Математическая биология и биоинформатика
2017. Т. 12. № 1. С. 102–119. doi: 10.17537/2017.12.102

==========ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ ==========
==========ТЕХНОЛОГИИ В БИОЛОГИИ И МЕДИЦИНЕ==========

УДК: 004.9:004.9:004.8:577.21

Большие данные в биоинформатике

Назипова Н.Н.[1], Исаев Е.А.[*2], Корнилов В.В.[2], Первухин Д.В.[2], Морозова А.А.[3], Горбунов А.А.[2], Устинин М.Н.[1]

[1]*Институт математических проблем биологии РАН – филиал Федерального государственного учреждения "Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук", Пущино*
[2]*Национальный исследовательский университет «Высшая школа экономики», Москва*
[3]*Союз предприятий Центральное научно-производственное объединение «КАСКАД», Москва*

Аннотация. Секвенирование человеческого генома началось в 1994 году. Понадобилось 10 лет работы многих научных коллективов для того, чтобы получить черновую последовательность ДНК человека. Современные технологии секвенирования позволяют получать геном конкретного человека за несколько дней. Обсуждаются успехи современной биоинформатики, связанные с появлением высокопроизводительных платформ секвенирования, которые не только способствовали расширению возможностей различных направлений биологии и других смежных наук, но и породили феномен больших данных. Обосновывается необходимость разработки новых технологий и методов для организации хранения, управления, анализа и визуализации больших данных. Современная биоинформатика столкнулась не только с проблемой больших данных, но и с огромным разнообразием методов обработки и представления, одновременным существованием различных программных средств и форматов данных. Обсуждаются пути решения возникших проблем, в частности путем использования наработок работы с большими данными из других областей современной жизни, таких как сетевой анализ и анализ деловых данных. Новые системы управления базами данных, отличные от реляционных, помогут решить проблему хранения больших данных и обеспечения приемлемого времени выполнения поисковых запросов. Новые технологии программирования, такие, как обобщенное программирование и визуальное программирование призваны решить проблему разнообразия форматов геномных данных и обеспечить возможность оперативного создания собственных скриптов для обработки данных.

Ключевые слова: *большие данные, Big Data, NGS, секвенирование генома, IT-технологии, биоинформатика, обобщенное программирование, визуальное программирование, нереляционные системы управления базами данных, NoSQL системы, Hadoop, MapReduce.*

ВВЕДЕНИЕ

В настоящее время понятие «большие данные» (англ., Big Data) стало общеупотребимым. Хотя до сих пор есть расхождения во мнениях по поводу строгого определения термина [1, 2], под большими данными понимают информацию огромного

[*]eisaev@hse.ru

объема, разнообразного состава, весьма часто обновляемую и находящуюся в разных источниках, а также специальные технологии хранения, передачи, обработки и анализа этой информации. Это понимание не только прочно вошло в лексикон специалистов по информационным технологиям, но и превратилось из модного технологического тренда в концепцию, включающую подходы, технологии и методики, активно используемые в самых различных областях жизни нашего общества. Более того, в октябре 2015 года из отчета аналитической компании Gartner «Цикл зрелости технологий» (Hype Cycle for Emerging Technologies), являющимся графическим отображением проникновения, адаптации и социального влияния новых технологических решений, вообще исчезло понятие больших данных, как отдельной технологии сбора и обработки больших массивов данных [3]. Свое решение аналитики компании объяснили тем, что в состав понятия «большие данные» входит большое количество активно используемых технологий, они частично относятся к другим популярным сферам и тенденциям и стали повседневным рабочим инструментом. Основной задачей работы с данными на сегодняшний день является извлечение из них ценных знаний. Наибольших успехов здесь достигли отрасли бизнеса, плотно взаимодействующие с потребителем и, соответственно, способные получить наибольшую выгоду от правильного анализа и предсказания поведения потенциальных клиентов. Это, прежде всего, относится к банкам, телекоммуникационной отрасли, розничной торговле, энергетике и отрасли ЖКХ. Сейчас речь идёт о грамотном использовании в бизнес-процессах компаний технологий хранения и обработки больших объемов данных, умения сделать их полезными бизнесу. Инструменты больших данных позволяют организациям эффективнее управлять ресурсами, предвидеть события, влияющие на их бизнес, и быстрее принимать обоснованные решения.

Информатика ответила на революционные изменения в жизни общества появлением новых научных направлений, самыми активно развивающимися среди которых являются сетевая аналитика и анализ бизнес-данных. В этих областях исследований уже существуют довольно развитые решения.

Сетевой анализ (Web Intelligence, WI) является областью научных исследований и разработок, которая занимается исследованием роли и практических последствий применения искусственного интеллекта (представлении знаний, планировании и организации процессов обнаружения знаний, интеллектуальном анализе данных, использовании интеллектуальных агентов) и передовых информационных технологий (беспроводных сетей, устройств электронной глобализации, социальных сетей, World Wide Wisdom Web (W4) и сетей (гридов) данных и знаний) на следующее поколение продуктов, систем, услуг и активностей для всемирной сети.

Бизнес-информация и аналитика (Business intelligence, BI) включает в себя технологические средства для сбора, обработки и анализа деловой информации, предназначенные для помощи руководителям корпораций, бизнес-менеджерам и другим конечным пользователям в принятии обоснованных бизнес-решений. Бизнес-анализ включает в себя широкий спектр инструментов, приложений и методик, которые позволяют организациям собирать данные из внутренних систем и внешних источников, готовить его для анализа, разработки и выполнения запросов к данным, а также создавать отчеты, информационные панели и другими способами визуализировать данные, чтобы сделать аналитические результаты доступными для лиц, принимающих корпоративные решения, а также оперативных работников.

Вместе с тем на пик популярности в области информационных технологий выходят новые тренды, базирующиеся на концепции больших данных, такие, как Интернет вещей (Internet of Things, IoT) [4]. Под этим понятием подразумевается революционное преобразование современного Интернета, когда множество устройств становятся «умными», способными собирать, анализировать информацию и обмениваться ею по

телекоммуникационным сетям, как с людьми, так и друг с другом. Особое значение приобретают машинное обучение, средства поиска правил и связей в очень больших объемах информации; средства интеллектуального анализа данных (Data Mining); развитые средства визуализации и самостоятельного анализа данных; системы поддержки принятия решений и искусственного интеллекта; системы распознавания естественных языков и др.

Однако всё вышесказанное, в основном, относится к индустриальной сфере жизни нашего общества. При этом, если вспомнить появление самого термина «большие данные» в 2008 году, то он тогда в первую очередь относился к научной сфере и, в значительной степени, к биоинформатике.

Считается, что термин «биоинформатика» впервые был употреблен в 1970 году Б. Хеспером (B. Hesper) и П. Хогевег (P. Hogeweg) в малоизвестной работе, опубликованной на голландском языке. Тогда он был определён как «изучение информационных процессов в биотических системах». Авторы считали, что определяющим свойством жизни является управление информацией в различных формах, например, накопление информации в процессе эволюции, обмен информацией между ДНК и участниками внутри- и межклеточных процессов, интерпретация информации на различных уровнях жизни. Авторы утверждают, что придумали этот термин, чтобы отделить биоинформатику от биофизики и биохимии [5].

