Оценивание расстояния между рёбрами в графе де Брюйна по неполным данным

Тарасов Артем Леонидович

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

Научный руководитель: к.ф.-м.н. Коробейников А.И.
Рецензент: м.н.с. Нурк С. Ю.

2015

Задача сборки генома (биоинформатика)

- $\mathcal{N} = \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ — алфавит *нуклеотидов*
- $g = g_1 g_2 \dots g_n$ — неизвестная строка ($g_i \in \mathcal{N}$) — *геном*
- длина генома n — 10^6–10^9 символов

Доступные данные

Выборка из пар подстрок g длины l — *парных чтений*

$$(g_\xi \dots g_{\xi+l-1}, \quad g_{\eta-l+1} \dots g_\eta), \text{ где}$$

- ξ, η — случайные позиции в строке g, $1 \le \xi < \eta \le n$
- ξ и $\eta - \xi$ независимы и не наблюдаемы
- $l \sim 10^2$–10^3; *размер вставки* $\eta - \xi + 1 \sim 10^2$–10^4
- размер выборки (количество парных чтений): 10^6–10^9

Общая задача

Восстановить как можно более длинные подстроки g

Граф де Брюйна

- входные данные: чтения $\mathcal{R} \subset \mathcal{N}^l$
- целочисленный параметр $k < l$

Несжатый граф

- вершины — все подстроки длины $k-1$
- рёбра — все подстроки длины k, встречающиеся в данных:

$$\{r_i r_{i+1} \dots r_{i+k-1} \ : \ r \in \mathcal{R},\ 1 \le i \le l-k+1\}$$

- ребро соединяет префикс и суффикс:

$$r_i \dots r_{i+k-1} \ : \ r_i \dots r_{\mathbf{i+k-2}} \ \rightarrow \ r_{\mathbf{i+1}} \dots r_{i+k-1}$$

Сжатый граф

- удаляются все вершины, в которые входит и выходит по единственному ребру
- инцидентные им рёбра „склеиваются“

Пример графа де Брюйна (k=4)

Входные данные:

ATGCTA
CTATGT
ATGTGC
GTGCGT

Рёбра:

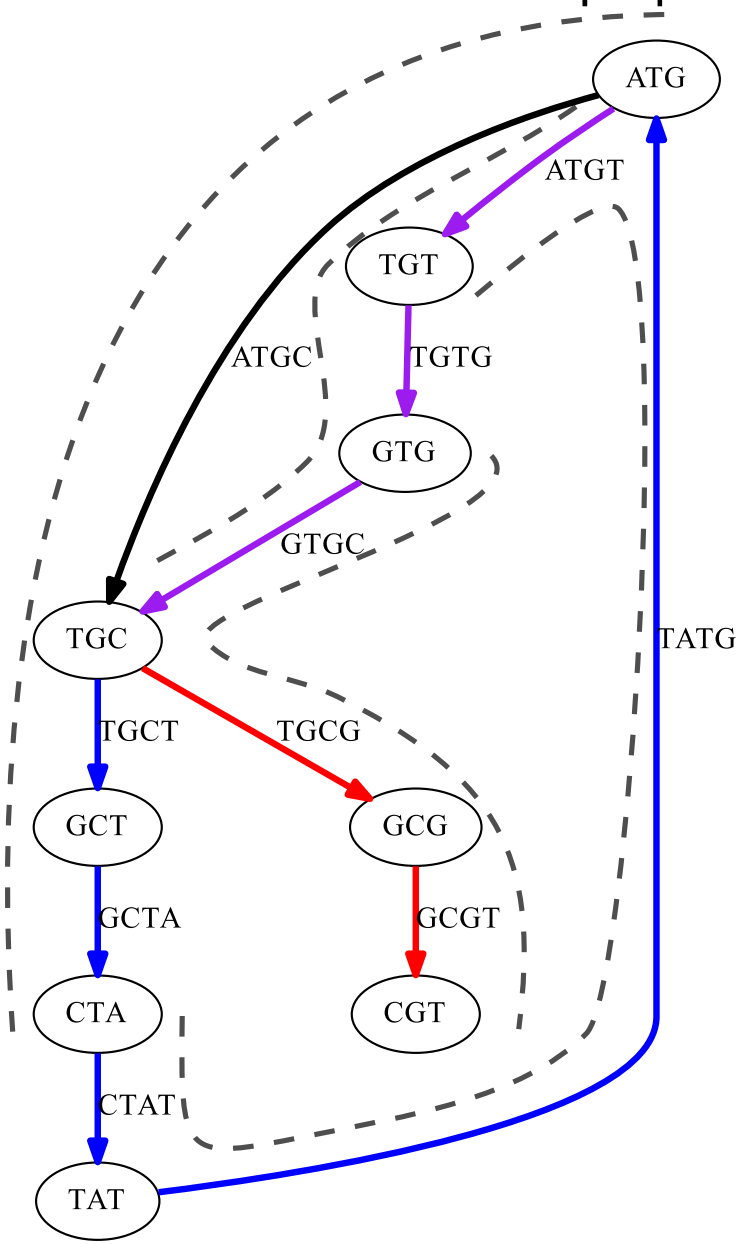
ATGC: ATG → TGC
TGCT: TGC → GCT
GCTA: GCT → CTA

