НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2025 Том 25 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
Jule–August 2025 Vol. 25 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-4-718-726
УДК 004.02

Метод сравнительного анализа временных серий наборов данных, заданных в виде множества строк, с использованием графов де Брейна

Артем Борисович Иванов[1]✉, Анатолий Абрамович Шалыто[2], Владимир Игоревич Ульянцев[3]

[1] Федеральный научно-клинический центр физико-химической медицины им. академика Ю. М. Лопухина Федерального медико-биологического агентства, Москва, 119435, Российская Федерация
[1,2,3] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] abivanov@itmo.ru✉, https://orcid.org/0000-0002-7997-0637
[2] anatoly.shalyto@gmail.com, https://orcid.org/0000-0002-2723-2077
[3] ulyantsev@itmo.ru, https://orcid.org/0000-0003-0802-830X

Аннотация

Введение. Рассмотрена задача сравнительного анализа временных серий образцов в наборах данных, заданных в виде множества строк. Предложен способ повышения точности определения различий между двумя образцами. На его основе разработан метод анализа временных серий из трех образцов, позволяющий повысить точность классификации изменений между образцами. Использование в анализе трех образцов обусловлено спецификой решаемой практической задачи обработки данных секвенирования метагеномных образцов, получение большего числа которых является очень ресурсоемким. **Методы.** Для классификации строк из одного образца на обнаруженные и необнаруженные строки в другом образце применяется метод сравнения двух образцов с использованием *k*-меров и графа де Брейна. В нем реализованы решающие правила, основывающиеся на статистиках частоты встречаемости *k*-меров, разных значениях параметра *k* и информации о возможных ошибках в строках. Для анализа временных серий из трех образцов (исходного и итогового образца для одного объекта и модифицирующего образца для другого объекта) разработан метод, на основе попарного сравнения образцов. Он применяется для разбиения строк каждого из образцов на группы в зависимости от обнаружения строк в других образцах. **Основные результаты.** Разработанный метод анализа временных серий протестирован на двух типах сгенерированных метагеномных данных, заданных в виде множества строк. Показано, что метод позволяет различать геномы организмов, имеющие отличия хотя бы в одном символе на каждые 10 000 символов. Показана высокая (больше 80 %) точность и полнота результатов классификации строк при анализе моделированных сложных данных большого объема, сопоставимых с реальными данными. **Обсуждение.** Предложенный метод позволяет сравнивать метагеномные образцы, заданные в виде множества строк, используя только сами данные и не требуя дополнительной информации. Это дает возможность осуществлять более точный анализ по сравнению с существующими методами, которые сравнивают образцы на основе результатов классификации строк в базах данных таксономической аннотации. Представленные методы могут найти применение в различных областях обработки строковых данных, например, для анализа изменений авторского стиля при написании серии текстов.

Ключевые слова

временные серии, граф де Брейна, *k*-меры, сравнительный анализ, классификация строк, метагеномные образцы

Ссылка для цитирования: Иванов А.Б., Шалыто А.А., Ульянцев В.И. Метод сравнительного анализа временных серий наборов данных, заданных в виде множества строк, с использованием графов де Брейна // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 4. С. 718–726. doi: 10.17586/2226-1494-2025-25-4-718-726

Comparative analysis method for time series data objects represented as sets of strings based on de Bruijn graphs

Artem B. Ivanov[1]✉, Anatoly A. Shalyto[2], Vladimir I. Ulyantsev[3]

[1] Lopukhin Federal Research and Clinical Center of Physical-Chemical Medicine of Federal Medical Biological Agency (FRCC PCM), Moscow, 119435, Russian Federation
[1,2,3] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] abivanov@itmo.ru✉, https://orcid.org/0000-0002-7997-0637
[2] anatoly.shalyto@gmail.com, https://orcid.org/0000-0002-2723-2077
[3] ulyantsev@itmo.ru, https://orcid.org/0000-0003-0802-830X

Abstract

The paper considers the comparative analysis of string datasets represented as time series of samples. We propose a method to increase the accuracy of determining differences between two samples. Based on this method, a method for analyzing time series of three samples has been developed, which allows for more accurate changes investigation between samples. The use of three samples in the analysis is due to the specific nature of the practical task of processing metagenomic sample sequencing data, obtaining a larger number of which is very resource-intensive. To classify strings from one sample into detected and undetected strings in another sample, a method of comparing two samples using *k*-mers and the de Bruijn graph is proposed. It implements decision rules based on statistics of the frequency of *k*-mers occurrence, different values of the parameter *k*, and information about possible errors in the strings. To analyze time series of three samples (the original and final sample for one object and the modifying sample for another object), a method based on pairwise comparison of samples is developed. It is used to divide the strings of each sample into groups depending on the detection of strings in other samples. The developed method for analyzing time series has been tested on two types of generated metagenomic data, represented as a set of strings. It was shown that the method allows distinguishing organisms that have differences in genomes in at least one symbol for every 10,000 symbols. High (more than 80 %) recall and precision of the results of string classification were demonstrated when analyzing simulated complex data, with properties comparable to real data. The developed method allows comparing metagenomic samples represented as a set of strings using only the data itself and without requiring additional information. This allows for a more accurate analysis compared to existing methods that compare samples based on the results of string classification in taxonomic annotation databases. The developed methods can also be used in other areas of string data processing such as analyzing changes in author style when writing a series of texts.

