Моделирование MDA

Корчемкин Дмитрий Александрович, гр. 422

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

научный руководитель: к.ф.-м.н., доцент А. И. Коробейников
рецензент: м.н.с. С. Ю. Нурк

Санкт-Петербург
2015 г.

Мотивация: ДНК

ДНК – полимер состоящий из нуклеотидов

- В клетке большую часть времени находится в виде двойной спирали из нуклеотидов
- Для многих приложений достаточно рассматривать как последовательность символов из алфавита $\{A, T, G, C\}$
- Нити двойной спирали комплементарны
- ДНК хранит информацию о синтезе белков

Мотивация: проблемы секвенирования

- Для многих задач нужно знать исходную последовательность целиком
- «Прочтение» ДНК целиком невозможно, необходимо большое количество перекрывающихся кусков
- Увеличение количества ДНК обычным путём (делением клеток) не всегда возможно без дополнительных условий
- Процедура MDA позволяет добиться многократного увеличения количества ДНК, но разные участки увеличиваются в разной мере

Мотивация: проблемы секвенирования

Процедура MDA позволяет добиться многократного увеличения количества ДНК, но разные участки увеличиваются в разной мере

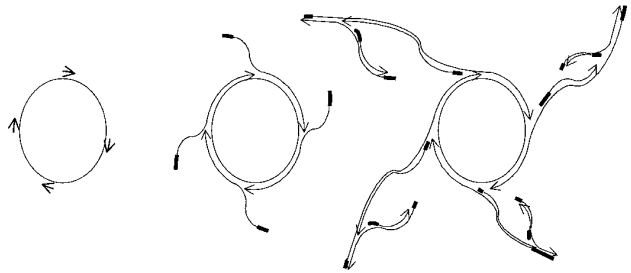


Рис.: Результат применения MDA

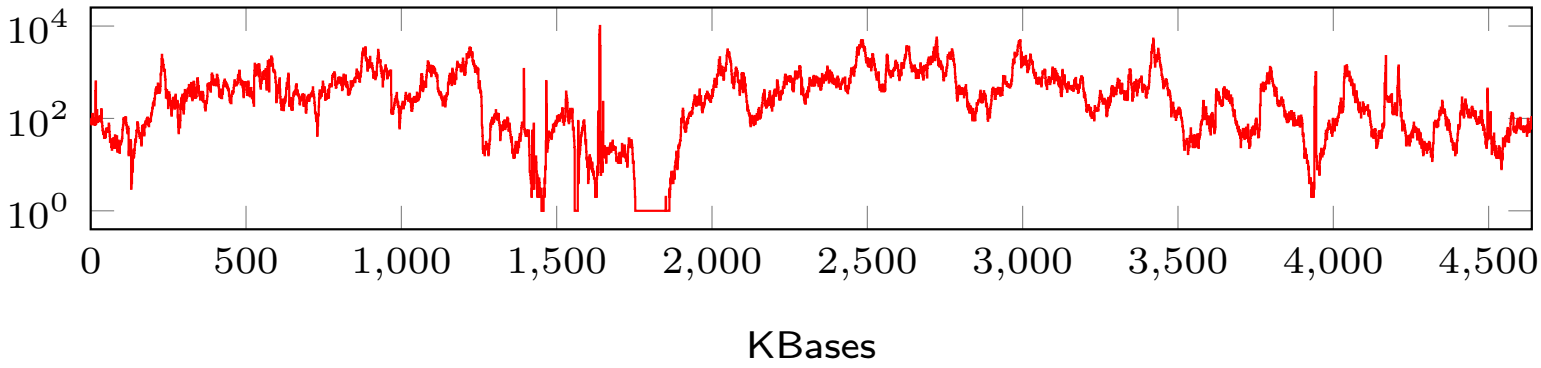


Рис.: Пример увеличения количества ДНК в зависимости от позиции

Постановка задачи

- Моделирование MDA с учётом возникающих ошибок и параметров эксперимента
- Изучение влияния параметров на «качество» увеличения количества ДНК

Свойства процесса

Важные свойства MDA:

- Вероятность ошибки в каждом копировании отдельного нуклеотида $< 10^{-3}$
- Strand displacement: разрыв части связей со старыми цепочками при образовании новых

Покрытие

- Цель применения MDA — увеличение количества «копий» исходной ДНК для её «прочтения»
- Разумная характеристика «качества» увеличения ДНК — количество пар связанных вхождений нуклеотида исходной цепочки в результат

Формализация и моделирование

Предлагается рассматривать процесс применения MDA как марковскую цепь; процесс одного этапа наращивания цепочек также рассматривается как марковская цепь.

