

**Разработка метода поиска распространяющихся микроорганизмов,
основанного на анализе представленности k-меров в данных метагеномного
секвенирования.**

Научный руководитель – Манолов Александр Иванович

Гуров Сергей Анатольевич

Студент (бакалавр)

Московский физико-технический институт, Москва, Россия

E-mail: gurov3@icloud.com

Распространение новых инфекций обусловлено деятельностью человека, включая его взаимодействие с дикой природой и циркулирующими вирусами. В данной работе предлагается метод обнаружения ранее неизвестных микроорганизмов на основе анализа динамики k-меров в метагеномных данных секвенирования. K-мер – это участок геномного текста длиной k. Предложенный подход использует динамику отдельных k-меров для достижения поставленной цели.

Для проверки данного подхода были взяты метагеномные данные секвенирования тотальной РНК назофарингеальных мазков из базы данных SRA. Исследование включало по 330 образцов, собранных до и после 2020 года (до и после пандемии COVID-19).

В рамках данной работы важной задачей для решения была проблема поиска оптимальной схемы для хранения данных. Для решения данной задачи была предложена идея хранения последовательности k-мера в индексе элемента массива, а значение элемента массива по этому индексу представляет собой логарифм представленности k-мера. Для хранения данных о распространенности k-меров также создается массив.

Предпринятые меры по оптимизации данных привели к двум ключевым улучшениям: снижению требований к памяти для хранения и существенному ускорению обработки. Ранее, обработка 15-меров занимала до 1,5 недель, а 17-меры не поддавались обработке. В настоящее время, те же этапы обработки занимают не более 12 часов.

Благодаря полученным данным о представленности и распространенности отдельных k-меров можно решить задачу фильтрации и идентификации мало представленных и мало распространённых k-меров, подбирая наиболее оптимальные параметры для обработки. Подбор данных параметров производился на основе количества k-меров соотнесенных с геномом вируса SARS-CoV-2 после фильтрации.

Анализ выявил 16 569 035 k-меров, уникальных для образцов, полученных после 2020 года. Из них 29 855 k-меров принадлежат новому коронавирусу.

Таким образом, разработанный нами подход позволяет идентифицировать геномные фрагменты нового, широко распространившегося вируса. Основной целью дальнейших исследований станет повышение точности и избирательности метода.

Источники и литература

- 1) C. Moeckel, M. Moreboina [et al.]. A survey of k-mer methods and applications in bioinformatics // Comput. Struct. Biotechnol. J. 2024 V. 23 P. 2289-2303
- 2) D. E. Wood, J. Lu, B& Langmead Improved metagenomic analysis with Kraken 2 // Genome Biol. 2019 V. 20: 257
- 3) Operating manual for Kraken 2 // Oregon State University : [Электронный портал]. URL: <https://software.cqls.oregonstate.edu/updates/docs/kraken2/MANUAL.html> (дата обращения: 24.02.2025)