

Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
GRADUATION THESIS

Разработка методов построения и визуализации геномного контекста в метагеномных данных с использованием Hi-C связей на графах де Брейна

Обучающийся / Student Шостина Анастасия Дмитриевна
Факультет/институт/кластер/ Faculty/Institute/Cluster факультет информационных технологий и программирования
Группа/Group M42381c
Направление подготовки/ Subject area 01.04.02 Прикладная математика и информатика
Образовательная программа / Educational program Программирование и искусственный интеллект 2020
Язык реализации ОП / Language of the educational program Русский
Статус ОП / Status of educational program
Квалификация/ Degree level Магистр
Руководитель ВКР/ Thesis supervisor Ульянцев Владимир Игоревич, кандидат технических наук, Университет ИТМО, факультет информационных технологий и программирования, доцент (квалификационная категория "ординарный доцент")

Обучающийся/Student

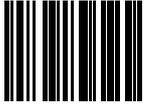
Документ подписан	
Шостина Анастасия Дмитриевна	
26.05.2022	

(эл. подпись/ signature)

Шостина
Анастасия
Дмитриевна

(Фамилия И.О./ name
and surname)

Руководитель ВКР/
Thesis supervisor

Документ подписан	
Ульянцев Владимир Игоревич	
21.05.2022	

(эл. подпись/ signature)

Ульянцев
Владимир
Игоревич

(Фамилия И.О./ name
and surname)

**Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University**

**АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
SUMMARY OF A GRADUATION THESIS**

Обучающийся / Student Шостина Анастасия Дмитриевна
Факультет/институт/кластер/ Faculty/Institute/Cluster факультет информационных технологий и программирования
Группа/Group M42381c
Направление подготовки/ Subject area 01.04.02 Прикладная математика и информатика
Образовательная программа / Educational program Программирование и искусственный интеллект 2020
Язык реализации ОП / Language of the educational program Русский
Статус ОП / Status of educational program
Квалификация/ Degree level Магистр
Тема ВКР/ Thesis topic Разработка методов построения и визуализации геномного контекста в метагеномных данных с использованием Hi-C связей на графах де Брейна
Руководитель ВКР/ Thesis supervisor Ульянов Владимир Игоревич, кандидат технических наук, Университет ИТМО, факультет информационных технологий и программирования, доцент (квалификационная категория "ординарный доцент")

**ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
DESCRIPTION OF THE GRADUATION THESIS**

Цель исследования / Research goal

Совершенствование методов построения и визуализации геномного контекста.

Задачи, решаемые в ВКР / Research tasks

1) Разработка метода построения геномного контекста вокруг анализируемых генов с учетом Hi-C связей при помощи приложения MetaCherchant и утилит картирования ридов BWA-MEM и SAMTools. 2) Разработка и реализация возможности использования данных Hi-C секвенирования и таксономического анализа при визуализации графов де Брейна в приложении Bandage.

Краткая характеристика полученных результатов / Short summary of results/findings

В данной работе был разработан метод построения геномного контекста при помощи приложения MetaCherchant и утилит картирования ридов BWA-MEM и SAMTools. Приложение Bandage было модифицировано для поддержания возможности использования Hi-C связей и результатов таксономического анализа при визуализации графов де Брейна. Помимо этого, была реализована возможность сжатия графа де Брейна. Полученные реализации были протестированы на реальных и сгенерированных данных, что позволило удостовериться в работоспособности нового функционала. Таким образом, все поставленные задачи были выполнены.

Наличие публикаций по теме выпускной работы / Publications on the topic of the thesis

1. Иванов А.Б., Шостина А.Д. Разработка методов построения и визуализации геномного контекста с учетом Hi-C связей в метагеномных данных//Сборник тезисов докладов конгресса молодых ученых. Электронное издание. – СПб: Университет ИТМО - 2022 (Тезисы)

Наличие выступлений на конференциях по теме выпускной работы / Conference reports on the topic of the thesis

1. XI Всероссийский конгресс молодых ученых , 04.04.2022 - 08.04.2022 (Конгресс, статус - всероссийский)
2. 51-ая научная и учебно-методическая конференция Университета ИТМО, 02.02.2022 - 05.02.2022 (Конференция, статус - университетский)

Обучающийся/Student

Документ подписан	
Шостина Анастасия Дмитриевна	
26.05.2022	

(эл. подпись/ signature)

Шостина
Анастасия
Дмитриевна

(Фамилия И.О./ name
and surname)

Руководитель ВКР/
Thesis supervisor

Документ подписан	
Ульянцев Владимир Игоревич	
21.05.2022	

(эл. подпись/ signature)

Ульянцев
Владимир
Игоревич

(Фамилия И.О./ name
and surname)

**Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University**

**ЗАДАНИЕ НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ /
OBJECTIVES FOR A GRADUATION THESIS**

Обучающийся / Student Шостина Анастасия Дмитриевна
Факультет/институт/кластер/ Faculty/Institute/Cluster факультет информационных технологий и программирования
Группа/Group M42381c
Направление подготовки/ Subject area 01.04.02 Прикладная математика и информатика
Образовательная программа / Educational program Программирование и искусственный интеллект 2020
Язык реализации ОП / Language of the educational program Русский
Статус ОП / Status of educational program
Квалификация/ Degree level Магистр
Тема ВКР/ Thesis topic Разработка методов построения и визуализации геномного контекста в метагеномных данных с использованием Hi-C связей на графах де Брейна
Руководитель ВКР/ Thesis supervisor Ульянов Владимир Игоревич, кандидат технических наук, Университет ИТМО, факультет информационных технологий и программирования, доцент (квалификационная категория "ординарный доцент")

Основные вопросы, подлежащие разработке / Key issues to be analyzed

В области здравоохранения есть различные проблемы, связанные с анализом геномного контекста, в том числе проблема устойчивости некоторых видов бактерий к антибиотикам. Когда антибиотико-резистентный ген содержится в плазмиде, то бактерия, в чьих клетках содержится данная плазида также приобретает свойство устойчивости к антибиотикам, при этом хромосомная ДНК бактерии никак не меняется. Поэтому для поиска таких бактерий нужно использовать и данные WGS секвенирования (для определения нуклеотидной последовательности генома) и данные Hi-C секвенирования (для определения пространственной организации генома).

При анализе генома интересен не только список генов, содержащихся в данном геноме, но и их взаимное расположение. Такую информацию предоставляет только граф де Брейна, поэтому при анализе геномного контекста часто используется визуализация графа де Брейна. Для визуализации графа де Брейна используется приложение Bandage [1], которое в отличии от программ общего назначения для визуализации графов (например, Cytoscape), отображает такие особенности графа, как длина нуклеотидной последовательности или глубина покрытия контигов. Однако приложение Bandage не визуализирует Hi-C связи и не использует данные таксономического анализа при визуализации графа.

Целью данной работы является использование Hi-C связей при построении геномного контекста. А также использование Hi-C связей и результатов таксономического анализа при визуализации графа де Брейна.

Для достижения данных целей ставится задача модификации приложений MetaCherchant

[2] и Bandage[1] для реализации возможности использования Hi-C связей при построении геномного контекста, а также использование данных Hi-C секвенирования и таксономического анализа при визуализации графа де Брейна в приложении Bandage. Таким образом необходимо выполнить следующее:

- 1) Получить базовые представления о геномах и их секвенировании.
- 2) Изучить способ построения геномного контекста при помощи приложения Metacherchant, а также способ картирование Hi-C ридов на список контигов при помощи утилит BWA-MEM и SAMTools.
- 3) Реализовать возможность использования Hi-C связей для построения контекста при помощи приложения MetaCherchant и утилит картирование Hi-C ридов.
- 4) Изучить способ визуализации графов де Брейна, используемый приложением «Bandage».
- 5) Реализовать в приложении «Bandage» возможность визуализации Hi-C связей, возможность использования таксонов при визуализации графов, а также возможность сжатие графа для упрощения его структуры.
- 6) Провести тестирование нового функционала на реальных и сгенерированных метагеномных данных.

Список литературы:

1. Bandage: Ryan R. W., Mark B. S., Justin Z., Kathryn E. H. Bandage: interactive visualization of de novo genome assemblies. Notes Bioinformatics — 2015 — Vol. 31, Pp. 3350–3352
2. MetaCherchant: Evgenii I Olekhovich, Artem T Vasilyev, Vladimir I Ulyantsev, Elena S Kostryukova, Alexander V Tyakht. MetaCherchant: analyzing genomic context of antibiotic resistance genes in gut microbiota. Bioinformatics — 2018 — Vol 34, Issue 3 — Pp. 434–444

Форма представления материалов ВКР / Format(s) of thesis materials:

программный код, презентация

Дата выдачи задания / Assignment issued on: 01.09.2021

Срок представления готовой ВКР / Deadline for final edition of the thesis 27.05.2022

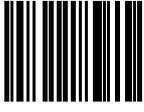
Характеристика темы ВКР / Description of thesis subject (topic)

Тема в области фундаментальных исследований / Subject of fundamental research: нет / not

Тема в области прикладных исследований / Subject of applied research: да / yes

СОГЛАСОВАНО / AGREED:

Руководитель ВКР/
Thesis supervisor

Документ подписан	
Ульянцев Владимир Игоревич	

Ульянцев
Владимир

20.05.2022

(эл. подпись)

Игорович

Задание принял к
исполнению/ Objectives
assumed BY

Документ подписан	
Шостина Анастасия Дмитриевна	
20.05.2022	

(эл. подпись)

Шостина
Анастасия
Дмитриевна

Руководитель ОП/ Head
of educational program

Документ подписан	
Парфенов Владимир Глебович	
23.05.2022	

(эл. подпись)

Парфенов
Владимир
Глебович

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	6
ГЛАВА 1. ОБЗОР МЕТОДОВ АНАЛИЗА ГЕНОМНОГО КОНТЕКСТА	9
1.1. МЕТОДЫ СЕКВЕНИРОВАНИЯ	9
1.2. ПОСТРОЕНИЕ ГЕНОМНОГО КОНТЕКСТА.....	12
1.3. ВИЗУАЛИЗАЦИЯ ГЕНОМНОГО КОНТЕКСТА	14
1.4. БИОЛОГИЧЕСКАЯ ПРОБЛЕМА.....	16
1.5. ПОСТАНОВКА ЗАДАЧИ.....	17
Выводы по главе 1.....	19
ГЛАВА 2. ПОСТРОЕНИЕ ГЕНОМНОГО КОНТЕКСТА С УЧЕТОМ HI-C СВЯЗЕЙ.....	20
2.1. ЭТАП 1: ПОСТРОЕНИЕ ИСХОДНОГО ГЕНОМНОГО КОНТЕКСТА	20
2.2. ЭТАП 2: ПОИСК HI-C РИДОВ	21
2.3. ЭТАП 3: ПОСТРОЕНИЕ РАСШИРЕННОГО ГЕНОМНОГО КОНТЕКСТА.....	23
2.4. ЭТАП 4: ВИЗУАЛИЗАЦИЯ РАСШИРЕННОГО ГЕНОМНОГО КОНТЕКСТА	24
2.5. РЕАЛИЗАЦИЯ	26
Выводы по главе 2.....	27
ГЛАВА 3. ВИЗУАЛИЗАЦИЯ ГРАФА ДЕ БРЕЙНА С HI-C СВЯЗЯМИ.....	28
3.1. ИСПОЛЬЗОВАНИЕ HI-C РЕБЕР ПРИ УКЛАДКЕ ГРАФА ДЕ БРЕЙНА.	29
3.2. ВЫБОР HI-C РЕБЕР ДЛЯ ОТОБРАЖЕНИЯ НА ГРАФЕ ДЕ БРЕЙНА	32
3.3. ВИЗУАЛИЗАЦИЯ РЕАЛЬНЫХ МЕТАГЕНОМНЫХ ДАННЫХ.....	35
3.4. АВТОМАТИЧЕСКИЙ ПОДБОР ПАРАМЕТРОВ ОТРИСОВКИ	38
3.5. ТАКСОНОМИЧЕСКИЙ АНАЛИЗ	41
3.6. СЖАТИЕ ГРАФА	48
Выводы по главе 3.....	50
ГЛАВА 4. ТЕСТИРОВАНИЕ НОВОГО СПОСОБА ПОСТРОЕНИЯ И ВИЗУАЛИЗАЦИИ ГЕНОМНОГО КОНТЕКСТА	51

4.1. БАКТЕРИЯ САЛЬМОНЕЛЛА И ЕЕ ПЛАЗМИДА	51
4.2. КИШЕЧНАЯ ПАЛОЧКА, БАКТЕРОИД ФРАГИЛИС И ИХ ПЛАЗМИДЫ	53
4.3. ОБРАЗЕЦ МИКРОБИОТЫ КИШЕЧНИКА ПРИ АНТИБАКТЕРИАЛЬНОЙ ТЕРАПИИ	56
Выводы по главе 4.....	60
ЗАКЛЮЧЕНИЕ	61
БИБЛИОГРАФИЧЕСКИЙ СПИСОК.....	62

ВВЕДЕНИЕ

В биоинформатике существуют различные методы сбора метагеномных данных. В данной работе рассматривается метод полногеномного секвенирования (WGS) и метод определения конформации хромосом Hi-C. WGS используется для определения нуклеотидной последовательности генома. Анализ WGS данных основан на построении графа де Брейна. Именно граф де Брейна дает наиболее полную информацию о рассматриваемом геноме, так как он предоставляет не просто список генов, содержащихся в данном геноме, но и их взаиморасположение. Именно поэтому при анализе генома часто рассматривается визуализация его графа де Брейна. Для визуализации графов используется приложение Bandage, которое в отличие от программ общего назначения для визуализации графов, отображает такие особенности графа де Брейна, как длина нуклеотидной последовательности или глубина покрытия контигов.

При анализе конкретных генов, как правило интересен не весь граф де Брейна, а только его часть, которая содержит информацию о генах, расположенных рядом с исследуемым геном. Поэтому анализ графа де Брейна можно заменить на анализ геномного контекста, то есть подграфа графа де Брейна, построенного вокруг анализируемого гена. Геномный контекст значительно меньше полного графа де Брейна, поэтому анализ геномного контекста предпочтительнее анализа всего гена целиком. Построить геномный контекст можно при помощи приложения MetaCherchant.

В отличие от метода полногеномного секвенирования Hi-C секвенирование используется не для определения линейной структуры генома, а для определения пространственной организации генома. Результатом Hi-C секвенирования является список пар Hi-C ридов, которые в пространстве расположены близко друг к другу. Если в паре Hi-C ридов риды принадлежат разным контигам, то между данными контигами есть Hi-C связь.

В области здравоохранения есть различные проблемы, связанные с анализом геномного контекста, в том числе проблема устойчивости некоторых видов бактерий к антибиотикам. В плаزمидях и других транспортных элементах часто бывают расположены антибиотико-резистентные гены. Когда плазида расположена в клетке бактерии, данная бактерия приобретает свойство устойчивости к антибиотикам, при этом ее хромосомная ДНК никак не меняется. Однако в этом случае при помощи метода Hi-C секвенирования можно найти такие пары Hi-C ридов, в которых один рид лежит внутри генома плазмиды, а другой рид внутри хромосомной ДНК бактерии. Поэтому возможным решением задачи поиска бактерий, которые приобрели свойство устойчивости к антибиотикам, является использование не только данных полногеномного секвенирования, но и данных Hi-C секвенирования при построении геномного контекста вокруг анализируемого гена. Таким образом, построение геномного контекста с учетом Hi-C связей позволит получить информацию о взаимосвязях мобильных элементов и их носителей.

Как было отмечено ранее, именно граф де Брейна предоставляет наиболее полную информацию о геноме, поэтому для анализа геномного контекста также используется его визуализация. Визуализация графа влияет на восприятие графа пользователем, поэтому крайне важно разрабатывать и модифицировать приложения по визуализации графов де Брейна. Например, приложение Bandage визуализирует только один тип ребер — WGS ребра и не поддерживает визуализацию Hi-C связей. Это приводит к тому, что при рассмотрении графа де Брейна не используется информация о его пространственном расположении. Кроме того, данное приложение также не поддерживает визуализацию данных таксономического анализа, однако понимание видовой принадлежности контигов крайне важно при анализе графов де Брейна, содержащих геномы разных бактерий, плазмид и вирусов.

Таким образом, в данной работе ставится цель совершенствования методов построения и визуализации геномного контекста. Для достижения данной цели выделяются две основные задачи, а именно: задача разработки метода построения геномного контекста с учетом Hi-C связей при помощи приложения MetaCherchant и утилит картирования ридов BWA-MEM и SAMTools и задача разработки способа отображения Hi-C связей и результатов таксономического анализа при визуализации графов де Брейна в приложении Bandage.

ГЛАВА 1. ОБЗОР МЕТОДОВ АНАЛИЗА ГЕНОМНОГО КОНТЕКСТА

1.1. МЕТОДЫ СЕКВЕНИРОВАНИЯ

Геном — совокупность наследственной информации организма, закодированной в молекулах дезоксирибонуклеиновой кислоты (ДНК). **Ген** — участок ДНК. **Метагеном** — набор генов всех микроорганизмов, находящихся в образце среды. **Хромосомой** будем считать отдельную молекулу ДНК. Молекула ДНК представляет собой полимер, в котором чередуются остатки сахара дезоксирибозы и фосфата. К остаткам сахара присоединено одно из четырех азотистых оснований: аденин (А), цитозин (Ц), гуанин (Г) и тимин (Т). Эти структурные единицы — мономеры — называются **нуклеотидами**. Нуклеотиды соединены между собой в нуклеотидные цепочки. Таким образом, последовательность мономеров цепи ДНК описывается строкой в алфавите из четырех символов {А, С, G, Т}.

Плазмиды — небольшие молекулы ДНК, физически обособленные от хромосом. Главным образом плазмиды встречаются у бактерий. В природе плазмиды обычно содержат гены, повышающие приспособленность бактерий к окружающей среде. Например, гены, которые обеспечивают устойчивость к антибиотикам. Нередко они могут передаваться от одной бактерии к другой того же вида, рода или семейства, являясь таким образом средством горизонтального переноса генов. Схематичное изображение клетки с плазмидой приведено на рисунке 1.1.

