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

Сборка генома

Геном — строка над конечным алфавитом $\{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$.

- Размеры геномов у различных биологических видов варьируются в диапазоне от 100 тыс. до 150 млрд. символов.
- Не существует метода, позволяющего прочитать геном целиком.
- Вместо этого из генома случайным образом считываются подстроки, называемые *ридами*.
- Исходный геном затем должен быть восстановлен по этим подстрокам.

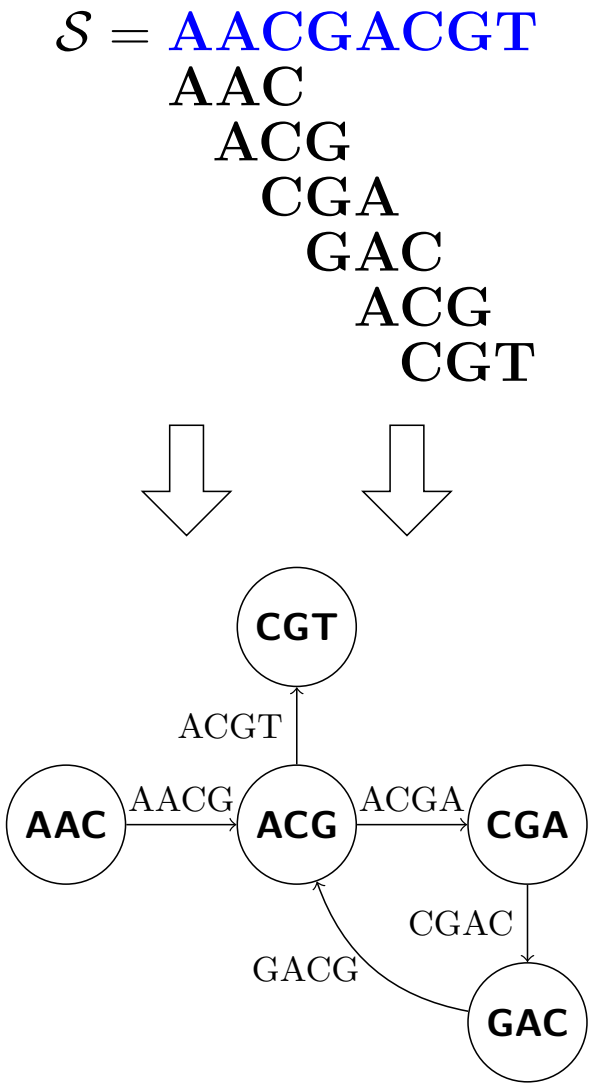
Граф де Брёйна

k-мер строки $\mathcal{S}$ — это её подстрока длины k.

Граф де Брёйна G, $k \in \mathbb{N}$:

1. Вершины — k-меры строки $\mathcal{S}$.
2. u и v соединены ребром кратности N, если $\mathcal{S}$ содержит N вхождений $k+1$-мера, имеющего префикс u и суффикс v.

Неформально говоря, граф де Брёйна состоит из всех подстрок длины k генома $\mathcal{S}$, которые соединены в том порядке, в котором они встречаются в $\mathcal{S}$.

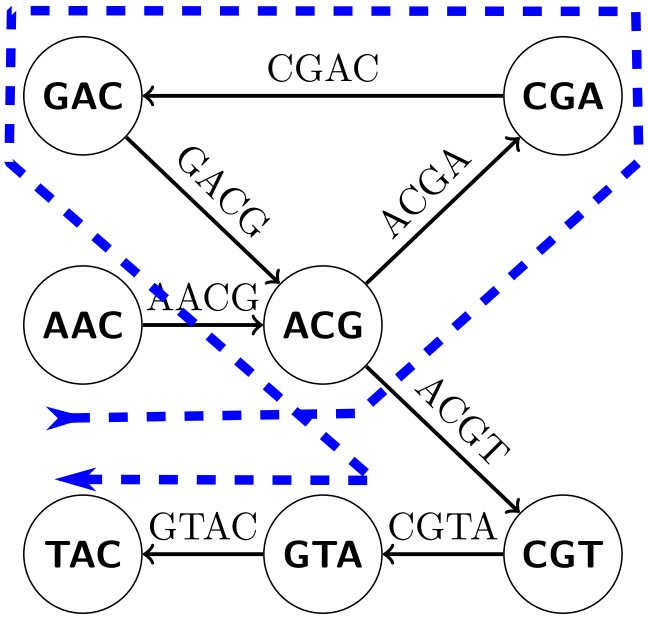


Сборка генома при помощи графа де Брёйна

Хорошо известно, что в графе де Брёйна G существует *эйлеров* (проходящий по всем рёбрам столько раз, какова их кратность) путь $\mathbf{p}$, который соответствует $\mathcal{S}$.

