ПРИБОРЫ, СИСТЕМЫ И ИЗДЕЛИЯ МЕДИЦИНСКОГО НАЗНАЧЕНИЯ

MEDICAL DEVICES, SYSTEMS AND PRODUCTS

УДК 004.89, 004.413.5, 519.724
doi: 10.21685/2307-5538-2024-4-16

СОВРЕМЕННАЯ КЛАССИФИКАЦИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ АНАЛИЗА БИОИНФОРМАЦИОННЫХ ДАННЫХ ГЕНОМНОЙ ПРИРОДЫ И НОВЫЕ КОНЦЕПЦИИ ИНТЕРПРЕТАЦИИ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

М. В. Спринджук[1], В. И. Берник[2], А. П. Кончиц[3]

[1] Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь
[2] Институт математики Национальной академии наук Беларуси, Минск, Республика Беларусь
[3] Институт леса Национальной академии наук Беларуси, Гомель, Республика Беларусь
[1] stepanenkomatvei@yandex.ru, [2] bernik@im.bas-net.by, [3] konchits@yandex.ru

Аннотация. *Актуальность и цели.* Актуальность темы обусловлена колоссальным ростом объемов накопленной и недообработанной информации медицинского назначения, эволюцией математического и алгоритмического аппарата, необходимостью усовершенствования существующих конвейеров обработки данных. Целью работы было сообщение и рассмотрение предложенной авторами новой классификации и концепций интерпретации данных, а также опыта разработки программного обеспечения. *Материалы и методы.* Рассматриваются вопросы классификации программного обеспечения для автоматизации процесса анализа биоинформационных данных геномной природы и новые принципы интерпретации вычислительных экспериментов в геномике. *Результаты и выводы.* Как результат в статье представлен краткий обзор литературы по теме современной вычислительной геномики, биоинформатики, математической биологии, медицинской кибернетики, больших данных для медицины и биологии. секвенированию нуклеиновых кислот, опыт разработки конвейеров для анализа геномных данных, новая классификация программного обеспечения и концепции интерпретации данных в этой сложной междисциплинарной многокомпонентной предметной области.

Ключевые слова: медицинская кибернетика, системы медицинского назначения, геномика, представление и отображение информации, интерпретация данных, большие данные, биоинформатика, прикладная математика, искусственный интеллект, машинное обучение, судебная экспертиза ДНК (дезоксирибонуклеиновой кислоты)

Финансирование: исследование выполнено при поддержке гранта ISTC PR150 "Belarus TB Database and TB Poland".

Для цитирования: Спринджук М. В., Берник В. И., Кончиц А. П. Современная классификация программного обеспечения для анализа биоинформационных данных геномной природы и новые концепции интерпретации вычислительных экспериментов // Измерение. Мониторинг. Управление. Контроль. 2024. № 4. С. 139–148. doi: 10.21685/2307-5538-2024-4-16

MODERN CLASSIFICATION OF SOFTWARE FOR THE ANALYSIS OF BIOINFORMATION DATA OF GENOMIC NATURE AND NOVEL CONCEPTS FOR THE INTERPRETATION OF COMPUTATIONAL EXPERIMENTS

M.V. Sprindzuk[1], V.I. Bernik[2], A.P. Konchits[3]

[1] United Institute of Informatics Problems of the Belarus National Academy of Sciences, Minsk, Republic of Belarus
[2] Institute of Mathematics of the National Academy of Sciences of Belarus, Minsk, Republic of Belarus
[3] Forest Institute of the National academy of sciences of Belarus, Gomel, Republic of Belarus
[1] stepanenkomatvei@yandex.ru, [2] bernik@im.bas-net.by, [3] konchits@yandex.ru

Abstract. *Background.* The emergence and importance of the research topic is substantiated by the colossal growth in the volume of accumulated and underprocessed medical information, the evolution of mathematical and algorithmic apparatus, and the need to improve existing data processing pipelines. The purpose of the work was to review and report the new classification and concepts of data interpretation proposed by the authors, as well as an experience in software development. *Materials and methods.* The questions of classification of software for automating the process of analyzing bioinformational data of the genomic nature and new the principles for interpreting computational experiments in genomics are discussed. *Results and conclusions.* As a result, the article presents a brief review of the literature on the topic of modern computational genomics, bioinformatics, mathematical biology, medical cybernetics, big data for medicine and biology, nucleic acid sequencing technologies, experience in developing pipelines for the analysis of genomic data, the new classification of software and concepts for interpreting data in this complex interdisciplinary multi-component subject area.