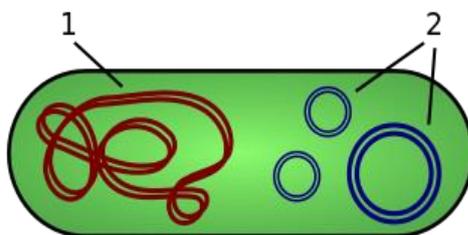


Рисунок 1.1. Хромосомная ДНК (1) и плазмиды (2) в бактериальной клетке.

Секвенированием называется процесс реконструкции последовательности нуклеотидов в молекулах ДНК. При этом термин «**секвенирование генома**» может использоваться в нескольких контекстах. С одной стороны, секвенирование генома — это непосредственно процесс «чтения» последовательности нуклеотидов сравнительно коротких фрагментов цепи ДНК с использованием одного из биотехнологических методов. Получаемые последовательности называют **ридами**. С другой стороны, говоря о **секвенировании организма**, имеют в виду восстановление нуклеотидной последовательности его хромосом.

Базовой методологией реконструкции генома ранее не изученного вида является **полногеномное секвенирование (WGS)**. При полногеномном секвенировании риды считываются со случайных позиций генома. При этом риды, прочитанные с близких позиций, перекрываются, и одна и та же позиция генома оказывается «покрыта» несколькими ридами. Вычислительный этап обработки набора ридов с целью восстановления продолжительных фрагментов генома называют **сборкой генома**. Все инструменты для осуществления сборки — **ассемблеры** — анализируют перекрытия между ридами с целью их объединения в более продолжительные нуклеотидные последовательности. Выделяют три класса ассемблеров [1; 2]):

- 1) Использующие процедуры “жадного” объединения ридов.
- 2) Использующие графы перекрытий.
- 3) Использующие графы де Брейна.

В данной работе будет рассматриваться стратегия сборки, основанная на графах де Брейна [3], используемая для сборки коротких ридов [4].

Пусть k — целое положительное число. Последовательности, состоящие из k нуклеотидов, будем называть k -мерами. **Граф де Брейна** — это ориентированный граф с петлями и кратными ребрами, вершины которого

соответствуют различным k -мерам, а рёбра — различными $k+1$ -мером. При этом ребро, отмеченное $k+1$ -мером, соединяет два k -мера, которые являются его префиксом и суффиксом длины k .

Идеальный граф де Брейна должен содержать один отдельный путь для каждой базовой последовательности, но такие особенности генома, как повторяющиеся участки, обычно препятствуют этому. Самые длинные последовательности в графе, которые можно определить однозначно, сохраняются как **контиги**. **Сжатый граф де Брейна** — это граф де Брейна, вершинами которого являются контиги. Для построения сжатого графа де Брейна нужно в обычном графе де Брейна объединить вершины, имеющие ровно одно входящее и одно исходящее ребро, со своими соседями. На рисунке 1.2 приведен пример графа де Брейна и сжатого графа де Брейна для нуклеотидной последовательности: *ACGTCCCGTCAA*. В данном графе используется $k = 3$.

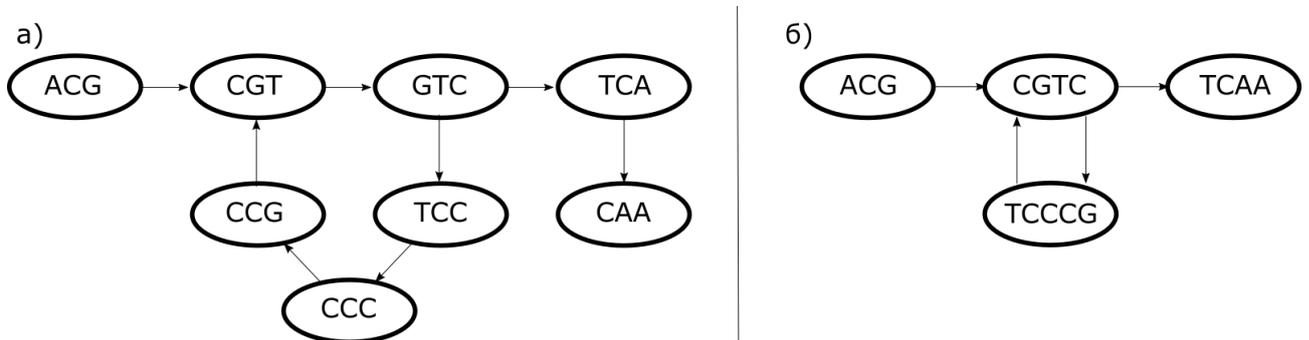


Рисунок 1.2. а) Граф де Брейна и б) сжатый граф де Брейна для последовательности *ACGTCCCGTCAA*.

При анализе генома уделяется внимание не только восстановлению его нуклеотидной последовательности, но и его пространственному расположению. Например, **Hi-C секвенирование** используется для определения пространственной организации генома [5]. Результатом работы метода Hi-C секвенирования является набор пар Hi-C ридов. Парами Hi-C ридов являются

участки ДНК, которые в пространстве расположены близко друг к другу. Это могут быть риды из одного генома или риды из двух различных геномов. Однако пары Hi-C ридов всегда лежат внутри одной клетки. Следует отметить, что большинство пар Hi-C ридов соединяют участки ДНК одного генома, однако если в клетке бактерии присутствует плазида, то в результате Hi-C секвенирования будут найдены такие пары Hi-C ридов, в которых один рид лежит внутри нуклеотидной последовательности плазмиды, а другой — внутри хромосомной ДНК бактерии.

Для объединения результатов полногеномного и Hi-C секвенирования пары Hi-C ридов картируются на контиги графа де Брейна, то есть каждому Hi-C риду будет соответствовать некоторый контиг, в котором данный рид расположен. Картировать Hi-C риды можно при помощи утилит BWA-MEM и SAMTools. Если в паре Hi-C ридов риды принадлежат разным контигам графа де Брейна, то между данными контигами есть Hi-C связь. Вес Hi-C связи равен числу таких пар Hi-C ридов.

1.2. ПОСТРОЕНИЕ ГЕНОМНОГО КОНТЕКСТА

Геномный контекст — это подграф графа де Брейна, построенный вокруг исследуемого гена. Граф де Брейна, построенный на метагеномных данных, имеет большой размер и сложную структуру, поэтому для анализа его части удобно использовать геномный контекст. Геномный контекст можно построить при помощи приложения MetaCherchant [6]. Данное приложение позволяет построить подграф графа де Брейна вокруг одного или нескольких генов. Подграф ограничивается максимальным радиусом, то есть максимальной допустимой длиной пути между любым k-мером и исследуемым геном, или максимальным количеством k-меров в подграфе. Приложение MetaCherchant строит геномный контекст следующим образом:

- 1) Декомпозиция WGS ридов и нуклеотидной последовательности исследуемого гена на k -меры. K -меры хранятся в хэш таблице вместе с их покрытием. **Покрытие** — это то, сколько раз k -мер встретился в прочтениях. Все k -меры, имеющие покрытие ниже фиксированного порога, считаются ошибочными. Из-за ошибок чтения в WGS секвенировании мы не можем доверять всем k -мерам.
- 2) Построение геномного контекста при помощи обхода графа в ширину.
 - a) На первой итерации алгоритма очередь инициализируется мерами из одного или нескольких исследуемых генов.
 - b) Обработка вершины из очереди происходит следующим образом: для текущего k -мера определяются его соседи. Все соседи являются суффиксом или префиксом друг друга. Длина суффикса (префикса) равна $k - 1$. Нуклеотидные последовательности состоят из четырех нуклеотидов, значит всего возможно четыре соседа справа и четыре соседа слева, которые получаются при помощи дописывания одного нуклеотида к суффиксу или префиксу справа или слева соответственно. В таблице 1.1 приведен пример поиска соседних k -меров. Затем в хэш таблице проверяется, какие из возможных соседних k -меров присутствуют в данном графе.
 - c) Если соседняя вершина не была посещена ранее, то она записывается в конец очереди.
 - d) Пункты 2.b и 2.c повторяются пока очередь не закончится или пока не будет достигнуто одно из следующих условий остановки: достижение максимальной дистанции до исследуемого гена или достижение максимального числа k -меров.
- 3) Построение сжатого графа де Брейна для упрощения его визуализации. Каждый путь без ответвлений в подграфе графа де Брейна сжимается в одну вершину.

- 4) Сохранение подграфа в файл в формате GFA (Graphical Fragment Assembly).
Граф, записанный в файл формата GFA, может быть визуализирован приложением Bandage [7].

Следует отметить, что приложение MetaCherchant при построении геномного контекста использует только пары WGS ридов и не учитывает Hi-C связи.

Таблица 1.1. Пример нахождения соседнего k-мера в графе де Брейна.

К-мер	Суффикс	Соседи справа	Префикс	Соседи слева
ACG	CG	CGA CGC CGG CGT	AC	AAC CAC GAC TAC

1.3. ВИЗУАЛИЗАЦИЯ ГЕНОМНОГО КОНТЕКСТА

При анализе геномного контекста используется его визуализация. Геномный контекст — это граф, а граф можно определить как набор вершин и набор ребер таких, что ребро показывает существование отношения между двумя вершинами. Визуализация графа может помочь понять структуру этих отношений. Однако, недостаточно просто нарисовать граф, визуализация графа оказывает значительное влияние на то, как граф воспринимается. Вершины, расположенные близко друг к другу, будут интерпретироваться пользователем как состоящие в более сильном отношении. То есть, чем меньше длина ребра между двумя вершинами, тем сильнее связь между этими вершинами (с точки зрения пользователя). Это означает, что расположение вершин в пространстве сильно влияет на восприятие данного графа.

Существует множество различных методов укладки графа, которые подразделяются на силовые алгоритмы (force-directed) [8] и многоуровневые алгоритмы (MultilevelLayout) [9, 10, 11].

Силовые алгоритмы визуализации графов хорошо подходят для визуализации графов средних размеров (50–500 вершин). Их цель — расположить узлы графа в двумерном или трехмерном пространстве так, чтобы все ребра имели одинаковую длину, и свести к минимуму число пересечений ребер. Помимо этого, этот алгоритм часто бывает неинформативен для больших сильно-загруженных или кластерных графов. Таким образом, силовые алгоритмы плохо подходят для отрисовки графов де Брейна (так как они, как правило, имеют большой размер).

Многоуровневый подход визуализации графов хорошо работает для больших графов с модульной структурой. Многоуровневые алгоритмы основаны на силовых методах, однако делают их более эффективными для больших графов. Многоуровневый алгоритм укладки графа — это итеративный алгоритм, на каждой итерации которого выделяется подграф текущего графа и делается укладка данного подграфа. Укладка подграфа осуществляется или при помощи следующей итерации алгоритма или при помощи силового алгоритма. После чего делается укладка текущего графа при помощи силового алгоритма, однако координаты части вершин текущего графа уже известны и не меняются.

В качестве средства визуализации графа де Брейна в данной работе используется приложение Bandage. Приложение Bandage позволяет приближать выделенные области графа, перемещать вершины, добавлять подписи к вершинам, менять их цвет и многое другое. Однако приложение Bandage при текущей реализации не позволяет отображать Hi-C связи и учитывать их при укладке графа. Также приложение Bandage не позволяет использовать данные таксономического анализа при визуализации графа де Брейна.

В приложении Bandage вершины отображаются как цветные «ленты», длина которых зависит от длины нуклеотидной последовательности соответствующих им контигов. Концы «лент» соединены между собой ребрами, которые изображены как черные прямые линии фиксированной длины. Пример визуализации графа де Брейна приведена на рисунке 1.4.

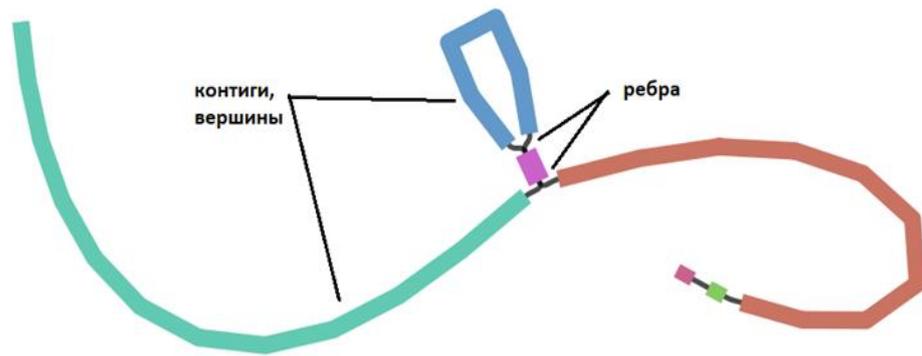


Рисунок 1.4. Визуализация графа де Брейна в приложении Bandage.

Укладка и отрисовка графов в приложении Bandage выполняется при помощи фреймворка “The Open Graph Drawing Framework” (OGDF) [12]. OGDF фреймворк является автономной библиотекой на языке C++ для отрисовки графов, включающей в себя алгоритмы укладки графов. В приложении Bandage для укладки графов используется способ многоуровневой укладки, а именно алгоритм «fast multipole multilevel» (FMMM) [11].

1.4. БИОЛОГИЧЕСКАЯ ПРОБЛЕМА

В области здравоохранения существуют проблемы, связанные с анализом геномного контекста, например, проблема устойчивости бактерий к некоторым

видам антибиотиков. Антибиотико-резистентные гены (АРГ) могут располагаться в плазмидах. Если плазида с АРГ находится в клетке бактерии, то данная бактерия приобретает свойство устойчивости к антибиотику. При этом хромосомная ДНК бактерии никак не меняется, поэтому при помощи метода WGS нельзя выделить бактерии, клетки которых содержат плазмиды с исследуемым АРГ. Однако в этом случае между хромосомной ДНК бактерии и хромосомной ДНК плазмиды есть Hi-C связи, поэтому такие бактерии можно найти при помощи Hi-C секвенирования. Для этого необходимо проанализировать Hi-C связи в подграфе графа де Брейна, построенном вокруг исследуемого гена.

На данный момент существуют методы поиска антибиотико-резистентных генов, находящихся в метагеноме, однако нет возможности определить в клетках каких бактерий расположены плазмиды с найденными антибиотико-резистентными генами. Возможным решением данной проблемы является анализ графа де Брейна с учетом Hi-C связей, что позволит определить хромосомные ДНК бактерий и плазмид, которые могут располагаться в одной клетке. Следует отметить, что одна и та же плазида может располагаться в клетках различных бактерий. Таким образом, с точки зрения биологии необходимо решить задачу получения информации о взаимосвязях мобильных элементов (в том числе плазмид) и их носителей.

1.5. ПОСТАНОВКА ЗАДАЧИ

Поставленную биологическую задачу можно трансформировать в техническую **цель совершенствования методов построения и визуализации геномного контекста**. Для достижения цели совершенствования методов построения геномного контекста предлагается разработать возможность

использования данных Hi-C секвенирования при построении геномного контекста, что позволит учитывать взаимосвязи между мобильными элементами и их носителями при построении геномного контекста.

Для визуализации геномного контекста используется приложение Bandage. Однако данное приложение не поддерживает визуализацию Hi-C связей. В качестве альтернативы для визуализации графа де Брейна сразу с WGS и Hi-C ребрами можно использовать программы общего назначения для визуализации графов (например, Cytoscape). Приложение Cytoscape позволяет на одном графе изобразить два типа ребер, однако в этом случае будет утрачена вся биологическая информация про контиги, а именно: длина нуклеотидной последовательности и покрытие контига. Кроме того, при анализе геномного контекста ключевую роль играет видовая принадлежность контигов, которую можно определить при помощи таксономического анализа. Однако результаты таксономического анализа не используются при визуализации графа де Брейна в приложении Bandage, поэтому пользователям приходится вручную при помощи таксономического отчета определять таксоны, которым соответствуют те или иные контиги, и отмечать эти данные на графе (в виде специального цвета или подписей к контигам). Таким образом, для улучшения способа визуализации геномного контекста ставится задача отображения Hi-C связей и результатов таксономического анализа при визуализации графа де Брейна в приложении Bandage.

Подводя итог, можно сформулировать две основные задачи, а именно:

- 1) Задача разработки метода построения геномного контекста с учетом Hi-C связей при помощи приложения MetaCherchant и утилит картирования ридов BWA-MEM и SAMTools.
- 2) Задача разработки способа отображения Hi-C связей и результатов таксономического анализа при визуализации графа де Брейна в приложении Bandage.

ВЫВОДЫ ПО ГЛАВЕ 1

В этой главе введены основные понятия генетики, описаны методы WGS и Hi-C секвенирования. Также приведен метод построения геномного контекста, используемый приложением MetaCherchant. Рассмотрен способ визуализации графа де Брейна в приложении Bandage. Отмечены недостатки рассмотренного способа построения и визуализации геномного контекста, которые приводят к необходимости поиска новых решений, а именно: разработки метода построения геномного контекста с учетом Hi-C связей и расширения функционала приложения Bandage.

ГЛАВА 2. ПОСТРОЕНИЕ ГЕНОМНОГО КОНТЕКСТА С УЧЕТОМ NI-C СВЯЗЕЙ

В данной главе описан разработанный алгоритм построения геномного контекста с учетом NI-C связей. Для использование пар NI-C ридов при построение геномного контекст предлагается выполнить следующие этапы:

2.1. ЭТАП 1: ПОСТРОЕНИЕ ИСХОДНОГО ГЕНОМНОГО КОНТЕКСТА

На первом этапе происходит построение геномного контекста вокруг исследуемого гена. На рисунке 2.1 схематично изображен построенный исходный геномный контекст. Контиги на рисунке представлены в виде прямоугольников. Анализируемый ген обозначен красным цветом. Контиги, вошедшие в геномный контекст, имеют синий цвет, а контиги, не вошедшие в геномный контекст, — серый. Построение геномного контекста будем производить при помощи приложения MetaCherchant. Данное приложение выделит все k-меры из WGS ридов, переданных во входном файле, а затем при помощи обхода графа в глубину построит подграф графа де Брейна вокруг переданного гена. Нуклеотидная последовательность исследуемого гена также передается приложению в файле формата *fasta* или *fastq*. Результатом работы приложения MetaCherchant является граф де Брейна, записанный в файл *graph.gfa* и список контигов с их нуклеотидными последовательностями, сохраненными в файле *seqs.fasta*.

