

Применение ДНК-языковых моделей для предсказания связывания транскрипционных факторов

Научный руководитель – Пензар Дмитрий Дмитриевич

Стрекаловских Вадим Владимирович

Студент (специалист)

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

E-mail: vadimstrek@yandex.ru

IBIS Challenge, разработанный в нашей лаборатории, предоставляет уникальную возможность сравнить различные методы предсказания сайтов связывания транскрипционных факторов на широком спектре данных, полученных различными методами (от ChIP-Seq и до SELEX).

Одним из популярных подходов для моделирования регуляторной грамматики в наши дни является обучение языковых моделей на геномной ДНК с последующим дообучением на целевой задаче. Поскольку участники конкурса не предложили решений, основанных на ДНК-языковых моделях, мы поставили задачу применить модели этого типа к данным IBIS Challenge.

Цель: оценка эффективности ДНК-языковых моделей в задаче предсказания связывания ТФ, сравнение их производительности на данных IBIS друг с другом и с конкурсными моделями.

Задача: предсказание мест связывания транскрипционных факторов с использованием предобученных моделей GENA-LM, Nucleotide Transformer и DNABERT-2 на данных IBIS Challenge.

Методология:

1. Подготовка данных: формирование датасетов положительных (сайты связывания ТФ) и отрицательных (случайные последовательности) примеров для каждого ТФ и каждого эксперимента IBIS. Обеспечение баланса классов.

2. Fine-tuning: дообучение каждой из трех выбранных моделей (GENA-LM, Nucleotide Transformer, DNABERT-2) для каждого ТФ на данных каждого эксперимента. Использование стратегии обучения с учителем.

3. Тестирование: получение предсказаний каждой модели на данных о связывании тех же факторов в экспериментах gHTS и CHS.

4. Подведение итогов: оценивание, ранжирование в рейтинге конкурсных моделей.

IBIS Challenge предполагает обнаружение мотивов сайтов связывания ТФ на основе данных 5 разных экспериментов (CHS, gHTS, PBM, HTS, SmileSeq) in vitro и in vivo с 40 транскрипционными факторами.

Тестирование дообученных моделей проводилось на датасетах, полученных естественным путём (эксперименты genomic HT-SELEX и ChIP-Seq) в рамках конкурсной дисциплины artificial-to-genomic.

Результаты тестирования и ранжирования показали, что дообученные ДНК-языковые модели уступают в качестве предсказаний другим моделям, основанным на рекуррентных и свёрточных нейронных сетях, обученных с нуля.

Иллюстрации

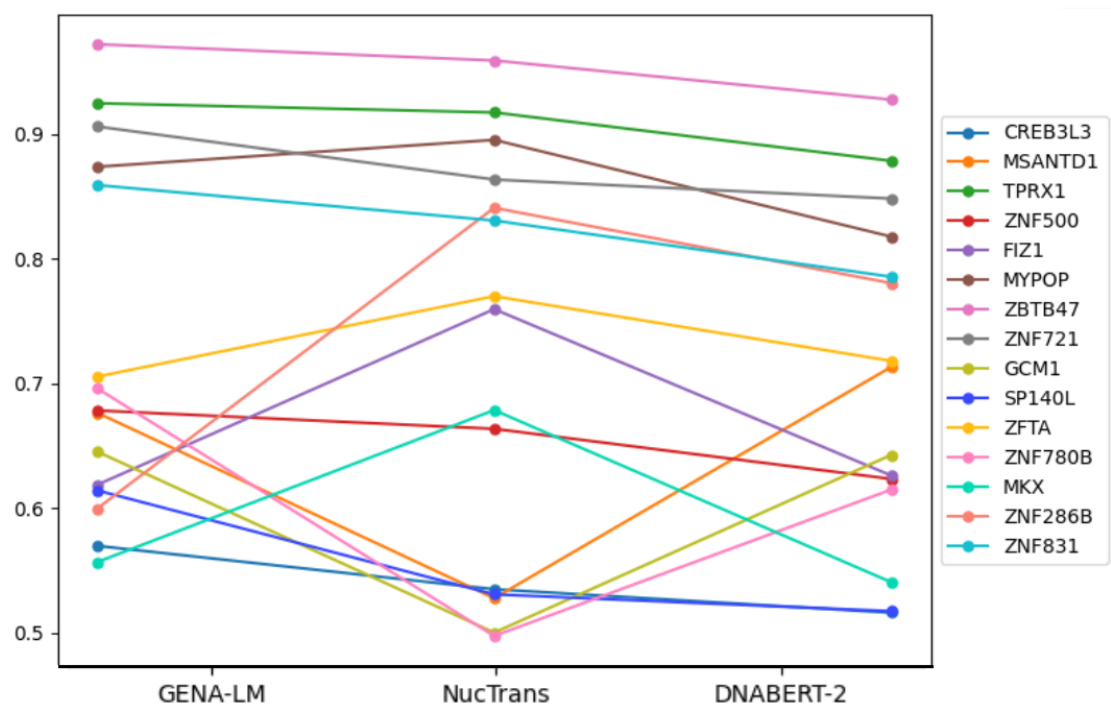


Рис. : Распределение значений AUROC-score по результатам валидации трёх моделей.