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

Факультет информационных технологий и программирования
Кафедра компьютерных технологий

Муравьёв Сергей Борисович

Разработка метода оценки сборки генома на основе принципа максимального правдоподобия

Научный руководитель: кандидат технических наук, доцент кафедры компьютерных технологий
Ф. Н. Царев

Санкт-Петербург
2014

Содержание

Введение **5**

Глава 1. Обзор предметной области **7**

1.1 Основные понятия 7
1.1.1 Строение ДНК 7
1.1.2 Секвенирование генома 7
1.1.3 Сборка генома 8
1.2 Постановка задачи 8
1.3 Обзор существующих методов и их проблем 9
1.3.1 CGAL 9
1.3.2 de Novo 10
1.3.3 ALE 10
1.4 Выводы к главе 1 11

Глава 2. Описание используемого подхода **12**

2.1 Особенности ALE 12
2.1.1 Описание метода 12
2.1.1.1 Нормализационная константа Z 12
2.1.1.2 Оценка сборки при отсутствии чтений 12
2.1.1.3 Корректность чтений 13
2.1.1.4 Расстояние между парными чтениями 13
2.1.2 Оценка глубины покрытия 13
2.1.3 Проблема оценки глубины покрытия 14
2.2 Оценка P_{depth} 17
2.2.1 Общая идея 17
2.2.2 Выбор GC-контента 17
2.2.3 Учёт ошибок 18
2.2.4 Анализ производительности 18
2.3 Выводы к главе 2 18

Глава 3. Применение используемого метода и результаты работы на реальных тестах 19

3.1 E.coli 19
3.2 Генерация ошибок 19

Глава 4. Заключение 20

Список литературы 21

Введение

Задача корректной оценки качества сборки геномной последовательности является важной частью биоинформатики. Задача состоит в том, чтобы по набору чтений и сборке цепи ДНК геномной последовательности определить качество сборки, не имея эталонной сборки. Основными проблемами в этом процессе являются наличие большого числа ошибок в исходных данных, а также большой объем входных данных, исчисляющийся сотнями гигабайт.

Последние достижения в области секвенирования следующего поколения резко снизили стоимость секвенирования. С развитием сборщиков появилась возможность использовать большие объёмы секвенируемой информации. Благодаря технологии секвенирования методом дробовика стало появляться всё больше геномов различных организмов, от маленьких бактерий до млекопитающих. Несмотря на это, некоторые генетические последовательности получаются напрямую из материи, содержащей в себе несколько организмов, при помощи single-cell секвенирования и метагеномного секвенирования.

При сборке конкретного организма возникают ошибки, обусловленные шумах в информации, большим объёмом данных и особенностями сборщика. В случае с метагеномной сборкой возникают дополнительные трудности: глубина покрытия чтениями распределена неравномерано, неоднозначность в анализе повторяющихся областей в случае метагеномной сборки усугубляется. Было разработано несколько инструментов для обнаружения ошибок при неметагеномной сборке. Однако такие средства используют эталонную или близкую к эталонной сборку организма. В случае отсутствия эталонной сборки подобные инструменты предоставляют косвенные оценки качества сборки генома, такие как N50, глубина покрытия, оценка расстояния между парными чтениями. Такие метрики предостав-

ляют информацию о производительности сборщика, но не обеспечивают внутренних прямых измерений качества сборки.

В данной работе была разработана статистическая модель для оценки качества сборки генома в условиях отсутствия эталонной сборки. Данная модель представляет из себя формулу на основе формулы Байеса, которая вычисляет вероятность того, что сборка правильная, то есть соответствует эталону.

Глава 1. Обзор предметной области

1.1. Основные понятия

Биоинформатика [1] — это наука на стыке двух дисциплин: биологии и информатики. Многие задачи биологии требуют обработки колоссального объема данных, что и привело к возникновению дисциплины. Одной из важных задач биоинформатики является задача корректной оценки собранной геномной последовательности.

