Задачи оценивания геномного расстояния на графах де Брёйна

Константинов Антон Владимирович, гр. 15.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к.ф.-м.н., доцент Коробейников А. И.
Рецензент: м.н.с. Шлемов А. Ю.

Санкт-Петербург
2019

Задача сборки генома

Основные термины

- **Геномом** будем называть строку $\mathcal{S}$ над четырёхбуквенным алфавитом $\{A, T, G, C\}$.
- **Рид** (или **прочтение**) — короткая подстрока $\mathcal{S}$.
- k-**мер** — подстрока $\mathcal{S}$, имеющая длину k.

Пусть $\mathfrak{R}$ — набор ридов для генома $\mathcal{S}$.

Задача сборки генома

По набору строк $\mathfrak{R}$ восстановить («собрать») как можно более длинные **контиги** — непрерывные подстроки исходной строки $\mathcal{S}$. В идеале хочется получить всю строку $\mathcal{S}$ целиком.

Граф де Брёйна

Пусть k — положительное целое число.

Сжатый граф де Брёйна строки $\mathcal{S}$ — направленный мультиграф следующей конструкции:

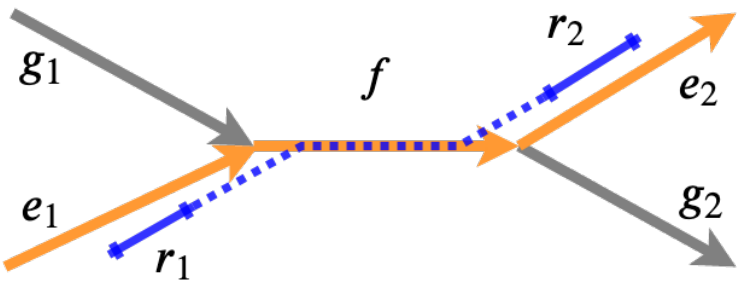
1. Множество вершин графа — множество всех k-меров строки $\mathcal{S}$.
2. Для каждого $(k+1)$-мера, содержащегося в $\mathcal{S}$, в граф добавляется ребро $v_1 \to v_2$, где v_1 и v_2 — его префикс и суффикс длины k соответственно. Кратность такого ребра равна количеству вхождений соответствующего $(k+1)$-мера в геном.
3. Пути, не имеющие разветвлений, заменяются рёбрами путём конкатенации соответствующих $(k+1)$-меров.

Проблема повторов

Геномный путь

В графе де Брёйна строки $\mathcal{S}$ *существует* соответствующий исходной строке $\mathcal{S}$ *эйлеров путь*, то есть путь, проходящий по каждому ребру мультиграфа ровно столько раз, какова его кратность. Будем называть этот путь **геномным**.

Проблема: повторы последовательностей (длины больше k).



Пусть $\mathcal{S} = \mathbf{e_1 f e_2} \ldots \mathbf{g_1 f g_2}$, где $\mathbf{e}_i$, $\mathbf{g}_i$ и $\mathbf{f}$ — некоторые строки.

Как должен проходить геномный путь?

Идея: будем сравнивать геномные и графовые расстояния между ридами.

Постановка задачи

Зафиксируем пару $\mathbf{e}_1, \mathbf{e}_2$ рёбер графа де Брёйна. Будем предполагать, что

1. $\mathbf{e}_1 = \mathcal{S}[a, b]$ и $\mathbf{e}_2 = \mathcal{S}[c, d]$, где $a < c$;
2. $\mathbf{e}_1$ и $\mathbf{e}_2$ соединяет путь $\boldsymbol{p} = \mathbf{e}_1 \to p_1 \to \ldots \to p_m \to \mathbf{e}_2$.

Графовое расстояние: $d_{\text{graph}}(\mathbf{e}_1, \mathbf{e}_2; \boldsymbol{p}) = \sum_{i=1}^{m} |p_i| - (m+1)k$,

Геномное расстояние: $d_{\text{genome}}(\mathbf{e}_1, \mathbf{e}_2) = c - b$.

