УДК 004.91:577.2.01
DOI 10.25205/1818-7900-2018-16-2-95-103

А. А. Цхай [1, 2], **С. В. Мурзинцев** [3]

[1] *Институт водных и экологических проблем СО РАН*
ул. Молодежная, *1*, *Барнаул*, *656038*, *Россия*

[2] *Алтайский государственный технический университет*
пр. Ленина, *46*, *Барнаул*, *656038*, *Россия*

[3] *Алтайский государственный университет*
пр. Ленина, *61*, *Барнаул*, *656049*, *Россия*

taa1956@mail.ru, *o.l00@yandex.ru*

ИСПОЛЬЗОВАНИЕ ГОРИЗОНТАЛЬНО МАСШТАБИРУЕМОЙ ИНФРАСТРУКТУРЫ ПРИ ПОИСКЕ СХОДСТВА В ГЕНОМНЫХ ДАННЫХ ЭКОСИСТЕМ

Рассмотрена проблема выявления генетического сходства при анализе баз данных (БД) геномов организмов. Такая проблема возникает с развитием методов метагеномики, сравнительной геномики, технологий высокопроизводительного секвенирования ДНК, а также инструментов оценки и прогнозирования состояния экосистем. Для быстрого сравнения геномов с целью выявления повторяющихся наборов нуклеотидов разработана специализированная компьютерная система. Из-за большого объема данных, возникающих при обработке исходной информации, осуществлен переход к нереляционным БД, как к более гибким и масштабируемым. В качестве основы подхода использованы распределенная нереляционная БД MongoDB и алгоритм обработки данных Winnowing. При использовании нереляционной БД для выявления генетического сходства предложен вариант представления отпечатков структурных вариаций геномов в виде «ключ – значение». Выполнена программная реализация разработанной модели. Проведены вычислительные эксперименты: 1) загрузка данных в БД с использованием одной и трех шард (серверов, где хранятся данные и осуществляются поиск и обработка информации); 2) поиск совпадений выбранных наборов нуклеотидов с БД геномов с использованием одной и трех шард; 3) расчет скорости поиска геномов в БД; 4) расчет скорости загрузки геномов в БД. Результатом экспериментов стало подтверждение возможности использования предложенного способа поиска генетического сходства. Продолжение работы может быть в направлениях: 1) решения задачи об определении момента, когда необходимо добавлять узел к кластеру при возрастании рассматриваемого количества выбранных наборов нуклеотидов и увеличении числа геномов в БД организмов; 2) практического наполнения создаваемой БД как можно большим количеством реальных геномов организмов; 3) исследования геномных нарушений с целью оценки вероятности генетических отклонений на этапе распознавания потенциально возможного неблагоприятного развития организма.

Ключевые слова: сравнение геномов, большие данные, нереляционные базы данных, алгоритмы поиска повторений, биоинформатика.

Введение

Изучение структурно-функциональной организации живых организмов продолжает оставаться актуальным направлением, развивающимся на стыке биологии и информатики [1]. В этой связи использование компьютерных технологий при исследовании геномов (см., например, [2–5]) получило широкое распространение в мире.

Цхай А. А., *Мурзинцев С. В.* Использование горизонтально масштабируемой инфраструктуры при поиске сходства в геномных данных экосистем // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 2. С. 95–103.

В числе современных крупных геномных проектов - несколько БД: ENSEMBL [1] – совместный проект организаций EMBL-EBI (Германия) и Sanger Centre (Великобритания) с целью создания программной системы для автоматического создания аннотаций геномов эукариотов. Проект ENSEMBL ориентирован на соответствие следующим критериям: точный, автоматический анализ данных генома; аннотации, основанные на текущих, своевременно обновляемых данных; доступность полученных данных через Интернет. GenBank [2] – БД нуклеотидных последовательностей, которая поддерживается NCBI (Национальным центром биотехнологической информации США). Крупнейшая интегрированная поисковая система ENTREZ, которая создана и поддерживается NCBI, используется для анализа нуклеотидных и аминокислотных последовательностей, библиографии (PubMed), полных геномов (Genomes), а также трехмерных структур белков (MMDB). При этом поиск данных о ДНК и белках не ограничивается только ресурсами GenBank, но распространяется и на другие доступные по сети хранилища информации. International Nucleotide Sequence Database Collaboration объединяет три крупнейшие коллекции нуклеотидных последовательностей: EMBL-EBI и GenBank (NCBI) и DDBJ (Япония). Информационный ресурс KEGG (Kyoto Encyclopedia of Genes and Genomes) [3] создается Институтом химических исследований (Kyoto University, Japan). Эта база знаний имеет обширные возможности для работы со всеми крупными мировыми информационными ресурсами. Обновление KEGG происходит ежедневно.

