Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики

Факультет информационных технологий и программирования
Кафедра компьютерных технологий

Васин Андрей Юрьевич

Сборка скэффолдов геномной последовательности на основе принципа максимального правдоподобия

Научный руководитель: доцент кафедры КТ Ф. Н. Царев

Санкт-Петербург
2014

Содержание

Введение

Сборка геномной последовательности одна из важнейших задач биоинформатики. Задача заключается в восстановлении цепи ДНК по набору ее чтений. В данном процессе приходится столкнуться с такими проблемами как большое количество ошибок в исходных данных и большой их объем, составляющий в некоторых случаях десятки и сотни гигабайт.

Обычно процесс сборки геномной последовательности разбивается на три части:

- исправление ошибок в чтениях,
- сборка **контигов**,
- сборка **скэффолдов**, наборов упорядоченных и ориентированных контигов с оценками дистанции между ними.

В идеальном случае результатом сборки будет являться единственных скэффолд, состоящий из одного контига, что будет соответствовать исходной цепи ДНК. Такого результата не удается достичь из-за вышеупомянутых ошибок в чтениях и строения геномной цепи, включающей в себя множество повторяющихся фрагментов.

Построение скэффолдов также состоит из трех частей:

- оценка расстояния между контигами,
- определение ориентации контигов,
- определение положений контигов в скэффолде.

Для этого часто используют информацию о **парных чтениях** — парах небольших фрагментов геномной цепи с произведенной оценкой дистанции между ними. Одной из существующих проблем методов сборки скэффолдов является низкая точность оценки расстояния между ними.

В данной работе был разработан метод более точной оценки расстояния между множеством контигов на основе принципа максимального

правдоподобия, а также усовершенствован метод сборки геномной последовательности, предложенный в работе [1].

Сборка скэффолдов начинается с построения графа контигов, вершины которого соответствуют контигам, а ребра — парам чтений, соединяющих контиги. Затем производится фильтрация ребер, после чего скэффолды создаются с помощью нахождения путей в данном графе.

Для определения ориентации скэффолдов используется информация о картировании пар чтений.

Метод, представленных в данной работе, делает возможной сборку скэффолдов геномной последовательности, базируясь на информации о парных чтениях. Результирующие скэффолды содержат улучшенные оценки на расстояния между контигами по сравнению с существующими аналогами.

Глава 1. Обзор предметной области

1.1. Основные определения

Как не трудно догадаться, биоинформатика включает в себя проблемы, относящиеся сразу к информатике и биологии. Отличительной чертой таких проблем является работа с большими объемами данных. Одной из таких задач и является сборка геномной последовательности.

1.1.1. Строение геномной последовательности

Геном — совокупность наследственного материала, заключенного в клетке организма [2]. Геном содержит биологическую информацию, необходимую для построения и поддержания организма. Большинство геномов, в том числе геном человека и геномы всех остальных клеточных форм жизни, построены из **дезоксирибонуклеиновой кислоты** (ДНК). ДНК в свою очередь является полимерной молекулой, состоящей из нуклеотидов [3]. Нуклеотиды, содержащиеся в ДНК, принадлежат одной из четырех групп в зависимости от азотистого основания: **аденин** (A), **гуанин** (G), **цитозин** (C) и **тимин** (T).

Молекула ДНК в регулярном случае представляет собой спираль [4], состоящую из двух цепей, расположенных азотистыми основаниями друг к другу. Соединения азотистых оснований соответствуют **принципу комплиментарности**, согласно которому цитозин соединяется с гуанином, а аденин с тимином.

Геном в биоинформатике представляется последовательностью нуклеотидов одной из комплементарных цепей молекулы ДНК, из которой можно получить вторую цепь воспользовавшись принципом комплиментарности. Строка, состоящая из символов A, G, C и T, соответствующих типам нуклеотидов, является наиболее удобным представлением геномной цепи.

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) [5]. Метод состоит в выделении из молекулы ДНК коротких участков (порядка нескольких сотен последовательных нуклеотидов), после чего происходит посимвольное считывание концов выделенных участков (Рис. 1.1). 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.

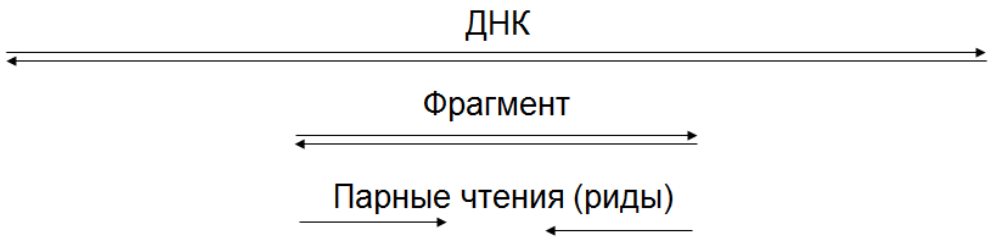

Рис. 1.1. Процесс выделения пары чтений из молекулы ДНК

1.2. Постановка задачи

Как было упомянуто ранее, процесс сборки геномной последовательности состоит из трех этапов: исправление ошибок в парных чтениях, построение контигов, длинных последовательных частей геномной последовательности, и построение скэффолдов, наборов упорядоченных ориентированных контигов с оценками расстояний между соседними контигами. Из выше сказанного вытекает задача данной работы: построение скэффолдов по множеству контигов и набору парных чтений.

