МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

ВОССТАНОВЛЕНИЕ ПОСЛЕДОВАТЕЛЬНОСТИ ПО ЕЁ ПОДПОСЛЕДОВАТЕЛЬНОСТЯМ В ЗАДАЧЕ СБОРКЕ ГЕНОМА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Диева Ильи Николаевича

Научный руководитель:
зав. кафедрой ИиП, к. ф.-м. н, доцент _____ Огнева М. В.
подпись, дата

Зав. кафедрой:
к. ф.-м. н, доцент _____ Огнева М. В.
подпись, дата

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы. В последние годы наука биоинформатика активно развивается, становясь ключевым направлением в исследованиях биологии и медицины. Рост объема биологических данных — от геномных последовательностей до данных о структуре белков и метаболических путей — требует высоких вычислительных мощностей и новых алгоритмов для их обработки и анализа. Одной из наиболее актуальных задач в современной биоинформатике является задача восстановления геномных последовательностей по имеющимся подпоследовательностям, известная как задача сборки генома. Эта задача возникает из-за особенностей секвенирования, при котором геном изучаемого организма «разбивается» на множество коротких фрагментов, или ридов. Эти фрагменты могут содержать ошибки и дублирующуюся информацию, что усложняет задачу их объединения в единую и точную последовательность, отражающую исходный геном [1].

Сборка генома — это не только техническая, но и важная биологическая задача, так как точность сборки напрямую влияет на последующий анализ генетической информации. Достоверное восстановление генома необходимо для понимания биологических функций, изучения эволюционных связей, идентификации генов, ответственных за определенные черты и заболевания [2].

В данной работе будут рассматриваться различные подходы к решению задачи сборки генома и возможности использования методов машинного обучения для восстановления геномных последовательностей.

Цель бакалаврской работы — реализация и сравнительный анализ различных способов для сборки генома.

Поставленная цель определила **следующие задачи:**

1. ознакомиться с процессами секвенирования и ассемблирования, изучить строение и состав генома;

2. рассмотреть методики для сборки генома с помощью картирования и сборки de novo, проанализировать их области применимости;
3. изучить два ключевых подхода к сборке генома de novo: на основе графа перекрытий и графа де Брюйна, оценить их преимущества и недостатки;
4. провести сравнение существующих сборщиков, основанных на графах, а также методов, использующих машинное обучение, для восстановления геномных последовательностей;
5. реализовать структуры графа перекрытий и графа де Брюйна и осуществить с их помощью сборку генома;
6. разработать и обучить модель машинного обучения для отделения ложных перекрытий от реальных;
7. провести сравнительный анализ качества сборки при использовании классического алгоритма и алгоритма с ML-компонентом;
8. оценить влияние внедрения методов машинного обучения на структуру графа и точность финальной сборки.

Методологические основы исследования в области обработки геномных последовательностей представлены в работах П. Певзнера [3], Ж. Сетубала [4], Ф. Крика [5], Е. Сидорова [6], С.В. Казакова [7].

Практическая значимость бакалаврской работы заключается в исследовании основных подходов к сборке генома на основе графов и машинного обучения, сравнительному анализу существующих инструментов для сборки генома, реализации структур графов перекрытий и графов де Брюйна и разработке классификатора перекрытий.

Структура и объём работы. Бакалаврская работа состоит из введения, 7 разделов, заключения, списка использованных источников и 4 приложений. Общий объем работы — 88 страниц, из них 65 страниц — основное содержание, включая 28 рисунков и 16 таблиц, список использованных источников информации – 33 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе «Обзор источников» представлено краткое описание использованных в работе книг и статей, среди которых содержатся исследования в области обработки биологических последовательностей и алгоритмы сборки генома в работах П. Певзнера [4], Ж. Сетубала [5], Ф. Крика [6], перечислены статьи из журналов BMC Bioinformatics, Journal of Clinical Medicine.

Второй раздел «Состав ДНК» посвящен структуре и составу геномных последовательностей.