CTAT: CTA → TAT
TATG: TAT → ATG
ATGT: ATG → TGT

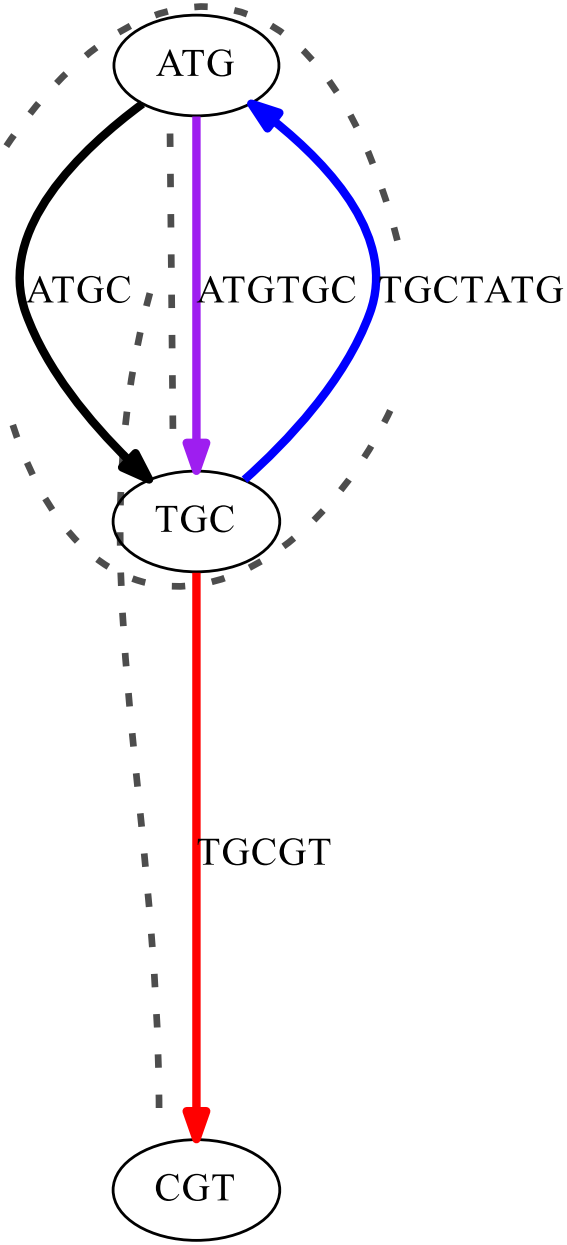
ATGT: ATG → TGT
TGTG: TGT → GTG
GTGC: GTG → TGC