Задача сборки скэффолдов разбивается на три этапа (Рис. 1.2): оценка расстояния между контигами, определение взаимного порядка контигов и определение ориентации контигов. Но в отличии от классической схемы построения скэффолдов в данной работе производится дополнитель-

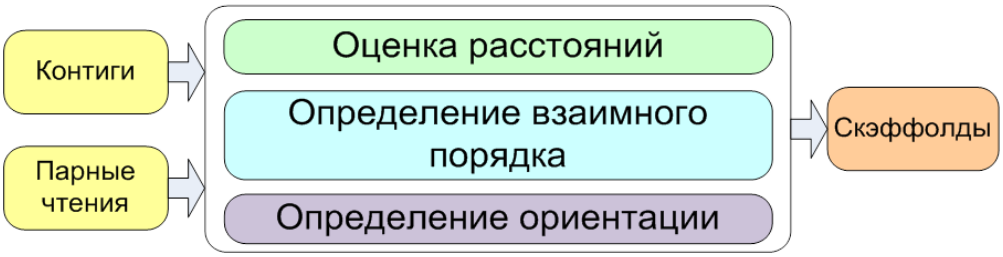

Рис. 1.2. Схема работы методов построения скэффолдов геномных последовательностей по набору контигов и парных чтений

ное уточнение расстояний между контигами в уже собранных скэффолдах, что позволяет повысить точность оценок.

1.3. Используемые обозначения

За геномную последовательность в данной работе предлагается считать последовательность символов A, C, G и T, сопоставленных нуклеотидам и расположенных в порядке, соответствующем следованию соответствующих нуклеотидов в восстанавливаемой цепи ДНК.

Для определения качества сборки часто применяется величина **N50** [6] — максимальное число такое, что общая длина всех контигов, имеющих длину не менее данного числа превышает половину суммы длин всех контигов. По аналогии также используют величину **N90**, только в данном случае сумма контигов длиннее данного числа должна составить не менее 90от суммы длин всех контигов.

Величина N50 для скэффолдов определяется схожим образом, где за длину скэффолда принимается сумма длин всех его контигов.

1.4. Обзор существующих методов

В данном разделе будут вкратце описаны существующие методы, с которыми в дальнейшем будет произведено сравнение результатов. Данная работа концентрируется вокруг оценки расстояний между контигами, что является причиной чуть более подробного рассмотрения способов оценки

расстояний в уже существующих методах.

1.4.1. SOPRA

Сборщик скэффолдов геномной последовательности SOPRA [10] в процессе сборки осуществляет следующие шаги:

1. Определение взаимной ориентации контигов;
2. Оценка расстояний между контигами;
3. Определение взаимного расположения контигов в скэффолдах.

Расстояние в данном методе оценивается изходя из распространенного предположения о нормальности распределения длин фрагментов, из которых были получены парные чтения. Согласно свойствам нормального распределения оптимальным значением расстояния будет являться среднее арифметическое по оптимальным расстояниям для каждой пары чтений.

$$l_{ij} = \frac{1}{J_{ij}} \sum_{k=1}^{J_{ij}} (\mu - d_{ik} - d_{jk} - 2r) \tag{1.1}$$

В данной формуле l_{ij} — расстояние между контигами i и j, J_{ij} — количество связывающих чтений, μ — средняя длина фрагмента, r — длина чтения, d_{ik} — расстояние от места картирования k-го чтения на i-й контиг до края контига.

1.4.2. GAPEST

GAPEST [9] является методом оценки расстояний между контигами, основанном на методе максимального правдоподобия. Данный метод также использует распространенное предположение о нормальности длин фрагментов. Особенность метода заключается в учете некоторой смещенности распределения длин фрагментов связывающих чтений.

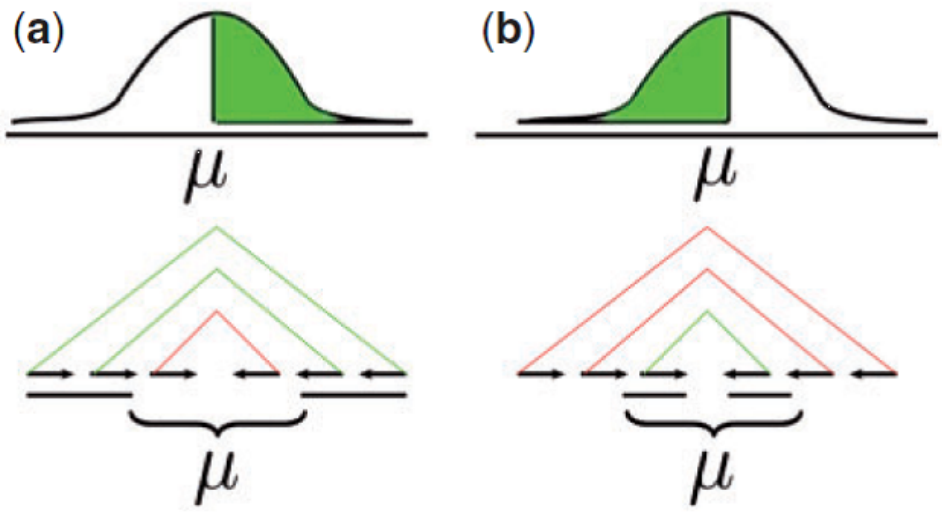

Рис. 1.3. Наблюдаемая смещенность в условном распределении длин фрагментов. a) наблюдаются только пары чтений с большой длинной фрагмента; b) наблюдаются пары чтений только с маленькой длинной фрагмента.