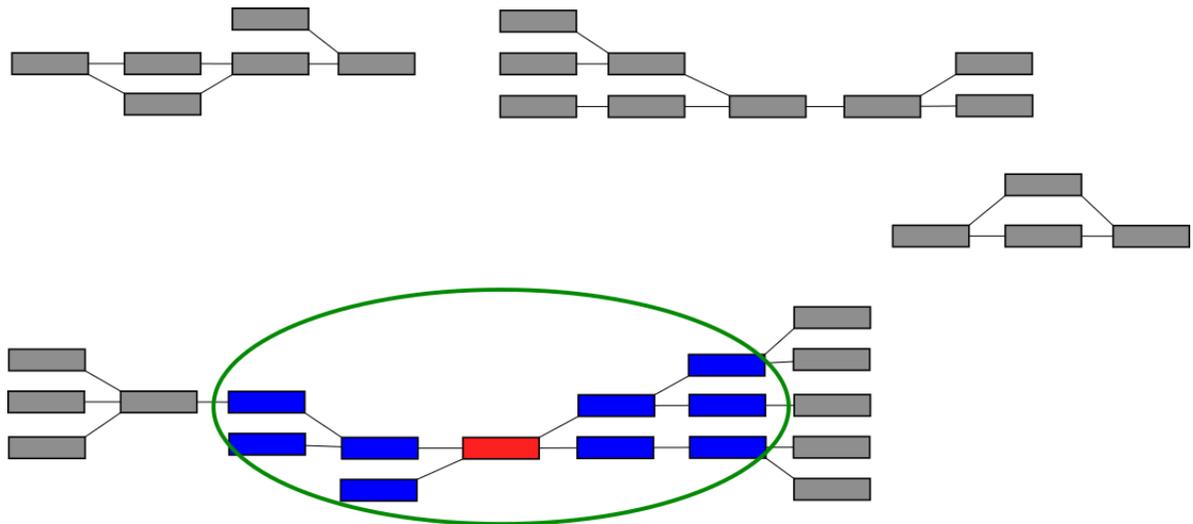


Рисунок 2.1. Схематическое изображение исходного геномного контекста.

2.2. ЭТАП 2: ПОИСК Hi-C РИДОВ

На втором этапе происходит поиск Hi-C ридов, которые лежат вне построенного геномного контекста, однако имеют Hi-C связи с ним. Именно такие Hi-C риды расширяют исходный геномный контекст. На рисунке 2.2 приведен пример Hi-C связей, используемых для расширения геномного контекста. Интересующие нас Hi-C связи изображены красной пунктирной линией, а остальные Hi-C связи отмечены черной пунктирной линией.

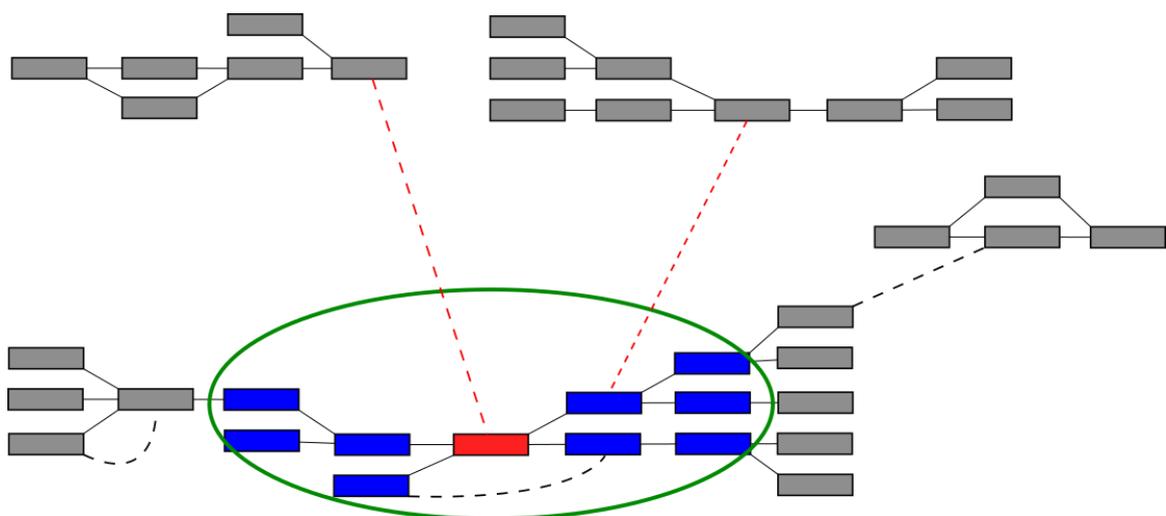


Рисунок 2.2. Схематическое изображение выделения Hi-C связей, расширяющих геномный контекст.

Выделение интересующих Hi-C ридов происходит при помощи утилит BWA [13] и SAMTools [14]. Утилита BWA используется для картирования пар Hi-C ридов на геномный контекст, полученный на первом этапе. Для этого используются файлы в формате *fasta* с парами Hi-C ридов и файл *seqs.fasta*, полученный на первом этапе. В результате картирования каждому Hi-C риду будет сопоставлен контиг, в котором он находится. Если такого контига нет, то рид будет отмечен как неизвестный. Результат картирования будет записан в файл *all_hic_reads.sam*. Затем при помощи утилиты SAMTools пары Hi-C ридов из файла *all_hic_reads.sam* будут отфильтрованы таким образом, чтобы остались только те Hi-C риды, которые не принадлежат построенному контексту (то есть отмечены как неизвестные), а их парные риды лежат в исходном контексте (то есть для них определен контиг, которому они принадлежат). Все найденные Hi-C риды будут записаны в файл *selected_reads.fasta*. В листинге 2.1 приведен фрагмент кода *bash* скрипта, который реализует описанный выше функционал. В этом коде используются следующие переменные: *hi_c_r1*, *hi_c_r2* и *seqs*. Переменные *hi_c_r1*, *hi_c_r2* содержат путь до двух файлов, содержащих пары Hi-C ридов. Ожидается, что *i*-ый рид в файле *hi_c_r1* и *i*-ый рид в файле *hi_c_r2* являются парными друг для друга. Переменная *seqs* содержит путь до файла *seqs.fasta*, который был создан на первом этапе, в результате вызова приложения *MetaCherchant*.

Листинг 2.1 – Поиск Hi-C ридов, расширяющих исходный геномный контекст.

```

1 bwa index $seqs
2 bwa mem $seqs $hi_c_r1 $hi_c_r2 > all_hic_reads.sam
3 samtools view -f 0x5 -F 0x908 -o filteredHiC_1.bam all_hic_reads.sam"
4 samtools view filteredHiC_1.bam | awk ' {print ">"1"\n"$10} ' -> selected_reads.fasta

```

2.3. ЭТАП 3: ПОСТРОЕНИЕ РАСШИРЕННОГО ГЕНОМНОГО КОНТЕКСТА

На данном этапе происходит построение геномного контекста с учетом Hi-C связей. Для этого строится объединенный геномный контекст вокруг исходного гена, а также всех выделенных на втором этапе Hi-C ридов. На рисунке 2.3 схематично изображен расширенный геномный контекст. Этот этап также происходит при помощи приложения Metacherchant, которое позволяет выделять объединенный геномный контекст вокруг нескольких нуклеотидных последовательностей. В приложении MetaCherchant все нуклеотидные последовательности, переданные в параметре *seq*, считаются целевыми генами и выделяются в отдельные контиги в графе де Брейна. Однако Hi-C риды не являются целевыми генами, поэтому не нужно выделять их в отдельные вершины в графе де Брейна, это только усложнит граф. В связи с этим, мной в приложении Bandage был добавлен новый параметр *hicseq*. В качестве данного параметра передается путь до файла в формате fasta, содержащего список Hi-C ридов. Приложение MetaCherchant будет строить контекст вокруг нуклеотидных последовательностей, переданных в параметрах *seq* и *hicseq*, но в отдельные контиги будут выделяться только нуклеотидные последовательности, переданные в параметре *seq*. Результатом работы приложения Metacharchant на этом этапе является геномный контекст, построенный с учетом Hi-C связей, записанный в файл *graph.gfa*.

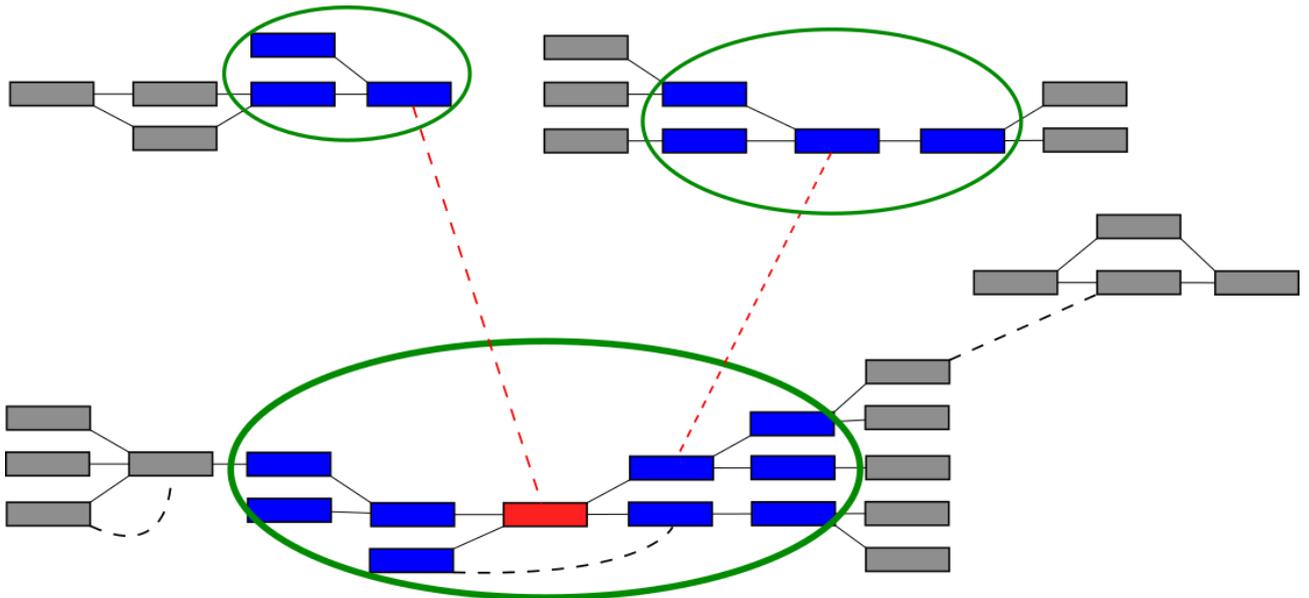


Рисунок 2.3. Схематическое изображение расширенного геномного контекста.

2.4. ЭТАП 4: ВИЗУАЛИЗАЦИЯ РАСШИРЕННОГО ГЕНОМНОГО КОНТЕКСТА

Визуализация полученного геномного контекста производится при помощи приложения Bandage. Данное приложение было модифицировано для поддержания визуализации Hi-C связей в графе де Брейна. Подробнее о способе визуализации Hi-C связей и новом функционале приложения Bandage рассказано в главе 3. На третьем этапе в результате работы приложения MetaCherchant был получен файл *graph.gfa*, который подходит для использования приложением Bandage. Однако этот файл не содержит информации о Hi-C связях. Для отображения Hi-C связей в Bandage необходимо подготовить файл в формате txt, состоящий из трех колонок: идентификаторы двух контигов, имеющих Hi-C связь друг с другом, а также вес данной Hi-C связи. Для получения этих данных необходимо при помощи утилиты BWA картировать все пары Hi-C ридов на контиги расширенного контекста. Затем отфильтровать пары Hi-C ридов, оставив только те пары, в которых оба риды

были сопоставлены с различными контигами. Это можно осуществить при помощи утилиты SAMTools. Результатом фильтрации является файл *filtered_HiC_diff_chr.sam*, содержащий пары контигов, с которыми были сопоставлены исходные пары Hi-C ридов. В листинге 2.2 приведен фрагмент кода bash скрипта, который реализует картирование Hi-C ридов на разные контиги геномного контекста, построенного на третьем этапе. Далее нужно преобразовать полученный файл в формат, принимаемый приложением Bandage. Для этого был написан скрипт на языке Python. Псевдокод алгоритма форматирования файла *filtered_HiC_diff_chr.sam* в формат, принимаемый приложением Bandage приведен в листинге 2.3.

Листинг 2.2 – Картирование Hi-C ридов на контиги из расширенного геномного контекста.

```

1 bwa index seqs.fasta
2 bwa mem seqs.fasta $hi_c_r1 $hi_c_r2 > filteredHiC_2.sam
3 samtools view -f 1 -F 2060 -b -o filteredHiC_pair_map.bam filteredHiC_2.sam
4 samtools view filteredHiC_pair_map.bam | awk '($3!=$7 && $7!="")' - >
  filtered_HiC_diff_chr.sam
5 python hic_map.py filtered_HiC_diff_chr.sam

```

Для создания списка Hi-C связей в соответствие с форматом, который поддерживает приложение Bandage, необходимо из файла *filtered_HiC_diff_chr.sam* загрузить различные пары контигов (это столбцы 3 и 7 в файле *filtered_HiC_diff_chr.sam*) и посчитать число пар Hi-C ридов, соответствующие каждой паре различных контигов. Пара контигов (A, B) и (B, A) — это одна и та же пара, поэтому контиги каждой пары сортируются по возрастанию перед добавлением в словарь, то есть словарь содержит пары контигов, в которых первый элемент пары всегда меньше второго. Каждая пара Hi-C ридов содержит комплементарную ей пару ридов, поэтому в файле

filtered_HiC_diff_chr.sam каждая пара Hi-C ридов учтена дважды, из-за этого число пар ридов для каждой пары различных контигов делится на два.

Листинг 2.3 – Форматирование файла *filtered_HiC_diff_chr.sam*.

```
data = dict()
for line in filtered_HiC_diff_chr.sam:
    l = line.strip().split()
    pair = tuple(sorted((l[2], l[6])))
    data[pair] = data.get(pair, 0) + 1
for k, v in data.items():
    print(k[0], k[1], v // 2, sep="\t")
```

2.5. РЕАЛИЗАЦИЯ

Предложенный способ построения геномного контекста был реализован в виде `bash` скрипта. Данный скрипт принимает следующие параметры:

- 1) `reads` — список файлов со всеми парами WGS ридов. Поддерживаются файлы в формате `fasta` и `fastq`. Это обязательный параметр.
- 2) `seq` — `fasta` или `fastq` файл, содержащий нуклеотидную последовательность, вокруг которой будет построен геномный контекст. Это обязательный параметр.
- 3) `hi-c-r1` и `hi-c-r2` — два входных файла в формате `fasta` или `fastq`, содержащие пары Hi-C ридов. Ожидается, что *i*-ый рид в файле `hi-c-r1` и *i*-ый рид в файле `hi-c-r2` являются парными друг для друга (то есть имеют одну Hi-C связь между собой). Это обязательный параметр.
- 4) `work-dir` — путь до рабочей директории, в которой будут храниться промежуточные результаты, а также результаты работы всех этапов предложенного алгоритма. Это обязательный параметр.
- 5) `metacherchant` — `jar` файл приложения `MetaCherchant`.

- 6) k — размер k -мера, используемого при построении графа де Брейна. Это опциональный параметр. Значение по умолчанию равно 31.
- 7) coverage — минимальный порог покрытия k -меров, используемых при построении графа де Брейна. Значение по умолчанию равно 50.
- 8) maxradius — максимально допустимая дистанция между любым k -мером и исследуемым геном. Значение по умолчанию равно 3000.

Bash скрипт состоит из 4 шагов, описанных ранее:

- 1) Построение исходного геномного контекста.
- 2) Поиск Hi-C ридов, расширяющих исходный геномный контекст.
- 3) Построение расширенного геномного контекста.
- 4) Картирование Hi-C ридов на расширенный геномный контекст.

Для успешного выполнения скрипта нужны следующие prerequisites: Java с версией не менее 1.8, приложение MetaCherchant, утилиты BWA, SAMTools, Python с версией не менее 3.0 и библиотека Pandas для языка Python. Исходный геномный контекст записан в файл `${work-dir}/output/1/graph.gfa`. Расширенный геномный контекст записан в файл `${work-dir}/output/2/graph.gfa`. Файл `${work-dir}/2/hic_map.txt` содержит информацию о Hi-C связях, записанных в формате, подходящем для использования в приложении Bandage.

ВЫВОДЫ ПО ГЛАВЕ 2

В данной главе был предложен новый алгоритм построения геномного контекста с учетом пар Hi-C ридов. Предложенный способ был реализован в виде bash скрипта. В реализации используется приложение MetaCherchant, утилиты BWA и SAMTools. Приложение MetaCherchant также было модифицировано для поддержания использования пар Hi-C ридов при построении геномного контекста.

ГЛАВА 3. ВИЗУАЛИЗАЦИЯ ГРАФА ДЕ БРЕЙНА С НІ-С СВЯЗЯМИ

В приложении Bandage граф де Брейна представляется в виде списка вершин — контигов и ребер — WGS связей. Так как контиги являются нуклеотидными последовательностями разной длины, то при укладке графа каждый контиг представляется в виде подграфа вида бамбук, число вершин которого зависит от длины контига. Таким образом, граф де Брейна имеет два вида ребер: ребра между контигами (WGS ребра) и вспомогательные ребра, ребра внутри контига. WGS ребра всегда соединяют концы контигов.