Существует ряд геномных браузеров, в том числе: NCBI, UCSC Genome Browser [4], ENSEMBL. Эти браузеры используются для получения и визуализации детальной справочной информации о геномах. Разработанная же авторами специализированная система предназначена для обработки справочной информации, а именно для быстрого сравнения геномов организмов с целью выявления повторяющихся наборов нуклеотидов.

В связи с развитием технологий высокопроизводительного секвенирования ДНК продолжается рост объема геномной информации в мире. Таким образом, несмотря на развитие международных БД и браузеров, остается важной задача сравнения протяженных геномных последовательностей, процессинга данных, поиска совпадений.

К настоящему времени предложены специальные форматы геномных данных (например, FASTA [5]), для поиска сходств в определенных классах геномных последовательностей разработаны компьютерные средства (например, BLAST [6]), идет работа над переносом рабочих процессов аппаратной оптимизации быстрого поиска геномных данных на виртуальные мощности облаков [6].

Нельзя не упомянуть вклад российских специалистов в разработку компьютерных методов решения задач метагеномики, сравнительной геномики, определения полиморфизмов, скрининга мутаций, транскриптомного профилирования и т. д. (например, [7–9]).

Данное исследование посвящено решению задачи сравнения выбранных наборов нуклеотидов с геномами организмов в реально существующем и постоянно актуализирующемся информационном ресурсе. Для этого необходимо сравнить содержащиеся в БД геномы организмов {G} с выбранными наборами нуклеотидов {N} в виде символьных последовательностей произвольной длины.

В работе источником элементов используемой БД являлась KEGG GENOME, включающая расшифрованные представления (около пяти тысяч организмов), находящиеся в свободном доступе. В настоящее время объем данных этого информационного ресурса можно оценить приблизительно в пять ТБ, что представляет технический вызов для существующих вычислительных мощностей.

В связи с большим объемом данных, возникающим при обработке исходной информации, был осуществлен переход от реляционных БД к нереляционным как к более гибким и масс-

[1] ENSEMBL. URL: http://www.ensembl.org (дата обращения 23.04.2018).
[2] GenBank. URL: http://www.ncbi.nlm.nih.gov/GenBank/ GenBankOverview.html (дата обращения 23.04.2018).
[3] KEGG GENOME. URL: http://www.genome.jp/kegg/catalog/org_list.html (дата обращения 23.04.2018).
[4] UCSC Genome Browser. URL: https://genome.ucsc.edu/ (дата обращения 23.04.2018).
[5] What is FASTA format? URL: https://zhanglab.ccmb.med.umich.edu/FASTA/ (дата обращения 23.04.2018).
[6] Basic Local Alignement Search Tool (BLAST). URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi/ (дата обращения 23.04.2018).

штабируемым. Такой путь решения подобных проблем был обоснован в работе [10], в том числе для поиска нуклеотидных полиморфизмов и сходства геномных последовательностей [11]. В этом случае необходим инструментарий, позволяющий осуществлять поиск в горизонтально масштабируемой информационной системе (далее ИС) с возрастающими ресурсами в процессе развития. Данная ИС должна характеризоваться достаточной «эластичностью», т. е. способностью к расширению, без существенных инвестиций в инфраструктуру ИС, а также доработку программного обеспечения и алгоритмов.

Один из вариантов решения данной задачи – это использование структуры хранения данных и разработка поискового механизма на основе средств нереляционной распределенной БД MongoDB и алгоритма Winnowing [12–14]. В статье представлены результаты, ориентированные на реализацию такого подхода для сравнения геномов и поиска отклонений в их структуре.

Схема распределенной информационной системы

Распределенная ИС, созданная в работе на основе использования семи серверов, включает нереляционную БД MongoDB. Три сервера (UbuntuSlaveA, UbuntuSlaveB, UbuntuSlaveC) отвечают за хранение данных. На серверах находятся две распределенные БД: база данных нуклеотидных последовательностей и БД организмов. Три сервера: UbuntuSlaveD, UbuntuSlaveE, UbuntuSlaveF – это управляющие сервера, отвечающие за запись данных и их хранение на серверах UbuntuSlaveA, UbuntuSlaveB, UbuntuSlaveC. Сервер UbuntuMaster – это управляющий сервер, который предназначен для управления всем кластером, созданным с использованием технологии MongoDB.

