Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем
Системное программирование

Бзикадзе Александр Важевич

Применение Hi-C к метагеномике

Бакалаврская работа

Научный руководитель:
к.т.н., доц. Литвинов Ю.В.

Научный консультант:
к.ф.-м.н., доц. Коробейников А.И.

Рецензент:
Научный сотрудник Центра алгоритмической биотехнологии СПбГУ
Пржибельский А.Д.

Санкт-Петербург
2019

SAINT-PETERSBURG STATE UNIVERSITY

Software and Administration of Information Systems
Software Engineering

Aleksandr Bzikadze

Application of Hi-C to metagenomics

Bachelor's Thesis

Scientific supervisor:
C.Sc., Associate Professor Yurii Litvinov

Scientific consultant:
Ph.D., Associate Professor Anton Korobeynikov

Reviewer:
Research Fellow at the Center for Algorithmic Biotechnology SPbU
Andrey Prjibelski

Saint-Petersburg
2019

Оглавление

Введение

Биоинформатика – одна из наиболее быстро развивающихся областей науки на текущий момент, объединяющая достижения биологии и Computer Science. Одной из важнейших проблем биоинформатики является сборка генома. Первым частично собранным геномом стал геном человека. Проект «Геном человека» (англ. *The Human Genome Project*) [19] продолжался на протяжении более десяти лет, и на его осуществление было израсходовано более трех миллиардов долларов. Современное научное сообщество считает достижение успехом, продвинувшим область далеко вперед, привлекшим колоссальные инвестиции и оказавшим огромное влияние на биоинформатику.

Особый интерес представляет изучение геномов микроорганизмов. Несмотря на малые размеры, данные организмы зачастую обладают невероятными по современным меркам биологическими инструментами, некоторые из которых адаптированы человеком в своих целях. Так ретровирусы обладают возможностью встраивать гены в клетки организма-носителя и таким образом самовоспроизводиться, что было использовано учеными для развития генотерапии [4]. Некоторые бактерии обладают адаптивной имунной системой [29], основанной на встраивании в собственный геном участков геномов враждебных вирусов или иных микроорганизмов для их идентификации и последующего уничтожения. На основе этого учеными были разработаны средства для редактирования генома [5], которые в частности планируется использовать для борьбы с раком [7] и ВИЧ-инфекцией [6] и создания генно-модифицированных продуктов без добавления чужеродной ДНК [11].

Однако наличие повторных участков в геноме усложняет его восстановление, добавляя неоднозначность, из-за чего требуются новые или вспомогательные методы. Одной из таких технологий является технология Hi-C, позволяющая исследовать пространственное строение генома, что может быть применено к восстановлению его линейного строения.

1. Постановка задачи

Требуется найти применение Hi-C данных к разрешению повторов в графе сборки микробиома, разработав расширение геномного ассемблера SPAdes [28] поддержкой Hi-C. Таким образом, планируется решить следующие подзадачи:

- провести анализ физического устройства Hi-C протокола и его биологических свойств;
- исследовать свойства изучаемых данных, типичные проблемы и ограничения;
- разработать прототип разрешения повторов на основе Hi-C информации;
- разработать расширение геномного ассемблера SPAdes.

2. Обзор предметной области

Геном

Геном – это совокупность наследственной информации, обычно представленной в виде молекул дезоксирибонуклеиновой кислоты (ДНК). Молекула ДНК представляет собой сдвоенную закрученную спираль из двух полимеров (стрендов) [34], состоящих из более простых молекул, называемых нуклеотидами: аденин (A), цитозин (C), гуанин (G), тимин (T). Химическая структура стренда позволяет определить его направление, что дает представление о структуре расположения стрендов в ДНК: нуклеотиды располагаются друг напротив друга, а направления стрендов – строго противоположные. Помимо этого для каждого нуклеотида известна его пара: аденин комплементарен тимину, цитозин – гуанину. Длина участков ДНК измеряется в спаренных основаниях (пара, англ. *base pair*, **bp**).

Таким образом, молекулы ДНК представляют собой две строки четырехбуквенного алфавита, реверс-комплементарные друг другу.

Молекулы ДНК называются хромосомами. Стоит отметить, что количество хромосом в клетках варьируется в зависимости от организма. Например, человеческий организм представлен двадцатью тремя парами хромосом, в то же время существуют бактерии, геном которых состоит из одной единственной хромосомы.

Проекты по определению генома вида или даже конкретного организма представляют огромный научный и практический интерес в таких областях, как персонализированная медицина [3], синтезирование антибиотиков, изучение наследственной информации и межвидовых связей, а также генная инженерия.

Сборка генома

Современные технологии не позволяют считывать последовательности ДНК напрямую. В связи с этим были разработаны подходы, преодолевающие данную проблему за счет генерирования (секвенирования) множества сравнительно коротких фрагментов генома, после чего возникает задача сборки первоначальной последовательности из предоставленных фрагментов. Задача осложнена тем, что секвенирование сопряжено с физическими ошибками оборудования. Наиболее популярный на текущий момент времени метод секвенирования – секвенирование следующего поколения (англ. *Next Generation Sequencing technologies*, **NGS**) [2]. Данный метод существенно снизил материальные затраты на восстановление генома, но обладает значительными ограничениями на длину считываемых последовательностей. Наиболее распространенные методы предоставляют фрагменты длины 100-200 нуклеотидов. В дополнение, существуют методы, позволяющие секвенировать фрагменты (риды, англ. *reads*) как пары, для которых известно расстояние, на котором они расположены в обрабатываемой последовательности ДНК.

Для ассемблирования генома из полученных NGS данных применяется граф де Брюина [26]. Если известны все подстроки фиксированной длины k (k-меры) исходного генома с учетом их кратности, то геном представляется одим из Эйлеровых путей в мультиграфе, ребрами которого являются k-меры, соединяющие уникально размеченные вершины, маркирующие $(k-1)$-меры. В связи с физическими ошибками и наличием повторов в геноме Эйлеров путь в полученном графе часто неуникальный.

В действительности невозможно гарантировать наличие всех подстрок генома фиксированной длины, но идеи графа де Брюина находят свое применение для построения графа сборки, ребра которого являются восстановленными последовательностями нуклеотидов, а некоторые пути – исходными геномными последовательностями. Так возникает задача разрешения повторов в графе сборки.

Технология Hi-C

Технология Hi-C [9] является частью семейства протоколов фиксации форм хромосом (англ. *Chromosome Conformation Capture*, **3C**) [35], направленных на изучение пространственного строения ДНК. В отличие от других, данный протокол не связан с конкретным участком генома и применяется по всей протяженности ДНК, а не только к изначальным или близким в пространстве участкам.

Заключительным этапом процедуры является парное секвенирование. Таким образом, результатом работы является библиотека парных ридов, полученных с помощью Hi-C протокола.

Применение Hi-C данных является перспективной задачей. Это сравнительно новая технология с потенциально широким спектром приложения. В силу специфики технологии в ней отсутствует ограничение на линейное расстояние в геноме между фрагментами, секвенированными как пара (расстояние вставки, англ. *insert size*), что представляет большой интерес для разрешения длинных повторов, неразрешимых с помощью обычных парных чтений.

Hi-C и метагеномика

Особый интерес вызывает применение Hi-C к метагеномике. Геномы микроорганизмов часто оказываются проще по строению, что облегчает процесс их изучения и сборки. В то же время подавляющее большинство микроорганизмов сосуществуют в постоянном взаимодействии с другими, отличными от себя видами, что делает невозможным секвенирование каждого из таких видов по отдельности. Задача, являющаяся основным предметом данного исследования – задача скаффолдинга, заключающаяся в определении верного порядка следования в геноме восстановленных участков. Объединенные в порядке следования восстановленные последовательности называются скаффолдами.

3. Обзор Hi-C протокола

Для генерации Hi-C данных предоставляется набор клеток, обрабатываемый специальным образом, после чего передаваемый для парного секвенирования. Далее приводится краткий обзор того, как именно происходит данный процесс, лишенный подробностей проведения самой процедуры генерации Hi-C данных.

Генерация Hi-C данных состоит из нескольких стадий (Рис. 1).

- Создание поперечных связей (англ. *crosslink*). Внутри входных клеток создаются поперечные связи, что является общей чертой всех 3C методов. Во время соединения клетки остаются неразрушенными, что должно предотвратить появление случайных связей между разными микроорганизмами. Результатом являются последовательности ДНК, некоторые сегменты которых оказываются соединены.