Такая структура графа позволяет добавить в граф де Брейна новый тип ребер. Данные ребра будут соединять середины контигов, если между этими контигами есть Ні-С связь. Изображаться Ні-С ребра будут в виде серых пунктирных линий.

Насыщенность серого цвета Ні-С ребер будет зависеть от веса Ні-С связи. Весом Ні-С связи между различными контигами будет называть число пар Ні-С ридов между данными контигами. Чем меньше вес Ні-С связи, тем светлее Ні-С ребро. Минимальный вес Ні-С связи равен одному, максимальный вес в данном графе будет определяться в приложение Bandage. В системе RGB черный цвет задается тройкой чисел (0, 0, 0), а светло-серый — (200, 200, 200). Таким образом, необходимо нормировать диапазон весов Ні-С связей на отрезок от 0 до 200. Самые тяжелые Ні-С связи будут сопоставляться с числом близким к 0, а самые легкие — с числом близком к 200. Предлагается делать нормировку веса Ні-С связи (w) по следующей формуле:

$$x = 200 - 200 \frac{\log(w)}{\log(w_{max})}.$$

На рисунке 3.1 приведен пример визуализации геномного контекста, содержащего Ні-С связи различного веса.

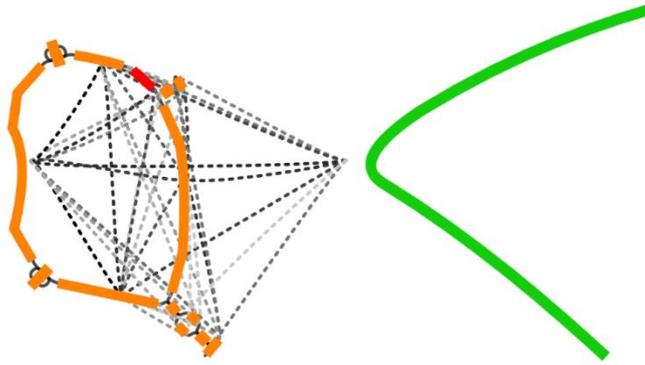


Рисунок 3.1 Визуализация геномного контекста в приложение Bandage.

3.1. ИСПОЛЬЗОВАНИЕ Hi-C РЕБЕР ПРИ УКЛАДКЕ ГРАФА ДЕ БРЕЙНА.

Hi-C ребра можно не учитывать при укладке, а просто отображать на WGS графе, соединяя середины контигов. Это самый простой способ отображения Hi-C связей, однако этот способ укладки графа приводит к большому числу пересечений между ребрами и плохой читаемости графа. На рисунке 3.2 приведен пример визуализации графа де Брейна, при укладке которого не использовались Hi-C ребра.

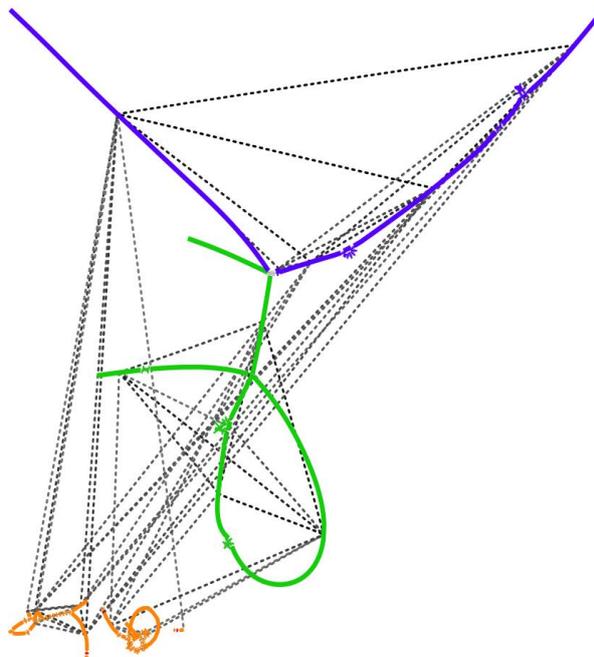


Рисунок 3.2 Укладка графа де Брейна без использования Hi-C ребер.

При укладке графа де Брейна используется алгоритм FMMM из библиотеки OGDF. Этот метод предполагает фиксированную длину ребер, поэтому все WGS ребра и все вспомогательные ребра имеют фиксированную длину. Следовательно, для использования Hi-C ребер при укладке графа Hi-C ребрам тоже нужно задать длину, поэтому в приложении была добавлена возможность задания пользователем длины Hi-C ребра. На рисунке 3.3 приведен пример, в котором в укладке графа использовалось ровно одно Hi-C ребро между каждой парой компонент связности. В данном варианте укладки связанные Hi-C ребрами компоненты связности расположены близко друг к другу, что упрощает восприятие графа. Между Hi-C ребрами стало меньше пересечений, однако все равно довольно много.

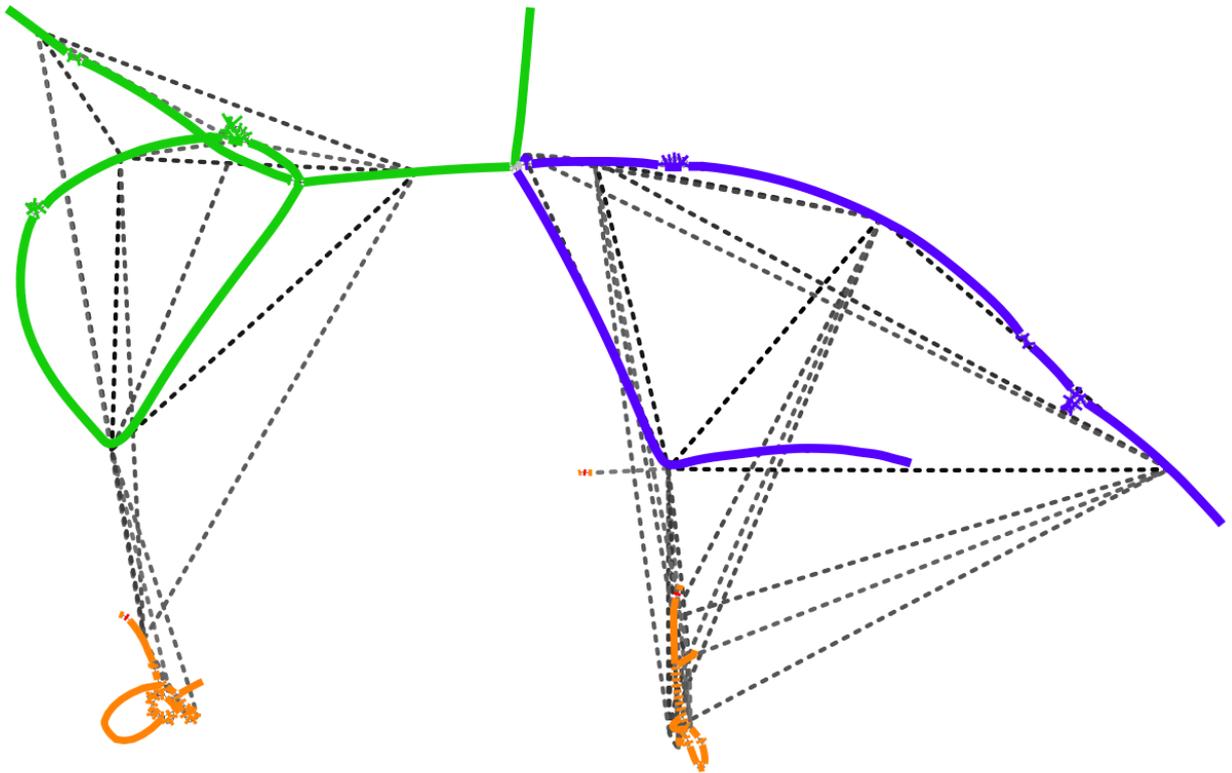


Рисунок 3.3 Укладка графа с использованием одного Hi-C ребра между различными компонентами связности.

На рисунке 3.4 приведен пример визуализации графа, в котором все отображенные Hi-C ребра участвуют в укладке графа. Это наиболее удачный способ укладки графа, в котором связанные Hi-C ребрами компоненты также расположены близко друг другу, а пересечений между Hi-C ребрами стало еще меньше. Из минусов данного вариант укладки можно выделить то, что некоторые контиги загибаются в середине, это связано с тем, что середины этих контигов стянуты Hi-C ребрами, которые имеют фиксированную длину.

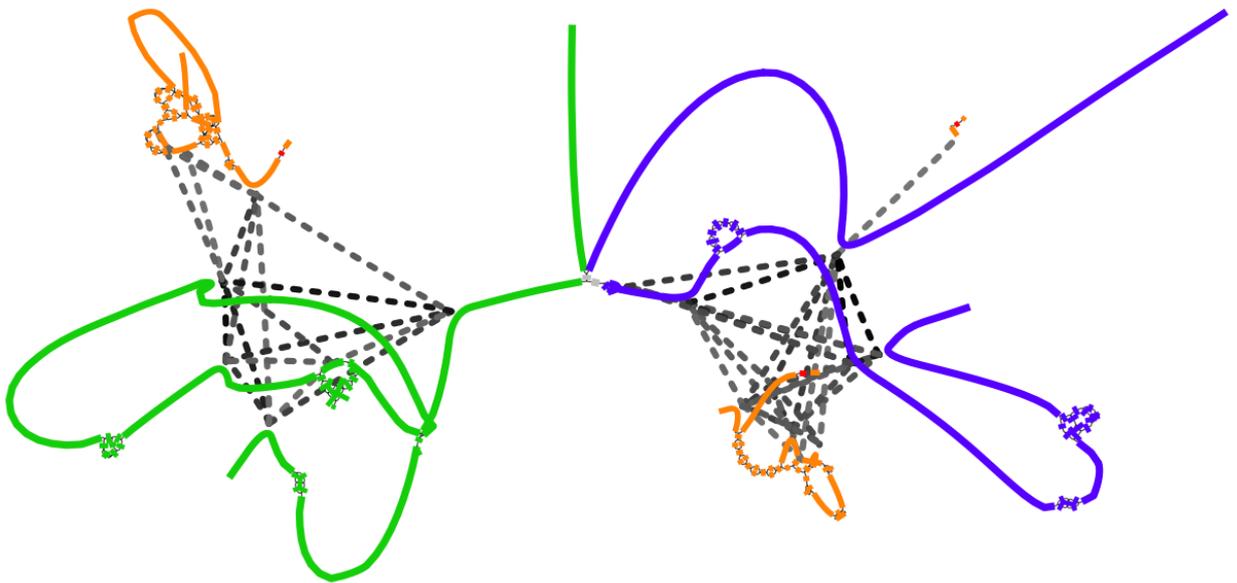


Рисунок 3.4. Все Hi-C ребра участвуют в укладке графа.

На всех рисунках 3.2–3.4 использовался один и тот же граф де Брейна и одинаковые настройки фильтрации Hi-C ребер в приложении Bandage, отличался только способ укладки графа. Как было замечено, во всех трех случаях есть недостатки, которые возникают при большом числе визуализированных Hi-C связей, однако визуализация графа, в укладке которого использовались все Hi-C связи, выглядит наиболее читаемой. Несмотря на это, пользователь может выбрать в программе любой из предложенных вариантов укладки графа.

3.2. ВЫБОР Hi-C РЕБЕР ДЛЯ ОТОБРАЖЕНИЯ НА ГРАФЕ ДЕ БРЕЙНА

Большое число ребер не только загромождает граф де Брейна и усложняет его анализ, но также увеличивает время его укладки и отрисовки. Для уменьшения числа визуализированных Hi-C ребер были предложены различные способы фильтрации Hi-C ребер. Во-первых, предлагается визуализировать только те Hi-C ребра, чей вес больше некоторого порогового значения, заданного пользователем. При WGS и Hi-C секвенировании могут происходить ошибки чтения, из-за чего мы не можем доверять всем Hi-C связям, поэтому в приложении была реализована возможность задания минимального веса Hi-C связей. Во-вторых, предлагается отображать только те Hi-C связи, которые соединяют контиги, чья длина нуклеотидной последовательности превышает заданное пользователем значение минимальной длины контига. Это также обусловлено тем, что при анализе требуется обращать внимание только на наиболее значимые Hi-C связи, из-за их большого числа.

На рисунке 3.5 приведен пример визуализации геномного контекста с минимальным весом Hi-C связей равным одному и минимальной длиной контига равной сто пар нуклеотид (п.н).

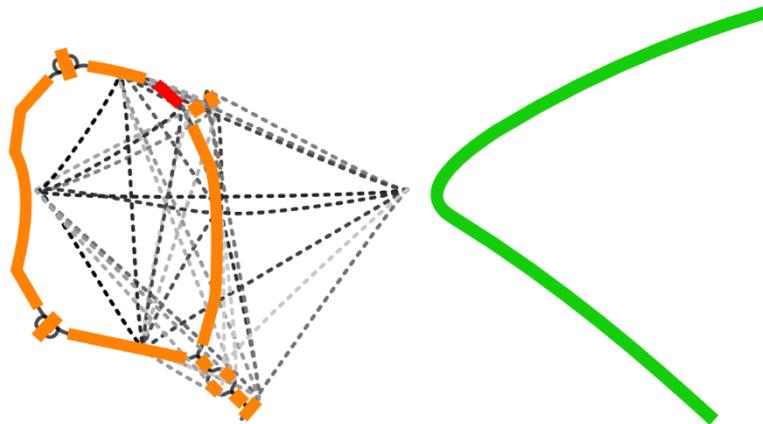


Рисунок 3.5. Геномный контекст: минимальный вес Hi-C связи = 1 и минимальная длина контига = 100 п.н.

На рисунке 3.6 приведен пример визуализации того же геномного контекста, что и на рисунке 3.5, но минимальный вес был увеличен до семидесяти, а минимальная длина контига осталась неизменной.

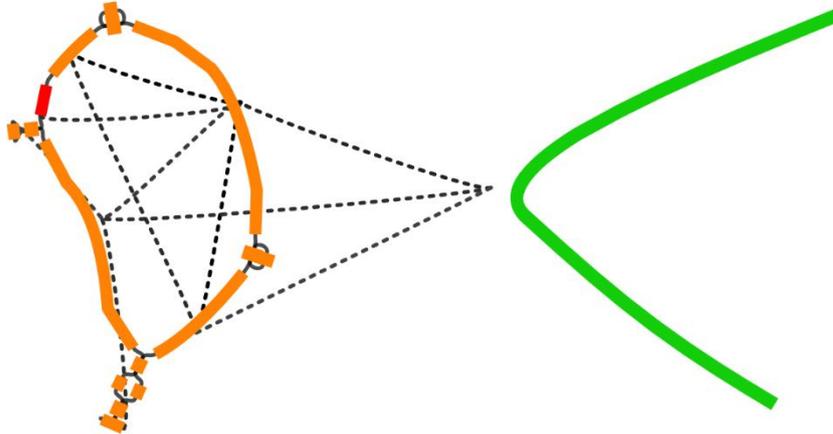


Рисунок 3.6. Геномный контекст: минимальный вес Hi-C связи = 70 и минимальная длина контига = 100 п.н.

На рисунке 3.7 наоборот минимальный вес Hi-C связи остался равным одному, а минимальная длина нуклеотидной последовательности была увеличена до 5000 пар нуклеотид.

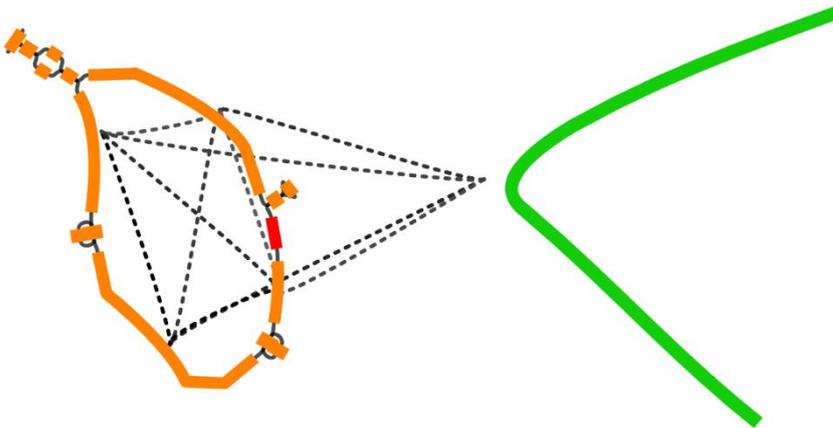


Рисунок 3.7. Геномный контекст: минимальный вес Hi-C связи = 1 и минимальная длина контига = 5000 п.н.

Следует отметить, что для дальнейшего анализа, как правило, интересны только Hi-C связи, соединяющие разные компоненты связности. Компоненты

связности определяются по графу де Брейна без Hi-C ребер. В листинге 3.1 приведен псевдокод определения компонент связности при помощи обхода графа в глубину.

Листинг 3.1 – Поиск компонент связности.

```

def findGraphComponents (allNodes):
    newComponentId = 1;
    for node in nodes:
        if (node.componentId is empty):
            dfs(node, newComponentId)
            componentId += 1

def dfs(node, newComponentId):
    if (node.componentId is empty):
        node.componentId = newComponentId
        for (nextNode in node.neighbours):
            dfs(nextNode, newComponentId)

```

Часто бывает важен только факт наличия Hi-C связей между двумя компонентами связности и не важно между какими именно контигами есть Hi-C связи, поэтому в приложении Bandage мной было реализовано три способа визуализации Hi-C ребер:

- 1) Отображаются все Hi-C ребра (рисунок 3.8.а).
- 2) Отображаются все Hi-C ребра между различными компонентами связности (рисунок 3.8.б).
- 3) Отображается ровно одно ребра с максимальным весом между каждой парой компонент связности (рисунок 3.8.в).