На рис. 1 представлен персональный компьютер пользователя, с которого осуществляются первоначальная загрузка данных для формирования БД геномов организмов и БД нуклеотидных последовательностей, а также последующие запросы к созданным БД.

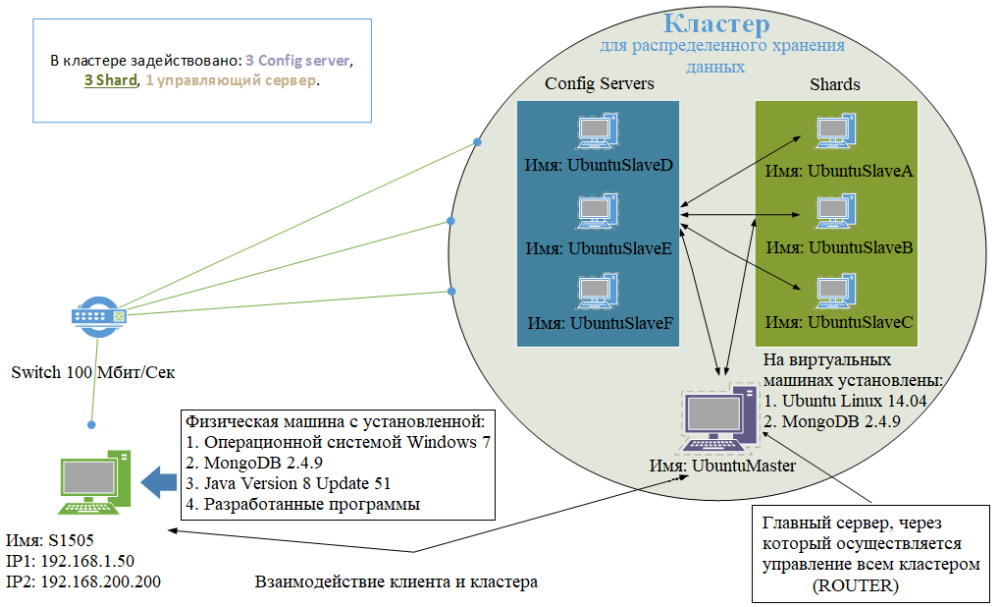

Рис. 1. Схема распределенной информационной системы для поиска выбранных наборов нуклеотидов

Данные о геномах, полученные из ENSEMBL, представляли собой наборы нуклеотидов, например, для мыши типа [CTAAAGTATA TATGAGTAAA CTTGGTCTGA CAGTTACCAA TGCTTAATCA GTGAGGCACC...] и т. п. В этом же виде они переносились в создаваемую в работе «вторичную» БД, со структурой, соответствующей описанной выше информационной системе.

Модель хранения геномов и их отклонений в нереляционной базе данных MongoDB

В исследованиях жизнедеятельности организмов, входящих в экосистемы различного уровня, используется представление о сходстве объектов (например, [15]). В литературе встречается достаточно много вариантов поиска сходства текстов, которые основаны на *n*-граммах (например, [16]). Методы поиска, использующие *n*-граммы, основаны на создании отпечатков документов, которые позволяют идентифицировать части при попарном сравнении символьных последовательностей. Подробный обзор сравнения алгоритмов с точки зрения производительности выполнен в работе [13].

Одним из алгоритмов, который основан на применении *n*-грамм, является алгоритм Winnowing [12]. Пример применения алгоритма Winnowing – для решения задачи сравнения текстов - рассмотрен в работе [14]. Показано, что выбранная *n*-грамма помещается в таблицу БД в виде триады *<хеш> – <документ> – <позиция хеша в документе>* и представляется записями как минимум в двух реляционно-связанных таблицах. Выполненный анализ демонстрирует высокую производительность при росте количества сравниваемых документов, но, как следствие, возникают скоростные ограничения в процессе использования реляционной модели.

В качестве решения проблемы производительности в работе предложена нереляционная модель типа *ключ – значение*, где хеш подстроки (*n*-граммы) документа выступает в качестве ключа, а значение представляет собой двумерную таблицу, содержащую значения *<документ> – <позиция хеша в документе>* для всех документов, имеющих совпадающую символьную последовательность. Таким образом, для поиска используется только одно ключевое значение, представленное значением хеш-функции от *n*-граммы. Благодаря этому возникает возможность выполнения параллельных запросов к нескольким узлам кластера одновременно.