Keywords: medical cybernetics, medical systems, genomics, presentation and display of information, data interpretation, big data, bioinformatics, applied mathematics, artificial intelligence, machine learning, forensic DNA

Financing: the study was supported by the ISTC PR150 grant "Belarus TB Database and TB Poland".

For citation: Sprindzuk M.V., Bernik V.I., Konchits A.P. Modern classification of software for the analysis of bioinformation data of genomic nature and novel concepts for the interpretation of computational experiments. *Izmerenie. Monitoring. Upravlenie. Kontrol' = Measuring. Monitoring. Management. Control.* 2024;(4):139–148. (In Russ.). doi: 10.21685/2307-5538-2024-4-16

Введение

Биоинформатика как междисциплинарная ветвь генетики и кибернетики претерпела долгий путь эволюционного развития от недоверия и обвинений в отсутствии истины до современных успехов в изучении генома человека, микробов и растений (рис. 1).

Рис. 1. Основные сферы применения технологий секвенирования следующего поколения (ССП; NGS – next generation sequencing), ключевых технических средств современной генетики и биоинформатики

Наиболее яркими примерами успешного применения геномики и биоинформатики является получение доказательств и последующее раскрытие преступлений против жизни и здоровья на основе вычислительного анализа ДНК (дезоксирибонуклеиновой кислоты), идентификации по биологическим следам непосредственно со сцены преступления, как правило, убийства

и прочего насилия [1]. Актуальность геномических исследований также связана с ростом опасных вирусных инфекций, наследственной патологии, онкологии (рис. 2, 3).

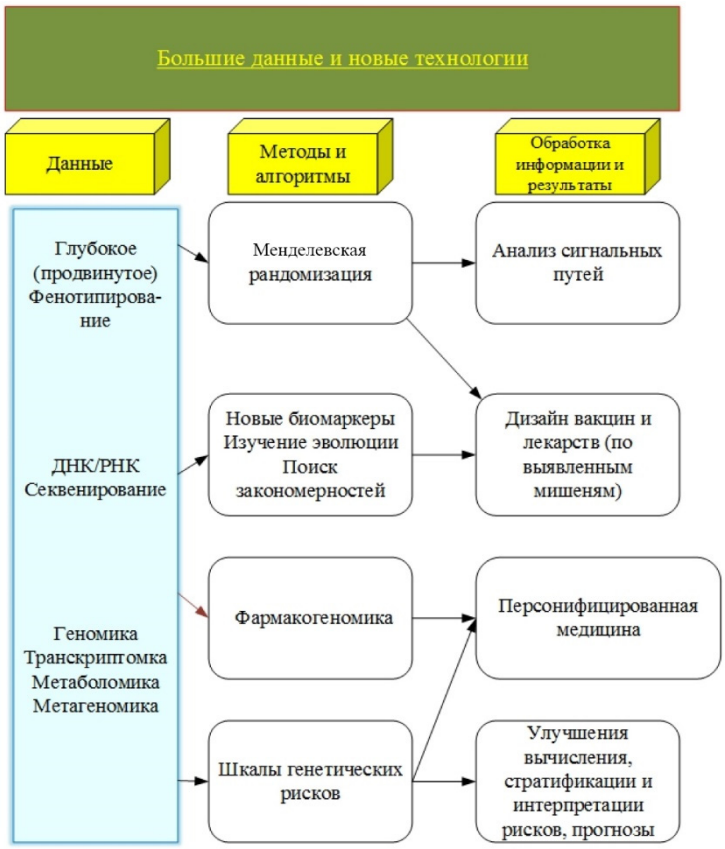

Рис. 2. Сферы применения технологий анализа больших биоинформационных данных геномной природы для персонифицированной медицины и фармакологии [2]

Рис. 3. Слайд-диаграмма применения технологий искусственного интеллекта для разработки и оценки эффективности новых вакцин и лекарственных средств [3]

Секвенирование (рис. 1) – это процесс определения порядка положения нуклеотидов в молекуле ДНК. Открытия, совершенные в 70-х гг. прошлого века Сэнгером и Максамом – Гилбертом, позволили осуществить технологический прорыв, сделав секвенирование рабочим инструментом для большого количества исследователей в биологии, медицине, криминалистике, лесном и сельском хозяйствах, пищевой промышленности и т.д. За последние 60 лет секвенирование ДНК значительно продвинулось с точки зрения приборостроения, программных инструментов и методик анализа информации. Секвенирование ДНК – экспериментальный метод определения последовательного расположения оснований нуклеиновых кислот (A, T, G и C) в полинуклеотиде, кодирующем различные белки, которые функционируют в клетке. Полный набор кодирующих и некодирующих последовательностей в ДНК живого организма называется геномом. Геном имеет информацию обо всех белках, необходимых для существования и функционирования жизни организма. Биологические последовательности (секвенции нуклеотидов, аминокислот, белков) показывают некоторое сходство между собой. Это можно определить путем поиска гомологичности между последовательностями.

