Глубинное обучение для выявления ассоциаций между функциональными геномными элементами

А. Д. Цветкова,

Международная лаборатория биоинформатики, НИУ ВШЭ, a.tsvetkova@hse.ru

М. С. Попцова,

Международная лаборатория биоинформатики, НИУ ВШЭ, mpoptsova@hse.ru

Аннотация

Архитектура глубинного обучения, трансформер, отличается механизмом внимания, который позволяет проинтерпретировать вывод моделей. DNABERT – одна из таких моделей, она обучена на геноме человека, то есть в ней уже заложена информация о некоторых взаимосвязях в ДНК. Модель можно настроить для других задач путем дообучения на небольшом датасете. В ходе работы была поставлена цель – проинтерпретировать мотивы, выявленные при помощи внимания трансформера, обученного распознавать гистоновые метки (H2A.Z, H4ac, H2A.XS139ph, H3K18ac, H3ac, H3K9ac, H3K9K14ac) и фактор транскрипции AIRE. Данные геномные элементы были выбраны, как ассоциирующиеся с Z-ДНК. Их более подробное изучение может привести к лучшему пониманию связи эпигенетических факторов и вторичных структур ДНК. Модель DNABERT дообучается на полногеномных данных ChIP-seq изучаемых объектов. Среднее значение AUC для распознавания гистоновых меток составило 0,93, для AIRE – 0,86. Для каждой модели были визуализированы карты внимания и определены регионы повышенного внимания, что соответствовало статически значимым участкам связывания с известными факторами транскрипции, которые локализуются с исследуемыми объектами. На данных CTCF было доказано, что модель DNABERT может использоваться для поиска транскрипционных факторов-партнеров. Для 5 геномных элементов удалось валидировать некоторые из найденных факторов транскрипции с помощью опубликованных исследований. В целом такой метод выявления мотивов может применяться для транскрипционных факторов, для которых мотивы неизвестны.

Ключевые слова: DNABERT, глубинное обучение, интерпретируемость моделей