Современная биоинформатика является наукой, занимающейся развитием и использованием компьютерных методов для анализа разнообразных геномных данных. Огромную роль в развитии биоинформатики сыграло стремительное развитие компьютерной техники и вычислительных методов обработки данных, появление современных телекоммуникационных технологий. Биоинформатика является одной из тех областей науки, которые в большей степени зависимы от Интернета и успешно могут развиваться только благодаря Интернету. Очень важное для биологии и медицины политическое решение об открытости сложнейшего биологического текста современности – генома человека – сделало эту информацию доступной для ученых всего мира и закончило формирование биоинформатики как коллективной науки, в которой достижения отдельных коллективов сразу же делаются достоянием всего научного сообщества, где принято свободно распространять созданные программные разработки и данные.

Ботаник Г Винклер (H. Winkler) очевидно не подозревал, что ввел моду на образование все новых и новых «омов», когда в 1920 году предложил термин «геном» для обозначения набора хромосом [6]. Тогда уже существовали понятия биома (множества живых существ) и ризома (системы корней), а в настоящее время ученые насчитывают тысячи различных «омов» [7]. Многие из этих терминов основаны на греческом суффиксе 'ome, что значит «имеющий природу». Одновременное развитие компьютерных мощностей и новых технологий получения данных в различных направлениях биологии, связанных с изучением геномов, привело к тому, что в биоинформатике выделились различные направления, называющиеся «омиками» (англ., 'omics), рассматривающие всю совокупность соответствующих объектов организма (ДНК, РНК, белки, метаболиты и т.д.) в структурно-функциональной взаимосвязи. Геномика, метагеномика, транскриптомика, протеомика, метаболомика, интерактомика и др. разделы биоинформатики занимаются изучением геномов, метагеномов, транскриптомов, протеомов, метаболомов, интерактомов и других совокупностей объектов [8].

Каждое из направлений биоинформатики имеет свои объекты для изучения и свои технологии получения данных. Но все они порождают огромные объемы данных в разных форматах и на различных уровнях, которые необходимо хранить, систематизировать, осмысливать и визуализировать с тем, чтобы углублять имеющиеся знания и стимулировать открытия.

СОВРЕМЕННАЯ БИОИНФОРМАТИКА: ПРОБЛЕМЫ И РЕШЕНИЯ

Геномика является исторически одной из первых омик биоинформатики, это направление занимается изучением генома. Геном представляет собой всю совокупность ДНК, имеющуюся в одной клетке, это генетический материал всех хромосом живого организма. Геномика занимается вопросами структуры, функции, эволюции и картирования геномов. Под картированием понимают определение места и функции различных функциональных элементов геномов (генов, кодирующих областей, промоторных и регуляторных участков, мигрирующих генетических элементов и др.) и структурных элементов (повторов, гомополимерных трактов и др.). В отличие от генетики, которая занимается изучением роли генов в наследственности, геномика изучает строение и механизмы функционирования генов. Основным методом получения данных является секвенирование геномной ДНК. До 2001 года, когда в результате международного научного сотрудничества был впервые «прочитан» и опубликован геном человека [9,10], секвенирование осуществлялось методом Сенгера, именно эта технология породила геномику. Новые технологии секвенирования (Next-Generation Sequencing, NGS, или high-throughput sequencing), появившиеся в конце первого десятилетия 21-го века, радикально снизили стоимость обработки одного генома по сравнению со 100 млн. долларов в 2001 году до 10 тыс. долларов в 2011 году, можно сказать, что они породили метагеномику. Метагеномика занимается изучением генетического материала исторически сложившейся совокупности видов живых организмов, объединённых общей областью распространения, называемой биотой.

Результатом проекта «Геном человека» стала одна последовательность более трёх миллиардов нуклеотидов, содержащихся в гаплоидном человеческом геноме. Однако все люди имеют в той или иной степени отличающиеся геномные последовательности, и для того, чтобы умение расшифровывать последовательности начало приносить реальную пользу, потребовалось создание и развитие методов, позволяющих одновременно прочитывать некоторое количество копий одного и того же генома. Этих методов в настоящее время насчитывается несколько десятков, каждый из которых имеет свои преимущества, ограничения и недостатки. Однако все они достаточно дёшевы и достаточно быстры для использования не только в специализированных лабораториях, но и медицинских клиниках, с их внедрением появилась возможность оперативно определять протяженные геномные последовательности конкретных пациентов, например, с целью выявления мутаций в генах, приводящих к развитию различных заболеваний. Новые высокопроизводительные технологии секвенирования позволяют решать задачи секвенирования *de novo* и ресеквенирования одного генома (DNA-Seq), изучения транскриптома всего организма (RNA-Seq) и транскриптома одной клетки (scRNA-Seq), характера ДНК-белок взаимодействий (ChIP-Seq), эпигенома, метагеномов и др. [8]. На одном и том же секвенаторе можно получать огромное количество разнообразных данных для различных омик, отличаются только способы (протоколы) получения и подготовки образцов для библиотек секвенирования.

Машины для секвенирования (секвенаторы) позволяют параллельно обрабатывать миллионы фрагментов нуклеиновых кислот, повторяя сотни раз однотипные операции, каждая из которых приводит к получению больших массивов данных для анализа. Для их обработки каждый секвенатор снабжается мощным серверным оборудованием, которое помогает решать задачи собственно прочтения до сотен миллиардов нуклеотидов в час, в результате чего биологи получают фрагменты заданной длины (т. нзв. прочтения, reads). После этого перед исследователями встает задача сборки и аннотации геномной последовательности. Существуют различные платформы и различные производители машин для секвенирования [11]. Основными платформами NGS второго поколения являются разработки Illumina Inc., Thermo Fisher Scientific, Roche и Pacific Biosciences. Своим появлением все они обязаны открытию полимеразной цепной реакции и автоматизация основных этапов чтения ДНК и

основываются на распараллеливании процесса чтения ДНК, за один прогон работы секвенатора можно определить первичные структуры нескольких участков генома. У каждой из этих технологий есть свои ограничения по длине и количеству прочтений, по цене, по доступности программного обеспечения и другим параметрам, но все они активно развиваются с тем, чтобы обеспечить быстрое и дешевое секвенирование любых геномных данных. Объемы этих данных огромны, их нужно в первую очередь хранить, анализировать и представлять в виде, полезном для биологов. Из-за разнообразия машин для секвенирования, у каждой из которых своя область применения, биологи должны разбираться в большом числе методов, отдельных компьютерных программах, созданных собственными силами, и совместных разработках членов сообщества, а также с разнообразием форматов данных и баз данных. Революционное развитие технологий секвенирования ставит новые задачи перед разработчиками программного обеспечения и специалистами по информационным технологиям.

