НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2025 Том 25 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2025 Vol. 25 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-3-545-553
УДК 004.02

Методы извлечения *k*-меров и признаков из наборов метагеномных графов де Брейна на основе информации о классах образцов

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 наборов метагеномных образцов с использованием графов де Брейна. Для повышения точности работы классификационных моделей разработаны методы автоматического извлечения признаков на основе результатов сравнительного анализа метагеномных образцов, экспертных метаданных и статистических тестов. Под признаками в данной работе понимаются связные подграфы графа де Брейна. **Методы.** Первый метод *unique_kmers* применяется для извлечения из данных строк длины *k* (*k*-меров), которые встречаются только в образцах одного класса. Второй метод *stats_kmers* применяется для извлечения *k*-меров, частота встречаемости которых статистически отличается между классами образцов. Для извлечения интерпретируемых признаков разработан третий метод, в котором реализовано выделение подграфов из графов де Брейна на основе опорных вершин, получаемых в результате применения одного из первых двух методов. Анализ данных состоит из двух этапов: вначале применяется метод *unique_kmers* или *stats_kmers* для предварительной обработки данных, затем к полученным результатам применяется третий метод для получения интерпретируемых признаков. **Основные результаты.** Апробация методов проведена на четырех сгенерированных тестовых наборах данных, которые моделируют параметры реальных метагеномных сообществ, такие как наличие похожих видов (штаммов) или разницу в частоте встречаемости бактерии. Разработанные методы были применены для извлечения признаков, которые использовались для классификации образцов из тестовых наборов. Для сравнения в качестве признаков использовались результаты таксономической аннотации образцов с помощью программы Kraken2. Показано, что качество классификации образцов повысилось при применении в классификационных моделях признаков, полученных с помощью предложенных методов, по сравнению с классификационными моделями, обученными на таксономических признаках. **Обсуждение.** Разработанные методы полезны при сравнительном анализе данных метагеномного секвенирования и могут служить основой систем поддержки принятия решений, например, при диагностировании заболеваний людей на основе данных секвенирования микробиоты кишечника.

Ключевые слова

извлечение признаков, граф де Брейна, *k*-меры, классификация, метагеномика

Ссылка для цитирования: Иванов А.Б., Шалыто А.А., Ульянцев В.И. Методы извлечения *k*-меров и признаков из наборов метагеномных графов де Брейна на основе информации о классах образцов // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 3. С. 545–553. doi: 10.17586/2226-1494-2025-25-3-545-553

Feature extraction methods for metagenome de Bruijn graphs collections based on samples classification information

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

[1] Lopukhin 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 metagenomic samples collections using de Bruijn graphs. We propose methods for automatic feature extraction based on the results of comparative sample analysis, expert metadata, and statistical tests to improve the accuracy of classification models. In this paper features are connected subgraphs of the de Bruijn graph. The first method, named *unique_kmers*, is used to extract strings of length *k* (*k*-mers) that occur only in samples of the certain class. The second method, named *stats_kmers*, is used to extract *k*-mers whose frequency of occurrence statistically differs between sample classes. To extract interpretable features, a third method has been developed that implements the extraction of subgraphs from de Bruijn graphs based on the selected nodes obtained as a result of applying one of the first two methods. Data analysis consists of two stages: firstly, *unique_kmers* or *stats_kmers* method is applied for data preprocessing, secondly, the third method is applied to obtain interpretable features. The methods were tested on four generated datasets that model the properties of real metagenomic communities such as the presence of similar species (strains) or differences in the relative abundance of bacteria. The developed methods were used to extract features. Machine learning model was trained in extracted features to classify samples from the test datasets. For comparison, the results of taxonomic annotation of samples using the Kraken2 program were used as features. It was shown that the accuracy of samples classification increased when using features obtained using the proposed methods in classification models compared to classification models trained on taxonomic features. The developed methods are useful for comparative analysis of metagenomic sequencing data and can form the basis of decision support systems, for example, in human diseases diagnostics based on gut microbiota sequencing data.

Keywords

feature extraction, de Bruijn graph, *k*-mers, classification, metagenomics

For citattion: Ivanov A.B., Shalyto A.A., Ulyantsev V.I. Feature extraction methods for metagenome de Bruijn graphs collections based on samples classification information. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 3, pp. 545–553 (in Russian). doi: 10.17586/2226-1494-2025-25-3-545-553

Введение

Метагеномика — раздел вычислительной биологии, который изучает в совокупности сообщества микроорганизмов, населяющих определенные экологические ниши: почва [1], водоемы [2], кожные покровы и кишечник человека [3]. Анализ метагеномных данных включает в себя определение видового состава образца, установление функций и роли отдельных таксонов (бактерий, вирусов) во взаимодействиях внутри сообщества и с окружающей средой, а также сравнение образцов между собой. Сравнительная метагеномика исследует сходства и различия метагеномных образцов с целью выявления закономерностей между микробным составом и свойствами окружающей среды или хозяина. Например, анализ микробиоты кишечника человека играет важную роль в современных медицинских исследованиях: для диагностирования заболеваний, прогноза успешности лечения и подбора персонализированной терапии [4–10].