Нуклеотиды (или азотистые основания) — это базовые структурные элементы, из которых целиком состоит дезоксирибонуклеиновая кислота (ДНК). Всего есть четыре азотистых основания — аденин (А), гуанин (Г), цитозин (Ц), тимин (Т). В задачах биоинформатики, связанных с геномом, ДНК принято считать обыкновенной строкой, которая состоит из четырехбуквенного алфавита. ДНК представляет собой две сцепленные нити из последовательностей азотистых оснований, которые строго упорядочены по отношению к друг другу — они сцеплены водородными связями по особому правилу — правилу комплиментарности: напротив аденина (А) всегда находится тимин (Т) и наоборот, а напротив гуанина (Г) всегда находится цитозин (Ц). Таким образом, количество аденина в цепи ДНК всегда соответствует количеству тимина, а количество цитозина всегда соответствует количеству гуанина. Имея одну цепочку ДНК, по ней всегда можно восстановить вторую.

Третий раздел «Процесс секвенирования» посвящен особенностям выявления генетического материала, которые важно учитывать при сборке генома.

Секвенированием называется процесс точного восстановления порядка следования нуклеотидов в ДНК. Для выявления фрагментов из экстракта используют специальное оборудование, которое называется секвенатором.

Эти фрагменты могут перекрывать друг друга, то есть конец одного участка соответствует началу другого. Тогда, совмещая данные строки по перекрытиям, можно получить целостную цепь ДНК. В результате него получаются контиги (один контиг или несколько) — достаточно длинные участки генома, по которым можно сделать вывод об истинной геномной последовательности. Финальная последовательность называется консенсусной («consensus» — соглашение), то есть такая последовательность, которая, по исходным данным, наиболее вероятно похожа на истинную референсную последовательность («reference» — эталон). Консенсус сравнивается с референсом для получения метрики качества ассемблирования.

Важно понимать, что в результате сборки почти никогда не удается получить единственный контиг — длинную геномную последовательность, так как это идеальный случай. Обычно ассемблирование дает набор из нескольких длинных контигов, которые уже на таком этапе можно анализировать для решения биоинформатических задач или попытаться собрать их в еще более длинные последовательности — скаффолды, которые образуются путем добавления вставок между контигами, которые, предположительно, должны находиться рядом.

Различают два основных подхода к сборке генома: картирование и сборка de novo (лат. «с самого начала», «из нового»). Картирование коротких последовательностей предполагает, что референсная последовательность заранее известна и задача сводится к сопоставлению каждому риду его позиции в референсном геноме. Подход de novo к сборке генома предполагает, что референсной последовательности нет, и сборка осуществляется путем поиска перекрытий и повторяющихся частей в ридах. К алгоритмам сборки генома относятся алгоритмы на графах перекрытий и алгоритмы на графах де Брюйна.

В четвертом разделе «Подходы к сборке генома» описаны структуры графа перекрытий и графа де Брюйна.

Граф перекрытий — это структура, которая используется для представления ридов, полученных при секвенировании, в виде ориентированного графа. В этом графе каждая вершина представляет собой уникальный рид, а ребра соединяют вершины, если два рида перекрываются на некоторое число нуклеотидов, что позволяет моделировать связи между ними. Длина перекрытий может варьироваться. Задача поиска консенсусной последовательности сводится к задаче нахождения гамильтонова пути.

Граф де Брюйна — это структура, в которой каждое ребро представляет собой нуклеотидную последовательность фиксированной длины (k-мер), а вершины ставятся на концах ребер, которые перекрываются на k-1 нуклеотид. Это означает, что, если два k-мера имеют общую подпоследовательность длиной k-1, они будут выходить из одной вершины. Граф де Брюйна позволяет моделировать структуру геномов, упрощая выявление повторяющихся и уникальных участков последовательности. Для построения графа де Брюйна необходимо сначала разбить входные риды на k-меры. Затем для каждого k-мера проверяется, существуют ли другие k-меры, которые перекрываются с ним, и при наличии таких последовательностей создаются соответствующие ребра в графе. Задача сводится к построению эйлерова пути, причем эйлеров путь в таких графах всегда существует, так как к любому (k-1)-меру можно четным числом способов приписать слева и справа нуклеотид, то есть в любую вершину всегда входит четное число ребер и всегда выходит четное число ребер, а значит выполняется критерий существования эйлерового пути.

Пятый раздел «Обзор существующих инструментов» посвящен описанию таких инструментов для сборки генома, как Abyss, Velvet, SPAdes. Каждый инструмент обладает своими достоинствами и недостатками, а также областью решаемых задач. Так, инструмент Abyss выдает более фрагментированную сборку, но с меньшим числом ошибок, а SPAdes и Velvet собирают более длинные контиги, но с большим числом ошибок.