- Клетки разрушаются, освобождая ДНК, после чего она измельчается за счет рестриктаз (англ. *restriction enzymes*), задачей которых является разделение ДНК на два фрагмента при нахождении определенной последовательности нуклеотидов. Оригинальные рестриктазы, задействованные в Hi-C протоколе – «HindIII», реагируют на последовательность из шести нуклеотидов. Последние модификации позволили также применять «MluCI», «Sau3aI», чьи последовательности составляют четыре нуклеотида в длину, что значительно повышает вероятность их нахождения, и, следовательно, разреза. Таким образом соединенные фрагменты оказываются отделены от общей массы ДНК.

- Места разрезов восстанавливаются, оказываясь помечены специальным биотином, после чего соединяются обратно. Во время обратного соединения возможны ошибки и присоединения фрагментов не к своей паре, а к себе или чужеродному участку.

- После соединения соединенные фрагменты разворачиваются, очищаются от первоначальных соединительных материалов и притяги-

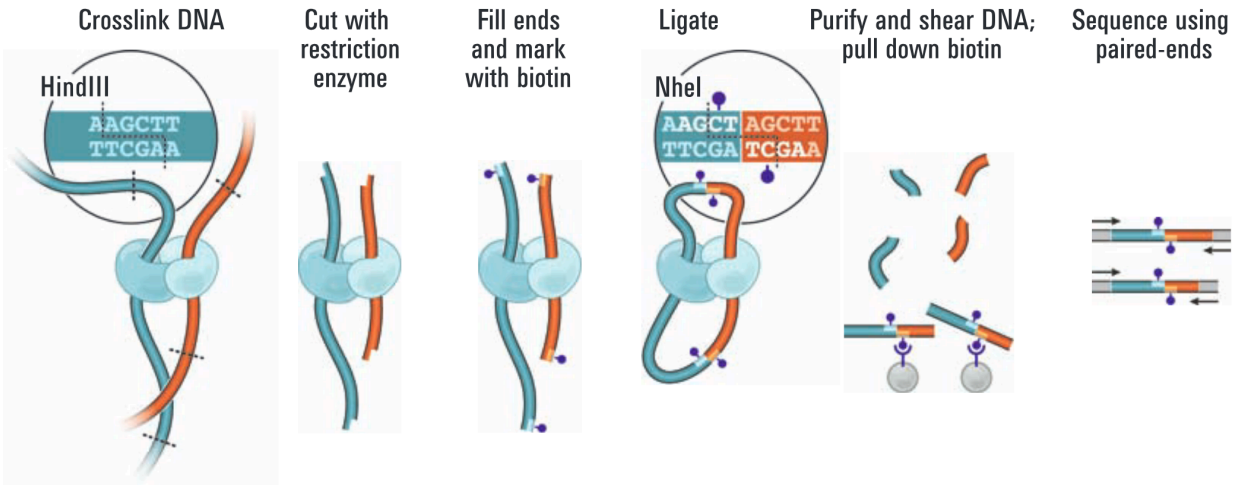


Рис. 1: Устройство Hi-C протокола [9]

ваются за биотин, вставленный во время восстановления разрезов. Таким образом, два различных фрагмента ДНК, соединенные на первом шаге, оказываются присоединены друг к другу последовательно, помечены биотином и притянуты для дальнейшего изучения.

- Полученные соединения сокращаются до размеров, пригодных для парного секвенирования, после чего передаются секвенатору.

Секвенирование естественного микробиома сопряжено с наличием в нем мертвых организмов, что осложняет интерпретацию результатов, так как в зависимости от качества исходного биома и проведения процедуры в Hi-C данных будут присутствовать ложные связи, отделение которых является сложноразрешимой задачей на вычислительном уровне.

Очевидно, что Hi-C данные зависят от выбора рестриктазы. Стандартной рекомендацией считается использование двух Hi-C библиотек, созданных при помощи различных рестриктаз, при игнорировании чего повышается риск отсутствия Hi-C информации в некоторых относительно длинных фрагментах ДНК.

4. Обзор инструментов

Далее приведен обзор доступных инструментов, связанных с Hi-C и задачей скаффолдинга.

4.1. Средства выравнивания Hi-C

Для поиска сходства между геномными последовательностями используется выравнивание, которое можно неформально характеризовать как поиск неточной подстроки. В случае парных ридов результатом выравнивания является пара выравниваний концов рида.

Для изучения Hi-C данных существуют средства, позволяющие выравнивать входные Hi-C риды на известный референсный геном, фильтровать их и реорганизовывать в удобном виде. Референсный геном – это известное человечеству объединение геномов организмов одного вида, не обязательно соответствующее некому конкретному живому организму, но достаточно близкий ко всем представителям данного вида. Инструмент такого типа требуется для анализа свойств Hi-C данных.

Одним из таких средств является Arima Hi-C Mapping Pipeline [1]. Данный инструмент использует bwa mem [23] для выравнивания Hi-C ридов на референсный геном, самостоятельно фильтрует их, после чего задействует Picard Tools [20] для отделения PCR дубликатов [25] – дубликатов специального вида, возникающих из-за особенностей методов NGS. Помимо прочего данное средство анализа предоставляет краткую статистику по обработанным Hi-C связям.

Другим инструментом, решающим ту же задачу, является distiller-nf [22]. Данное средство расширяет функциональность Arima Hi-C Mapping Pipeline, реорганизуя данные о межхромосомных связях в pairix архивы [10] и контактные карты. Контактными картами в контексте Hi-C технологии называется способ представления распределения Hi-C связей по геному в виде двумерного изображения, каждый пиксель которого соответствует паре участков генома и окрашен в цвет, иллюстрирующий

количество связей между выбранной парой участков. Для дальнейшего анализа Hi-C данных был выбран именно этот инструмент.

4.2. Симулятор Hi-C данных

Существует симулятор Hi-C библиотек на основе предоставленного референсного генома sim3C [13]. Он позволяет симулировать в том числе микробиомы, задавая соотношение количества микроорганизмов и указывая используемую рестриктазу. Распределение расстояния вставки симулированных Hi-C ридов представляет собой смесь равномерного и геометрических распределений. Данный инструмент требуется для симуляции данных, на которых можно проводить эксперименты в условиях достоверно известного референсного генома, то есть с заранее известным правильным ответом.

4.3. Средства скаффолдинга

Одним из ранних иструментов скаффолдинга является Lachesis [8], принцип работы которого заключается в следующем:

- строится полный взвешенный граф, вершинами которого являются восстановленные участки генома;
- исключая предположительные повторные участки и короткие, длина которых настраивается вручную в зависимости от изучаемого организма, проводится иерархическая кластеризация по количеству Hi-C связей до тех пор, пока число кластеров не совпадет с ожидаемым;
- внутри кластеров строится минимальное остовное дерево, где весом ребер выступает величина, обратная количеству Hi-C связей и нормированная количеством искомых рестриктазами участков;
- для каждого дерева выбирается самый длинный путь, после чего оставшиеся ребра пытаются добавить в него, максимизируя количество связей между соседними вершинами.

Главный недостаток данного метода заключается в требовании предварительного знания об исследуемых данных, и невозможность работы одновременно с несколькими организмами, представленными в образце неравномерно, то есть он непригоден для работы с метагеномами.

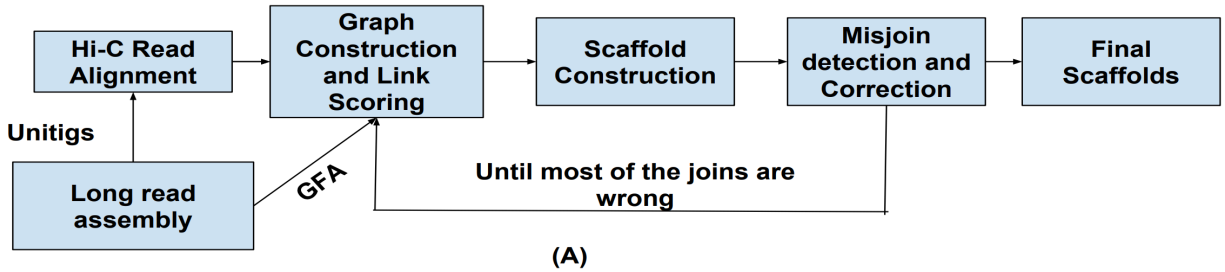


Рис. 2: Алгоритм работы Salsa2 [21]

Другим инструментом скаффолдинга на основе Hi-C является Salsa2 [21], работающий по следующему алгоритму (Рис. 2):

- с помощью графа сборки и Hi-C данных генерируется взвешенный граф, вершинами которого являются начала и концы ребер исходного графа сборки, а ребра задаются на основе графа сборки и Hi-C данных;
- в данном графе по приближенному жадному алгоритму выбирается максимальное взвешенное паросочетание;
- выбранное паросочетание служит основой для составления скаффолда, и если большая часть его соединений подтверждается с помощью Hi-C, то процесс генерации продолжается.