1.1.1. Строение ДНК

Геном — это совокупность информации, передаваемой всеми живыми существами по наследству. Геномы большинства живых организмов состоят из молекул *ДНК* — дезоксирибонуклеиновой кислоты. ДНК — это полимерная молекула, представляющая из себя две закрученных цепочки, состоящие из соединённых в последовательность *нуклеотидов*. Нуклеотиды, входящие в ДНК, разделяют по азотистым основаниям на четыре группы: *аденин* (A), *цитозин* (C), *гуанин*(G), и *тимин* (T). Цепи ДНК соединены между собой по *принципу комплиментарности*: аденин с тимином, цитозин с гуанином.

Геном в биоинформатике представляется одной из двух комплементарных цепей молекулы ДНК. Строка, состоящая из символов A, G, C и T, соответвующих типам нуклеотидов,является наиболее удобным представлением генома в биоинформатике.

1.1.2. Секвенирование генома

Для определения линейной последовательности нуклеотидов в молекуле ДНК геном подвергают секвенированию. Одним из популярных методов секве-нирования является метод дробовика (Shotgun Sequencing). Метод состоит в выделении из молекулы ДНК коротких участков (порядка

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

1.1.3. Сборка генома

Традиционно процесс сборки генома состоит из трех этапов: исправление ошибок в парных чтениях, сборка контигов, длинных последовательных частей геномной последовательности, и сборка скэффолдов, наборов упорядо-ченных 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.

1.2. Постановка задачи

В последнее время появилось множество сборщиков генома. Возникла проблема сравнения качества их работы. В прошлом это делалось при помощи таких популярных метрик как *N50/N90*, *NG50/NG90*, длина наибольшего контига или скэффолда. Хотя исследования показали, что простые метрики коррелируют с качеством сборки, используемые в настоящее время метрики являются грубыми и не обеспечивают полной информации о результате сборки. Например сборка, состоящая из просто склеенных конец-в-конец чтений, имеет очень большой *N50*, но, очевидно, плохое качество сборки.

Требуется разработать метрику, качественно оценивающую сборку на основе принципа максимального правдоподобия.

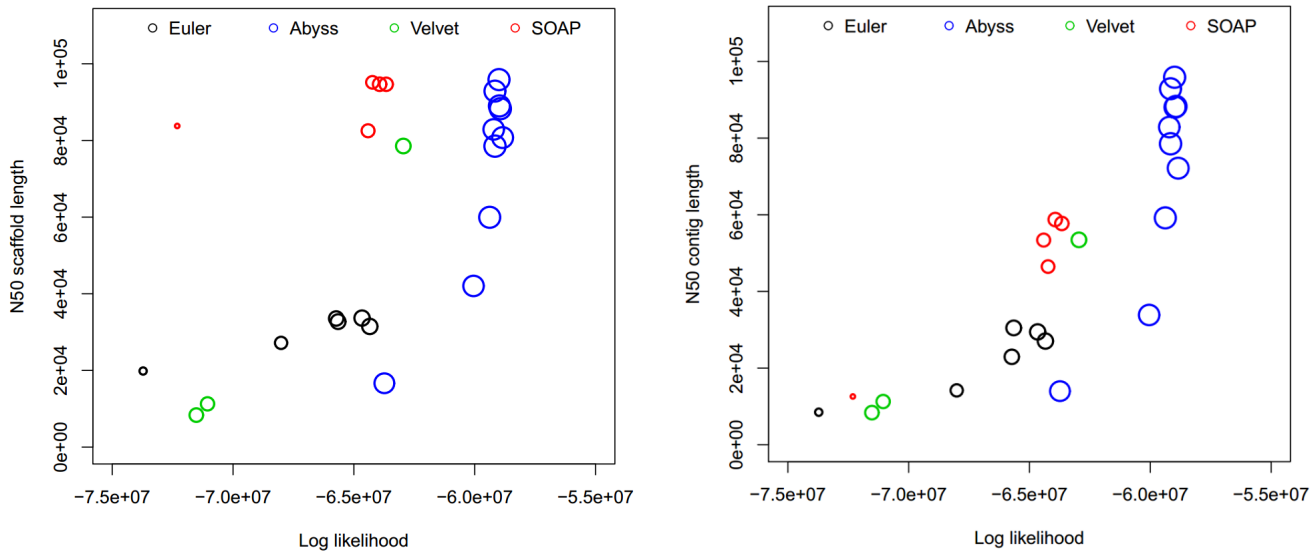

Рис. 1.1. Графики корреляции N50 и loglikelihood, посчитанным с помощью CGAL, для *E.coli*. Окружности соответствуют сборкам с различными длинами k-меров. Размер окружности соответствует схожести с эталонной сборкой.