- В реальной ситуации G строится по k- и $k+1$-мерам, полученным из ридов.
- Если эйлеровых путей в полученном графе несколько, то неизвестно, какой из них соответствует $\mathcal{S}$.

$$\mathcal{S} = \mathbf{AACGACGTAC}$$
$$k = 3$$



Сборка генома $\iff$ поиск $\mathbf{p}$ среди всех эйлеровых путей в G.

Проблема повторов

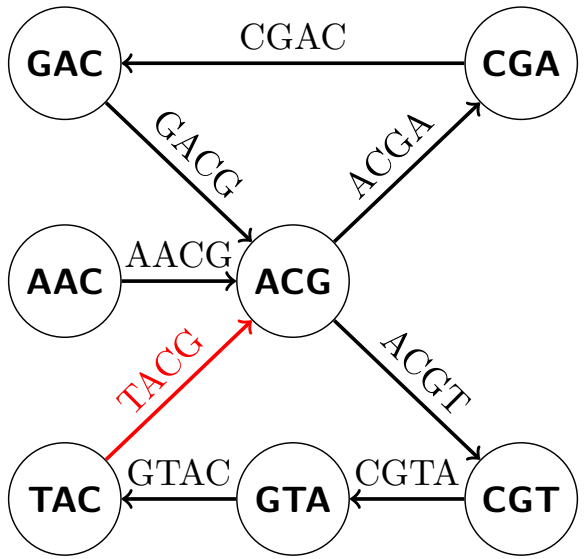
Проблема: повторы последовательностей (длины $\geq k$) приводят к образованию циклов.

$$\mathcal{S} = \mathbf{AACGACGTACG}$$
$$k = 3$$

Добавим к строке $\mathcal{S}$ из прошлого примера один символ **G** в конец.

Как теперь должен проходить геномный путь:

- По верхней петле, затем по нижней?
- Наоборот, сначала по нижней петле, затем по верхней?



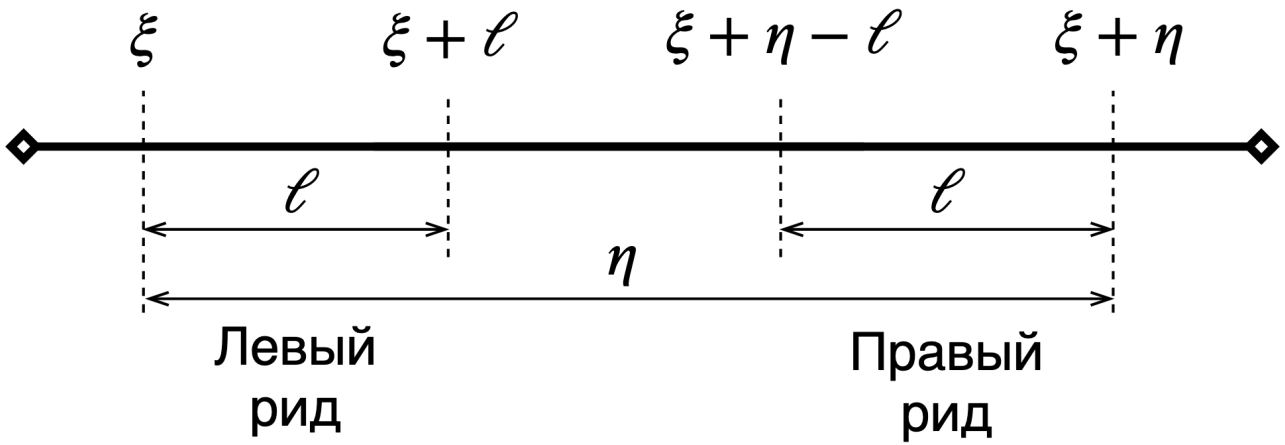
По графу ответить на этот вопрос невозможно!