GTGC: GTG → TGC
TGCG: TGC → GCG
GCGT: GCG → CGT

несжатый граф

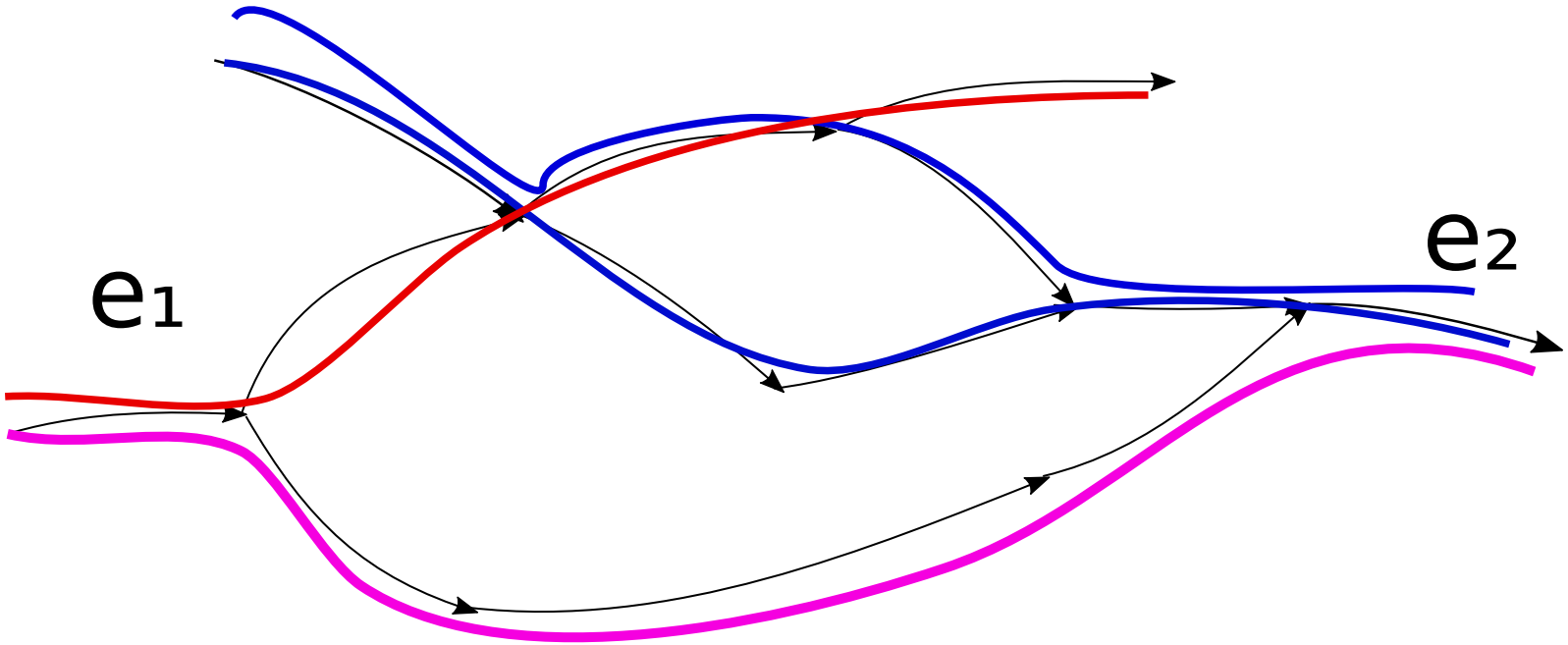


сжатый граф



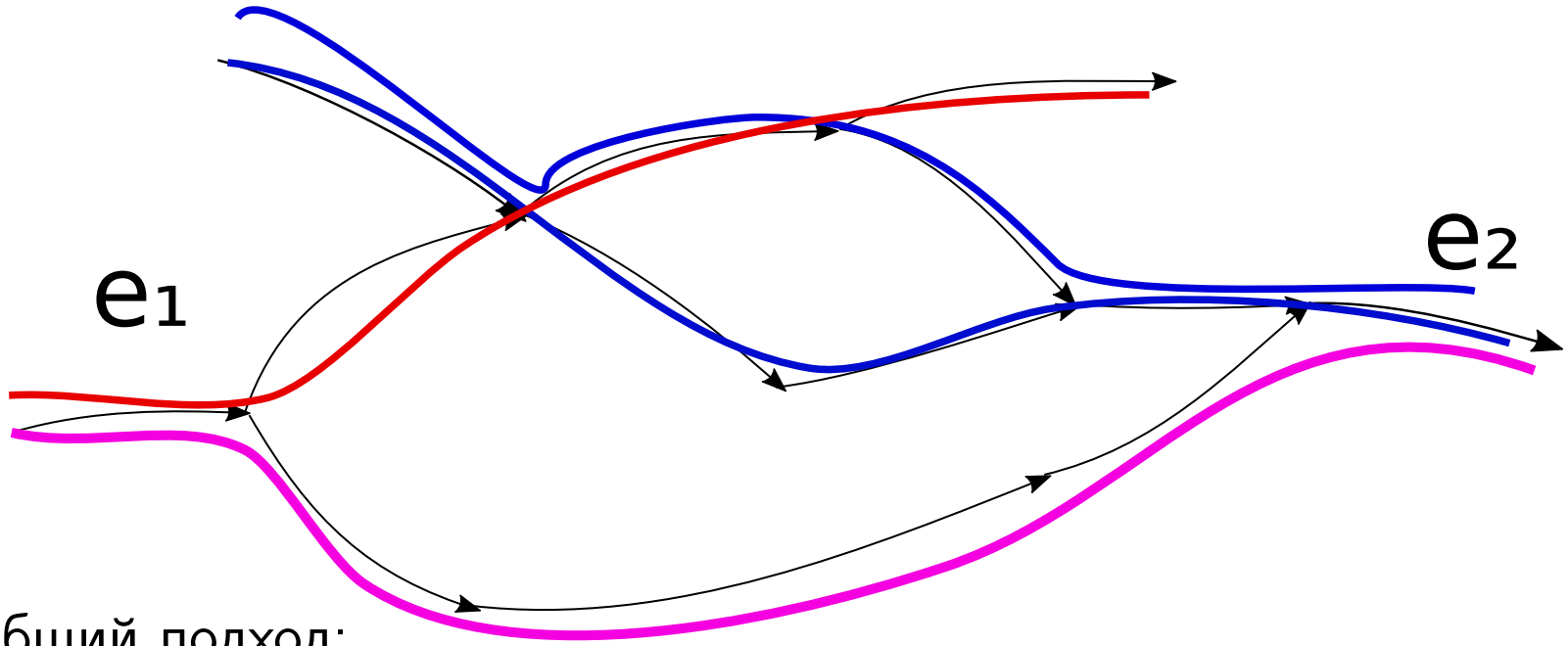
Назначение правдоподобий путям между двумя рёбрами

- Рассмотрим всевозможные пути между рёбрами e_1 и e_2
- Через какие проходит строка g?



Назначение правдоподобий путям между двумя рёбрами

- Рассмотрим всевозможные пути между рёбрами e_1 и e_2
- Через какие проходит строка g?



Общий подход:

- Получим множество длин путей $\mathcal{L}(e_1, e_2)$
- Сопоставим его с приложившимися к e_1 и e_2 чтениями
- Оценим вероятности прохождения строки g через пути

Простой метод назначения правдоподобий

Предложенный в [Bankevich et al. 2012] метод NAIVE:

- В качестве правдоподобия длины $l \in \mathcal{L}(e_1, e_2)$ берётся доля парных чтений, для которых вычисленный при этой длине размер вставки — ближайший к среднему

Недостатки:

- Метод статистически необоснован
- Используется крайне мало информации о распределении размера вставки (только среднее значение)