1.3. Обзор существующих методов и их проблем

За последние два года было разработано несколько новых способов качественной оценки сборки геномной последовательности на основе принципа максимального правдоподобия. Их способ заключается в получении *loglikelihood* – логарифма вероятности того, что сборка является верной при наличии заданного набора чтений.

1.3.1. CGAL

Данный метод является первым методом оценки на основе принципа максимального правдоподобия. В её основе лежит слудующая формула:

$$l(A,R) = \ln \prod_{i=1}^{n} p(r_i|A) \approx \sum_{i=1}^{N} \ln \sum j = 1_i^M p_F(l_{i,j}) p_S(s_{i,j}) p_E(r_i|a_{i,j}, e_{i,j})$$

- R – множество чтений;
- A – сборка;
- M_i – количество возможных "соответствий"в сборке чтения r_i;

- $l_{i,j}$, $s_{i,j}$, $a_{i,j}$ и $e_{i,j}$ – соответственно длина чтения, позиция чтения, подпоследовательность сборки и ошибки для "соответствия"j для чтения i.

Проблема данного метода заключается в том, что согласно авторам данного метода чтения должны быть распределены равномерно по всей длине генома, что является существенным недостатком, поскольку в реальной ситуации чтения распределены чаще всего неравномерно.

1.3.2. de Novo

В основе данного метода лежит формула Байеса:

$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

A – событие, при котором сборка является эталонной геномной последовательностью R – событие, при котором исследуется определённый набор чтений. $P(A)$ и $P(R)$ являются константами.

$$P(R|A) = \prod_{r \in R} P(r|A)$$

Задача сводится к получению оценки $P(r|A)$ с использованием динамического программирования. Данный метод достаточно точно вычисляет вероятности ошибок для больших наборов чтений. Однако существенными недостатками являются большой объём потребляемой памяти ($O(n^2)$, где n – длина сборки) и сложность в реализации данного алгоритма. Кроме того на сегодняшний день нет ни одной работающей реализации данного подхода.

1.3.3. ALE

На данный момент этот метод является наиболее совершенным. В его основе также лежит формула Байеса, как и в *de Novo*.

$$P(A|R) = \frac{P(R|A)P(A)}{Z}$$

Z – константа. $P(A)$ описывает качество сборки в отсутствие какой-либо информации о чтениях. $P(R|A)$ оценивается по следующей формуле:

$$P(R|A) = P_{placement}(R|A)P_{insert}(R|A)P_{depth}(R|A)$$

$P_{placement}$ оценивает, насколько хорошо содержание чтений совпадает со сборкой, $P_{insert}(r|A)$ оценивает, насколько хорошо априорные расстояния между парными чтениями (insert length) совпадают с получившимися в результате сборки, $P_{depth}(r|A)$ оценивает, насколько априорная глубина покрытия в каждой позиции совпадает с получившейся в результате сборки на основе GC-контента. *Глубина покрытия* – это количество чтений, которое покрыло данную позицию в сборке. *GC-контент* – это процентный состав суммы всех нуклеотидов, являющихся гуанином(G) или цитозином(C) по отношению к длине исследуемого участка генома.

Более подробно о данном методе и о его проблемах будет изложено в следующей главе.

1.4. Выводы к главе 1

В условиях быстро появляющихся новых сборщиков, возникла задача качественного сравнения их между собой, поэтому разработка метрики оценки качества сборки геномной последовательности является важной задачей биоинформатики. Существующие популярные метрики, такие как *N50/N90*, не дают полной информации о результате сборки, поэтому был предложен новый подход, заключающийся в вычислении *loglikelihood* – логарифма вероятности того, что сборка является верной при заданном наборе чтений. В последнее время было предложено несколько способов получения *loglikelihood*, однако они не лишены своих недостатков. Наиболее совершенным инструментом оценки на сегодняшний день является ALE. В данной работе представлен способ оценки качества на базе ALE.