Одной группой методов анализа метагеномных данных является сборка полных геномов с их последующей аннотацией — определением видового состава образца. Для этих целей широкое применение нашли графы де Брейна [11–14]. Несмотря на наличие эффективных методов для построения графа де Брейна и разбиения его на компоненты [15–17], задача сборки является сложной в связи с объемом данных, их комплексностью (образец является смесью геномов многих видов) и зашумленностью. Сборка генома на 90 % из метагеномного образца требует в среднем десятикратного покрытия при секвенировании [18], но большой объем получаемых данных трудно обрабатывать, а стоимость ультраглубокого секвенирования не позволяет использовать его повсеместно. Аннотация применяется с целью установить видовой состав микроорганизмов в образце, однако она ограничена неполнотой баз данных, существующей из-за неизученной части мира микробов и их быстрой естественной эволюции в настоящем времени. При этом данные, полученные со стандартной глубиной секвенирования, позволяют обнаруживать только виды с относительной представленностью больше одного процента в образце, что приводит к потере большой доли информации о разнообразии слабо представленных видов [19].

Другой группой методов анализа метагеномных данных является обработка, фильтрация и классификация «сырых данных» — прочтений, поступающих из секвенатора. Это позволяет учитывать всю извлеченную из образца информацию и не зависеть от баз данных, однако требует разработки эффективных алгоритмов. В задачах сравнения метагеномных последовательностей используются методы на основе *k*-меров — подстроках длины *k*. Некоторые алгоритмы позволяют эффективно манипулировать *k*-мерами и применять их для оценки похожести исходных образцов [20–22]. Другие методы используют статистические тесты для отбора подмножества *k*-меров, которые являются клю-

чевыми для разбиения исходных образцов на категории [23, 24]. Однако все методы анализа «сырых данных» позволяют провести только сравнение метагеномных образцов.

Аннотация и интерпретация признаков в виде *k*-меров является сложной задачей в связи с неполнотой баз данных и короткой длиной *k*-меров (от 11 до 91 символа), что не позволяет точно установить их биологическую роль. С целью повышения точности интерпретации в данной работе в качестве признаков используются связные подграфы графа де Брейна. Строковые последовательности для подграфов обладают достаточной длиной (от 100 до нескольких тысяч символов), что позволяет точнее установить их биологическую роль по сравнению с *k*-мерами. Методы извлечения подграфов используются в задаче сборки геномов. Заметим, что предположительно не существует методов извлечения подграфов с использованием результатов сравнительного анализа метагеномных образцов, экспертных метаданных и статистических тестов.

Сложность анализа метагеномных данных состоит также в их объеме. С одной стороны, каждый образец метагеномного секвенирования занимает 5–10 ГБ, поэтому одновременная обработка даже 100 образцов для сравнительного анализа и извлечения признаков требует большого объема оперативной памяти и дискового пространства, как правило, доступных только на вычислительных серверах. С другой стороны, типичное число образцов в метагеномных исследованиях находится в диапазоне от 100 до 200 из-за сложности сбора данных, особенно если речь идет о работе с данными метагеномов людей, например пациентов с определенным заболеванием. Таким образом, возникает проблема большого числа признаков, которые могут быть извлечены из данных, и малого числа образцов, что является критической проблемой для методов машинного обучения. По этой причине актуальной является задача разработки методов извлечения признаков из графов де Брейна, в том числе и больших (10^8–10^{10} вершин), при наличии требования интерпретируемости (описание свойств признаков, например, установление биологической функции, соответствующей извлеченной последовательности).

В настоящей работе разработаны методы извлечения признаков из наборов метагеномных графов де Брейна, которые на основе результатов сравнительного анализа метагеномных образцов, экспертных метаданных и статистических тестов извлекают из указанных графов ветвящиеся связные подграфы. Новизна предложенных методов состоит в использовании информации о классах образцов на этапе извлечения признаков и извлечении признаков достаточной для интерпретации длины. Существующие решения удовлетворяют только одному из двух указанных свойств: или они используют метаданные, но извлекают признаки в виде коротких неинтерпретируемых последовательностей [21–24]; или они извлекают длинные интерпретируемые признаки из графов сборки, которые, однако, не позволяют с высокой точностью классифицировать образцы на группы [15, 16]. Показано, что признаки, полученные с помощью предложенных методов, повышают точность работы классификационных моделей по сравнению с моделями, обученными на признаках таксономической аннотации.

Для извлечения признаков в виде подграфов из наборов метагеномных графов де Брейна необходимо научиться выделять опорные *k*-меры. Опорными *k*-мерами будем называть подмножество всех *k*-меров из образцов, встречаемость которых отличается между образцами разных классов. Эти *k*-меры будут сопоставлены опорным вершинам графа, на основе которых будут строиться подграфы. Для извлечения опорных *k*-меров было разработано два метода *unique_kmers* и *stats_kmers*.

Метод *unique_kmers* для извлечения опорных *k*-меров на основе их уникальной представленности в данных

В настоящей работе разработан метод извлечения опорных *k*-меров на основе их уникальной представленности в данных. Этот метод назван *unique_kmers* и состоит из следующих этапов.

Этап 1. Каждый метагеномный образец представляется в виде множества *k*-меров. Полученные файлы с *k*-мерами группируются на основании экспертных метаданных, например классов метагеномных образцов.

Этап 2. Фиксируется один из классов, назовем его *X*. Для него выбирается пороговое значение *G*. Из всех образцов класса *X* отбираются *k*-меры, которые присутствуют не менее чем в *G* файлах данного класса. Полученные *k*-меры являются кандидатами на специфичность для класса *X*.

Этап 3. Производится фильтрация *k*-меров из множества кандидатов, полученных на этапе 2. *k*-мер исключается из множества, если он встречается хотя бы в одном образце не из класса *X*. Полученное множество образцов является уникальным для класса *X*.

Этап 4. Этапы 2 и 3 выполняются независимо по одному разу для каждого класса из набора данных.

