Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики
Кафедра «Компьютерные технологии»

А.А. Ахи

Разработка метода сборки скэффолдов геномных последовательностей на основе принципа максимального правдоподобия

Магистерская работа

Научный руководитель: Ф.Н. Царев

Санкт-Петербург
2013

Оглавление

Введение

Задача сборки геномной последовательности является важной частью биоинформатики. Задача состоит в том, чтобы по набору чтений цепи ДНК восстановить исходную цепь. Основными трудностями в этом процессе являются наличие большого числа ошибок в исходных данных, а также большой объем входных данных, исчисляющийся сотнями гигабайт.

Процесс сборки геномной последовательности обычно состоит из трех частей: исправление ошибок в чтениях, построение *контигов*, достаточно длинных фрагментов исходной цепи ДНК, и построение *скэффолдов*, наборов упорядоченных и ориентированных контигов с известными оценками расстояний между ними. Идеальным результатом сборки геномной последовательности является единственный скэффолд, состоящий из одного контига, соответствующего исходной цепи ДНК. Однако достигнуть подобного результата не удается из-за наличия ошибок во входных данных, а также сложной структуры генома, содержащего большое число повторяющихся частей.

Сборка скэффолдов состоит обычно из трех частей: оценка расстояния между контигами, определение взаимной ориентации контигов и определение взаимного порядка контигов в скэффолде. Для этого часто используют информацию о *парных чтениях* – пары небольших кусочков ДНК с известной оценкой расстояния между ними. Одной из проблем существующих методов сборки скэффолдов является низкая точность оценки расстояния между ними, которая часто производится на основе среднего арифметического. Также проблемой существующих методов является большое число ошибочных связей в результирующей сборке – ситуаций, когда в скэффолде рядом поставлены контиги, располагающиеся далеко друг от друга в исходной цепи ДНК.

В данной работе был разработан метод более точной оценки расстояния между контигами на основе принципа максимального правдоподобия, а также создан метод сборки скэффолдов геномной последовательности, использующий полученные оценки на расстояния и совершающий небольшое число ошибок при сборке. В основе улучшенной оценки расстояния лежит использование

функции правдоподобия, учитывающей не только информации о парах чтений, связывающих пару контигов, но и число несвязывающих чтений, а также длины контигов.

Сборка скэффолдов основывается на построении графа контигов, вершины которого соответствуют контигам, а ребра – парам чтений, соединяющим контиги. Отсеивание подозрительных ребер и вырезание из графа сложных структур позволяет построить первое приближение скэффолдов. Далее скэффолды объединяются с помощью нахождения путей в графе скэффолдов – графе, вершинами которого являются скэффолды, а ребрами – парные чтения, соединяющие концы скэффолдов. В графе скэффолдов каждая вершина имеет два «конца», соответствующие двум концам скэффолда. Объединяющий скэффолды путь должен всегда входить в вершину с одного конца, а выходить с другого.

Ориентация скэффолдов производится с помощью информации на ребрах, соединяющих контиги в одном скэффолде – используется информация о картировании соответствующих пар чтений на соответствующие пары контигов.

Описанный в настоящей работе метод позволяет осуществлять сборку скэффолдов геномных последовательностей на основе информации о парных чтениях. Результирующие скэффолды имеют более точные оценки расстояния между контигами, а также содержат меньшее число ошибочных связей, чем у других существующих методов, не уступая при этом по другим показателям.

Глава 1. Обзор предметной области и существующих методов

1.1. Основные понятия

Биоинформатика является наукой на стыке двух дисциплин: биологии и информатики. Многие задачи биологии требуют обработки большого объема данных, что и привело к возникновению дисциплины. Одной из важных задач биоинформатики является задача сборки геномной последовательности.

1.1.1. Строение и представление молекулы ДНК

Геном — совокупность наследственного материала, заключенного в клетке организма [1]. Геном большинства живых организмов состоит из молекул *дезоксирибонуклеиновой кислоты* (ДНК). ДНК представляет собой полимерную молекулу, состоящую из повторяющихся блоков — *нуклеотидов* [2]. Нуклеотиды, входящие в молекулы ДНК, разделяют на четыре группы согласно азотистым основаниям: *аденин* (A), *гуанин* (G), *цитозин* (C) и *тимин* (T).

В большинстве случаев молекула ДНК состоит из двух цепей, закрученных в спираль [3] и ориентированных азотистыми основаниями нуклеотидов друг к другу. Азотистые основания цепей соединены водородными связями согласно *принципу комплиментарности*: аденин соединяется с тимином, а гуанин — с цитозином.

В биоинформатике геном представляется последовательностью нуклеотидов одной из комплементарных цепей молекулы ДНК. Наиболее удобным представлением является строка, состоящая из символов A, G, C и T, соответствующих типам нуклеотидов. Обратно-комплиментарная цепь ДНК получается путем разворота строки и замены символов-нуклеотидов на комплиментарные.

1.1.2. Секвенирование генома

Для определения линейной последовательности нуклеотидов в молекуле ДНК геном подвергают секвенированию. Одним из популярных методов секвенирования является метод дробовика (Shotgun Sequencing) [4]. Метод состоит в выделении из молекулы ДНК коротких участков (порядка нескольких сотен по-

следовательных нуклеотидов), после чего происходит посимвольное считывание концов выделенных участков (рис. 1). Таким образом, получаются *парные чтения* (*mate-pairs*). В силу множества различных факторов при прочтении отдельных нуклеотидов могут быть допущены ошибки. Также неизвестно точное расстояние между чтениями, известно лишь распределение длин фрагментов.

Рис. 1. Процесс выделения пары чтений из молекулы ДНК.

1.2. Постановка задачи

Традиционно процесс сборки генома состоит из трех этапов: исправление ошибок в парных чтениях, сборка *контигов*, длинных последовательных частей геномной последовательности, и сборка *скэффолдов*, наборов упорядоченных ориентированных контигов с оценками расстояния между соседними контигами. Таким образом, задачей сборки скэффолдов является построение вышеупомянутых наборов по множеству контигов и библиотекам парных чтений. Про библиотеки парных чтений известны математическое ожидание и стандартное отклонение длин фрагментов, из которых были получены парные чтения.

Рис. 2. Схема работы методов построения скэффолдов геномных последовательностей. На вход сборщик получает набор контигов и парных чтений, выходом сборщика является набор скэффолдов.

Задача сборки скэффолдов естественным образом разбивается на три этапа (рис. 2): оценка расстояния между контигами, определение взаимного порядка контигов в исходной геномной цепи и определение ориентации контигов, принадлежность к одной из двух комплементарных цепей молекулы ДНК.

1.3. Используемые обозначения

Под геномной последовательностью в данной работе понимается строка из символов A, C, G и T, соответствующих нуклеотидам и расположенных в порядке, соответствующем следованию соответствующих нуклеотидов в восстанавливаемой цепи ДНК.

Часто для оценки качества сборки геномной последовательности используют величину *N50* [5]. Для набора контигов эта величина определяется следующим образом – это максимальное число *x* такое, что сумма длин контигов, имеющих длину не менее *x* превышает половину суммы длин всех контигов. Например, N50 набора длин (2, 2, 2, 3, 3, 4, 8, 8) составит 8, так как сумма всех длин составляет $2 \cdot 3 + 3 \cdot 2 + 4 + 8 \cdot 2 = 32$, а сумма длин не менее 8 составляет $8 \cdot 2 = 16$, при этом, средняя длина в наборе – 4, а медианное значение длины – 3. По аналогии также часто используют величину N90, в этом случае сумма контигов длиннее *x* должна составить хотя бы 90% от общей длины набора.

Для набора скэффолдов величину N50 определяют аналогичным образом, при этом в качестве длины скэффолда используется сумма длин контигов, в него входящих.

1.4. Обзор существующие методов и их проблем

1.4.1. *OPERA*

Метод сборки скэффолдов геномных последовательностей *OPERA* [6] производит сборку в следующем порядке: вначале определяется взаимная ориентация контигов, затем происходит разбиение контигов на скэффолды и определение взаимного порядка контигов в скэффолдах, финальным этапом является оценка расстояния между контигами в скэффолдах.

Определение взаимной ориентации контигов производится на основании пар чтений, связывающих контиги – парных чтений таких, что одно чтение картируется на один контиг, а другое на другой. При картировании чтений на контиги известно не только местоположение чтения в контиге, но и входит ли оно в контиг прямым или обратно-комплиментарным образом. Так как известна взаимная ориентация чтений в паре и то, как чтения легли на контиги, то представляется возможным определить взаимную ориентацию контигов. Однако пара контигов может быть связана более чем одной парой чтений, которые могут указывать на различные взаимные ориентации. В таком случае, в методе *OPERA* решение о взаимной ориентации принимается согласно методу большинства. Пары чтений, противоречащие принятой ориентации удаляются из набора чтений.

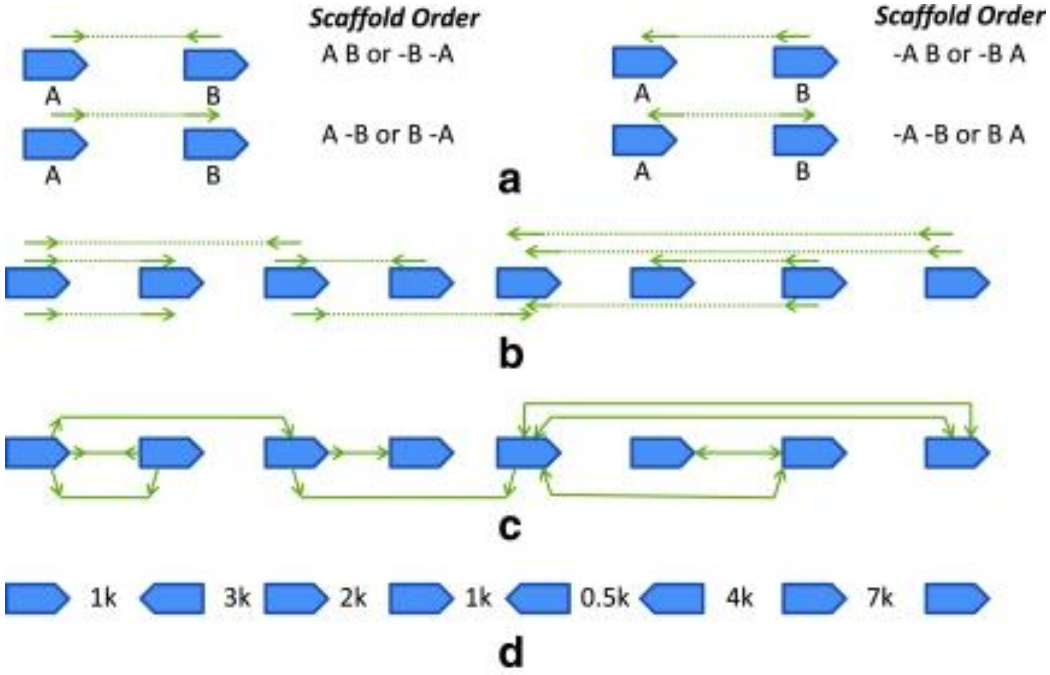

Рис. 3. Описание работы метода сборки скэффолдов *OPERA*. a) Определение взаимной ориентации контигов; b) набор контигов и парных чтений; c) граф скэффолдов; d) результирующий скэффолд. [6]

Для определения взаимного порядка контигов в методе *OPERA* строится *граф скэффолдов* (рис. 3) [6]. Рассматриваемый граф состоит из вершин, соот-

ветствующих контига, и ребер, соединяющих контиги и соответствующих наборам чтений. Граф строится аналогично графу, описанному в работе [7].