Описанная модель была положена в основу структуры БД и протестирована в ряде экспериментов, результаты которых представлены ниже.

Экспериментальные результаты

Проведены четыре эксперимента по оценке следующих характеристик:

1) скорость загрузки геномов организмов и выбранных наборов нуклеотидов в БД MongoDB на одном компьютере в сравнении со случаем, когда БД распределена между тремя компьютерами в кластере;

2) зависимость скорости загрузки от размера выбранного набора нуклеотидов;

3) скорость загрузки геномов организмов на один компьютер в сравнении со случаем, когда БД распределена между тремя компьютерами в кластере;

4) скорость сравнения геномов организмов с БД выбранных наборов нуклеотидов на одном компьютере в сравнении со случаем, когда БД распределена между тремя компьютерами в кластере.

Применимость подхода оценивалась через два параметра. Первый параметр – это масштабируемость инфраструктуры, а второй – скорость работы. Скорость работы распределенной ИС оценивалась при сравнении выбранного набора нуклеотидов с геномами организмов по времени загрузки последних в БД. Объем данных, использованных в экспериментах, составлял порядка 20 Гб.

Оценка масштабирования инфраструктуры выполнялась при построении прототипа ИС с использованием различных конфигураций. Были протестированы конфигурации без разделения записей по индексу, а также с разделением записи по индексу между одной шардой и тремя. Шарды – это сервера, где хранятся данные и осуществляются поиск и обработка информации.

При тестировании ПО на всех конфигурациях не потребовалось изменения программного обеспечения, что позволяет утверждать, что прототип ИС может работать при разделении БД между неограниченным количеством рабочих станций. Данное свойство придает распределенной ИС эластичность при росте объема данных.

Оценка скорости загрузки

Первый эксперимент был предназначен для оценки скорости загрузки массива геномов организмов в БД. Данная операция не является критической для поисковой системы. Средняя скорость загрузки для различных конфигураций (мс/КБ текста): без шард – 76,2; 3 шарды – 17,9. Зависимость скорости загрузки от размера геномов организмов показана на графиках, которые представлены на верхней панели рис. 2. При загрузке данных производится обновление индексов. Производительность операции добавления данных зависит от объема уже загруженной информации. Из этих графиков видно, что время загрузки документов уменьшается в зависимости от количества шард в кластере.