Технологии метода химической деградации, предложенные Максамом и Гилбертом, метод дидезокси-терминации цепи, разработанный командой Sanger в 1977 г., и автоматическое секвенирование с помеченной флуоресценцией в 1990-х гг. вместе сформировали первое поколение секвенирования (FGS – first generation sequencing). Благодаря своей сравнительной простоте метод Сэнгера стал доминирующим методом в FGS. Секвенирование Сэнгера позволило секвенировать бактериофаг PhiX 174, который содержит приблизительно 5375 нуклеотидов. Это исследование стало первым полностью секвенированным геномом в 1977 г. В 2003 г. международный проект консорциума «Геном человека» (HGP – human genome project) успешно просеквенировал и картировал весь геном человека, что произошло после 13 лет исследований во многих лабораториях мира. Секвенирование второго поколения (SGS, second generation sequencing) или секвенирование следующего поколения (NGS – next generation sequencing) относится к высокопроизводительным технологиям секвенирования ДНК и РНК (рибонуклеиновой кислоты), которые могут секвенировать даже миллиарды нитей нуклеиновых кислот. Процесс идентификации последовательности происходит с использованием репликации или амплификации, обеспечивающей значительную пропускную способность и многократное секвенирование целевых участков. Секвенирование третьего поколения (TGS – third generation sequencing) характеризуется обработкой нуклеотидов по одному для получения длинных и точных результатов секвенирования, в то время как технология амплификации не используется. Одноклеточное секвенирование также относится к технологии TGS [4]. Секвенирование ДНК и РНК на сегодняшний день выполняется всего несколькими технологиями, описание которых приводится далее в тексте статьи.

Принцип, преимущества и недостатки технологии Illumina/Solexa

Технология секвенирования с помощью синтеза (SBS – synthesis based sequencing) Illumina/Solexa основана на методе реверсивной терминации (reversible termination). Секвенирование Illumina путем синтеза состоит из четырех основных этапов: подготовка образца, генерация кластеров, секвенирование и анализ данных [4]. Одним из главных преимуществ технологии SBS является то, что со стандартными реагентами она позволяет секвенировать до 100 образцов за цикл работы. Также SBS-технология имеет явное преимущество при секвенировании гомополимерных последовательностей по сравнению с 454 или Ion Torrent, поскольку позволяет включать один нуклеотид на реакцию. Значительным преимуществом данной платформы является возможность парноконцевого чтения последовательностей ДНК. Одним из основных недостатков технологии SBS остается ограничение длины прочтений, особенно когда речь идет о задачах расшифровки последовательности de novo. Ошибки замещения появляются из-за увеличения фонового шума в каждом цикле, гомополимеров, смещение GC-ratio (соотношения гуанина и цитозина) в ходе мостиковой амплификации. Эта технология секвенирования на основе синтеза ДНКа нитей. Отличается высокой пропускной способностью с высокой точностью и воспроизводимостью [5].

Технология Ion Torrent

Платформа Ion Torrent – это первая технология, которая не применяет оптические датчики. В этом методе используется полупроводниковая система обнаружения при секвенировании последовательности, основанная на обнаружении ионов водорода, которые являются

побочными продуктами добавления нуклеотидов к шаблонной цепи во время полимеризации. Бусины, содержащие обогащенную ДНК, добавляются в микроячейку на чипе. Основным преимуществом Ion Torrent секвенирования является то, что оно использует относительно простую химию в процессе секвенирования и требует сравнительно очень небольшого размера образца для анализа. Отсюда – высокая скорость секвенирования при низких эксплуатационных расходах. В качестве недостатков следует отметить наличие значительного количества ошибок секвенирования в виде однонуклеотидных вставок и делеций. Для решения этой проблемы Life Technologies выпустила обновление программного продукта Ion Reporter. Вторым недостатком этой системы является короткая длина читаемого фрагмента по сравнению с другими методами секвенирования, такими как секвенирование по Сэнгеру или пиросеквенирование. Большие длины читаемых фрагментов полезны особенно для сборки генома de novo [26], что обеспечивает быстрое, достаточно точное и удобное в использовании секвенирование. Качество секвенирования может быть ухудшено наличием гомополимеров и повторов [5].