Определим множества

$$\mathbf{D}_{\text{graph}} = \left\{ d_{\text{graph}}(\mathbf{e}_1, \mathbf{e}_2; \boldsymbol{p}) \mid \boldsymbol{p} \text{ — путь, соединяющий } \mathbf{e}_1 \text{ с } \mathbf{e}_2 \right\},$$

$$\mathbf{D}_{\text{genome}} = \left\{ d_{\text{genome}}(\mathbf{e}_1^{(i)}, \mathbf{e}_2^{(j)}) \mid \mathbf{e}_s^{(t)} \text{ — } t\text{-ое вхождение } \mathbf{e}_s \text{ в геном } \mathcal{S} \right\},$$

Задача: Найти пересечение $\mathbf{D} = \mathbf{D}_{\text{graph}} \cap \mathbf{D}_{\text{genome}}$.

Из чего состоит библиотека ридов $\mathfrak{R}$?

Пусть

1. ξ — дискретная случайная величина с носителем $\{1, \ldots, |\mathcal{S}|\}$, имеющая смысл координаты в геноме,
2. η — независимая от ξ неотрицательная целочисленная случайная величина (т. н. **длина вставки**),
3. ℓ — положительное целое число (**длина рида**).

1. Фрагмент — подстрока генома, имеющая вид $\mathcal{S}[\xi, \xi + \eta]$;
2. Левый рид — префикс длины ℓ фрагмента, т. е. подстрока $\mathcal{S}[\xi, \xi + \ell]$;
3. Правый рид — суффикс длины ℓ фрагмента, т.е. подстрока $\mathcal{S}[\xi + \eta - \ell, \xi + \eta]$.

Вероятностный подход к задаче

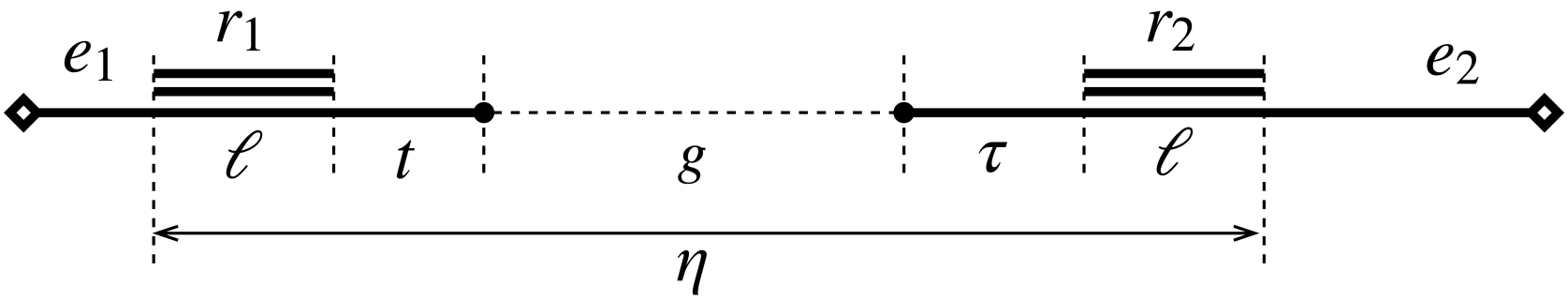


Рис. 1: Расположение ридов на рёбрах графа

Пусть $(r_1, r_2) \in \mathfrak{R}$, и r_i является подстрокой ребра $\mathbf{e}_i$ $(i = 1, 2)$.

Введём обозначения:

1. g — геномное расстояние между $\mathbf{e}_1$ и $\mathbf{e}_2$,
2. t — расстояние от конца r_1 до конца $\mathbf{e}_1$,
3. τ — координата начала r_2 на $\mathbf{e}_2$.

Вероятностный подход к задаче

Рассмотрим формально выборку $\Big((t_1, \tau_1, g_1), \ldots, (t_n, \tau_n, g_n)\Big)$.

1. Реализации (t, τ) наблюдаются только при условии $A_{\mathbf{e}_2}(r_2) = \{\text{рид } r_2 \text{ приложен к } \mathbf{e}_2\}$ (будем считать, что r_1 уже приложен);
2. Реализации g не наблюдаются вовсе.

При этом

1. Совместное распределение вектора (t_i, τ_i) зависит от g_i как от параметра.
2. t_i, τ_i и g_i связаны соотношением $\tau_i = \eta_i - t_i - g_i - 2\ell$, где $g_i \in \mathbf{D}$.

Получаем набор реализаций $\mathbb{T} = \Big((t_1, \tau_1), \ldots, (t_n, \tau_n)\Big)$.