Для разрешения повторов требуется внешняя информация.

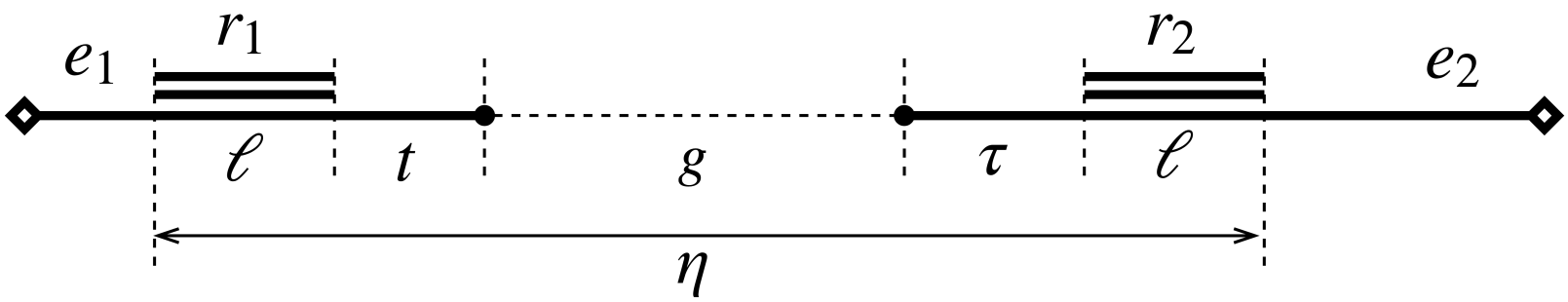
Вероятностная модель парных ридов

Пусть $\ell > 0$ — целое число. Рассмотрим независимые случайные величины $\xi \sim \mathrm{U}(\{1, \ldots, |\mathcal{S}|\})$ и $\eta > 0$.

1. Фрагмент — подстрока генома, имеющая вид $\mathcal{S}[\xi, \xi + \eta]$;
2. Парный рид — это пара из префикса $\mathcal{S}[\xi, \xi + \ell]$ (*левый* рид) и суффикса $\mathcal{S}[\xi + \eta - \ell, \xi + \eta]$ (*правый* рид) фрагмента.



Парные риды как внешняя информация



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 $\mathbf{e}_2$.

Тогда

$$\eta = t + g + \tau + 2\ell.$$

Зная распределение η, мы получаем информацию о расстоянии между рёбрами графа, «прикладывая» к ним риды.

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: Предложить алгоритм, определяющий элементы множества $\mathbf{D} = \mathbf{D}_{\text{graph}} \cap \mathbf{D}_{\text{genome}}$ при помощи набора (*библиотеки*) ридов $\mathfrak{R}$.

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 $g^{(i)} \in \mathbf{D}_{\text{genome}}$ для некоторого i.
- Поэтому нельзя напрямую сделать переход к повторной независимой выборке, как это обычно бывает в статистике.

Приходим к модели смеси:

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

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

Данные

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

Были проведены эксперименты на:

Синтетических ридах с длиной вставки $\eta \sim \mathrm{N}(\mu, \sigma^2)$:

- $\mu = 1000$, $\sigma = 30$.
- $\mu = 400$, $\sigma = 30$.

Реальных ридах. Были рассмотрены две библиотеки:

- Первая имеет близкое к нормальному распределение η. Использовалась ф. р. нормального распределения с оценёнными параметрами ($\mu \approx 215$, $\sigma \approx 10$).
- Для второй библиотеки в качестве F использовалась эмпирическая ф. р. ($\mathrm{med}\ \eta \approx 480$).

Пример: описание библиотеки

Рассмотрим условия, максимально приближенные к реальным:

- Медианная длина вставки: 480.
- В качестве функции распределения F, требуемой для получения оценок, мы будем использовать эмпирическую ф. р., полученную по всем имеющимся ридам.