Данный инструмент является лучшим из существующих на данный момент. Помимо перечисленных существуют скаффолдеры 3D-DNA [12], instaGRAAL [17], результаты которых хуже чем результаты Salsa2 [21].

4.4. Средства биннинга

Не менее важной задачей метагеномики является задача биннинга, заключающаяся в кластеризации восстановленых последовательностей ге-

нома по принадлежности к одному микроорганизму. Биннинг используется учеными для исследования естественных данных на присутствие известных или новых микроорганизмов в изучаемых образцах, возможно опуская полное восстановление их геномов.

Задачи скаффолдинга и биннинга тесно связаны. Так, кластеризация скаффолдов упрощается в связи с повышением количества доступной информации. В то же время скаффолдинг внутри кластеров обладает большими шансами на успешность.

Одним из инструментов биннинга является Metaphase [30], использованный авторами для успешного исследования бактерий и дрожжей, содержащихся в изучаемом ими образце бельгийского ламбика [18]. Он обладает визуализацией кластеризуемого графа, но в открытом доступе отсутствует и полная версия исходного кода, доступная для тестирования, и уже скомпилированная бинарная версия.

Другой инструмент – bin3C [14] – был разработан авторами симулятора Hi-C данных sim3C и успешно работает на симулированных данных.

5. Исследование Hi-C данных

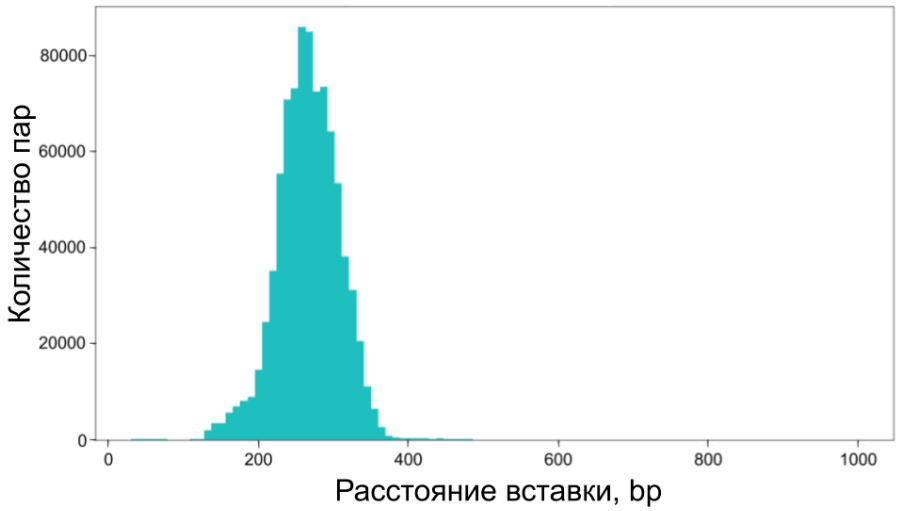
Стоит отметить, что выбор публично доступных данных ограничен в связи с новизной технологии. Для исследования свойств Hi-C данных были взяты три набора данных.

- Beitel [31]. Синтетический датасет (искусственно выращенный) из пяти хорошо изученных микроорганизмов с известными референсными геномами высокого качества. Двое из представленных организмов являются разными штаммами кишечной палочки, другой состоит из двух хромосом, а также присутствует бактерия, содержащая помимо хромосомы две плазмиды. Плазмида – это небольшая молекула ДНК, отделенная от хромосом и способная к самостоятельной репликации, нередко сменяющая носителей в ходе своего существования.

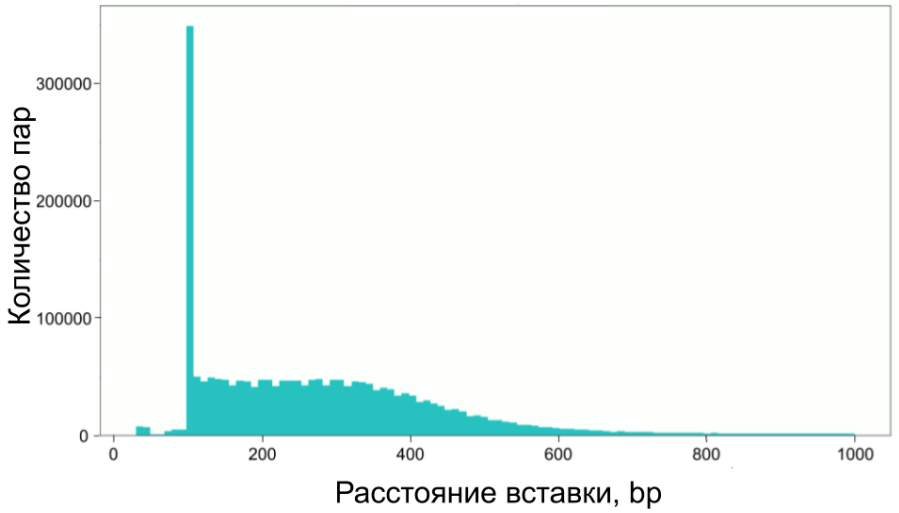
- Beer [18]. Данный естественный набор данных, состоящий предположительно из восьми бактерий и дрожжей, был получен из бельгийского ламбика. Данный датасет представляет интерес из-за естественный среды, так как обилие пищи предоставляет возможность микроорганизмам усиленно размножаться, изменяя соотношение мертвых и живых организмов в биоме.

- Fecal [16]. Естественный датасет, полученный из стула здорового человека. В сравнении с предыдущими, данный датасет много больше и содержит предположительно более пятидесяти различных микроорганизмов, а также использует рестриктазы «MluCI», «Sau3aI» вместо «HindIII».

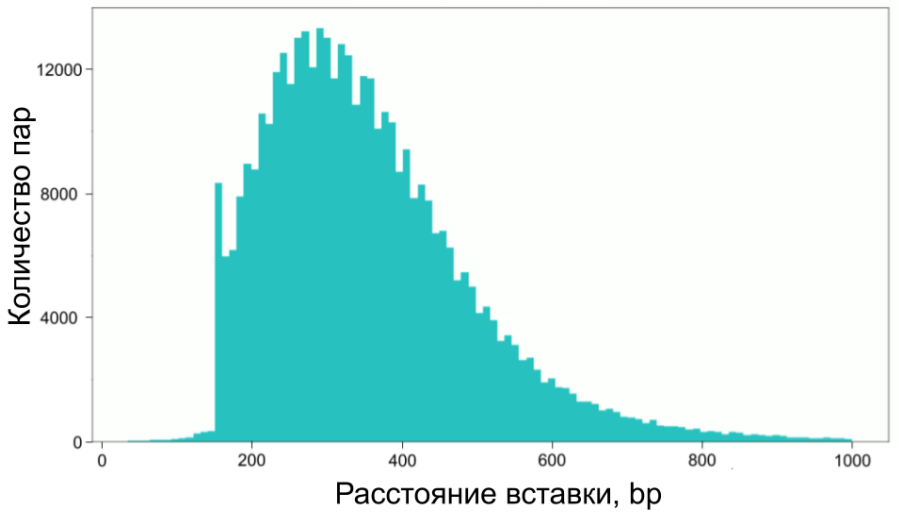
Для каждого естественного набора данных было выбрано несколько референсных геномов, обладающих высоким качеством, для дальнейшего выравнивания Hi-C библиотеки на них. Главное свойство, вызывающее интерес – распределение расстояния вставки как основное качество данных, позволяющее разрешать длинные повторы (Рис. 3). В результате было отмечено, что большая часть пар обладает расстоянием вставки менее 1Kbp.



(a) Распределение расстояния вставки 88% пар в Beitel



(b) Распределение расстояния вставки 55% пар в Beer



(c) Распределение расстояния вставки 96% пар в Fecal

Рис. 3: Распределения расстояния вставки, ограниченное 1Kbp

Рассматривая распределение расстояния вставки и исключая первую тысячу нуклеотидов, следует отметить, что с ростом расстояния количество соединений снижается. Начиная с десяти тысячи нуклеотидов падение количества соединений замедляется и остается ненулевым на протяжении всего генома.

Таким образом, количество Hi-C связей имеет явную зависимость от длин рассматриваемых участков генома, потому простое количество не может корректно характеризовать расстояние между различными участками. Естественным подходом является нормирование Hi-C информации на длину рассматриваемых участков, но Hi-C связи распределены по геному неравномерно, что особенно влияет при рассмотрении малых по длине участков.