В результате применения этого метода для каждого класса извлекается множество уникальных опорных *k*-меров. Он обеспечивает высокую точность и скорость на данных с сильными различиями между образцами разных классов, однако может приводить к ошибкам из-за вариабельности и зашумленности в данных.

Метод *stats_kmers* для извлечения опорных *k*-меров на основе статистических тестов

Для извлечения опорных *k*-меров из наборов метагеномных образцов недостаточно найти только уникальные *k*-меры, поскольку метагеномы обладают естественной вариацией и зашумленностью, а в данных содержатся ошибки. В рамках работы разработан метод извлечения опорных *k*-меров на основе статистических тестов. Этот метод назван *stats_kmers* и состоит из следующих этапов.

Этап 1. Каждый метагеномный образец представляется в виде множества *k*-меров. Полученные файлы с *k*-мерами группируются на основании экспертных метаданных, например классов метагеномных образцов.

Этап 2. Отбираются специфические *k*-меры, встречаемость которых статистически значимо различается между образцами разных классов. Для каждого *k*-мера формируется таблица из двух строк и *M* столбцов, где *M* — число классов образцов. В таблицу для каждого класса записывается число образцов, в которых встречается и не встречается *k*-мер. Затем с помощью критерия χ^2 [25] для таблиц сопряженности проверяется гипотеза о равенстве частоты встречаемости *k*-мера в классах (использована поправка Йейтса [26]). В результате остаются только те *k*-меры, для которых гипотеза отвергается при заданном уровне значимости.

Этап 3. Поскольку на этапе 2 используется только информация о присутствии *k*-меров в образцах, выполняется дополнительная фильтрация на основе представленности *k*-меров в классах образцов. В качестве выборки для каждого класса образцов рассматриваются частоты встречаемости *k*-меров в образцах этого класса. Проводится попарное сравнение выборок с использованием критерия Манна–Уитни [27], который проверяет гипотезу о том, что обе выборки получены из одного распределения. В результате удаляются *k*-меры, для которых ни одна гипотеза о равенстве распределений не отвергается: удаляются *k*-меры, присутствующие примерно одинаково во всех образцах всех классов. При наличии трех и более классов образцов используется поправка Бонферрони [28] для множественных сравнений.

Этап 4. Для прошедших все этапы фильтрации *k*-меров рассчитывается среднее значение его представленности в каждом из классов. *k*-мер признается специфичным для того класса, в которой этот показатель наибольший.

В результате применения этого метода для каждого класса извлекается множество специфичных опорных *k*-меров. Он показывает высокую точность при наличии в образцах разных классов одинаковых организмов (бактерий, вирусов) с разной частотой встречаемости.

Метод извлечения признаков в виде подграфов из графа де Брейна на основе опорных *k*-меров

Для удобства дальнейшей работы с извлеченными множествами опорных *k*-меров был предложен метод, который позволяет на их основе получать более длинные и интерпретируемые признаки. Недостатками использования *k*-меров в качестве признаков является их большое число и малая длина. Большое число признаков в виде *k*-меров затрудняет их применение в моделях машинного обучения случайного леса и линейной регрессии, особенно учитывая факт малого числа размеченных метагеномных образцов. Кроме того, малая длина *k*-меров существенно затрудняет их биологическую аннотацию, что ограничивает дальнейшую интерпретируемость модели и применимость в реальных задачах.

Для решения указанных проблем разработан метод, который на основе опорных *k*-меров позволяет извлекать признаки в виде подграфов графа де Брейна. Метод основывается на работе одного из двух методов извлечения опорных *k*-меров (*unique_kmers* или *stats_kmers*) и позволяет получить независимые признаки. Метод извлечения опорных *k*-меров, на основе которых будут строиться подграфы, выбирается экспериментатором в зависимости от входных данных (результаты экспериментов показали, что метод *stats_kmers* работает дольше, но признаки, построенные на основе его опорных *k*-меров, позволяют точнее классифицировать более похожие метагеномные образцы). Предлагаемый метод извлечения признаков в виде подграфов из графа де Брейна запускается отдельно для каждого класса образцов и соответствующего множества опорных *k*-меров, и состоит из следующих этапов.

Этап 1. Построение общего графа де Брейна из *k*-меров всех образцов (вершины — *k*-меры, ребра — пересечения длины $k-1$). Вершины графа де Брейна, соответствующие опорным *k*-мерам, полученным с помощью метода *unique_kmers* или *stats_kmers*, помечаются соответствующим образом.

Этап 2. В графе производится поиск связных компонент на основе опорных вершин. Запускается алгоритм обхода в ширину, начиная от случайно выбранной опорной вершины. Алгоритм состоит из следующих этапов.

Этап 2.1. Если из текущего *k*-мера существует путь только в одну вершину в графе, то следующая вершина присоединяется к текущей компоненте.

Этап 2.2. Если на пути в графе встречается развилка, то алгоритм выполняет предпросмотр в глубину в каждой ветке до следующей развилки. В случае обнаружения в какой-либо из веток опорной вершины, вся ветка добавляется к текущей компоненте.

Этап 2.3. Далее этапы 2.1 и 2.2 повторяются для последней вершины каждой добавленной ветки.

Этап 2.4. Если для текущего *k*-мера нет исходящих путей или на развилке не найдено ни одной ветки с опорными *k*-мерами, то обход заканчивается.

Этап 3. Все просмотренные на этапе 2 вершины отмечаются как посещенные. Этап 2 повторяется для еще не просмотренных опорных вершин. В результате получается набор подграфов де Брейна, содержащий все опорные *k*-меры. Каждый подграф может рассматриваться как отдельный признак, специфичный для данного класса образцов.