В этом случае исходная задача сводится к статистическому выводу для g_i по $\mathbb{T}$.

Апостериорное распределение для одной реализации

Было получено выражение для функции вероятности $p(g \mid t, \tau, A_{\mathbf{e}_2})$.

Предложение

Пусть длина вставки η имеет распределение $\mathcal{P}_\eta$ с функцией распределения $F(x) = \mathbb{P}(\eta < x)$. Будем считать, что априорно g равномерно распределена на $\mathbf{D}_{\text{graph}}$.

Тогда

$$p(g \mid t, \tau, A_{\mathbf{e}_2}) = \frac{q(\tau, g, t)}{\sum_{j=1}^{k} q(\tau, g^{(j)}, t)},$$

где

$$q(x, y, z) = \frac{F\big(x + y + z + 2\ell + 1\big) - F\big(x + y + z + 2\ell\big)}{F(y + z + \ell + M) - F(y + z + 2\ell)}.$$

Переход к случаю нескольких реализаций

- На практике для каждого рида $(r_1, r_2) \in \mathfrak{R}$ реализуется собственное расстояние $g^{(i)} \in \mathbf{D}_{genome}$ для некоторого i.
- Поэтому нельзя напрямую сделать переход к повторной независимой выборке, как это обычно бывает в статистике.

Приходим к модели смеси:

$$(t, \tau) \sim \sum_{i=1}^{k} \pi_i \mathcal{L}_{\tau,t}\left(g^{(i)}\right), \text{ где } \pi_i \geq 0 \text{ и } \sum_{i=1}^{k} \pi_i = 1.$$

Здесь π_i мы можем оценить, усредняя апостериорную вероятность $p(g^{(i)} \mid t, \tau, A_{\mathbf{e}_2})$ по всем имеющимся реализациям.

Апостериорное распределение для n реализаций

Получено следующее утверждение, дающее апостериорное распределение g при условии набора реализаций (t, τ).

Предложение

Пусть длина вставки η имеет распределение $\mathcal{P}_\eta$ с функцией распределения $F(x) = \mathbb{P}(\eta < x)$. Будем считать, что априорно g равномерно распределена на $\mathbf{D}_{\text{graph}}$.

Тогда

$$p(g \mid \mathbb{T}, A_{\mathbf{e}_2}) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{q(\tau_i, g, t_i)}{\sum_{j=1}^{k} q(\tau_i, g^{(j)}, t_i)} \right],$$

где

$$q(x, y, z) = \frac{F\big(x + y + z + 2\ell + 1\big) - F\big(x + y + z + 2\ell\big)}{F(y + z + \ell + M) - F(y + z + 2\ell)}.$$

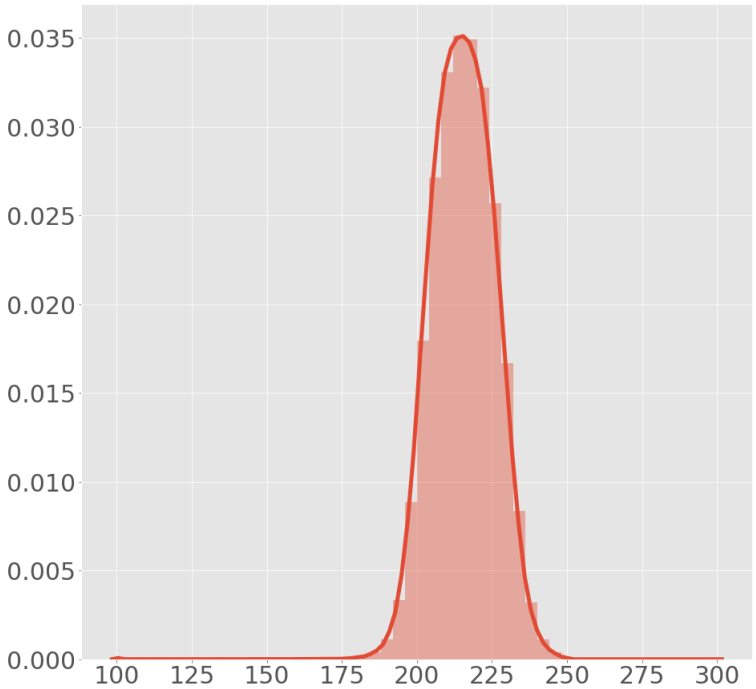
Данные

Во всех следующих примерах используются графы де Брёйна, построенные по различным библиотекам ридов для первых 400 тысяч нуклеотидов генома *E.coli* (штамм *K12 MG1655*).