Шестой раздел «Применение методов машинного обучения» посвящен подходам, решающим некоторые проблемы сборки генома с помощью машинного обучения. Существует подход, который сочетает в себе графы де Брюйна и скрытую марковскую модель (СММ). По сравнению с другими ассемблерами этот подход позволяет получить больше контигов с более широким охватом генов. Машинное обучение может быть использовано для группировки ридов на этапе предварительной сборки: рекуррентная нейронная сеть (RNN) может моделировать последовательность оснований в заданном фрагменте, затем оцениваются реальные и моделируемые наборы ридов. Также можно использовать классификатор для обнаружения ошибок и ложных путей в графе, содержащихся в данных секвенирования. Поиск ошибок основывается на анализе базовой частоты и предположении, что редкие основания непосредственно связанны с ошибками.

В седьмом разделе «Практическая часть» представлено подробное описание этапов практической части работы.

Для практической части в качестве референсной последовательности, на которой будет проводиться оценка качества, был выбран геном организма Salmonella enterica. Длина генома Salmonella enterica составляет 4857450 нуклеотидов. В качестве данных с подпоследовательностями был выбран результат эксперимента над Salmonella, файл содержит 2870709 ридов, длина каждой подпоследовательности варьируется в диапазоне от 90 до 110 нуклеотидов.

В практической части работы использовались инструменты:

— сервис usegalaxy — онлайн-платформа, предоставляющая множество инструментов для обработки и анализа геномных последовательностей;

— инструменты Abyss, SPAdes, Velvet для сборки геномов, инструмент Quast для проверки качества сборки по рефренсному геному;

— библиотека Biopython языка программирования Python, предоставляющая различные методы для обработки геномных последовательностей;

— библиотека NetworkX — пакет Python для создания, управления и изучения структуры и функций сложных графов.

В работе были реализованы графы перекрытий и граф де Брюйна. Для собственной реализации графа перекрытий была использована библиотека networkx для Python, предоставляющая инструменты и методы для создания, обработки и анализа графов.

Граф перекрытий строится по набору ридов, для которых попарно вычисляется перекрытие. Также были реализованы методы для упрощения линейных путей графа, удаления тупиков и пузырей (разветвляющиеся пути, сливающиеся в одной вершине), что позволяет упростить граф для дальнейшей обработки. В связи с ограничением графа перекрытий, для демонстрации работы графа на практике использовались другие данные — длинные риды (long reads), извлеченные из непосредственно референсого генома. Всего таким образом было получено 96 ридов, средняя длина которых составляет 50 тысяч пар нуклеотидов с областью перекрытий 1000-5000 пар нуклеотидов. Для приближения к задаче обработки реальных данных, в риды был добавлен шум — некоторые отдельные нуклеотиды были заменены, вставлены новые и удалены имеющиеся. В результате сборки графом перекрытий было получено 4 крупных рида с показателем N50 равным 1177000, что является хорошим результатом для сборки на длинных ридах.

Для реализации графа де Брюйна использовалась библиотека networkx для работы с графами, которая применялась для реализации графа перекрытий. Методы для удаления тупиков, пузырей и упрощения линейных путей остались такими же почти без изменений, но изменилось построение графа — теперь вершинами выступают не сами риды, а (k-1)-меры каждого рида; ребра графа интерпретируются как k-меры; количество одинаковых k-

меров, полученных из различных ридов складывается в вес ребра. В реализацию был добавлен новый метод для генерации k-мер из ридов. В результате сборки всего было получено 374911 контигов, из них 26951 имею длину большую 160 п.н. — эти очень большие значения свидетельствует о том, что сборка крайне фрагментированна, а сами контиги получились очень короткими — самая большая длина составляет всего 657 п.н.