Программное обеспечение

В настоящее время существует огромное множество компьютерных программ для работы с большими данными NGS [12]. Большая часть из них является коммерческими продуктами, среди них наиболее известными и широко используемыми являются BaseSpace (Illumina, Inc.) [13], CLCBio [14], Lasergene (DNASTAR, Inc.) [15], Geneious [16]. Среди свободных и гибридных (частично коммерческих, частично свободных) популярными являются программные комплексы Galaxy [17, 18], Globus Genomics [19], PATRIC [20], UGENE [21, 22]. Они обладают широчайшими возможностями, большинство из этих программных пакетов могут быть установлены на локальный компьютер или сервер, они работают под разными операционными системами. Однако, несмотря на огромный выбор программного обеспечения NGS, выпущенного в последние годы, до сих пор нет баланса между требованиями пользователей и возможностями предлагаемых им инструментов.

Современные программные решения для анализа лавины данных NGS должны быть снабжены удобной средой разработки. Время биологов, которые не имели дела с IT-технологиями, уходит в прошлое. Отсутствие возможности для исследователя быстро создать свое инновационное приложение может затормозить и усложнить анализ данных. Современные задачи аналитики геномных данных требуют компьютерной подготовки, которая имеет уровень выше среднего. Ввиду того, что ученые, занятые в больших проектах, над выполнением которых работают различные лаборатории, нуждаются в совершенно новых интерактивных средствах, активно использующих высокопроизводительные вычислительные инфраструктуры для анализа наборов данных NGS. Будущее в условиях, когда данные продолжают умножаться в объемах и в сложности, будет за средствами онлайн-анализа и хранения для обеспечения совместной работы исследователей, с методами, которые обеспечивают высокую степень интерактивного анализа данных и визуализации. Примерами программного обеспечения для управления научными рабочими процессами, созданными специально для работы с большими научными данными и более или менее используемыми для задач биоинформатики, являются KNIME [23], Pipeline Pilot [24], TAVERNA [25].

Система управления рабочими процессами представляет собой системное программное обеспечение для набора параметров производительности и мониторинга определенной последовательности задач, расположенных в виде рабочего процесса. Система представляет собой часть программного обеспечения, которое обеспечивает инфраструктуру для запуска, выполнения и мониторинга научных рабочих потоков. Такие разработки заменили собой библиотеки программ, снабженные возможностями командной строки, которые подходят для разработчиков программного обеспечения и обученных программированию пользователей и служат основой для разработки

пакетов программ высокого уровня, которые обладают развитым графическим интерфейсом пользователя (GUI) с меню и заранее запрограммированными рабочими процессами или конвейерным анализом данных.

Наличие легко осваиваемого и понятного интерфейса является ключевым аспектом поступательного движения в исследованиях NGS, поскольку он может обеспечить успешную работу пользователей без знания основ программирования. Программное обеспечение крупных программных пакетов высокого уровня тщательно разработано для того, чтобы пользователи не делали ошибок, вызванных отсутствием специальных знаний. В дополнение к GUI в этих пакетах есть возможности использования отдельных функций и готовых рабочих процессов, а также визуальные конструкторы рабочих процессов. Визуальные конструкторы имеют свои собственные графические интерфейсы, они позволяют комбинировать более или менее гибко из существующих программных функций новые конвейеры обработки данных, которые не могут быть доступны в раскрывающихся меню или иконках в общем графическом интерфейсе.

Когда пользователь вынужден пользоваться высокопроизводительными вычислительными инфраструктурами для выполнения проектов, связанных, например, с большим числом человеческих геномов, он должен иметь возможность использовать уже знакомую ему программу, с которой он уже имел дело на своем настольном компьютере, а не осваивать новую. Примером правильного проектирования прикладного программного обеспечения является программная система которая обеспечивает неизменный синтаксис программы для всех возможных конфигураций (настольный компьютер, сервер или распределенная среда).

Несмотря на избыток программ для NGS, имеет место недостаток возможностей для программирования низкого уровня для работы со специфическими структурами данных, например, с графами де Брёйна, используемыми в задачах сборки генома, и выполнением специфических функций, например, преобразования Барроуза-Уиллера используемого при сжатии данных. Этих возможностей нет в языках C++ или Java. Программные пакеты BAMTools [25], htslib (SAMtools/bcftools) [27], NGS++ [28], Bioclojure [29], libStatGen [30] используют стандартные форматы данных и предоставляют мало возможностей для использования специфических структур данных и разработки новых алгоритмов, необходимых для анализа данных NGS. Несмотря на наличие разработок в технологии обобщенного программирования [31], их применение к NGS проблематично из-за гигантского скачка в объемах данных. Это справедливо и для расширений языков программирования с открытым исходным кодом BioPerl [32], BioRuby [33], BioJava [34], BioPython [35], которые были созданы благодаря усилиям Open Bioinformatics Foundation [36] для крупных пакетов обработки данных типа Bioconductor [37, 38].

Обобщённое программирование (англ. *generic programming*) – парадигма программирования, заключающаяся в таком описании данных и алгоритмов, которое можно применять к различным типам данных, не меняя само это описание. В том или ином виде оно поддерживается разными языками программирования. Возможности обобщённого программирования впервые появились в виде обобщённых функций в 1970-х годах в языках Клу и Ада, затем в виде параметрического полиморфизма в ML и его потомках, а затем во многих объектно-ориентированных языках, таких как C++, Java, Object Pascal, D, Eiffel, языках для платформы .NET и других.

Обобщённое программирование рассматривается как методология программирования, основанная на разделении структур данных и алгоритмов через использование абстрактных описаний требований. Абстрактные описания требований являются расширением понятия абстрактного типа данных. Вместо описания отдельного типа в обобщённом программировании применяется описание семейства типов, имеющих общий интерфейс и семантическое поведение (англ. *semantic behavior*). Набор требований, описывающий интерфейс и семантическое поведение,

называется *концепцией* (англ. *concept*). Написанный в обобщённом стиле алгоритм может применяться для любых типов, удовлетворяющих его концепциям, такая возможность называется полиморфизмом. В C++ объектно-ориентированное программирование реализуется посредством виртуальных функций и наследования, а обобщенное программирование – с помощью шаблонов классов и функций.

Для обеспечения всех потребностей исследователей, способных к созданию низкоуровневых программных разработок и программ высокого уровня для расширения возможностей анализа данных NGS предлагается использование визуального программирования. Визуальное программирование не является новой концепцией [39]; начиная с начала 1960-х годов, оно является предметом философской дискуссии. Это способ создания компьютерной программы путём манипулирования графическими объектами вместо написания текста. Визуальное программирование часто представляют как следующий этап развития текстовых языков программирования. В последнее время визуальному программированию стали уделять больше внимания в связи с развитием мобильных сенсорных устройств. Визуальное программирование в основном используется для создания программ с графическим интерфейсом пользователя. Среда визуального программирования позволяет создавать веб-приложения и консольные приложения.

Форматы данных

Большой проблемой является гетерогенность биологических данных. Каждый производитель нового прибора разрабатывает свой собственный формат данных, делая задачу унификации данных сложнее и сложнее. Это подтверждает необходимость наличия хороших навыков в программировании для людей, работающих с различными приборами с тем, чтобы они могли модифицировать существующие или создавать новые скрипты для синтаксического разбора данных (парсинга) и преобразования одного формата в другой.