- Исходная ДНК длиной порядка 10^6 увеличивается в 10^6 раз (т.е. $\sim 10^{12}$ нуклеотидов)
- Распределения покрытия не достаточно для идентификации полного состояния
- Позиции гибридизации праймеров присутствуют не только на цепочках, появившихся на предыдущем этапе

Проблема:
Необходимость хранения описания полной структуры цепочек и связей.

Формализация и моделирование

Проблема:
Необходимость хранения описания полной структуры цепочек и связей.
В работе предлагается алгоритм моделирования и структуры данных решающие проблему:

- Моделирования многократных успешных исходов
- Использования графоподобных структур для хранения «похожих» фрагментов
- Параллелизация за счёт уменьшения «зависимости» моделирования цепочек

Параметры моделирования

Цель: изучить зависимость результата применения MDA от параметров.
Параметры, варьируемые в экспериментах:

- Последовательность ДНК:*Escherichia coli*, *Rhodobacter sphaeroides*, *Staphylococcus aureus*
- Средняя продолжительность жизни полимераз: $\{10^4, 2 \cdot 10^4, 4 \cdot 10^4\}$
- «Плотность» праймеров: $\{\frac{1}{2000}, \frac{1}{1000}, \frac{1}{500}\}$ (праймеров на общую длину ДНК на начало этапа)
- Наборы вероятностей событий:
  - Моделирование с ошибками
  - Моделирование без ошибок

(54 набора параметров)