Глава 2. Описание используемого подхода

2.1. Особенности ALE

2.1.1. Описание метода

Как уже было изложено выше, ALE оценивает качество сборки при помощи следующей формулы:

$$P(A|R) = \frac{P_{placement}(R|A)P_{insert}(R|A)P_{depth}(R|A)P(A)}{Z}$$

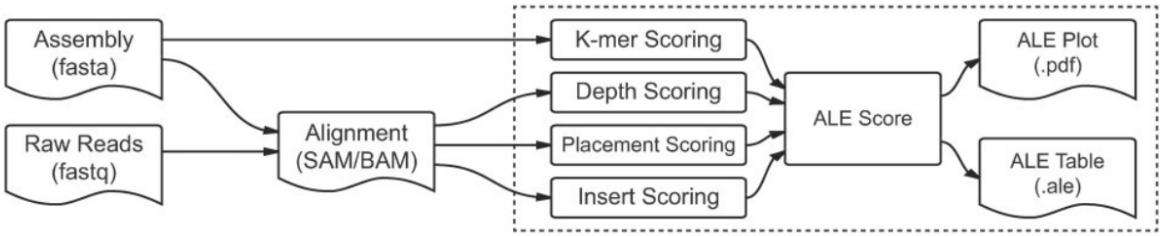

Рис. 2.1. Общая схема работы ALE

2.1.1.1. Нормализационная константа Z

Z является константой, вычисляется как:

$$Z = \sum_{A'} P(R|A')P(A')$$

Данную величину вычисляют приближённо, поскольку явно её вычислить невозможно из-за слишком большого множества сборок A' (4^L, где L – длина сборки).

2.1.1.2. Оценка сборки при отсутствии чтений

$$P(A) = ZP_{kmer}(A)$$

$P_{kmer}(A) = \prod_{i \in K} f_i^{n_i}$, где K – множество уникальных k-меров, n_i – количество раз, когда k-мер i встречается в текущем контиге в сборке. f_i – частота появление k-мера i в контиге: $f_i = n_i / \sum_{j \in K} n_j$.

2.1.1.3. Корректность чтений

$P_{placement}(R|A)$ – оценивает насколько чтения соответствуют сборке, выражается следующим образом:

$$P_{placement}(R|A) = \prod_{r_i \in R} P_{placement}(r_i|A) =$$

$$= \prod_{r_i \in R} P_{matches}(r_i|A) P_{orientation}(r_i|A)$$

$P_{matches}(r_i|A)$ оценивает, насколько содержимое чтения соответствует тому участку сборки, на который было произведено картирование данного чтения. Каждый нуклеотид j, считанный при помощи секвенатора имеет качество считывания Q_j, $Q_j \subseteq [0; 1]$, тогда $P_{matches}(r_i|A) = \prod_{j \in r_i} P_{base_j|A}$, где $P_{base_j|A} = Q_j$ когда нуклеотид j совпадает со сборкой и $P_{base_j|A} = (1 - Q_j)/4$ в противном случае. Если в сборке встречается неизвестный нуклеотид N, то считаем, что $P_{base_j|A} = 1/4$ Если инструмент картирования сопоставил чтение более чем в одно место, ALE выбирает позицию, у которой $P_{placement}(R|A)$ наибольший.

$P_{orientation}(r_i|A)$ оценивает корректность ориентации в случае парных чтений. Величина вычисляется эмпирически.

2.1.1.4. Расстояние между парными чтениями

$P_{insert}(R|A)$ оценивает расстояние между парными чтениями, вычисляется как $P_{insert}(R|A) = \prod_{r_i \in R} P_{insert}(r_i|A)$. $P_{insert}(r_i|A) = Normal(L_i; \mu, \sigma^2)$

2.1.2. Оценка глубины покрытия

Рассмотрим более подробно вычисление $P_{depth}(R|A)$. Эта величина описывает, насколько глубина покрытия в каждой позиции, соответствует

глубине, которую мы бы ожидали увидеть, если бы картирование производилось на эталонную сборку.

$$P_{depth}(R|A) = \prod_i P_{depth}(d_i|A)$$

d_i – глубина в позиции i. Предполагается, что глубины распределены по Пуассону с центром, вычисленным из независимого гамма-распределения с центром в ожидаемой глубине в данной позиции и GC-контентом в качестве второго параметра. Рассмотрим это утверждение более подробно.