Метод состоит в итеративном увеличении скэффолда, путем присоединения к нему все новых контигов. Во время работы метода поддерживаются два множества: *множество висящих ребер* – множество ребер графа, идущих из контигов, уже находящихся в скэффолде, в контиги, не находящиеся в скэффолде на данный момент, и множество, называемое *активным регионом* – множество контигов скэффолда, являющееся минимальным суффиксом скэффолда, содержащим все вершины скэффолда, инцидентные множеству висящих ребер (рис. 4) [6].

Рис. 4. Множество висящих ребер (пунктир) и активный регион (обведено пунктиром). Контиги, уже входящие в скэффолд, выделены оранжевым. [6]

Шаг метода состоит в выборе очередного контига для присоединения к скэффолду. Перебираются все возможные контиги, еще не входящие в строящийся скэффолд, а также обе их возможные ориентации в геноме. Если при присоединении рассматриваемого контига к скэффолду, получаемые множества висящих ребер и активный регион ранее встречались в процессе работы алгоритма, то перебор отсекается, и прибавление контига к скэффолду не рассматривается. Метод заканчивает работу, когда к скэффолду нет возможности добавить новые контиги, и скэффолд не имеет в себе внутренних противоречий.

Для определения расстояния между контигами используется метод максимального правдоподобия. Метод *OPERA* использует распространенное предположение о нормальности распределения длин фрагментов. Используется функция правдоподобия, являющаяся произведением вероятностей получения длин фрагментов для связывающих контиги пар чтений. Такая функция правдоподобия максимизируется средним арифметическим.

Метод *OPERA* способен собирать скэффолды с большой величиной N50. Однако при этом часто наблюдается довольно большое число ошибок в сборке. Наиболее распространенной ошибкой является расположение в скэффолде рядом двух контигов, располагающихся на значительном расстоянии друг от друга в исходной геномной последовательности. Это происходит из-за возникновения ошибочных ребер в графе скэффолдов, ставшего результатом ошибок в чтениях и/или ошибок в используемых контигах. Еще одной причиной возникновения таких ребер может являться наличие повторяющихся частей в геномной последовательности. Более редким видом ошибок в скэффолдах, построенных методом *OPERA*, является неправильное расположение контигов в скэффолде, например два контига поменяны местами.

1.4.2. *GAPEST*

GAPEST [8] является методом оценки расстояния между контигами. Как и многие другие, этот метод основан на принципе максимального правдоподобия и использует предположение о нормальности распределения длин фрагментов пар чтений. Отличие метода состоит в использовании новой функции правдоподобия, которая учитывает некоторую смещенность распределения длин фрагментов чтений, которые связывают контиги.

Прежде всего, стоит отметить, что предположение о нормальности распределения неоднократно подтверждалось на практике. Однако, в данном случае речь идет о всех парах чтений в библиотеке. При рассмотрении пар чтений, связывающих контиги, можно заметить, что пар с большой длинной фрагмента будет несколько больше, чем пар с меньшей длинной. Это связано с тем, что пары чтений с длинной фрагмента, недостаточной для покрытия расстояния между парой контигов в геномной последовательности, просто не способны связать рассматриваемую пару (рис. 5a) [8]. Более того, в случае небольших контигов наблюдается и обратный эффект, когда пары чтений с большой длинной фрагмента не способны связать контиги, так как покрывают значительно большее расстояние (рис. 5b) [8].

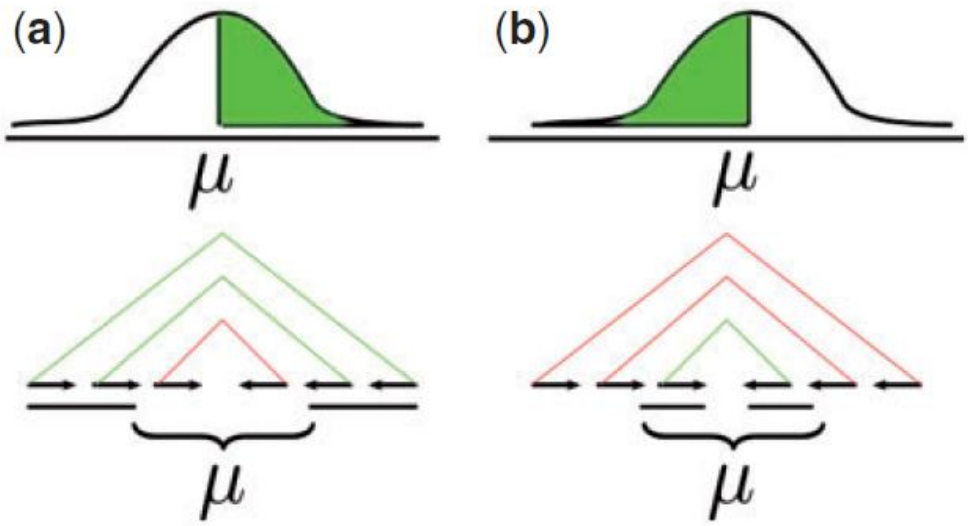

Рис. 5. Наблюдаемая смещенность в условном распределении длин фрагментов. a) наблюдаются только пары чтений с большой длинной фрагмента; b) наблюдаются пары чтений только с маленькой длинной фрагмента. [8]

Для учета смещения в методе *GAPEST* вводят распределение длин фрагментов парных чтений, которые связывают контиги длин c_1 и c_2, находящиеся на расстоянии d (1).

$$h(x \mid d, c_1, c_2) = \frac{p(x \mid d, c_1, c_2) f(x)}{\int_{-\infty}^{+\infty} p(y \mid d, c_1, c_2) f(y) dy} \quad (1)$$

Знаменатель формулы (1) необходим для нормализации, чтобы получить именно вероятностное распределение. За $f(x)$ обозначена вероятность генерации пары чтений с длинной фрагмента x. Согласно общепринятому предположению эта величина подчиняется нормальному закону. Величина $p(x \mid d, c_1, c_2)$ является вероятностью парного чтения с длинной фрагмента x связать пару контигов длин c_1 и c_2, находящиеся на расстоянии d.

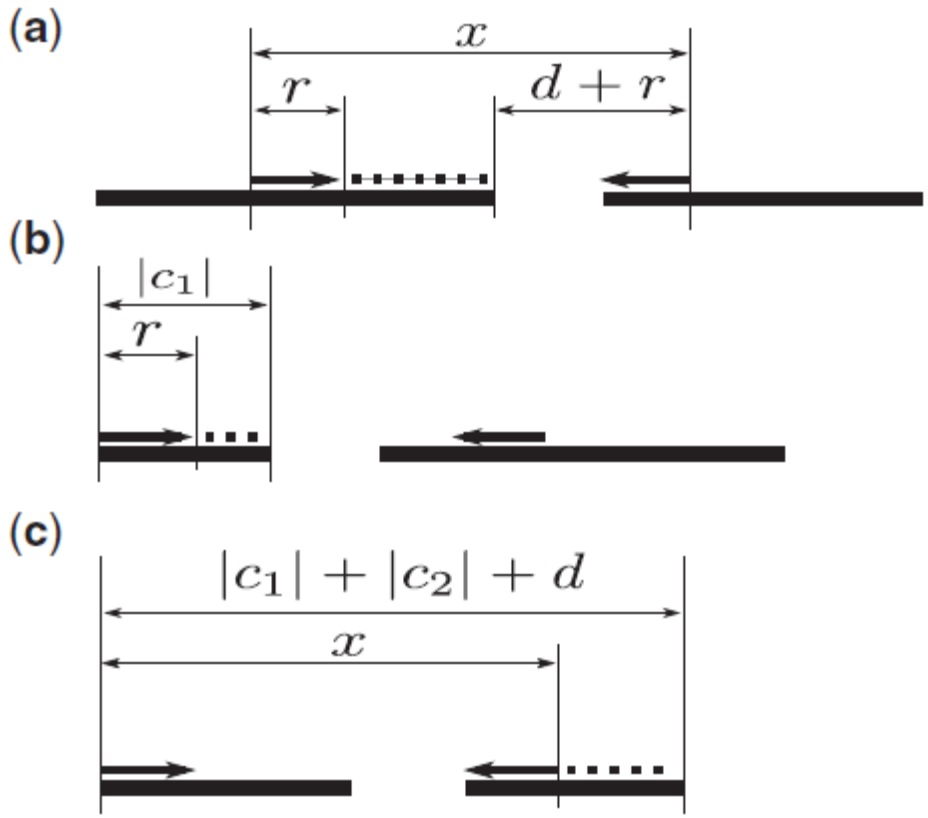

Рис. 6. Возможные случаи местоположения связывающей пары чтений. [8]

Для нахождения вероятности $p(x \mid d, c_1, c_2)$ рассматриваются три случая.

- Первый случай (рис. 6a) [8] рассматривает ситуацию, когда контиги достаточно длинны, а потому пара чтений может закрывать промежуток между контигами любым образом. В таком случае существует $x - d - 2r + 1$ способ размещения фрагмента.
- Второй случай (рис. 6b) [8] рассматривает ситуацию, когда один из контигов настолько короток, что уменьшает число позиций для размещения фрагмента. При предположении того, что $c_1 \leq c_2$, получаем $c_1 - r + 1$ возможных позиций чтения.
- Последний случай (рис. 6c) [8] рассматривает ситуацию, когда длина фрагмента сравнима с величиной $c_1 + c_2 + d$, из-за чего число позиций сокращается до $c_1 + c_2 + d - x + 1$.

