На правах рукописи

Казаков Сергей Владимирович

Автоматизация сборки генома и сравнительного анализа метагеномов для обучения геномной биоинформатике

Специальность 05.13.06 – Автоматизация и управление технологическими процессами и производствами (образование)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2016

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики».

Научный руководитель: доктор технических наук, профессор
Шалыто Анатолий Абрамович

Официальные оппоненты: **Садовский Михаил Георгиевич**,
доктор физико-математических наук, ведущий научный сотрудник отдела вычислительной математики Института вычислительного моделирования ФГБНУ «Федеральный исследовательский центр «Красноярский научный центр Сибирского отделения РАН»

Спирин Сергей Александрович,
кандидат физико-математических наук, старший научный сотрудник отдела математических методов в биологии Научно-исследовательского института физико-химической биологии имени А.Н. Белозерского ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова».

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет»

Защита состоится 22 декабря 2016 г. в 15 часов 30 минут на заседании диссертационного совета Д 212.227.06 при Санкт-Петербургском национальном исследовательском университете информационных технологий, механики и оптики по адресу: 197101, г. Санкт-Петербург, Кронверкский просп., д. 49., ауд. 431.

С диссертацией можно ознакомиться в библиотеке Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики по адресу: 197101, г. Санкт-Петербург, Кронверкский просп., д. 49. и на сайте: `fppo.ifmo.ru/?page1=16&page2=52&page_d=1&page_d2=145757`.

Автореферат разослан 15 ноября 2016 года.

Ученый секретарь
диссертационного совета Д 212.227.06,
к.ф.-м.н., доцент  Холодова Светлана Евгеньевна

Общая характеристика работы

Актуальность темы исследования. За последние 40 лет основным методом получения информации о клетке живого существа и процессах, протекающих в ней, стало **секвенирование**. *Секвенирование дезоксирибонуклеиновой кислоты* (*ДНК*) – процесс определения последовательности нуклеотидов в молекуле ДНК. Эта молекула обеспечивает хранение и передачу генетической информации. Иными словами, секвенирование позволяет получить по физической субстанции ДНК или РНК (рибонуклеиновая кислота) ее нуклеотидную последовательность в цифровом (электронном) виде. При этом процесс секвенирования состоит из двух частей – физико-химической (непосредственный процесс «чтения» нити ДНК или РНК) и компьютерной (обработка полученных «сырых» данных). Компьютерная часть обычно называется «*сборка генома*». Ее наличие обусловлено тем, что физико-химическая часть секвенирования не позволяет получить всю цепочку ДНК целиком, которая необходима для изучения генома, а только маленькие ее фрагменты (*чтения*). Компьютерная часть позволяет решить эту проблему. Таким образом, *сборка генома* – процесс получения больших фрагментов генома (ДНК) из небольших чтений. *Сборка генома de novo* – задача сборки еще неизвестного генома. Методы компьютерного анализа составляют основу *геномной биоинформатики*, которая является составной частью *биоинформатики* и ориентирована на изучение геномов живых организмов.

По мере развития технологий секвенирования развивались и программы для сборки генома. Они становились более сложными, и, как правило, строились на основе модульной архитектуры – состояли из набора модулей, каждый из которых ответственен за выполнение своей задачи (этапа). Эта архитектура обычно является иерархической – каждый этап может состоять из подэтапов. Другими особенностями программ по сборке генома являются их ориентированность на специалистов узкой направленности (в основном биоинформатиков), возможность работы только под операционной системой *Linux* и требование больших объемов оперативной памяти для работы. Именно поэтому такие программы обычно запускают на серверах или на кластерах. При этом описанные особенности затрудняют использование таких программ для обучения, так как их установка, настройка и запуск на компьютерах обучающихся, которые обычно являются персональными, плохо осуществимы.

Со временем развитие технологий секвенирования привело к расширению границ его применимости. Секвенирование стало применяться не только для получения генома отдельного организма, но и для анализа набора геномов (метагеном). *Метагеном* – совокупность геномов микроорганизмов (бактерий, архей, вирусов) из одной среды обитания (почва, водные ресурсы, кишечник человека и т. п.). Компьютерный анализ таких данных включает в себя методы сравнительного анализа набора метагеномов, методы определения таксономического состава метагенома (какие бактерии находится в метагеноме) и другие. Этому направлению в биоинформатике также необходимо обучать. При этом существующие программы для анализа метагеномов, как и в случае с программами сборки генома, плохо подходят для такого использования, так как являются сложными и труднонастраиваемыми.

Таким образом, разработка автоматизированных методов сборки генома *de novo* и сравнительного анализа метагеномов, которые применимы в образовательном процессе, является **актуальной задачей**.

В соответствии с паспортом специальности 05.13.06 «Автоматизация и управление технологическими процессами и производствами (образование)» диссертация относится к следующей области исследований: «20. Разработка автоматизированных систем научных исследований».

Цель диссертационной работы – разработка автоматизированных методов сборки генома *de novo* и сравнительного анализа метагеномов, оптимизированных по объему используемой оперативной памяти, а также расширение их области применимости для использования при обучении.

Для этого решаются следующие **основные задачи**:

1. Произвести анализ существующих методов сборки генома *de novo* и сравнительного анализа метагеномов по возможности их применения в образовательном процессе.
2. Разработать автоматизированный метод сборки генома *de novo* на основе совместного применения графа де Брейна и графа перекрытий, оптимизированный по объему используемой памяти.
3. Разработать автоматизированный метод сравнительного анализа метагеномов на основе анализа графа де Брейна, оптимизированный по вычислительным ресурсам.
4. Произвести экспериментальные сравнения программ, реализующих предлагаемые и существующие методы, по метрикам качества получаемых результатов и необходимым вычислительным ресурсам.

