

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

Основные термины

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

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

Задача сборки генома

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

Пусть k — положительное целое число.

Сжатый граф де Брёйна строки S — направленный мультиграф следующей конструкции:

1. Множество вершин графа — множество всех k -меров строки S .
2. Для каждого $(k + 1)$ -мера, содержащегося в S , в граф добавляется ребро $v_1 \rightarrow v_2$, 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.

Зафиксируем пару e_1, e_2 рёбер графа де Брёйна. Будем предполагать, что

1. $e_1 = \mathcal{S}[a, b]$ и $e_2 = \mathcal{S}[c, d]$, где $a < c$;
2. e_1 и e_2 соединяет путь $p = e_1 \rightarrow p_1 \rightarrow \dots \rightarrow p_m \rightarrow e_2$.

Графовое расстояние: $d_{\text{graph}}(e_1, e_2; p) = \sum_{i=1}^m |p_i| - (m+1)k$,

Геномное расстояние: $d_{\text{genome}}(e_1, e_2) = c - b$.

Определим множества

$$\mathbf{D}_{\text{graph}} = \{d_{\text{graph}}(e_1, e_2; p) \mid p \text{ — путь, соединяющий } e_1 \text{ с } e_2\},$$

$$\mathbf{D}_{\text{genome}} = \{d_{\text{genome}}(e_1^{(i)}, e_2^{(j)}) \mid e_s^{(t)} \text{ — } t\text{-ое вхождение } e_s \text{ в геном } \mathcal{S}\},$$

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 $\{1, \dots, |\mathcal{S}|\}$, 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 $\mathcal{S}[\xi + \eta - \ell, \xi + \eta]$.

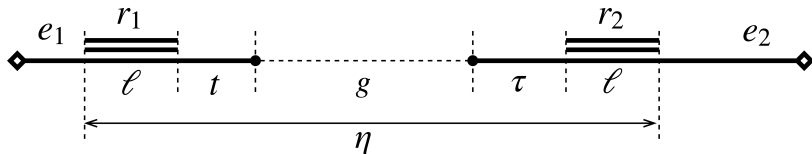


Рис. 1: Расположение ридов на рёбрах графа

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

Введём обозначения:

1. g — геномное расстояние между e_1 и e_2 ,
2. t — расстояние от конца r_1 до конца e_1 ,
3. τ — координата начала r_2 на e_2 .

Рассмотрим формально выборку $\left((t_1, \tau_1, g_1), \dots, (t_n, \tau_n, g_n)\right)$.

1. Реализации (t, τ) наблюдаются только при условии $A_{e_2}(r_2) = \{\text{рид } r_2 \text{ приложен к } e_2\}$ (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 $\mathbb{T} = \left((t_1, \tau_1), \dots, (t_n, \tau_n)\right)$.

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

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

Предложение

Пусть длина вставки η имеет распределение \mathcal{P}_η 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

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

- На практике для каждого ряда $(r_1, r_2) \in \mathfrak{R}$ 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**:

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

Здесь π_i мы можем оценить, усредняя апостериорную вероятность $p(g^{(i)} \mid t, \tau, A_{e_2})$ по всем имеющимся реализациям.

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

Предложение

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

Тогда

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

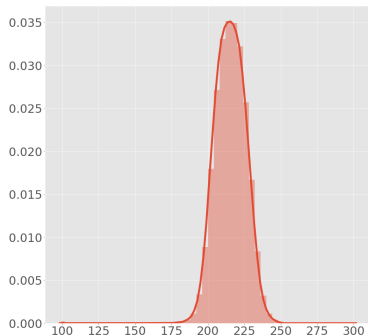
где

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

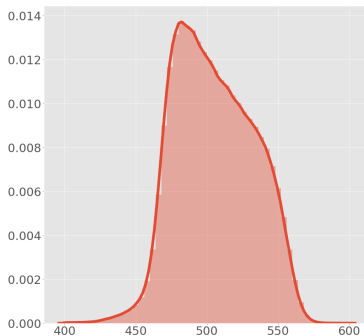
Во всех следующих примерах используются графы де Брёйна, построенные по различным библиотекам ридов для первых 400 тысяч нуклеотидов генома *E.coli* (штамм *K12 MG1655*).

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. ($\text{med } \eta \approx 480$).