Для учета смещений в методе GAPEST вводят распределение длин фрагментов парных чтений, которые связывают контиги длин c_1 и c_2, находящиеся на расстоянии d.

$$h(x|d,c_1,c_2) = \frac{p(x|d,c_1,c_2)f(x)}{\int\limits_{-\infty}^{+\infty} p(y|d,c_1,c_2)f(y)dy} \tag{1.2}$$

Знаменатель в данной формуле необходим для нормализации, чтобы получить именно вероятностное распределение. За $f(x)$ обозначена вероятность получения пары чтений с длиной фрагмента x. Согласно общепринятому предположению эта величина подчиняется нормальному распределению. Величина $p(x|d,c_1,c_2)$ является вероятностью парного чтения с длинной фрагмента x связать пару контигов длин c_1 и c_2, находящиеся на расстоянии d.

Поиск наиболее вероятного расстояния между контигами осуществляется путем максимизации функции правдоподобия. Функция правдоподобия представляет собой произведение вероятностей наблюдения чтений с имеющимися длинами фрагментов.

$$d_{GAPEST} = \underset{d}{\operatorname{argmax}} \prod_{i=1}^{n} h(x_i|d, c_1, c_2) \tag{1.3}$$

Одним из недостатков данного метода является сложность его реализации, которая заключается в том, что числитель и знаменатель в формуле (1.2) на практике оказываются крайне малыми величинами, что может приводить к большой ошибке в вычислениях. Также оптимизация данной функции представляется затруднительной, так как она имеет множество локальных максимумов.

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

1.4.3. OPERA

Сборщик скэффолдов геномной последовательности OPERA [7] в процессе сборки осуществляет следующие шаги:

1. Определение взаимной ориентации контигов;
2. Разбиение контигов на скэффолды;
3. Определение взаимного расположения контигов в скэффолдах;
4. Оценка расстояний между контигами.

При оценки расстояний OPERA использует метод максимального правдоподобия. Опять же используется распространенное предположение о нормальности распределения длин фрагментов, из которых получаются парные чтения. Используемая функция правдоподобия представляет собой произведение вероятностей получения длин фрагментов, из которых были получены связывающие чтения. Затем эта функция правдоподобия максимизируется средним арифметическим.

Глава 2. Разработка сборщика скэффолдов геномных последовательностей

Представленный метод сборки скэффолдов, в отличии от многих других, состоит из четырех частей:

1. Первичная оценка расстояний между контигами;
2. Определение взаимного порядка контигов в скэффолдах;
3. Определение взаимной ориентации контигов;
4. Уточнение оценки расстояний между контигами в скэффолдах.

Так как первые три пункта сборки скэффолдов повторяют соответствующие пункты из работы [1], здесь будет приведено их сжатое изложение. Первый пункт будет описан чуть более подробно для упрощения понимания четвертого пункта.

2.1. Первичная оценка расстояния на основе принципа максимального правдоподобия

Метод оценки расстояния основан на принципе максимального правдоподобия. Функция учитывает не только пары чтений, связывающих пару контигов, а также длины контигов и число несвязывающих чтений.

Метод 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

Одним из элементов, легших в основу формулы правдоподобия, является распространенное предположение о нормальности распределения длин фрагментов, из которых получаются парные чтения. Также задей-

ствовано предположение, что парные чтения получаются следующим образом:

1. Выбирается длина фрагмента согласно нормальному распределению, из которого будет получена пара чтений;
2. Равновероятно выбирается положение фрагмента в геномной последовательности;
3. Считываются концы выбранного фрагмента, которые и являются итоговой парой чтений.

Экспериментальные данные подтверждают высокую степень достоверности данных предположений. Базируясь на данной модели, становится возможно получить функцию правдоподобия, обеспечивающую более точную оценку расстояния между парой контигов, чем у имеющихся аналогов.

2.1.2. Параметры модели

Параметры нормального распределения длин фрагментов можно получить из парных чтений, при условии, что оба чтения были скартированны на один и тот же контиг. В данном случае оценка параметров производится стандартными методами статистики. Математическим ожиданием длины фрагмента является среднее арифметическое длин фрагментов, а стандартным отклонением среднеквадратичное отклонение длин фрагментов от математического ожидания.

2.1.3. Учет связывающих чтений

Связывающее чтение — парное чтение, обе части которого попадают на разные контиги. Для их учета в функцию правдоподобия включается вероятность их получения. Про каждую пару связывающих чтений известно место их картирования на контиги (Рис. 2.1).