Научная новизна. В работе получены следующие новые научные результаты, которые выносятся на защиту:

1. Автоматизированный метод сборки генома *de novo* на основе совместного применения графа де Брейна и графа перекрытий. Программа, разработанная на основе этого метода, позволяет производить сборку малых и средних по размеру геномов на персональных компьютерах под управлением трех самых распространенных операционных систем (*Windows, macOS/OS X, Linux*), что отличает ее от существующих программ.
2. Автоматизированный метод сравнительного анализа метагеномов, основанный на анализе компонент связности в графе де Брейна. Разработанный метод отличается от существующих тем, что он выполняет «упрощенную» сборку метагеномов вместо стандартной, позволяя значительно сократить требуемые вычислительные ресурсы.

Методы исследования. В работе используются методы теории графов, дискретной математики, теории сложности и математической статистики.

Положения, выносимые на защиту. На защиту выносятся:

1. Автоматизированный метод сборки генома *de novo* на основе совместного применения графа де Брейна и графа перекрытий.
2. Автоматизированный метод сравнительного анализа метагеномов, основанный на анализе компонент связности в графе де Брейна.

Отличия разработанных методов от существующих указаны в разделе *Научная новизна*.

Достоверность научных положений и выводов, полученных в диссертации, подтверждается корректным обоснованием постановок задач, точной формулировкой критериев, результатами экспериментов по использованию предложенных в диссертации методов и их статистическим анализом.

Теоретическое значение работы состоит в том, что показана применимость алгоритмов сборки генома *de novo* и сравнительного анализа метагеномов для работы на персональных компьютерах, обычно применяемых при обучении геномной биоинформатике.

Практическое значение работы состоит в том, что разработанные методы реализованы в виде исполняемых программ с открытым исходным кодом, которые позволяют производить сбоку генома *de novo* и сравнительный анализ метагеномов на персональных компьютерах обучающихся под управлением трех самых распространенных операционных систем (*Windows, macOS/OS X, Linux*). При этом предлагаемые методы позволяют существенно уменьшить объем используемой оперативной памяти по сравнению с существующими решениями.

Использование и внедрение результатов работы. Результаты диссертационной работы были использованы в учебном процессе в Санкт-Петербургском политехническом университете Петра Великого в рамках магистерской программы «Прикладная математика и информатика. Биоинформатика» (имеется акт внедрения) и в Университете ИТМО при проведении занятий по биоинформатике на кафедре «Компьютерные технологии» (имеется акт внедрения). Результаты работы также использовались в Казанском (Приволжском) Федеральном Университете в лаборатории масс-спектрометрии при выполнении научно-исследовательских работ по анализу геномов шести малоизученных бактерий (имеется акт внедрения).

Апробация результатов работы. Основные результаты работы докладывались на следующих международных и российских конференциях, семинарах и школах:

- VIII-я Всероссийская межвузовская конференция молодых ученых (2011, Санкт-Петербург);
- Вторая Международная научно-практическая конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине: геномика, протеомика, биоинформатика» (2011, Новосибирск);
- XIX-я Всероссийская научно-методическая конференция «Телематика'2012» (2012, Санкт-Петербург);
- Международная научно-практическая конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине» (2012, 2014, Казань);
- Всероссийская научная конференция по проблемам информатики СПИСОК (2012, 2016, Матмех СПбГУ);
- Первый всероссийский конгресс молодых ученых (2012, Санкт-Петербург);
- Первая Международная школа-конференция студентов, аспирантов и молодых ученых «Биомедицина, материалы и технологии XXI века» (2015, Казань);
- Летняя школа по биоинформатике (2015, Москва);
- VII-я Международная научная конференция «Компьютерные науки и информационные технологии» (2016, Саратов);
- de novo Genome Assembly Assessment Project workshop (dnGASP) (2011, Барселона);

– «Bioinformatics 2012» Conference (2012, Стокгольм);
– 8th International Conference on Intelligent Systems and Agents (2014, Лиссабон);
– Moscow Conference on Computational Molecular Biology (2015, Москва).

Личный вклад автора. Автором лично разработаны: идея совместного использования графа де Брейна и графа перекрытий для сборки генома, методы исправления ошибок и сборки контигов при сборке генома, метод выполнения «упрощенной сборки» при анализе метагеномов, а также реализация всех предложенных методов.

Публикации. Основные результаты по теме диссертации изложены в 19 публикациях, четыре из которых изданы в российских журналах, рекомендованных ВАК, четыре – в изданиях, индексируемых в международных базах цитирования *Web of Science* и *Scopus*. Доля диссертанта в работах, выполненных в соавторстве, указана в списке публикаций.

Свидетельства о регистрации программ для ЭВМ. В рамках диссертационной работы получено пять свидетельств о регистрации программ для ЭВМ:

– № 2011614454 от 06.06.2011 г. «Программное средство для удаления ошибок из набора чтений нуклеотидной последовательности»;
– №2012616774 от 27.07.2012 г. «Программное средство для сборки квазиконтигов из парных чтений»;
– № 2013616471 от 09.07.2013 г. «Программное средство, реализующее алгоритм поиска перекрытий между квазиконтигами»;
– № 2013619155 от 26.09.2013 г. «Программное средство, реализующее запуск этапов сборки генома через графический интерфейс пользователя»;
– № 2013660881 от 21.11.2013 г. «Программное средство, реализующее алгоритм упрощения графа перекрытий при сборке геномных последовательностей».