Результаты

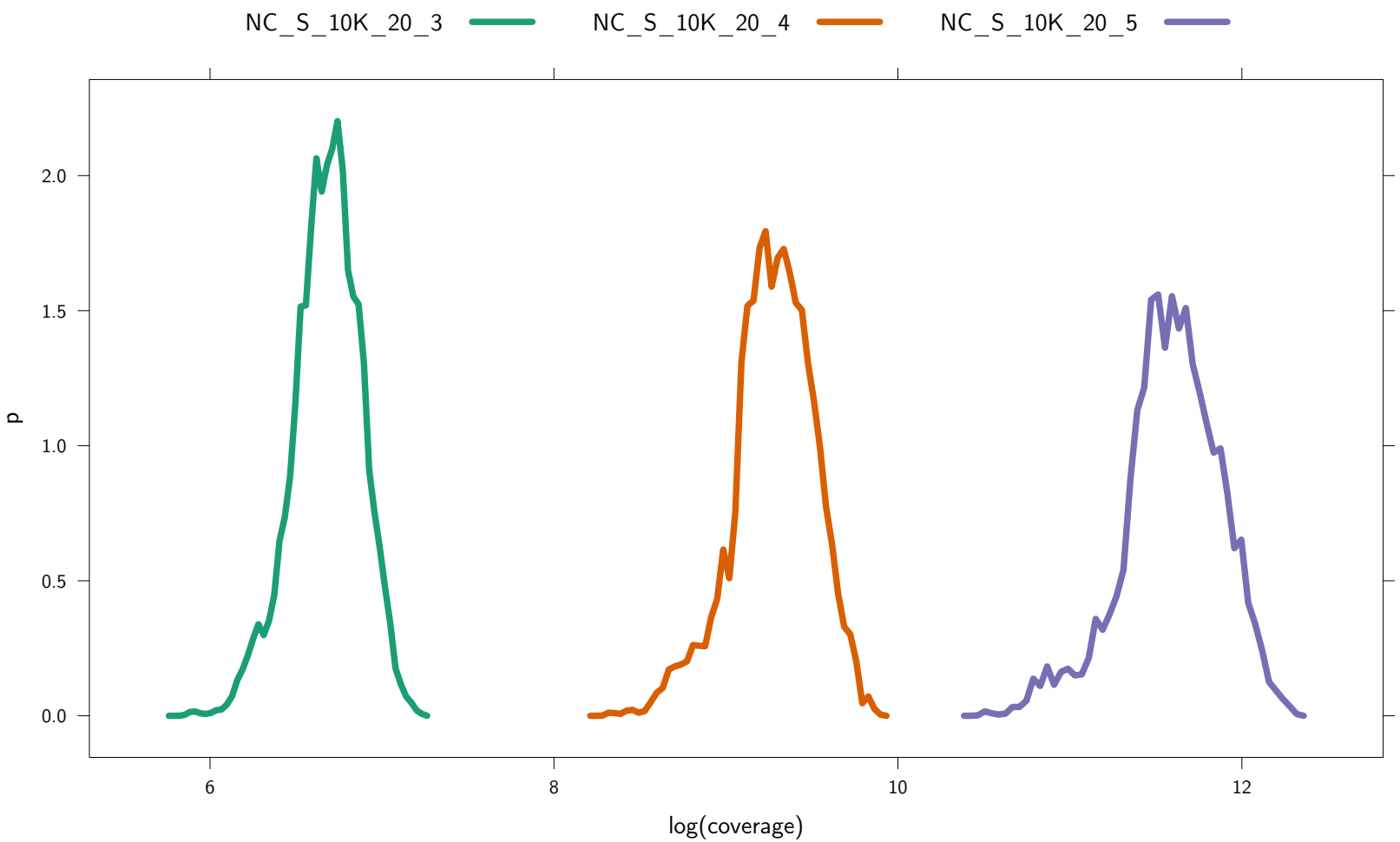


Рис.: «Плотность» распределения логарифма покрытия (на разных итерациях)

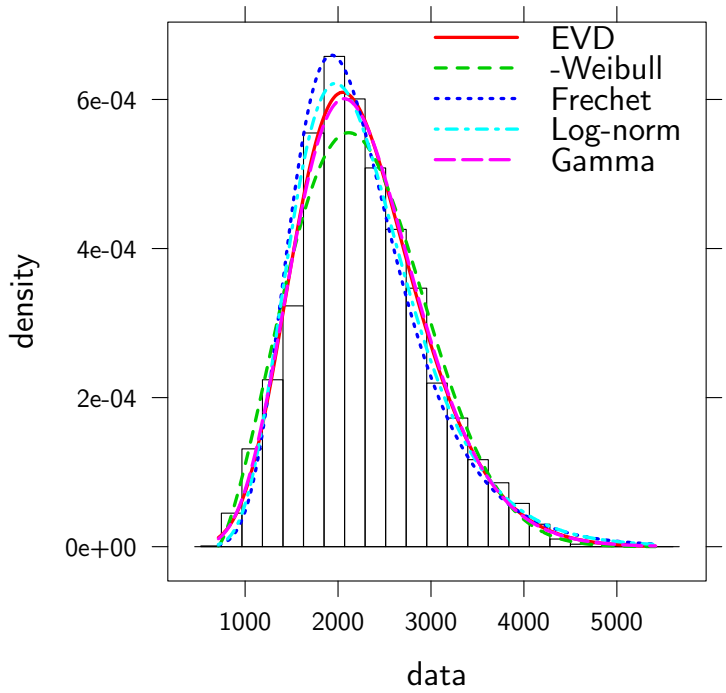
Сравнение с «классическими» распределениями

Сравним распределения покрытия, полученные путём моделирования, с несколькими распределениями:

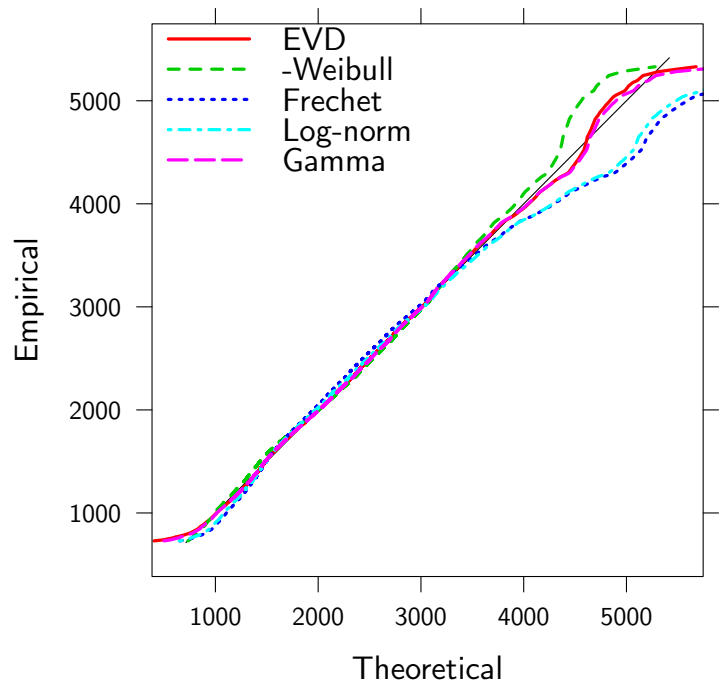
- Гамма
- Лог-нормальное
- EVD-семейство
  - Распределение Вейбулла
  - Распределение Фреше

Параметры распределений получены из численных MLE оценок.