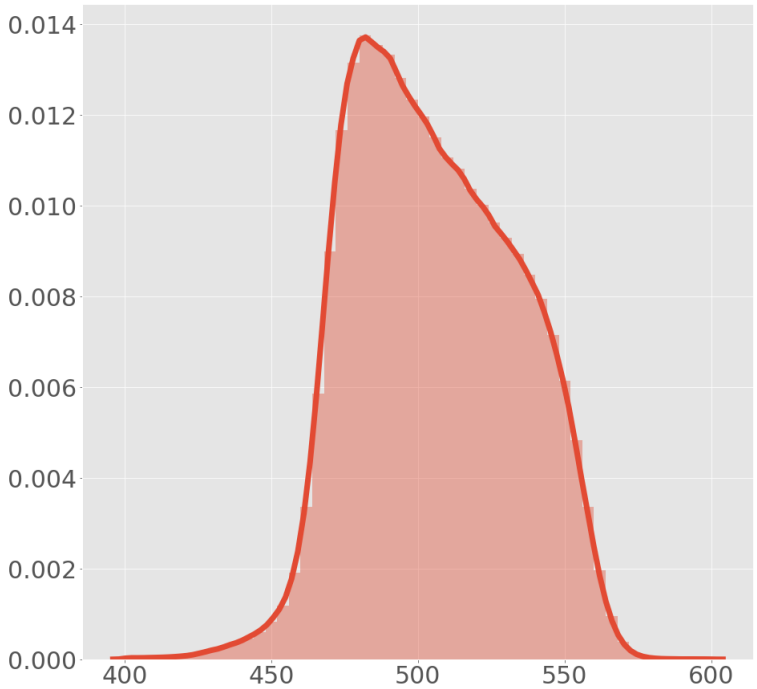
Реальные риды. Были рассмотрены две библиотеки:

1. Первая («Библиотека А») имеет близкое к нормальному распределение η. Использовалась ф. р. нормального распределения с оценёнными параметрами ($\mu \approx 215$, $\sigma \approx 10$).
2. Для второй библиотеки («Библиотека Б») в качестве F использовалась эмпирическая ф. р. ($\operatorname{med} \eta \approx 480$).

Распределения длины вставки



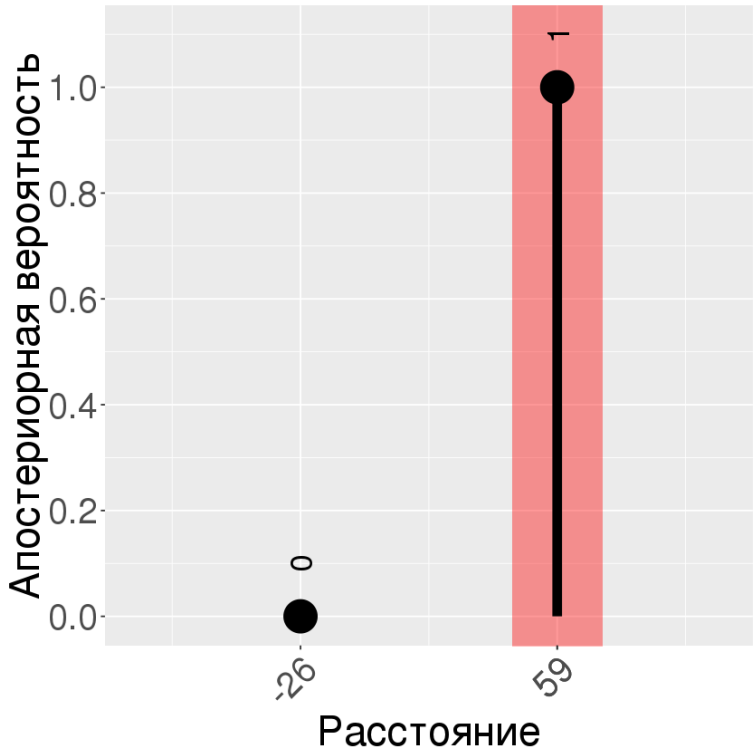
(a) Библиотека А. Хорошо аппроксимируется нормальным с параметрами $\mu = 215$, $\sigma = 10$