Участие в научно-исследовательских работах. Некоторые результаты диссертации были получены при выполнении следующих научно-исследовательских работ: «Разработка методов сборки генома, сборки транскриптома и динамического анализа протеома» (Государственный контракт № 14.B37.21.0562, 2012–2013 гг.) и «Разработка метода сборки геномных последовательностей на основе восстановления фрагментов по парным чтениям» (Государственный контракт № 16.740.11.0495, 2011–2013 гг.). Автор является победителем конкурса грантов для студентов вузов, расположенных на территории Санкт-Петербурга, аспирантов вузов, отраслевых и академических институтов, расположенных на территории Санкт-Петербурга 2013 и 2014 гг., темы проектов: «Разработка алгоритма упрощения графа перекрытий при сборке геномных последовательностей» (2013 г.) и «Сборка контигов геномных последовательностей на основе принципа максимального правдоподобия» (2014 г.).

Объем и структура диссертации. Диссертация состоит из введения, четырех глав, заключения и двух приложений. Объем диссертации составляет 171 страницу, с 37 рисунками, 16 таблицами и тремя листингами. Список источников содержит 119 наименований.

Содержание работы

В **первой главе** приводится обзор предметной области, включающий в себя определения базовых понятий, существующих методов секвенирования ДНК, метагеномного секвенирования, подходов к сборке геномных последовательностей *de novo* и анализу данных метагеномного секвенирования. Производится анализ возможности использования существующих программ для обучения биоинформатике, приводятся таблицы сравнения. На основе результатов обзора формулируются задачи, решаемые в диссертации.

Дезоксирибонуклеиновая кислота (*ДНК*) – химическое вещество, биологический полимер, обеспечивающий хранение и передачу из поколения в поколение генетической информации. *Секвенирование ДНК* – определение нуклеотидной последовательности по имеющемуся образцу ДНК. *Биоинформатика* – совокупность методов и подходов, применяемых для обработки биологических данных. *Геномная биоинформатика* – область биоинформатики, ориентированная на изучение геномов живых организмов.

Благодаря существенным технологическим прорывам в секвенировании ДНК за последние десятилетия, сам процесс секвенирования стал достаточно быстрым (от нескольких часов до двух недель) и относительно дешевым (цена приближается к $1000 за секвенирование генома человека). Все это значительно увеличило число исследований по анализу данных секвенирования. При этом за последние 20 лет появилось новое направление в биологии – *метагеномика,* которое посвящено получению и исследованию геномных последовательностей непосредственно из образцов среды (почв, океанов, кишечника человека) без изолирования конкретных микроорганизмов из них. Совокупность геномных последовательностей из таких сред получила название *метагеном*, а набор организмов, присутствующих в среде, также называют *сообществом микроорганизмов* из данной среды обитания.

Сборка геномной последовательности (*сборка генома*) – процесс получения больших фрагментов генома из небольших чтений, который выполняется с использованием компьютеров. Задача *de novo сборки генома* – задача сборки генома живого существа, для которого геном еще неизвестен. Сложность задачи сборки геномной последовательности обусловлена следующими факторами:

- большой объем исходных данных (может достигать сотен гигабайт);
- сложность структуры генома – наличие в нем повторов и полиморфизмов;
- наличие ошибок в исходных данных, полученных с устройств-секвенаторов.

Задачу сборки геномной последовательности можно поставить следующим образом. Будем рассматривать геном и чтения как строки над алфавитом $\Sigma = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Элементы этого алфавита (символы) обозначают нуклеотиды (азотистые основания) геномной последовательности: аденин (A), цитозин (C), гуанин (G) и тимин (T). Исходными данными для задачи сборки генома является набор строк $S = \{s_i\}$, где каждое чтение $s_i \in \Sigma^*$, $i \in [1, N]$, N – число чтений, L_i – длина чтения s_i. Задача сборки состоит в нахождении такой строки $g \in \Sigma^*$, что для каждого чтения s_i выполняется условие: s_i входит в строку g как подстрока. Дополнительным условием оптимизации может выступать требование минимизации длины строки g.

Активное развитие методов сборки генома происходило в течение последних 20 лет. Первоначально для этих целей использовались простые жадные алгоритмы (например, алгоритм *GREEDY*), которые обладали существенными недостатками.

Современные сборщики являются сложными программами с модульной архитектурой. Процесс сборки при этом разбит на несколько этапов:

- исправление ошибок в исходных данных;
- восстановление непрерывных фрагментов генома по парным чтениям; такие фрагменты называются *квазиконтигами*;
- сборка расширенных непрерывных фрагментов (*контигов*) из квазиконтигов;
- построение *скэффолдов* – упорядоченных контигов с оценкой расстояния между ними.

Квазиконтиги отличаются от контигов тем, что они имеют небольшую длину (обычно распределение длин известно), тогда как контиги – максимальные по длине непрерывные фрагменты генома, которые не удается расширить.

Процесс сборки традиционно базируется на основе одной из двух структур данных – на графе перекрытий или на графа де Брейна.

Графом перекрытий (*overlap graph*) называется взвешенный ориентированный граф, каждой вершине которого сопоставлена строка s_i, а две вершины соединяются ребром, если соответствующие строки перекрываются (вес ребра при этом равен длине перекрытия). Пример графа перекрытий показан на рисунке 1.