В эру больших данных меняются привычные общепринятые форматы хранения данных. Для обмена данными были разработаны и долгое время считались классическими основные форматы данных, такие как pdb для пространственных структур белков, fasta для нуклеотидных и аминокислотных последовательностей. Сейчас появляются новые форматы данных. Примером может служить формат базы данных PDB [40]. Появившийся в 1970-х годах этот формат долгое время служил для хранения и обмена данными о структурах небольших белков. Однако формат PDB невозможно использовать для больших комплексов, которые состоят из тысяч аминокислот, поэтому теперь введен формат PDBx/mmCIF, который объединил формат PDB и формат хранения кристаллографических данных mmCIF [41] и официально заменил PDB-формат в 2014-м году.

Биоинформатика всегда была связана с большим количеством баз данных, в которых хранятся разнообразные геномные данные. Самая серьезная коллекция биомедицинских баз данных, которую ведет журнал Nucleic Acids Research (NAR), насчитывает около 1900 описаний ресурсов, опубликованных в ежегодных тематических выпусках журнала, посвященных базам данных. Коллекция разбита на 15 тематических категорий, в которых выделена еще 41 подкатегория [42]. Эта систематизация в большой степени условна, потому что хорошим тоном считается делать информационные ресурсы политематическими, организовывать перекрестные ссылки между всеми другими базами данных, где может содержаться полезная информация. Коллекция собирается уже почти четверть века, в ней есть «золотое» ядро из 105 ресурсов, которые существуют давно и постоянно обновляются. Крупнейшие базы данных, которые содержат объемные экспериментальные данные, такие как нуклеотидные и белковые последовательности, данные о структурах биологических макромолекул, кристаллографические данные и др., пополняются главным образом

самими экспериментаторами с использованием возможностей электронной подачи. Это обусловлено требованиями всех журналов, публикующих результаты подобного рода. Авторы до подачи статьи, описывающей факт определения структуры биомолекулы, должны разместить данные на общедоступном ресурсе, сделав их достоянием всего научного сообщества.

Хранение и обмен данными

Проблема полноты баз данных решена путем создания концорциумов баз данных, которые существуют для хранения последовательностей ДНК и белков. Нуклеотидные последовательности загружаются в одну из трех баз данных GenBank [43], ENA [44] или DDBJ [45] либо авторами, либо центрами секвенирования. Между этими тремя базами данных налажен ежедневный обмен данными, так что ежедневные обновления на серверах NCBI, где хранится GenBank, включают в себя самые последние доступные данные о последовательностях из всех трех источников.

Однако в новую эру больших данных привычные системы управления базами данных, основанные на реляционном принципе, перестали соответствовать большим объемам данных, разнообразию форматов, необходимости совместного доступа к данным из разных уголков земного шара и разнообразию поисковых запросов [45]. Реляционная организация хранения данных предполагает наличие заданных заранее поисковых полей и логической структуры запросов. Для хранения данных в реляционной схеме строятся таблицы для каждого поискового поля, в которых хранятся значения полей. При разрешении запросов строятся временные таблицы, что при огромных объемах данных делает работу неэффективной, а в ряде случаев, и невозможной. Поэтому ключевые игроки Интернета, такие как Amazon, Inc. и Google, Inc. в начале 2000-х годов начали разработку новых систем управления базами данных.

Одним из таких решений является NoSQL (“Not Only SQL”) – класс нереляционных систем управления базами данных (СУБД), разработанных для работы с большими данными [45]. Системы NoSQL обеспечивают быстрое время отклика на поисковые запросы при высокой пропускной способности обработки потока запросов. По типу организации хранения данных эти системы можно поделить на 2 группы.

Первая группа использует логический принцип «ключ-значение», это, по сути, ассоциативная таблица, каждому ключу соответствует уникальное значение. Вторая группа систем ориентирована на документ. Это не полностью систематизированное хранение, здесь не используются таблицы.

Другими нереляционными системами управления данными являются графовые базы данных, которые используют для семантических запросов структуры графов с узлами, ребрами и свойствами узлов для представления и хранения данных. Ключевой концепцией системы является граф (или ребро или связь), которые имеют непосредственное отношение к элементам данных. Связи позволяют связывать данные друг с другом непосредственно, и во многих случаях, извлекать за одну операцию.

Это отличает графовые системы от традиционных реляционных баз данных, где связи между данными реализованы с помощью таблиц, сложные поисковые запросы решаются путем объединения таблиц, удовлетворяющих элементарным запросам. Графовые базы данных обеспечивают простое и быстрое извлечение сложных иерархических структур, которые трудно искать в реляционных системах [47]. Принцип хранения графовых баз данных постоянно меняется. Некоторые системы используют элементы реляционной организации, то есть хранят графы в виде таблиц, в то время как другие используют принцип ключ-значение или документ-ориентированную концепцию для хранения данных, что делает их по сути NoSQL структурами.

Извлечение данных из графовой базы данных требует специального языка запросов, отличного от SQL, который был разработан для реляционных баз данных и не способен

элегантно осуществлять обход графа. Ни один из языков запросов не стал общепринятым, как SQL использовался для реляционных баз данных, существует большое разнообразие систем, жестко привязанных к конкретному продукту. Однако некоторые усилия по стандартизации были осуществлены, что привело к появлению языков запросов, таких как Cypher, который может стать стандартным [48, 4948]. В дополнение к наличию языков запросов, некоторые графовые базы данных доступны через API.

Другим примером нереляционной модели хранения данных является HBase [50]. Она разработана для работы с файловой системой распределенной операционной системы Hadoop (HDFS, Hadoop Distribited File System). Система Hadoop была специально разработана для работы с большими данными. Обычная файловая система состоит из таблицы файловых дескрипторов и области данных. В HDFS вместо таблицы используется сервер имён (NameNode), а данные распределены по серверам данных (DataNode). Информация о том, на каких машинах расположены блоки данных, позволяет запустить на них же вычислительные процессы и выполнить большую часть вычислений локально, т.е. без передачи данных по сети. Именно эта идея лежит в основе парадигмы MapReduce и её конкретной реализации в Hadoop. Классическая конфигурация кластера Hadoop состоит из одного сервера имён, одного мастера MapReduce (т.н. JobTracker) и набора рабочих машин, на каждой из которых одновременно работает сервер данных (DataNode) и воркер (TaskTracker). Каждая MapReduce работа состоит из двух фаз.

– Фаза map выполняется параллельно и (по возможности) локально над каждым блоком данных. Вместо того, чтобы доставлять терабайты данных к программе, небольшая, определённая пользователем программа копируется на сервера с данными и делает с ними всё, что не требует перемешивания и перемещения данных (shuffle).

– Фаза reduce дополняет фазу map агрегирующими операциями.