Этап 4. Каждый подграф сохраняется как множество строк, соответствующих линейным путям в графе. Полученные строки могут использоваться для аннотации и интерпретации признаков.

Метод позволяет для каждого класса метагеномных образцов (например, образцы здоровых и больных пациентов) получить набор признаков, специфичных для каждого класса. Для возможности использования признаков в классификационных моделях необходимо сопоставить им численные значения для каждого образца. Для этого определяется покрытие извлеченных подграфов *k*-мерами каждого образца. Оно рассчитывается как отношение числа *k*-меров в образце, попавших в данный подграф, к общему числу вершин в подграфе. Полученные значения объединяются в таблицу признаков с числом строк, равным числу образцов, и числом столбцов, равным суммарному числу признаков во всех классах. Эта таблица признаков используется в дальнейшем в моделях искусственного интеллекта для

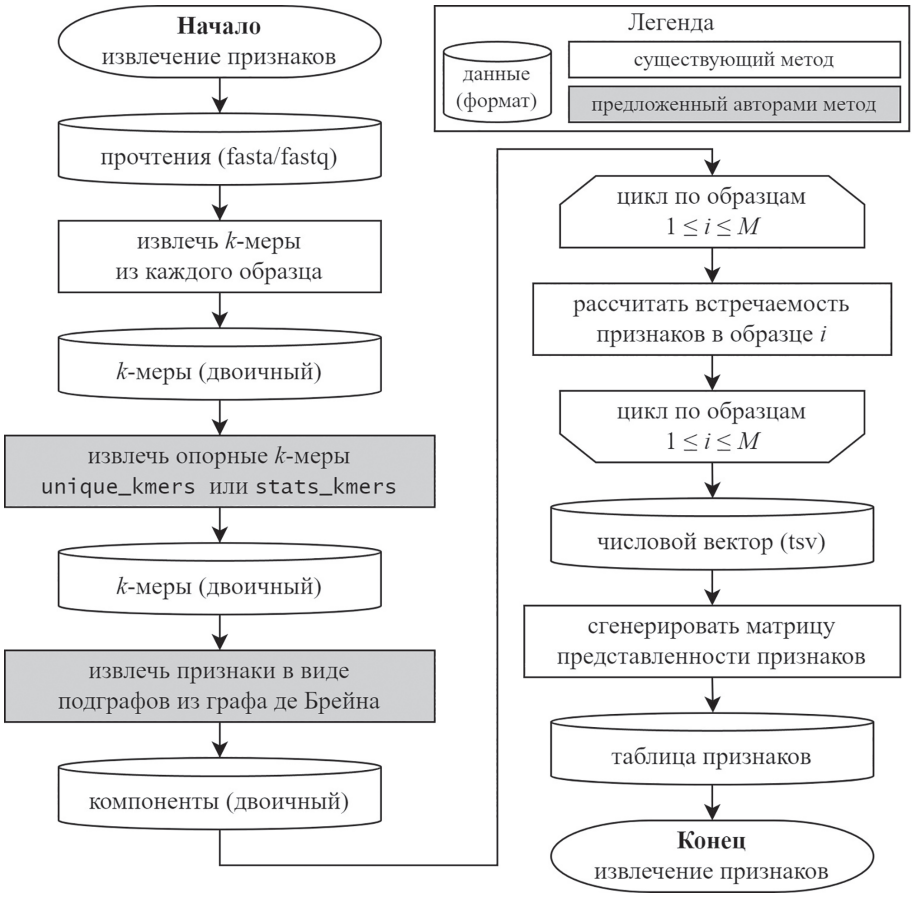

Рисунок. Схема алгоритма извлечения признаков из метагеномных данных

Figure. Algorithmic pipeline for feature extraction from metagenomic data

обучения классификационных моделей и предсказании классов для новых неразмеченных метагеномных образцов. Кроме того, для возможности биологического анализа, аннотации и интерпретации извлеченных признаков они преобразовываются из внутреннего двоичного формата подграфов в набор строковых последовательностей. В дальнейшем производится их поиск в базах данных для установления видовой принадлежности и выполняемых функций. Схема алгоритма (рисунок) извлечения признаков из метагеномных данных содержит предложенные в данной работе методы и показывает этапы обработки метагеномных данных от файла с прочтениями до интерпретируемых признаков.

Вычислительные эксперименты

Для валидации разработанных методов проведены вычислительные эксперименты с использованием сгенерированных данных, моделирующих метагеномные сообщества различной степени сложности.

Наборы данных. Для тестирования предложенных методов извлечения признаков из наборов метагеномных образцов были сгенерированы наборы метагеномных данных с известным распределением бактерий и известными различиями между двумя классами образцов. Были симулированы четыре набора данных, отличающихся похожестью содержащихся в них бактерий, каждый из которых состоял из 40 метагеномных образцов — по 10 образцов в двух классах для тренировочной и тестовой выборок. Видовой состав бактерий, содержащихся в разных наборах, приведен в табл. 1. Отбор видов бактерий для образца проводился следующим образом:

— для каждого образца случайным образом выбираются 10 бактерий из фиксированного списка из 20;
— для наборов данных 2 и 4 в каждый образец добавляется общий штамм кишечной палочки *E. Coli NZ_CP007265.1*;
— для наборов данных 1 и 2 в образцы классов А и Б добавляются различные штаммы кишечной палочки: *E. Coli NC_000913.3* в класс А и *E. Coli NC_002695.2* в класс Б;
— для наборов данных 3 и 4 в образцы классов А и Б добавляется одинаковый штамм кишечной палочки *E. Coli NC_000913.3*, причем гарантируется, что относительная частота его встречаемости в образцах класса А находится в интервале [0,01; 0,05], а в образцах класса Б в интервале [0,05; 0,1].