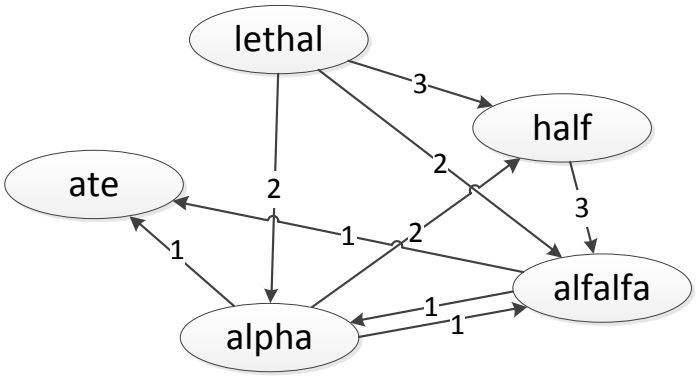


Рисунок 1 – Граф перекрытий

Графом де Брейна (*de Bruijn graph*) степени k над алфавитом Σ называется ориентированный граф, вершинами которого являются строки фиксированной длины k (k-меры) из Σ^k, а ребра – строки e из Σ^{k+1}, причем две вершины соединены ребром, если из первого k-мера можно получить второй путем добавления одного символа в конец первого k-мера и удаления одного символа из начала. Пример графа де Брейна для $k = 3$ показан на рисунке 2.

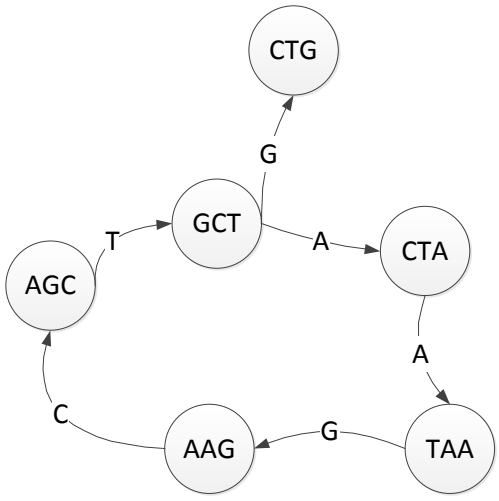


Рисунок 2 – Пример графа де Брейна

В первой главе также приведено сравнение (таблица 1) существующих сборщиков по следующим критериям, определяющим возможность их использования в учебном процессе:

1. Возможность свободного использования.
2. Возможность работы на персональном компьютере (при небольшом объеме оперативной памяти).
3. Кроссплатформенность.
4. Наличие графического интерфейса пользователя.
5. Возможность простого запуска (одной командой, без выбора дополнительных параметров сборки).

Таблица 1 – Сравнение существующих программ *de novo* сборки генома

Название сборщика	**1**	**2**	**3**	**4**	**5**
ABySS	+	–	–	–	+
Allpaths-LG	+	–	–	–	–
CLC Genomics Workbench	–	+	+	+	+
MaSuRCA	+	–	–	–	–
Minia	+	+	–	–	–
Newbler	±	–	–	+	–
SPAdes	+	–	–	–	+
SparseAssembler	+	+	–	–	–
Velvet	+	–	–	–	–

Из таблицы 1 следует, что наиболее подходящим для образовательного процесса является сборщик *CLC Genomics Workbench*, который, однако, **не распространяется свободно, является платным** и весьма дорогим. Остальные сборщики обычно не являются кроссплатформенными, не имеют графического интерфейса или требуют «специальных» знаний для их запуска. Эти особенности приводят к усложнению их использования при обучении геномной биоинформатике.

При анализе данных метагеномного секвенирования выделяют несколько главных вопросов, которые обычно вызывают интерес. Среди них – определение *таксономического состава* (набора организмов, которые присутствуют в среде), *функционального состава* (набора генов, содержащихся в геномах организмов из исследуемой среды), а также вопросы сравнения двух и более метагеномов. При сравнении метагеномов также ставится задача выделения конкретных признаков (*features*), которые отличают одно сообщество микроорганизмов от другого. Такими признаками могут быть, например, конкретные организмы (или их ДНК), присутствующие в одном метагеноме и отсутствующие в другом, или различия в представленности ДНК в разных сообществах. Благодаря таким признакам удается научиться различать несколько метагеномов между собой и выделять характерные особенности каждого из них, что крайне важно при анализе группы сред обитания микроорганизмов или при регулярном наблюдении за одной средой.

Существует много подходов для сравнительного анализа метагеномов. Их можно разделить на несколько классов по базовому принципу, на котором основан подход. Среди них – традиционные методы, основанные на «выравнивании» (поиске соответствий) метагеномных данных на каталог референсных (известных) геномов; методы, основанные на совместной сборке метагеномных данных; абстрактные

методы, базирующиеся на выделении и анализе k-мерного спектра (набора всех k-меров) метагеномов. Методы из каждой группы имеют существенные недостатки, которые ограничивают область их применимости. Среди них – требование наличия репрезентативной базы геномов, необходимость больших вычислительных ресурсов для работы, малоинформативность получаемых признаков и сложный запуск имеющихся программ. Указанные недостатки приводят к необходимости искать другие решения для анализа набора метагеномов, а требование больших вычислительных ресурсов и сложный запуск программ усложняют их использование в учебном процессе.

Первая глава завершается формулировкой задач, решаемых в диссертационной работе.

Во **второй главе** приводится описание предлагаемого метода сборки генома *de novo*. Метод основан на совместном применении графа де Брейна и графа перекрытий и позволяет использовать преимущества обеих структур данных.