Мобильное оборудование секвенирования Oxford Nanopore

Принцип работы секвенаторов ONT (Oxford nanopore technology) основан на измерении электрического тока при прохождении молекулы нуклеиновой кислоты через нанопору. Обнаружение осуществляется в камере с разделенной мембраной, содержащей нанопору. К камере прикладывается электрическое напряжение, заставляющее ДНК или РНК двигаться через пору. По мере прохождения молекулы сечение поры уменьшается, в результате чего уменьшается сила тока. Таким образом, измеряя ток, можно определить тип нуклеотида, проходящего через пору в заданный интервал времени. По сравнению с существующими методами секвенирования использование данного метода секвенирования имеет такие преимущества, как низкая стоимость и доступность использования, высокая чувствительность, высокая длина считывания (до десятков тысяч оснований), высокая портативность, быстрый анализ и отображение результатов в реальном времени. К недостаткам можно отнести такие свойства, как низкое качество считываний по сравнению с другими платформами [4]. Частота ошибок секвенирования может быть высокой, особенно при повторах прочтений и наличии высокогомологичных регионов. Кроме того, качество секвенирования может быть затронуто наличием примесей [5].

PacBio (Pacific Bioscience)

Высококачественная технология секвенирования, которая может генерировать длинную цепочку ДНК последовательности с высоким качеством, позволяет обеспечить распознавание сложных геномных регионов и точное выявление генетических вариантов. Способна осуществлять обнаружение эпигенетических модификаций. Самое дорогое по цене секвенирование требует большого количества данных для генерации надежных результатов [5].

Классификация программного обеспечения для анализа биоинформационных данных геномной природы

Компьютерная программа представляет собой набор команд электронной машине и набор обрабатываемых данных. На сегодняшний день в медицинской кибернетике и биоинформатике доминируют скриптовые языки программирования Python, R, Shell. Тем не менее крупные программно-вычислительные комплексы создаются на основе ресурсов языков программирования Java, C++, C#, часто используют несколько языков программирования и имеют привязку к ресурсам интернета.

Современное программное обеспечение, предназначенное для анализа биоинформационных данных геномной природы (рис. 4) [6–8], можно классифицировать, обобщая свои знания и опыт изучения, разработки и применения автоматизированных систем анализа данных, следующим образом:

1. Программное обеспечение, исполняемое внутри специальной облачной платформы Galaxy, KNIME (Konstanz information miner), наиболее эффективно для быстрой разработки конвейеров анализа данных при условии владения разработчиком соответствующей технологии.

2. Настольное программное обеспечение имеет графический интерфейс пользователя к выполнению команд для манипуляции данными. Оно может иметь привязку к Интернету (вариант А) или быть независимым в своих функциях от Интернета (вариант Б).

3. Полностью веб-приложение, работающее на высокопроизводительном сервере хостинга, имеющее соответствующий веб-интерфейс из HTML (язык гипертекстовой разметки, hypertext markup language) форм, позволяющее загружать данные и выводить результаты обработки информации с функциями загрузки отчетов, пересылки, визуализации результатов в браузере.

4. Безынтерфейсные скрипты. Они могут представлять собой многофункциональную самостоятельную компьютерную программу или быть элементом программно-вычислительного комплекса. Широко используются для учебных и практических задач.

5. Конвейерные технологии, имеющие специальный семантический язык программирования. Наиболее распространены для целей вычислительной геномики и транскриптомики SnakeMake [9] и NextFlow [10].

Рис. 4. Программное обеспечение для анализа биоинформационных данных геномной природы

Дополнительно можно номенклатурно классифицировать программное обеспечение для анализа биоинформационных данных геномной природы по следующим критериям-признакам:

– применение технологий контейнеризации (да/нет);

– применение и реализации технологий искусственного интеллекта (да/нет);

– по количеству используемых в программной реализации языков программирования;

– по наличию коммуникации с мобильными устройствами связи и передачи и информации (да/нет);

– по кроссплатформенности (да/нет);

– по возможности обрабатывать так называемые большие и грязные данные (да/нет).

На рис. 5 представлен пример современного программного обеспечения: разработанный авторами конвейер обработки геномных текстов микобактерии туберкулеза, предназначенный для обнаружения устойчивости к современным лекарственным антитуберкулезным медикаментам.