Keywords

time series, de Bruijn graph, *k*-mers, comparative analysis, strings classification, metagenomic samples

For citation: Ivanov A.B., Shalyto A.A., Ulyantsev V.I. Comparative analysis method for time series data objects represented as sets of strings based on de Bruijn graphs. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 4, pp. 718–726 (in Russian). doi: 10.17586/2226-1494-2025-25-4-718-726

Введение

Информация может быть представлена в различных форматах: в текстовых данных, в виде изображений, аудио- и видеофайлов. Примерами текстовых данных являются литературные произведения, информация в статьях в сети Интернет и текстовые посты пользователей в социальных сетях. Такая информация сохраняется в виде наборов строк различной длины в алфавите, который содержит строчные и прописные буквы одного или нескольких языков, цифры и специальные символы. Несмотря на поддержку большого числа специальных символов в современных текстовых редакторах, многие тексты используют ограниченный набор символов. Кроме текстов общего назначения, формат строковых данных может использоваться в предметных областях для описания данных по определенным правилам. Например, в области вычислительной биологии данные о генетической информации, получаемой из организмов при помощи секвенирования, сохраняются в файлы в виде прочтений — строк небольшой длины (100–300 символов) в алфавите из четырех символов — нуклеотидов (A, C, G, T).

Наиболее часто в методах для анализа наборов данных, заданных множеством строк, первым этапом является извлечение из строк коротких признаков. Для этого текст разбивается на слова фиксированной длины — *n*-граммы [1]. Такие слова применяются для сравнения текстов [2], ранжирования документов в сети Интернет [3], исправления опечаток [4, 5] и обнаружения плагиата [6]. Однако интерпретация признаков в виде *n*-грамм — сложная задача, в связи с тем, что они являются короткими (как правило, от двух до 10 символов). Отдельные *n*-граммы не обладают смыслом с точки зрения естественного языка. Для получения осмысленных слов или словосочетаний необходимо объединять *n*-граммы в более длинные последовательности.

Для решения этой задачи применим граф специального вида — граф де Брейна [7]. Он получил название в честь голландского математика Николаса де Брейна, который активно изучал свойства графов, построенных по множеству строк. Вершинами в графе де Брейна являются последовательности длины n из алфавита размера t (всего не более t^n вершин), а ориентированное ребро ведет из вершины a в вершину b, если суффикс строки a длины $n-1$ совпадает с префиксом строки b длины $n-1$.

Для получения интерпретируемых признаков *n*-граммы записываются в вершинах графа де Брейна, а ребра соединяют вершины, *n*-граммы которых пересекаются. Линейные пути, извлеченные из такого графа, соответствуют длинным текстовым строкам, которые

могут быть осмыслены с точки зрения предметной области, например, как ключевые слова или фразы на естественном языке.

Однако наиболее широкое применение графы де Брейна нашли в области вычислительной биологии для сборки геномных последовательностей и сравнения образцов геномного и метагеномного секвенирования [8–12]. Для построения графа де Брейна из метагеномных данных все прочтения образца разбиваются на k-меры — строки фиксированной длины k. Они являются аналогами n-грамм, однако извлекаются из строковых данных с пересечениями, что часто не верно для n-грамм. Каждому k-меру сопоставляется вершина в графе де Брейна, в котором ориентированные ребра соединяют k-меры, пересекающиеся на $k-1$ символах. Также в вершинах может быть записана дополнительная информация о k-мерах, например, частота их встречаемости в строках образца.

Актуальной является задача сравнения наборов строковых данных, состоящих из временной серии образцов из одного объекта. Для получения временной серии образцов производится фиксация состояний объекта в начальный и конечный моменты времени, а также при необходимости в нескольких промежуточных точках. В случае с текстовыми данными это могут быть версии статьи, в которую вносились правки, или посты от одного пользователя на заданную тему с похожим содержанием. В области анализа метагеномных данных изучаются сообщества образцов, взятые из одного источника в разные моменты времени. Различия, проявляющиеся в разные моменты времени, могут как быть связаны с сезонными изменениями, так и быть результатом влияния различных внешних факторов. Примерами могут быть анализ образцов воды в разные времена года или анализ микробного состава кишечника человека в динамике в ответ на терапию или трансплантацию микробиоты.

Существующие методы анализа метагеномных данных не позволяют извлекать интерпретируемые признаки, которые содержали бы информацию только о различиях в данных. Потому в настоящей работе предложен метод обработки временных серий образцов, заданных множеством строк, который позволит классифицировать строки по их встречаемости в данных и извлечь из каждой временной точки изучаемого образца общие и уникальные части. Этот метод может использоваться не только для анализа метагеномных данных, но для данных из любых областей, которые представляются в виде наборов строк. Например, с его помощью может проводиться анализ изменения постов человека в социальной сети на заданную узкую тему с течением времени.

Метод сравнения двух образцов на основе графа де Брейна для детекции изменений

Разработанный метод для сравнения двух образцов, представленных в виде набора строк, состоит в следующем. Один из образцов назовем *эталонным*, он соответствует исходному состоянию объекта, с которым проводится сравнение. Второй образец определим *анализируемым*, он соответствует новому состоянию исследуемого объекта, строки которого будут анализироваться. Метод состоит из следующих этапов.

Этап 1. Строки эталонного образца разбиваются на k-меры, из которых строится граф де Брейна. Вершинам соответствуют k-меры, а ребрам — пересечения длиной $k-1$.