Сравнение с «классическими» распределениями



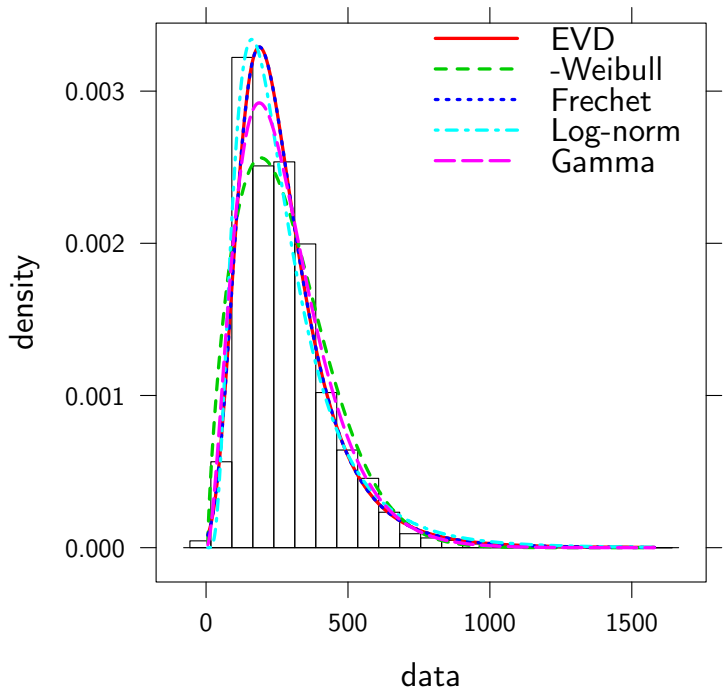
(a) Сравнение плотностей



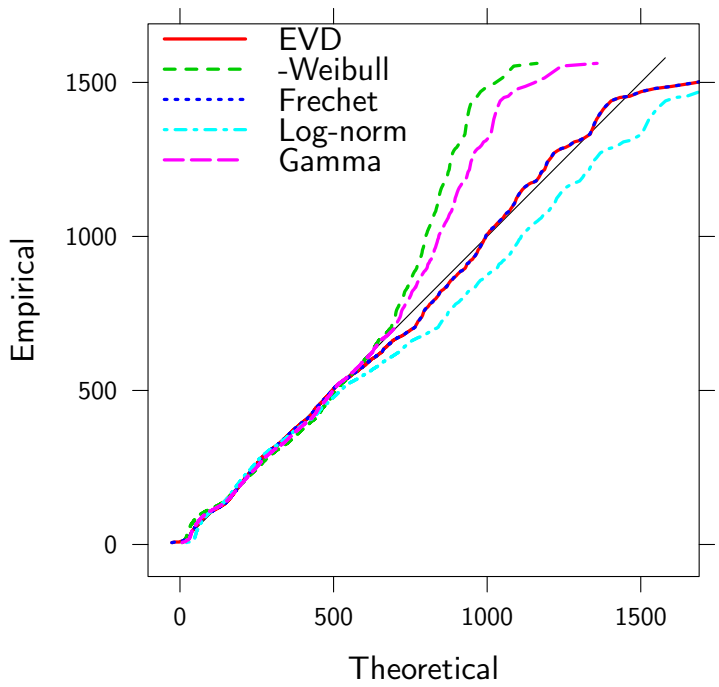
(b) QQ-plot

Рис.: Сравнение с «классическими» распределениями

Сравнение с «классическими» распределениями



(a) Сравнение плотностей



(b) QQ-plot

Рис.: Сравнение с «классическими» распределениями

Заключение

В работе рассмотрено моделирование процесса MDA, в частности:

- Предложен эффективный алгоритм моделирования и сопутствующие структуры данных, позволяющие моделировать MDA в достаточно общей модели
- Путём моделирования показано, что распределение покрытия можно рассматривать как распределение из EVD семейства

В то же время, существуют задачи, решение которых продолжит начатую работу:

- Формальное доказательство принадлежности предельного распределения к какому-либо семейству
- Изучение различия между распределениями при отсутствии и наличии ошибок
- Исследование влияния параметров на хвосты распределения