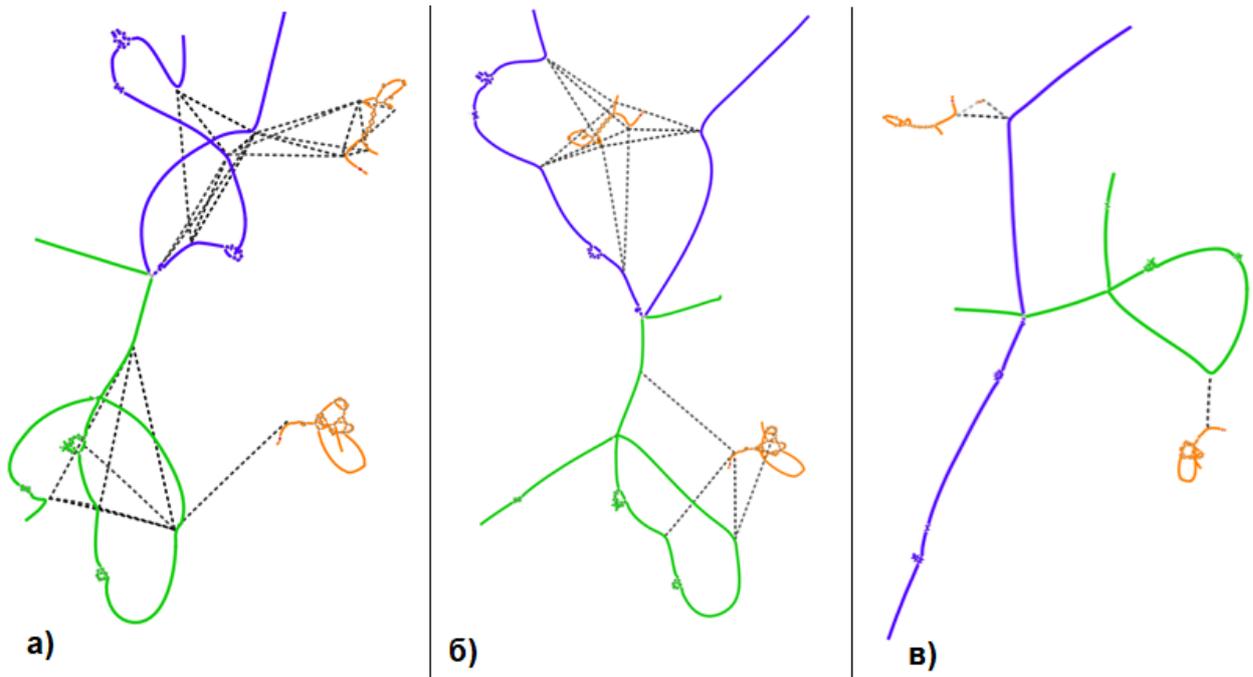


Рисунок 3.8. Визуализация геномного контекста с отображением всех Hi-C ребер (а), всех Hi-C ребер между различными компонентами связности (б), одного Hi-C ребра максимального веса между различными компонентами (в).

В первом случае будет отрисовано максимальное число Hi-C ребер, а в третьем случае — минимальное. Во всех трех вариантах отрисовки будут отображены только те Hi-C ребра, которые имеют вес больше порогового значения минимального веса, а также соединяющие контиги, чья длина в п.н. больше заданного значения минимальной длины контига.

3.3. ВИЗУАЛИЗАЦИЯ РЕАЛЬНЫХ МЕТАГЕНОМНЫХ ДАННЫХ

При работе с реальными данными, а именно при работе с геномным контекстом, построенным вокруг гена TEM на образце микробиоты кишечника человека во время антибактериальной терапии, было выявлено, что геномный контекст может иметь большой размер и большое число компонент связности, однако интерес представляет лишь его часть. На рисунке 3.9 приведен пример

визуализации геномного контекста вокруг гена TEM с отображением всех компонент связности и только тех Hi-C ребер, которые соединяют различные компоненты связности с целевой компонентой. Целевой компонентой будем называть компоненту связности, содержащую целевой ген.

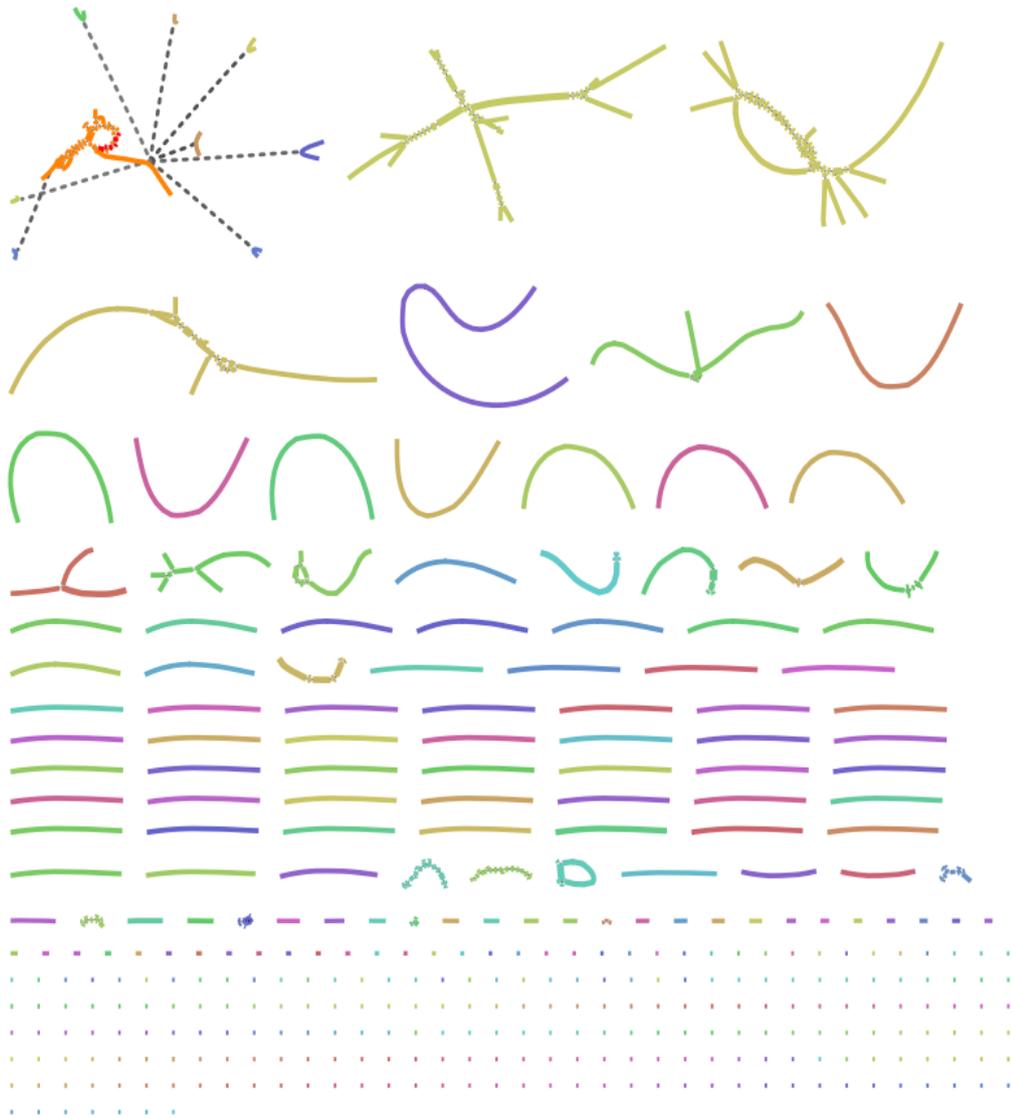


Рисунок 3.9. Геномный контекст вокруг гена TEM.

Данный геномный контекст содержит большое число компонент связности, чей размер меньше 1000 пар нуклеотид. Однако нас интересуют только те компоненты связности, чей суммарный размер более 1000 пар нуклеотид. Если в графе присутствует много коротких фрагментов генома,

которые не были соединены в более длинные нуклеотидные цепочки, то короткие фрагменты генома не вызывают доверия, так как могли быть получены в результате ошибок секвенирования или могут быть биологическим мусором.

Для сокращения размера графа предлагается не визуализировать слишком маленькие компоненты связности. Также предлагается отображать только те компоненты связности, которые имеют Hi-C связи с целевой компонентой связности. Таким образом будет отфильтровано большинство компонент связности, а оставшиеся компоненты связности будет удобно анализировать. Два предложенных выше способа сокращения числа визуализированных компонент связности были мной реализованы и возможность их использования была добавлена в пользовательский интерфейс приложения Bandage. Следует отметить, что WGS и Hi-C ребра отображаются, только если вершины, которые они соединяют визуализированы.

На рисунке 3.10 приведена визуализация геномного контекста вокруг гена TEM с отображением только тех компонент связности, которые имеют Hi-C связи с целевой компонентой. При этом были визуализированы только те Hi-C ребра, которые соединяют различные компоненты связности с целевой компонентой.

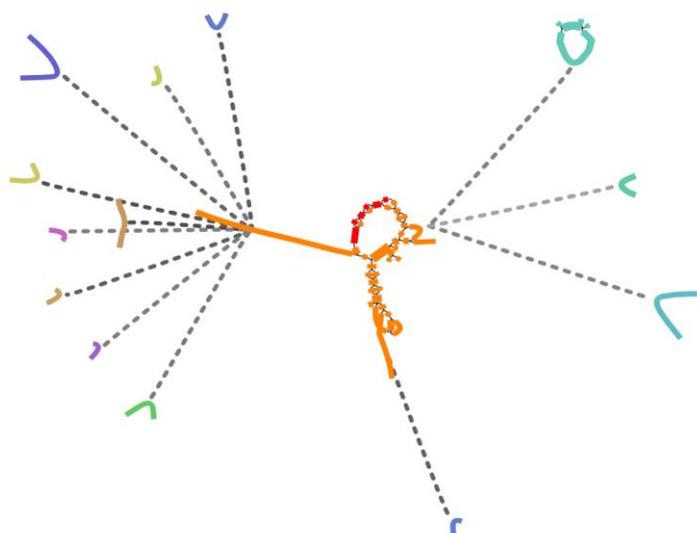


Рисунок 3.10. Геномный контекст с сокращением числа вершин.

3.4. АВТОМАТИЧЕСКИЙ ПОДБОР ПАРАМЕТРОВ ОТРИСОВКИ

Для упрощения работы пользователя с приложением Bandage был разработан алгоритм автоматического подбора параметров визуализации Hi-C связей. В автоматическом режиме отображаются только те вершины, которые расположены в компонентах связности, которые имеют Hi-C связи с целевыми компонентами. Также в данном режиме все Hi-C связи участвуют в укладке графа. Пороговое значение минимальной длины контига задается в каждой компоненте связности отдельно и равно максимальному значению из 1000 и средней длины контигов в данной компоненте связности. Если хотя бы один из двух контигов, соединенных Hi-C связью, не удовлетворяет требованию минимальной длины контига в свой компоненте, тогда данное Hi-C ребро не визуализируется. При определении компонент связности, которые имеют Hi-C связи с целевой компонентой, используются только те Hi-C связи, которые удовлетворяют требованию минимальной длины контига. Как было сказано ранее есть четыре варианта визуализации графа де Брейна с Hi-C связями:

- 1) Визуализация всех Hi-C ребер.
- 2) Визуализация всех Hi-C ребер, соединяющих различные компоненты связности.
- 3) Визуализация одного Hi-C ребра между каждой парой компонент связности.
- 4) Визуализация ровно одного ребра между каждой компонентой связности и целевой компонентой связности.

В автоматическом режиме выбор способа отображение Hi-C ребер будет зависеть от числа компонент связности N :

- 1) Если $N \leq 2$, тогда будут отображены все Hi-C ребра.
- 2) Если $N \leq 5$, тогда будут отображены все ребра между различными компонентами связности (рисунок 3.11).

- 3) Если $5 < N \leq 10$, тогда между каждой парой компонент связности, будет отображено одно H_i -C ребро с максимальным весом (рисунок 3.12).
- 4) Если $N > 10$, то будут отображены только H_i -C ребра, соединяющие компоненты связности с целевой компонентой, при этом между каждой парой компонент связности будет отображено не более одного H_i -C ребра (рисунок 3.13).

На рисунках 3.11–3.13 приведены примеры визуализации графа де Брейна в автоматическом режиме для пяти, девяти и четырнадцати компонент связности соответственно.

На рисунке 3.11 изображен геномный контекст, построенный вокруг гена *sfxA* и содержащий пять визуализированных компонент связности. Это средний по размеру геномный контекст содержащий примерно 47 тысяч контигов.

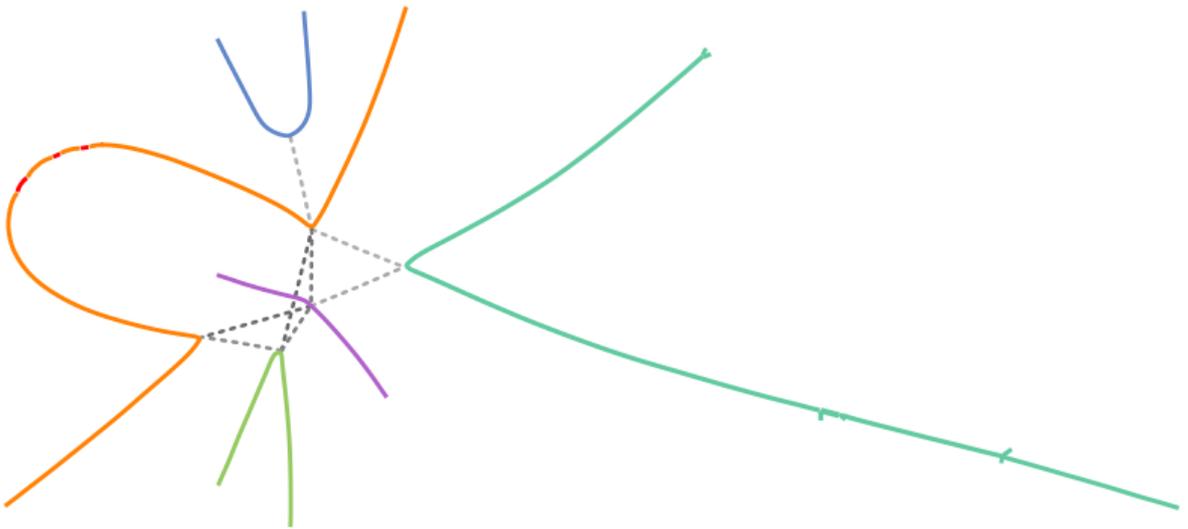


Рисунок 3.11. Граф де Брейна, содержащий 5 компонент связности.

На рисунке 3.12 изображен геномный контекст, построенный вокруг гена *OXA* и отображающий 9 компонент связности. Это большой геномный контекст, содержащий около 205 тысяч контигов.

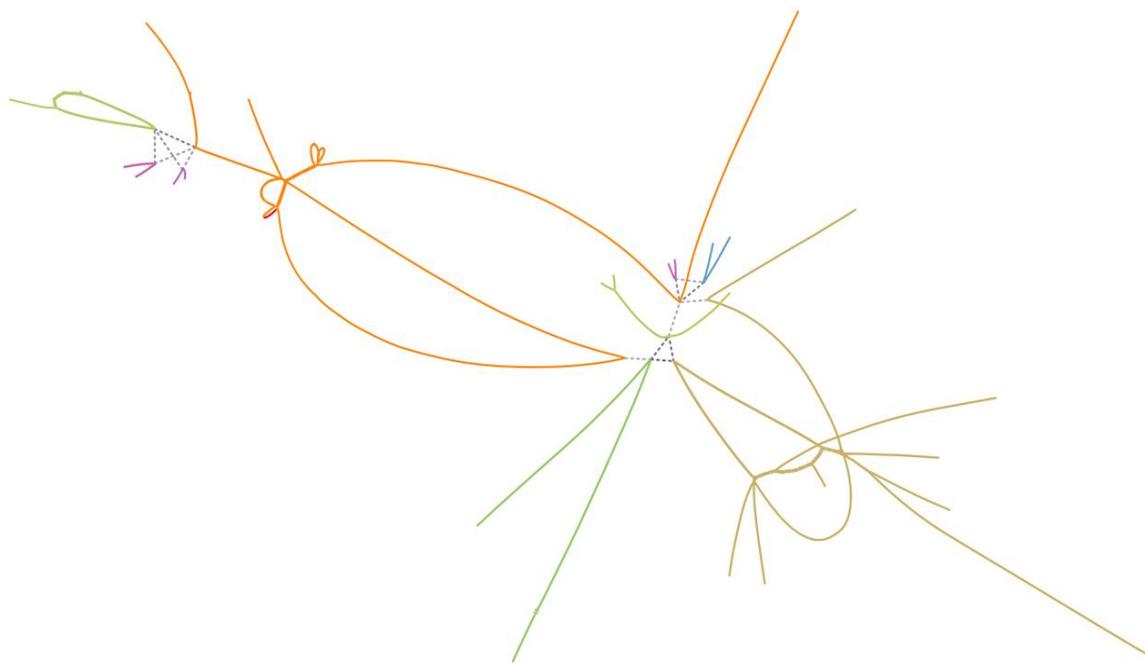


Рисунок 3.12. Граф де Брейна, состоящий из 9 компонент связности.

На рисунке 3.13 изображен геномный контекста вокруг гена TEM, содержащий 14 отображенных компонент связности. Несмотря на самое большое число визуализированных компонент связности, геномный контекст имеет самый маленький размер: около 650 контигов.