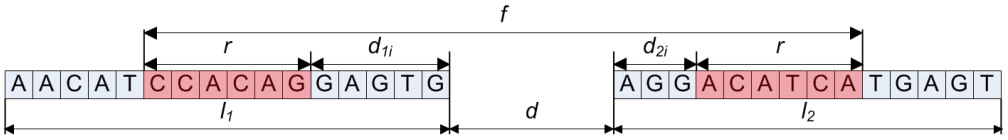

Рис. 2.1. Пример парного чтения, связывающего контиги, где l_1, l_2 – длины контигов; d_1, d_2 – расстояния от мест картирования чтений до краев контигов; r – длина чтения; f – длина фрагмента, из которго получена пара чтений; d – предполагаемое расстояние между контигами

Вклад каждого чтения в используемую функцию правдоподобия состоит из двух частей:

1. Вероятность получить фрагмент имеющейся длины

$$p(fragment\_size|d) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(d_{1i} + d_{2i} + d + 2r - \mu)^2}{2\sigma^2}) \quad (2.1)$$

2. Вероятность получить имеющееся местоположение фрагмента

$$p(fragment\_position|d) = \frac{1}{L} \quad (2.2)$$

где L предполагаемая длина геномной последовательности.

Суммарный вклад каждого чтения состоит из произведения эти двух вероятностей. Вклад всех связывающих чтений состоит из произведения вероятностей получения чтений, где n количество связывающих чтений:

$$p(connecting|d) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma L} exp(-\frac{(d_{1i} + d_{2i} + d + 2r - \mu)^2}{2\sigma^2}) \quad (2.3)$$

2.1.4. Учет несвязывающих чтений

Несвязывающие чтения — любое парное чтение, не являющиеся связывающим. Кроме связывающих чтений, имеются также чтения, которые рассматриваемые контиги не связывают. Данный метод отличается тем, что он учитывает вклад и таких чтений в используемой функции правдоподобия.

Для учета несвязывающих парных чтений вычисляется вероятность случайной пары чтений связать рассматриваемую пару контигов. Данная вероятность вычисляется с использованием формулы полной вероятности,

где параметром служит длина фрагмента, из которой получается пара чтений.

$$p(random\_connect|d) = \sum_f p(f|d)p(connect|f,d) \tag{2.4}$$

Зафиксировав длину фрагмента, вероятность получения такой длины фрагмента вычисляется согласно нормальному распределению. Для нахождения вероятности пары чтений с длиной фрагмента f связать пару контигов, необходимо найти количество позиций, в которых пара чтений будет связывать рассматриваемую пару контигов.

Обозначим число таких позиций за $w_d(f)$ по аналогии с работой [9], где она использовалась. Таким образом вероятность получения парного чтения в одной из таких позиций $\frac{w_d(f)}{L}$. В итоге получаем результирующую формулу для вероятности случайной пары чтений связать пару контигов.

$$p(random\_connect|d) = \sum_f \frac{w_d(f)}{\sqrt{2\pi}\sigma L} exp(-\frac{(\mu - f)^2}{2\sigma^2}) \tag{2.5}$$

Но для учета вклада несвязывающих чтений нужна не вероятность случайной пары чтений связать пару контигов, а обратная ей величина, то есть вероятность случайной пары чтений не связать пару контигов $1 - p(random\_connect|d)$. Обозначим суммарное количество парных чтений за R, количество связывающих чтений n и получим результирующий вклад всех несвязывающих чтений в функцию правдоподобия:

$$p(non - connecting|d) = (1 - p(random\_connect|d))^{R-n} \tag{2.6}$$

2.1.5. Нахождение наиболее вероятного расстояния

Функция правдоподобия, составленная из частей (2.3) и (2.6), максимизируется для нахождения наиболее вероятного расстояния.

Для упрощения работы с функцией правдоподобия используется ее логарифм. Часть, отвечающая за учет связывающий чтений, преобразуется

в многочлен от d. В данном многочлене возможно предпосчитать коэффициенты, что позволит вычислять вклад связывающего чтения за $O(1)$ для произвольного d.

$$\log p(connecting|d) = n \log \frac{1}{\sqrt{2\pi}\sigma L} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (d_{1i} + d_{2i} + d + 2r - \mu)^2 \quad (2.7)$$

При логарифмировании формула, отвечающая за вклад несвязывающих чтений, не приобретает простого вида. Для упрощения вычисления вклада несвязывающих чтений сумма из формулы (2.5) аппроксимируется интегралом.

$$p(random\_connect|d) \approx \int_0^\infty \frac{w_d(f)}{\sqrt{2\pi}\sigma L} exp(-\frac{(\mu - f)^2}{2\sigma^2}) df \quad (2.8)$$

Функция $w_d(f)$ имеет вид трапеции и почти всюду равна нулю, что позволяет разбить данный интеграл на три. Каждый из них вычисляется за $O(1)$ при использовании табличных значений для нормального распределения.