Результирующая вероятность получается путем комбинирования трех случаев (2). За G обозначена предполагаемая длина геномной последовательности. За r обозначена длина чтения.

$$p(x \mid d, c_1, c_2) = \frac{1}{G} \min \{\max \{x - d - 2r + 1, 0\},$$

$$\min \{c_1, c_2\} - r + 1, \max \{c_1 + c_2 + d - x + 1, 0\}\} \quad (2)$$

Поиск расстояния осуществляется путем максимизации функции правдоподобия (3). Функция правдоподобия состоит из произведения вероятностей наблюдать чтения с именно такими длинами фрагмента. Длина фрагмента получается путем сложения предполагаемого расстояния с расстояниями от местоположения чтений в контигах до концов контигов.

$$d_{GAPEST} = \arg \max_{d} \prod_{i=1}^{n} h(x_i \mid d, c_1, c_2) \quad (3)$$

Одним из недостатков указанного метода является сложность его реализации. Числитель и знаменатель, используемые в формуле (1), на практике оказываются крайне малыми величинами, что может приводить к большой ошибке в вычислениях. Более того, при небольших длинах контигов функция правдоподобия имеет более одного локального максимума, что делает нахождение оптимального расстояния алгоритмически более сложным.

Метод был разработан для борьбы с заниженной оценкой расстояний между контигами, наблюдаемой у других методов. На практике, хоть результаты оценки с помощью *GAPEST* и оказываются в среднем ближе к реальным значениям, чем у других методов, они также оказываются в среднем больше их. Это указывает на необходимость совершенствования функции правдоподобия.

1.4.3. *SOPRA*

Метод сборки скэффолдов геномных последовательностей *SOPRA* [9] производит сборку в следующем порядке: вначале определяется взаимная ориентация контигов, затем производится оценка расстояния между контигами, финальным этапом является определение взаимного расположения контигов в скэффолдах.

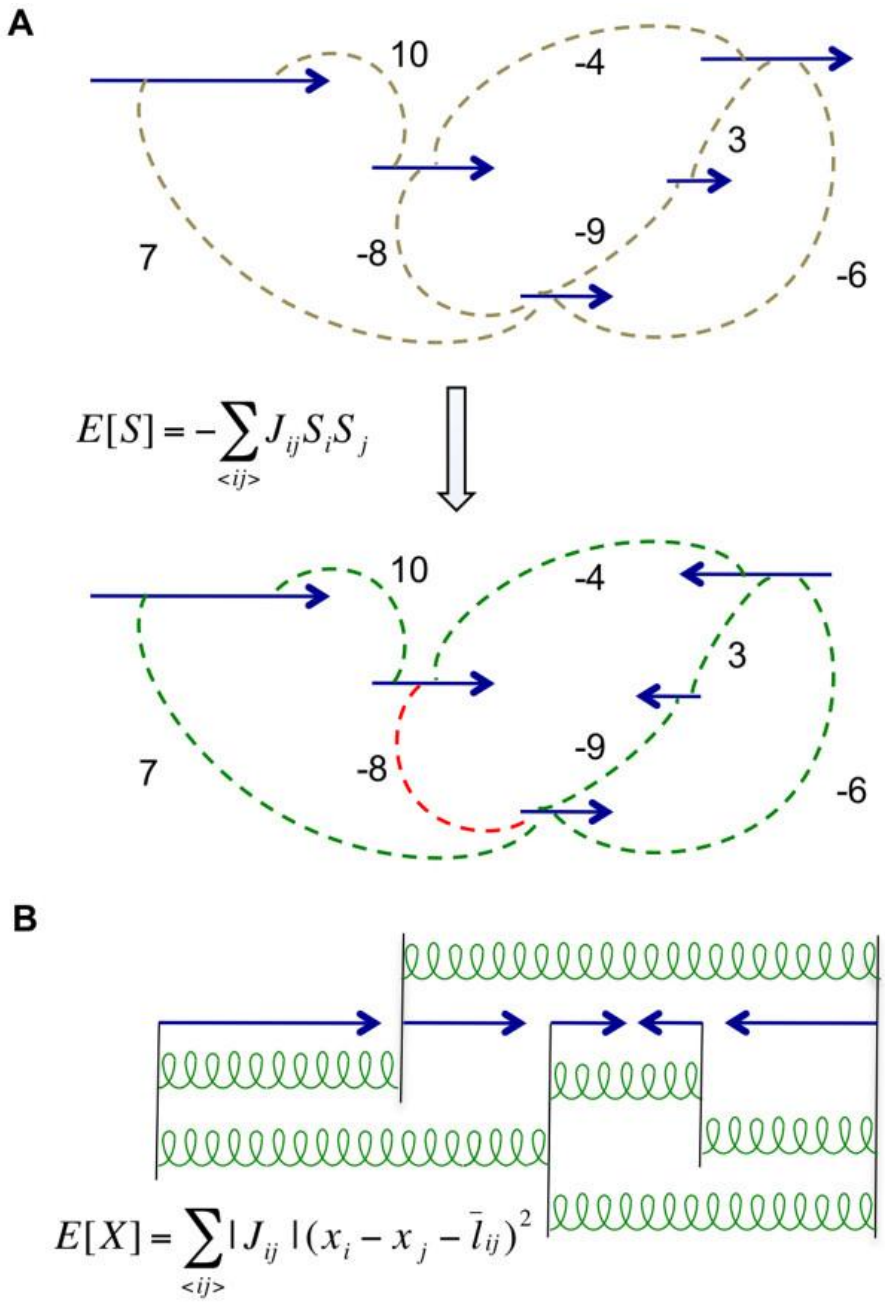

Рис. 7. Преобразование графа контигов в систему пружинок, применяемое в методе сборки скэффолдов SOPRA. [9]

Определение взаимной ориентации производится на основе информации о парных чтениях. В случае возникновения конфликтов – ситуаций, когда разные пары чтений говорят о разной взаимной ориентации пары контигов, они разрешаются по принципу большинства. Если перевес большинства небольшой, то все чтения, участвующие в конфликте, убираются из рассмотрения. Затем выбирается ориентация контигов таким образом, чтобы она не входила в про-

тиворечие с большинством оставшихся в рассмотрении чтений. Так как такая задача является NP-полной [10, 11], для ее быстро решения применяются эвристические жадные методы.

Следующим этапом сборки скэффолдов в методе *SOPRA* является определение расстояния между контигами. Расстояние определяется исходя из распространенного предположения о нормальности распределения длин фрагментов пар чтений. Согласно свойствам нормального распределения оптимальным значением расстояния будет являться среднее арифметическое по оптимальным расстояниям для каждой пары чтений (4). За J_{ij} обозначено число чтений, связывающих контиги i и j, за μ – средняя длина фрагмента пар чтений, за r – длина чтений, за d_{ik} расстояние от места картирования k-ого чтения на i-ый контиг до края контига.

$$l_{ij} = \frac{1}{J_{ij}} \sum_{k=1}^{J_{ij}} \left(\mu - d_{ik} - d_{jk} - 2r \right) \tag{4}$$

Заключительным этапом в методе *SOPRA* является определение расположения контигов в геноме. Для этого каждому контигу сопоставляется величина x_i координата i-го контига в геномной последовательности. Для определения местоположения контигов, связывающие пары чтений представляются в виде пружинок (рис. 7) [9]. При этом пружинка, соединяющая i и j, имеет расслабленную длину l_{ij}, а за жесткость пружинки берется J_{ij}. Далее находится позиция равновесия в получившейся физической задаче. Для нахождения равновесия используются эвристические и численные методы. Нахождение равновесной ситуации в данном случае эквивалентно задачи минимизации энергии системы (5).

$$E[X] = \sum_{<ij>} \frac{J_{ij}}{2} \left(x_i - x_j - l_{ij} \right) \tag{5}$$

Недостатками метода *SOPRA* являются низкое качество оценки расстояний между контигами, так как не учитывается смещенность распределения длин фрагментов пар чтений, связывающих пару контигов. Также стоит отме-

тить сравнительно долгое время работы метода. Это связано с большой сложностью возникающей физической системы пружинок. Другой существенный вклад во время работы метод вносят эвристические методы определения взаимной ориентации контигов. Несмотря на указанные недостатки, метод *SOPRA* собирает достаточно длинные скэффолды с большой величиной *N50*.

Глава 2. Разработка сборщика скэффолдов геномных последовательностей

Представленный метод сборки скэффолдов, как и многие другие, состоит из трех частей:

1. Оценка расстояния между контигами;
2. Определение взаимного порядка контигов в скэффолдах;
3. Определение взаимной ориентации контигов.

В отличие от многих других методов, представленный метод использует именно указанный порядок решения подзадач.

2.1. Оценка расстояния на основе принципа максимального правдоподобия

Разработанный метод оценки расстояния основан на принципе максимального правдоподобия. Была разработана новая функция правдоподобия, обеспечивающая более точную оценку расстояния. Новая функция учитывает не только пары чтений, которые связывают пару контигов, но и длины контигов, и число несвязывающих чтений.

Метод оценки расстояния между контигами состоит из двух шагов: картирование чтений на контиги и оценка расстояния с использованием метода максимального правдоподобия на основе результатов картирования.

2.1.1. Модель получения чтений

Для построения функции правдоподобия предлагаемый метод оценки расстояния использует распространенное предположение о нормальности распределение длин фрагментов, из которых получаются парные чтения. Также используется предположение, что парные чтения получаются согласно следующей процедуре:

1. Вначале согласно нормальному распределению выбирается длина фрагмента, из которого будет получена пара чтений;
2. Затем равновероятно выбирается местоположение фрагмента в геномной последовательности;

3. Далее происходит считывание концов выбранного фрагмента, и получается пара чтений.

Экспериментальные данные показывают высокую степень достоверности предложенной модели. Использование данной модели позволяет построить функцию правдоподобия, обеспечивающую более точную оценку расстояния между контигами.

2.1.2. Определение параметров модели

Параметры нормального распределения длин фрагментов можно получить из парных чтений, оба чтения которых картируются на один и тот же контиг. В таком случае можно производить оценку параметров стандартными методами статистики. В качестве математического ожидания длины фрагмента в таком случае выступает среднее арифметическое длин фрагментов разных пар чтений. Стандартным отклонением является среднеквадратичное отклонение длин фрагментов от этой величины.