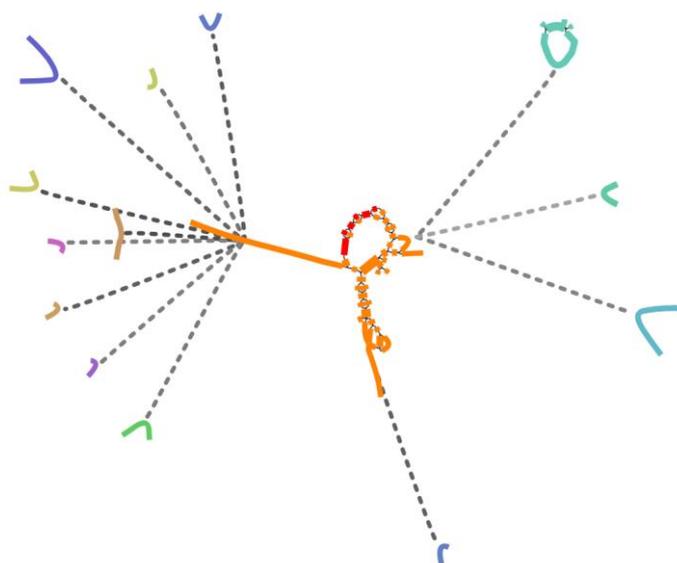


Рисунок 3.13. Граф де Брейна, состоящий из 14 компонент связности.

Все три визуализированных графа де Брейна выглядят читаемыми, содержат небольшое число вершин и Hi-C ребер. Также следует отметить, что время отрисовки самого большого геномного контекста (205000 контигов) в автоматическом режиме составило примерно 7 минут, при попытке визуализировать геномный контекст целиком, время отрисовки было около двух часов.

3.5. ТАКСОНОМИЧЕСКИЙ АНАЛИЗ

В биоинформатике для определения видовой принадлежности контига используется таксономический анализ. Классификацию контигов можно сделать при помощи программы Kraken 2 [15]. Таксоны являются иерархической структурой и содержат 8 уровней: домен, царство, тип, класс, порядок, семейство, род и вид. Каждый следующий по уровню таксон является уточнением своего предка. Например, царство — бактерии, тип — протеобактерии.

Чем больше номер уровня, тем больше таксонов содержит данный уровень, поэтому таксоны предлагается хранить в виде дерева, что позволит использовать меньше памяти. Пример такого дерева приведен на рисунке 3.14. Каждой вершине графа де Брейна будет соответствовать список таксонов, который является путем в дерево таксонов от таксона максимального уровня до корня дерева, поэтому для каждой вершины достаточно хранить указатель на таксон с максимальным уровнем. Например, если вершина графа де Брейна ссылается на таксон уровня класс *Gamma*proteobacteria, тогда данной вершине соответствует иерархия таксонов: Gamma*proteobacteria*, Proteobacteria и Bacteria.

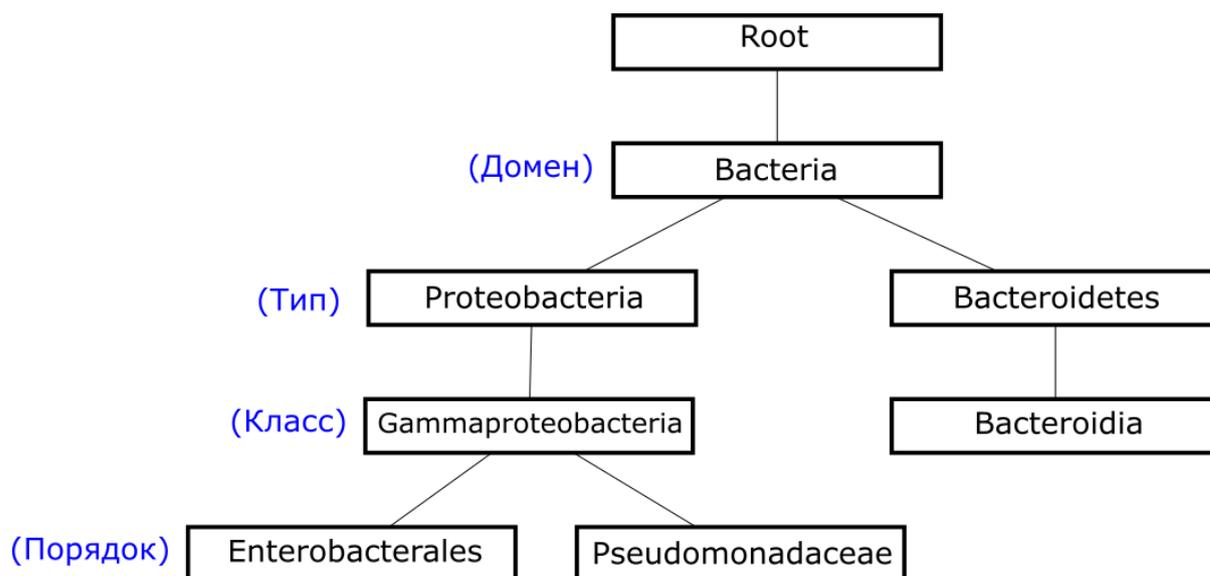


Рисунок 3.14. Дерево таксонов.

Для визуализации результатов таксономического анализа при визуализации графа де Брейна мной была реализована раскраска вершин по таксонам с возможностью выбора уровня таксонов, используемых в раскраске. Одинаковым таксонам соответствует одинаковый цвет. Также добавлена возможность подписи названия или идентификатора таксона для каждой вершины графа с выбором уровня таксона или же подпись таксона с максимальным уровнем. Помимо этого, при выделении вершины на визуализированном графе в специальном окне отображается вся иерархия таксонов для выбранной вершины. Следует отметить, что таксоны определены не для всех вершин. Для отображения вершин без таксонов используется серый цвет. На рисунке 3.14 приведен пример раскраски графа де Брейна по таксонам уровня «класс».

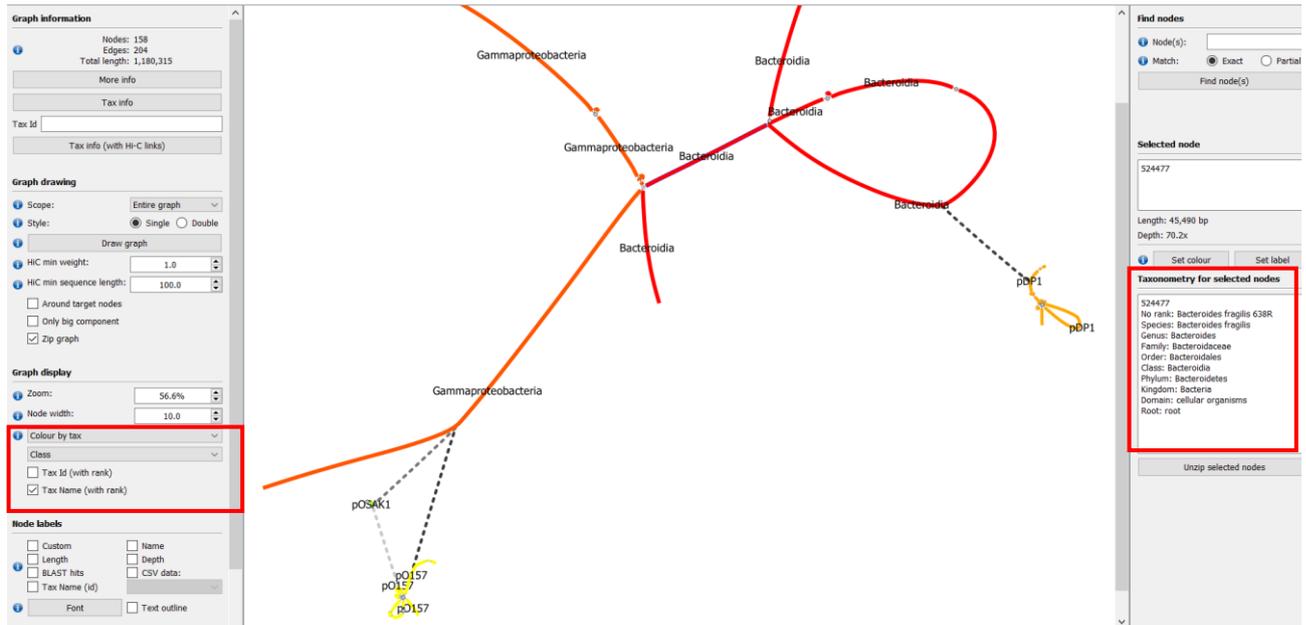


Рисунок 3.14 Использование таксономической раскраски в графе де Брейна.

В результате таксономического анализа таксон определен не для всех вершин, однако с высокой вероятностью таксон вершина, имеющая в соседях слева и справа только вершины с одним и тем же или неопределенным таксоном, тоже имеет таксон, совпадающий с таксонами ее соседей. В таких случаях предлагается продлевать раскраску на неизвестные таксоны, в соответствии с таксонами их соседей. То есть вершине с неизвестным таксоном будет присвоен таксон A , если в ее соседях справа и слева, есть как минимум один таксон A , а для остальных соседей таксон не определен. Такая вершина будет заштрихована в цвет таксона ее соседей, пример приведен на рисунке 3.15.

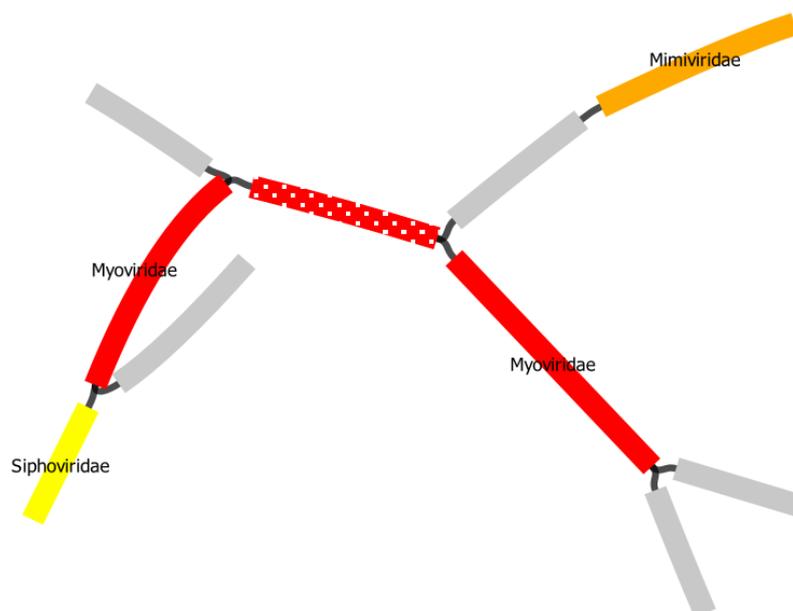


Рисунок 3.15. Граф де Брейна с продлением раскраски по таксонам на вершины с неизвестными таксонами.

Граф де Брейна, построенный на реальных данных и содержащий геномы различных бактерий, может быть очень запутанным (рисунок 3.16.а), при этом часто возникает необходимость исследовать геном каждой бактерии отдельно. Для этого мной была реализована возможность фильтрации вершин графа де Брейна по их таксонам. В этом случае будут изображены только те вершины, которые имеют указанный пользователем таксон или не имеют таксонов, однако связаны WGS ребрами с уже визуализированными вершинами. Для этого из каждой не посещенной ранее вершины, которая имеет указанный таксон, запускается алгоритм обхода графа в глубину, который посещает только те вершины, которые не имеют таксона или их таксон совпадает с заданным пользователем таксоном, будут визуализированы только посещенные вершины. На рисунке 3.16.б приведена визуализация генома, соответствующая одной бактерии, выбранной из графа де Брейна, приведенного на рисунке 3.16.а.

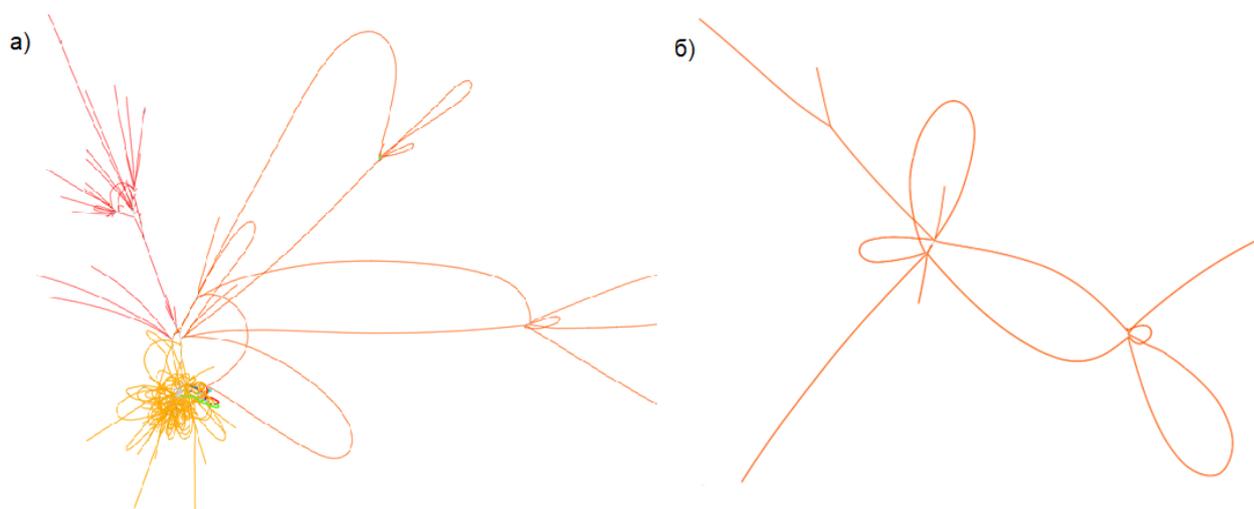


Рисунок 3.15 а) Граф де Брейна, содержащий геномы нескольких различных бактерий, б) Граф де Брейна, содержащий геном бактерии *Levilactobacillus*, выделенной из графа де Брейна под пунктом «а».

Помимо этого, была реализована возможность использования Hi-C связей при визуализации вершин с заданным таксоном. В этом случае будут визуализированы только те, вершины, которые имеют заданный пользователем таксон или таксон, имеющий Hi-C связи с заданным. То есть при обходе графа в глубину будем проходить не только по WGS ребрам, но и по Hi-C ребрам, но только тем, которые соединяют вершины с известными таксонами, один из которых совпадает с указанным пользователем таксоном. Вершины с неизвестными таксонами будут отображаться на графе только если они связаны WGS ребрами с уже отображенными вершинами. Этот режим визуализации, примененный к плазмиде *pO157*, содержащейся в графе де Брейна, изображенном на рисунке 3.14, позволит отобразить плазмиду *pO157*, плазмиду *pOSAK1* и *Кишечную палочку*. Именно этот способ визуализации позволяет определить в клетках какой из двух бактерий может находиться плаزمид *pO157*. Пример визуализации плазмиды и бактерии, полученной при помощи данного способа, приведен на рисунке 3.17

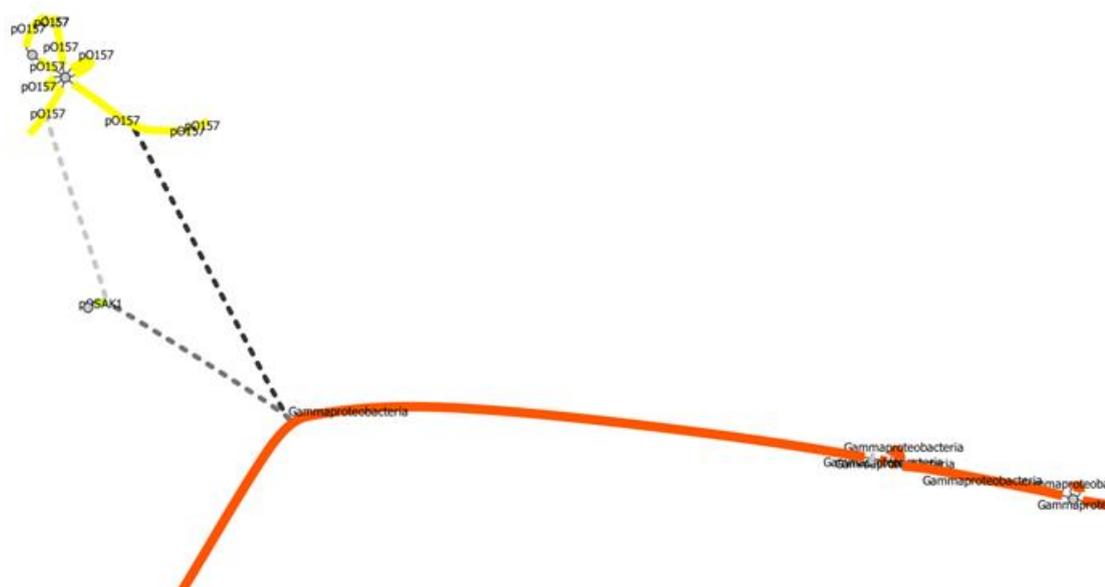


Рисунок 3.17. Визуализации плазмиды *pO158* с учетом Hi-C связей.

Для удобства пользователя мной была реализована возможность получить статистический отчет о представленности таксонов в графе Де Брейна. Этот отчет разделен на уровни таксонов. Каждый блок содержит список таксонов под данным уровнем. Таксоны отсортированы по убыванию их представленности. Представленность таксона равна суммарной длине контигов, имеющих данный таксон. Помимо представленности таксона в данной таблице хранится его название, идентификатор и число контигов. Фрагмент такого отчета приведен в таблице 2.1.

Таблица 2.1 Представленность таксонов на разных уровнях.

Название	Id	Длина контигов	Число контигов
Kingdom			
Bacteria	2	1178009	142
Other sequences	28384	195707	73
Phylum			
Bacteroidetes	976	591053	79
Proteobacteria	1224	586771	60
Plasmids	36549	195707	73

Class			
Bacteroidia	200643	591053	79
Gamma proteobacteria	1236	586771	60
pDP1	365495	99789	46
pO157	365491	92508	24
pOSAK1	365499	3410	3

Помимо общего отчета можно получить отчет по Hi-C связям для заданного пользователем таксона. Этот отчет содержит список таксонов, которые имеют Hi-C связи с выбранным таксоном. Отчет содержит название таксона, его идентификатор, суммарный размер контигов данного таксона, суммарный вес Hi-C связей между выбранным таксоном и текущим, а также процентное соотношение веса Hi-C связей с текущим таксоном и веса всех Hi-C связей выбранного таксона. Данные отсортированы по проценту Hi-C связей между таксонами. Таксон может иметь Hi-C связи сам с собой, а также Hi-C связи с контигами без таксонов. Пример такого отчета приведен в таблице 2.2.