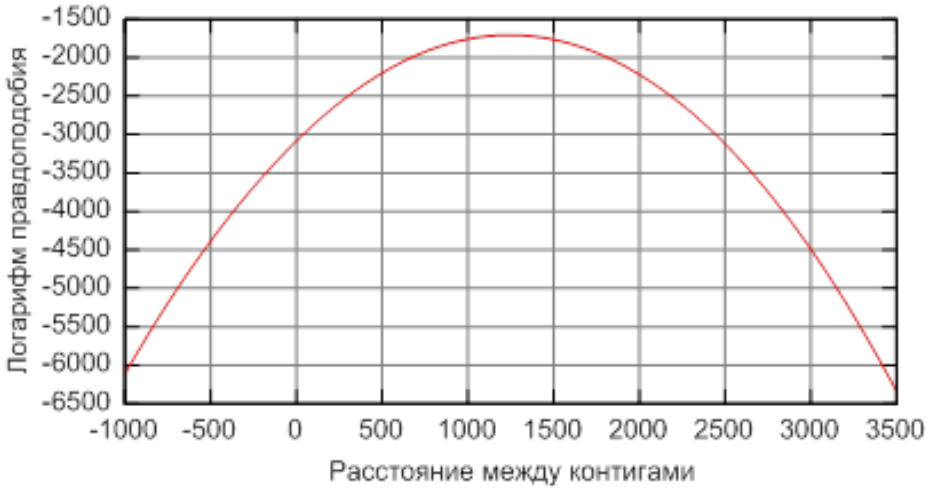

Рис. 2.2. Пример графика логарифма функции правдоподобия в зависимости от расстояния между контигами

В итоге логарифм функции правдоподобия вычисляется за $O(1)$ при условии предподсчета, проводимого за $O(n)$. Для поиска максимума правдоподобия используется тернарный поиск, осуществляющий вычисление логарифма функции правдоподобия$O(\log \mu)$ раз. Суммарное время поиска оптимального расстояния между парой контигов на данном этапе составляет $O(n + \log \mu)$.

2.2. Определение взаимного порядка

После первичной оценки расстояний происходит разбиение контигов на скэффолды, для чего сначала строится граф контигов, по которому строится первое приближение скэффолдов. Далее строится граф скэффолдов, который позволяет объединить сравнительно малые скэффолды, полученные при первом приближении, в большие.

При построении графа контигов в качестве вершин берутся сами контиги, а ребра соответствуют парным чтениям, связывающих контиги.

Каждому ребру соответствует его вес, который представляет собой оценку на расстояние между парой контигов. Первое приближение скэффолдов строится с использованием жадного алгоритма, который находит простые пути, состоящие из максимально возможного числа вершин, при этом имеющие наименьший суммарный вес ребер.

Дальше строится граф скэффолдов, где ребра соединяют концы скэффолдов, то есть крайние контиги в скэффолдах. Данный граф состоит из скэффолдов, полученных на предыдущем этапе, а также скэффолдов, состоящих из одного контига, который не удалось включить в какой-либо скэффолд на предыдущем этапе. Объединение скэффолдов опять происходит согласно принципу кратчайшего пути, после каждой итерации объединения граф перестраивается.

2.3. Определение ориентации

Определение ориентации проходит с использованием ребер из графа контигов, соединяющих контиги из одного и того же скэффолда, который рассматривается в данный момент. Контиги скэффолда обходятся в глубину с использованием данных ребер и выбирается такая ориентация контигов, с которой согласуются большинство парных чтений, расположенных на ребрах графа.

2.4. Уточнение оценки расстояний в скэффолде на основе принципа максимального правдоподобия

Для уточнения оценки расстояний в уже готовых скэффолдах ранее предложенная формула правдоподобия для оценки расстояния между двумя контигами оказывается непригодной в большинстве случаев. Ее основным недостатком является учет наличия лишь двух контигов, в то время как наличие других контигов игнорируется. Возможна более точная оценка расстояний между контигами, если включить в формулу правдоподобия

одновременный учет расстояний сразу между множеством контигов.

Для учета множества контигов формула правдоподобия, использовавшаяся для случая двух контигов, распространяется на случай множества контигов, а значит метод оценки будет состоять из тех же шагов, а также будет подразумеваться та же модель получения чтений.

В начале расстояние между двумя контигами d нужно заменить на вектор расстояний между соседними контигами $\overline{d}$, в котором координата $\overline{d_i}$ соответствует расстоянию между контигами i и $i+1$. Данное изменения позволит положить начало распространению формулы правдоподобия на случай скэффолда, в котором уже определен порядок контигов.

2.4.1. Учет связывающих чтений

В случае связывающих чтений распространение формулы правдоподобия не встречает существенных трудностей. События, соответствующие получению двух связывающих парных чтений, являются независимыми, причем какие именно контиги связываются не играет значения. Соответственно вероятность данных событий может быть рассчитана как вероятность независимых событий.