Hbase – это распределенная, колоночно-ориентированная, мультиверсионная база типа «ключ-значение», сделанная по образцу BigTable [51], разработанной Google. Данные организованы в строки, проиндексированные первичным ключом, который в HBase называется RowKey. Для каждого ключа RowKey может храниться неограниченный набор атрибутов (или колонок). Колонки организованы в группы колонок, называемые Column Family. Как правило, в одну Column Family объединяют колонки, для которых одинаковы паттерн использования и хранения. Для каждого атрибута может храниться несколько различных версий. Разные версии отличаются по метке времени timestamp. Записи физически хранятся в отсортированном по ключу RowKey порядке. При этом данные, соответствующие разным элементам типа Column Family, хранятся отдельно, что позволяет при необходимости читать данные только из нужного семейства колонок. Атрибуты, принадлежащие одной группе колонок и соответствующие одному ключу, физически хранятся как отсортированный список. Любой атрибут может отсутствовать или присутствовать для каждого ключа, при этом если атрибут отсутствует это не вызывает накладных расходов на хранение пустых значений. Четырехмерную модель данных HBase можно сформулировать как отношение типа «ключ-значение» следующего вида:

<table, RowKey, Column Family, Column, timestamp> -> значение.

По результатам исследований по применимости HBase в биоинформатике для NGS данных [50] было признано, что масштабируемость и надежность работы ориентированной на большие данные HBase достаточно велика. Было также показано, что эта архитектура позволяет быструю интеграцию и анализ больших и гетерогенных данных, используя для их хранения небольшое количество таблиц.

Визуализация

Новые технологии NGS получения данных в биологии открывают перед исследователями новые горизонты в формулировании новых представлений и концепций, однако большие данные трудны для анализа и визуализации. Визуализация играет ключевую роль в обнаружении новых паттернов и трендов, недостаток специализированных средств представления является лимитирующим фактором в интерпретации данных.

Поскольку визуализация данных является фундаментом в интерпретации данных секвенирования, многие разработчики ПО занимаются созданием программных средств для визуального анализа данных. Эти разработки более специализированы, чем другие пакеты анализа данных NGS, каждый из них имеет свой объект визуализации и свою область применимости [45]. Некоторые из них (ngs.plot [52] и Integrative Genomics Viewer [53]) позволяют интегрировать гетерогенные данные, такие как аннотации генов, клиническая информация и фенотипические данные, из разных источников. Пакет Girafe [54] может использоваться для визуализации процесса выравнивания прочтений (reads) с геномными фрагментами, он удобен тем, что он работает вместе с пакетом R/Bioconductor [38]. Bioconductor – это масштабный проект с открытым ПО, предоставляющий множество отдельных пакетов для биоинформатических исследований, использует язык программирования R, благодаря чему поддерживает кросплатформенность (Linux, большинство UNIX-подобных систем, Mac OS X, Windows). Несмотря на наличие возможностей работать с динамическими графиками [55] на веб-страницах, эти решения недостаточны для управления динамическими графиками для больших данных и достаточной интерактивности.

Главная особенность визуализации огромных данных состоит в том, что нужно показывать на экранах мониторов с ограниченным числом пикселей многие миллионы точек. Важная техническая проблема заключается в необходимости динамически взаимодействовать с графиками, например, для изменения типа графика, уменьшения и увеличения масштаба изображения, просмотра и изменения параметров и мгновенного получения новой картинки. К тому же, революция технологий геномики и большое количество имеющихся данных приводит к увеличению значения совместной работы. Для анализ NGS обычно необходимо тесное оперативное сотрудничество ученых из разных коллективов и разных мест. Это сотрудничество обеспечивается созданием веб-приложений, которые должны поддерживать возрастающий потенциал интенсивности исследования данных. Это вызывает необходимость использования веб-технологий для обмена данными, инструментов анализа и результатов через веб-приложения, доступные из Интернета. Тем не менее, веб-приложения имеют некоторые технические ограничения, связанные с возможностями современных веб-браузеров. Многочисленные проблемы возникают при создании веб-приложений, которые включают в себя интерактивные средства визуализации для анализа BigData. Например, веб-браузеры не могут поддерживать огромные интерактивные графики и таблицы с тысячами фрагментов данных. Таким образом, разработчики веб-приложений для анализа и визуализации NGS данных должны создавать приложения с поддержкой больших объемов данных и увеличивающихся потребностей ученых в интерактивности их данных в сочетании с ограничениями и разнообразием веб-браузеров.

IT-компании имеют готовые инновационные решения для визуализации больших данных для визнес-анализа (Business Intelligence). Эти продукты за последнее десятилетие стали значительно более мощными и менее дорогостоящими. В качестве примеров можно привести пакеты Tibco Spotfire [56] или SAS [57], оба из которых используются в науках о жизни и способны значительно помочь визуализировать и исследовать данные NGS. Главное преимущество этих решений состоит в том, что они предлагают мощную визуализацию с многочисленными типами графиков для

представления данных и обеспечивают высокий уровень интерактивности для изменения параметров изображений или масштабирования.

МЕТОДЫ ОРГАНИЗАЦИИ ВЫЧИСЛЕНИЙ С БОЛЬШИМИ ОБЪЕМАМИ ДАННЫХ

В современной биоинформатике появилась возможность массово проводить анализ генома разных по сложности живых систем от микроорганизмов до человека. Накапливаемые данные потенциально содержат до сих пор неизвестную для исследователей и крайне важную информацию о функциях и работе генетического кода. Однако объемы генерируемых данных настолько велики (к примеру, сохранение одной геномной последовательности на компьютерном носителе информации потребует около сотни гигабайт), что проблемой становится не только хранение и передача этих данных, ещё большие трудности возникают при обработке данных. Сама сборка генома из миллиардов просеквенированных олигонуклеотидов является сложной задачей. А решение прикладных задач биоинформатики с использованием геномных последовательностей требует применения новейших разработок распределенного и параллельного программирования.

Таким образом, по мере сбора огромных объемов геномной информации критичными становятся более эффективные, точные и специфичные методики выполнения анализа этой информации, опирающихся на современные методы анализа больших объемов данных. Для их создания можно позаимствовать IT-решения, которые были созданы для управления большими данными и их анализа в таких областях информационных наук, как сетевой искусственный интеллект и бизнес-информация и аналитика.

Прежде всего – это математические и статистические методы анализа и обработки данных и алгоритмы поиска информации, применимые для огромных наборов данных. К таким методам относятся средства углублённой аналитики класса Data Mining (напр., кластерный и регрессионный анализ), методики обработки текстов на естественном языке (включая тональный анализ), прогнозная аналитика, алгоритмы статистического анализа данных (такие как A/B-тестирование и анализ временных рядов), алгоритмы машинного обучения и другие.

Во-вторых, это инструментальные и программно-технические средства информационных технологий, позволяющие хранить и обрабатывать сверхбольшие объемы данных. Основным способом решения этих задач является организация распределённых вычислений с использованием большого числа вычислительных узлов, чаще всего объединённых в параллельную вычислительную систему [58]. К наиболее известным реализациям модели выполнения распределённых вычислений в больших параллельных кластерах вычислительных узлов относят платформу программирования «MapReduce», предложенную компанией Google [59] и свободно-распространяемый фреймворк «Hadoop», созданный и поддерживаемый фондом Apache Software Foundation [60] и предназначенный для разработки и поддержки выполнения программ распределённых вычислений. Hadoop основывается на реализации модели MapReduce и распределённой файловой системе Hadoop (англ., HDFS), предназначенной для хранения файлов больших размеров, распределённых между узлами вычислительного кластера. К инфраструктуре Hadoop также относят различные программные средства обработки больших данных, такие как Apache Pig – высокоуровневый процедурный язык программирования, программную систему организации запросов к данным Apache Hive, распределённую базу данных HBase, Apache Spark – программный фреймворк для реализации распределённой обработки неструктурированных и слабоструктурированных данных и др. [61]. Кроме того, существует большое число коммерческих реализаций программных систем, основанных на технологии Hadoop – разного класса, возможностей и назначения, предлагаемых как крупными вендорами,

такими как IBM или Microsoft, так и относительно небольшими компаниями – Cloudera, Hortonworks, MapR и др.