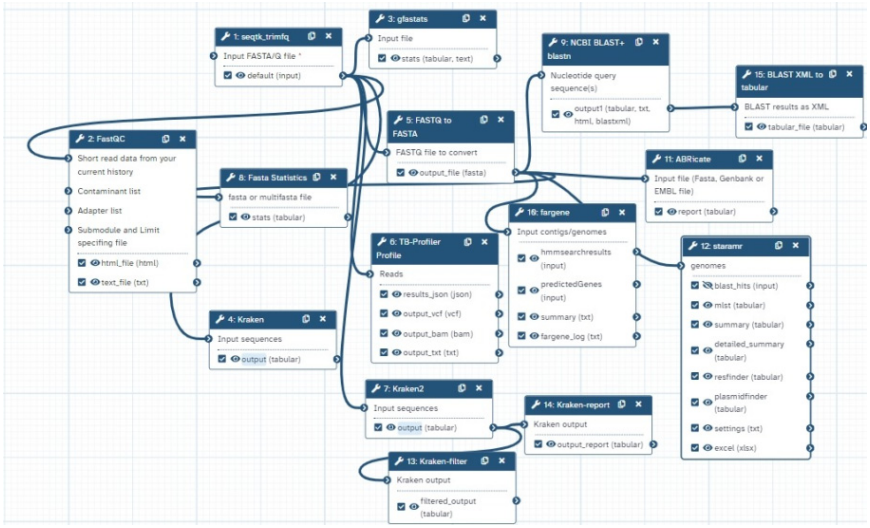

Рис. 5. Автоматический конвейер, предназначенный для быстрой оценки геномов *M. Tuberculosis* на предмет лекарственной устойчивости и вирулентности

Биоинформационная технология Galaxy – это свободно доступная платформа, предоставляющая сегодня возможность инженерам-программистам, биоинформатикам выбирать и включать лучшее доступное программное обеспечение для геномики в собственные конвейеры и в конечном итоге интегрировать конвейеры в автоматизированные облачные системы обработки данных. В таких конвейерах обработки геномной информации можно использовать современные инструменты вычислительной микробиологии, предназначенные для анализа данных секвенирования следующего поколения (коротких и длинных прочтений, их гибридов), включая TbProfiler [11, 12] Snippy [https://github.com/tseemann/snippy], Staramr [13], ResFinder [14–18], VirulenceFinder [19, 26], Flye [20], Medaka [https://github.com/nanoporetech/medaka], Kraken [7, 12], BLAST (basic local alignment search tool, базовый инструмент поиска сходства последовательностей, основанный на локальном выравнивании) [23], Prokka [26] и многие другие.

Концепции интерпретации данных

Биоинформационные вычислительные эксперименты генерируют множество разнородной по составу и большой по объему информации. Поэтому имеется интерес и необходимость изучения теоретических основ интерпретации таких данных. На основе многолетнего личного опыта и данных литературы [24, 25] авторы сформулировали концепции интерпретации биоинформационных данных геномной природы (рис. 6).

Рис. 6. Концепции интерпретации биоинформационных данных геномной природы

1. Концепция генерации или присвоения терминов на основе сравнения и сопоставления с базами данных. На этом принципе построены системы номенклатуры NextClade SARS-CoV-2 (Severe acute respiratory syndrome-related coronavirus-2; Коронавирус-2, связанный с тяжелым острым респираторным синдром) и программное обеспечение для аннотирования геномных или транскриптомных текстов. Новые термины могут генерироваться по заданным правилам накопления мутаций как отличий от референсного генома или их множества.

2. Концепция выделения доминирующих сигналов из множества генерируемых по заданным правилам фильтрации, объединения, элиминации, трансформации. На таком принципе, например, построено программное обеспечения генерации отчетов о мутационном профиле в результате выравнивания и запроса вариантов.

3. Концепция презумции ошибочности и несоответствия. Под этим подразумевается необходимость понимания и критического анализа каскада генерации данных и вероятность наличия в данных ошибок различной природы, в том числе по причине наличия ошибок в референсных базах данных и экспертных знаний, в рандомной природе вычислительных экспериментов с нейронными сетями и другими алгоритмами машинного обучения. История генетики и биоинформатики полна фактов первоначального недоверия и известных экспертных ошибок, особенно в криминалистике ДНК-идентификации, но с развитием технологий появилось множество доказательных фактов об успешности применения биоинформационного анализа данных геномной природы.