Известно, что формирование Hi-C связей происходит вблизи участков, разрезаемых рестриктазами. Следовательно, можно нормировать на количество таких участков в рассматриваемых участках генома, но это требует большой уверенности в их качестве, так как в связи с малой длиной (4-6 нуклеотидов) искомых фрагментов нельзя искать их неточно, а даже один измененный нуклеотид приведет к провалу его нахождения. В это же время нормирование малых по длине участков, содержащих малое количество искомых сегментов, крайне чувствительно к изменению нормирующего значения.

Другим подходом является использование отсечки по длине, то есть, рассматривая два геномных участка, следует учитывать лишь связи, расположенные на крайних, например, пяти тысячах нуклеотидов. Также необходимо учитывать наличие четырех вариантов выбора концов геномных участков. Данный метод был избран для учета количества связей в дальнейших вычислениях.

6. Прототип

На основе полученных результатов был разработан прототип решения задачи скаффолдинга на основе Hi-C данных. В связи с трудностью прикладывания парных ридов к коротким ребрам, первым шагом граф сборки сжимается, после чего Hi-C риды требуется выравнить на получившиеся длинные ребра. Помимо этого для коротких ребер велика вероятность отсутствия участков, на которые реагируют рестриктазы, и как следствие, отсутствие Hi-C связей, что делает разрешение повторов на основе Hi-C данных невозможным для данного ребра. Затем, начиная с каждого ребра, алгоритм пытается нарастить путь.

6.1. Подготовка данных

Исходными данными являются библиотеки обычных парных ридов и библиотеки Hi-C. Первые собираются с помощью геномного ассемблера SPAdes в метагеномном режиме (опция --meta), результатом чего является граф сборки. После этого Hi-C риды выравниваются на граф сборки и после преобразования информации записываются как список: {ребро}-{ребро}-{количество Hi-C связей}. Данный процесс является неоптимизированным и временным, так как для разработки и тестирования прототипа предложенные действия требуется выполнить единожды для каждого датасета, а впоследствии станут излишними при интеграции в SPAdes.

6.2. Порог длины на сжатие ребер

В связи с последними модификациями Hi-C протокола, позволяющими использовать рестриктазы, реагирующие на последовательности длины четыре нуклеотида вместо шести, частота мест, потенциально способных быть соединенными Hi-C связями, увеличивается в шестнадцать раз. Если считать геном случайной строкой, каждая буква которой распределена независимо и равномерно на четырехбуквенном алфавите, то фиксированная последовательность из четырех нуклеотидов встречается в среднем единожды на 4^4 нуклеотидов. Стоит отметить, что геномы далеко не случайны, потому в качестве ограничения на длину ребер было взя-

то пятьсот нуклеотидов, примерно в два раза превосходящее ожидаемое расстояние между двумя соседними по геному встречами последовательностей, на которые реагируют рестриктазы.

6.3. Сжатие графа

Для удаления коротких ребер и сохранения информации о графе применяется метод сжатия графа, результатом которого является сжатый граф сборки. По заданному значению K все ребра, длины которых не превосходят его, должны быть сжаты (Рис. 4). Введем отношение R наличия короткого ребра на вершинах, то есть $\forall(v,u) \Rightarrow vRu$, где (v,u) – короткое ребро. Вершины сжатого графа представляют из себя классы эквивалентности вершин исходного графа, где отношение эквивалентности – транзитивное замыкание R, а ребра задаются на основе ребер исходного графа сборки.

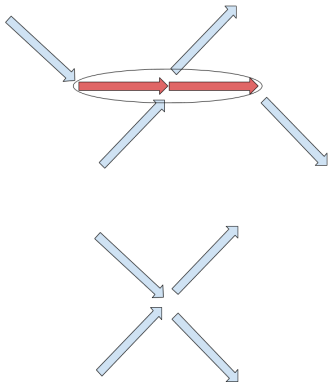


Рис. 4: Сжатие двух ребер

Для сжатия графа сборки следует представить вершины как систему дизъюнктных одноэлементных множеств, которые следует объединять при обнаружении ребра длины меньше заданного K. В результате почти за линейное время работы от количества ребер в графе – в среднем $O(V\alpha(V))$, где $\alpha(n)$ является обратной функцией Аккермана [32] – образуются метавершины, представляющие собой множества вершин исходного графа, подлежащие объединению. На основе исходного графа сборки и полученных метавершин осуществляется построение нового графа.

6.4. Наращивание пути

Рассмотрим процесс генерации путей. Каждое ребро рассматриваем как начало потенциального пути. Алгоритм пытается увеличить выбранный путь, применяя далее рассматриваемый критерий наращивания пути. Если такая возможность отсутствует, путь расширяется в обратную сторону аналогичным способом.

В текущей реализации существует вероятность зацикливания пути. На экспериментальных данных возникающие циклы не превышали по длине два ребра. Проблема была решена для данного случая, так как существует множество способов борьбы с данной проблемой, уже реализованных в SPAdes, позволяя использовать их при дальнейшей интеграции.

6.5. Критерий наращивания пути

Для продолжения текущего пути была рассмотрена композиция простейших критериев. Рассмотрим очередное ребро e в сжатом графе, являющееся концом текущего пути.

$weight(e, e')$ – количество Hi-C пар, выравненных на ребра e, e' таким образом, что один из ридов лежит в последних четырех тысячах нуклеотидах ребра e, а второй – в первых четырех тысячах нуклеотидов ребра e'. В качестве предполагаемого продолжения пути выбираем $e_m = \underset{e' \in out(e) \wedge e' \neq e}{\arg\max} weight(e, e')$.

- Проверяем, что $weight(e, e_m) > 1$ – абсолютный критерий.
- Проверяем, что $\forall e'(e' \in out(e) \wedge e' \notin \{e, e_m\} \rightarrow \frac{weight(e,e')}{weight(e,e_m)} \leq T)$ – относительный критерий, где T – входной параметр.
- Проверяем, что $\forall e'(e' \in in(e_m) \rightarrow weight(e', e_m) \leq weight(e, e_m))$ – обратный критерий.

Если ребро e_m удовлетворяет всем условиям, оно выбирается как продолжение ребра e и текущего пути, становясь его новым концом.

7. Эксперименты над прототипом

Для оценки результатов работы прототипа был осуществлен его запуск на двух наборах данных. Параметры запусков: $K = 500,\ T = 0.8$, то есть рассматриваются ребра графа сборки длины 500 нуклеотидов и более, и относительный критерий наращивания пути требует превосходства предполагаемого продолжения над остальными кандидатами по количеству Hi-C связей в $\frac{5}{4}$ раз.

7.1. Эксперименты на симулированных данных

Для тестирования прототипа скаффолдера был использован публично доступный симулированный с помощью sim3C датасет [14] из референсных геномов высокого качества шестидесяти трех бактерий. По аналогии с датасетом Fecal данный датасет был просимулирован с использованием рестриктаз «MluCI» и «Sau3aI». После сборки и сжатия граф сборки содержал более девяноста шести тысяч конфликтов, то есть вершин сжатого графа, количество исходящих ребер из которых превышает один, что означает отсутствие однозначного продолжения.

48.5% конфликтов было разрешено, из которых 97% разрешено корректно. Конфликт определяется разрешенным корректно, если предложенное продолжение является ближайшим в геноме по линейному расстоянию, либо расстояние до него не превышает две тысячи нуклеотидов.

Основная причина неразрешения конфликтов (77% случаев) – неудовлетворение абсолютному критерию, то есть недостаток Hi-C информации, что является естественным ограничением на применение метода. В качестве возможного пути обхода данного ограничения следует рассматривать информацию не только с конца пути, но со всего, возможно, с помощью весовой функции.

7.2. Эксперименты на естественных данных

Для тестирования на естественных данных был взят описанный выше датасет Fecal. Он состоит из двух Hi-C библиотек, созданных с помощью рестриктаз «MluCI» и «Sau3aI», которые реагируют на последовательности из четырех нуклеотидов. Были выбраны три референсных генома высокого качества, для которых получены свидетельства присутствия соответствующим им микроорганизмов в исследуемом биоме с использованием mash screen [24].

После сжатия граф содержал более 463 тысяч конфликтов, из которых 19% были разрешены. Из успешно приложенных разрешены корректно были 88%.

Как и в случае с симулированным микробиомом, большинство неразрешенных конфликтов отвергнуты по причине недостатка Hi-C информации (71%).