Для генерации метагеномных прочтений на основе файлов с видовым составом образцов использовалась программа InSilicoSeq [29]. В качестве параметров запуска была выбрана модель прочтений *--model hiseq*, число прочтений *--n-reads 5 000 000* и модель относительной встречаемости видов *--abundance exponential*.

Методика эксперимента. Для отбора опорных *k*-меров из сгенерированных данных были использо-

Таблица 1. Параметры сгенерированных наборов данных
Table 1. Generated datasets properties

Параметр		Класс А		Класс Б	
Выборка		Тренировочная	Тестовая	Тренировочная	Тестовая
Число образцов		10	10	10	10
Номер набора данных	1	10 случайных бактерий из 20			
		E. Coli NC_000913.3		*E. Coli NC_002695.2*	
	2	10 случайных бактерий из 20 + *E. Coli NZ_CP007265.1*			
		E. Coli NC_000913.3		*E. Coli NC_002695.2*	
	3	10 случайных бактерий из 20			
		E. Coli NC_000913.3 встречаемость ∈ [0,01; 0,05]		*E. Coli NC_000913.3* встречаемость ∈ [0,05; 0,1]	
	4	10 случайных бактерий из 20 + *E. Coli NZ_CP007265.1*			
		E. Coli NC_000913.3 встречаемость ∈ [0,01; 0,05]		*E. Coli NC_000913.3* встречаемость ∈ [0,05; 0,1]	

ваны два разработанных метода *unique_kmers* и *stats_kmers*. Затем применен метод извлечения признаков в виде подграфов из графа де Брейна на основе опорных *k*-меров и построены таблицы численных признаков. Сравнение проводилось с признаками, полученными с помощью метода таксономической аннотации Kraken2 [30]. Метод таксономической аннотации также в качестве результата выполнения возвращает таблицу численных признаков представленности различных таксонов в образцах. Полученные признаки были использованы для построения классификатора на основе метода машинного обучения — случайный лес [31], который был обучен по тренировочным образцам. Выбор модели обусловлен работой с табличными признаками и хорошей интерпретируемостью результатов работы случайного леса (для каждого дерева решений можно извлечь информацию, на основе каких признаков был классифицирован образец). При реализации использовался класс RandomForestClassifier из библиотеки scikit-learn [32] версии 1.3.0 для языка программирования Python версии 3.9.5. Для параметра числа решающих деревьев было установлено значение *--n-estimators 100*, для остальных параметров использовались значения по умолчанию. Признаки были применены для разметки тестовых образцов.

Поскольку в сгенерированных данных размер классов сбалансирован, то, как метрики качества были выбраны Precision и Recall [33]. Точность (Precision) определяется как доля верно классифицированных образцов среди всех классифицированных образцов. Чувствительность (Recall) определяется как доля верно классифицированных образцов среди истинных меток. В случае двухклассовой задачи (две категории образцов) *истинный класс* определяется как класс объектов, которые должны обнаруживаться с помощью классификатора (например, при разработке классификатора для диагностирования заболевания, истинным классом будет образец с заболеванием). К недостаткам метрики Recall относится необходимость выбора истинного класса, а также сложность обобщения на многоклассовые задачи (если требуется распознавать более чем два класса образцов). Однако эти метрики являются популярными для оценки качества классификации в моделях искусственного интеллекта.

Результаты. Результаты классификации приведены в табл. 2. Из полученных результатов видно, что для трех из четырех сгенерированных наборов данных, включая самый сложный — набор данных 4, с помощью разработанных методов извлечения признаков было получено более высокое качество классификации по сравнению с классическим методом таксономической аннотации. Для оставшегося набора данных метод *stats_kmers* показал максимально возможную точность, как и таксономическая аннотация. Метод *unique_kmers* допускал ошибки классификации в наборах данных 3 и 4, что объясняется отличиями между классами только в относительной представленности, в то время как этот метод направлен на поиск различных организмов. Полученные результаты подтверждают работоспособность и полезность разработанных методов. Показано повышение точности работы классификационных моделей при использовании извлеченных признаков по сравнению с моделями, обученными на признаках таксономической аннотации.

Обсуждение

Сравнение наборов метагеномных образцов, которые разделены на классы, является актуальной для решения прикладных биомедицинских задач. Примерами таких данных могут быть образцы, взятые при исследовании микробиоты кишечника здоровых людей и пациентов с заболеванием. Извлечение признаков может использоваться для выявления отдельных видов бактерий или продуцируемых ими метаболитов, которые связаны с развитием заболевания. Также на основании извлеченных признаков могут обучаться классификационные модели для ранней диагностики и скрининга людей на наличие заболеваний без необходимости проходить инвазивные процедуры.

В настоящей работе предложены методы для извлечения признаков из наборов метагеномных образцов.

Таблица 2. Результаты классификации наборов данных с помощью разработанных методов извлечения признаков и таксономической аннотации на сгенерированных данных

Table 2. Classification results based on developed feature extraction methods and taxonomic annotation for generated metagenomic datasets

Номер набора данных	Метрика	Метод извлечения признаков		
		Таксономия Kraken2	*unique_kmers*	*stats_kmers*
1	Precision	0,74	**0,91**	**0,91**
	Recall	0,70	**1,00**	**1,00**
2	Precision	0,55	**1,00**	**1,00**
	Recall	0,55	**1,00**	**1,00**
3	Precision	**1,00**	0,55	**1,00**
	Recall	**1,00**	0,60	**1,00**
4	Precision	0,45	0,67	**0,83**
	Recall	0,45	**1,00**	**1,00**

Примечание. Полужирным шрифтом выделено значение лучшего метода для каждой задачи и метрики.