Этап 2. Строки анализируемого образца разбиваются на k-меры. Осуществляется их поиск в графе де Брейна: каждый k-мер ищется в множестве k-меров, соответствующих вершинам графа де Брейна. В результате для каждой строки из анализируемого образца собирается информация о встречаемости его k-меров в графе де Брейна эталонного образца.

Этап 3. Строки из анализируемого образца разделяются на группы на основе рассчитываемых по частоте встречаемости k-меров статистик. Рассмотрим подробнее четыре решения данной задачи.

Все решения, используемые на этапе 3, направлены на установление факта встречаемости строки из анализируемого образца в множестве строк эталонного образца. Выбор решения зависит от анализируемых данных, доступных ресурсов и необходимости дальнейшего анализа данных. В таблице представлено сравнение решений на основе следующих параметров: парные прочтения (использование информации о связях между парами строк в метагеномных данных), исправление ошибок (использование информации о точности данных метагеномного секвенирования для исправления ошибок в строках), число категорий (число групп, на которые классифицируются строки анализируемого образца).

Решение 1 является базовым, на котором основаны решения 2–4, и состоит в следующем. Вначале для строки вычисляется глубина покрытия k-мерами как

Таблица. Сравнение четырех предложенных решений для разбиения строк на группы в методе сравнения двух образцов

Table. Comparison of four proposed solutions for strings classification in method for two samples comparison

Номер решения	Число значений k	Парные прочтения	Исправление ошибок	Число категорий	Условия применения
1	1	нет	нет	2	Первичное сравнение образцов
2	2	нет	нет	3	Сравнение похожих образцов
3	1	да	нет	2	Парная информация сохраняется для анализа и сборки
4	2	да	да	6	Лучшая детализация различий, наиболее ресурсоемкий вариант

средняя частота встречаемости *k*-меров из строки в графе де Брейна. Затем для строки вычисляется ширина покрытия как отношение числа покрытых в строке символов *k*-мерами к длине строки. Далее наблюдаемая ширина покрытия сравнивается с теоретической шириной покрытия, рассчитываемой на основе глубины покрытия и распределения Пуассона [13]. Если наблюдаемая ширина покрытия попадает в доверительный интервал теоретической ширины с уровнем доверия 95 %, и при этом ширина покрытия больше задаваемого при запуске пользователем параметра (по умолчанию 0,9), то она помещается в группу обнаруженных строк. В противном случае строка помещается в группу необнаруженных строк. Группа обнаруженных строк соответствует части данных анализируемого образца, которая сохранилась из эталонного образца. Группа необнаруженных строк соответствует части данных анализируемого образца, которой не было в эталонном образце. В дальнейшем обе группы обрабатываются по отдельности для интерпретации данных в предметной области и получения информации о том, какая часть данных сохраняется между образцами, а какая появляется в результате внешнего воздействия на образец.

Решение 2 модифицирует решение 1 с помощью добавления возможности выбора второго значения *k* для построения графа де Брейна и извлечения *k*-меров из строк. Выбор значения параметра *k* может влиять на обнаружение строк, так как слишком маленькие его значения могут привести к случайным совпадениям, а слишком большие — к построению несвязного графа. Пользователь задает два различных значения параметра *k* и строка считается обнаруженной, только если она удовлетворяет условиям из решения 1 при обоих значениях. Если строка удовлетворяет условиям только для одного значения *k*, то она помещается в группу *частично обнаруженных* строк. Оставшиеся строки помещаются в группу необнаруженных строк.

Решения 3 и 4 используют особенности строковых данных, получаемых в результате метагеномного секвенирования, для которого характерно наличие информации о связях между парами строк, а также информации о точности определения каждого символа в строке.

Решение 3 обрабатывает данные, которые имеют парную структуру: для каждой строки во входных данных существует другая строка (известная на основе метаданных). Обе строки из пары должны либо присутствовать в анализируемом образце, либо отсутствовать в нем. Исходя из этого, при получении разных результатов классификации для двух прочтений из одной пары, оба прочтения попадают в группу необнаруженных в эталонном образце. Прочтение попадает в группу обнаруженных строк, только если оно само и парное к нему удовлетворяют условиям из решения 1.

Решение 4 использует информацию о наличии ошибок в данных. Современные технологии получения строковых данных метагеномного секвенирования обычно обеспечивают не более одной ошибки в строке, а также сохраняют дополнительную информацию о качестве определения каждой позиции в строке. В случае неудачной попытки обнаружения прочтения в эталонном образце предпринимается попытка исправить один символ в строке с худшим качеством (если его значение ниже задаваемого порога). Если исправленный вариант строки обнаруживается в эталонном образце, то считается, что в исходной строке была ошибка, и она относится к категории обнаруженных.

Метод сравнительного анализа временных серий наборов данных из трех образцов

На основе метода сравнения двух образцов, представленных в виде множества строк, в настоящей работе разработан метод для анализа временных серий из трех образцов с использованием графов де Брейна. Рассматривается ситуация, в которой два образца соответствуют одному *исследуемому* объекту, имеющему *исходное* и *итоговое* состояния. Третий образец аналогичен *модифицирующему* объекту, который вносит изменения в набор строк исследуемого объекта. Метод используется для отслеживания изменений, которые добавляются в исследуемый объект с помощью другого объекта, и состоит из следующих этапов.

Этап 1. Для каждого образца строится отдельный граф де Брейна. Вершинам соответствуют *k*-меры, а ребрам — пересечения длиной $k - 1$.