Третьей составляющей технологического инструментария для обработки больших данных являются специализированные сверхпроизводительные программно-аппаратные комплексы для работы с большими данными. Это продукты, использующие парадигму обработки данных in-memory analytics (технологии, максимально использующие оперативную память для работы аналитических систем), такие как, например, недавно анонсированный компанией IBM мейнфрейм z13s [62]. Этот суперкомпьютер, ориентированный на решение задач из области больших данных, имеет до 20-ти восьмиядерных, 5.0 ГГц специализированных процессоров, сверхбыструю оперативную память объёмом до 10 ТБ, поддерживает до 8000 виртуальных серверов и обладает расчётной максимальной производительностью более 111000 MIPS (миллионов операций в секунду). Подобные специализированные суперкомпьютеры также предлагаются различными поставщиками таких решений, как Teradata, Oracle, SAP, SAS и другими.

Таким образом, в настоящее время мы имеем достаточно продвинутый набор методов аналитики больших данных, позволяющих извлекать из них необходимые знания. Однако, как уже было сказано ранее, подавляющее число технологий и методов работы с большими данными реализовано и успешно применяется практически исключительно в различных отраслях бизнеса.

ЗАКЛЮЧЕНИЕ

Революционные изменения, которые высокопроизводительное секвенирование (NGS) вызвало в биологических науках, способствуют появлению с большой скоростью огромного количества данных. Очевидно, что большие данные имеют разнообразные форматы и представляют большой интерес для различных групп исследователей. Нет общего мнения, какие данные можно считать большими, общим является мнение, что это коллекции данных, которые слишком велики, чтобы управлять и анализировать их с использованием традиционных подходов. Согласно этой модели, масштаб и способ получения больших данных являются спецификой каждой области исследований. Данные, которые подходят под это определение в биологии и медицине, генерируются из многочисленных источников, в том числе, из лабораторных экспериментов, и доступны через онлайновые базы данных. Медико-биологические большие данные появляются в результате слияния небольших источников данных.

Например, ученые делают достоянием коллег лабораторные эксперименты по изучению экспрессии генов, помещая их в базу данных ArrayExpress [63], которая является одной из трех коллекций результатов дорогостоящих экспериментов наряду с Omnibus в NCBI (США) и DDBJ Omics Archive (Япония). Эти хранилища сообща хранят обработанные данные и метаданные, описывающие свойства образцов и технические детали эксперимента, в том числе экспериментальные показатели и протоколы, в то время как необработанные последовательности помещаются в Европейскую нуклеотидную базу данных (ENA), попадая сразу в связанные с этим ресурсом коллекции данных GenBank и DDBJ.

Эти общие большие данные экономически выгодны, поскольку стоимость получения данных делится на многие лаборатории, а для использования этих общих данных разработаны специальные вычислительные методы. Медико-биологические большие данные дают возможность развивать предсказания, основанные на фактических данных, которые дополняют гипотезы, основанные на знаниях поколений. Поскольку эти данные поставляются в коллекции различными исследовательскими коллективами из различных учреждений, а изучаемые системы разнообразны, то и открытия будут с большей вероятностью обобщенными.

Большие данные не только открывают новые возможности, но и ставят новые задачи. Становится необходимой адаптация учебных программ по биоинформатике к новым реалиям [64]. Необходима разработка учебных программ, которые обеспечивают навыки эффективного и уверенного использования больших данных и способности критически оценивать результаты.

Проблемы, поднятые большими данными, требуют, чтобы программы обучения готовили студентов для решения задач, включающих в себя объединение данных, преодоление вычислительных трудностей и ограничений хранения, навыки многократного тестирования гипотез и работу со смещенными и смешанных данными. Объединение данных охватывает проблемы получения необходимых данных в соответствующем формате и их нормализацию, чтобы сделать их сопоставимыми по источникам. Вычислительные ограничения относятся к трудностям и затратам, связанным с хранением данных, перемещением данных и анализом данных. Многократная проверка гипотез относится к задаче статистической вероятности обнаружения паразитные ассоциаций в больших наборах данных. расхождения и противоречивость данных относятся к проблемам, связанным с тем, какие эксперименты были выполнены или какие процессы наиболее часто анализировали.

Эта область знаний развивается быстро, и проблемы, сформулированные здесь, не являются статичными, они тоже быстро меняются. Биоинформатики в эру больших данных должны быть в состоянии разбираться в вычислительной среде и в том, как в этой среде наиболее эффективно анализировать и получить аналитические выводы из крупномасштабных данных. Кроме того, они должны хорошо разбираться в алгоритмах сборки геномов, чтобы выбрать из огромного множества существующих наиболее подходящий.

Значительные ресурсы выделяются в мире для подготовки ученых к анализу крупномасштабных данных. Правительство США выделило 200 млн. долларов на финансирование программ Big Data и NIH Big Data to Knowledge (BD2K) initiative [65]. Программа Big Data в частности, направлена на значительное усовершенствование инструментов и методов доступа, организации и сбора данных об открытиях, связанных с огромными объемами цифровых данных.

Реальность такова, что использование привычных реляционных систем управления базами данных уже не может удовлетворять возросшим нагрузкам на объемы данных. Реляционная база не может адаптироваться к большому количеству запросов; объемы таблиц, необходимых для реализации этой модели, растут слишком быстро для большого количества хранимых данных. Реляционная модель больше не соответствует критерию производительности, так как эта модель данных оперирует большим числом временных таблиц, в которых хранятся промежуточные результаты. Необходимы другие системы управления базами больших данных, отвечающие возросшим требованиям современной жизни.

Большой проблемой является гетерогенность форматов данных и программных средств для их обработки. Каждый производитель машин для секвенирования разрабатывает, как правило, свой собственный формат данных, что делает унификацию данных актуальной задачей. Эта ситуация требует, чтобы биологи-исследователи имели серьезный уровень знаний в языках программирования, для того, чтобы они могли использовать существующие или создавать новые скрипты для анализа данных и извлечения полезной информации. В Интернете выложено множество инструментов для преобразования и анализа данных. Все они написаны на разных языках программирования и предназначены для различных компьютерных платформ. Трудность состоит в понимании уровня согласования между собой разных инструментов и организации из них рабочего процесса, а также в обновлении и сопровождении программного обеспечения.

Нужны новые распределенные вычислительные технологии. В областях бизнес-анализа уже создан ряд решений, которые можно было бы применить и для биоинформатики. Эти задачи под силу новому поколению биоинформатиков.