Для анализа требуемых ресурсов на хранение исходных данных автором были получены следующие зависимости объема памяти (в байтах) для хранения графа перекрытий (*ov-graph*) и графа де Брейна (*db-graph*):

$$M_{ov-graph} = N \cdot L \cdot \frac{2}{8} + N \cdot \left\lfloor \frac{N(L-L_{ov})}{G} \right\rfloor \cdot 6; \quad (1)$$

$$M_{db-graph} = \left(K_g + \lfloor N \cdot L \cdot p_{er} \rfloor \cdot \frac{k(L-k+1)}{L}\right) \cdot 8 \cdot \frac{4}{3}, \quad (2)$$

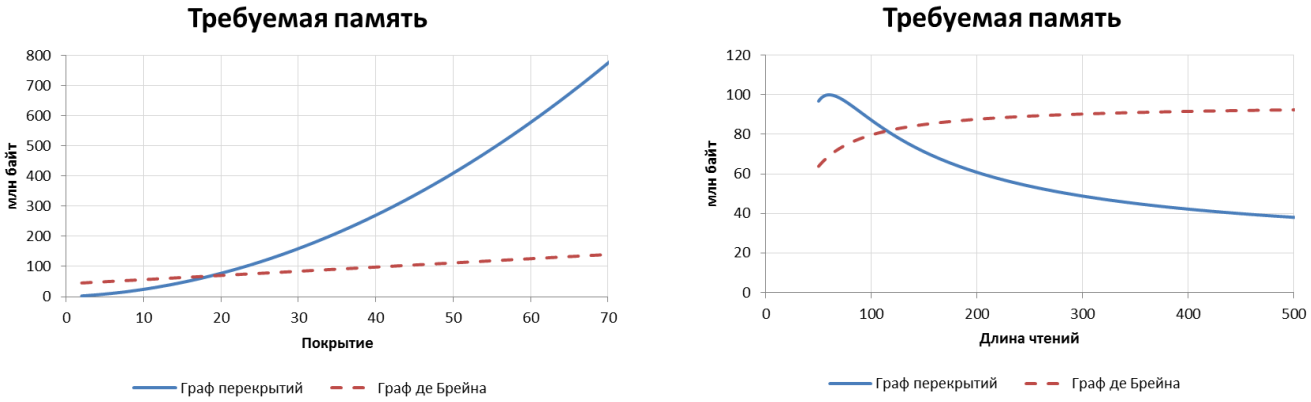
где N – число чтений, L – средняя длина чтения, G – длина генома, L_{ov} – длина перекрытия, k – размер k-мера, K_g – число безошибочных k-меров, p_{er} – вероятность ошибки в одной позиции чтения.

Покрытие генома чтениями, обозначаемое c, определяется как среднее число чтений, покрывающих некоторую позицию в геноме, и вычисляется по формуле:

$$c = \frac{N \cdot L}{G} = \frac{\sum_{i=1}^{N} L_i}{G}.$$

Полученные зависимости (1) и (2) отображаются на двух графиках, представленных на рисунке 3. На первом из них фиксируется длина чтения $L = 100$ нукл., а переменными являются число чтений N и покрытие генома чтениями c. На втором графике при фиксированном покрытии $c = 20$ переменными являются средняя длина чтения L и число чтений N (с сохранением равенства $c = 20$).

Из представленных графиков следует, что при большом покрытии генома чтениями ($c > 40$) граф перекрытий требует значительно больше памяти, чем граф де Брейна. При небольшом покрытии ($c = 20$) и длине чтений $L = 100$ нукл. оба графа используют одинаковый объем памяти, однако при увеличении длины чтения граф перекрытий требует меньше памяти. Поэтому для достижения лучших результатов автором было предложено применять граф де Брейна на этапе сборки квазиконтигов из парных чтений, где данных может быть много, а длина чтения небольшой (обычно 35–100 нукл.), и граф перекрытий на этапе сборки контигов, где исходные данные (квазиконтиги) обычно достаточно длинные (в среднем по 500 нукл.).



а) б)

Рисунок 3 – Зависимости требуемой памяти для хранения графов перекрытий и де Брейна при а) изменении покрытия генома чтениями, б) изменении длины чтений

Сборку генома *de novo* предлагается проводить следующим образом:

– исправление ошибок секвенирования (основано на анализе *k*-мерного спектра);
– сборка квазиконтигов из парных чтений (выполняется на графе де Брейна);
– сборка контигов из квазиконтигов (выполняется на графе перекрытий);

Схема предлагаемого решения показана на рисунке 4.

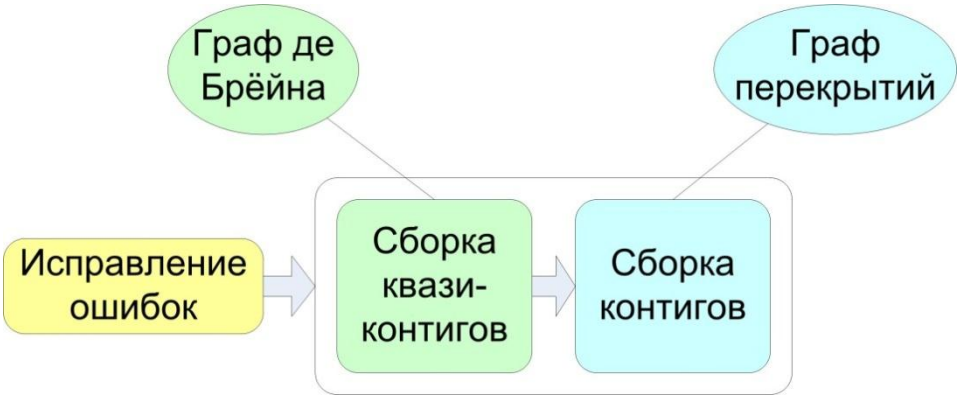


Рисунок 4 – Архитектура предлагаемого решения

Алгоритм исправления ошибок основан на частотном анализе содержащихся в чтениях *k*-меров (подстрок длины *k*) и состоит в следующем. Сначала для каждого *k*-мера вычисляется частота его присутствия в исходных чтениях. На основании этих данных все *k*-меры разделяются на две группы – «надежные» *k*-меры (встречаются чаще некоторого эвристически выбранного порога) и «подозрительные» (редко встречаемые). Метод исправления ошибок состоит в замене каждого «подозрительного» *k*-мера на «близкий» надежный *k*-мер, который может быть получен из подозрительного путем изменений небольшого числа нуклеотидов в нем.