Этап 2. *k*-меры исходного образца выравниваются по итоговому образцу для того, чтобы получить разбиение на вытесненную и сохранившуюся в результате модификаций часть данных.

Этап 3. *k*-меры модифицирующего образца выравниваются по итоговому образцу для того, чтобы понять, какая часть изменений была принята, а какая — отклонена.

Этап 4. *k*-меры итогового образца поочередно выравниваются по исходному и модифицирующему образцам. Таким образом, получаются четыре группы строк: сохранившиеся от исходного состояния прочтения, приобретенные от модифицирующего объекта прочтения, общие для исследуемого и модифицирующего объектов прочтения, а также уникальные для итогового состояния прочтения, связанные с естественной вариацией в метагеномных данных.

На основе использования этого метода извлекаются 8 групп строк. Схема разбиения исходных данных для каждого образца временной серии из трех метагеномов представлена на рис. 1. Иногда возникает необходимость анализа исследуемого образца, для которого доступно более чем две временные точки (например, до изменений, сразу после изменений и через фиксированные промежутки времени после изменений) для определения закрепления изменений в образце. В таком случае в качестве итогового образца поочередно используются временные точки после изменений и делается вывод о наличии стабильных изменений.

Вычислительные эксперименты

Метод был реализован в виде пайплайна на языке программирования Shell и размещен в открытом доступе[1]. Для валидации разработанного метода были прове-

[1] [Электронный ресурс]. Режим доступа: https://doi.org/10.5281/zenodo.15570956 (дата обращения: 16.06.2025).

Рис. 1. Схема разбиения строк на 8 групп с применением графов де Брейна при анализе временной серии из трех образцов, представленных в виде множества строк

Fig. 1. Scheme of strings classification into 8 categories using de Bruijn graphs in time series analysis of three samples represented as sets of strings

дены вычислительные эксперименты с использованием сгенерированных данных, моделирующих строковые данные, применяемые в прикладных задачах.

Наборы данных. Для тестирования предложенного метода анализа временных серий были рассмотрены строковые данные на примере секвенирования метагеномных образцов. Сгенерированы два типа модельных данных. Первый тип данных позволяет установить, насколько близкие организмы может различить представленный метод. Второй тип дает возможность понять точность метода анализа временных серий из трех образцов в зависимости от числа видов бактерий в образце и объема данных.

Первый тип модельных данных использовался для оценки точности обнаружения различий между метагеномными данными, содержащими похожие виды бактерий, в зависимости от степени подобия геномов и средней глубины покрытия образца.

Различение близких видов и штаммов является важной задачей, поскольку некоторые штаммы одного вида могут быть патогенными для организма носителя, в то время как другие штаммы этого вида — безвредны. Для генерации образцов применено множество из 12 геномов штаммов кишечной палочки (*E.Coli*). Было сгенерировано 27 модельных наборов данных с различными параметрами, выбор которых описан в подразделе «Методика эксперимента». Каждый набор состоит из трех образцов: исходного, модифицирующего и итогового. Для исходного и модифицирующего образцов случайным образом выбиралось по одному штамму кишечной палочки. Итоговый образец состоит из двух штаммов: объединения из исходного и модифицирующего образцов. Из референсных геномов штаммов с помощью программы InSilicoSeq [14] производилась генерация строковых данных: метагеномных прочтений, моделирующих выходные данные от секвенатора IlluminaHiSeq.

Второй тип модельных данных использовался для определения точности работы метода при большом объеме данных и большом числе видов организмов в образце. Итоговый образец содержал как геномы бактерий из исходного и модифицирующего образцов, так и новые геномы. Было сгенерировано 15 модельных наборов данных с различными параметрами, выбор которых описан в подразделе «Методика эксперимента». Каждый набор состоит из трех образцов: исходного, модифицирующего и итогового. Для генерации образцов применен открытый набор данных из 1520 видов референсных геномов бактерий, присутствующих в кишечнике человека [15]. Число геномов в каждом образце зависело от выбранных параметров эксперимента. Из референсных геномов штаммов с помощью программы InSilicoSeq с параметрами запуска по умолчанию производилась генерация строковых данных: метагеномных прочтений, моделирующих выходные данные от секвенатора IlluminaHiSeq. Относительная представленность видов бактерий внутри образца подчинялась экспоненциальному распределению.

Методика эксперимента. В первом типе данных моделировалась ситуация, когда в исходном и модифицирующем образцах присутствуют разные штаммы, которые затем обнаруживаются вместе в итоговом образце. Было проведено 27 экспериментов с девятью различными парами геномов и тремя уровнями глубины покрытия. Частота совпадений между геномами оценивалась с помощью программы Mash [16] (ось абсцисс на рис. 2). Для каждого эксперимента генерировался набор из трех образцов, которые обрабатывались с помощью разработанного метода анализа временных серий.

Для второго типа данных было проведено 15 экспериментов, которые отличались тремя видами сложности и пятью видами глубины покрытия. Сложность задавалась числом различных геномов в образце, выбранных случайным образом из 1520 референсных геномов: низкая сложность — 30 геномов, средняя — 100 геномов, высокая — 300 геномов. Для каждого эксперимента генерировался набор из трех образцов, которые обрабатывались с помощью разработанного метода анализа временных серий.

Для оценки точности классификации строк использовались метрики полноты и точности [17]. Точность определяется как доля верно классифицированных строк среди всех классифицированных. Полнота определяется как доля верно классифицированных строк среди истинных, которые должны принадлежать данной группе.