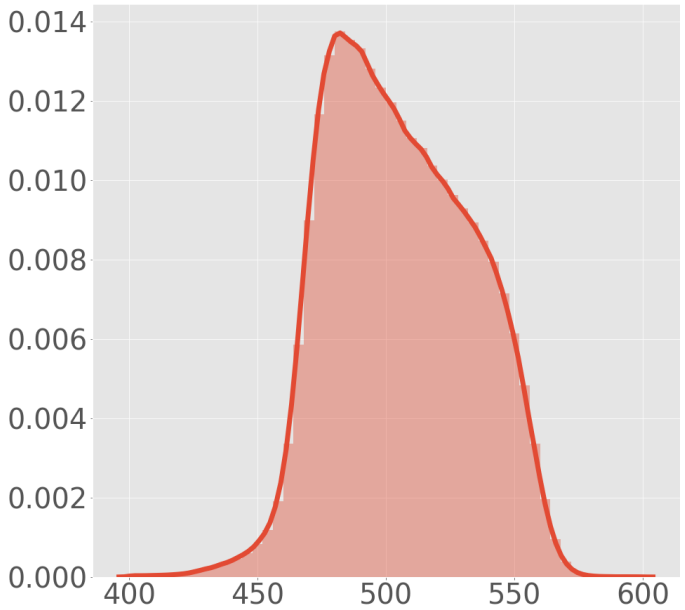
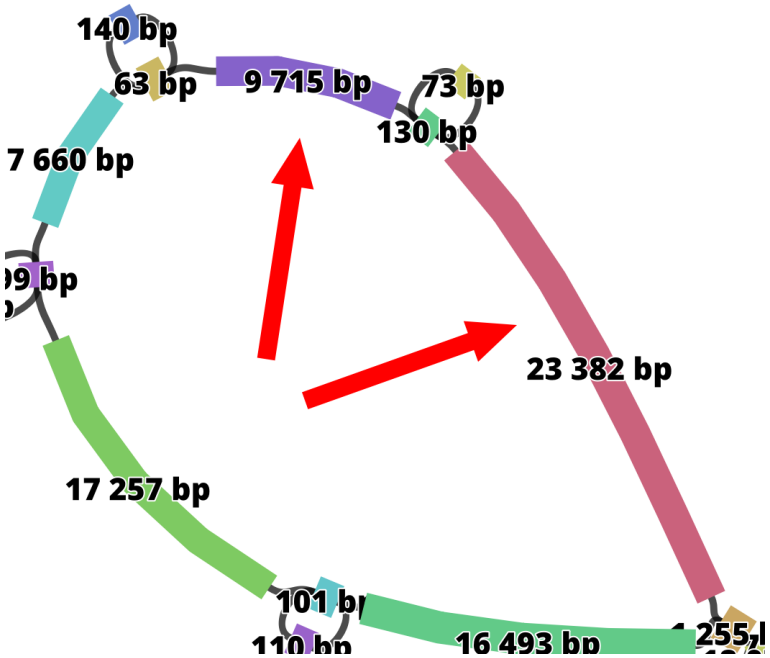
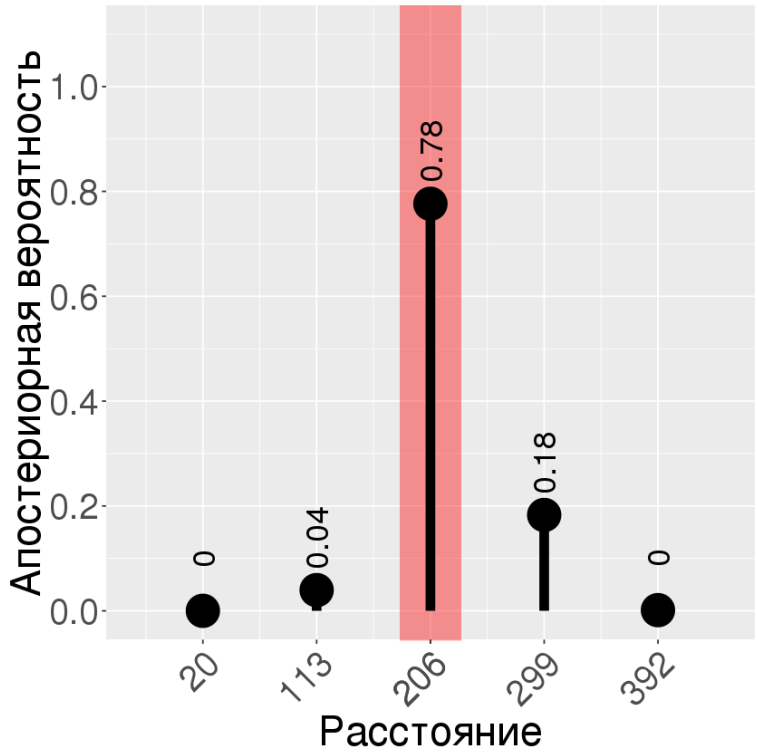