Сборка квазиконтигов использует **граф де Брейна**, который строится из «надежных» *k*-меров, полученных на предыдущем этапе. Непрерывный фрагмент генома (квазиконтиг), с концов которого получены парные чтения, восстанавливается на данном этапе с помощью графа де Брейна. Для этого производится поиск пути в графе между вершинами, соответствующими *k*-мерам из парных чтений. При этом используется подход *meet-in-the-middle*, при котором поиск происходит одновременно с двух сторон, для каждой из которых применяется обход в ширину. Этот метод имеет меньшую временную сложность, чем

стандартные подходы обхода в глубину или ширину, позволяя быстрее найти подходящий путь в графе. Оценка сложности приведена в тексте диссертации.

Сборка контигов из квазиконтигов основана на подходе *overlap-layout-consensus* (*OLC*) с применением **графа перекрытий**. Используемый подход состоит в последовательном выполнении нескольких этапов – поиск перекрытий между квазиконтигами (*overlap*), построение графа перекрытий по ним, упрощение полученного графа, поиск подходящих путей в графе (*layout*) и нахождение консенсуса для них (*consensus*). Под консенсусом понимается процесс построения контигов путем выбора наиболее часто встречаемого нуклеотида в чтениях для каждой позиции отдельно.

Отличительной особенностью применения *OLC* в предлагаемом решении является то, что в случае нехватки оперативной памяти при выполнении некоторого этапа, все чтения разбиваются на несколько групп, и этап выполняется отдельно для каждой из них. При этом на такое выполнение требуется меньше памяти, чем при обработке всех чтений сразу.

Новизной предлагаемого сборщика является то, что он использует разные графы на разных этапах сборки, что позволило сократить объем используемой памяти. Схожий подход используется и в сборщике *MaSuRCA*, однако автором он был предложен в 2011 г. на конференции *de novo Genome Assembly Assessment Project workshop* (*dnGASP*), в то время как сборщик *MaSuRCA* начал разрабатываться в 2012 г., а первая статья о нем вышла в 2013 г.

Предложенный подход реализован на языке программирования *Java*, что позволяет использовать программу на трех самых распространенных операционных системах (*Windows, macOS/OS X, Linux*). Исходный код размещен в открытом доступе в сети Интернет по адресу `http://genome.ifmo.ru/ru/assembler`. Разработанная программа имеет графический интерфейс пользователя и возможность простого запуска (достаточно указать только исходные файлы). Благодаря этому сборщик удовлетворяет всем критериям, используемым в таблице 1, и, следовательно, может быть использован при обучении.

Автоматизация сборки генома состоит в том, что все этапы и подэтапы, необходимые для проведения сборки, выполняются автоматически (как единый процесс). При этом программа имеет модульную архитектуру, что позволяет независимо изменять отдельные модули.

В этой главе также описывается экспериментальное исследование, которое было проведено для сравнения разработанной программы с существующими. Было использовано три набора данных, свободно доступных в сети Интернет, – стандартные данные секвенирования бактериального генома, данные секвенирования с большим покрытием (~500) бактериального генома и данные секвенирования среднего по размеру генома – 14-ой хромосомы человека (длина хромосомы – 107,4 млн. нуклеотидов).

Экспериментальное исследование показало, что предлагаемый сборщик *ITMO Genome Assembler* отработал эффективнее по памяти, чем сборщики *SPAdes, Velvet, MaSuRCA, Newbler* и *Minia*, однако использовал больше памяти, чем сборщик *SparseAssembler*. При этом по числу найденных генов в сборке предлагаемый подход незначительно уступает *SPAdes* (разница от **0.1%** до **0.6%**), но показал лучшие результаты по сравнению со всеми остальными сборщиками (разница может достигать **81.7%**). При этом сборщики *SPAdes, Velvet, MaSuRCA* и *Newbler* не

смогли произвести сборку среднего по размеру генома при ограничении в памяти в 16 Гб в отличие от предлагаемого решения, которое уложилось в 4 Гб памяти. Полное описание результатов сравнения приведено в тексте диссертационной работы.

В **третьей главе** приводится описание разработанного метода сравнительного анализа метагеномов *MetaFast*, основанного на анализе компонент связности в графе де Брейна. Предлагаемый метод частично основан на описанном в главе 2 методе *de novo* сборки генома, однако, в отличие от него, использует непарные чтения. Метод состоит из четырех этапов:

1. Выполнение «упрощенной» сборки для каждого метагенома отдельно (выполняется на графе де Брейна).
2. Построение общего графа де Брейна для всех метагеномов и выделение компонент связности в нем (каждая компонента далее используется как единичный признак, определение признака дано в главе 1).
3. Вычисление характеристического вектора для каждого метагенома. Для каждой компоненты подсчитывается сколько раз *k*-меры, которые образуют компоненту, присутствуют в чтениях рассматриваемого метагенома.
4. Вычисление попарного расстояния между метагеномами на основе полученных характеристических векторов (с использованием индекса Брея-Кертиса).

На **первом этапе** для каждого метагенома отдельно выполняются построение графа де Брейна и выделение неветвящихся путей. Найденные короткие фрагменты получаются в несколько раз короче, чем обычные контиги, но дальше такие фрагменты используются как контиги.