Результаты. Результаты верного обнаружения прочтений для референсных геномов штаммов для каждой из 8 получаемых групп строк (заголовки графиков) оценивались с помощью метрик полноты (рис. 2, *а*) и

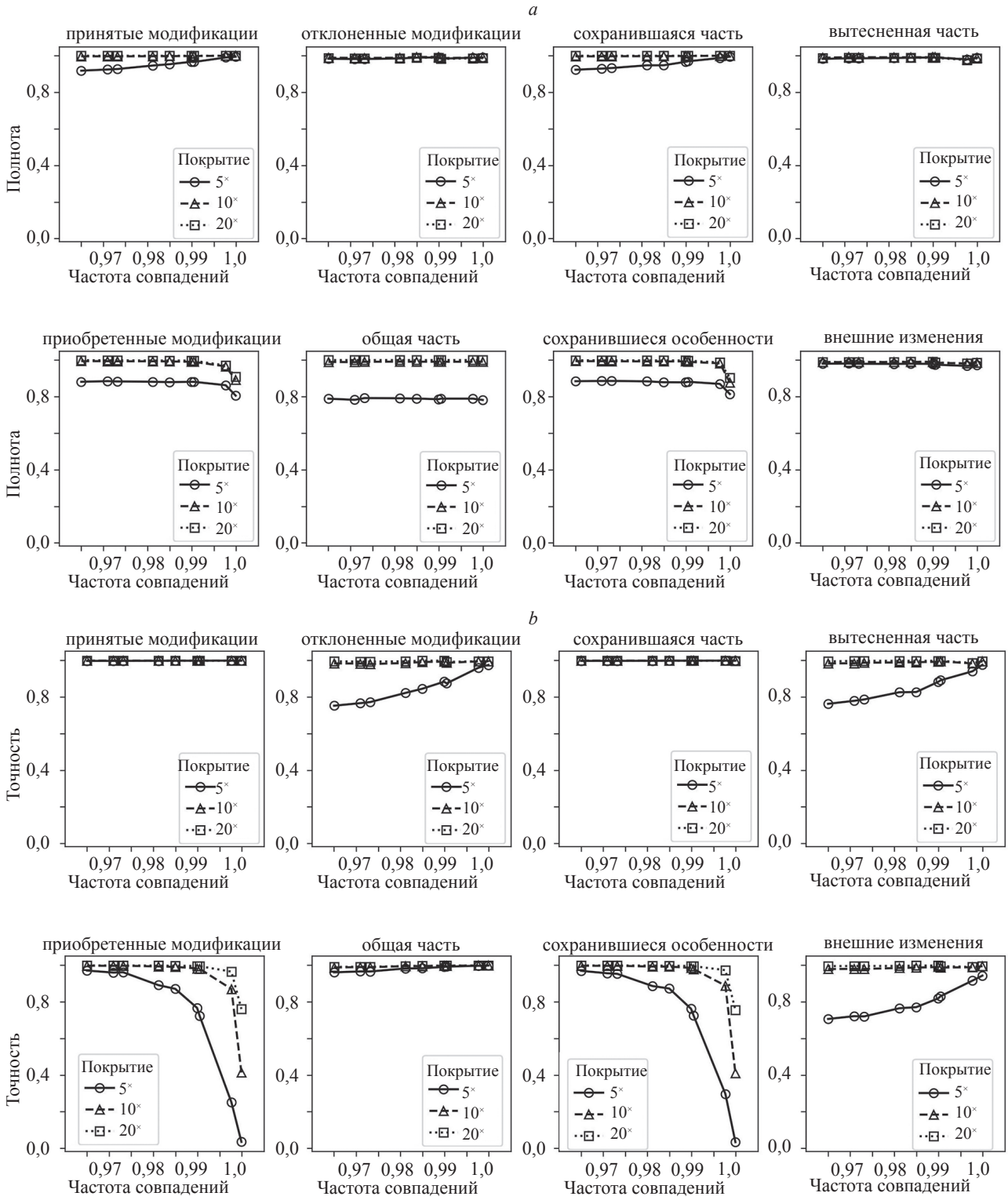

Рис. 2. Полнота (*a*) и точность (*b*) обнаружения прочтений из штаммов исследуемого и модифицирующего объектов, сосуществующих в итоговом объекте.

Частота совпадений соответствует схожести геномов

Fig. 2. Recall (*a*) and Precision (*b*) of reads detection from strains of the studied and modifying object coexisting in the final object

точности (рис. 2, *b*). Из графиков видно, что уже при глубине покрытия образца увеличением в $10^{\times}$ разработанный метод позволяет достоверно различить штаммы с частотой несовпадений до одного на 10 000 нуклеотидов, что равно максимальной точности современных секвенаторов коротких прочтений. Дальнейшее увеличение точности нецелесообразно, поскольку невозможно будет отличить истинную классификацию строк на группы от классификации, получаемой в результате наличия ошибок в данных.