(b) Библиотека Б. Используем эмпирическую ф. р., $\mathrm{mode}\ \eta = 480$

Рис. 2: Распределения длины вставки для библиотек А и Б

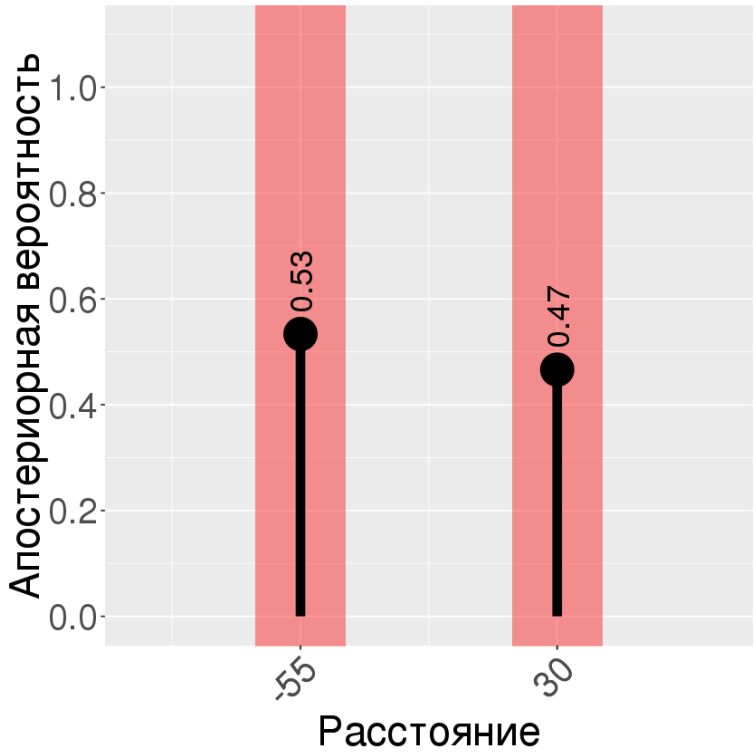
Библиотека A



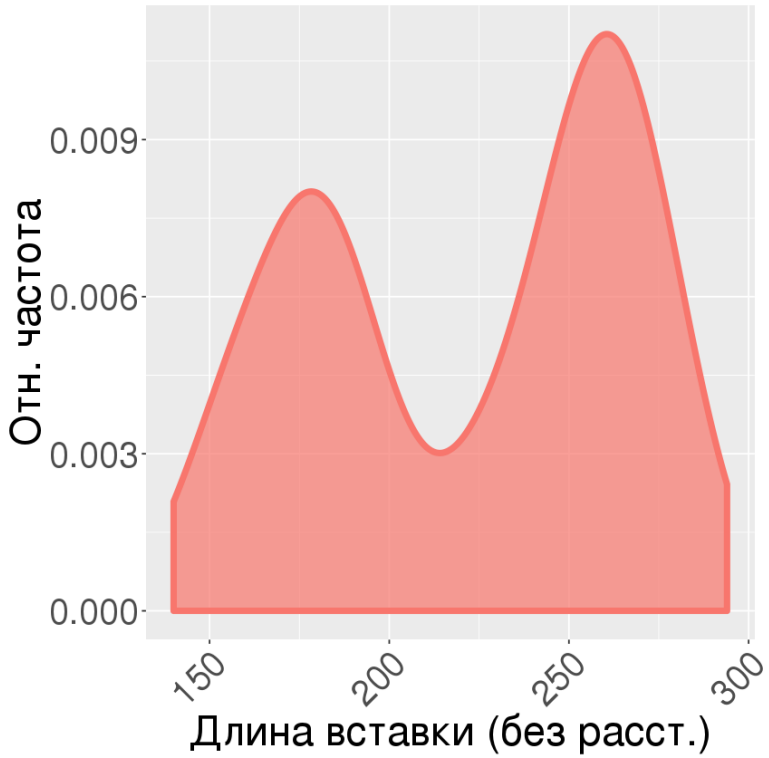
(a) Апостериорное распределение