Сначала рассчитываются средние глубины для каждого из 100 множеств GC-контента: 0-1, 1-2 ... 99-100%. Пусть X_i – это средний GC-контент по всем чтениям, которые покрывают позицию i. $\mu_{depth}(X_i)$ – средняя глубина для того множества GC-контентов, которому соответствует X_i. Минимальное значение $\mu_{depth}(X_i)$ устанавливается равным 10. В итоге для каждой позиции i в сборке оценка глубины будет производиться по следующей формуле:

$$P_{depth}(d_i|A, X_i) = \int_0^\infty Poisson(d_i, Y_i)Gamma(Y_i, max(10, \mu_{depth}(X_i), 1))dY_i =$$
$$= NegBinom(d_i, \mu_{depth}(X_i), 0.5)$$

Подытоживая вышесказанное, получаем, что глубины распределены по отрицательному биномиальному распределению $NB(r, p)$ с параметрами $r = \mu_{depth}(X_i)$ и $p = 0.5$.

2.1.3. Проблема оценки глубины покрытия

В ходе выполнения данной работы было выявлено, глубины покрытий распределены не по отрицательному биномиальному распределению.

Для того, чтобы убедиться в этом был для каждого из 100 множеств GC-контента было построено эмпирическое распределение глубин покрытий. В итоге получилась таблица $100 \times maxdepth$, где $maxdepth$ – это максимальное значение глубины покрытия позиции чтениями в данной сборке.