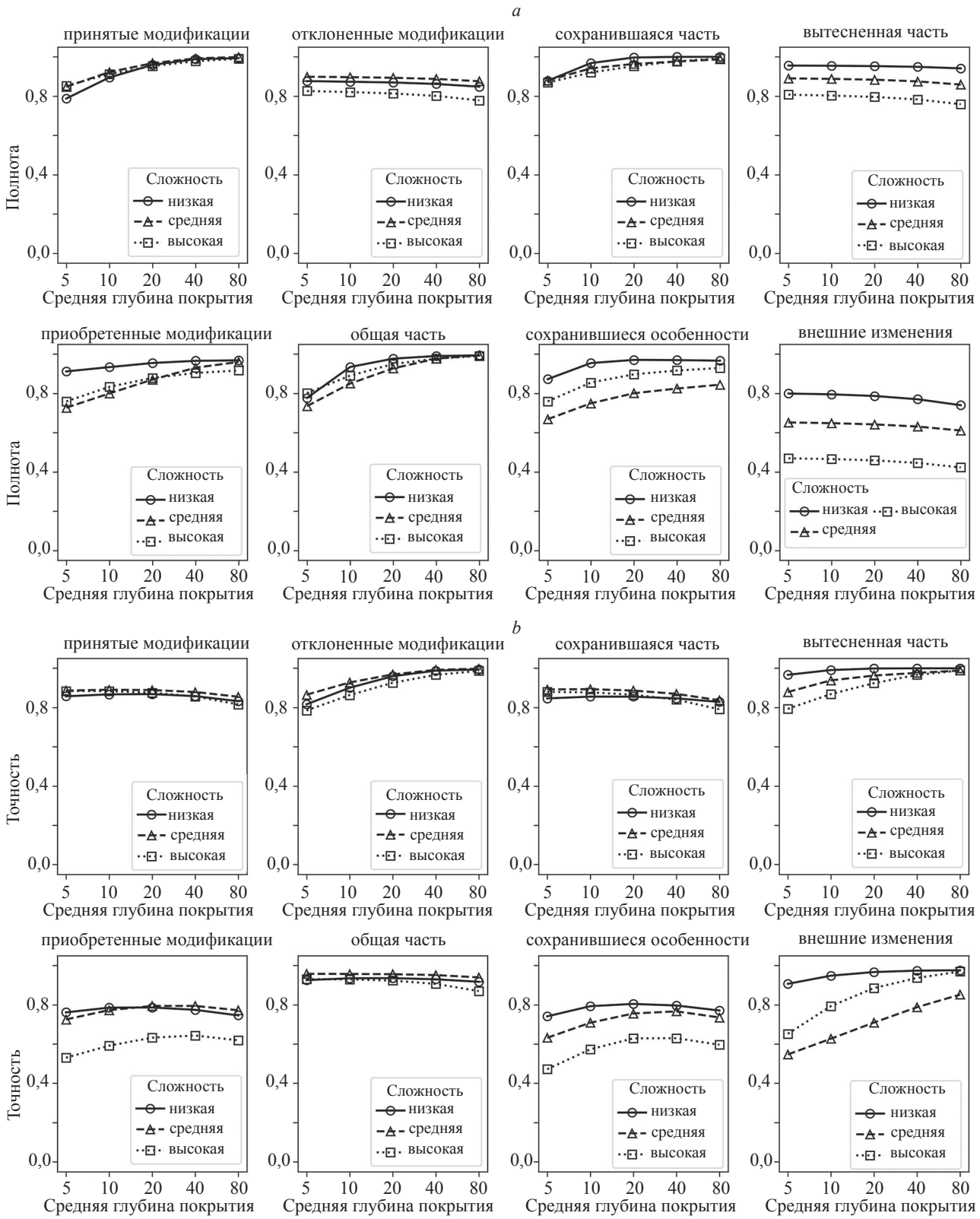

Рис. 3. Полнота (*a*) и точность (*b*) обнаружения прочтений для наборов данных в зависимости от числа геномов и глубины покрытия (в кратах)

Fig. 3. Recall (*a*) and Precision (*b*) of reads detection for datasets vs. the number of genomes and depth coverage

Результаты верного обнаружения прочтений для геномов различных видов для каждой из 8 получаемых групп строк (заголовки графиков) оценивались с помощью метрик полноты (рис. 3, *a*) и точности (рис. 3, *b*). Результаты группировки строк для исходного и модифицирующего образцов, обнаруживаемых в итоговом образце, стремится по метрике полноты к единице с ростом глубины покрытия независимо от сложности

набора данных. Худшее качество получается в группе строк, соответствующих внешним изменениям. Однако в прикладных задачах процент таких строк от общего объема данных является как правило небольшим (от 1 до 5 %), поэтому точность данной группы не является критичной. Также показано, что глубина покрытия играет важную роль независимо от сложности образца, поэтому на этапе сбора данных и их предварительной обработки необходимо учитывать данный параметр: для получения качественных результатов анализа необходимо запускать секвенирование с достаточным покрытием (хотя бы 20×), а также валидировать результаты, полученные на данных с невысоким покрытием.

Обсуждение

Задача сравнения наборов строковых данных, которые изменяются со временем, актуальна во многих прикладных областях. Примерами являются статьи или страницы в сети Интернет, в которые вносятся правки; посты в социальных сетях от одного пользователя или в одной группе на заданную тему. Также строковые данные используются для описания метагеномных образцов из сообществ бактерий, например, населяющих кишечник человека. Изучение таких сообществ помогает выявлять взаимосвязи между набором бактерий и здоровьем человека, работой его иммунной системы и реакцией на различные заболевания и методы лечения.