8. Расширение SPAdes

Входными данными пайплайна SPAdes являются обыкновенные парные риды и поддерживаемая дополнительная информация, к которой относятся Hi-C данные. Скаффолдинг осуществляется одной из завершающих стадий после ассемблирования генома и получения графа сборки. Данная стадия пайплайна несет название ExSPAnder [15].

8.1. Архитектура модуля ExSPAnder

Результирующие скаффолды строятся с помощью путей в графе сборки. ExSPAnder состоит из трех основных этапов: рост путей, коррекция путей и полировка путей (Рис. 5).

Основным элементом этапа роста путей является интерфейс PathExtender, позволяющий выбрать продолжение для переданного пути. Для объединения работы нескольких реализаций данного интерфейса использует CompositeExtender, запускающий хранимые внутри него экземпляры PathExtender в заданном порядке и принимающий за ответ первый успешный результат.

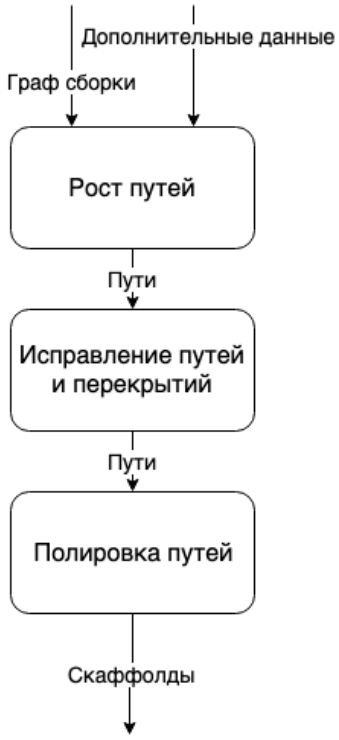


Рис. 5: ExSPAnder

Простой реализацией интерфейса PathExtender является класс SimpleExtender, содержащий в себе экземпляр интерфейса ExtensionChooser, который по переданному пути возвращает возможные продолжения данного пути. Если количество возвращаемых кандидатов равно единице, то при прохождении дополнительных проверок, зависящих от конфигурации пайплайна SPAdes, единственный кандидат добавляется к текущему пути.

Коррекция путей заключается в поиске пересечений путей для их последующего дробления, таким образом исправляя возможные ошибки на

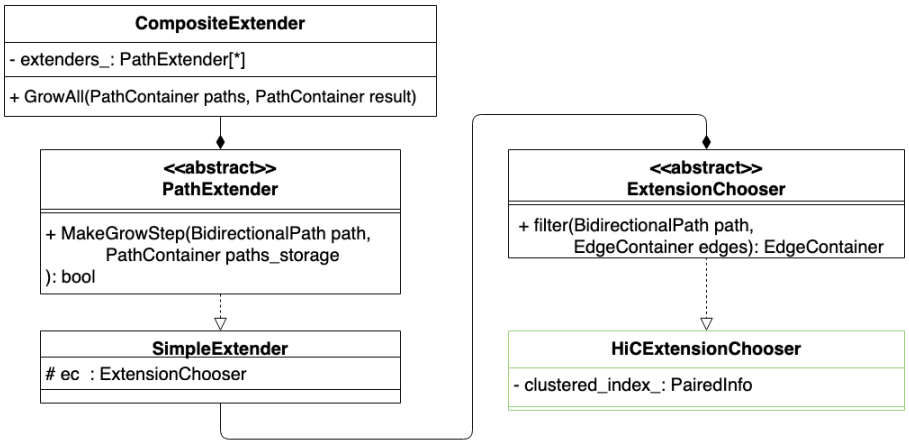


Рис. 6: Фрагмент архитектуры роста путей с поддержкой Hi-C

предыдущей подстадии. Следует отметить, что для верно соединенного пути наличие другого, ошибочно собранного пути, пересекающегося с первым, означает дробление обоих путей.

Возможна ситуация, когда для заданного пути возможно определить продолжение, расположенное на некотором малом растоянии в геноме. В таком случае путь расширяется этим продолжением с учетом дополнительной информации о наличии пробела. Одной из задач полировки пути является заполнение данных пробелов на основе имеющейся информации.

8.2. HiCExtensionChooser

В качестве первоначального подхода поддержки Hi-C данных был создан HiCExtensionChooser, реализация ExtensionChooser (Рис. 6). Результатом его работы является список кандидатов на продолжение – пустой, если продолжение не найдено, и содержащий единственный элемент в обратном случае. Содержащий HiCExtensionChooser экземпляр SimpleExtender добавляется в CompositeExtender в последнюю очередь, таким образом предполагая его наименьшую надежность.

Заметим, что в отличие от прототипа здесь невозможно игнорировать присутствие коротких ребер в пути. Тем не менее в связи с малым доверием выравнивания Hi-C ридов на короткие ребра их следует исключить из рассмотрения. Короткими ребрами в подсекции 6.2 было принято счи-

тать ребра короче пятисот нуклеотидов. Следует отметить, что последнее ребро в продолжаемом пути не всегда бывает достаточной длины, потому первым этапом работы HiCExtensionChooser является поиск последнего длинного ребра в пути. Данное ребро назовем опорным ребром.

Для формирования начального списка кандидатов начиная с опорного ребра запускается оптимизированный алгоритм Дейкстры, не запускающий обход далее длинных ребер, то есть при нахождении длинного ребра оно помещается в результирующий список, минуя приоритетную очередь, используемую для работы алгоритма Дейсктры. Помимо прочего на данном этапе оценивается предположительное расстояние до искомых кандидатов.

В связи с наличием значительной доли парных ридов, расстояние вставки которых крайне мало, количество связей ребер самих с собой многократно превышает все остальные. В концепции расширения пути новым ребром принятие ребра, уже присутствующего в пути, вероятнее всего приведет к образованию ошибочного цикла, потому такие ребра следует исключить из рассмотрения.

Для нормализации количества Hi-C связей в секции 5 было принято решение использовать абсолютную отсечку по длине, на которой Hi-C связи будут учитываться. Таким образом рассматриваются связи с последних пяти тысяч нуклеотидов пути на первые пять тысяч нуклеотидов ребра, рассматриваемого как кандидат.

В качестве критериев выбора используются разработанные для прототипа критерии. В первую очередь выбирается максимум по Hi-C связям, после чего проверяется, что количество Hi-C связей с ним превышает один и превосходит остальных кандидатов в $\frac{5}{4}$ раз. Такой кандидат называется вероятным продолжением.

Так как вероятное продолжение должно быть продолжением текущего пути, мы удостоверяемся, что оно достижимо по коротким ребрам с конца

пути, после чего запускается обратный критерий.

Algorithm 1 HiCExtensionChooser

```
1: procedure FILTER(path)
       ▷ Последнее длинное ребро в пути
2:     pivot ← FindLastLong(path)
       ▷ Запуск алгоритма Дейкстры
3:     edges_within_range ← FindCandidates(pivot)
       ▷ Фильтрация кандидатов по пересечению с path
4:     filtered ← FilterCandidates(edges_within_range, path)
       ▷ Выбор максимума по Hi-C с учетом абсолютного и относительных
   критериев
5:     answer ← ChooseByHiC(path, filtered)
       ▷ Проверка достижимости вероятного продолжения с конца пути
6:     Verify: CheckConsistencyWithPath(path, answer)
7:     Verify: CheckBackHiC(path, answer)
8:     return answer
9: end procedure
```

В случае отсутствия опорного ребра, вероятного прохождения, пустоты множества кандидатов до или после фильтрации, непрохождении обратного критерия, результат работы функции (Алг. 1) – отрицательный.

Результаты тестирования HiCExtensionChooser оказались неудовлетворительными. Во-первых, недостаток Hi-C информации часто приводит к неоднозначности и невозможности выбрать продолжение. Во-вторых, недостаток Hi-C информации допускает неверные соединения, приводящие к дроблению уже существующей сборки последующим этапом работы ExSPAnder.

8.3. HiCPathToPathExtensionChooser

Для решения данных проблем было положено отойти от концепции роста путей и применить Hi-C данные уже после формирования путей. Для этого была использована иная реализация интерфейса PathExtender – SimplePathToPathExtender, повторяющая идеологию SimpleExtender, но поддерживающая прибавление кандидатов путей к расширяемому.

Выбор путей-продолжений с помощью Hi-C данных реализован в HiC-PathToPathExtensionChooser, являющемся модифицированной версией HiCExtensionChooser. В связи с изменением концепции возникает целый ряд проблем, требующий разрешения.

Количество Hi-C информации между путями учитывается с первых пяти тысяч нуклеотидов, таким образом увеличив количество Hi-C информации в сравнении с предыдущей реализацией.