Идея нового метода

Использовать всё распределение размера вставки

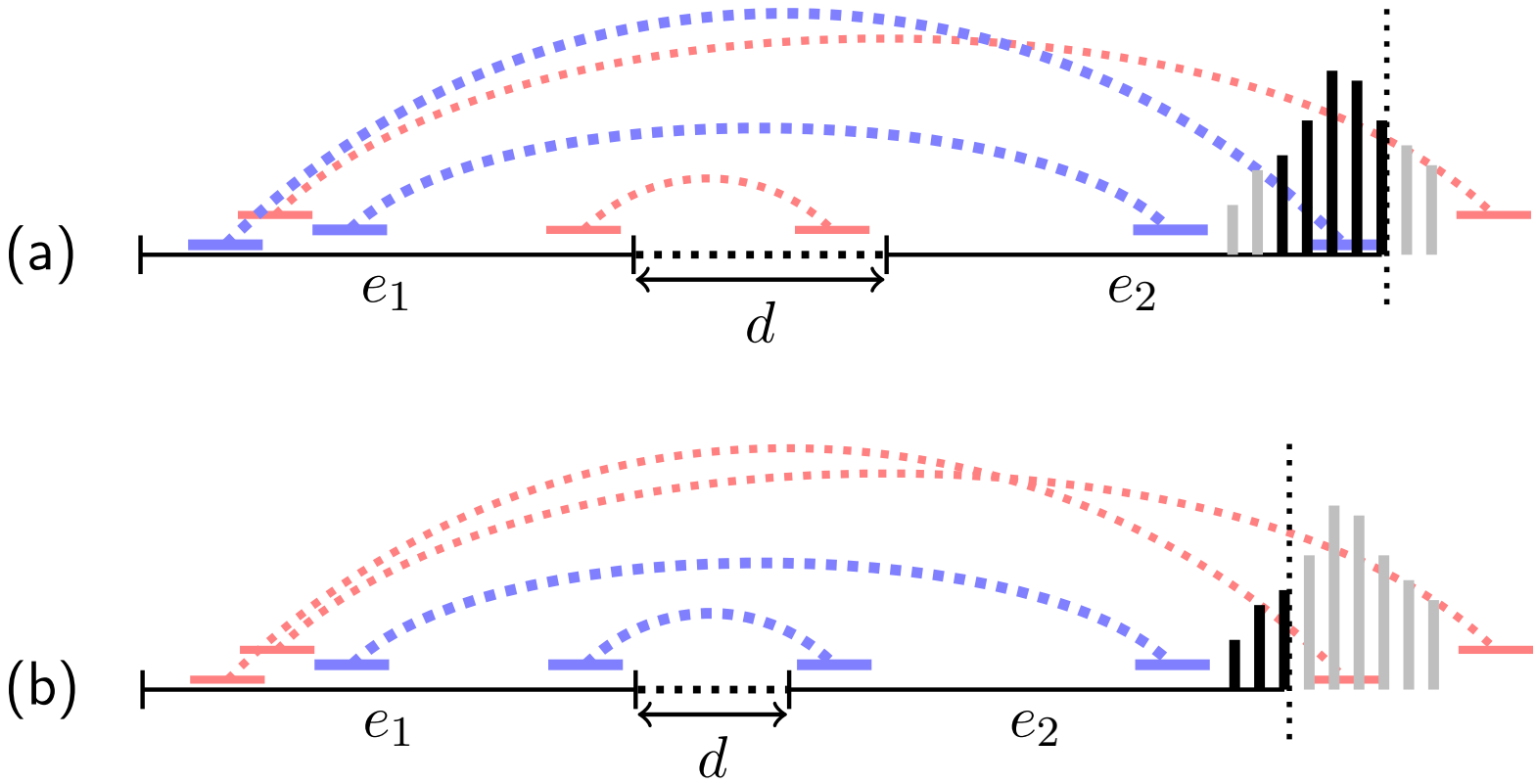
- Ф. р. размера вставки $f(m) = \mathcal{P}(x = m)$ хорошо оценивается с помощью парных чтений, приложившихся к длинным рёбрам графа
- Найдём ф. р. *видимого* размера вставки для фиксированных рёбер и длины пути l между ними $f_l(m) = \mathcal{P}\{x = m \mid \text{размер вставки } m \text{ наблюдаем}\}$
- Распределение видимого размера вставки — это смесь

$$\sum_{l \in \mathcal{L}(e_1, e_2)} \pi(l) f_l$$

- В качестве правдоподобий возьмём оценки параметров $\hat{\pi}(l)$, полученные применением EM-алгоритма