Стоит отметить, что рассматриваемое распределение длин фрагментов пар чтений, попадающих на один контиг, является смещенным, так как на короткие контиги могут быть картированы только пары чтений с небольшой длиной фрагмента Другой проблемой является наличие повторяющихся частей в геноме. Из-за повторов пара чтений оказываться в контиге на расстоянии многократно превышающем среднюю длину фрагмента.

Для борьбы с этими трудностями оценка параметров распределения проводится в два этапа. На первом этапе происходит примерное определение параметров – берется среднее арифметическое длин фрагментов по все парам чтений, попадающим на один контиг. Обозначим первичную оценку средней длины фрагмента за μ'. После первого этапа из рассмотрения убираются все контиги, длина которых не превышает $2\mu'$. Также из рассмотрения убираются пары чтений, для которых длина фрагмента превышает $2\mu'$. После этого происходит второй подсчет параметров распределения длин фрагментов пар чтений. Такой двухэтапный метод вычисления параметров позволяет

уменьшить влияние ошибок и смещенности наблюдаемого распределения на оценку параметров.

2.1.3. Учет связывающих чтений

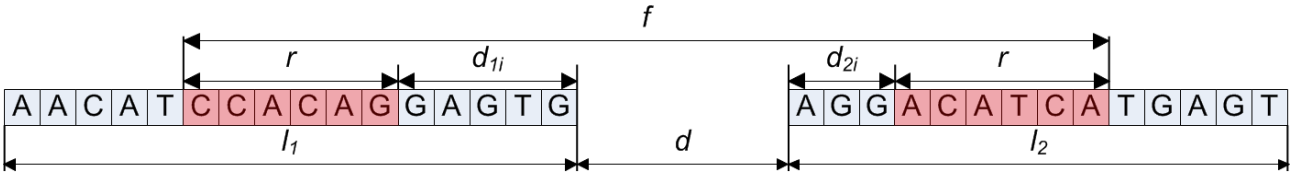

Рис. 8. Пример парного чтения, связывающего контиги. l_1, l_2 – длины контигов; d_{1i}, d_{2i} – расстояния от мест картирования чтений до краев контигов; r – длина чтения; f – длина фрагмента, из которго получена пара чтений; d – предполагаемое расстояние между контигами.

Для учета связывающих чтений в функцию правдоподобия включаются вероятности получения таких чтений, согласно используемой модели получения чтений. Про каждую пару чтений, связывающую пару контигов, известно место их картирования на контиги и, соответственно, расстояние до этих мест от краев контигов (рис. 8). Нужный край определяется согласно ориентации чтения при картировании.

Вклад каждого связывающего чтения в используемую функцию правдоподобия состоит из двух частей: вероятности получить фрагмент желаемой длины (6) и вероятности получить наблюдаемое местоположение фрагмента в геномной последовательности (7). За L обозначена предполагаемая длина геномной последовательности.

$$p(\ fragment\quad size\ \mid d\ ) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\left(d_{1i} + d_{2i} + d + 2r - \mu\right)^2}{2\sigma^2} \right) \tag{6}$$

$$p(\ fragment\quad position\quad \mid d\ ) = \frac{1}{L} \tag{7}$$

Результирующий вклад каждого чтения состоит из произведения этих вероятностей (8). Суммарный вклад связывающих чтений в функцию правдоподобия состоит из произведения вероятностей получения чтений (9). За n

обозначено число парных чтений, связывающих рассматриваемую пару контигов.

$$p(mate\quad pair\ \mid d) = \frac{1}{\sqrt{2\pi}\sigma L}\exp\left(-\frac{(d_{1i} + d_{2i} + d + 2r - \mu)^2}{2\sigma^2}\right) \tag{8}$$

$$p(connecting\quad \mid d) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma L}\exp\left(-\frac{(d_{1i} + d_{2i} + d + 2r - \mu)^2}{2\sigma^2}\right) \tag{9}$$

2.1.4. Учет несвязывающих чтений

Помимо n пар чтений, связывающих контиги, в библиотеке содержатся еще $R - n$ пар чтений, которые рассматриваемые контиги не связывают. За R обозначено общее число пар чтений в библиотеке. Предлагаемый метод отличается тем, что и такие чтения учитываются в используемой функции правдоподобия.

Для учета несвязывающих пар чтений вычисляется вспомогательная величина – вероятность случайной пары чтений связать рассматриваемую пару контигов. Вероятность вычисляется при условии предполагаемого расстояния между контигами d и согласно предлагаемой модели получения чтений. Данная вероятность вычисляется с использованием формулы полной вероятности, где параметром служит длина фрагмента, из которого получается пара чтений (10).

$$p(random\quad connect\quad \mid d) = \sum_{f} p(f \mid d)\, p(connect\quad \mid f, d) \tag{10}$$

При фиксированной длине фрагмента, вероятность получения такой длины фрагмента вычисляется согласно нормальному распределению. Для нахождения вероятности пары чтений с длиной фрагмента f связать пару контигов, расположенных на расстоянии d друг от друга, необходимо найти число «хороших» позиций пары чтений – позиций в геномной последовательности, попадая на которые пара чтений будет связывать рассматриваемую пару контигов.

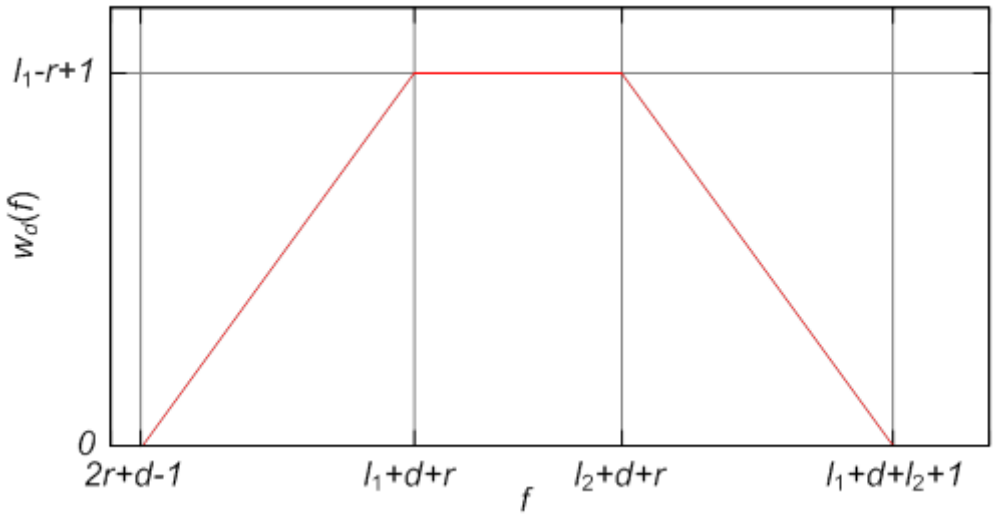

Рис. 9. Число «хороших» позиций фрагмента в зависимости от его длины.

Будем обозначать число хороших позиций как $w_d(f)$. Данная функция использовалась в работе [8] и описывается формулой (2). График этой функции в зависимости от параметра f показан на рисунке 9. Таким образом, вероятность «хорошего» местоположения фрагмента равняется $w_d(f)/L$. Комбинируя полученные результаты, получаем вероятность случайной пары чтений связать пару контигов (11).

$$p(random \quad connect \quad | d) = \sum_f \frac{w_d(f)}{\sqrt{2\pi}\sigma L} \exp\left(-\frac{(\mu - f)^2}{2\sigma^2}\right) \tag{11}$$

Однако, для учета несвязывающих чтений используется не вероятность случайной пары чтений связать пару контигов, а обратная величина – вероятность случайной пары чтений не связать пару контигов, то есть $1 - p(random \quad connect \quad | d)$. Так как каждая из $R - n$ пар чтений не связала рассматриваемую пару контигов, то она дает вклад в функцию правдоподобия $1 - p(random \quad connect \quad | d)$. Перемножая вклады всех чтений получаем суммарный вклад всех несвязывающих чтений (12).

$$p(non - connecting \quad | d) = (1 - p(random \quad connect \quad | d))^{R-n} \tag{12}$$

2.1.5. Нахождение наиболее вероятного расстояния

Используемая функция правдоподобия (13) получается комбинированием частей, учитывающих связывающие (9) и несвязывающие (12) пары чтений. Наиболее вероятное расстояние находится с помощью максимизации функции правдоподобия (14).

$$p(reads \mid d) = (1 - p(random\ \ connect \mid d))^{R-n}\, p(connecting \mid d) \quad (13)$$

$$d_{opt} = \arg\max_{d} p(reads \mid d) \quad (14)$$

Однако, получившаяся функция правдоподобия неудобна для использования, так как ее вычисление требует слишком большого числа действий, а получающиеся значения чрезвычайно малы. Для удобства работа ведется не с самой функцией правдоподобия, а с ее логарифмом. В таком случае часть, отвечающая за учет связывающих пар чтений, преобразуется в квадратный трехчлен от d (15). Если предпосчитать коэффициенты этого многочлена, то вычисление рассматриваемой части логарифма функции правдоподобия будет производиться за $O(1)$ для любого d.

$$\log p(connecting \mid d) = n \log\left(\frac{1}{\sqrt{2\pi\sigma L}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(d_{1i} + d_{2i} + d + 2r - \mu\right)^2 \quad (15)$$

Часть функции правдоподобия, отвечающая за учет несвязывающих чтений, при взятии логарифма не приобретает простого вида. Остается необходимость эффективного вычисления вероятности случайной пары чтений связать рассматриваемую пару контигов. Для этого сумма в формуле (11) аппроксимируется интергралом (16).

$$p(random\ \ connect \mid d) \approx \int_{-\infty}^{+\infty} \frac{w_d(f)}{\sqrt{2\pi\sigma L}} \exp\left(-\frac{(\mu - f)^2}{2\sigma^2}\right) df \quad (16)$$

Так как функция $w_d(f)$ имеет вид трапеции и почти всюду равна нулю (рис. 9), вычисление такого интергала можно разбить на вычисление трех интегралов. Каждый из этих интрегралов вычисляется за $O(1)$ с помощью табличных значений для нормального распределения при любом значении d.