Необходимы новые методы визуализации, помогающие человеческому сознанию осмыслить данные различных омик. Поэтому в эпоху технологий быстрого и дешевого получения любых данных, возможно, нужно изменить методы работы. Например, взять за правило не «множить сущности», то есть не накапливать данные, а тщательно планировать исследования и наглядное представление результатов. И привлекать дизайнеров для визуализации на этапе планирования экспериментов, а не после того, как экспериментальные данные уже получены. Тогда структура данных будет более продуманной и оптимальной.

Размышления о способах визуализации могут привести к разработке альтернативных представлений для одних и тех же данных. Это может повлечь за собой разработку других подходов к сбору, организации и поиску данных, которые способствуют максимальной осмысленности экспериментальных данных, что в свою очередь стимулирует интуицию, открытия и информационное взаимодействие.

Наш мир изменился, наше общество становиться информационным, полностью управляемым информационными потоками; знание и умение становятся главной ценностью. Инструменты больших данных позволяют эффективнее управлять ресурсами, предвидеть будущие события, быстрее принимать обоснованные решения.

В биоинформатике и вычислительной биологии данных также стало слишком много для их анализа «по старинке», причём скорость их появления всё более нарастает, а сложность их анализа очень высока из-за их специфической структуры и организации. При этом применение технологий больших данных в биоинформатике, биомедицине и здравоохранении [66] способно не просто улучшить, а кардинально, революционно изменить ситуацию в этой области. Однако, несмотря на некоторые успехи в развитии методов анализа и в практическом применении новых технологий работы с большими данными, в биоинформатике и биомедицине имеется огромный неиспользованный потенциал для их развития.

Исследование частично поддержано грантами РФФИ №15-07-05783 (Н.Н.Н.), 16-07-00937 и 16-07-01000 (У.М.Н.) и Программой фундаментальных научных исследований президиума РАН I.33П. (У.М.Н.).

СПИСОК ЛИТЕРАТУРЫ

1. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A.H. *The Next Frontier for Innovation, Competition, and Productivity*. San Francisco: McKinsey Global Institute, 2011. URL: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (дата обращения: 17.02.2017).
2. Jacobs A. The Pathologies of Big Data. *Communications of the ACM*. 2009. V. 52. No. 8. doi: 10.1145/1536616.1536632
3. What's New in Gartner's Hype Cycle for Emerging Technologies, 2015. *Gartner*. URL: http://www.gartner.com/smarterwithgartner/whats-new-in-gartners-hype-cycle-for-emerging-technologies-2015/ (дата обращения: 17.02.2017).
4. Chui M., Löffler M., Roberts R. The Internet of Things. *McKinsey Quarterly*. 2010. URL: http://www.mckinsey.com/industries/high-tech/our-insights/the-internet-of-things (дата обращения: 17.02.2017).
5. Hogeweg P. The Roots of Bioinformatics in Theoretical Biology. *PLOS Computational Biology*. 2011. V. 7. No. 3. Article No. e1002021.
6. Winkler H. *Verbreitung und Ursache der Parthenogenesis im Pflanzen - und Tierreiche*. Jena: Verlag Fischer, 1920.
7. Baker M. The 'Oms Puzzle. *Nature*. 2013. V. 494. P. 416–419.

8. Ohashi H., Hesegawa M., Wakimoto K., Miyamoto-Sato E. Next-generation technologies for multiomics approaches including interactome sequencing. *BioMed Research International*. 2015. V. 2015. Article No. 104209.
9. International Human Genome Sequencing Consortium. Human genome. *Nature*. 2001. V. 409. P. 860–921.
10. Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., et al. The sequence of the human genome. *Science*. 2001. V. 291. No. 5507. P. 1304–1351.
11. Buermans H.P.J., den Dunnen J.T. Next generation sequencing technology. Advances and applications. *BBA – Molecular Basis of Disease*. 2014. V. 1842. No. 10. P. 1932–1941.
12. Bioinforx Inc. *Next Generation Sequencing Software*. URL: http://bioinfo.wisc.edu/knowledge_base/next-gen-seq_software.php (дата обращения: 17.02.2017).
13. *BaseSpace Sequence Hub*. URL: https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_basespace.pdf (дата обращения: 17.02.2017).
14. *CLCBio*. URL: http://www.clcbio.com (дата обращения: 17.02.2017).
15. *DNASTAR Lasergene*. URL: https://www.dnastar.com/t-allproducts.aspx (дата обращения: 17.02.2017).
16. Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012. V. 28. No. 12. P. 1647–1649.
17. Giardine B., Riemer C., Hardison R.C., Burhans R., Elnitski L., Shah P., Zhang Y., Blankenberg D., Albert I., Taylor J., et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*. 2005. V. 15. No. 10. P. 1451–1455.
18. Goecks J., Nekrutenko A., Taylor J., Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010. V. 11. No. 8. Article No. R86.
19. Madduri R.K., Sulakhe D., Lacinski L., Liu B., Rodriguez A., Chard K., Dave U.J., Foster I.T. Experiences Building Globus Genomics: A Next-Generation Sequencing Analysis Service using Galaxy, Globus, and Amazon Web Services. *Concurr. Comput*. 2014. V. 26. No. 13. P. 2266–2279.
20. Wattam A.R., Abraham D., Dalay O., Disz T.L., Driscoll T., Gabbard J.L., Gillespie J.J., Gough R., Hix D., Kenyon R., et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*. 2014. V. 42. P. D581–D591.
21. Golosova O., Henderson R., Vaskin Y., Gabrielian A., Grekhov G., Nagarajan V., Oler A.J., Quinones M., Hurt D., Fursov M., Huyen Y. Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses. *PeerJ*. 2014. V. 2. Article No. e644.
22. Okonechnikov K., Golosova O., Fursov M., UGENE Team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*. 2012. V. 28. No. 8. P. 1166–1167.
23. Jagla B., Wiswedel B., Coppree J.-Y. Extending KNIME for next-generation sequencing data analysis. *Bioinformatics*. 2011. V. 27. No. 20. P. 2907–2909.
24. Warr W.A. Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of Computer-Aided Molecular Design*. 2012. V. 26. No. 7. P. 801–804.
25. Oinn T., Addis M., Ferris J., Marvin D., Senger M., Greenwood M., Carver T., Glover K., Pocock M.R., Wipat A., Li P. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*. 2004. V. 20. No. 17. P. 3045–3054.