Нахождение кандидатов осуществляется с помощью ранее реализованного измененного алгоритма Дейсктры, завершающего работу на длинных ребрах. Далее из доступной информации каждое найденное ребро добавляет все пути, которым оно принадлежит. Пути, аналогично ребрам, имеют ориентацию, следовательно, требуется определить условия, при которых добавляемый путь может называться кандидатом. Например, расположение найденного ребра в конце очень длинного пути свидетельствует, что данный путь, вероятно, следует считать недостижимым. Для решения данной проблемы было принято считать добавляемые пути достижимыми, если найденные ребра, присутствующие в данных путях, расположены в первых пяти тысячах нуклеотидах. Использование отсечки суммирования Hi-C связей в путях на ближайшие пять тысяч нуклеотидов позволяет гарантировать, что найденное ребро будет использовано при расчете количества Hi-C связей между рассматриваемым продолжением и продолжаемым путем.

Ранее в HiCExtensionChooser мы отбрасывали ребра-кандидаты, уже содержащиеся в пути, так как в концепции роста путей это в лучшем случае приводит к завершению работы, а в худшем – образованию цикла и вероятной ошибке сборки. В текущей ситуации мы не ожидаем, что добавляемый путь может полностью содержаться в продолжаемом. В то же время, какое-либо пересечение по ребрам между ними следует признать допустимым, так как главная проблема, которую требуется решить – разрешение повторов, а следовательно, некоторые повторы допустимы. Во-первых, можно ограничить исследование пересечений длинными реб-

рами, но данное ограничение все еще слишком сильное, так как повторы могут достигать и много большей длины, чем пятьсот нуклеотидов [33]. Во-вторых, невозможно не исключать некоторых кандидатов из-за угрозы образования цикла. В качестве гибридного решения были добавлены несколько фильтров.

- Добавляемый путь исключается из пересечения, если среди учитываемых в ходе суммирования Hi-C связей ребер в нем и продолжаемом пути встречаются одинаковые. В связи с обилием Hi-C связей с ребер на самих себя невозможно объективно судить, должен ли такой путь являться продолжением.

- Добавляемый путь исключается из пересечения, если суммарная длина повторяющихся длинных ребер превышает 10% длины добавляемого или продолжаемого путей.

Согласно предыдущим результатам, обратный критерий – очень сильное условие. Одной из ранее исследованных ситуаций, безусловно, требующей его наличия, является присоединение к короткому, малонадежному пути, длинного ребра, что ранее было критично, так как приводило к дроблению существующих верных путей. Для ослабления обратного критерия его запуск производится только в том случае, если отношение длины продолжаемого пути к добавляемому превышает $\frac{5}{4}$.

Возвращаемое значение (Алг. 2) – список кандидатов путей – обеспечено существующей архитектурой, так как в общем случае ExtensionChooser может возвращать отличное от одного количество кандидатов. Выполнение алгоритма происходит в соответствии с концепцией монады Maybe, то есть в случае, если на каком-либо из этапов выполнения алгоритм не может предоставить ответ, будь то выбор максимума по Hi-C связям, не выполненный из-за абсолютного или относительного критериев, или пустой список отфильтрованных кандидатов, алгоритм завершается с отрицательным результатом, представленным пустым списком. Иначе возвращается список с единственным элементом.

Algorithm 2 HiCPathToPathExtensionChooser

procedure FILTER($path$)
    ▷ Последнее длинное ребро в пути
    $pivot \leftarrow FindLastLong(path)$
    ▷ Запуск алгоритма Дейкстры
    $edges\_within\_range \leftarrow FindCandidates(pivot)$
    ▷ Фильтрация кандидатов по пересечениям с $path$
    $filtered \leftarrow FilterCandidates(edges\_within\_range, path)$
    ▷ Выбор максимума по Hi-C с учетом абсолютного и относительных критериев
    $answer \leftarrow ChooseByHiC(path, filtered)$
    if then $\frac{answer.Length}{path.Length} >= 0.8$
        Verify: $CheckBackHiC(path, answer)$
    end if
    return answer
end procedure

Соединение путей происходит после их формирования на этапах роста и корректировки путей. По аналогии с этапом роста путей, формируется CompositeExtender, содержащий единственный SimplePathToPathExtender, после чего вновь происходит запуск корректировки полученных путей. Завершающим этапом является полировка путей.

9. Результаты

9.1. Доступные данные

Несмотря на достаточное время существования Hi-C технологии, основным фокусом её применения были эукариотические организмы, а не метогеномы. Более того, недавняя модификация Hi-C протокола, позволяющая использовать рестриктазы, реагирующие на более короткие последовательности, сужает выбор доступных данных до последних двух лет. Помимо прочего для оценки качества требуется наличие референсных геномов высокого качества присутствующих в данных организмов, что еще сильнее ограничивает спектр выбора данных.

Для экспериментов был использован публично доступный симулированный с помощью sim3C датасет из референсных геномов высокого качества шестидесяти трех бактерий [14]. Оригинальные референсные геномы являются совокупностью скаффолдов, совмещенных авторами в единый референсный геном для простоты симуляции. Рестриктазами, использованными при проведении симуляции, являются «MluCI» и «Sau3aI», реагирующие на последовательности длины четыре нуклеотида.

Несмотря на то что алгоритм изначально разрабатывался для прокариотических организмов, в связи с недостатком данных было было решено использовать естественные данные, полученные из образца *Epichloe festucae Fl1* [27]. Известный референсный геном *Epichloe festucae Fl1* содержит восемь хромосом, длины которых лежат в промежутке от трех до восьми миллионов нуклеотидов, а суммарная длина составляет тридцать пять миллионов нуклеотидов. Использованной рестриктазой является «Sau3AI», реагирующая на последовательность длиной четыре нуклеотида.

9.2. Оценка качества сборки

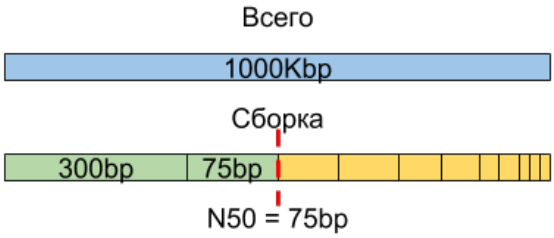
Для оценки качества сборки существует ряд важных метрик. Рассмотрим некоторые из них. Часть метрик доступна лишь при наличии рефе-

ренсного генома.

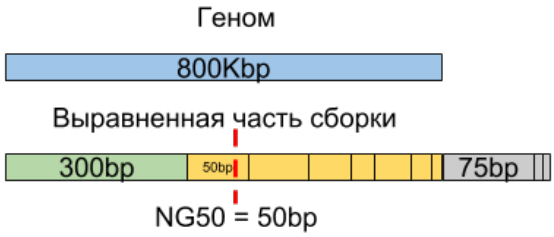
- *Genome fraction* – процент покрытых сборкой нуклеотидов референсного генома. Величина данной метрики лежит в пределах от нуля до ста процентов. В хорошей сборке значение данной метрики близко к ста процентам.

- *Duplication ratio* – суммарная длина выравненной на референсный геном сборки, нормированная длиной референсного генома. Данная метрика демонстрирует степень, в которой сборка дублирует референсный геном. Её значение является неотрицательным рациональным числом, и идеальным результатом является единица.

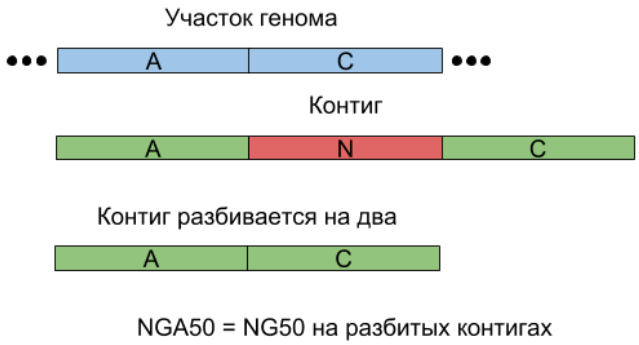
- *N50* – такая длина присутствующего в сборке скаффолда, что суммарная длина всех превышающих данное значение по длине скаффолдов составляет более половины суммарной длины сборки.



- *NG50* – такая длина присутствующего в сборке скаффолда, что суммарная длина всех превышающих данное значение по длине скаффолдов, выравненных на данный референсный геном, составляет более половины длины этого референсного генома.



- *NGA50* – в случае, если в скаффолде присутствует ошибка сборки, например, вставка значительной длины, данный скаффолд разбивается на части, полностью выравненные на референсный геном. На данном множестве запускается метрика *NG50*.