В работе предложены методы анализа временных серий строковых данных из двух и трех образцов. Отметим, что второй метод, является основным и базируется на первом. Вычислительные эксперименты показали, что второй метод может быть применен для анализа данных микробиоты кишечника, взятых в трех временных точках (например, до лечения, в начале лечения и после окончания лечения). Данные были промоделированы таким образом, чтобы соответствовать сложности реальных микробных сообществ кишечника человека. Результаты экспериментов подтверждают, что предложенный второй метод позволяет с высокой точностью характеризовать изменения, которые происходят во временной серии образцов. Метод позволяет выделить группы прочтений, которые соответствуют как стабильным на протяжении времени бактериям, так и появляющимся или исчезающим в результате терапии. В дальнейшем каждую из полученных групп прочтений возможно анализировать отдельно. Для этого может применяться аннотация с целью установления видов бактерий и их функций, а также сборка геномов и анализ отдельных генов для выявления принципов, которые отвечают за стабильность микробиоты кишечника. Разработанный метод трех образцов может быть применен к открытым данным для формулирования биологических гипотез, которые затем могут быть экспериментально проверены.

Заключение

В работе предложен метод сравнения двух образцов, представленных в виде набора строк. Метод позволяет производить классификацию строк на группы с использованием графов де Брейна и информации о частоте встречаемости *k*-меров. На его основе разработан метод для сравнительного анализа временных серий, состоящих из трех образцов: исходного и итогового состояния одного объекта и другого модифицирующего объекта. В этом методе производится попарное сравнение образцов для разбиения строк на подмножества, соответствующие закрепившимся и не закрепившимся изменениям. В дальнейшем полученные подмножества обрабатываются по отдельности для интерпретации данных в предметной области и получения информации о том, какая часть данных сохраняется между образцами, а какая появляется в результате внешнего воздействия на образец.

Показана высокая точность предложенных методов при анализе модельных данных метагеномного секвенирования. Предложенные методы могут быть использованы для анализа временных серий образцов, например, для изучения динамики изменений микробиоты кишечника пациентов в ответ на терапию или трансплантацию микробиоты. Полученные результаты могут помочь понять принципы изменений в микробных сообществах и лечь в основу систем поддержки принятия решений для модификаций сообществ.

Литература

1. Brown P.F., Della Pietra V.J., Desouza P.V., Lai J.C., Mercer R.L. Classbased *n*-gram models of natural language // Computational Linguistics. 1992. V. 18. N 4. P. 467–480.
2. Cavnar W.B., Trenkle J.M. N-gram-based text categorization // Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94). 1994. P. 161–175.
3. Rajalakshmi R., Aravindan C. Web page classification using n-gram based URL features // Proc. of the Fifth International Conference on Advanced Computing (ICoAC). 2013. P. 15–21. https://doi.org/10.1109/icoac.2013.6921920
4. Riseman E.M., Hanson A.R. A contextual postprocessing system for error correction using binary *n*-grams // IEEE Transactions on Computers. 1974. V. C-23. N 5. P. 480–493. https://doi.org/10.1109/T-C.1974.223971
5. Sidorov G., Gupta A., Tozer M., Catala D., Catena A., Fuentes S. Rule-based system for automatic grammar correction using syntactic *n*-grams for English language learning (L2) // Proc. of the 17th

References

1. Brown P.F., Della Pietra V.J., Desouza P.V., Lai J.C., Mercer R.L. Classbased *n*-gram models of natural language. *Computational Linguistics*, 1992, vol. 18, no. 4, pp. 467–480.
2. Cavnar W.B., Trenkle J.M. N-gram-based text categorization. *Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94)*, 1994, pp. 161–175.
3. Rajalakshmi R., Aravindan C. Web page classification using n-gram based URL features. *Proc. of the Fifth International Conference on Advanced Computing (ICoAC)*, 2013, pp. 15–21. https://doi.org/10.1109/icoac.2013.6921920
4. Riseman E.M., Hanson A.R. A contextual postprocessing system for error correction using binary *n*-grams. *IEEE Transactions on Computers*, 1974, vol. C-23, no. 5, pp. 480–493. https://doi.org/10.1109/T-C.1974.223971
5. Sidorov G., Gupta A., Tozer M., Catala D., Catena A., Fuentes S. Rule-based system for automatic grammar correction using syntactic *n*-grams for English language learning (L2). *Proc. of the 17th*