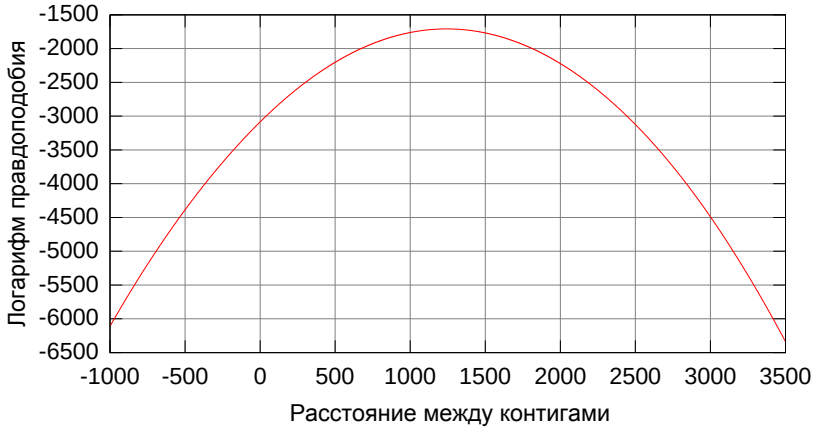

Рис. 10. Пример графика логарифма функции правдоподобия в зависимости от расстояния между контигами.

Модифицированный логарифм функции правдоподобия вычисляется за $O(1)$ для любого значения d при выполненом за $O(n)$ предподсчете коэффициентов многочлена от d. Экспериментальные данные показывают, что логарифм функции правдоподобия практически всегда имеет выпуклый вид (рис. 10), что позволяет находить его максимум с помощью тернарного поиска. В процессе тернарного поиска требуется осуществить $O(\log \mu)$ вычислений логарифма функции правдоподобия. Так как вычисление логарифма функции правдоподобия осуществляется за $O(1)$, для нахождения наиболее вероятного расстояния между парой контигов требуется $O(n + \log \mu)$ операций.

2.1.6. Работа с несколькими библиотеками парных чтений

Часто оказываются доступны несколько библиотек парных чтений с различными параметрами. Считается, что это позволяет увеличить качество сборки как контигов, так и скэффолдов. Разработанный метод оценки расстояния между контигами легко распространить на такой случай. С этой целью необходимо внести изменения в используемую функцию правдоподобия.

Для учета связывающих чтений необходимо в формулу (9) подставить параметры распределения длин фрагментов, характерные для той библиотеки, которой принадлежат чтения (17).

$$p(connecting \quad | d) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_{l_i} L} \exp\left(- \frac{\left(d_{1i} + d_{2i} + d + 2r - \mu_{l_i}\right)^2}{2\sigma_{l_i}^2} \right) \quad (17)$$

Учитывать несвязывающие чтения следует для каждой из библиотек в отдельности. Для этого для каждой из библиотек находится вероятность случайной пары чтений из этой библиотеки связать рассматриваемую пару контигов. Затем в функцию правдоподобия для каждой несвязывающей пары чтений включаются вероятность того, что эта пара не свяжет пару контигов при условии, что для получения чтения используются параметры распределения ее библиотеки (18).

$$p_l(random \quad connect \quad | d) = \sum_{f} \frac{w_d(f)}{\sqrt{2\pi}\sigma_l L} \exp\left(- \frac{(\mu_l - f)^2}{2\sigma_l^2} \right) \quad (18)$$

$$p(reads \quad | d) = p(connecting \quad | d) \prod_{l=1}^{m} (1 - p_l(random \quad connect \quad | d))^{R_l - n_l} \quad (19)$$

Полученная таким образом функция правдоподобия (19) для нескольких библиотек так же подвергается логарифмированию и изменениям для повышения эффективности вычислений. Время работы метода в случае использования m библиотек увеличивается, так как теперь вычисления логарифма функции правдоподобия осуществляется за $O(m)$, а не за $O(1)$, как было ранее. Таким образом, метод оценки расстояния при использовании m библиотек совершает $O(n + m \log \mu)$ действий.

2.2. Определение взаимного порядка

За фазой оценки расстояния между контигами следует фаза определения взаимного порядка контигов, а также объединения контигов в скэффолды. Определение взаимного порядка состоит из нескольких частей: в начале

стоится граф контигов, по которому затем сторится первое приближение скэффолдов – скэффолды небольшой длины, имеющие малую вероятность ошибки. Далее стоится граф скэффолдов, который позволяет объединять небольшие скэффолды в значительно большие. На финальном этапе в скэффолдах заполняются пропуски, возникшие по тем или иным причинам на предыдущих этапах.

2.2.1. Построение графа контигов

Первоначальным этапом определения взаимного порядка контигов является построения графа контигов. Вершинами в таком графе являются контиги, поступающие на вход сборщику скэффолдов, а ребра соответствуют наборам пар чтений, связывающих пару контигов. Каждое ребро графа имеет длину – оценку на расстояние между соответствующими контигами.

К сожалению, построенный граф содержит большое число ошибочных ребер, которые возникают из-за того, что геномная последовательность содержит большое число повторов, а также в используемых парных чтениях допущенно большое число ошибок. Однако, часть таких ребер все таки можно выявить и удалить на этапе построения графа контигов.

Прежде всего, удаляются ребра, которые соответствуют множествам пар чтений меньше порогового размера. Вероятность возникновения такого ребра в результате появления ошибок в чтения достаточно велика, а оценка расстояния между контигами, соеденеными таким ребром, является крайне неточной. В любом из этих случаев, удаление ребра должно положительно повлиять на конечный результат сборки скэффолдов.

$$n_{\exp\ ect} = p(random\quad connect\quad | d)R \tag{20}$$

Другой способ выявления ошибочных ребер основан на использование описанной в разделе 2.1.4 вероятности случайной пары чтений связать пару контигов (16). Вычисление этой вероятности позволяет также вычислить математическое ожидание числа пар чтений, связывающих рассматриваемую пару контигов (20). Большое отличие ожидаемого числа связывающих пар чтений от фактического может являться индикатором ошибочности ребра.

Стоит однако отметить, что из-за ошибок в чтениях фактическое число связывающих пар чтений в среднем несколько ниже ожидаемого. Данные преположения находят свое подтверждение в результатах экспериментальных проверок. Назначение правильного порогового отклонения фактической величины от ожидаемой позволяет значительно снизить как число ошибочных ребер, так и соотношение числа ошибочных ребер к числу правильных.

2.2.2. Построение первого приближения скэффолдов

Для построения первого приближения скэффолдов граф контигов подвергается дальнейшему упрощению. Как известно, геномные последовательности даже простейших организмов содержат большое число достаточно продолжительных повторов. Этот факт сильно затрудняет сборку генома как на этапе построения контигов, так и на этапе построения скэффолдов. Чтобы упростить нахождение первоначального приближения из графа контигов удаляются контиги, соответствующие повторяющимся частям геномной последовательноти.

Модель получения парных чтений содержит предполоржение о том, что геном покрыт чтениями равномерно. Величиной покрытия будем называть соотношение суммарной длины чтений, падающих на какую-либо последовательность нуклеотидов, к длине этой последовательности. Если рассмотреть последовательность нуклиотидов, встречающуюся в геноме более одного раза, то при картировании на такую последовательность чтений, покрытие будет значительно превышать покрытие генома. Таким образом, контиги, имеющие покрытие генома, значительно превышающее покрытие восстанавливаемой геномной последовательности, скорее всего соответствуют повторам. По этой причиене, такие контиги удаляются из графа контигов.

Однако, даже после удаления вершин, соответствующих контигам с высоким покрытием, граф контигов часто все равно имеет слишком сложную структуру. Чтобы еще сильнее упростить граф контигов, из него удаляются все вершины степени, превышающей три. Полученный граф имеет достаточно простую структуру. Для нахождения первого приближения скэффолдов в нем

используется жадный алгоритм. В качестве скэффолдов в графе выбираются простые пути, содержащие максимальное число вершин, имеющие при этом наименьшую возможную суммарную длину ребер. Найденный путь представляет из себя новый скэффолд, а его вершины удаляются из графа контигов. Если граф остается не пуст, то в нем продолжается поиск новых путей и соответственно скэффолдов.

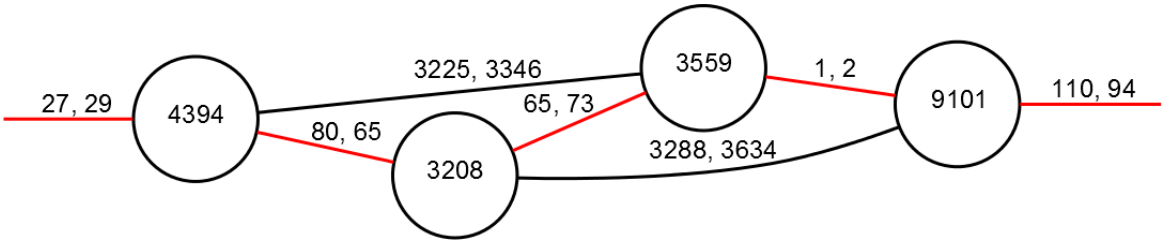

Рис. 11. Пример работы жадного метода при построении первого приближения скэффолдов. В вершинах указаны длины соответствующих контигов. На ребрах обозначены полученные оценки расстояния, а также истинное значение расстояния. Красным отмечены ребра, входящие в путь, выбранный жадным методом.

Рассмотрим более подробно оправдонность использования такого жадного алгоритма. Для этого расмотрим крайний случай, когда все ребра, оставшиеся в графе не являются ошибочными, а расстояния между контигами оценены абсолютно точно. Также будем считать, что геномная последовательность не содержит достаточно длинных повторов. В таком случае, именно ребра минимальной длины связывают контиг с его соседями. Исходя из этого соображения именно по кратчайшим ребрам стоит «входить» и «выходить» из контига при проходе по скэффолду в порядке, соответствующему исходной геномной последовательности. Полученный путь будет являться кратчайшим, среди всех путей, проходящих через все контиги скэффолда, так как он покрывает каждый нуклеотид исходной геномной подпоследовательности, соответствующей рассматриваемому скэффолду, ровно один раз, в то время как любой другой путь будет покрывать некоторые нуклеотиды дважды, а значит его суммарная длина будет больше.

К сожалению, графы, получаемые на реальных данных, далеки от рассмотренного выше идеального случая. Однако, использование тех же жадных методов выделения скэффолдов позволяет построить скэффолды, содержащие малое число ошибочных связей. Полученные таким образом скэффолды часто имеют не очень большую длину, поэтому требуюся следующие этапы сборки.

2.2.3. Объединение скэффолдов с помощью эвристики кратчайшего пути

Для увеличения длины скэффолдов, на следующем этапе сборки производится попытка определить контиги, находящиеся между уже построенными скэффолдами и, таким образом, эти скэффолды объединить в один большой.