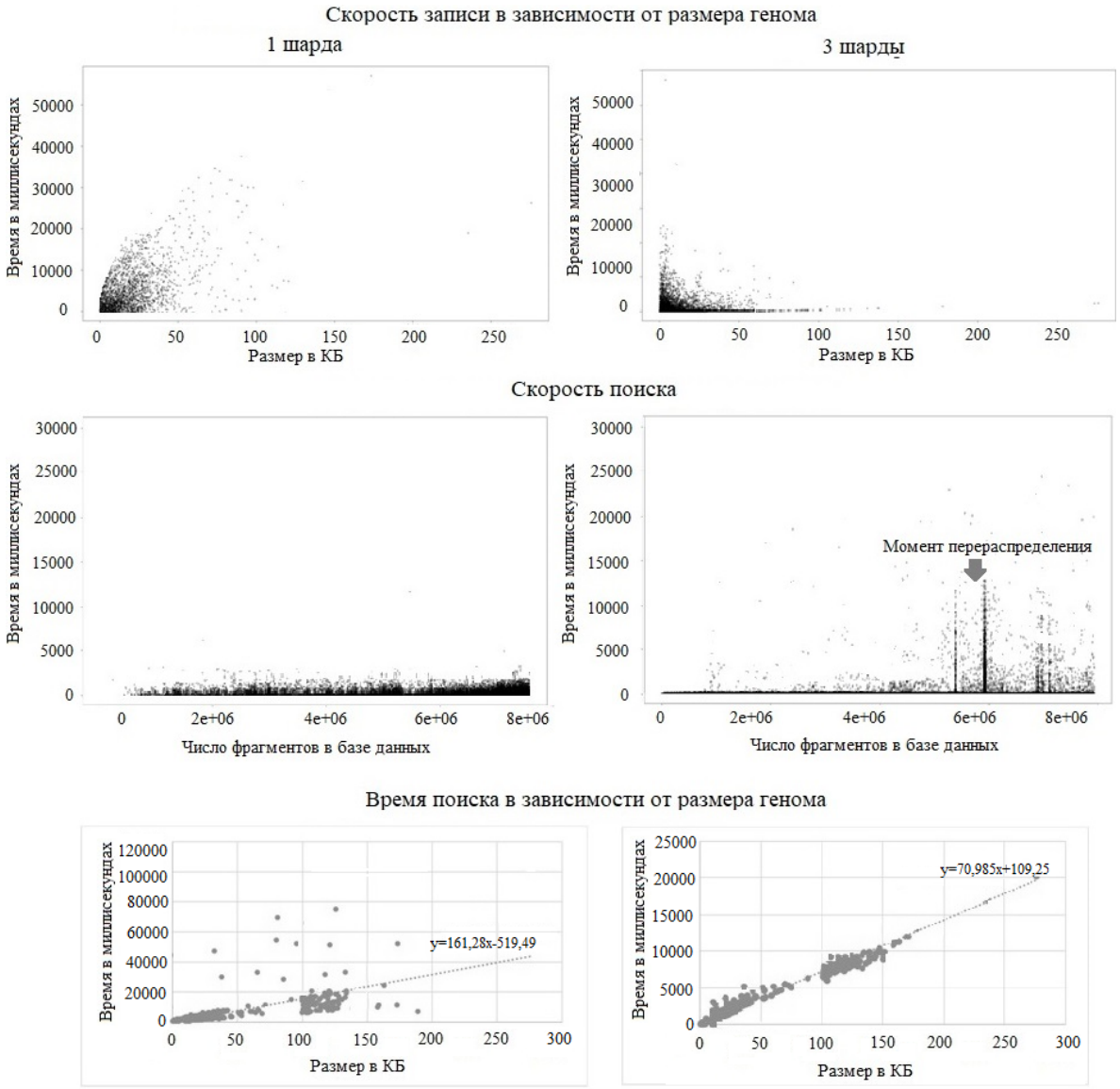

Рис. 2. Скорость поиска информации (одна и три шарды)

Из графиков нижней панели рис. 2 следует, что зависимость относительного увеличения времени, требуемого на загрузку, от объема уже загруженных геномов организмов близка к линейной. Также видно, что скорость поиска уменьшается. Отсюда можно сделать вывод, что скорость работы информационной системы с тремя шардами выше, чем с одной. Увеличение количества узлов хранения с одного до трех дает возможность оценки относительной

скорости загрузки, показывающей меньший рост при увеличении числа узлов хранения. Это позволяет утверждать, что с точки зрения загрузки данных система является горизонтально масштабируемой.

На графике, расположенном в средней панели рис. 2, видно, что в момент перераспределения индекса по шардам скорость загрузки геномов организмов в БД уменьшается (стрелкой показан момент перераспределения).

Оценка скорости поиска отклонений в геномах организмов

Целью второго эксперимента была оценка скорости поиска по БД геномов организмов. Для проведения оценки массив из геномов организмов был загружен в БД MongoDB, которая была установлена на загруженные ранее архитектуры. При осуществлении поиска случайным образом из списка всех рассматриваемых наборов нуклеотидов были выбраны некоторые с определенными нарушениями в порядке нуклеотидов и выполнено сравнение отобранных вариантов с геномами заполненной БД. Результатом поиска было совпадение набора хешей.

Эксперимент оценивал не качество, а скорость поиска. Графики зависимости времени поиска от размера приведены на рис. 2 (нижняя часть). Средняя скорость поиска (мс/КБ) составила: без шард – 121,8; 3 шарды – 72,3. Из этого можно сделать вывод о том, что средняя скорость загрузки без использования распределенной инфраструктуры как минимум в два раза выше, чем при хранении данных в распределенных узлах.

Результаты экспериментов свидетельствуют о принципиальной пригодности разработанной модели для поиска сходства в созданной БД, заполненной реальной информацией о геномах. Оценка скорости работы созданной модели, говорит о ее приемлемости с точки зрения производительности для поиска сходства геномных последовательностей организмов и, как следствие, дает возможность выявлять отклонения в развитии на ранних этапах диагностики.

Алгоритм, основанный на использовании *n*-грамм, в процессе экспериментов с созданной программной платформой показал достаточно хорошие результаты при поиске сходства наборов нуклеотидов в БД геномов.