На **втором этапе** все найденные контиги объединяются в один граф де Брейна, после этого производится анализ компонент связности в нем. Компонента связности отличается от пути тем, что она позволяет объединить одинаковые части генома, которые могут иметь небольшие вариации у разных микроорганизмов. Части путей в таких компонентах обычно соответствуют генам или другим важным частям геномов микроорганизмов, присутствующих в среде. Небольшие по размеру компоненты не используются для анализа, так как они обычно малоинформативны, а слишком большие компоненты разбиваются на меньшие путем выделения наиболее представленных из них *k*-меров, присутствующих в большем числе метагеномов. После этого каждая компонента используется как единичный признак, который помогает выявить отличительные особенности отдельных метагеномов.

На **третьем этапе** для каждого метагенома вычисляется характеристический вектор, каждый элемент которого – сколько раз *k*-меры, которые образуют некоторую компоненту, присутствуют в чтениях рассматриваемого метагенома. На **четвертом этапе** производится расчет матрицы расстояния по вычисленным характеристическим векторам на основе индекса Брея-Кертиса (*Bray-Curtis dissimilarity*).

Новизна предлагаемого подхода состоит в применении «частичной» сборки метагеномов вместо стандартной. Благодаря таким изменениям получаются относительно короткие контиги вместо длинных, что положительно сказывается на последующем анализе, позволяя сохранить разнообразие в геномных последовательностях микроорганизмов, а также увеличить скорость работы алгоритма и уменьшить требуемую память.

Благодаря такому подходу становится возможным устранить недостаток в требовании больших вычислительных ресурсов, который присущ методам, основанным на совместной сборке метагеномных данных. При этом удается сохранить их преимущества в осмысленности получаемых признаков и независимости от базы референсных геномов по сравнению с другими методами.

Предлагаемый подход реализован на языке программирования *Java* и является кроссплатформенным. Исходный код размещен в открытом доступе в сети Интернет по адресу `https://github.com/ctlab/metafast`. Реализация также имеет графический интерфейс пользователя и возможность простого запуска, что делает ее удобной для использования при обучении.

В этой главе также приводятся результаты экспериментального исследования, целью которого было сравнение предлагаемого подхода с существующими. Исследования проводились на четырех наборах данных, свободно доступных в сети Интернет. Они включали в себя симулированные данные (набор № 1), стандартный реальный набор данных секвенирования из 30 метагеномов (набор № 2), большой набор реальных данных (157 метагеномов, набор № 3) и реальный набор данных секвенирования из плохо изученной среды – виромы озер (набор № 4).

Сравнение предложенного подхода проводилось с четырьмя традиционными методами, основанными на использовании каталога референсных геномов, и с методом, основанным на анализе совместной сборки метагеномов.

Экспериментальная проверка подтвердила высокую точность получаемых результатов предлагаемого решения (корреляция Спирмена $r = 0.96$ с истинной матрицей расстояния на первом наборе данных, $r = 0.81–0.91$ на наборах № 2,3,4), высокую производительность метода (время работы сравнимо со временем работы традиционных методов, основанных на выравнивании на каталог известных геномов, и в 40–185 раз быстрее метода, основанного на совместной сборке). Требуемая при этом память сравнима с необходимой памятью для традиционных методов и в 5–13 раз меньше, чем у метода, основанного на совместной сборке (на наборах данных № 2, 4). Также подтвердилось преимущество предложенного метода в независимости от базы референсных геномов по сравнению с традиционными подходами, что особенно важно при изучении малоисследованных сообществ микроорганизмов. Подробное описание результатов сравнения приведено в тексте диссертационной работы.

В **четвертой главе** приводится описание внедрений результатов диссертационной работы.

Результаты диссертационной работы были использованы в учебных процессах в Санкт-Петербургском политехническом университете Петра Великого и в Университете ИТМО при проведении лекционных занятий и лабораторных работ, а также использовались в Казанском (Приволжском) Федеральном Университете при выполнении научно-исследовательских работ по анализу геномов бактерий.

Заключение

В диссертационной работе получены следующие результаты:

1. Предложен автоматизированный метод сборки генома *de novo* на основе совместного применения графа де Брейна и графа перекрытий.
2. Предложен автоматизированный метод сравнительного анализа метагеномов, основанный на анализе компонент связности в графе де Брейна.

3. Предложенные методы были реализованы на языке программирования *Java*. Разработанные программы свободно доступны в сети Интернет.
4. Проведены экспериментальные сравнения разработанных программ с существующими на разных наборах данных секвенирования. Результаты сравнений подтверждают применимость разработанных подходов.

Статьи в журналах из перечня ВАК

1. *Казаков С. В., Шалыто А. А.* Анализ геномных и метагеномных данных в образовательных целях // Компьютерные инструменты в образовании. – 2016. – 3. – С. 5–15. – 0,688 п. л. / 0,65 п. л.
2. *Сергушичев А. А.*, *Александров А. В., Казаков С. В., Царев Ф. Н., Шалыто А. А.* Совместное применение графа де Брейна, графа перекрытий и микросборки для *de novo* сборки генома // Известия Саратовского университета. Новая серия. Серия Математика. Механика. Информатика. – 2013. – Т. 13, вып. 2, ч. 2. – С. 51–57. – 0,438 п. л. / 0,06 п. л.
3. *Александров А. В., Казаков С. В., Мельников С. В., Сергушичев А. А., Царев Ф. Н.* Метод сборки контигов геномных последовательностей на основе совместного применения графов де Брюина и графов перекрытий // Научно-технический вестник информационных технологий, механики и оптики. – 2012. – 6(82). – С. 93–98. – 0,375 п. л. / 0,06 п. л.
4. *Александров А. В., Казаков С. В., Мельников С. В., Сергушичев А. А., Царев Ф. Н., Шалыто А. А.* Метод исправления ошибок в наборе чтений нуклеотидной последовательности // Научно-технический вестник Санкт-Петербургского государственного университета информационных технологий, механики и оптики. – 2011. – 5(75). – С. 81–84. – 0,25 п. л. / 0,042 п. л.

