEX в ChIP-seq — LegNet с инициализацией весами PWM (рис. 1).

Иллюстрации

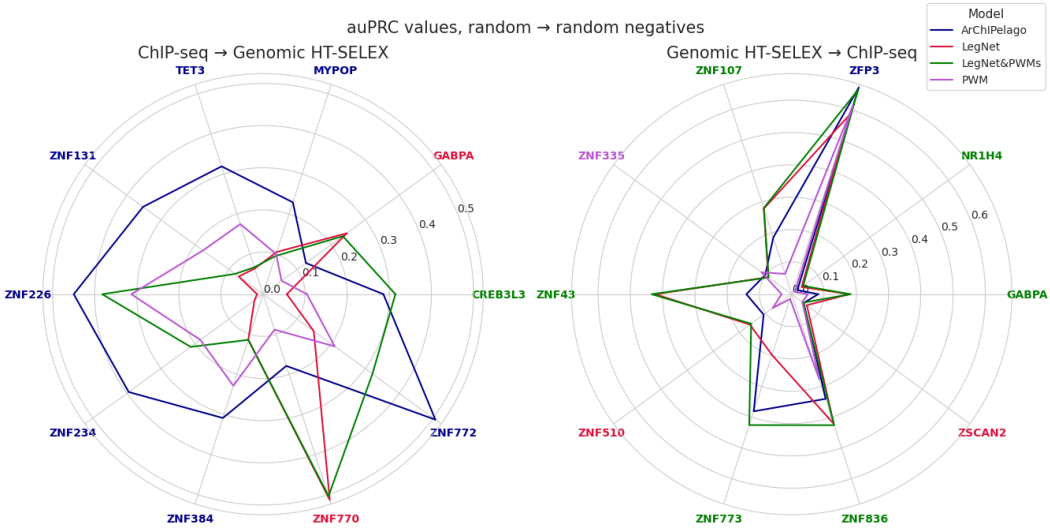

Рис. : Сравнение значений метрики auPRC моделей PWM, ArChIPelago и LegNet при переносе классификации между двумя экспериментальными подходами по выявлению областей связывания ТФ.