Неполнота данных

- Рассмотрим все парные чтения, приложившиеся к e_1 и e_2
- Не все размеры вставки наблюдаются



- В работе получено аналитическое выражение для функции распределения наблюдаемого размера вставки

Оценка вкладов расстояний

- Размер вставки i-го парного чтения равен $\psi_i + l$, где ψ_i — известно, $l \in \mathcal{L}(e_1, e_2)$ — неизвестная длина пути
- Для $\mathcal{L}(e_1, e_2) = \{l_1, \ldots, l_m\}$ вычисляются $\hat{f}^{(1)}, \ldots, \hat{f}^{(m)}$
- Ф.р. видимого размера вставки ищется в виде $\sum_{i=1}^{m} \pi_i \hat{f}^{(i)}$

EM

- Входные данные: $\{\psi_i\}_{1 \le i \le N}$
- Скрытая переменная: длина пути $l \in \mathcal{L}(e_1, e_2)$
- Размер вставки вычисляется как $\psi_i + l$
- $\prod_{i=1}^{N} \sum_{j=1}^{m} \pi_j \hat{f}^{(j)}(\psi_i + l_j) \to \max, \ \sum_{j=1}^{m} \pi_j = 1, \ \pi_j \ge 0$

Параметры смеси берутся в качестве правдоподобий:

$$\mathrm{EM}_{e_1,e_2}(l_j) = \hat{\pi}_j, \ 1 \le j \le m\,.$$

Оценка вкладов расстояний

- Размер вставки i-го парного чтения равен $\psi_i + l$, где ψ_i — известно, $l \in \mathcal{L}(e_1, e_2)$ — неизвестная длина пути
- Для $\mathcal{L}(e_1, e_2) = \{l_1, \ldots, l_m\}$ вычисляются $\hat{f}^{(1)}, \ldots, \hat{f}^{(m)}$
- Ф.р. видимого размера вставки ищется в виде $\sum_{i=1}^{m} \pi_i \hat{f}^{(i)}$

EM

- Входные данные: $\{\psi_i\}_{1 \le i \le N}$
- Скрытая переменная: длина пути $l \in \mathcal{L}(e_1, e_2)$
- Размер вставки вычисляется как $\psi_i + l$
- $\prod_{i=1}^{N} \sum_{j=1}^{m} \pi_j \hat{f}^{(j)}(\psi_i + l_j) \to \max, \ \sum_{j=1}^{m} \pi_j = 1, \ \pi_j \ge 0$

Параметры смеси берутся в качестве правдоподобий:

$$\mathrm{EM}_{e_1,e_2}(l_j) = \hat{\pi}_j, \ 1 \le j \le m\,.$$

Метод реализован на C++ (как часть программы для сборки геномов SPAdes [Bankevich et al. 2012])

Модификации базового метода

Метод EM использует не всю доступную информацию.
Предложены модификации, которые можно комбинировать (на практике не реализованы):

External-EM

использует информацию о том, к каким рёбрам приложились *внешние* концы парных чтений

Internal-EM, Narrowing-EM

используют информацию о том, к каким рёбрам приложились *внутренние* концы парных чтений

Extensive-EM

увеличивает выборку засчёт парных чтений, приложившихся к некоторым из соседних рёбер

Достоверность метода на паре рёбер

- *метод* $\mathcal{M}$ — это семейство функций $\mathcal{M}_{e_1,e_2} : \mathbb{N} \to [0, +\infty)$, где e_1, e_2 — рёбра графа де Брюйна
- ранее введены методы $\textsc{Naive}$, $\textsc{EM}$ и его модификации
- пути, через которые проходит g, назовём *истинными*
- введём „идеальный" метод:

$$\textsc{Ideal}_{e_1,e_2}(l) = \begin{cases} 1, & \exists \text{ истинный путь длины } l \text{ из } e_1 \text{ в } e_2 \\ 0 & \text{иначе} \end{cases}$$