Рис. 1: Распределение длины вставки η

Пример: рёбра без повторов



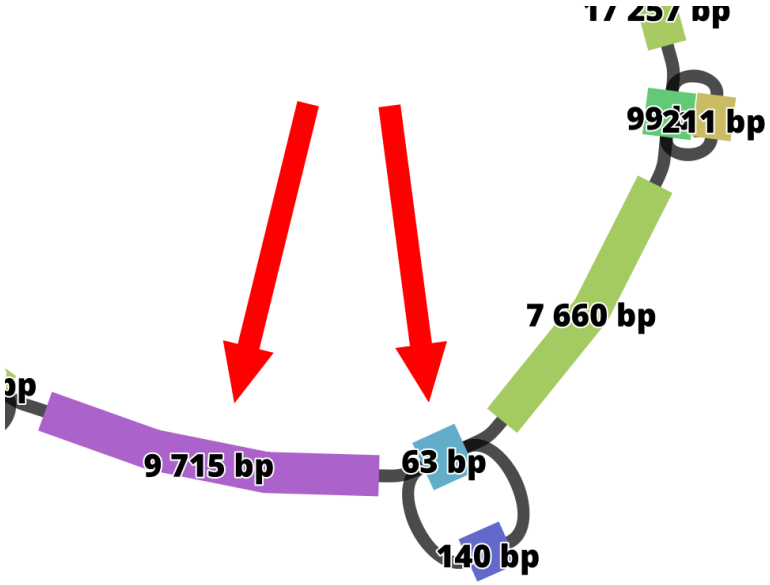
(a) Фрагмент графа



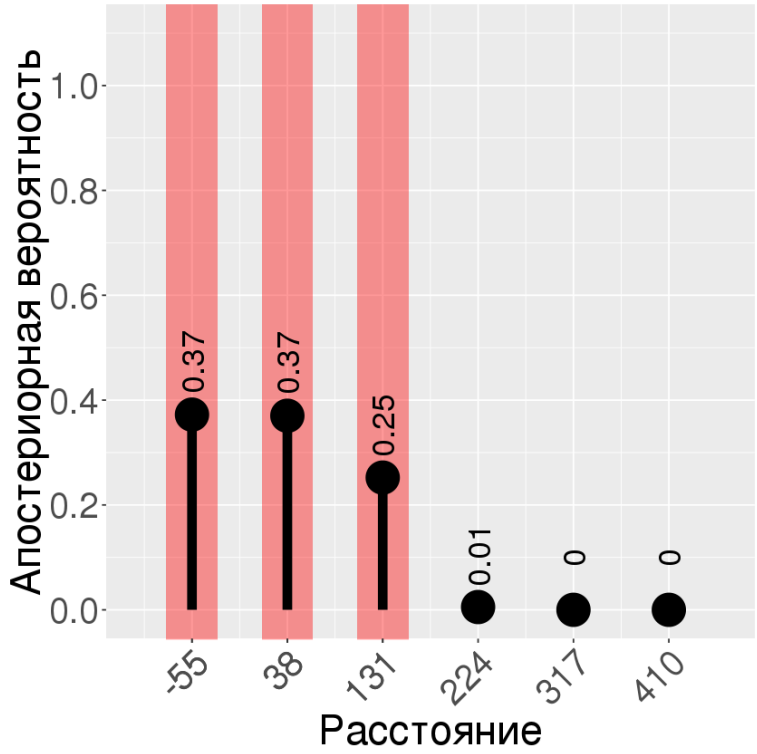
(b) Апостериорное распределение расстояния g

Рис. 2: Два длинных ребра, не имеющих повторов

Пример: повторное ребро



(a) Фрагмент графа



(b) Апостериорное распределение расстояния g

Рис. 3: Длинное ребро без повторов и короткое ребро (63 bp), имеющее повтор тройной кратности

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 как на синтетических, так и на реальных геномных данных.

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