

Использование LLM для поиска распространяющихся микроорганизмов в метагеномных данных

Н.С. Попов^{1,2*}, В.В. Панова¹, А.Н. Лукашев², А.И. Манолов¹

¹ *НИИ Системной биологии и медицины Роспотребнадзора, Москва, Россия*

² *Институт медицинской паразитологии, тропических и трансмиссивных заболеваний им. Е.И. Марциновского, МГМУ им. И.М. Сеченова, Москва, Россия*

* *nikoraisp@gmail.com*

Ключевые слова: Болезнь X; метагеномика; LLM

Application of LLMs for detecting spreading microorganisms in metagenomic data

N.S. Popov^{1,2*}, V.V. Panova¹, A.N. Lukashev², A.I. Manolov¹

¹ *Research Institute for Systems Biology and Medicine (RISBM) of Rosпотребнадзор, Moscow, Russia*

² *Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov First Moscow State Medical University, Moscow, Russia*

* *nikoraisp@gmail.com*

Key words: Disease X; metagenomics; LLM

«Болезнь X» – термин ВОЗ, введенный в 2018 году для обозначения потенциального неизвестного патогена, способного вызвать пандемию. Такой патоген может иметь скрытое течение и не иметь значимой гомологии с известными последовательностями. Языковые модели, обученные на нуклеотидных последовательностях, могут выявлять функциональные и структурные закономерности геномов без выравнивания с известными последовательностями, что особенно полезно для анализа изменчивых или новых данных. В этом исследовании мы применили большие языковые модели (LLM) для мониторинга изменений представленности последовательностей в метагеномных данных до и после начала пандемии COVID-19. Из базы NCBI отобраны метагеномы: препандемический набор до 2019 года и постпандемический с 2020 года, по 367 SRR-запусков в каждом. Нуклеотидные последовательности преобразуются в эмбединги моделью ViraLM – из каждого из 12 слоев трансформера извлекаются векторы и усредняются по токенам. Затем данные обрабатываются методом главных компонент (PCA), после чего их распределение анализируется с помощью разбиения пространства на ячейки и сравнения их заполненности ячеек с помощью теста Колмогорова–Смирнова для выявления различий между двумя датасетами. Метод успешно обнаружил присутствие SARS-CoV-2 в постпандемическом наборе с чувствительностью до 70%.

Финансирование: Исследование поддержано государственным заданием «Разработка алгоритмов выявления новых уникальных последовательностей ДНК или РНК в метагеномах и их фенотипических характеристик *in vitro*», номер 122030900069-4.