Два метода позволяют извлекать из данных опорные *k*-меры, которые затем используются для построения графа де Брейна и извлечения интерпретируемых признаков. Третий метод извлекает признаки из наборов графов де Брейна, которые используются для построения классификационных моделей. Для вычислительных экспериментов были промоделированы данные, которые отражают специфику реальных метагеномных образцов. Часто отличия между классами образцов заключаются в штаммах одного вида бактерии, один из которых является нейтральным для человека, а второй — условно-патогенным и содержит мутации, которые коррелируют с наличием заболевания. Также отличия между классами образцов могут заключаться в относительной представленности видов внутри образца, что было промоделировано в наборах данных 3 и 4.

Результаты показали, что использование признаков, полученных с помощью предложенных методов, повышает качество классификации образцов по сравнению с моделями, обученными с использованием классических признаков таксономической аннотации Kraken2. Разработанные методы могут быть применены к открытым данным метагеномного секвенирования для формулирования биологических гипотез о взаимосвязях между составом микробиоты кишечника человека и различными заболеваниями, которые затем могут быть экспериментально проверены.

Заключение

В работе предложены методы извлечения *k*-меров и признаков из наборов метагеномных данных, использующие информацию о классах образцов. Предложено три метода, два из которых отбирают опорные для наборов метагеномных образцов *k*-меры, а третий объединяет опорные *k*-меры в признаки в виде подграфов графа де Брейна. Использование информации о классах образцов на этапе построения признаков позволяет повысить точность работы классификационных моделей, обучаемых на извлекаемых признаках. Объединение *k*-меров в более длинные признаки и извлечение ветвящихся подграфов позволяет повысить интерпретируемость признаков.

Предложенные методы имеют важное значение — они поддерживают работу с наборами метагеномных данных с неограниченным числом классов. Метод извлечения признаков может быть использован для классификации метагеномных данных различной природы и быть основой методов поддержки принятия решений, например, при диагностировании заболеваний людей на основе данных секвенирования микробиоты кишечника.

Литература

1. Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome // Nature Reviews Microbiology. 2017. V.15. N 10. P. 579–590. https://doi.org/10.1038/nrmicro.2017.87
2. Garner R.E., Kraemer S.A., Onana V.E., Fradette M., Varin M.P., Huot Y., Walsh D.A. A genome catalogue of lake bacterial diversity and its drivers at continental scale // Nature Microbiology. 2023. V. 8. N 10. P. 1920–1934. https://doi.org/10.1038/s41564-023-01435-6
3. Huttenhower C., Gevers D., Knight R., et al. Structure, function and diversity of the healthy human microbiome // Nature. 2012. V. 486. N 7402. P. 207–214. https://doi.org/10.1038/nature11234

References

1. Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology*, 2017, vol. 15, no. 10, pp. 579–590. https://doi.org/10.1038/nrmicro.2017.87
2. Garner R.E., Kraemer S.A., Onana V.E., Fradette M., Varin M.P., Huot Y., Walsh D.A. A genome catalogue of lake bacterial diversity and its drivers at continental scale. *Nature Microbiology*, 2023, vol. 8, no. 10, pp. 1920–1934. https://doi.org/10.1038/s41564-023-01435-6
3. Huttenhower C., Gevers D., Knight R., et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 2012, vol. 486, no. 7402. pp. 207–214. https://doi.org/10.1038/nature11234