4. Концепция отображения результата в виде графической информации в двумерной или трехмерной плоскостях. Примером может послужить точечный график сравнения двух геномных текстов.

5. Концепция управляемости результатами анализа данных. В настоящее время выбор алгоритмов и программного обеспечения для анализа геномных данных огромен, он уменьшается при конкретизации целей и задач исследования, но остается еще и следующая ступень выбора комбинации настроек программного обеспечения.

6. Концепция недостатка и несбалансированности наборов данных, ее неполноты и избыточности, грязных данных. Имеет общее с концепцией № 3, но в первую очередь подразумеваются дефекты не баз данных, а входной информации. Исследователю необходимо осознавать дефекты информации, подлежащей критическому анализу.

7. Концепция эволюции, нестабильности и изменчивости информации.

8. Концепция математического представления данных. С появлением нового математического аппарата расширяются реальные возможности представления и интерпретации.

Заключение

Технологии секвенирования ДНК и РНК развиваются вместе с достижениями биохимии, биофизики, кибернетики, математики и биологии и сегодня представляют современный арсенал технических средств для изучения основы живой природы для решения междисциплинарных задач. Снижение стоимости секвенирования и компьютеров приводит к широкому распространению и доступности данной технологии. В данной статье приводится классификация программного обеспечения для вычислительной биологии и геномики, а также обсуждаются сформулированные авторами концепции интерпретации результатов вычислительных экспериментов в данной предметной области. Надеемся, что предложенные концепции найдут свое применение и развитие не только в биоинформатике и медицинской кибернетике. Литературы по разработке и классификации программного обеспечения для науки не так много, правильная организация и упорядочивание такой информации будут полезны для современных программистов, занимающихся усовершенствованием информационных технологий. Аналитическое обобщение и номенклатура современных компьютерных технологий необходимы для возможности адекватного выбора реальных технических средств разработки информационных систем.

Список литературы

1. Primorac D., Schanfield M. Forensic DNA applications: An interdisciplinary perspective. CRC Press, 2023. 506 p.
2. Flynn J. T. Pediatric hypertension. N. Y. : Springer Berlin Heidelberg, 2018. 1005 p.
3. Singh T. R., Saini H., Junior M. C. Bioinformatics and Computational Biology: Technological Advancements, Applications and Opportunities. CRC Press, 2023. 350 p.
4. Бородинов А., Манойлов В., Заруцкий И. [и др.]. Поколения методов секвенирования ДНК (обзор) // Научное приборостроение. 2020. Vol. 30, № 4. P. 3–20.
5. Martinez-Carranza J., Inzunza-Gonzalez E., Garcia-Guerrero E. E., Tlelo-Cuautle E. Machine Learning for Complex and Unmanned Systems. Boca Raton : CRC Press, 2024. 382 p.
6. Leipzig J. A review of bioinformatic pipeline frameworks // Briefings in bioinformatics. 2017. Vol. 18, № 3. P. 530–536.
7. Wratten L., Wilm A., Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers // Nature methods. 2021. Vol. 18, № 10. P. 1161–1168.
8. Yang A., Troup M., Ho J.W. Scalability and validation of big data bioinformatics software // Computational and structural biotechnology journal. 2017. Vol. 15. P. 379–386.
9. Köster J., Rahmann S. Snakemake–a scalable bioinformatics workflow engine // Bioinformatics. 2012. Vol. 28, № 19. P. 2520–2522.
10. Di Tommaso P., Chatzou M., Floden E. W. [et al.]. Nextflow enables reproducible computational workflows // Nature biotechnology. 2017. Vol. 35, № 4. P. 316–319.
11. Mahe P., El Azami M., Barlas P., Tournoud M. A large scale evaluation of TBProfiler and Mykrobe for antibiotic resistance prediction in Mycobacterium tuberculosis // Peer J. 2019. Vol. 7. P. e6857.
12. Verboven L., Phelan J., Heupink T. H., Van Rie A. Correction: TBProfiler for automated calling of the association with drug resistance of variants in Mycobacterium tuberculosis // PLoS One. 2023. Vol. 18, № 10. P. e0293254.