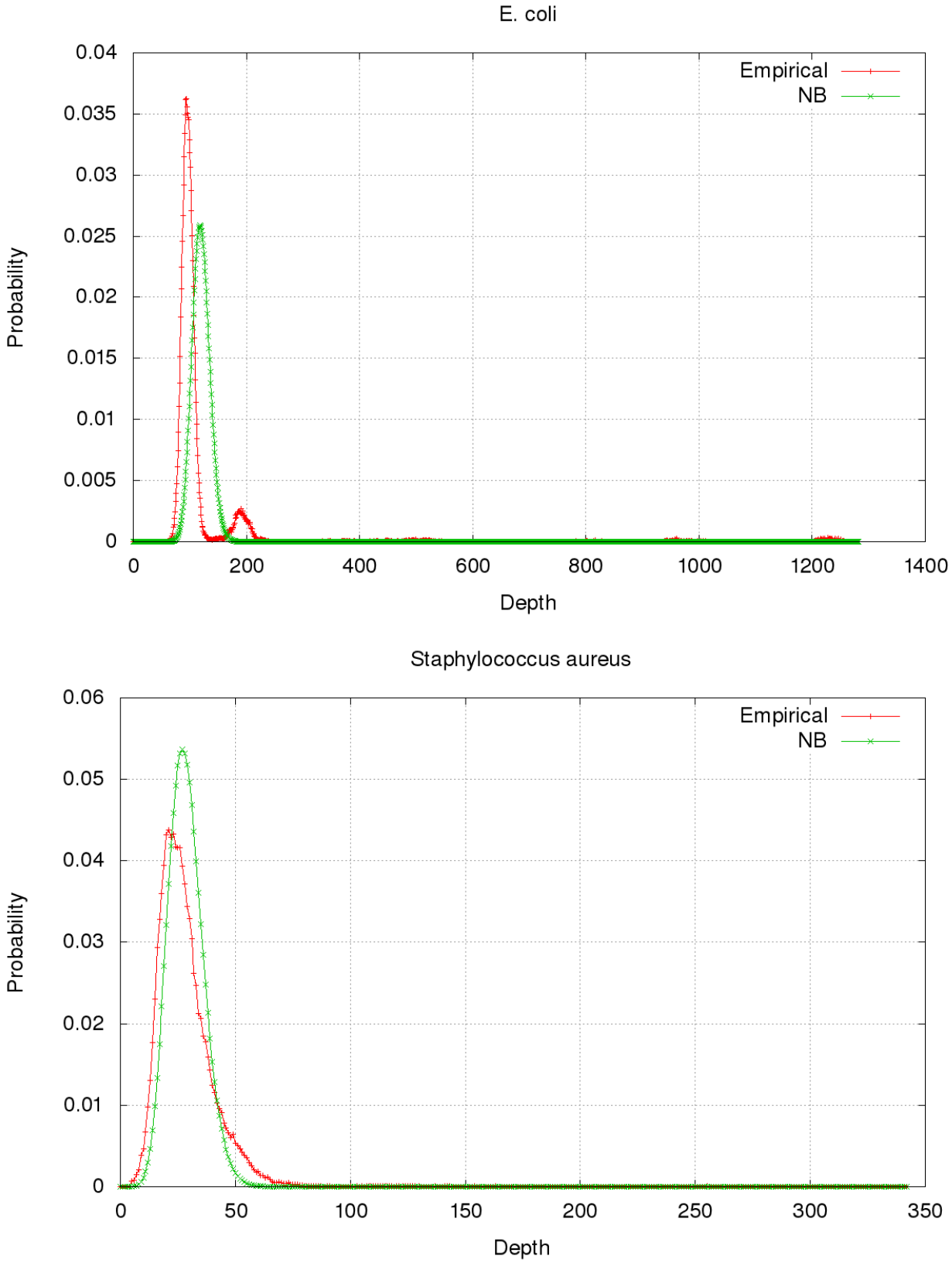

Рис. 2.2. Графики распределения глубин покрытий для E.coli и для Staphylococcus aureus, посчитанные эмпирически и при помощи отрицательного биномиального распределения.

