УДК 571.27, 004.93, 004.8

ПРЕДСТАВЛЕНИЕ И АНАЛИЗ ЭМБЕДДИНГОВ Т-КЛЕТОЧНЫХ РЕЦЕПТОРОВ В ЛАТЕНТНЫХ ПРОСТРАНСТВАХ МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ

Кабалина А.А. (Санкт-Петербургский государственный университет)
Научный руководитель – преподаватель факультета информационных технологий и программирования Власова Е.К.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО», Федеральное государственное автономное образовательное учреждение высшего образования "Российский национальный исследовательский медицинский университет имени Н.И. Пирогова" Министерства здравоохранения Российской Федерации);
Научный руководитель – кандидат биологических наук, Шугай М. А.
(Федеральное государственное автономное образовательное учреждение высшего образования "Российский национальный исследовательский медицинский университет имени Н.И. Пирогова" Министерства здравоохранения Российской Федерации)

Введение. Т-клеточные рецепторы (TCRs) - важнейшие компоненты адаптивной иммунной системы, отвечающие за распознавание и связывание со специфическими антигенами. Высокая вариабельность последовательностей TCR и низкий уровень перекрытия иммунных репертуаров между индивидуумами осложняют их классификацию, предсказание связывания с антигенами и поиск совпадений интересующих нас TCR в существующих базах данных [1].

Основная часть. Подходы на основе машинного обучения позволяют анализировать разнообразие TCR. Использование латентных пространств в моделях глубокого обучения позволяет находить скрытые закономерности, выявлять биологически значимые паттерны и классифицировать рецепторы по их функциональным свойствам.
В рамках данного проекта была разработана генеративно-состязательная сеть (GAN). Архитектура модели включает два основных компонента: генератор и дискриминатор, которые совместно обучаются в процессе состязательной оптимизации [2].
Генератор отвечает за процесс генерации новых последовательностей. В процессе обучения генератор стремится создавать последовательности, наименее отличимые от реальных, чтобы повысить вероятность отнесения данной последовательности к естественным последовательностям.
Дискриминатор оценивает вероятность того, что данный сиквенс встречается в организме. Он обучается различать истинные последовательности от сгенерированных, передавая значение ошибки обратно к генератору для улучшения качества генерируемых последовательностей.
При совместном обучении формируется состязательный процесс, в результате которого генератор улучшает качество воспроизводимых последовательностей, а дискриминатор учится более точно их оценивать.
Модель была обучена на датасете из 15000 последовательностей TCR длинной 15 аминокислотных остатков.
Рассмотренный подход также позволяет проводить анализ TCR в латентном пространстве. Данный подход может быть использован для группировки рецепторов с похожими свойствами и выявления сложных закономерностей для предсказания взаимодействия с эпитопами антигенов

Выводы. Разработанная модель позволяет генерировать новые последовательности TCR, потенциально встречающиеся в организме. Это может быть использовано для

расширения существующих баз данных. Представление TCR в латентном пространстве позволяет выявлять структурные закономерности и формировать биологически осмысленные эмбеддинги, что открывает перспективы для более глубокого анализа иммунного репертуара.

Список использованных источников:

1. Elhanati, Yuval & Murugan, Anand & Jr, Curtis & Mora, Thierry & Walczak, Aleksandra. (2014). Quantifying selection in immune receptor repertoires. Proceedings of the National Academy of Sciences of the United States of America. 111. 10.1073/pnas.1409572111

2. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. Commun. ACM 63, 11 (November 2020), 139–144. https://doi.org/10.1145/3422622