Разработанная программная модель позволила протестировать алгоритм Winnowing и распределить «отпечатки» геномов организмов и выбранных наборов нуклеотидов по кластеру в нереляционной БД MongoDB. Разработанный программный комплекс позволил провести эксперименты, которые способствуют выработке новых стратегий и алгоритмов по улучшению поиска выбранных наборов нуклеотидов в геномах организмов.

Заключение

Оценим объем данных, которые возникнут в геномике при развитии постгеномных технологий секвенирования. Международный проект «1 000 геномов» уже привел к секвенированию порядка 100 000 индивидуальных геномов, и рост продолжается. Если оценить размер генома человека в 3 ГБ (без вспомогательной информации и аннотации), и население планеты в 7 миллиардов, то получим порядка 210 тысяч петабайт информации.

Рост геномной информации по секвенированию геномов лабораторных животных – крыс и мышей – как результат экспериментов в биомедицине, при тестировании фармпрепаратов, приводит к росту БД различных организмов. Так, например, следует отметить важность охарактеризованных ресурсов и системы быстрого поиска в них для развития средств мониторинга сообществ гидробионтов в водных экосистемах. Модели водных экосистем нового пятого поколения содержат в качестве основополагающих внутренних параметров геномные характеристики видов планктона [17].

Вышесказанное делает актуальным продолжение работы в нескольких направлениях.

Первое направление – это решение задачи об определении момента, когда необходимо добавлять узел к кластеру при возрастании рассматриваемого количества выбранных наборов нуклеотидов и увеличении числа геномов в БД организмов.

Второе – это практическое наполнение БД как можно большим количеством реальных геномов организмов. Использование полученных результатов в междисциплинарных геномных системных исследованиях позволило бы говорить о детализации и развитии модели в перспективном плане.

Третье - это исследование геномных нарушений с целью оценки вероятности генетических отклонений на этапе распознавания потенциально возможного неблагоприятного развития организма.

При индустриальном использовании БД геномов человека, животных и растений, в сотрудничестве со специалистами-генетиками, вышесказанное выглядит все реальнее с учетом продолжающегося «бума» исследований в данной области [18].

Благодарность. Авторы благодарны коллективу, выпускающему журнал, за внимание к работе и ценные советы.

Список литературы

1. Биоразнообразие и динамика экосистем: информационные технологии и моделирование / Отв. ред. В. К. Шумный, Ю. И. Шокин, Н. А. Колчанов, А. М. Федотов. Новосибирск: Изд-во СО РАН, 2006. 648 с.

2. *Lesk A. M.* Introduction to Genomics. 3rd ed. New York: Oxford University Press, 2017. 544 p.

3. *Dankar F. K., Ptitsyn A., Dankar S. K.* The development of large-scale de-identified biomedical databases in the age of genomics-principles and challenges // Hum. Genomics. 2018. Vol. 12 (1). P. 19. DOI 10.1186/s40246-018-0147-5.

4. *Langmead B., Nellore A.* Cloud computing for genomic data analysis and collaboration // Nat. Rev. Genet. 2018. Vol. 19 (4). P. 208–219. DOI 10.1038/nrg.2017.113.

5. *Nakagawa H., Fujita M.* Whole genome sequencing analysis for cancer genomics and precision medicine // Cancer Sci. 2018. Vol. 109 (3). P. 513–522. DOI 10.1111/cas.13505.

6. *Hong D., Rhie A., Park S. S., Lee J., Ju Y. S., Kim S. et al.* FX: an RNA-Seq. analysis tool on the cloud // Bioinformatics. 2012. Vol. 28. P. 721–723.

7. *Орлов Ю. Л., Брагин А. О., Медведева И. В., Гунбин И. В., Деменков П. С., Вишневский О. В., Левицкий В. Г., Ощепков В. Г., Подколодный Н. Л., Афонников Д. А., Гроссе И., Колчанов Н. А.* ICGenomics: программный комплекс анализа символьных последовательностей геномики // Вавиловский журнал генетики и селекции. 2012. Т. 16 (4/1). С. 732–741.

8. *Boekhorst R., Naumenko F. M., Orlova N. G., Galieva E. R., Spitsina A. M., Chadaeva I. V., Orlov Y. L., Abnizova I. I.* Computational problems of analysis of short next generation sequencing reads // Вавиловский журнал генетики и селекции. 2016. Т. 20 (6). С. 746–755. DOI 10.18699/VJ16.191.