Публикации в рецензируемых изданиях, индексируемых Web of Science или Scopus

5. *Ulyantsev V., Kazakov S., Dubinkina V., Tyakht A., Alexeev D.* MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data // Bioinformatics. – 2016. – 32 (18). – P. 2760–2767. – 0,5 п. л. / 0,109 п. л.
6. *Kazakov S., Shalyto A.* Overlap graph simplification using edge reliability calculation / Proceedings of the 8th International Conference on Intelligent Systems and Agents 2014 (ISA 2014). – 2014. – P. 222–226. – 0,313 п. л. / 0,3 п. л.
7. *Bradnam K., Fass J., Alexandrov A., Baranay P., Bechner M.* и др., всего 91 человек. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species // GigaScience. – 2013. – 2(10). – P. 1–31. – 1,938 п. л. / 0,02 п. л.
8. *Alexandrov A., Kazakov S., Melnikov S., Sergushichev A., Shalyto A., Tsarev F.* Combining de Bruijn graph, overlap graph and microassembly for de novo genome assembly / Proceedings of the 12th annual conference in bioinformatics "Bioinformatics 2012". Stockholm, Sweden. – 2012. – P. 72. – 0,063 п. л. / 0,01 п.л.

Другие публикации

9. *Kazakov S., Ulyantsev V., Dubinkina V., Tyakht A., Alexeev D.* MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data / Proceedings of the International Moscow Conference on Computational Molecular Biology 2015 (MCCMB'15). – Moscow, 2015. – 0,25 п. л. / 0,125 п. л.

10. *Казаков С. В., Шалыто А. А.* Сборка генома *de novo* на персональном компьютере / Материалы Всероссийской научной конференции по проблемам информатики СПИСОК-2016. – 2016. – С. 220–222. – 0,188 п. л. / 0,178 п. л.
11. *Казаков С. В., Шалыто А. А.* Сборка генома *de novo* из данных высокопроизводительного секвенирования на персональном компьютере / Сборник трудов VII Международной научной конференции «Компьютерные науки и информационные технологии». – 2016. – С. 178–181. – 0,25 п. л. / 0,22 п. л.
12. *Казаков С. В., Ульянцев В. И., Дубинкина В. Б., Тяхт А. В., Алексеев Д. Г.* MetaFast: высокопроизводительный сравнительный анализ метагеномов на основе графа де Брейна / Сборник тезисов I международной школы-конференции «Биомедицина, материалы и технологии XXI века». – Казань, 2015. – С. 98. – 0,063 п. л. / 0,05 п. л.
13. *Ульянцев В. И., Казаков С. В., Дубинкина В. Б., Тяхт А. В., Алексеев Д. Г.* MetaFast – программное средство для высокопроизводительного сравнительного анализа метагеномов / Сборник трудов IV международной научно-практической конференции «Постгеномные методы анализа в биологии, лабораторной и клинической медицине». – Казань, 2014. – С. 103. – 0,063 п. л. / 0,021 п. л.
14. *Александров А. В., Казаков С. В., Мельников С. В., Сергушичев А. А., Царев Ф. Н., Шалыто А. А.* Метод сборки контигов геномных последовательностей на основе совместного применения графов де Брюина и графов перекрытий / Материалы Всероссийской научной конференции по проблемам информатики СПИСОК-2012. – 2012. – С. 415–418. – 0,125 п. л. / 0,021 п. л.
15. *Александров А. В., Казаков С. В., Мельников С. В., Сергушичев А. А., Царев Ф. Н.* Метод сборки контигов геномных последовательностей на основе совместного применения графов де Брюина и графов перекрытий / Сборник тезисов докладов I всероссийского конгресса молодых ученых. – 2012. – 1. – С. 235–237. – 0,094 п. л. / 0,019 п. л.
16. *Александров А. В., Казаков С. В., Мельников С. В., Сергушичев А. А.* Метод *de novo* сборки контигов геномных последовательностей на основе совместного применения графов де Брюина и графов перекрытий / Труды XIX Всероссийской научно-методической конференции «Телематика'2012». – 2012. – Т. 1. – С. 183–185. – 0,094 п. л. / 0,023 п. л.
17. *Александров А. В., Казаков С. В., Мельников С. В., Сергушичев А. А., Царев Ф. Н., Шалыто А. А.* Совместное применение графов де Брейна, графов перекрытий и микросборки для *de novo* сборки генома / Сборник тезисов III международной конференции «Постгеномные методы анализа в биологии, лабораторной и клинической медицине». – Казань, 2012. – С. 45–46. – 0,125 п. л. / 0,021 п. л.
18. *Александров А. В., Казаков С. В., Мельников С. В., Сергушичев А. А., Царев Ф. Н.* Метод сборки геномных последовательностей на основе совместного применения графов де Брюина и графов перекрытий / Тезисы II Международной научно-практической конференции «Постгеномные методы анализа в биологии, лабораторной и клинической медицине: геномика, протеомика, биоинформатика». – Новосибирск, 2011. – Т. 2. – С. 188. – 0,063 п. л. / 0,013 п. л.
19. *Александров А. В., Исенбаев В. В., Казаков С. В., Сергушичев А. А., Мельников С. В., Царев Ф. Н.* Метод сборки генома с помощью восстановления его фрагментов по парным чтениям / Сборник тезисов докладов конференции молодых ученых. – 2011. – 1. – С. 220. – 0,063 п. л. / 0,01 п. л.

Тиражирование и брошюровка выполнены в учреждении «Университетские телекоммуникации»
197101, Санкт-Петербург, Саблинская ул., 14. Тел. (812) 233 46 69.
Объем 1,0 у.п.л. Тираж 100 экз.