13. Bharat A., Petkau A., Avery B. P. [et al.]. Correlation between Phenotypic and In Silico Detection of Antimicrobial Resistance in Salmonella enterica in Canada Using Staramr // Microorganisms. 2022. Vol. 10, № 2. P. 292.
14. Bortolaia V., Kaas R. S., Ruppe E. [et al.]. ResFinder 4.0 for predictions of phenotypes from genotypes // J Antimicrob Chemother. 2020. Vol. 75, № 12. P. 3491–3500.
15. Florensa A. F., Kaas R. S., Clausen P. [et al.]. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes // Microb Genom. 2022. Vol. 8, № 1. P. 000748.
16. Kleinheinz K. A., Joensen K. G., Larsen M. V. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences // Bacteriophage. 2014. Vol. 4, № 1. P. e27943.
17. Verschuuren T., Bosch T., Mascaro V. [et al.]. External validation of WGS-based antimicrobial susceptibility prediction tools, KOVER-AMR and ResFinder 4.1, for Escherichia coli clinical isolates // Clin Microbiol Infect. 2022. Vol. 28, № 11. P. 1465–1470.
18. Zankari E. Comparison of the web tools ARG-ANNOT and ResFinder for detection of resistance genes in bacteria // Antimicrob Agents Chemother. 2014. Vol. 58, № 8. P. 4986.
19. Roer L., Kaya H., Tedim A. P. [et al.]. VirulenceFinder for Enterococcus faecium and Enterococcus lactis: an enhanced database for detection of putative virulence markers by using whole-genome sequencing data // Microbiol Spectr. 2024. № 1. P. e0372423.
20. Freire B., Ladra S., Parama J. R. Memory-Efficient Assembly Using Flye // IEEE/ACM Trans Comput Biol Bioinform. 2022. Vol. 19, № 6. P. 3564–3577.
21. Gomi R., Wyres K. L., Holt K. E. Detection of plasmid contigs in draft genome assemblies using customized Kraken databases // Microb Genom. 2021. Vol. 7, № 4. P. 000550.
22. Lu J., Rincon N., Wood D. E. [et al.]. Metagenome analysis using the Kraken software suite // Nat Protoc. 2022. Vol. 17, № 12. P. 2815–2839.
23. Piccoli C., Munoz-Merida A., Crottini A. PARSID: a Python script for automatic analysis of local BLAST results for a rapid molecular taxonomic identification // BMC Res Notes. 2024. Vol. 17, № 1. P. 35.
24. Meakin G. E., Kokshoorn B., van Oorschot R. A., Szkuta B. Evaluating forensic DNA evidence: Connecting the dots // Wiley Interdisciplinary Reviews: Forensic Science. 2021. Vol. 3, № 4. P. e1404.
25. Pope, S., Puch-Solis, R. Interpretation of DNA data within the context of UK forensic science investigation // Emerging Topics in Life Sciences. 2021. Vol. 5, № 3. P. 395–404.
26. Seemann T. Prokka: rapid prokaryotic genome annotation // Bioinformatics. 2014. Vol. 30, № 14. P. 2068–2069.

References

1. Primorac D., Schanfield M. *Forensic DNA applications: An interdisciplinary perspective*. CRC Press, 2023:506.
2. Flynn J.T. *Pediatric hypertension*. New York: Springer Berlin Heidelberg, 2018:1005.
3. Singh T.R., Saini H., Junior M.C. *Bioinformatics and Computational Biology: Technological Advancements, Applications and Opportunities*. CRC Press, 2023:350.
4. Borodinov A., Manoylov V., Zarutskiy I. et al. Generations of DNA sequencing methods (review). *Nauchnoe priborostroenie = Scientific instrumentation*. 2020;30(4):3–20.
5. Martinez-Carranza J., Inzunza-Gonzalez E., Garcia-Guerrero E.E., Tlelo-Cuautle E. *Machine Learning for Complex and Unmanned Systems*. Boca Raton: CRC Press, 2024:382.
6. Leipzig J. A review of bioinformatic pipeline frameworks. *Briefings in bioinformatics*. 2017;18(3): 530–536.
7. Wratten L., Wilm A., Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature methods*. 2021;18(10):1161–1168.
8. Yang A., Troup M., Ho J.W. Scalability and validation of big data bioinformatics software. *Computational and structural biotechnology journal*. 2017;15:379–386.
9. Köster J., Rahmann S. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520–2522.
10. Di Tommaso P., Chatzou M., Floden E.W. et al. Nextflow enables reproducible computational workflows. *Nature biotechnology*. 2017;35(4):316–319.
11. Mahe P., El Azami M., Barlas P., Tournoud M. A large scale evaluation of TBProfiler and Mykrobe for antibiotic resistance prediction in Mycobacterium tuberculosis. *Peer J*. 2019;7:e6857.
12. Verboven L., Phelan J., Heupink T.H., Van Rie A. Correction: TBProfiler for automated calling of the association with drug resistance of variants in Mycobacterium tuberculosis. *PLoS One*. 2023;18(10):e0293254.
13. Bharat A., Petkau A., Avery B.P. et al. Correlation between Phenotypic and In Silico Detection of Antimicrobial Resistance in Salmonella enterica in Canada Using Staramr. *Microorganisms*. 2022;10(2):292.
14. Bortolaia V., Kaas R. S., Ruppe E. et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother*. 2020;75(12):3491–3500.