С этой целью строится граф скэффолдов, вершинами которого являются уже построенные скэффолды. В отличие от обычных графов, вершины графа скэффолдов имеют концы, соответствующие концам скэффолда. Ребра соединяют концы скэффолдов, если есть чтения, соединяющие контиги, являющиеся крайними на соответствующих концах своих скэффолдов. Если крайний контиг скэффолда сильно меньше средней длины фрагмента, из которого получаются парные чтения, то допускаются и чтения, связывающие второй с края контиг.

Полученный граф скэффолдов содержит как скэффолды, полученные на предыдущем этапе сборки, так и множество скэффолдов, состоящих из одиночных контигов. Опять рассмотрим идеальный случай, когда уже собранные скэффолды не содержат ошибок, а построенные между концами скэффолдов ребра имеют точные оценки расстояния. Более того, будем считать, что геномная последовательность не содержит достаточно длинных повторов. В рассматриваемом случае, чтобы определить, какие контиги располагаются в исходной последовательности между уже построенными скэффолдами, вновь можно действовать согласно принципу кратчайшего пути.

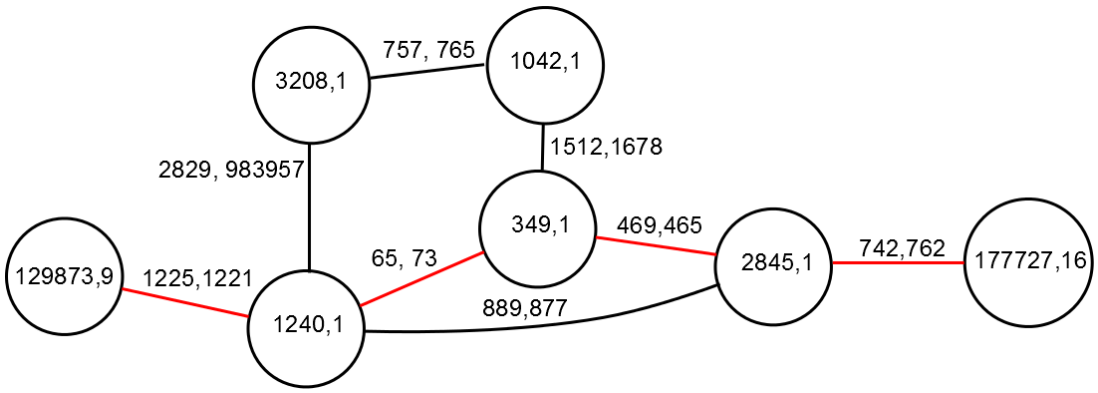

Рис. 12. Пример работы жадного метода при объединении скэффолдов. В вершинах обозначена суммарная длина контигов в скэффолде, а также число контигов. На ребрах обозначены оценки расстояний и реальные расстояния. Красным отмечены ребра пути, выбранного жадным эвристическим методом.

Рассмотрим чуть более подробно, как именно будет выглядеть принцип кратчайшего пути в данном случае. Помимо определения контигов, расположенных между скэффолдами, также необходимо по возможности понять, в каком порядке идут скэффолды в исходной геномной последовательности. Принцип кратчайшего пути позволяет ответить на оба вопроса сразу. Для этого будем находить кратчайшие пути из концов скэффолдов, до концов других скэффолдов, проходящие по скэффолдам, состоящим из одного контига. В данном случае, длина пути – это суммарная длина оценок на ребрах. В идеальном случае, из каждого конца любого скэффолда будет достижимо не более одного другого конца скэффолда. Из этого следует, что соединенные путем скэффолды расположены в геноме рядом, причем направленны друг к другу концами, которые связал путь. Контиги, по которым прошел путь, и есть контиги, расположенные между скэффолдами в исходной геномной последовательности. Нахождение именно кратчайшего пути позволяет не пропускать контиги между скэффолдами, так как покрытие нуклеотида исходной геномной цепи проходом по ребру вносит вклад в длину пути, равный единице, в то время как покрытие этого нуклеотида соответствующим контигом обходится «бесплатно». Скэффолды, соединенные

путем, а также все контиги на пути объединяются в новый большой скэффолд, после чего граф скэффолдов перестраивается.

Такой же, как и в идеальном случае, метод используется для объединения скэффолдов и в рельном случае. Однако, для приспособления описанного метода к реальной ситуации в него необходимо внести некоторые изменения. К сожалению, граф скэффолдов содержит большое число ошибочных ребер, оценки расстояний на ребрах не идеально точны, а также геномная последовательность содержит достаточно большие повторы, что приводит в возникновению контигов, которые полностью входят в повторяющуюся часть, а потому имеют много связей. Для того, чтобы избавиться от последней проблемы, на начальном этапе фазы объединения скэффолдов из графа удаляются все ребра, хотя бы один конец которых падает на контиг с слишком большим покрытием. Таким образом, контиги, соответствующие повторам, практически полностью исключаются из графа скэффолдов, что позволяет находить скэффолды, не содержащие в себе повторяющихся частей исходной геномной последовательности.

Для того, чтобы избежать использования ошибочных ребер, на ранних этапах объединения скэффолдов вводятся дополнительные ограничения на пути, соединяющие концы скэффолдов. Такое ограничение состоит в том, что каждое из ребер пути должно быть единственным ребром соответствующего конца для хотя бы одного из скэффолдов. Поскольку ребро является единственным с этого конца, то любой путь, содержащий этот скэффолд, будет использовать это ребро. Путь, построеный с использованием такого ограничения, является почти единственным возможным.

Когда новые пути, удовлетворяющие ограничениям, более не находятся, ограничения несколько ослабляются. Однако стоит избегать полного избавления от ограничений, при поиске кратчайшего пути, так как это с большой долей вероятности приведет к использованию ошибочного ребра для объединения скэффолдов. Допущение ребер, соединяющих контиги с покрытием чтениями, повышающем покрытие геномной последовательности не

более чем в три раза, а также ослабление требований на допустимое число альтернативных ребер на концах скэффолдов до двух, позволяют производить дальнейшее объединение скэффолдов, не допуская при этом большого числа ошибок.

2.2.4. Восстановление пропусков

На этапе объединения скэффолдов отсутствует какая-либо возможность вставить контиги в середины уже существующих скэффолдов. Из-за этого на разных этапах сборки могут возникать пропуски в скэффолдах. Чаще всего это пропуски небольших контигов, возникшие из-за удаления таких контигов из графа контигов на первом этапе построения скэффолдов. Но пропуски могут возникать и на других этапах сборки.

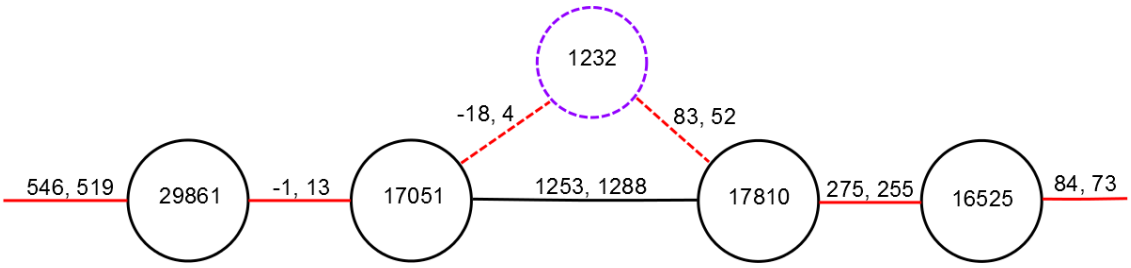

Рис. 13. Пример успешного восстановления пропущенного контига. В вершинах обозначена длина контигов. На ребрах обозначены оценки на расстояния и реальные расстояния между контигами. Красным обозначены ребра, соединяющие соседние контиги в результирующем скэффолде. Пунктиром выделен контиг, вставляемый в середину скэффолда.

Следующей фазой сборки скэффолдов является заполнение этих пропусков. Если есть контиг (или скэффолд), ребра с разных концов которого ведут к двум контигам, расположенным в одном скэффолде и являющихся соседними, то такой контиг вставляется в скэффолд между контигами, с которыми он связан. Такая процедура позволяет расставить по местам небольшие контиги, через которые иногда «перескакивают» парные чтения. Восстановление пропусков можно производить совместно с объединением скэффолдов, итеративно улучшая результат до тех пор, пока не удается внести изменения не противоречащие используемым ограничениям.

2.3. Определение ориентации

Определение взаимной ориентации контигов является заключительным этапом предлагаемого метода сборки скэффолдов. Определение ориентации осуществляется после разбиения контигов на скэффолды и с использованием только тех ребер графа контигов, которые соединяют контиги из одного скэффолда.

Ориентация контигов определяется ленивым образом: контиги скэффолда обходятся обходом в глубину с использованием ребер графа контигов, а ориентация контига выбирается таким образом, чтобы она согласовывалась с большинством чтений ребра, по которому алгоритм пришел в вершину этого контига. Так как скэффолды, полученные в результате работы предудущих этапов метода, содержат сравнительно небольшое число ошибочных связей, то определение взаимной ориентации контигов в скэффолдах с использованием ребер, входящих в результирующие скэффолды, также должно иметь высокую точность.

Глава 3. Применение сборщика скэффолдов геномных последовательностей

Предлагаемый метод сборки скэффолдов был имплементирован на языке программирования *Java*. Метод также был протестирован на различных данных, соответствующих геному бактерии *Escherichia coli*, геномная последовательность которой практически полностью известна [12]. Это позволяет сравнивать результаты сборки скэффолдов с результатами работы других сборщиков по различным параметрам.

В разделе 3.1.1 описаны данные, подающиеся сборщику скэффолдов на вход. В разделе 3.1.2 описаны используемые в работе оценки качества полученных скэффолдов. Раздел 3.2.1 содержит описание данных, на которых проводилась аппробация разработанного метода. Раздел 3.2.2 содержит анализ полученных результатов, а также сравнение их с результатами других сборщиков. Разделы 3.3 и 3.4 содержат рекомендации по внедрению и улучшению предлагаемого метода.

3.1. Тестирование сборщика скэффолдов

К сожелению, не существует стандартизированного метода тестирования сборщиков скэффолдов. Общепринятой оценки результатов сборки скэффолдов также не существует. В данном разделе описаны используемые сборщиком данные, а также использованные в работе методы оценки.