- *достоверностью* метода $\mathcal{M}$ на паре рёбер e_1, e_2 назовём долю правдоподобия, соответствующего истинным путям:

$$\kappa_{\mathcal{M}}(e_1, e_2) = \frac{\sum\limits_{l \in \mathcal{L}(e_1,e_2)} \textsc{Ideal}_{e_1,e_2}(l) \times \mathcal{M}_{e_1,e_2}(l)}{\sum\limits_{l \in \mathcal{L}(e_1,e_2)} \mathcal{M}_{e_1,e_2}(l)}$$

Точность метода

- Обозначим рёбра, к которым прикладываются внутренние концы парного чтения $(r_1, r_2) \in \mathcal{N}^l \times \mathcal{N}^l$, как $\mathcal{E}(r_1)$ и $\mathcal{E}(r_2)$
- Выделим из исходного набора парных чтений $\mathcal{R} \subset \mathcal{N}^l \times \mathcal{N}^l$ нетривиальные в смысле определения истинности путей: $\mathcal{R}' = \{(r_1, r_2) \in \mathcal{R} : e_1 = \mathcal{E}(r_1) \neq \mathcal{E}(r_2) = e_2,\ |\mathcal{L}(e_1, e_2)| > 1\}$

Определение

Точностью метода $\mathcal{M}$ на строке g и наборе парных чтений R' будем называть величину

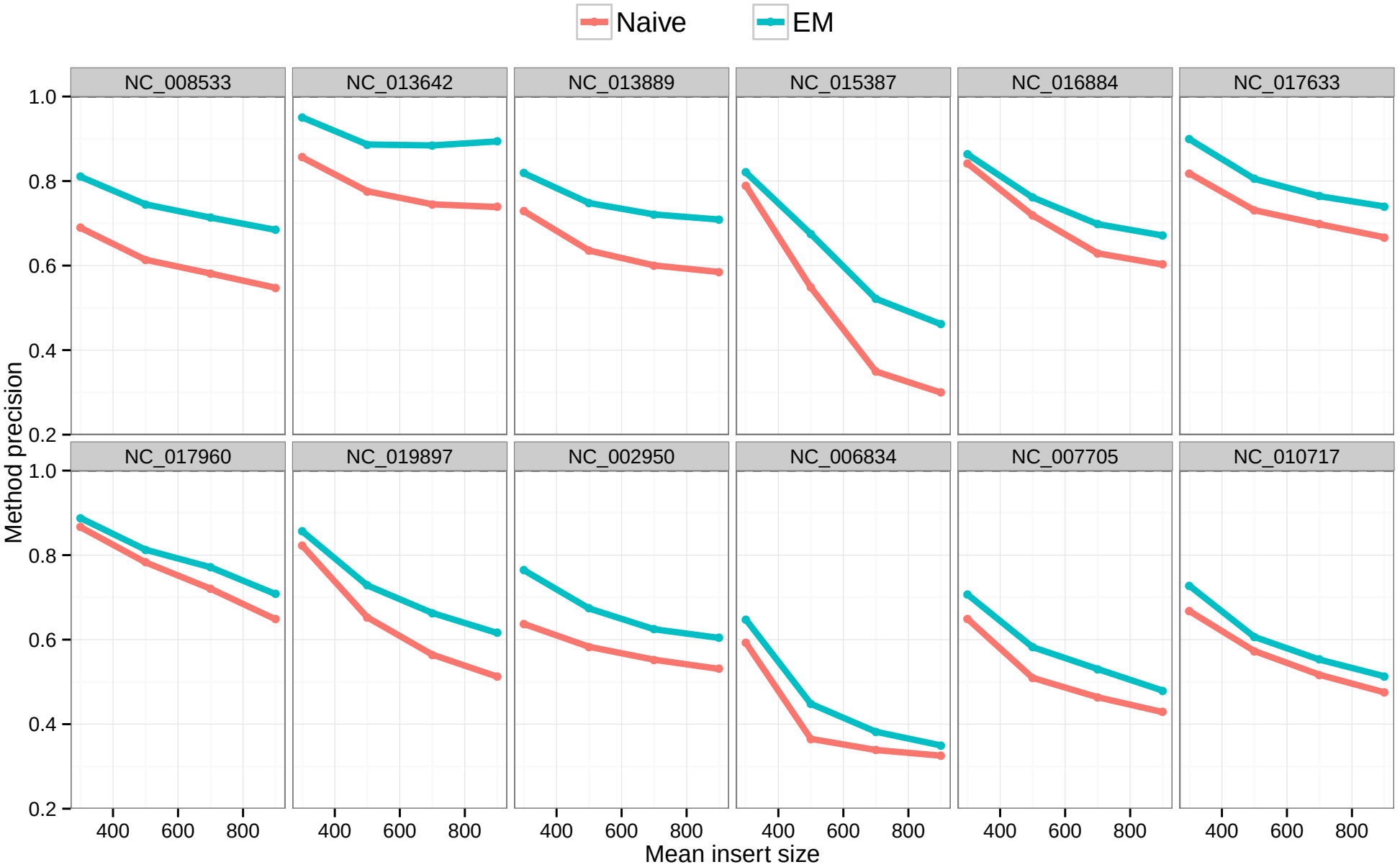
$$\mathrm{Prec}_{\mathcal{M}} = \frac{1}{|\mathcal{R}'|} \sum_{(r_1, r_2) \in \mathcal{R}'} \kappa_{\mathcal{M}}\left(\mathcal{E}(r_1), \mathcal{E}(r_2)\right)$$