Помимо метрик можно оценивать количество ошибок сборки, подразделяемых на четыре важных типа:

- *релокация* – наличие вставки или пропуска длиной более тысячи нуклеотидов при условии, что выравненные участки принадлежат одной хромосоме одного организма;
- *транслокация* – ошибка сборки, при которой один скаффолд разбивается на две части, выравненные на разные хромосомы одного организма;
- *межвидовая транслокация* – ошибка сборки, когда один скаффолд разбивается на две части, выравненные на референсные геномы разных организмов;
- *инверсия* – ошибка сборки, когда скаффолд разбивается на две части, одна из которых выравнена, а другая выравнена реверс-комплементарно.

9.3. Результаты на симулированных данных

Для экспериментов был использован датасет, симулированный из референсных геномов высокого качества шестидесяти трех бактерий [14]. Из них было выбрано четыре наиболее представленные в данных бактерии для выравнивания сборки на их референсные геномы. Сравнение проходит с базовой версией SPAdes 3.13.0.

Первый референсный геном (3Mbp) собирается базовой версией SPAdes в два фрагмента длиной 1847Kbp и 978Kbp, разделенные фрагментами суммарной длины 15Kbp. Hi-C информация позволяет собрать данный референс целиком, но в связи с наличием в середине пути размера 2Kbp допускается инверсия этого фрагмента, что связано с малой длиной и невозможностью определить с помощью Hi-C направление на малом расстоянии. За исключением инверсии в середине, скаффолд собран корректно.

Второй референсный геном (3.1Mbp) представляет собой девять крупных фрагментов в сборке базовой версии SPAdes, где длина двух наименьших 62Kbp и 133Kbp, а оставшаяся часть сборки не превышает в сумме 50Kbp. С помощью Hi-C удается собрать вторую половину референсного генома в один скаффолд, улучшая первую, но не добавляя серьезных изменений. Тем не менее в сборке присутствует ошибка, и заключительный фрагмент длиной 352Kbp большого скаффолда инвертируется. Было установлено, что это происходит в связи с тем, что количество Hi-C связей на ребра, расположенные на конце референса, превосходит ближайшие варианты, что мы связываем с особенностями симуляции, а не ошибками алгоритма или свойствами данных.

Третий референсный геном (3.3Mbp) собирается базовой версией SPAdes в восемь крупных фрагментов, длина наименьшего из которых – 52Kbp, а остаток составляет менее 60Kbp. Исключая 45Kbp в заключительной части референсного генома, его удается полностью собрать в один скаффолд абсолютно корректно.

Четвертый референсный геном (2.9Mbp) ранее уже восстанавливался полностью, но на стадии полировки путей, соединяющей два крупных сегмента. Так как применение Hi-C информации происходит раньше, то удается соединить фрагменты в скаффолд на более раннем этапе.

Значения метрик оценки качества сборки приведены в таблице 1. Сравнение происходит с базовой версией SPAdes (*базовый*), предварительной

Метрика	Референсный геном	*базовый*	*до Hi-C*	*после Hi-C*
Genome fraction	GCF_001645615	99.968	99.957	100
	GCF_001678755	99.94	99.94	99.91
	GCF_001888565	99.871	99.789	99.788
	GCF_900064425	99.976	99.955	99.965
Duplication ratio	GCF_001645615	1.001	1.001	1.052
	GCF_001678755	1.003	1.002	1.05
	GCF_001888565	1.186	1.022	1.045
	GCF_900064425	1.001	1.001	1.189
NG50	GCF_001645615	1847785	1847785	3069011
	GCF_001678755	380293	494002	1931715
	GCF_001888565	784850	784850	3269347
	GCF_900064425	2923228	2219132	2929422
NGA50	GCF_001645615	1847785	1847785	1848153
	GCF_001678755	352179	493859	540764
	GCF_001888565	784850	784850	3268693
	GCF_900064425	2923128	2219132	2929422

Таблица 1: Значения метрик на сборке симулированных данных

сборкой до полировки путей (*до Hi-C*), версией SPAdes с поддержкой Hi-C данных (*после Hi-C*).

Отличие значений метрик *NG50* и *NGA50* объясняются наличием ошибок сборки, подробно проиллюстрированных ранее.

9.4. Результаты на естественных данных

Для тестирования на естественных данных использовались результаты секвенирования *Epichloe festucae Fl1* [27].

Следует отметить, что улучшаемая сборка далека до полного восстановления генома. Применение Hi-C данных оказывает умеренное положительное влияние на качество сборки, в то же время добавляя значительное число ошибок. Стоит отметить, что большая часть из формально причисляемых к ошибкам ситуаций связана с отсутствием между длинными ребрами, соединяемыми HiCPathToPathExtensionChooser, других подходящих по длине кандидатов.

Для проведения подробного анализа были введены дополнительные типы ошибок, более узкие чем введенные ранее:

- *пропуск* – специализация релокации, в которой между двумя фрагментами скаффолда в геноме содержится фрагмент, отсутствующий в скаффолде;
- *корректный пропуск* – пропуски, не содержащие длинных ребер. В связи с тем, что Hi-C не позволяет работать с короткими ребрами, пропуски подобного рода невозможно разрешить с помощью Hi-C;
- *простая инверсия* – специализация инверсии, когда инвертированый фрагмент скаффолда расположен непосредственно после первого;
- *инверсия с пропуском* – специализация инверсии, при которой между первым и инвертированным фрагментами содержится пропуск;
- *инверсия с корректным пропуском* – специализация инверсии с пропуском, при которой пропуск между первым и инвертированным фрагментами является корректным.

Всего содержащих ошибки скаффолдов семьдесят восемь. Перечисление типов ошибок идет в порядке увеличения степени их угрозы качеству сборки. Тридцать два скаффолда из содержащих ошибки сборки характеризуются только корректными пропусками. Основным типом ошибки являются различные виды инверсий. Двадцать восемь скаффолдов содержат помимо корректных пропусков простые инверсии и инвесии с корректными пропусками. Из оставшихся восемнадцати скаффолдов помимо ошибок ранее перечисленных типов четыре содержат транслокацию, пять инверсий с пропусками и девять пропусков помимо ошибок уже перечисленных типов.

Таким образом, из общего количества ошибочных скаффолдов стоит выделить восемнадцать, ошибки в которых представляют наибольшую угрозу качеству сборки. Основным же типом ошибок является инверсия, что объясняется особенностями Hi-C технологии.

Метрика	*базовый*	*до Hi-C*	*после Hi-C*
Genome fraction	70.642	70.647	70.784
Duplication ratio	1.005	1.005	1.04
NG50	44857	42153	89507
NGA50	43968	41655	52397
Количество ошибочных скаффолдов	1	1	78
Количество скаффолдов со значительными ошибками	1	1	18

Таблица 2: Значения метрик на сборке естественных данных

Значения метрик оценки качества сборки, вычисленные для всего генома, без разделения на каждую хромосому, приведены в таблице 2.

10. Итоги работы

- Изучено физическое устройство протокола Hi-C;
- проведен анализ свойств Hi-C данных;
- разработан прототип решения задачи скаффолдинга;
- разработано расширение геномного ассемблера SPAdes;
- проведено тестирование расширения на симулированных и естественных данных.

Список литературы

[1] Arima Genomics Inc. Hi-C Mapping Pipeline. — GitHub, 2018. — URL: https://github.com/ArimaGenomics/mapping_pipeline (online; accessed: 9.05.2019).

[2] Behjati Sam, Tarpey Patrick S. What is next generation sequencing? // Arch Dis Child Educ Pract Ed. — 2013. — Dec. — Vol. 98, no. 6. — P. 236–238. — 23986538[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/23986538 (online; accessed: 9.05.2019).

[3] Bioinformatics challenges for personalized medicine / Guy Haskin Fernald, Emidio Capriotti, Roxana Daneshjou et al. // Bioinformatics. — 2011. — Jul. — Vol. 27, no. 13. — P. 1741–1748. — 21596790[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/21596790 (online; accessed: 9.05.2019).

[4] Bushman Frederic D. Retroviral integration and human gene therapy // J Clin Invest. — 2007. — Aug. — Vol. 117, no. 8. — P. 2083–2086. — 17671645[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/17671645 (online; accessed: 9.05.2019).

[5] CRISPR/Cas9 & Targeted Genome Editing: New Era in Molecular Biology / Ph.D. Alex Reis, Ph.D. Breton Hornblower, Ph.D. Brett Robb, Ph.D George Tzertzinis. — 2014. — URL: https://international.neb.com/tools-and-resources/feature-articles/crispr-cas9-and-targeted-genome-editing-a-new-era-in-molecular-biology (online; accessed: 9.05.2019).