Формула вероятности получения фрагмента длины d (2.1) почти соответствует желаемому, но вместо величины d нужно ввести другую, которая будет учитывать возможное наличие дополнительных контигов между теми контигами, которые оказались связанными парным чтением. Таким образом расстояние между контигами n и m будет равно сумме контигов $\sum_{k=n+1}^{m-1} l_k$, лежащих между ними, а также сумме расстояний между соседними контигами $\sum_{t=n}^{m-1} \overline{d_t}$. Итоговая формула для расчета вероятности получения фрагмента имеющейся длины:

$$p(fragment\_size|\overline{d}) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(d_{1i} + d_{2i} + 2r - \mu + \sum\limits_{t=n}^{m-1} \overline{d_t} + \sum\limits_{k=n+1}^{m-1} l_k)^2}{2\sigma^2}) \tag{2.9}$$

Формула вероятности получения фрагмента в заданной позиции не изменится и имеет следующий вид:

$$p(fragment\_position|\overline{d}) = \frac{1}{L} \tag{2.10}$$

Суммарный вклад каждого чтения состоит из произведения эти двух вероятностей. Вклад всех связывающих чтений состоит из произведения вероятностей получения чтений, где n количество связывающих чтений:

$$p(connecting|\overline{d}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma L} exp(-\frac{(d_{1i} + d_{2i} + 2r - \mu + \sum\limits_{t=n}^{m-1} \overline{d_t} + \sum\limits_{k=n+1}^{m-1} l_k)^2}{2\sigma^2}) \tag{2.11}$$

2.4.2. Учет несвязывающих чтений

В случае несвязывающих чтений распространение формулы правдоподобия встречает некоторые трудности, связанные с тем, что события парное чтение не связывает некоторые контиги под номерами один и два и это же чтение не связывает некоторые контиги под номерами три и четыре являются зависимыми.

Для устранения этой проблемы необходимо изменить вспомогательную формулу (2.5), которая позволяет подсчитать вероятность получения случайного связывающего чтения. Изменения коснутся функции $w_d(f)$, которая подсчитывает количество позиций, в которых может быть получено случайное связывающее чтение для пары контигов. Вводится новая функция $\tilde{w_d}(f)$, которая подсчитывает число позиций, в которых может быть

получено случайное связывающее чтение для всех контигов в скэффолде. В итоге измененная функция имеет вид:

$$p(random\_connect|\overline{d}) = \sum_f \frac{\tilde{w}_d(f)}{\sqrt{2\pi}\sigma L} exp(-\frac{(\mu - f)^2}{2\sigma^2}) \tag{2.12}$$

Опять же для учета вклада несвязывающих чтений нужна не вероятность случайной пары чтений связать пару контигов, а обратная ей величина $1 - p(random\_connect|d)$. Суммарное количество парных чтений R, количество связывающих чтений n, результирующий вклад всех несвязывающих чтений в функцию правдоподобия:

$$p(non - connecting|\overline{d}) = (1 - p(random\_connect|\overline{d}))^{R-n} \tag{2.13}$$

2.4.3. Нахождение наиболее вероятного расстояния

Функция правдоподобия, составленная из частей (2.11) и (2.13), максимизируется для нахождения наиболее вероятного расстояния.

Для упрощения работы с функцией правдоподобия используется ее логарифм. Часть, отвечающая за учет связывающий чтений, преобразуется в многочлен от $\overline{d}$. В данном многочлене возможно предпосчитать коэффициенты и построить дерево отрезков для поиска расстояния между контигами, что позволит вычислять вклад связывающий чтения за $O(\log|\overline{d}|)$ для произвольного $\overline{d}$.

$$\log p(connecting|\overline{d}) = n \log \frac{1}{\sqrt{2\pi}\sigma L} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (d_{1i} + d_{2i} + 2r - \mu + \sum_{t=n}^{m-1} \overline{d_t} + \sum_{k=n+1}^{m-1} l_k)^2 \tag{2.14}$$

При логарифмировании формула, отвечающая за вклад несвязывающих чтений, не приобретает простого вида. Для упрощения вычисления вклада несвязывающих чтений сумма из формулы (2.12) аппроксимируется интегралом.

$$p(random\_connect|\overline{d}) \approx \int_0^\infty \frac{\tilde{w}_d(f)}{\sqrt{2\pi}\sigma L} exp(-\frac{(\mu - f)^2}{2\sigma^2})df \tag{2.15}$$

Функция $\tilde{w}_d(f)$, к сожалению, не имеет простого вида в отличии от функции $w_d(f)$, что усложняет оптимизацию функции.

В итоге логарифм функции правдоподобия вычисляется за $O(n \log |\overline{d}|)$ при условии предподсчета, проводимого за $O(n + |\overline{d}| \log |\overline{d}|)$. Для поиска максимума правдоподобия используется градиентный спуск, каждая итерация которого занимает $O(n|\overline{d}| \log |\overline{d}|)$, что с связано с необходимость вычисления, как самой функции правдоподобия, так и ее градиента.

Глава 3. Результаты

Метод сборки, описанный в данной работе, был реализован на языке программирования Java. Метод был протестирован на различных наборах контигов, соответствующих геному бактерии кишечной палочки Escherichia coli, геном которой подвергся подробному изучению и на данный момент полностью известен [13]. Благодаря этому возможно сравнение результатов, полученных с помощью метода, предложенного в данной работе, с результатами работы других методов сборки скэффолдов геномной последовательности.