4. Olekhnovich E., Ivanov A., Babkina A., Sokolov A., Ulyantsev V., Fedorov D., Ilina E. Consistent stool metagenomic biomarkers associated with the response to melanoma immunotherapy // Msystems. 2023. V. 8. N 2. https://doi.org/10.1128/msystems.01023-22
5. Ivanova V., Chernevskaya E., Vasiluev P., Ivanov A., Tolstoganov I., Shafranskaya D., Ulyantsev V., Korobeynikov A., Razin S., Beloborodova N., et al. Hi-C metagenomics in the ICU: exploring clinically relevant features of gut microbiome in chronically critically ill patients // Frontiers in Microbiology. 2022. V. 12. P. 770323. https://doi.org/10.3389/fmicb.2021.770323
6. Olekhnovich E., Ivanov A., Ulyantsev V., Ilina E. Separation of donor and recipient microbial diversity allows determination of taxonomic and functional features of gut microbiota restructuring following fecal transplantation // Msystems. 2021. V. 6. N 4. P. e00811-21. https://doi.org/10.1128/msystems.00811-21
7. Lloyd-Price J., Arze C., Ananthakrishnan A.N., Schirmer M., Avila-Pacheco J., Poon T.W., Andrews E., Ajami N.J., Bonham K.S., Brislawn C.J., et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases // Nature. 2019. V. 569. N 7758. P. 655–662. https://doi.org/10.1038/s41586-019-1237-9
8. Jie Z., Xia H., Zhong S.-L., Feng Q., Li S., Liang S., Zhong H., Liu Z., Gao Y., Zhao H., et al. The gut microbiome in atherosclerotic cardiovascular disease // Nature Communications. 2017. V. 8. P. 845. https://doi.org/10.1038/s41467-017-00900-1
9. Yu J., Feng Q., Wong S.H., Zhang D., Liang Q., Qin Y., Tang L., Zhao H., Stenvang J., Li Y., et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer // Gut. 2017. V. 66. N 1. P. 70–78. https://doi.org/10.1136/gutjnl-2015-309800
10. Qin J., Li Y., Cai Z., Li S., Zhu J., Zhang F., Liang S., Zhang W., Guan Y., Shen D., et al. A metagenome-wide association study of gut microbiota in type 2 diabetes // Nature. 2012. V. 490. N 7418. P. 55–60. https://doi.org/10.1038/nature11450
11. 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
12. 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
13. Compeau P.E., 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
14. Компо Ф., Певзнер П. Алгоритмы биоинформатики. Москва: ДМК Пресс, 2023. 680 с.
15. Nurk S., Meleshko D., Korobeynikov A., Pevzner P.A. metaSPAdes: new versatile metagenomic assembler // Genome Research. 2017. V. 27. N 5. P. 824–834. https://doi.org/10.1101/gr.213959.116
16. Kolmogorov M., Bickhart D.M., Behsaz B., Gurevich A., Rayko M., Shin S.B., Kuhn K., Yuan J., Polevikov E., Smith T.P., et al. metaFlye: scalable long- read metagenome assembly using repeat graphs // Nature Methods. 2020. V. 17. N 11. P. 103–1110. https://doi.org/10.1038/s41592-020-00971-x
17. Bankevich A., Bzikadze A.V., Kolmogorov M., Antipov D., Pevzner P.A. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads // Nature Biotechnology. 2022. V. 40. N 7. P. 1075–1081. https://doi.org/10.1038/s41587-022-01220-6
18. Meyer F., Fritz A., Deng Z.-L., Koslicki D., Lesker T.R., Gurevich A., Robertson G., Alser M., Antipov D., Beghini F., et al. Critical assessment of metagenome interpretation: the second round of challenges // Nature Methods. 2022. V. 19. N 4. P. 429–440. https://doi.org/10.1038/s41592-022-01431-4
19. Pereira-Marques J., Hout A., Ferreira R. M., Weber M., Pinto-Ribeiro I., Van Doorn L.-J., Knetsch C. W., Figueiredo C. Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis // Frontiers in Microbiology. 2019. V. 10. P. 1277. https://doi.org/10.3389/fmicb.2019.01277
20. Marçais G., Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers // Bioinformatics. 2011. V. 27. N 6. P. 764–770. https://doi.org/10.1093/bioinformatics/btr011
21. Ondov B.D., Treangen T.J., Melsted P., Mallonee A.B., Bergman N., 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

4. Olekhnovich E., Ivanov A., Babkina A., Sokolov A., Ulyantsev V., Fedorov D., Ilina E. Consistent stool metagenomic biomarkers associated with the response to melanoma immunotherapy. *Msystems*, 2023, vol. 8, no. 2. https://doi.org/10.1128/msystems.01023-22
5. Ivanova V., Chernevskaya E., Vasiluev P., Ivanov A., Tolstoganov I., Shafranskaya D., Ulyantsev V., Korobeynikov A., Razin S., Beloborodova N., et al. Hi-C metagenomics in the ICU: exploring clinically relevant features of gut microbiome in chronically critically ill patients. *Frontiers in Microbiology*, 2022, vol. 12, pp. 770323. https://doi.org/10.3389/fmicb.2021.770323
6. Olekhnovich E., Ivanov A., Ulyantsev V., Ilina E. Separation of donor and recipient microbial diversity allows determination of taxonomic and functional features of gut microbiota restructuring following fecal transplantation. *Msystems*, 2021, vol. 6, no. 4. pp. e00811-21. https://doi.org/10.1128/msystems.00811-21
7. Lloyd-Price J., Arze C., Ananthakrishnan A.N., Schirmer M., Avila-Pacheco J., Poon T.W., Andrews E., Ajami N.J., Bonham K.S., Brislawn C.J., et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 2019, vol. 569, no. 7758, pp. 655–662. https://doi.org/10.1038/s41586-019-1237-9
8. Jie Z., Xia H., Zhong S.-L., Feng Q., Li S., Liang S., Zhong H., Liu Z., Gao Y., Zhao H., et al. The gut microbiome in atherosclerotic cardiovascular disease. *Nature Communications*, 2017, vol. 8, pp. 845. https://doi.org/10.1038/s41467-017-00900-1
9. Yu J., Feng Q., Wong S.H., Zhang D., Liang Q., Qin Y., Tang L., Zhao H., Stenvang J., Li Y., et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, 2017, vol. 66, no. 1, pp. 70–78. https://doi.org/10.1136/gutjnl-2015-309800
10. Qin J., Li Y., Cai Z., Li S., Zhu J., Zhang F., Liang S., Zhang W., Guan Y., Shen D., et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 2012, vol. 490, no. 7418, pp. 55–60. https://doi.org/10.1038/nature11450
11. 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
12. 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
13. Compeau P.E., 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
14. Compeau P., Pevzner P. *Bioinformatics Algorithms*. Active Learning Publishers, 2018, 728 p.
15. Nurk S., Meleshko D., Korobeynikov A., Pevzner P.A. metaSPAdes: new versatile metagenomic assembler. *Genome Research*, 2017, vol. 27, no. 5, pp. 824–834. https://doi.org/10.1101/gr.213959.116
16. Kolmogorov M., Bickhart D.M., Behsaz B., Gurevich A., Rayko M., Shin S.B., Kuhn K., Yuan J., Polevikov E., Smith T.P., et al. metaFlye: scalable long- read metagenome assembly using repeat graphs. *Nature Methods*, 2020, vol. 17, no. 11, pp. 103–1110. https://doi.org/10.1038/s41592-020-00971-x
17. Bankevich A., Bzikadze A.V., Kolmogorov M., Antipov D., Pevzner P.A. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nature Biotechnology*, 2022, vol. 40, no. 7, pp. 1075–1081. https://doi.org/10.1038/s41587-022-01220-6
18. Meyer F., Fritz A., Deng Z.-L., Koslicki D., Lesker T.R., Gurevich A., Robertson G., Alser M., Antipov D., Beghini F., et al. Critical assessment of metagenome interpretation: the second round of challenges. *Nature Methods*, 2022, vol. 19, no. 4, pp. 429–440. https://doi.org/10.1038/s41592-022-01431-4
19. Pereira-Marques J., Hout A., Ferreira R. M., Weber M., Pinto-Ribeiro I., Van Doorn L.-J., Knetsch C. W., Figueiredo C. Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Frontiers in Microbiology*, 2019, vol. 10, pp. 1277. https://doi.org/10.3389/fmicb.2019.01277
20. Marçais G., Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, 2011, vol. 27, no. 6, pp. 764–770. https://doi.org/10.1093/bioinformatics/btr011
21. Ondov B.D., Treangen T.J., Melsted P., Mallonee A.B., Bergman N., 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