9. *Спицина А. М., Орлов Ю. Л., Подколодная Н. Н., Свичкарев А. В., Дергилев А. И., Чен М., Кучин Н. В., Черных И. Г., Глинский Б. М.* Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования ДНК // Программные системы: теория и приложения. 2015. Т. 1 (24). С. 157–174.

10. Вычислительные методы, алгоритмы и аппаратурно-программный инструментарий параллельного моделирования природных процессов / М. Г. Курносов [и др.]; отв. ред. В. Г. Хорошевский; Рос. акад. наук, Сиб. отд-ние, Ин-т физики полупроводников им. А. В. Ржанова [и др.]. Новосибирск: Изд-во СО РАН, 2012. 335 с. (Интеграционные проекты СО РАН; вып. 33).

11. *Peise E., Fabregat-Traver D., Aulchenko Yu., Bientinesi P.* Algorithms for Large-scale Whole Genome Association Analysis. 2013. DOI 10.1145/2488551.2488577.

12. *Schleimer S., Wilkerson D., Aiken A.* Winnowing: Local Algorithms for Document Fingerprinting // International Conference on Management of Data (ACM SIGMOD. Proceedings). San Diego, 2003. P. 76–85.

13. *Faro S., Lecroq T.* The exact online string matching problem: A review of the most recent results // ACM Computing Surveys. 2013. Vol. 45, № 13. P. 42–50. http://dx.doi.org/10.1145/2431211.2431212.

14. *Цхай А. А., Бутаков С. В., Мурзинцев С. В., Ким Л. С.* Обнаружение плагиата с использованием нереляционных баз данных // Вестн. алтайской науки. 2015. № 1. С. 280–285.

15. *Федотов А. М., Чураев Р. Н.* О подходах к построению мер сходства между объектами // Математические модели эволюции и селекции: Сб. ст. Новосибирск, 1977. С. 120–131.

16. *Дягилев В. В., Цхай А. А., Бутаков С. В.* Архитектура сервиса определения плагиата, исключающая возможность нарушения авторских прав // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2011. Т. 9, № 3. С. 23–29.

17. *Jørgensen S. E.* Structurally dynamic models: a new promising model type // Environmental Earth Sciences. 2015. № 74. DOI 10.1007/s12665-015-4735-6.

18. *Park S. T., Kim J.* Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. // Int. Neurourol. J. 2016. Vol. 20, № 2. P. 76–83. http://doi.org/10.5213/inj.1632742.371

Материал поступил в редколлегию 28.02.2018

A. A. Tskhai [1, 2], **S. V. Murzintsev** [3]

[1] *Institute for Water and Environmental Sciences*
1 Molodezhnaya Str., *Barnaul*, *656038*, *Russian Federation*

[2] *I. I. Polzunov Altai State Technical University*
46 Lenin Ave., *Barnaul*, *656038*, *Russian Federation*

[3] *Altai State University*
61 Lenin Ave., *Barnaul*, *656049*, *Russian Federation*

taa1956@mail.ru, *o.l00@yandex.ru*

THE USE OF A HORIZONTALLY SCALABLE INFRASTRUCTURE IN THE SEARCH FOR GENETIC SIMILARITY IN BIODIVERSITY

The problem of rapid detection of genetic similarity in the analysis of databases (DB) of genomes of individuals of ecosystems at various levels is considered. The distributed non-relational DB MongoDB and the Winnowing data processing algorithm are used as the basis for creating the information system. Using a non-relational database to identify genetic similarity, a variant of representing the prints of the structural variations of the genomes in the form of «key-value» was proposed, a program implementation of the developed model was carried out, and computational experiments were carried out, which confirmed the possibility of using the proposed method of genetic similarity search, for example, in a personified analysis of deviations in the gene level.

Keywords: similarity of genomes, large data, nonrelational databases, search algorithms for repetitions.

References

1. Biodiversity and ecosystem dynamics: information technology and modeling / answering. Ed. V. C. Shumny, Yu. I. Shokin, N. A. Kolchanov, A. M. Fedotov. Novosibirsk, SB RAS Publishing House, 2006, 648 p. (in Russ.)

2. Lesk A. M. Introduction to Genomics. 3rd ed. New York: Oxford University Press, 2017, 544 p.

3. Dankar F. K., Ptitsyn A., Dankar S. K. The development of large-scale de-identified biomedical databases in the age of genomics-principles and challenges. *Hum. Genomics*, 2018, vol. 12 (1), p. 19. DOI 10.1186/s40246-018-0147-5.