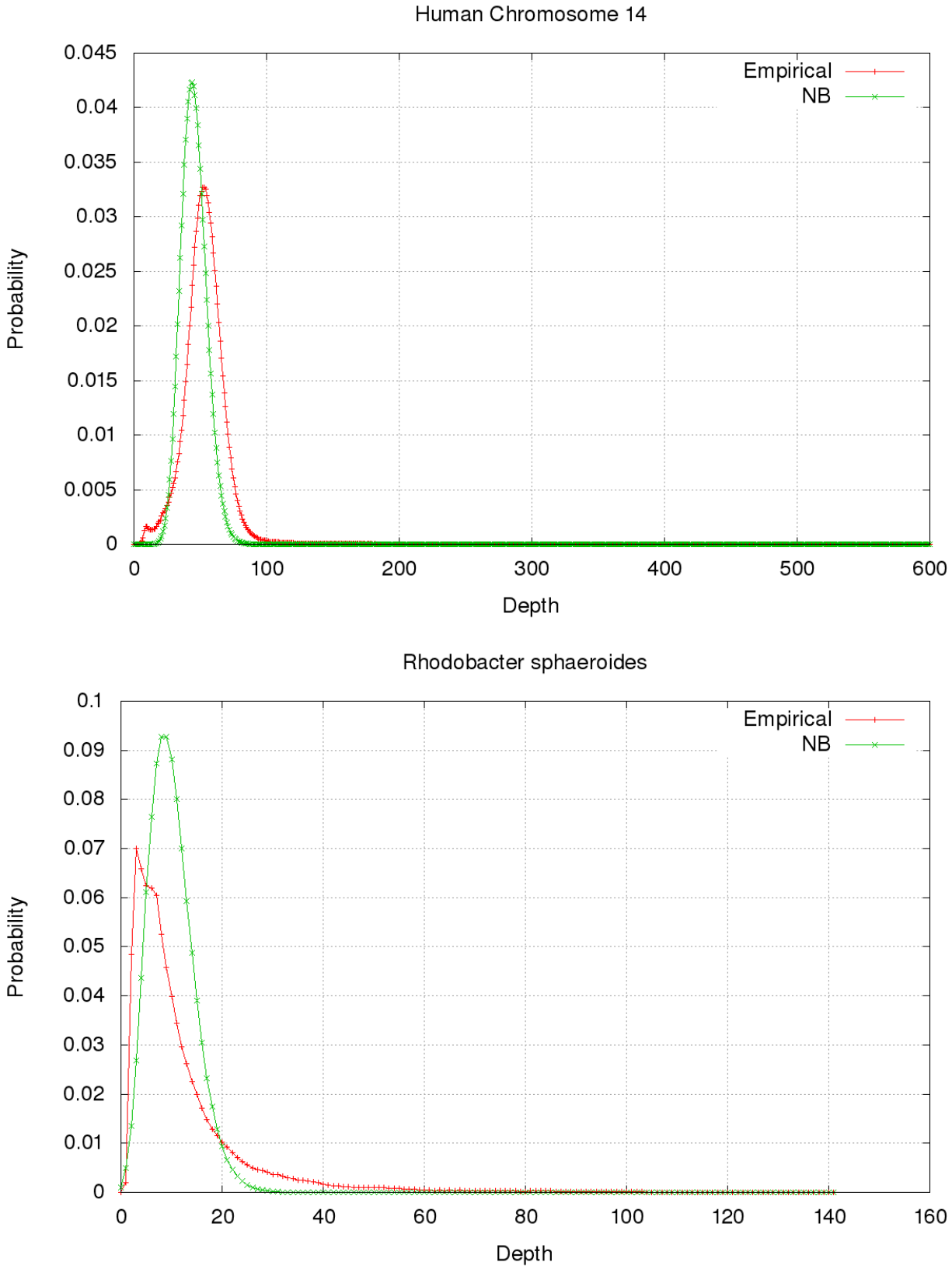

Рис. 2.3. Графики распределения глубин покрытий для 14-й хромосомы человека и для Rhodobacter sphaeroides, посчитанные эмпирически и при помощи отрицательного биномиального распределения.

Далее было выбрано такое множество GC-контента, которому соответствует максимальное число позиций в сборке, и построены графики сравнения эмпирического и отрицательного биномиального распределения, представленные выше.

Также были приведены статистические тесты, которые показали, что реальное распределение (эмпирическое) не соответствует теоретическому (отрицательному биномиальному). В частности на Staphylococcus aureus p-value после χ^2 теста было равно 2×10^{-5} (за $p = 1$ бралась нулевая гипотеза). Приближение на основе принципа максимального правдоподобия также показало отрицательный результат: логарифм вероятности того, что эмпирическое распределение соответствует отрицательному биномиальному ≈ -5.

2.2. Оценка P_{depth}

2.2.1. Общая идея

Предпринимались попытки приближения различными распределениями и комбинациями распределений, однако они не принесли результатов. p-value и логарифм вероятности соответствия двух распределений друг другу, описанный выше, либо не удавалось значительно улучшить, либо в случае видимого улучшения на сборке одного организма метод не работал на сборках других организмов.

В связи с этим для оценки глубины покрытия в каждой позиции в данной работе предлагается использовать эмпирическое распределение. Оно заведомо точнее отрицательного биномиального, поскольку учитывает реальное распределение при заданном наборе чтений.

2.2.2. Выбор GC-контента

Аналогично ALE будем выделять 100 множеств GC-контентов: 0-1, 1-2 ... 99-100%. Такой выбор связан с тем, у современных секвенаторов Illumina и Ion Torrent средняя длина чтений около 100 нуклеотидов. Были

проведены опыты с разбиением на меньшее и большее число множеств. В первом случае существенно падает точность оценки, во втором сильно возрастает время работы оценщика, при этом значительного выигрыша в точности нет.

2.2.3. Учёт ошибок

Эксперименты показали, что в большинстве случаев, нуклеотиды, не покрытые чтениями, являются ошибками сборки геномной последовательности, поэтому предлагается при подсчёте эмпирического рапределения не учитывать позиции в геноме с нулевыми глубинами покрытия. Эффективность такого подхода будет рассмотрена в следующей главе.

2.2.4. Анализ производительности

В отличие от ALE, метод не требует предварительного вычисления средней глубины покрытия для заданного отрезка GC-контента $\mu_{depth}(X_i)$. Для хранения вероятностей глубин в каждой позиции требуется таблица T размером $100 \times maxdepth$, где $maxdepth$ – это максимальное значение глубины покрытия в данной сборке. Нахождение вероятности производится производится обращением к соответствующей ячейке таблицы $T[gc_i][depth_i]$, где $depth_i$ и gc_i – это соответственно глубина в позиции i и множество, к которому принадлежит GC-контент в позиции i.

2.3. Выводы к главе 2

Глава 3. Применение используемого метода и результаты работы на реальных тестах

3.1. E.coli

3.2. Генерация ошибок

Глава 4. Заключение

Список литературы

1. Scala language. http://www.scala-lang.org/.
2. hueta language. http://www.scala-lang.org/.