22. Maillet N., Collet G., Vannier T., Lavenier D., Peterlongo P. COMMET: comparing and combining multiple metagenomic datasets // Proc. of the IEEE international conference on bioinformatics and biomedicine (BIBM). 2014. P. 94–98. https://doi.org/10.1109/BIBM.2014.6999135
23. Rahman A., Hallgrímsdóttir I., Eisen M., Pachter L. Association mapping from sequencing reads using *k*-mers // Elife. 2018. V. 7. P. e32920. https://doi.org/10.7554/eLife.32920
24. Wang Y., Chen Q., Deng C., Zheng Y., Sun F. KmerGO: a tool to identify group-specific sequences with *k*-mers // Frontiers in Microbiology. 2020. V. 11. P. 2067. https://doi.org/10.3389/fmicb.2020.02067
25. Greenwood P.E., Nikulin M.S. A Guide to Chi-Squared Testing. John Wiley & Sons, 1996. 304 p.
26. Крамер Г. Математические методы статистики. М.: Институт компьютерных исследований, 2019. 648 с.
27. Hettmansperger T.P., McKean J.W. Robust nonparametric statistical methods. CRC press, 2010. 554 p.
28. Dunn O.J. Multiple comparisons among means // Journal of the American Statistical Association. 1961. V. 56. N 293. P. 52–64. https://doi.org/10.1080/016f21459.1961.10482090
29. 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
30. Wood D.E., Lu J., Langmead B. Improved metagenomic analysis with Kraken 2 // Genome Biology. 2019. V. 20. N 1. P. 257. https://doi.org/10.1186/s13059-019-1891-0
31. Breiman L. Random forests // Machine Learning. 2001. V. 45. N 1. P. 5–32. https://doi.org/10.1023/A:1010933404324
32. Pedregosa F.,Varoquaux, G., Gramfort, A., Michel, V., et al. Scikit-learn: Machine learning in Python // Journal of Machine Learning Research. 2011. V. 12. P. 2825–2830.
33. 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

22. Maillet N., Collet G., Vannier T., Lavenier D., Peterlongo P. COMMET: comparing and combining multiple metagenomic datasets // *Proc. of the IEEE international conference on bioinformatics and biomedicine (BIBM)*. 2014. pp. 94–98. https://doi.org/10.1109/BIBM.2014.6999135
23. Rahman A., Hallgrímsdóttir I., Eisen M., Pachter L. Association mapping from sequencing reads using *k*-mers. *Elife*, 2018, vol. 7, pp. e32920. https://doi.org/10.7554/eLife.32920
24. Wang Y., Chen Q., Deng C., Zheng Y., Sun F. KmerGO: a tool to identify group-specific sequences with *k*-mers. *Frontiers in Microbiology*, 2020, vol. 11, pp. 2067. https://doi.org/10.3389/fmicb.2020.02067
25. Greenwood P.E., Nikulin M.S. *A Guide to Chi-Squared Testing*. John Wiley & Sons, 1996, 304 p.
26. Cramér H. *Mathematical Methods of Statistics*. Princeton University Press, 2019, 575 p.
27. Hettmansperger T.P., McKean J.W. *Robust nonparametric statistical methods*. CRC press, 2010, 554 p.
28. Dunn O.J. Multiple comparisons among means. *Journal of the American Statistical Association*, 1961, vol. 56, no. 293, pp. 52–64. https://doi.org/10.1080/016f21459.1961.10482090
29. 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
30. Wood D.E., Lu J., Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 2019, vol. 20, no. 1, pp. 257. https://doi.org/10.1186/s13059-019-1891-0
31. Breiman L. Random forests. *Machine Learning,* 2001, vol. 45, no. 1, pp. 5–32. https://doi.org/10.1023/A:1010933404324
32. Pedregosa F.,Varoquaux, G., Gramfort, A., Michel, V., et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011, vol. 12, pp. 2825–2830.
33. 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 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., Chief Researcher, Full Professor, 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 18.03.2025
Одобрена после рецензирования 26.04.2025
Принята к печати 27.05.2025

Received 18.03.2025
Approved after reviewing 26.04.2025
Accepted 27.05.2025