(b) Гистограмма $\eta - g$

Рис. 3: Неповторные рёбра

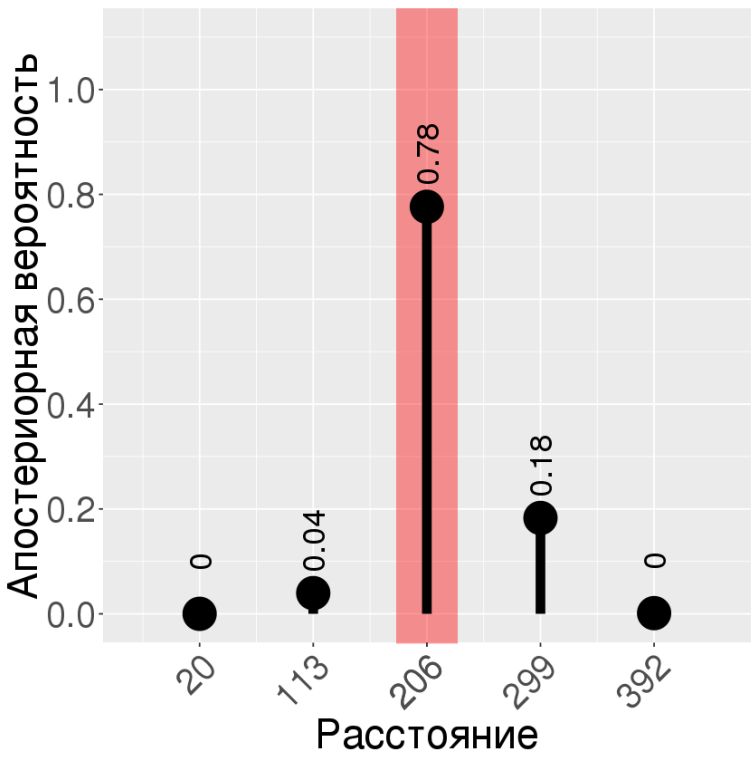


(a) Апостериорное распределение



(b) Гистограмма $\eta - g$

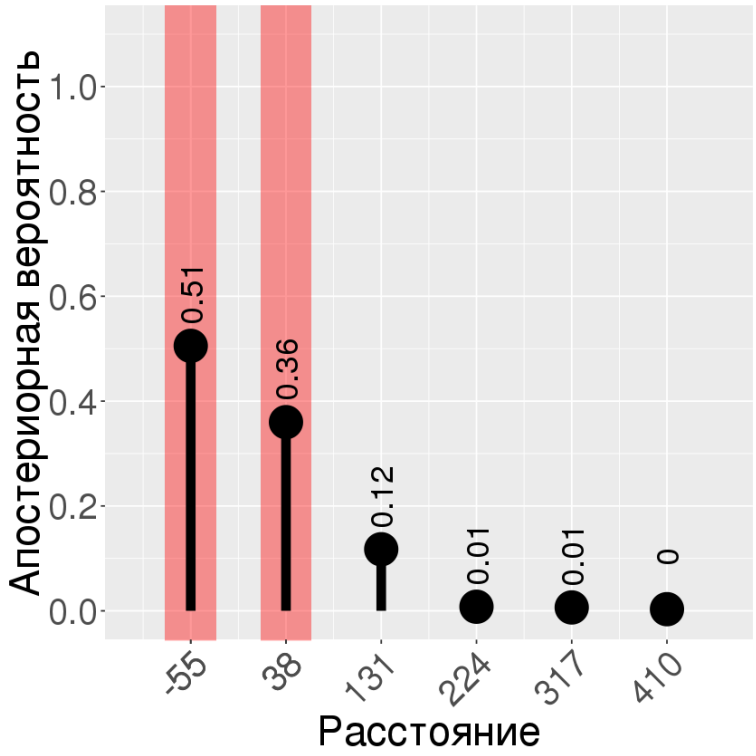
Рис. 4: Одно из рёбер имеет двойную кратность