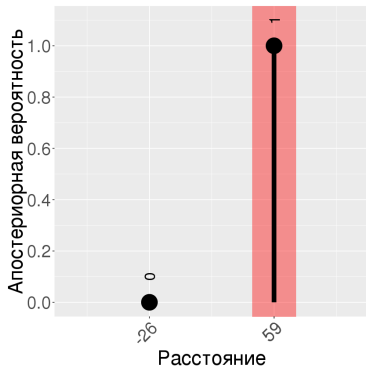


(a) Библиотека А. Хорошо аппроксимируется нормальным с параметрами $\mu = 215$, $\sigma = 10$

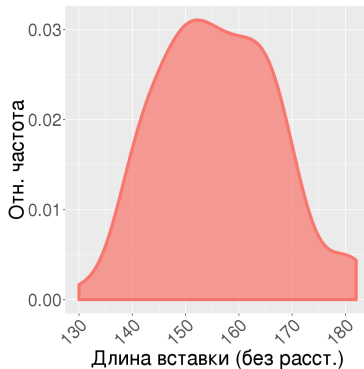


(b) Библиотека Б. Используем эмпирическую ф. р., $\text{mode } \eta = 480$

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

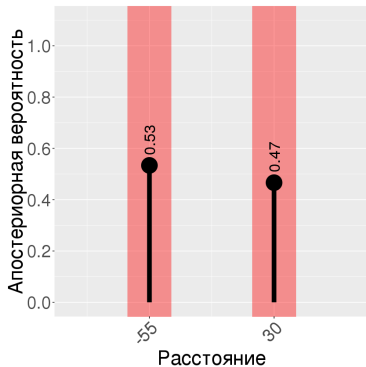


(a) Апостериорное распределение

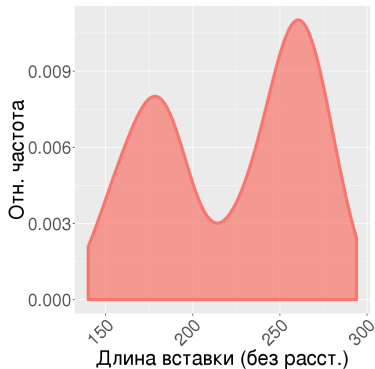


(b) Гистограмма $\eta - g$

Рис. 3: Неповторные рёбра

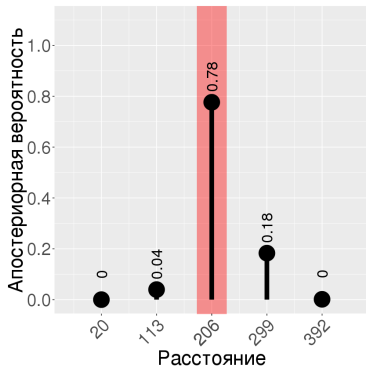


(a) Апостериорное распределение



(b) Гистограмма $\eta - g$

Рис. 4: Одно из рёбер имеет двойную кратность

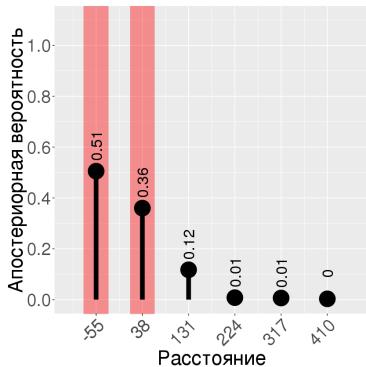


(a) Апостериорное распределение

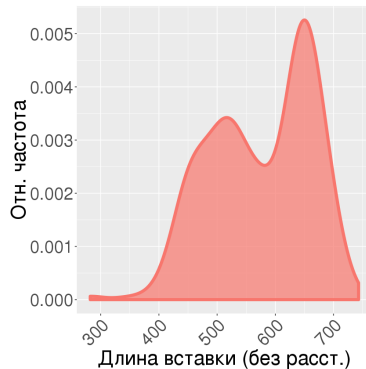


(b) Гистограмма $\eta - g$

Рис. 5: Неповторные рёбра

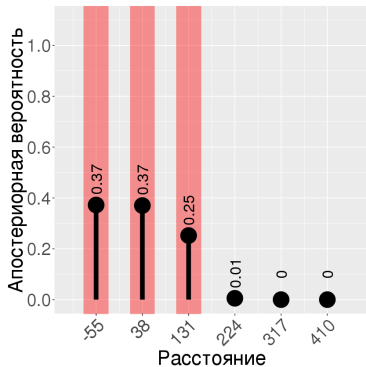


(a) Апостериорное распределение

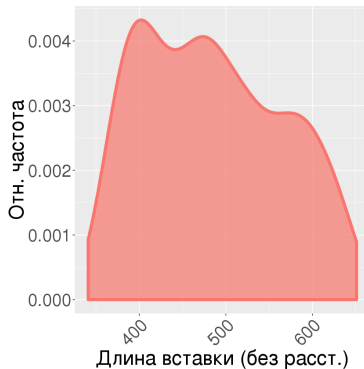


(b) Гистограмма $\eta - g$

Рис. 6: Одно из рёбер имеет двойную кратность



(a) Апостериорное распределение



(b) Гистограмма $\eta - g$

Рис. 7: Одно из рёбер имеет тройную кратность

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, имеющих 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.