4. Langmead B., Nellore A. Cloud computing for genomic data analysis and collaboration. *Nat. Rev. Genet.*, 2018, vol. 19 (4), p. 208–219. DOI 10.1038/nrg.2017.113.

5. Nakagawa H., Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci*., 2018, vol. 109 (3), p. 513–522. DOI 10.1111/cas.13505.

6. Hong D., Rhie A., Park S. S., Lee J., Ju Y. S., Kim S. et al. FX: an RNA-Seq. analysis tool on the cloud. *Bioinformatics*, 2012, vol. 28, p. 721–723.

7. Orlov Yu. L., Bragin A. O., Medvedeva I. V., Gunbin I. V., Demenkov P. S., Vishnevsky O. V., Levitsky V. G., Oshchepkov D. G., Podkolodny N. L., Afonnikov D. A., Grosse I., Kolchanov N. A. ICGenomics: a program complex for analysis of symbol sequences in genomics. *Vavilov Journal of Genetics and Breeding*, 2012, vol. 16, p. 732–741 (in Russ.)

8. Boekhorst R., Naumenko F. M., Orlova N. G., Galieva E. R., Spitsina A. M., Chadaeva I. V., Orlov Y. L., Abnizova I. I. Computational problems of analysis of short next generation sequencing reads. *Vavilov Journal of Genetics and Breeding*, 2016, vol. 20 (6), p. 746–755. DOI 10.18699/VJ16.191.

9. Spitsina A. M., Orlov Yu. L., Svichkarev A. V., Dergilev A. I., Ming C., Kuchin N. V., Chernykh I. G., Glinskij B. M. Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. *Program Systems: Theory and Applications*, 2015, vol. 6, p. 157–174. DOI 10.25209/2079-3316-2015-6-1-157-174. (in Russ.)

10. Computational methods, algorithms and hardware-software tools for parallel modeling of natural processes / M. G. Kurnosov [et al]; Resp. edited by V. G. Khoroshevsky; Russ. Akad. Sciences, Sib. Branch, In-t semiconductor physics named A. V. Rzhanov [et al.]. Novosibirsk, Publishing house of SB RAS, 2012,335 p. (Integration projects SB RAS; issue. 33) (in Russ.)

11. Peise E., Fabregat-Traver D., Aulchenko Yu., Bientinesi P. Algorithms for Large-scale Whole Genome Association Analysis. 2013. DOI 10.1145/2488551.2488577.

12. Schleimer S., Wilkerson D., Aiken A. Winnowing: Local Algorithms for Document Fingerprinting. *International Conference on Management of Data (ACM SIGMOD. Proceedings)*. San Diego, 2003, p. 76–85.

13. Faro S., Lecroq T. The exact online string matching problem: A review of the most recent results. *ACM Computing Surveys*, 2013, vol. 45, № 13, p. 42–50. http://dx.doi.org/10.1145/2431211.2431212.

14. Tskhai A. A., Butakov S. V., Murzincev S. V., Kim L. S. Obnaruzhenie plagiata s ispol'zovaniem nerelyatsionnykh baz dannykh [Plagiarism detection using non-relational databases]. *Vestnik altajskoj nauki*, 2015, no. 1, p. 280–285. (in Russ.)

15. Fedotov A. M., Churaev R. N. On approaches to the construction of measures of similarity between objects. *Mathematical models of evolution and selection*. Novosibirsk, 1977, p. 120–131. (in Russ.)

16. Dyagilev V. V., Tskhai A. A., Butakov S. V. Arkhitektura servisa opredeleniya plagiata, isklyuchayushhaya vozmozhnost' narusheniya avtorskikh prav [The architecture of the service definition of plagiarism, which excludes the possibility of copyright infringement]. Vestnik NSU. Series: Information Technologies, 2011, vol. 9, no. 3, p. 23–29. (in Russ.)

17. Jørgensen S. E. Structurally dynamic models: a new promising model type. *Environmental Earth Sciences*, 2015, no. 74. DOI 10.1007/s12665-015-4735-6.

18. Park S. T., Kim J. Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. *Int. Neurourol. J*., 2016, vol. 20, № 2, p. 76–83. http://doi.org/10.5213/inj.1632742.371

For citation:

Tskhai A. A., Murzintsev S. V. The Use of a Horizontally Scalable Infrastructure in the Search for Genetic Similarity in Biodiversity. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 2, p. 95–103. (in Russ.)

DOI 10.25205/1818-7900-2018-16-2-95-103