3.1.1. Входные данные

Разработанный и имплементированный метод получает на вход контиги, результаты картирования парных чтений на контиги, а также приблизительную длину геномной последовательности, которая обычно известна заранее. В случае, если приблизительная длина не известна, она аппроксимируется суммой длин контигов, поданных на вход. При тестировании метода для картирования чтений использовалась сторонняя программа *Bowtie* [13]. Также для оценки результатов, сборщика, в качестве

дополнительного параметра можно передать результаты картирования контигов на референсный геном. При тестировании предлагаемого метода использовалась программа *BLAST* [14] для картирования контигов на референсную геномную последовательность бактерии *E. Coli*. При передачи этого параметра происходит подсчет дополнительной статистики, позволяющей определить качество сборки на каждом этапе.

3.1.2. Параметры оценки результатов сборки

К сожелению, не существует единых общепринятых параметров, позволяющих адекватно оценить результат сборки скэффолдов. Довольно часто используется величина *N50*, которая, однако, не полностью отражает качество сборки. Так, например, объединив все скэффолды в один можно сильно увеличить *N50*, однако при этом в результат сборки будет привнесено существенное число грубых ошибок.

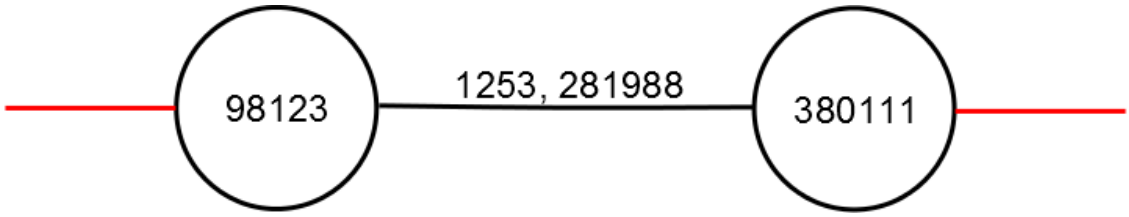

Рис. 14. Пример ошибки использования неправильного ребра. В вершинах обозначены местоположения контигов в геномной последовательности. На ребрах обозначены оценки на расстояния и реальные расстояния между контигами.

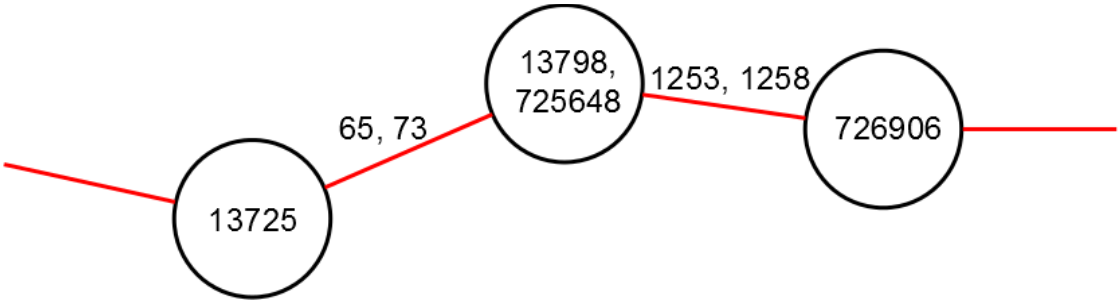

Рис. 15. Пример ошибки «перескок через повтор». В вершинах обозначены местоположения контигов в геномной последовательности. На ребрах обозначены оценки на расстояния и реальные расстояния между контигами.

При наличии референсного генома появляется возможность более точно оценить результат сборки. Прежде всего, можно вычислить число ошибок, допущенных при построении скэффолдов. При тестировании метода выявлялись два вида ошибок: использование ошибочной связи (рис. 14) и «перескок через повтор» (рис. 15). Первая ошибка соответствует тому, что в скэффолде рядом расположены контиги, находящиеся на большом расстоянии друг от друга в референсной геномной последовательности. Второй род ошибок возникает из-за наличия повторов в геномной последовательности. Если контиг картируется более чем в одно место генома, то скэффолд может прийти в этот контиг в одном месте его картирования, а уйти в другом, тем самым через такой контиг соединяются контиги располагающиеся на значительном расстоянии друг от друга. Отдельно стоит выделить еще один вид ошибки – нарушение порядка двух контигов – случай, когда два соседних контига идут в скэффоде в неправильном порядке (рис. 16). Такая ошибка обычно встречается из-за неправильного расположения в скэффолде короткого контига.

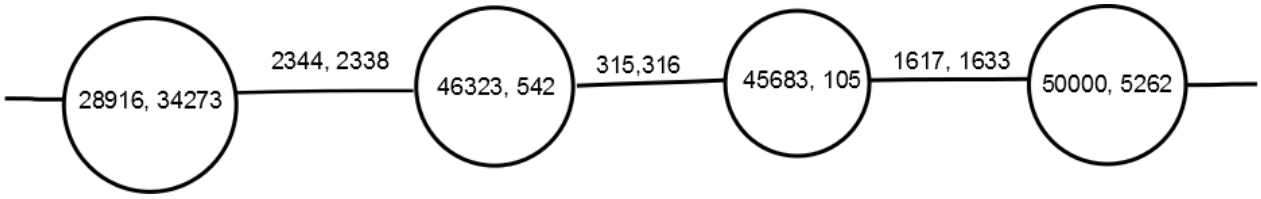

Рис. 16. Пример ошибки неправильно расположения контигов в скэффолде. В вершинах обозначены местоположения контигов в геномной последовательности и их длины. На ребрах обозначены оценки на расстояния и реальные расстояния между контигами. Два центральных контига должны идти в обратном порядке.

Помимо подсчета их числа, выявление ошибок позволяет посчитать величину *N50* для «исправленных» скэффолдов (*N50cut*), то есть величину *N50,* получающуюся при разрезании скэффолдов в местах ошибок. Эта величина лучше отображает реальное качество сборки скэффолдов, чем просто величина *N50*. Однако, стоит заметить, что для подсчета величины N50cut требуется

наличие референсной геномной последовательности, в то время как величина *N50* зависит лишь от набора скэффолдов.

Для оценки метода оценки расстояния считается среднее отклонение метода от истинного отклонение, а также среднеквадратичное отклонение отклонения. При сравнении с другими методами также возможно посчитать число связей, на которых оценка была точнее, чем у сравниваемого метода.

Для определения качества определения взаимной ориентации контигов в скэффолде подсчитывалось число ошибочно ориентированных контигов (из рассчета, что большинство контигов ориентированны правильно), а также число ошибочных смен ориентации – число ситуаций, когда ориентации соседних контигов в скэффолде не согласуются.

Все описанные параметры позволяют более комплексно оценивать результаты сборки скэффолдов и избежать простой «накачки» используемого параметра оценки.

3.2. Результаты сборки

В данном разделе описаны использованные для тестирования данные, а также произведена оценка результатов работы предложенного метода и сравнение этих результатов с результатами работы других сборщиков скэффолдов.

3.2.1. Тестовые данные

Для тестирования описанного метода использовался геном бактерии *Escherichia coli*. Геном этой бактерии полностью известен, что дает возможность использовать референсную геномную последовательность для оценки и сравнения результатов сборки скэффолдов. Геном *E.Coli* содержит ~4.65 миллионов нуклеотидов. Для тестирования использовались различные наборы контигов, полученных сборщиком контигов, разработанным в лаборатории «Алгоритмы сборки геномных последовательностей» в НИУ ИТМО, на основе реальных данных. Параметры наборов контигов можно увидеть в таблице 1. Для одного из наборов имелась референсная геномная

последовательность, что позволяет более детально проанализировать результаты работы сборщиков скэффолдов. Использовались как искусственные парные чтения, полученные с помощью стороннего средства *MetaSim* [15], так и реальные библиотеки парных чтений, полученные с помощью секвенатора *Illumina* [16]. Параметры наборов чтений приведены в таблице 2.

Таблица 1. Параметры тестовых наборов контигов.

Набор	Число контигов	Суммарная длина	Максимальная длина	*N50*
Reference	502	4581252	73908	18047
Illumina	424	4638964	195661	49106

Таблица 2. Параметры тестовых библиотек чтений.

Набор	Число пар чтений	Длина чтений	Средняя длина фрагмента	Стандартное отклонение
MetaSim Set1 – Set3	600000	36	3000	300
MetaSim Set5000	600000	36	5000	500
Illumina	2500000	35	6000	800

3.2.2. Оценка полученных результатов и сравнение с другими методами

Результаты работы предложенного метода на тестовых данных сравнивались с результатами работы других методов на тех же входных данных. Различные параметры получившихся наборов скэффолдов можно увидеть в таблицах 3–9.

Таблица 3. Среднее отклонение оцененного расстояния от реального, а также стандартное отклонение ошибки.

Набор	*SOPRA*	*GAPEST*	*OPERA*	Предложенный метод
MetaSim Set1	209±15	175±14	187±14	149±13
MetaSim Set2	139±13	217±17	152±15	135±13
MetaSim Set3	215±15	172±14	203±14	153±13
MetaSim Set5000	190±14	195±12	191±13	120±11
Illumina	Не известно	Не известно	Не известно	Не известно

Таблица 4. Процент ситуаций, когда оценка предложенным методом ближе к реальному значению расстояния между контигами и когда оценки методов совпадают.

Набор	Предложенный метод vs *SOPRA*	Предложенный метод vs *GAPEST*	Предложенный метод vs *OPERA*
MetaSim Set1	68% / 6%	48% / 5%	62% / 6%
MetaSim Set2	63% / 9%	65% / 2%	59% / 7%
MetaSim Set3	64% / 10%	62% / 3%	61% / 9%
MetaSim Set5000	63% / 10%	77% / 2%	64% / 5%
Illumina	Не известно	Не известно	Не известно

Качество оценки расстояния между контигами сравнивалось с распространненым методом среднего арифметического, применяемым в сборщиках скэффолдов *SOPRA* и *OPERA*. Также описанный метод оценки расстояния сравнивался с методом *GAPEST*. Стоит отметить, что метод *OPERA*, хоть и использует метод среднего арифметического для оценки расстояний между контигами, производит оценку расстояний между несколькими контигами одновременно, что позволяет получать более точные оценки.

Результаты показывают, что оценка расстояния, полученная предалаемым методом, в среднем меньше отклонена от истинного значения (табл. 3) и чаще находится ближе к реальному расстоянию (табл. 4), чем оценки других методов. Эти данные покакзывают, что разработанный метод точнее оценивает расстояние между контигами, что должно приводить к лучшему качеству скэффолдов.