[6] CRISPR/Cas9 Genome Editing to Disable the Latent HIV-1 Provirus / Amanda R. Panfil, James A. London, Patrick L. Green, Kristine E. Yoder // Front Microbiol. — 2018. — Dec. — Vol. 9. — P. 3107–3107. — 30619186[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/30619186 (online; accessed: 9.05.2019).

[7] CRISPR/Cas9 for Cancer Therapy: Hopes and Challenges / Marta Martinez-Lage, Pilar Puig-Serra, Pablo Menendez et al. //

Biomedicines. — 2018. — Nov. — Vol. 6, no. 4. — P. 105. — 30424477[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/30424477 (online; accessed: 9.05.2019).

[8] Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions / Joshua N. Burton, Andrew Adey, Rupali P. Patwardhan et al. // Nat Biotechnol. — 2013. — Dec. — Vol. 31, no. 12. — P. 1119–1125. — 24185095[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/24185095 (online; accessed: 9.05.2019).

[9] Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome / Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams et al. // Science. — 2009. — Oct. — Vol. 326, no. 5950. — P. 289. — URL: http://science.sciencemag.org/content/326/5950/289.abstract (online; accessed: 9.05.2019).

[10] Coordination 4D Nucleome Data, Center Integration. Pairix // Github. — 2016. — URL: https://github.com/4dn-dcic/pairix (online; accessed: 9.05.2019).

[11] DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins / Je Wook Woo, Jungeun Kim, Soon Il Kwon et al. // Nature Biotechnology. — 2015. — Oct. — Vol. 33. — P. 1162 EP –. — URL: https://doi.org/10.1038/nbt.3389 (online; accessed: 9.05.2019).

[12] De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds / Olga Dudchenko, Sanjit S. Batra, Arina D. Omer et al. // Science. — 2017. — Apr. — Vol. 356, no. 6333. — P. 92. — URL: http://science.sciencemag.org/content/356/6333/92.abstract (online; accessed: 9.05.2019).

[13] DeMaere Matthew Z., Darling Aaron E. Sim3C: simulation of Hi-C and Meta3C proximity ligation sequencing technologies // Gigascience. — 2017. — Nov. — Vol. 7, no. 2. — P. 1–12. — 29149264[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/29149264 (online; accessed: 9.05.2019).

[14] DeMaere Matthew Z., Darling Aaron E. bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes // Genome Biol. — 2019. — Feb. — Vol. 20, no. 1. — P. 46–46. — 30808380[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/30808380 (online; accessed: 9.05.2019).

[15] ExSPAnder: a universal repeat resolver for DNA fragment assembly / Andrey D. Prjibelski, Irina Vasilinetc, Anton Bankevich et al. // Bioinformatics. — 2014. — Jun. — Vol. 30, no. 12. — P. i293–i301. — URL: https://doi.org/10.1093/bioinformatics/btu266 (online; accessed: 9.05.2019).

[16] Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions / Maximilian O. Press, Andrew H. Wiser, Zev N. Kronenberg et al. // bioRxiv. — 2017. — Jan. — P. 198713. — URL: http://biorxiv.org/content/early/2017/10/05/198713.abstract (online; accessed: 9.05.2019).

[17] High-quality genome (re)assembly using chromosomal contact data / Hervé Marie-Nelly, Martial Marbouty, Axel Cournac et al. // Nat Commun. — 2014. — Dec. — Vol. 5. — P. 5695–5695. — 25517223[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/25517223 (online; accessed: 9.05.2019).

[18] Identification of a novel interspecific hybrid yeast from a metagenomic spontaneously inoculated beer sample using Hi-C / Caiti Smukowski Heil, Joshua N. Burton, Ivan Liachko et al. // Yeast. — 2018. — Jan. — Vol. 35, no. 1. — P. 71–84. — URL: https://doi.org/10.1002/yea.3280 (online; accessed: 9.05.2019).

[19] Initial sequencing and analysis of the human genome / International Human Genome Sequencing Consortium, Eric S. Lander, Lauren M. Linton et al. // Nature. — 2001. — Feb. — Vol. 409. — P. 860 EP –. — URL: https://doi.org/10.1038/35057062 (online; accessed: 9.05.2019).

[20] Institute Broad. Picard Tools. — GitHub. — URL: http://broadinstitute.github.io/picard/ (online; accessed: 9.05.2019).

[21] Integrating Hi-C links with assembly graphs for chromosome-scale assembly / Jay Ghurye, Arang Rhie, Brian P. Walenz et al. // bioRxiv. — 2018. — Jan. — P. 261149. — URL: http://biorxiv.org/content/early/2018/02/07/261149.abstract (online; accessed: 9.05.2019).

[22] Lab Mirny. distiller-nf. — GitHub, 2017. — URL: https://github.com/mirnylab/distiller-nf (online; accessed: 9.05.2019).

[23] Li Heng, Durbin Richard. Fast and accurate short read alignment with Burrows-Wheeler transform // Bioinformatics. — 2009. — Jul. — Vol. 25, no. 14. — P. 1754–1760. — 19451168[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/19451168 (online; accessed: 9.05.2019).

[24] Mash: fast genome and metagenome distance estimation using MinHash / Brian D. Ondov, Todd J. Treangen, Páll Melsted et al. // Genome Biology. — 2016. — Jun. — Vol. 17, no. 1. — P. 132. — URL: https://doi.org/10.1186/s13059-016-0997-x (online; accessed: 9.05.2019).

[25] Minikel Eric Vallabh. How PCR duplicates arise in next-generation sequencing. — 2012. — URL: http://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/ (online; accessed: 9.05.2019).

[26] Pevzner P. A., Tang H., Waterman M. S. An Eulerian path approach to DNA fragment assembly // Proc Natl Acad Sci U S A. — 2001. — Aug. — Vol. 98, no. 17. — P. 9748–9753. — 11504945[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/11504945 (online; accessed: 9.05.2019).

[27] Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus Epichloë festucae / David J. Winter, Austen R. D. Ganley, Carolyn A. Young et al. // PLOS Genetics. — 2018. — 10. — Vol. 14, no. 10. — P. 1–29. — URL: https://doi.org/10.1371/journal.pgen.1007467 (online; accessed: 9.05.2019).

[28] SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing / Anton Bankevich, Sergey Nurk, Dmitry Antipov

et al. // J Comput Biol. — 2012. — May. — Vol. 19, no. 5. — P. 455–477. — 22506599[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/22506599 (online; accessed: 9.05.2019).

[29] Sorek Rotem, Lawrence C. Martin, Wiedenheft Blake. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea // Annual Review of Biochemistry. — 2013. — Jun. — Vol. 82, no. 1. — P. 237–266. — URL: https://doi.org/10.1146/annurev-biochem-072911-172315 (online; accessed: 9.05.2019).

[30] Species-Level Deconvolution of Metagenome Assemblies with Hi-C-Based Contact Probability Maps / Joshua N. Burton, Ivan Liachko, Maitreya J. Dunham, Jay Shendure // G3: Genes|Genomes|Genetics. — 2014. — Jul. — Vol. 4, no. 7. — P. 1339. — URL: http://www.g3journal.org/content/4/7/1339.abstract (online; accessed: 9.05.2019).

[31] Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products / Christopher W. Beitel, Lutz Froenicke, Jenna M. Lang et al. // PeerJ. — 2014. — May. — Vol. 2. — P. e415–e415. — 24918035[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/24918035 (online; accessed: 9.05.2019).

[32] Tarjan Robert E., van Leeuwen Jan. Worst-case Analysis of Set Union Algorithms // J. ACM. — 1984. — mar. — Vol. 31, no. 2. — P. 245–281. — URL: http://doi.acm.org/10.1145/62.2160 (online; accessed: 9.05.2019).

[33] Treangen Todd J., Salzberg Steven L. Repetitive DNA and next-generation sequencing: computational challenges and solutions // Nat Rev Genet. — 2011. — Nov. — Vol. 13, no. 1. — P. 36–46. — 22124482[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/22124482 (online; accessed: 9.05.2019).

[34] WATSON J. D., CRICK F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid // Nature. — 1953. — Apr. —

Vol. 171. — P. 737 EP –. — URL: https://doi.org/10.1038/171737a0 (online; accessed: 9.05.2019).

[35] de Wit Elzo, de Laat Wouter. A decade of 3C technologies: insights into nuclear organization // Genes Dev. — 2012. — Jan. — Vol. 26, no. 1. — P. 11–24. — 22215806[pmid]. URL: https://www.ncbi.nlm.nih.gov/pubmed/22215806 (online; accessed: 9.05.2019).