Это оценка вероятности того, что для случайной пары чтений (r_1, r_2) будет выбрана такая строка s, что $r_1 s r_2$ — подстрока g.

Подход к моделированию

- несколько реальных геномов размера 10^6–10^8
- 10 миллионов парных чтений
  - длина чтения — 100 нуклеотидов
  - распределение позиции ξ — дискретное равномерное
  - размер вставки каждого чтения моделируется как $\lfloor \kappa \rfloor$, где $\kappa \sim \mathcal{N}(a, \sigma^2)$, $300 \leq a \leq 900$, $\sigma = 50$
- параметр графа де Брюйна $k = 55$

Точность методов на различных геномах



Заключение

Результаты:

- Предложен новый метод назначения правдоподобий (EM)
- Показано, что точность EM на практике выше, чем Naive
- Также предложены различные модификации нового метода

Дальнейшая работа:

- Программная реализация модификаций (Extensive-EM, External-EM, Narrowing-EM, Internal-EM)
- Исследование того, как эффективно использовать полученные веса для построения более длинных подстрок (назначение весов рёбрам, продолжающим путь)

EM-алгоритм

Параметры алгоритма: $\textsc{kMaxIter} = 50$, $\textsc{kMaxDiff} = 10^{-5}$
Входные данные: $\{\psi_i\}_{i=1}^{N}$

$\mathbf{w} \leftarrow \mathbf{1}_m/m$
repeat
  $\mathbf{w}' \leftarrow \mathbf{0}_m$
  for $1 \le i \le N$ **do**
    $\mathbf{r} \leftarrow \left(\mathrm{w}_1 \hat{f}^{(1)}(l_1 + \psi_i), \ldots, \mathrm{w}_m \hat{f}^{(m)}(l_m + \psi_i)\right)^T$
    $\mathbf{w}' \leftarrow \mathbf{w}' + \mathbf{r}/\|\mathbf{r}\|_1$
  $\mathbf{w}' \leftarrow \mathbf{w}'/\|\mathbf{w}'\|_1$
  $\delta \leftarrow \|\mathbf{w} - \mathbf{w}'\|_1$
  $\mathbf{w} \leftarrow \mathbf{w}'$
until $i > \textsc{kMaxIter}$ **or** $\delta < \textsc{kMaxDiff}$