Conference on Computational Natural Language Learning: Shared Task. 2013. P. 96–101.
6. Barrón-Cedeño A., Rosso P. On automatic plagiarism detection based on *n*-grams comparison // Lecture Notes in Computer Science. 2009. V. 548. P. 696–700. https://doi.org/10.1007/978-3-642-00958-7_69
7. Huang S., Zhang H., Bao E. A Comprehensive review of the de Bruijn graph and its interdisciplinary applications in computing // Engineered Science. 2024. V. 28. P. 1061. https://doi.org/10.30919/es1061
8. Idury R.M., Waterman M.S. A new algorithm for DNA sequence assembly // Journal of Computational Biology. 1995. V. 2. N 2. P. 291–306. https://doi.org/10.1089/cmb.1995.2.291
9. Pevzner P.A., Tang H., Waterman M.S. An Eulerian path approach to DNA fragment assembly // Proceedings of the National Academy of Sciences of the United States of America. 2001. V. 98. N 17. P. 9748–9753. https://doi.org/10.1073/pnas.171285098
10. Compeau P.E.C, Pevzner P.A., Tesler G. How to apply de Bruijn graphs to genome assembly // Nature Biotechnology. 2011. V. 29. N 11. P. 987–991. https://doi.org/10.1038/nbt.2023
11. Nurk S., Meleshko D., Korobeynikov A., Pevzner P.A. metaSPAdes: a new versatile metagenomic assembler // Genome Research. 2017. V. 27. N 5. P. 824–834. https://doi.org/10.1101/gr.213959.116
12. Компо Ф., Певзнер П. Алгоритмы биоинформатики. М.: ДМК-Пресс, 2023. 680 с.
13. Пуассон С.Д. Исследования о вероятности приговоров в уголовных и гражданских делах. Берлин: NG Verlag, 2013. 328 с.
14. Gourlé H., Karlsson-Lindsjö O., Hayer J., Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq // Bioinformatics. 2019. V. 35. N 3. P. 521–522. https://doi.org/10.1093/bioinformatics/bty630
15. Zou Y., Xue W., Luo G., Deng Z., Qin P., Guo R., et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses // Nature Biotechnology. 2019. V. 37. N 2. P. 179–185. https://doi.org/10.1038/s41587-018-0008-8
16. Ondov B.D., Treangen T.J., Melsted P., Mallonee A.B., Bergman N.H., Koren S., Phillippy A.M. Mash: fast genome and metagenome distance estimation using MinHash // Genome Biology. 2016. V. 17. P. 132. https://doi.org/10.1186/s13059-016-0997-x
17. Buckland M., Gey F. The relationship between recall and precision // Journal of the American Society for Information Science. 1994. V. 45. N 1. P. 12–19. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L

Conference on Computational Natural Language Learning: Shared Task, 2013, pp. 96–101.
6. Barrón-Cedeño A., Rosso P. On automatic plagiarism detection based on *n*-grams comparison. *Lecture Notes in Computer Science*, 2009, vol. 548, pp. 696–700. https://doi.org/10.1007/978-3-642-00958-7_69
7. Huang S., Zhang H., Bao E. A comprehensive review of the de Bruijn graph and its interdisciplinary applications in computing. *Engineered Science*, 2024, vol. 28, pp. 1061. https://doi.org/10.30919/es1061
8. Idury R.M., Waterman M.S. A new algorithm for DNA sequence assembly. *Journal of Computational Biology*, 1995, vol. 2, no. 2, pp. 291–306. https://doi.org/10.1089/cmb.1995.2.291
9. Pevzner P.A., Tang H., Waterman M.S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 2001, vol. 98, no. 17, pp. 9748–9753. https://doi.org/10.1073/pnas.171285098
10. Compeau P.E.C, Pevzner P.A., Tesler G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 2011, vol. 29, no. 11, pp. 987–991. https://doi.org/10.1038/nbt.2023
11. Nurk S., Meleshko D., Korobeynikov A., Pevzner P.A. metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 2017, vol. 27, no. 5, pp. 824–834. https://doi.org/10.1101/gr.213959.116
12. Compeau Ph., Pevzner P. Bioinformatics Algorithms: an Active Learning Approach. Active Learning Publishers, 2018, 684 p.
13. Poisson S.-D. *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile: Précédées des Règles Générales du Calcul des Probabilités*. Adegi Graphics LLC, 1999, 431 p. (in French)
14. Gourlé H., Karlsson-Lindsjö O., Hayer J., Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*, 2019, vol. 35, no. 3, pp. 521–522. https://doi.org/10.1093/bioinformatics/bty630
15. Zou Y.,Xue W., Luo G., Deng Z., Qin P., Guo R., et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology*, 2019, vol. 37, no. 2, pp. 179–185. https://doi.org/10.1038/s41587-018-0008-8
16. Ondov B.D., Treangen T.J., Melsted P., Mallonee A.B., Bergman N.H., Koren S., Phillippy A.M. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 2016, vol. 17, pp. 132. https://doi.org/10.1186/s13059-016-0997-x
17. Buckland M., Gey F. The relationship between recall and precision. *Journal of the American Society for Information Science*, 1994, vol. 45, no. 1, pp. 12–19. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L

Авторы

Иванов Артем Борисович — младший научный сотрудник, Федеральный научно-клинический центр физико-химической медицины им. академика Ю. М. Лопухина Федерального медико-биологического агентства, Москва, 119435, Российская Федерация; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57222438932, https://orcid.org/0000-0002-7997-0637, abivanov@itmo.ru

Шалыто Анатолий Абрамович — доктор технических наук, профессор, главный научный сотрудник, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 56131789500, https://orcid.org/0000-0002-2723-2077, anatoly.shalyto@gmail.com

Ульянцев Владимир Игоревич — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 55062303000, https://orcid.org/0000-0003-0802-830X, ulyantsev@itmo.ru

Authors

Artem B. Ivanov — Junior Researcher, Lopukhin Federal Research and Clinical Center of Physical-Chemical Medicine of Federal Medical Biological Agency (FRCC PCM), Moscow, 119435, Russian Federation; PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57222438932, https://orcid.org/0000-0002-7997-0637, abivanov@itmo.ru

Anatoly A. Shalyto — D.Sc., Professor, Chief Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 56131789500, https://orcid.org/0000-0002-2723-2077, anatoly.shalyto@gmail.com

Vladimir I. Ulyantsev — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 55062303000, https://orcid.org/0000-0003-0802-830X, ulyantsev@itmo.ru

Статья поступила в редакцию 22.04.2025
Одобрена после рецензирования 03.06.2025
Принята к печати 17.07.2025

Received 22.04.2025
Approved after reviewing 03.06.2025
Accepted 17.07.2025