Секция «Биоинформатика»

Детекция сайтов связывания транскрипционных факторов на основе результатов ChIP-Seq при помощи свёрточных нейросетей

Гуков Борис Сергеевич
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: jakewayd.ru@gmail.com

Одним из процессов, лежащих в основе реализации генетических инструкций в клетке, является транскрипция - перенос информации из ДНК в РНК. Во многих случаях транскрипция регулируется специальными белками - транскрипционными факторами (ТФ). С помощью ДНК-связывающих доменов (DNA-binding domain, DBD) ТФ распознают специфические последовательности ДНК (сайты связывания ТФ), образуя систему, которая управляет экспрессией генома, следствием чего является, например, дифференцировка тканей. При нарушении взаимодействий между ТФ и ДНК могут развиваться различные заболевания. Причиной нарушений этих взаимодействий может быть мутация в сайте связывания ТФ.

Для определения сайтов связывания используют метод ChIP-Seq[1] - иммунопреципитацию хроматина с последующим секвенированием. Данный метод позволяет получить последовательности сайтов связывания для конкретных ТФ по всему геному. Тем самым появляется возможность исследовать влияние отдельных замен внутри сайтов.

Данные ChIP-Seq можно использовать для обучения нейросетевой модели, которая могла бы для конкретного фактора транскрипции выдавать вероятность его связывания с произвольной последовательностью.

Для построения архитектур нейросетевых моделей и их обучения использовался библиотека PyTorch языка программирования Python.

Все последовательности полученных данных содержат внутри себя тот или иной сайт связывания того или иного ТФ, то есть принадлежать положительному классу. Для создания негативного класса в работе использовались 3 различных метода получения негативного класса: перемешивание, выбор последовательностей от других ТФ и последовательности из не принадлежащих пику регионов. Метод перемешивания заключается в следующем: для каждой последовательности каждого ТФ негативной последовательностью называется перемешанная исходная последовательность. Метод выбора последовательностей от других ТФ заключается в том, чтобы для каждой последовательности положительного класса случайным образом выбрать последовательность, которую не связывает данный ТФ. Третий метод заключается в использовании координат пика при картировании ридов на геном. Координата как бы смещается на N позиций, далее происходит вырезка последовательности негативного класса из генома.

В качестве baseline использовались PWM для каждого из ТФ. По данному методу предсказанию класса последовательности была рассчитана метрика ROC-AUC для каждого ТФ.

В рамках работы над архитектурой модели были протестированы различные варианты сверточных нейросетей. Было протестировано множество моделей. Одной из основных моделей является усложненная и адаптированная под задачу версия архитектуры ResNet[2]. Модель показала заметное увеличение качества классификации по сравнению с PWM (рисунок 1).

Источники и литература

1) Park, P. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet 10, 669–680 (2009). https://doi.org/10.1038/nrg2641
2) Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition", arXiv:1512.03385 [cs.CV], Dec. 2015.

Иллюстрации

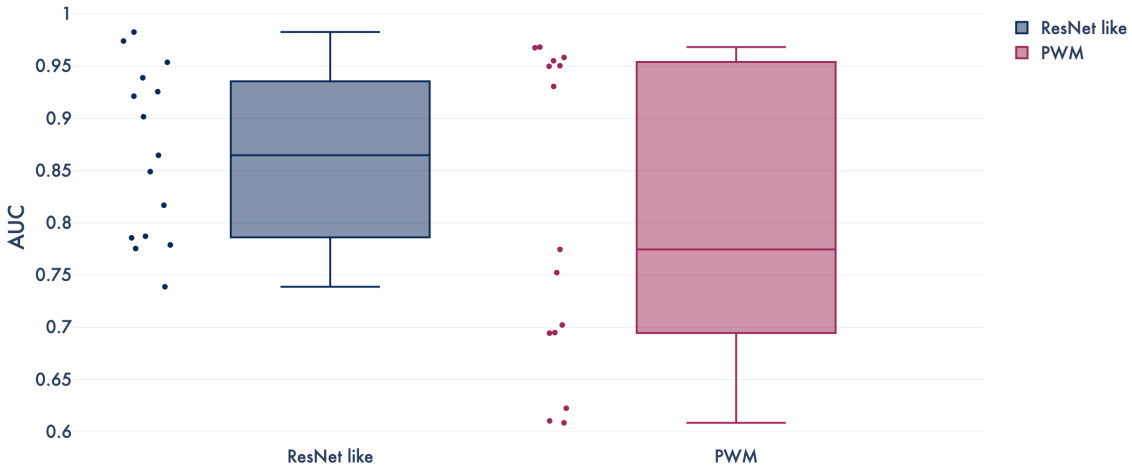

Рис. : Сравнение распределений метрики качества ROC-AUC для реализованной модели ResNet-like и PWM.