Таблица 5. Параметры собранных скэффолдов для метода *OPERA*.

Набор	*OPERA*		
	N50	*N50cut*	Число ошибок
MetaSim Set1	366.5k	215.7k	11
MetaSim Set2	428.6k	215.7k	11
MetaSim Set3	465.0k	294.0k	11
MetaSim Set5000	887.5k	618.6k	7
Illumina	315.2k	Не известно	Не известно

Анализ результатов сборки скэффолдов (табл. 5–7) показывает, что предлагаемый метод совершает меньше ошибок при сборке в сравнении с другими методами. Также результирующие наборы скэффолдов выигрывают по параметру *N50cut*. Однако, результаты показывают, что предлагаемый метод иногда проигрывает конкурентам по величине *N50*. Это объясняется крайне малым числом ошибок в наборе скэффолдов, полученном разработанным методом, и большим числом ошибок в скэффолдах других методов. Как было показано выше, за счет увеличения числа ошибок достаточно просто сильно увеличить величину *N50*, что, однако, негативно сказывается на общем качестве сборки. Провести сравнение числа ошибок для данных набора Illumina не представляется возможным из-за отсутствия референсной геномной последовательности. Стоит заметить, что на этих данных все методы показывают более слабые результаты, а метод *SOPRA* и вовсе не собирает ни одного скэффолда.

Таблица 6. Параметры собранных скэффолдов для метода *SOPRA*.

Набор	*SOPRA*		
	N50	*N50cut*	Число ошибок
MetaSim Set1	262.3k	262.3k	3
MetaSim Set2	630.4k	215.7k	13
MetaSim Set3	430.0k	282.3k	10
MetaSim Set5000	375.0k	320.6k	8
Illumina	49.1k	Не известно	Не известно

Для оценки определения взаимной ориентации подсчитывалось как число ошибочно ориентированных контигов , а также число ошибочных смен ориентации – число ситуаций, когда ориентации соседних контигов в скэффолде не согласуются (табл. 8). Как видно из результатов, предложенные метод совершает меньше ошибок неправильной смены ориентации. Чаще всего такие ошибки происходят при объединении контигов в скэффолды с помощью ошибочных ребер. При разрезании скэффолдов в местах ошибок сборки, у всех методов исчезают ошибки ориентации контигов.

Таблица 7. Параметры собранных скэффолдов для предлагаемого метода.

Набор	Предложенный метод		
	N50	*N50cut*	Число ошибок
MetaSim Set1	311.7k	311.7k	1
MetaSim Set2	605.3k	392.0k	7
MetaSim Set3	578.2k	322.6k	9
MetaSim Set5000	639.3k	639.3k	6
Illumina	120.7k	Не известно	Не известно

Помимо результирующих скэффолдов, алгоритмы также сравнивались по времени работы. Все алгоритмы запускались на одной и той же машине с параметрами *AMD Opteron(TM) Processor 6272 2x16 2.1 GHz, 128 GB RAM*.

Таблица 9 содержит времена работы методов на различных входных данных. Результаты показывают, что предложенный метод значительно выигрывает в быстро действии у других методов. При этом стоит отметить, что во время работы предлагаемого метода также включено время работы сторонней программы *Bowtie*, используемой для картирования парных чтения на контиги. Во время работы метода *SOPRA*, который также использует результаты работы программы *Bowtie*, время работы сторонней программы включено не было. Стоит отметить, что на наборе данных *Illumina* предложенный метод уступает методу *SOPRA* в быстродействии. Это связано с тем, что на этом наборе данных метод *SOPRA* не построил ни одной связи между контигами, а потому быстро завершил свою работу.

Таблица 8. Число ошибочно ориентированных контигов / число ошибочных смен ориентации.

Набор	*SOPRA*	*OPERA*	Предложенный метод
MetaSim Set1	3 / 1	83 / 8	0 / 0
MetaSim Set2	93 / 10	72 / 6	54 / 5
MetaSim Set3	82 / 8	103 / 9	89 / 7
MetaSim Set5000	27 / 3	63 / 6	70 / 6
Illumina	Не известно	Не известно	Не известно

Таблица 9. Время (в минутах) работы методов сборки скэффолдов.

Набор	*SOPRA*	*OPERA*	Предложенный метод
MetaSim Set1	48	18	1.5
MetaSim Set2	45	21	1.5
MetaSim Set3	47	23	1.5
MetaSim Set5000	48	21	2
Illumina	5	52	5.5

3.3. Рекомендации по внедрению

Предлагаемый метод рекомендуется применять для сборки скэффолдов геномных последовательностей, с использованием парных чтений с небольшой средней длинной фрагмента. К сожалению, в тех случаях, когда средняя длина фрагмента значительно превышает длины многих контигов, графы контигов и скэффолдов начинают иметь сложную форму с множеством тупиков, а также многие контиги, расположенные рядом в геномной последовательности, оказываются не связаны никакими чтениями. Все это может приводить к плохим результатам сборки.

Стоит отметить, что разработанный метод оценки расстояний способен работать с любыми библиотеками парных чтений, а также с разными сочетаниями библиотек с различными параметрами. Это позволяет использовать предложенный метод оценки совместно с другими методами сборки скэффолдов. Для сборщиков скэффолдов, чувствительных к оценке расстояния между контигами, использование предлагаемого метода оценки может существенно улучшить качество финального набора скэффолдов.

Введенная в работе вероятность случайной пары чтений связать пару контигов, а также разработанный и связанный с этой вероятностью критерий определения ошибочных ребер, могут быть использованы в других сборщиках для упрощения используемых графов контигов и/или скэффолдов, а также выявления ошибочных связей в уже собранных скэффолдах.

3.4. Рекомендацию по улучшению

На данном этапе предлагаемый метод сборки скэффолдов не всегда выдает хорошие результаты при использовании библиотек парных чтений с большой средней длинной фрагмента. Также, хоть сочитание различных библиотек и приводит к улучшению оценки расстояния, оно приводит к усложнению используемых при сборке графов, что может негативно сказываться на качестве результирующего набора скэффолдов. Поэтому, необходимо улучшить работу метода с библиотеками чтений с средней

длинной фрагмента, значительно превышающей длины некоторых контигов, а также научиться сочитать работу с парными чтениями таких библиотек с использованием парных чтений библиотек с другими параметрами.

Выводы

Был разработан и реализован метод сборки скэффолдов геномной последовательности на основе принципа максимального правдоподобия. Входными данными метода являются набор контигов – длинных частей генома, а также библиотека парных чтений – небольших фрагментов геномной последовательности, для которых известно примерное расстояние между ними. Результатом работы метода является набор скэффолдов – множества упорядоченных ориентированных контигов с известными расстояниями между ними.

Для оценки расстояния между контигами используется принцип максимального правдоподобия и новую функцию правдоподобия, учитывающую не только пары чтений, связывающие контиги, но и несвязывающие пары чтений, а также длины контигов. Для определения взаимного расположения контигов вначале строится упрощенный граф контигов, позволяющий построить первое приближение скэффолдов. Далее полученные скэффолды объединяются в большие скэффолды с помощью поиска кратчайших путей между ними в графе скэффолдов.

Разработанный метод был реализован и протестирован на примере генома бактерии *E.Coli*. Результаты работы метода были сравнены с результатами других распространенных сборщиков скэффолдов. Анализ полученных оценок на расстояния показывает, что разработанный метод оценки имеет меньшее среднее отклонение от истинного расстояния, чем другие распространенные методы. Сравнение различных параметров результирующих наборов скэффолдов показывает, что предложенный метод допускает меньшее число ошибок при сборке, а также конструирует скэффолды с большей величиной *N50*.

Предлагаемый метод собирает скэффолды обладающие лучшим качеством по сравнению с другими распространенными сборщиками на тестовых данных. Это указывает на перспективность развития и использования прелагаемого метода сборки скэффолдов геномных последовательностей.

Разработанный метод рекомендуется использовать как один из этапов сборки генома. Разработанный метод оценки расстояний рекомендуется использовать как для повышения качества уже построенных скэффолдов, так и для улучшения результатов работы других методов сборки.

Список используемых источников

1. Talking glossary of genetic terms: genome // National Human Genome Research Institute, http://www.genome.gov/Glossary/
2. *Alberts B., Johson A., Lewis J., Raff M., Roberts K., Walter P.* Molecular Biology of the Cell; Fourth Edition // New York: Garland Science, 2002
3. *Watson J., Crick F.* Molecular structure of nucleic acids; a structure of deoxyribose nucleic acid // Nature, no. 171, 1953
4. *Anderson S.* Shotgun DNA sequencing using cloned DNase I-generated fragments // Nucleic Acids Res., no. 9(13), 1981, pp 3015-3027
5. http://en.wikipedia.org/wiki/N50_statistic
6. *Gao S., Sung W.-K., Nagarajan N.* Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences // Journal of Computational Biology, Vol. 18, 2001, pp. 1681-1691
7. *Huson D.H., Reinert K., Myers E.W.* The greedy path-merging algorithm for contig scaffolding // Journal of ACM, vol. 49, No. 5, 2002, pp. 603-615
8. *Sahlin K., Street N., Lundeberg J., Arvestad L.* Improved gap size estimation for scaffolding algorithms // Bioinformatics, Vol. 28, No. 17, 2012, pp. 2215-2222
9. *Dayarian A., Michael T.P., Sengupta A.M.* SOPRA: Scaffolding algorithm for paired reads via statistical optimization // BMC Bioinformatics, No. 11, 2010, p. 345
10. *Garey M.R., Johnson D.S:* Computers and intractability: a guide to the theory of NP-completeness // San Francisco: W. H. Freeman, 1979
11. *Barahona F:* On the computational complexity of Ising spin glass models // Journal of Physics A: Mathematical and General, no. 15, 1982, pp. 3241-3253
12. *Blattner F.R., Plunkett G., Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B., Shao Y.* The complete genome sequence of Escherichia coli K-12 // Science, no. 277 (5331), 1997, pp. 1453–1462.

13. *Langmead B., Trapnell C., Pop M., Salzberg S.L.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome // Genome Biology, no. 10, 2009
14. *Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J.* Basic Local Alignment Search Tool // Journal of Molecular Biology, no. 215(3), 1990, pp. 403-410
15. *Richter D.C., Ott F., Auch A.F., Schmid R., Huson D.H.* MetaSim—A Sequencing Simulator for Genomics and Metagenomics // PLoS, vol. 3, no. 10, 2008
16. Illumina, Inc. http://www.illumina.com/