Таблица 2.2. Данные о Hi-C связях для таксона *pO157*.

Название таксона	Идентификатор таксона	Длина контигов	Вес Hi-C связей	Процент числа Hi-C связей от числа всех Hi-C связей выбранного таксона
pO157	365491	92508	16492	48.04%
Gamma proteobacteria	1236	586771	8910	5.99%
pOSAK1	365499	3410	10	1.04%

3.6. СЖАТИЕ ГРАФА

Граф де Брейна состоит из длинных контигов и небольших скоплений коротких контигов, которые могут быть вариативностью генома в различных штаммах одной бактерии, небольшими повторяющимися фрагментами генома или ошибками секвенирования.

При анализе метагеномных данных часто не нужен столько детальный граф де Брейна, скопления коротких контигов лишь зашумляют граф, поэтому мной была реализована возможность сжатия графа. Этот функционал позволяет объединять соседние короткие контиги в одну вершину.

Будем считать контиг коротким, если его длина меньше 1000 пар нуклеотид. Контиги, чья длина более тысячи пар нуклеотид, сжиматься не будут. Реализовать данный функционал предлагается следующим образом:

- 1) Найти длинный контиг, имеющий wgs ребра с короткими контигами, чья длина как минимум в 4 раза меньше длинного контига. Назовем данный длинный контиг опорным.
- 2) Запустим обход графа в глубину от всех коротких контигов, которые имеют входящее ребра от опорного контига.
- 3) Обход графа в глубину будет идти только по коротким контигам. Сразу запомним все длинные контиги, связанные с короткими.
- 4) Объединим в одну вершину все короткие контиги, найденные при помощи обхода графа в глубину, запущенного для всех детей опорного контига. Добавим ребра между новой вершиной и длинными контигами, найденными ранее.
- 5) Затем попытаемся сжать короткие контиги, которые имеют исходящие wgs ребра к опорному контигу. То есть один опорный контиг может быть связан с двумя сжатыми вершинами слева и справа.

Сжатые вершины будут изображены в виде серого круга, диаметр которого совпадает с шириной отрисованных контиггов. Пример сжатия графа приведен на рисунке 3.18.

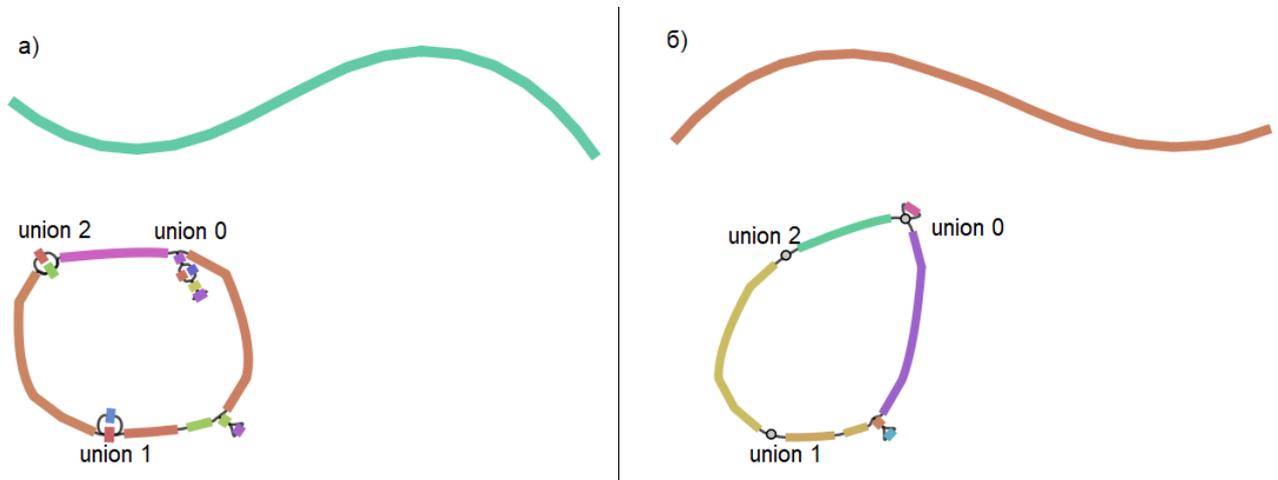


Рисунок 3.18. Сжатие графа де Брейна: а) не сжатый граф де Брейна, б) сжатый граф.

В приложении Bandage был добавлен функционал, который позволяет сжимать и разжимать весь граф целиком. При этом алгоритм поиска сжатых вершин и создания новых вершин, объединяющих сжатые, будет запущен только один раз.

Также мной была реализована возможность раскрывать по одной сжатой вершины. Для этого при укладке нового графа, в качестве начальных координат вершин будут использоваться координаты с предыдущей визуализации, так как отличия в числе и расположении вершин не большое, то при укладке графа будет использовано меньшее число итераций, благодаря чему вершины, присутствующие на предыдущей визуализации графа, практически не поменяют своего расположения. Также будет сохранена раскраска вершин с предыдущей визуализации, даже если использовалась рандомная раскраска вершин. Пример раскрытия одной сжатой вершины приведен на рисунке 3.19.

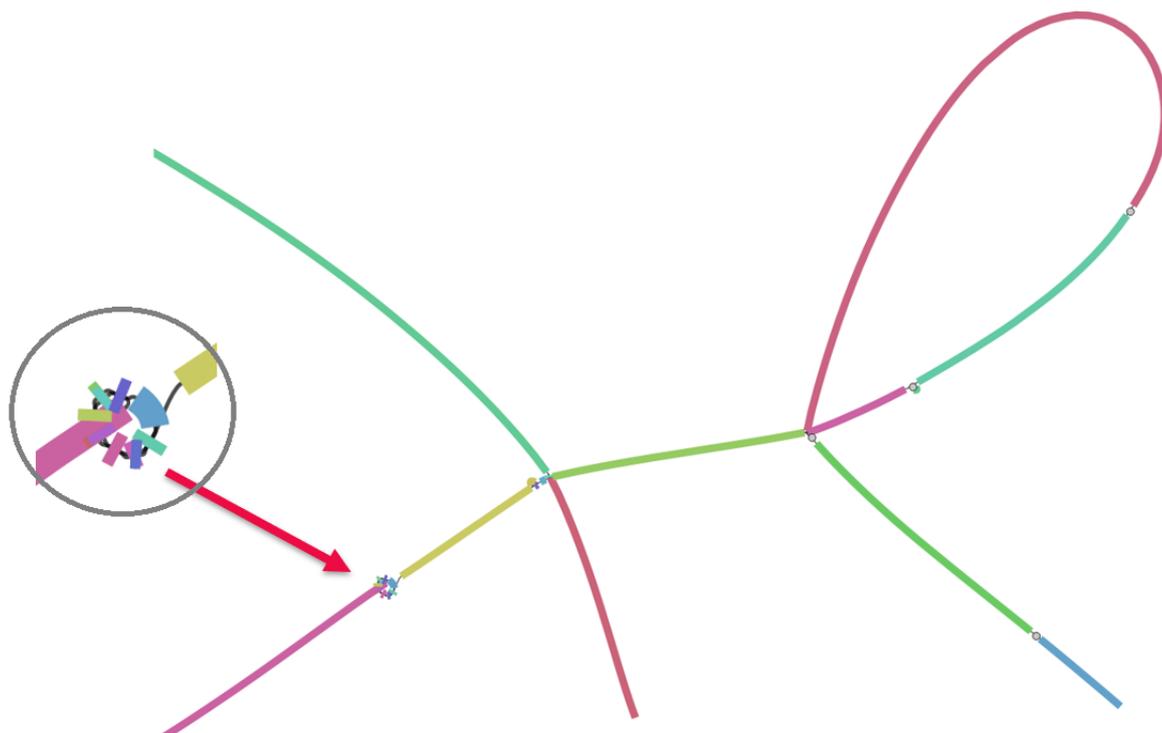


Рисунок 3.19. Раскрытие одной сжатой вершины в графе де Брейна.

ВЫВОДЫ ПО ГЛАВЕ 3

В данной главе был описан алгоритм визуализации Hi-C связей в графе де Брейна. Были разработаны различные способы сокращения числа визуализированных вершин и Hi-C ребер. Был разработан режим с автоматическим подбором параметров отрисовки Hi-C связей и показана эффективность его использования на реальных метагеномных данных. Был разработан способ использования результатов таксономического анализа при визуализации графа де Брейна. Помимо этого, была добавлена возможность выделения и отрисовки геномов выбранных существей (бактерий, плазмид и вирусов). Данные модификации приложения Bandage позволяют при анализе графа де Брейна учитывать не только данные полногеномного секвенирования, но и данные Hi-C секвенирования, и результаты таксономического анализа.

ГЛАВА 4. ТЕСТИРОВАНИЕ НОВОГО СПОСОБА ПОСТРОЕНИЯ И ВИЗУАЛИЗАЦИИ ГЕНОМНОГО КОНТЕКСТА

Тестирование реализованных методов построения и визуализации геномного контекста происходило на реальных и сгенерированных данных.

При построение геномного контекста на сгенерированных данных пары WGS и Hi-C ридов получаются с помощью утилит генерации InSilicoSeq [16] и sim3C [17], а не биологических методов Hi-C и WGS секвенирования.

4.1. БАКТЕРИЯ САЛЬМОНЕЛЛА И ЕЕ ПЛАЗМИДА

В данном примере происходило построение геномного контекста для фрагмента генома бактерии Сальмонелла и ее плазмиды. Следует отметить, что это сгенерированный пример. Hi-C риды генерировались исходя из предположения, что плазида лежит в клетке бактерии Сальмонелла. В таблице 4.1 приведены параметры для генерации данных и построения геномного контекста, а также некоторые статистические данные. Для построения геномного контекста использовался скрипт построения контекста с учетом Hi-C связей, предложенный в главе 2.

Таблица 4.1. Параметры геномного контекста бактерии Сальмонелла.

Длина фрагмента генома бактерии	~100 000 пар нуклеотид
Длина генома плазмиды	~100 000 пар нуклеотид
Количества пар WGS ридов	30 000 пар ридов
Количество пар Hi-C ридов	30 000 пар ридов
k	31
Покрытие	5
Максимальный радиус графа	100 000

Размер исходного геномного контекста	~100 000 пар нуклеотид
Размер расширенного геномного контекста	~200 000 пар нуклеотид
Минимальный вес Hi-C связи	30
Минимальная длина контига	100 п.н.

В качестве целевого гена использовался один из генов плазмиды. Также в данном примере максимальный радиус графа больше, чем длина нуклеотидной последовательности плазмиды, поэтому геном плазмиды вошел в исходный геномный контекст целиком. На рисунке 4.1 приведен исходный геномный контекст. Целевой ген отмечен красным цветом, а плазида — оранжевым.

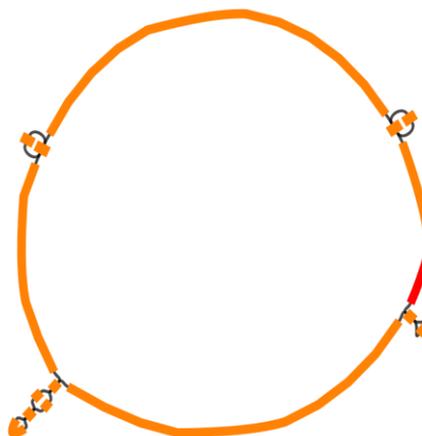


Рисунок 4.1 Исходный геномный контекст для бактерии Сальмонелла.

На рисунке 4.2 приведен примеры расширенного геномного контекста для бактерии сальмонелла и ее плазмиды. Фрагмент бактерии Сальмонелла окрашен в зеленый цвет, а плазида — в оранжевый цвет. Как мы видим, фрагмент генома бактерии Сальмонелла присутствует в расширенном геномном контекст, так как между геномами бактерии и плазмиды есть Hi-C связи. Полученная визуализация соответствует ожиданиям.

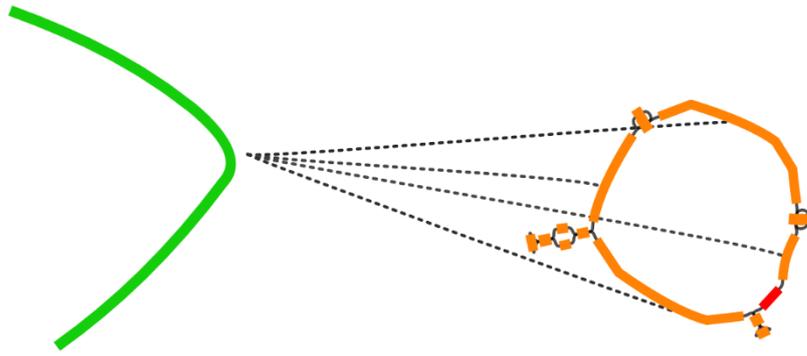


Рисунок 4.2. Расширенный геномный контекст бактерии Сальмонелла и ее плазмиды.

4.2. КИШЕЧНАЯ ПАЛОЧКА, БАКТЕРИОИД ФРАГИЛИС И ИХ ПЛАЗМИДЫ

Это также сгенерированный пример. Для его генерации использовался фрагмент генома Кишечной палочки длиной в 500000 пар нуклеотидов и фрагмент генома Бактероида Фрагилис длиной в 500000 пар нуклеотид. Была эмулирована ситуация, когда метагеномные данные содержат две клетки бактерии: одна клетка содержит хромосомную ДНК Кишечной палочки и две ее плазмиды pO157 и pOSAK1, а в другой клетке содержится хромосомная ДНК Бактероида и его плазмида. На рисунке 4.3 приведено схематическое изображение этих клеток.

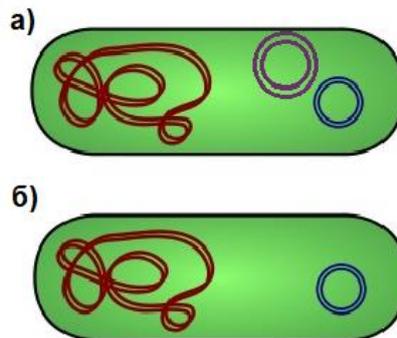


Рисунок 4.3. Клетка бактерии Кишечная палочка(а) и клетка бактерии Бактероид Фрагилис (б).

В таблице 4.2 приведены размеры геномов и другие параметры, используемые при генерации и построении геномного контекста для данного примера. Геномный контекст, как и в предыдущем примере, строился с использованием Hi-C связей.

Таблица 4.2. Параметры геномного контекста Кишечной палочки и бактерии Бактероид Фрагилиса.

Длина фрагмента генома Кишечной палочки	~500 000 пар нуклеотид
Длина фрагмента генома Кишечной палочки	~500 000 пар нуклеотид
Длина фрагмента генома pO157	~94 000 пар нуклеотид
Длина фрагмента генома pOSAK1	~3 300 пар нуклеотид
Длина фрагмента генома плазмиды 1	~99 000 пар нуклеотид
Количества пар WGS ридов	400 000 пар ридов
Количество пар Hi-C ридов	400 000 пар ридов
k	55
Покрытие	4
Максимальный радиус графа	100 000
Размер исходного геномного контекста	~203 000
Размер расширенного геномного контекста	~1 200 000
Минимальный вес Hi-C связей	40
Минимальная длина контигов, соединенных Hi-C ребром	1 000

В качестве целевых генов были взяты три гена из трех разных плазмид. Этот пример также демонстрирует возможность построения геномного контекста с учетом Hi-C связей сразу для нескольких целевых генов. Геномы Кишечной палочки и Бактероида пересекаются. Для контиг, вошедших в

расширенный геномный контекст, был проведен таксономический анализ при помощи программы Kraken 2. На рисунке 4.4 приведена визуализация геномного контекста, описанного ранее. Как и ожидалось, две плазмиды имеют Hi-C ребра с контигами, входящими в геном Кишечной палочки, а контиги, содержащиеся в геноме Бактероида, имеют Hi-C ребра только с одной плазмидой. При визуализации геномного контекста были использованы результаты таксономического анализа. Использование результатов таксономического анализа позволяет визуально разделить хромосомные ДНК Кишечной палочки (*Escherichia coli*) и Бактероида (*Bacteroides fragilis*). Следует отметить, что плазмида 1 (плазмида в клетке Бактероида) не была классифицирована, поэтому покрашена в серый цвет. Подписи на контигах совпадают с данными таксономического анализа.