15. Florensa A.F., Kaas R.S., Clausen P. et al. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb Genom*. 2022;8(1):000748.
16. Kleinheinz K.A., Joensen K.G., Larsen M.V. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage*. 2014;4(1):e27943.
17. Verschuuren T., Bosch T., Mascaro V. et al. External validation of WGS-based antimicrobial susceptibility prediction tools, KOVER-AMR and ResFinder 4.1, for Escherichia coli clinical isolates. *Clin Microbiol Infect*. 2022;28(11):1465–1470.
18. Zankari E. Comparison of the web tools ARG-ANNOT and ResFinder for detection of resistance genes in bacteria. *Antimicrob Agents Chemother*. 2014;58(8):4986.
19. Roer L., Kaya H., Tedim A. P. et al. VirulenceFinder for Enterococcus faecium and Enterococcus lactis: an enhanced database for detection of putative virulence markers by using whole-genome sequencing data. *Microbiol Spectr*. 2024;(1):e0372423.
20. Freire B., Ladra S., Parama J.R. Memory-Efficient Assembly Using Flye. *IEEE/ACM Trans Comput Biol Bioinform*. 2022;19(6):3564–3577.
21. Gomi R., Wyres K.L., Holt K.E. Detection of plasmid contigs in draft genome assemblies using customized Kraken databases. *Microb Genom*. 2021;7(4):000550.
22. Lu J., Rincon N., Wood D.E. et al. Metagenome analysis using the Kraken software suite. *Nat Protoc*. 2022;17(12):2815–2839.
23. Piccoli C., Munoz-Merida A., Crottini A. PARSID: a Python script for automatic analysis of local BLAST results for a rapid molecular taxonomic identification. *BMC Res Notes*. 2024;17(1):35.
24. Meakin G.E., Kokshoorn B., van Oorschot R.A., Szkuta B. Evaluating forensic DNA evidence: Connecting the dots. *Wiley Interdisciplinary Reviews: Forensic Science*. 2021;3(4):e1404.
25. Pope S., Puch-Solis R. Interpretation of DNA data within the context of UK forensic science investigation. *Emerging Topics in Life Sciences*. 2021;5(3):395–404.
26. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068–2069.

Информация об авторах / Information about the authors

Матвей Владимирович Спринджук
кандидат технических наук,
старший научный сотрудник,
Объединенный институт проблем информатики
Национальной академии наук Беларуси
(Республика Беларусь, г. Минск, ул. Сурганова, 6)
E-mail: stepanenkomatvei@yandex.ru

Matvey V. Sprindzhuk
Candidate of technical sciences,
senior scientist,
United Institute of Informatics Problems
of the Belarus National Academy of Sciences
(6 Surganova street, Minsk, Republic of Belarus)

Василий Иванович Берник
доктор физико-математических наук, профессор,
главный научный сотрудник,
Институт математики Национальной
академии наук Беларуси
(Республика Беларусь, г. Минск, ул. Сурганова, 11)
E-mail: bernik@im.bas-net.by

Vasiliy I. Bernik
Doctor of physical and mathematical sciences,
professor, chief researcher,
Institute of Mathematics of the National Academy
of Sciences of Belarus
(11 Surganova street, Minsk, Republic of Belarus)

Андрей Петрович Кончиц
кандидат биологических наук,
ведущий научный сотрудник,
Институт леса Национальной
академии наук Беларуси
(Республика Беларусь, г. Гомель,
ул. Пролетарская, 71)
E-mail: konchits@yandex.ru

Andrey P. Konchits
Candidate of biological sciences,
leading researcher,
Forest Institute of the National academy
of sciences of Belarus,
(71 Proletarskaya street, Gomel, Republic of Belarus)

Авторы заявляют об отсутствии конфликта интересов /
The authors declare no conflicts of interests.

Поступила в редакцию/Received 20.07.2024
Поступила после рецензирования/Revised 12.08.2024
Принята к публикации/Accepted 09.09.2024