В качестве компоненты машинного обучения был разработан классификатор, определяющий истинные и ложные пути в графе. Классификация будет проводиться по парам пересекающихся ридов, для которых будет выдан ответ — является ли перекрытие истинным или ложным. Для перекрытия считаются такие статистики, как: длина перекрытия, количество несоответствий (ошибок) в перекрытиях при выравнивании, GC-состав ридов, длины самих ридов, процент длины перекрытия от длины ридов, в качестве метки ставится 0 или 1 в зависимости от близости ридов друг к другу по индексу. В результате, на данном датасете была обучена логистическая регрессия для определения. Теперь, после упрощения графа, на этапе сборки контигов, прежде чем выстраивать очередной эйлеров путь, из каждой вершины, имеющей несколько соседей, будет запускаться классификатор для определения наиболее вероятного перекрытия данного фрагмента. Применение машинного обучения смогло улучшить сборку в плане уменьшения количества ошибок, однако, сборка получилась крайне фрагментированной, а сами контиги имеют короткую длину. Высокая фрагментация объясняется тем, что из графа было удалено большое количество ребер, в результате чего появилось множество компонент связности. Для дальнейшего улучшения результатов сборки можно предложить использование реккурентных нейронных сетей, способных обрабатывать последовательности элементов, чтобы «угадывать» продолжение рида, но это может сказаться на количестве ошибок сборки.

ЗАКЛЮЧЕНИЕ

В данной работе рассмотрены современные инструменты и методы, используемые для сборки геномов, а также описаны их эффективность и применение. Были изучены различные подходы, включая графы де Брюйна и графы перекрытий, а также готовые инструменты, такие как Abyss, SPAdes и Velvet, которые помогают в решении задач сборки и анализа геномных данных.

Среди готовых инструментов, наименьшее количество ошибок при сборке показал инструмент Abyss, но сама сборка имеет высокую фрагментацию. Инструменты SPAdes и Velvet собирают более длинные контиги, но при этом допускают больше крупных ошибок при ассемблировании.

В ходе работы было реализовано несколько подходов к сборке генома, включая графы де Брюйна и перекрытий, а также применены методы машинного обучения. Использование графов де Брюйна и графа перекрытий имеет свои преимущества и недостатки. Графы де Брюйна, с их возможностью эффективного моделирования последовательностей и работы с повторяющимися элементами, стали основой для многих современных ассемблеров. Однако они требуют тщательной настройки параметров, что может усложнить процесс сборки. В то же время графы перекрытий обеспечивают более интуитивное представление данных, однако могут сталкиваться с проблемами при увеличении сложности и объема данных, из-за чего их применение возможно только на небольшом количестве длинных ридов.

Внедрение методов машинного обучения, таких как классификация истинности перекрытий, уменьшает количество ошибок при ассемблировании, что может позволить улучшить сборку, повысить точность выявления геномных вариаций и адаптироваться к различным задачам. Несмотря на заметное снижение ошибок, сборка осталась фрагментированной,

что указывает на необходимость улучшения реализации алгоритмов по обработке графа. Для улучшения результатов предлагается использовать глубокие нейронные сети для предсказания наиболее вероятного продолжения ридов, но это может увеличить количество ошибок. В результате, дальнейшие исследования в этой области могут привести к созданию более совершенных и адаптивных инструментов, способных справляться с задачами, возникающими в геномной биоинформатике.

Основные источники информации:

1. Новикова Е. И. Секвенирование «Нового поколения» (ngs): применение для молекулярно-генетических исследований в онкологии / И. Е. Новикова // Вестник Российского научного центра рентгенорадиологии Минздрава России. — 2016. — Т. 1. — 16 с.
2. Ткачук Е. А. Роль генетики в современной медицине / А. Е. Ткачук // Байкальский медицинский журнал. — 2022. — Т. 1. — С. 81–87.
3. П. Певзнер, Ф. Компо,. Алгоритмы биоинформатики / Ф. Компо, П. Певзнер. — Москва: ДМК-Пресс, 2023. — 682 с.
4. Ж. Сетубал, Ж. Мейданис,. Введение в вычислительную молекулярную биологию / Ж. Сетубал, Ж. Мейданис. — Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. — 420 с.
5. Ф. Крик,. Что за безумное стремленье! / Москва: АСТ, 2020. — 320 с.
6. Сидоров, Е. А. Применение параллельных алгоритмов для построения графа де Брюйна в задаче сборки генома / Е. А. Сидоров // Научно-исследовательский центр «Мир науки». — 2018. — С. 14–18.
7. Казаков, С. В. Автоматизация сборки генома и сравнительного анализа метагеномов для обучения геномной биоинформатике / Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики. — Санкт-Петербург, 2016. — 171 с.