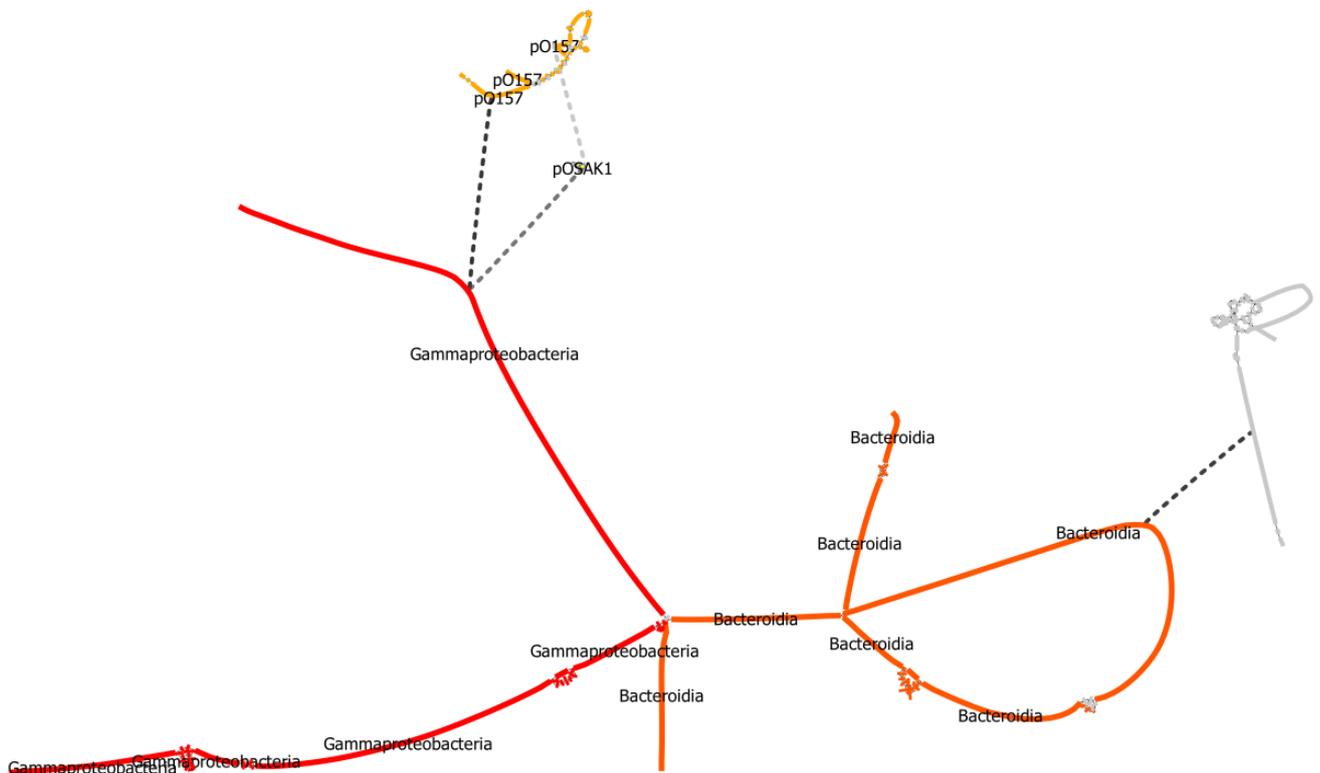


Рисунок 4.4. Геномный контекст Кишечной палочки и Бактероида.

4.3. ОБРАЗЕЦ МИКРОБИОТЫ КИШЕЧНИКА ПРИ АНТИБАКТЕРИАЛЬНОЙ ТЕРАПИИ

В качестве реальных данных был взят образец, исследуемый в статье «Hi-C Metagenomics in the ICU: Exploring Clinically Relevant Features of Gut Microbiome in Chronically Critically Ill Patients» [18]. Эти данные были получены в результате обсервационного исследования в отделении реаниматологии Федерального научно-клинического центра реаниматологии и реабилитологии, Москва, Российская Федерация. В исследовании участвовали два хронически тяжелобольных пациента: пациент А — женщина 75 лет после внутримозгового кровоизлияния и пациент Б — мужчина 74 лет после ишемического инсульта. Оба пациента проходили антибактериальную терапию.

В данной работе рассматривается образец IC6, взятый у пациента Б при подозрении на бактериальную инфекцию. Во время взятия образца пациент проходил терапию антибиотиками Цефоперазон и Сульбактам. В качестве целевых генов для тестирования построения и визуализации геномного контекста были выбраны ARG гены *cfxA*, *TEM* и *OXA*.

В таблице 4.3 приведены размеры геномов и другие параметры, используемые при построении геномного контекста для данного примера.

Таблица 4.3 Параметры для построения геномного контекста вокруг генов *TEM*, *cfxA* и *OXA*.

	TEM	cfxA	OXA
Количества пар WGS ридов	~109 000 000 пар ридов		
Количество пар Hi-C ридов	~69 000 000 пар ридов		
k	55	31	31
Покрытие	500	3 000	700
Максимальный радиус графа	3 000	3 000	3 000

В результате построения расширенного геномного контекста по предложенному алгоритму были получены графы де Брейна с параметрами, представленными в таблице 4.4. Также в таблице 4.4 представлен размер визуализированного графа де Брейна, полученный при использовании режима автоматического подбора параметров. Легко заметить, что автоматический подбор параметров уменьшает размер графа де Брейна более чем на три четверти.

Таблица 4.4. Размеры геномных контекстов для генов TEM, cfxA и OXA.

	TEM	cfxA	OXA
Число вершин	661	47 677	205 148
Число WGS ребер	388	48 871	209 234
Число Hi-C ребер с весом > 5	9184	154 156	136 418
Число компонент связности	347	106	330
Число отображенных вершин	58	20	187
Число отображенных WGS ребер	55	16	227
Число отображенных Hi-C ребер	13	8	14
Число отображенных компонент связности	14	5	9

На рисунке 4.5 приведен пример визуализации геномного контекста вокруг гена TEM (а), cfxA (б) и OXA (в). Для отрисовки данных контекстов использовался режим с автоматической подборкой параметров отрисовки Hi-C связей. Целевые компоненты покрашены в оранжевый цвет. Несмотря на различный размер геномных контекст, все визуализированные графы имеют небольшой размер, не загружены лишними вершинами и Hi-C ребрами, а также имеют малое число пересечений между ребрами. Данный эксперимент показывает работоспособность автоматического режима для визуализации

геномного контекста, построенного на реальных метагеномных данных различных размеров.

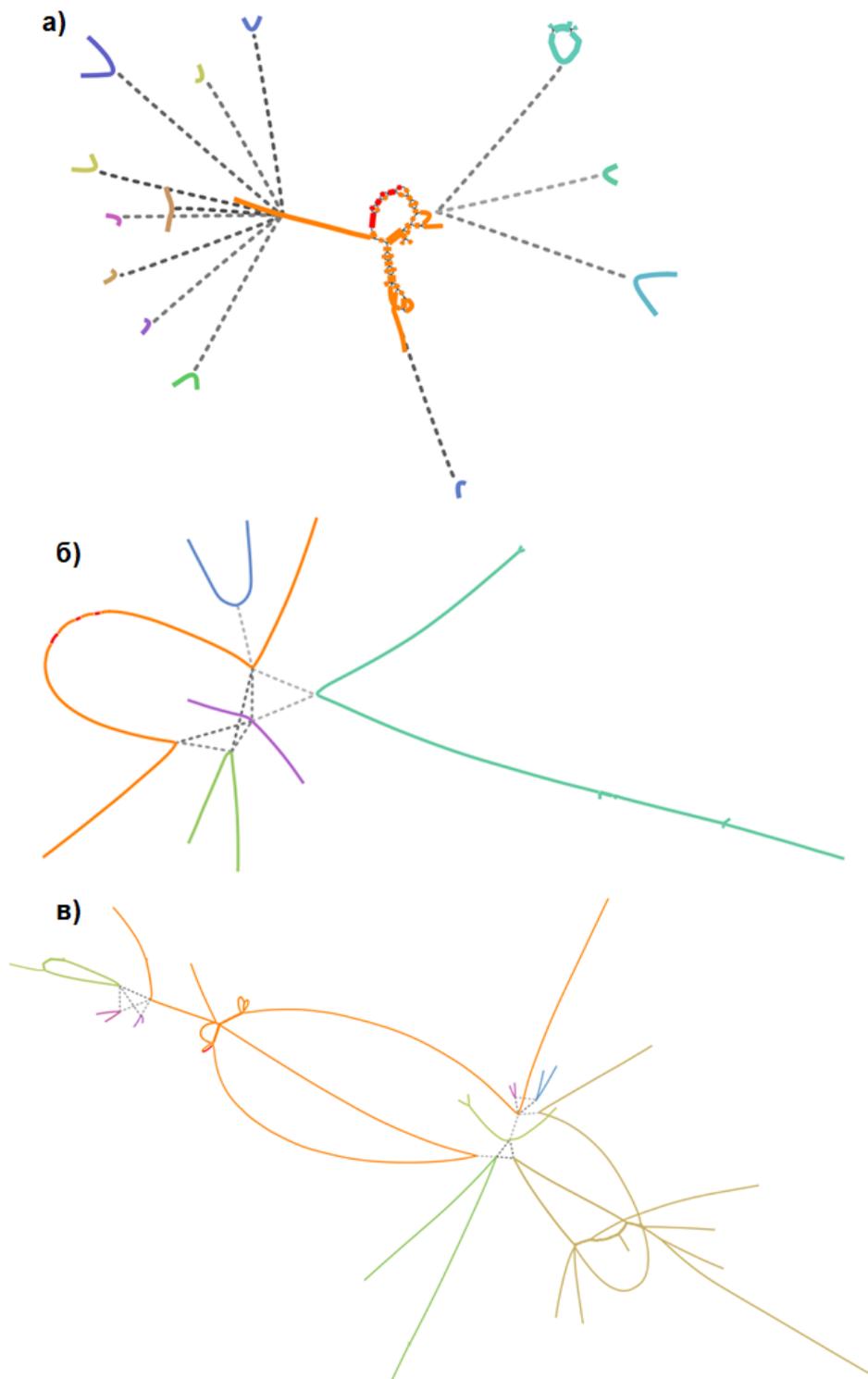


Рисунок 4.5. Визуализированный в автоматическом режиме геномный контекст вокруг гена TEM(a), *cfxA*(б) и *OXA*(в).

На рисунке 4.6 приведен пример визуализации геномного контекста вокруг гена *TEM* с учетом результатов таксономического анализа. Таксономический анализ был произведен при помощи программы Kraken 2. Полученная визуализация демонстрирует, что плазмида с геном *TEM* находится в клетках семейства бактерий *Enterobacteriaceae*, а именно: в клетках бактерии *Escherichia coli* и в клетках бактерии *Klebsiella pneumoniae*, что совпадает с данными, представленными в статье. Помимо этого, выявлены контиги, которые не были классифицированы программой Kraken 2, а потому интересны для дальнейшего изучения.

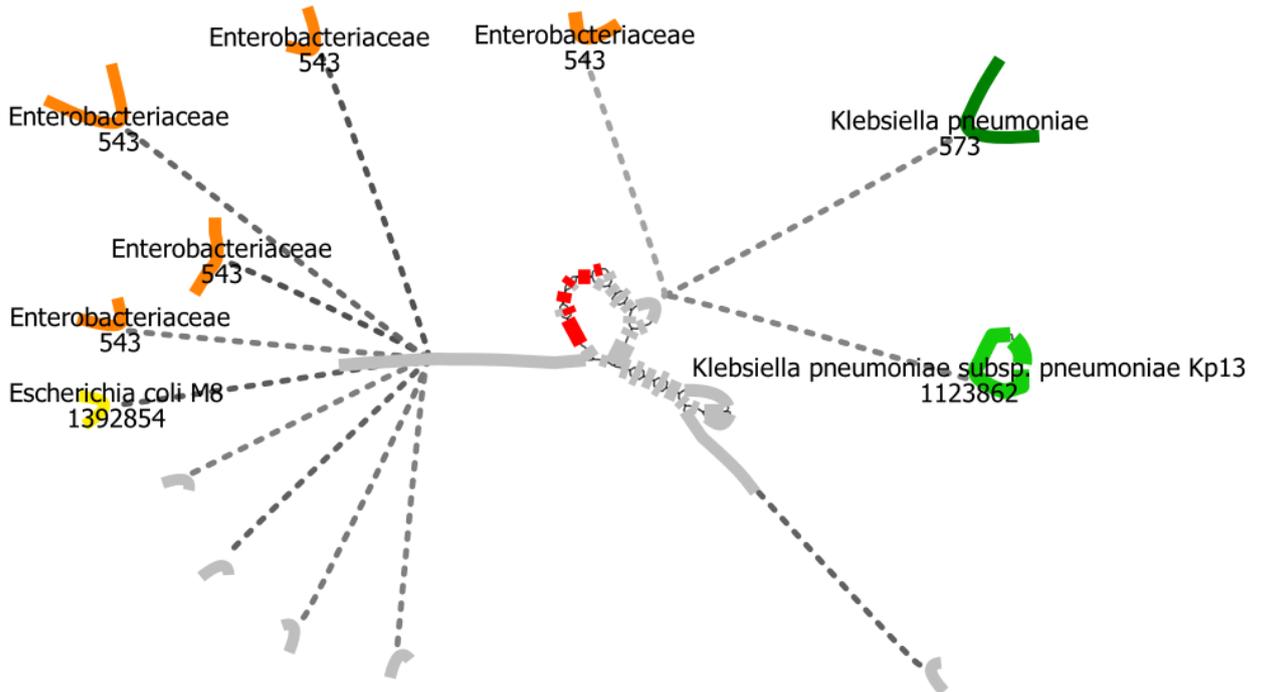


Рисунок 4.6 Визуализация геномного контекста вокруг гена *TEM* с использованием результатов таксономического анализа.

ВЫВОДЫ ПО ГЛАВЕ 4

В данной главе описано проведенное тестирование построения и визуализации геномного контекста с учетом Hi-C связей. Также была продемонстрирована возможность использования результатов таксономического анализа при визуализации графа де Брейна. Было проведено тестирование новой реализации приложения Bandage, как на сгенерированных данных, так и на реальных данных. В ходе тестирования было показано, что визуализированный геномный контекст соответствует ожиданиям. Также следует отметить, что полученные в ходе тестирования визуализации геномных контекстов подходят для дальнейшего ручного анализа пользователем, так как визуализированные графы имеют малый размер.

ЗАКЛЮЧЕНИЕ

В данной работе был проведен обзор методов построения и визуализации геномного контекста. Описаны их недостатки, связанные с получением информации о взаимосвязях мобильных элементов (например, плазмид) и их носителей. Поставлена цель совершенствования методов построения и визуализации геномного контекста.

В данной работе был реализован метод построения геномного контекста с учетом Hi-C связей. Также приложение Bandage было модифицировано для поддержания визуализации данных Hi-C секвенирования и результатов таксономического анализа. Полученные реализации позволяют находить бактерии, в клетках которых содержатся плазмиды с исследуемыми антибиотико-резистентными генами. Помимо этого, была реализована возможность выделения и визуализации геномов выбранных бактерий, плазмид и вирусов из графа де Брейна, содержащего несколько различных бактерий, на основе данных таксономического анализа. Также была реализована возможность получения информации о Hi-C связях между таксонами. Модификация приложения MetaCherchant и приложения Bandage была протестирована на сгенерированных и реальных данных. Было проверено, что отображенные Hi-C связи и отображенные данные о таксонах совпадают с данными Hi-C секвенирования и результатами таксономического анализа. Таким образом, все поставленные задачи были выполнены.

По результатам работы сделано выступление с докладом на XI Конгрессе молодых ученых в Университете ИТМО и 51-ой научной и учебно-методической конференции Университета ИТМО.

Также хочу выразить благодарность Иванову Артему Борисовичу, который консультировал меня в ходе данной работы в вопросах сбора и анализа метагеномных данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Nagarajan N., Pop M. Sequence assembly demystified //Nature Reviews Genetics. – 2013. – Т. 14. – №. 3. – С. 157-167.
2. Li Z. et al. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph //Briefings in functional genomics. – 2012. – Т. 11. – №. 1. – С. 25-37.
3. Li R. et al. De novo assembly of human genomes with massively parallel short read sequencing //Genome research. – 2010. – Т. 20. – №. 2. – С. 265-272.
4. Reuter J. A., Spacek D. V., Snyder M. P. High-throughput sequencing technologies //Molecular cell. – 2015. – Т. 58. – №. 4. – С. 586-597.
5. Lieberman-Aiden E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome //science. – 2009. – Т. 326. – №. 5950. – С. 289-293.
6. Olekhnovich E. I. et al. MetaCherchant: analyzing genomic context of antibiotic resistance genes in gut microbiota //Bioinformatics. – 2018. – Т. 34. – №. 3. – С. 434-444.
7. Wick R. R. et al. Bandage: interactive visualization of de novo genome assemblies //Bioinformatics. – 2015. – Т. 31. – №. 20. – С. 3350-3352.
8. Fruchterman T. M. J., Reingold E. M. Graph drawing by force-directed placement //Software: Practice and experience. – 1991. – Т. 21. – №. 11. – С. 1129-1164.
9. Walshaw C. et al. A multilevel algorithm for force-directed graph-drawing //Journal of graph algorithms and applications. – 2006. – Т. 7. – №. 3. – С. 253-285.
10. Gajer P., Kobourov S. G. GRIP: Graph drawing with intelligent placement //Journal of Graph Algorithms and Applications. – 2004. – Т. 6. – №. 3. – С. 203-224.
11. Hachul S., Jünger M. Drawing large graphs with a potential-field-based multilevel algorithm //International Symposium on Graph Drawing. – Springer, Berlin, Heidelberg, 2004. – С. 285-295.

12. Chimani M. et al. The Open Graph Drawing Framework (OGDF) //Handbook of graph drawing and visualization. – 2013. – Т. 2011. – С. 543-569.
13. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM //arXiv preprint arXiv:1303.3997. – 2013. — URL: <https://doi.org/10.48550/arXiv.1303.3997>(дата обращения: 26.01.2022)
14. Danecek P. et al. Twelve years of SAMtools and BCFtools //Gigascience. – 2021. – Т. 10. – №. 2. – С. giab008.
15. Wood D. E., Lu J., Langmead B. Improved metagenomic analysis with Kraken 2 //Genome biology. – 2019. – Т. 20. – №. 1. – С. 1-13.
16. Gourelé H. et al. Simulating Illumina metagenomic data with InSilicoSeq //Bioinformatics. – 2019. – Т. 35. – №. 3. – С. 521-522.
17. DeMaere M. Z., Darling A. E. Sim3C: simulation of Hi-C and Meta3C proximity ligation sequencing technologies //GigaScience. – 2018. – Т. 7. – №. 2. – С. gix103.
18. Ivanova V. et al. Hi-C Metagenomics in the ICU: Exploring Clinically Relevant Features of Gut Microbiome in Chronically Critically Ill Patients //Frontiers in microbiology. – 2022. – Т. 12. – С. 770323.