В данной главе сначала будут описаны данные, которые являются входными параметрами сборщика скэффолдов. Затем описаны критерии, по которым производилось сравнение скэффолдов. После этого следует подробное описание данных, использовавшихся для тестирования, а также анализ полученных результатов и их сравнение с остальными методами. Глава завершается рекомендациями по внедрению и улучшению разработанного метода.

3.1. Тестирование сборщика скэффолдов

На данный момент не существует стандартного способа оценки результатов работы сборщиков скэффолдов, ровно как и какой-либо общепринятой системы оценки. В данной секции описываются данные, которые подаются сборщику в качестве входных параметров, а также использованные способы оценки.

3.1.1. Входные данные

Сборщик в качестве входных параметров принимает контиги, результат картирование парных чтений на контиги в SAM формате и приблизительную длину геномной последовательности, которую обычно можно узнать заранее, но в случае, когда данный параметр неизвестен, он при-

ближается суммой длин всех контигов. Для картирования чтений была использована программа Bowtie [14]. При тестировании также была задействована программа BLAST [15] для картирования контигов на референсную геномную последовательной бактерии кишечной палочки.

3.1.2. Параметры оценки результатов сборки

Как было сказано ранее, к сожалению, общепринятые или стандартизированные методы оценки результатов сборки скэффолдов отсутствуют. Данная работа была направлена на улучшение оценки расстояний, а для сравнения качества оценки расстояний с другими сборщиками использовались следующие критерии:

1. Среднее отклонение расстояния от истинного значения, а также среднеквадратичное отклонение отклонения;
2. Процент ситуаций, когда оценка расстояния предложенного метода была ближе к истинному значению по сравнению с другим методов;
3. Процент ситуаций, когда оценка расстояния предложенного метода была равноудалена от истинного значения по сравнению с другим методов.

3.2. Результаты сборки

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

3.2.1. Тестовые данные

Для тестирования метода, описанного в данной работе, использовался геном бактерии кишечной палочки(Escherichia coli). Геном данной бактерии практически полностью известен, что дает возможность использовать референсную геномную последовательность для оценки и сравнению

результатов сборки скэффолдов. Данная геномная последовательность содержит приблизительно 4.6 миллионов нуклеотидов.

Для тестирования использовались четыре различных набора контигов, полученные различными сборщиками контигов, два из которых получены различными версиями сборщика контигов, разработанного в лаборатории «Алгоритмы сборки геномных последовательностей» в Университете ИТМО, один из которых получен с помощью сборщика контигов SPAdes и последний с помощью сборщика контигов ABySS. Все контиги были получены на основе реальных данных. Параметры наборов контигов можно увидеть в таблице 3.1. Также использовались реальные библиотеки парных чтений, из которых были собраны контиги, параметры которых приведены в таблице 3.2.

Таблица 3.1. Параметры тестовых наборов контигов.

Набор	Число контигов	Суммарная длина	Максимальная длина	N50
ITMO 1	592	4605995	77153	17546
ITMO 2	514	4532445	71310	15931
ABySS	632	4623758	167292	64280
SPAdes	305	4562415	167394	45241

Таблица 3.2. Параметры тестовых библиотек чтений.

Число пар чтений	Длина чтений	Средняя длина фрагмента	Стандартное отклонение
10000000	36	200	10

3.2.2. Оценка полученных результатов и сравнение с другими методами

Результаты работы исходного и улучшенного метода на тестовых данных сравнивались с результатами работы других методов на тех же исходных данных. Различные параметры получившихся наборов представлены в таблицах 3.3, 3.4 и 3.5.

Таблица 3.3. Среднее отклонение оцененного расстояния от реального и стандартное отклонение ошибки.

Набор	SOPRA	GAPEST	OPERA	Исходный метод	Улучшенный метод
ITMO 1	217 ± 13	211 ± 16	207 ± 14	161 ± 13	155 ± 11
ITMO 2	206 ± 14	195 ± 15	212 ± 13	155 ± 12	130 ± 11
ABySS	189 ± 13	173 ± 13	163 ± 13	142 ± 11	143 ± 10
SPAdes	195 ± 15	175 ± 14	177 ± 13	123 ± 13	117 ± 13

Таблица 3.4. Процент ситуаций, когда оценка исходным методом ближе к реальному значению расстояния между контигами и когда оценки методов совпадают.

Набор	Исходный метод vs SOPRA	Исходный метод vs GAPEST	Исходный метод vs OPERA
ITMO 1	67%/5%	65%/4%	60%/8%
ITMO 2	68%/8%	55%/7%	57%/7%
ABySS	64%/7%	70%/3%	59%/6%
SPAdes	63%/7%	68%/5%	61%/6%

Таблица 3.5. Процент ситуаций, когда оценка улучшенным методом ближе к реальному значению расстояния между контигами и когда оценки методов совпадают.