(a) Апостериорное распределение

(b) Гистограмма $\eta - g$

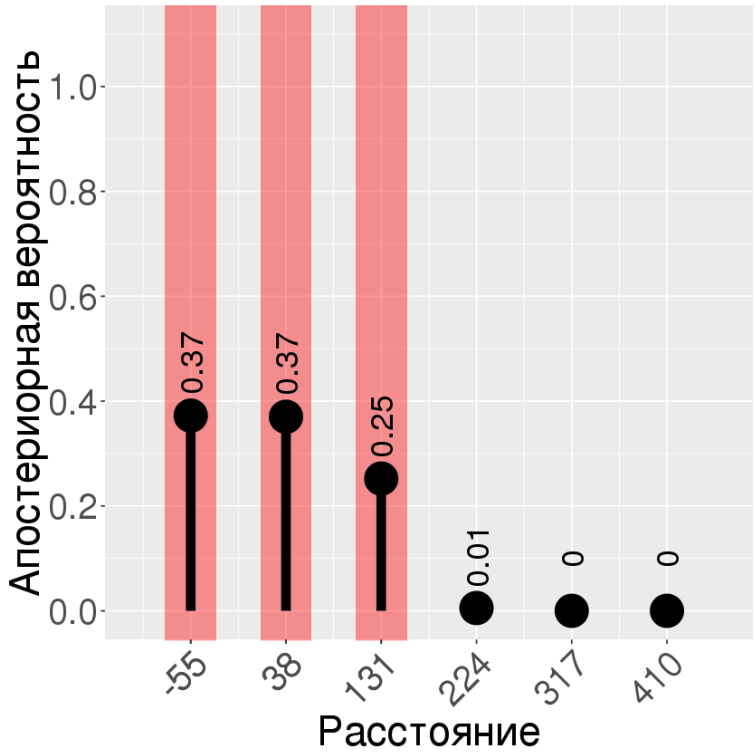
Рис. 5: Неповторные рёбра



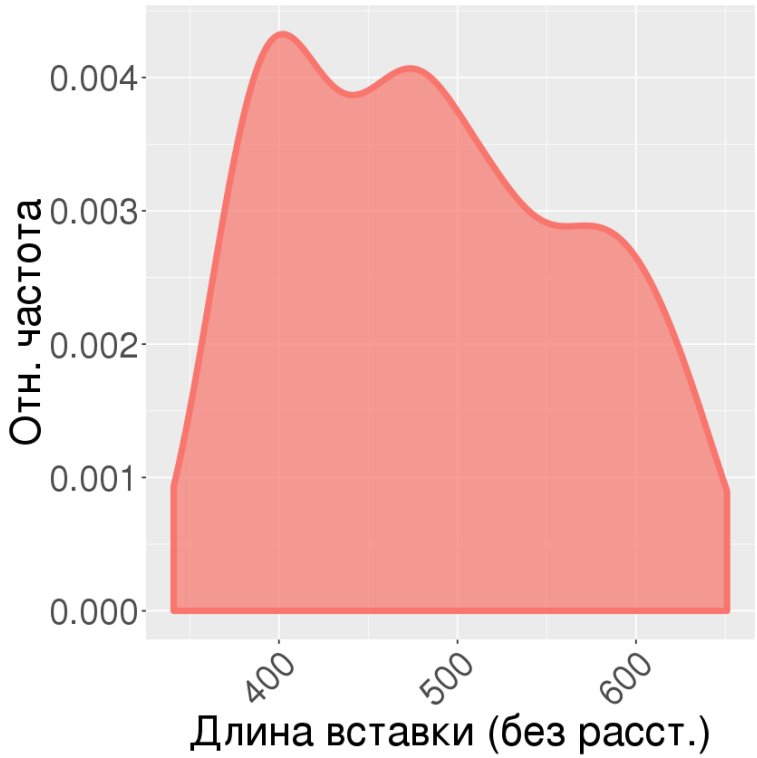
(a) Апостериорное распределение

(b) Гистограмма $\eta - g$

Рис. 6: Одно из рёбер имеет двойную кратность



(a) Апостериорное распределение



(b) Гистограмма $\eta - g$

Рис. 7: Одно из рёбер имеет тройную кратность

Заключение

В работе была рассмотрена задача оценки геномных расстояний между рёбрами в графе де Брёйна.

1. Построена вероятностная модель, позволяющая получать требуемые оценки в виде апостериорных вероятностей для расстояний, имеющихся в графе.
2. Построенная модель протестирована на реальных геномных данных.

В дальнейшем полученные оценки, например, могут быть применены в геномных ассемблерах для разрешения повторов в графе де Брёйна.