26. Barnett D.W., Garrison E.K., Quinlan A.R., Stromberg M.P., Marth G.T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*. 2011. V. 27. No. 12. P. 1691–1692.
27. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009. V. 25. No. 16. P. 2078–2079.
28. Nordell Markovits A., Joly Beauparlant C., Toupin D., Wang S., Droit A., Gevry N. NGS++: a library for rapid prototyping of epigenomics software tools. *Bioinformatics*. 2013. V. 29. No. 15. P. 1893–1894.
29. Plieskatt J., Rinaldi G., Brindley P.J., Jia X., Potriquet J., Bethony J., Mulvenna J. Bioclojure: a functional library for the manipulation of biological sequences. *Bioinformatics*. 2014. V. 30. No. 17. P. 2537–2539.
30. *libStatGen*. URL: https://github.com/statgen/libStatGen/ (дата обращения: 17.02.2017).
31. Pitt W.R., Williams M.A., Steven M., Sweeney B., Bleasby A.J., Moss D.S. The Bioinformatics Template Library – generic components for biocomputing. *Bioinformatics*. 2001. V. 17. No. 8. P. 729–737.
32. Stajich J.E., Block D., Boulez K., Brenner S.E., Chervitz S.A., Dagdigian C., Fuellen G., Gilbert J.G., Korf I., Lapp H., et al. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*. 2002. V. 12. No. 10. P. 1611–1618.
33. Goto N., Prins P., Nakao M., Bonnal R., Aerts J., Katayama T. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics*. 2010. V. 26. No. 20. P. 2617–269.
34. Holland R.C., Down T.A., Pocock M., Prlic A., Huen D., James K., Foisy S., Drager A., Yates A., Heuer M., et al. BioJava: an open-source framework for bioinformatics. *Bioinformatics*. 2008. V. 24. No. 18. P. 2096–2097.
35. Cock P.J., Antao T., Chang J.T., Chapman B.A., Cox C.J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009. V. 25. No. 11. P. 1422–1423.
36. *Open Bioinformatics Foundation*. URL: https://www.open-bio.org/wiki/Main_Page (дата обращения: 17.02.2017).
37. Huber W., Carey V.J., Gentleman R., Anders S., Carlson M., Carvalho B.S., Bravo H.C., Davis S., Gatto L., Girke T., et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*. 2015. V. 12. No. 2. P. 115–121.
38. Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004. V. 5. No. 10. Article No. R80.
39. Milicchio F., Rose R., Bian J., Min J., Prosperi M. Visual programming for next-generation data analytics. *BioData Mining*. 2016. V. 9. Article No. 16.
40. Bernstein F.C., Koetzle T.F., Williams G.J., Meyer E.F.Jr., Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T., Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*. 1977. V. 112. No. 3. P. 535–542.
41. Bourne P. E., Berman H.M., McMahon B., Watenpaugh K.D., Westbrook J.D., Fitzgerald P.M.D. Macromolecular crystallographic information file. *Methods in Enzymology*. 1997. V. 277. P. 571–590.
42. Galperin M.Y., Fernández-Suárez X.M., Rigden D.J. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. *Nucleic Acids Res*. 2017. V. 45. P. D1–D11.

43. Benson D., Lipman D.J., Ostell J. GenBank. *Nucleic Acids Res.* 1994. V. 22. P. 3441–3444.
44. Rice C.M., Fuchs R., Higgins D.G., Stoehr P.J., Cameron G.N. The EMBL Data Library. *Nucleic Acids Res.* 1993. V. 21. P. 2967–2971.
45. Tateno Y., Gojobori T. DNA Data Bank of Japan in the age of information biology. *Nucleic Acids Res*. 1997. V. 25. No. 1. P. 14–17.
46. de Brevern A.G., Meyniel J.-P., Fairhead C., Neuvéglise C., Malpertuy A. Trends in IT Innovation to Build a Next Generation Bioinformatics Solution to Manage and Analyse Biological Big Data Produced by NGS Technologies. *BioMed Research International*. V. 2015. Article No. 904541.
47. Lith A., Mattsson J. *Investigating Storage Solutions for Large Data. A comparison of well performing and scalable data storage solutions for real time extraction and batch insertion of data*: Master of Science Thesis. 2010. URL: http://publications.lib.chalmers.se/records/fulltext/123839.pdf (дата обращения: 17.02.2017).
48. Svensson J. Relational vs. graph databases: Which to use and when? *SD Times*. 2016. URL: http://sdtimes.com/guest-view-relational-vs-graph-databases-use/#sthash.yHI6aoDv.dpuf (дата обращения: 17.02.2017).
49. Have C.T., Jensen L.J. Are graph databases ready for bioinformatics? *Bioinformatics*. 2013. V. 29. No. 24. P. 3107–3108.
50. Taylor R.C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics*. 2010. V. 11. Article No. S1.
51. Chang F., Dean J., Ghemawat S., Hsieh W.C., Wallach D.A., Burrows M., Chandra T., Fikes A., Gruber R.E. *Bigtable: A Distributed Storage System For Structured Data*. In: The 7th Symposium on Operating System Design and Implementation Seattle, WA: Usenix Association, 2006. 14 p. URL: https://static.googleusercontent.com/media/research.google.com/ru//archive/bigtable-osdi06.pdf (дата обращения: 17.02.2017).
52. Shen L., Shao N., Liu X., Nestler E. Ngs.plot: quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*. 2014. V. 15. No. 1. Article No. 284.
53. Robinson J.T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E.S., Getz G., Mesirov J.P. Integrative genomics viewer. *Nature Biotechnology*. 2011. V. 29. No. 1. P. 24–26.
54. Toedling J., Ciaudo C., Voinnet O., Heard E., Barillot E. Girafe – an R/Bioconductor package for functional exploration of aligned next-generation sequencing reads. *Bioinformatics*. 2010. V. 26. No. 22. P. 2902–2903.
55. Nolan D., Lang D.T. Interactive and animated scalable vector graphics and R data displays. *Journal of Statistical Software.* 2012. V. 46. No. 1. P. 1–88.
56. *TIBCO Spotfire Homepage*. URL: http://spotfire.tibco.com/ (дата обращения: 17.02.2017).
57. Wexler J., Thompson W., Aponte K. Time Is Precious, So Are Your Models. SAS provides solutions to streamline deployment. In: *SAS Global Forum 2013*. Paper No. 086-2013. URL: https://support.sas.com/resources/papers/proceedings13/086-2013.pdf (дата обращения: 17.02.2017).
58. Таненбаум Э., ван Стеен М. *Распределенные системы. Принципы и парадигмы*. С.-П.: Питер, 2003. 877 с.
59. Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun. ACM*. 2008. V. 51. No. 1. P. 107–113.
60. White T. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 2015. 756 p.
61. *The Apache Software Foundation Home page*. URL: http://www.apache.org/ (дата обращения: 17.02.2017).

62. *IBM z Systems – z13s*. URL: http://www-03.ibm.com/systems/z/hardware/z13s.html/ (дата обращения: 17.02.2017).
63. Rustici G., Kolesnikov N., Brandizi M., Burdett T., Dylag M., Emam I., Farne A., Hastings E., Ison J., Keays M., et al. ArrayExpress update – trends in database growth and links to data analysis tools. *Nucleic Acids Res*. 2013. V. 41. P. D987–D990.
64. Greene A.C., Giffin K.A., Greene C.S., Moore J.H. Adapting bioinformatics curricula for big data. *Briefings in Bioinformatics*. 2016. V. 17. No. 1. P. 43–50.
65. Margolis R., Derr L., Dunn M., Huerta M., Larkin J., Sheehan J., Guyer M., Green E.D. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J. Am. Med. Inform. Assoc*. 2014. V. 21. P. 957–958.
66. Luo J., Wu M., Gopukumar D., Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomed. Inform. Insights*. 2016. V. 8. P. 1–10.

Рукопись поступила в редакцию 21.12.2016.
Дата опубликования 10.03.2017.