Набор	Улучшенный метод vs SOPRA	Улучшенный метод vs GAPEST	Улучшенный метод vs OPERA
ITMO 1	71%/6%	70%/6%	65%/5%
ITMO 2	73%/8%	72%/4%	60%/10%
ABySS	75%/8%	71%/3%	64%/6%
SPAdes	73%/5%	75%/7%	63%/8%

Результат оценки расстояния между контигами в скэффолдах сравнивался с распространенным методом среднего арифметического, применяемым в сборщиках скэффолдов SOPRA и OPERA. Также предложенный в данной работе метод сравнивался со сборщиком GAPEST, в котором также используется метод среднего арифметического, но производится оценка расстояний между несколькими контигами одновременно, что позволяет получать более точные оценки.

Сравнение показывает, что результаты, полученные с помощью предложенного метода, в среднем меньше отклонены от истинного значения и чаще находятся к реальному расстоянию, чем оценки других методов и исходного сборщика. Разработанный метод позволяет более точно оценивать расстояния между контигами в скэффолде, что показывают выше указанные результаты.

3.3. Рекомендации по внедрению

Представленный в данной работе метод рекомендуется применять для сборки скэффолдов геномных последовательностей на основе имеющихся контигов и парных чтений. Также, в силу структуры данной работы, описанный метод улучшения оценок расстояний в уже готовых скэффолдах рекомендуется применять для уточнения оценок расстояний в скэффолдах, полученных любым другим произвольным путем. Применения данного подхода, основанного на принципе максимального правдоподобия, позволит существенно улучшить оценки расстояний в скэффолдах, где данный подход при сборке оказался незадействованным.

3.4. Рекомендации по улучшению

В данный момент формула правдоподобия, учитывающая наличие множества контигов, позволяет получить лишь оценки на расстояния, в то время как ориентация контигов, а также их порядок в скэффолде по-

лучаются эвристическим путем. Рекомендуется улучшить формулу правдоподобия или предложить способ ее применения, который бы учитывал данные параметры скэффолдов наряду с оценкой расстояний.

Заключение

Было разработано и реализовано улучшение метода сборки скэффолдов геномной последовательности на основе принципа максимального правдоподобия. Входными данными метода являются набор контигов — длинных частей генома, а также набор парных чтений — небольших фрагментов геномной последовательности, для которых известно примерное расстояние между ними. Результатом работы улучшенного метода является набор скэффолдов — множества упорядоченных и ориентированных контигов с известными расстояниями между ними.

Для уточнения оценки расстояния между контигами используется принцип максимального правдоподобия и новая функция правдоподобия, которая учитывает не только пары чтений, но и несвязывающие пары чтений, а также наличие множества контигов, что позволяет повысить точное оценки расстояний между контигами по сравнению с исходной формулой. В ходе работы сборщика строится упрощенных граф контигов, из которого получается первое приближение скэффолдов. Затем скэффолды, полученные при первом приближении, объединяются в большие скэффолды с помощью построения графа скэффолдов и поиска кратчайщих путей в нем.

Предложенный улучшенный метод был реализован и протестирован на примере бактерии Escherichia coli. Результаты работы метода были сравнены с результатами работы других распространенных сборщиков скэффолдов. Анализ результатов сравнения явно указывает на то, что улучшенных метод превосходит в точности оценок расстояний не только другие распространенные сборщики, но и исходный сборщик как по среднему отклонению от истинного расстояния, так и по проценту ситуаций, в которых точность оценки оказалась выше у предложенного метода.

Предлагаемый метод собирает скэффолды, обладающие лучшим качеством по сравнению с другими распространенными сборщиками на те-

стовых данных. Это указывает на перспективность развития и использования сборки скэффолдов на основе принципа максимального правдоподобия. 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.

Источники

1. Ахи А. *Сборка скэффолдов геномной последовательности на основе принципа максимального правдоподобия.* 2013.
2. *Talking glossary of genetic terms: genome.* National Human Genome Research Institute.
3. Alberts B., Johson A., Lewis J., Raff M., Roberts K., Walter P. *Molecular Biology of the Cell.* Fourth Edition, New York: Garland Science, 2002.
4. Watson J., Crick F. *Molecular structure of nucleic acids; a structure of deoxyribose nucleic acid.* Nature, no. 171, 1953.
5. Anderson S. *Shotgun DNA sequencing using cloned DNase I-generated fragments.* Nucleic Acids Res., no. 9(13), 1981, pp 3015-3027.
6. http://en.wikipedia.org/wiki/N50_statistic
7. 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.
8. 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.
9. Sahlin K., Street N., Lundeberg J., Arvestad L. *Improved gap size estimation for scaffolding algorithms.* Bioinformatics, Vol. 28, No. 17, 2012, pp. 2215-2222.
10. Dayarian A., Michael T.P. Sengupta A.M., *SOPRA: Scaffolding algorithm for paired reads via statistical optimization.* BMC Bioinformatics, No. 11, 2010, p. 345.
11. Garey M.R., Johnson D.S *Computers and intractability: a guide to the theory of NP-completeness.* San Francisco: W. H. Freeman, 1979.
12. Barahona F. *On the computational complexity of Ising spin glass models*, Journal of Physics A: Mathematical and General, no. 15, 1982, pp. 3241-3253.
13. 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.

14. 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.

15. 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.

16. 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.

17. *Illumina, Inc.* http://www.illumina.com/