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: Шлемов А. Ю.

Санкт-Петербург
2019

Основные понятия, связанные с геномом

- **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.
- **Спектр** k-**меров** — множество всех k-меров, встречающихся в $\mathcal{S}$.

Задача сборки генома

Рассмотрим некоторый геном $\mathcal{S}$ и предположим, что имеется набор его ридов. Обозначим его через $\mathfrak{R}$ и будем называть **библиотекой ридов** для $\mathcal{S}$.

Задача сборки генома:

По набору строк $\mathfrak{R}$ восстановить как можно более длинные **контиги** — непрерывные подстроки исходной строки $\mathcal{S}$ (в идеале, конечно, всю строку целиком).

Граф де Брёйна

Граф де Брёйна строки $\mathcal{S}$:

1. В качестве вершин графа берётся спектр 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.
3. Количество таких рёбер равно количеству вхождений соответствующего $(k+1)$-мера в геном.

Замечание: На практике вместо кратных рёбер обычно используют взвешенные, а однозначно продолжимые рёбра склеивают вместе.

Свойства графа де Брёйна

Эйлеров путь в мультиграфе — это путь, проходящий по каждому ребру мультиграфа ровно столько раз, какова его кратность.

Пусть G — граф де Брёйна строки $\mathcal{S}$. Тогда

1. В этом графе **существует** соответствующий исходной строке $\mathcal{S}$ эйлеров путь. Будем называть этот путь **геномным**.
2. Если в графе всего один эйлеров путь, то мы получаем возможность однозначно восстановить исходную строку.

Сборка при помощи графа де Брёйна

Итак, есть библиотека ридов $\mathfrak{R}$.

Проблема: для построения графа де Брёйна требуется знать **все** $k+1$-меры **неизвестной** строки $\mathcal{S}$.

Решение: необходимо наложить на библиотеку ридов $\mathfrak{R}$ дополнительные условия.

Предположим, что риды из $\mathfrak{R}$ содержат все $(k+1)$-меры, имеющиеся в $\mathcal{S}$ (т. н. *модель плотных ридов*).

Тогда можно извлечь из $\mathfrak{R}$ спектр её $(k+1)$-меров и построить граф де Брёйна, используя их.

Проблемы подхода

Плохое качество сборки может быть следствием

1. Ошибок в ридах (неточных прочтений),
2. Нарушения предположения о плотности покрытия генома ридами,
3. Особенностей структуры генома — повторы последовательностей (имеющие длину больше k) в $\mathcal{S}$ приводят к неединственности эйлерова пути.

Повторы

Предположим, что геном имеет вид $\mathbf{e_1 f e_2 \ldots g_1 f g_2}$, где $\mathbf{e}_i$, $\mathbf{g}_i$ и $\mathbf{f}$ — некоторые строки.

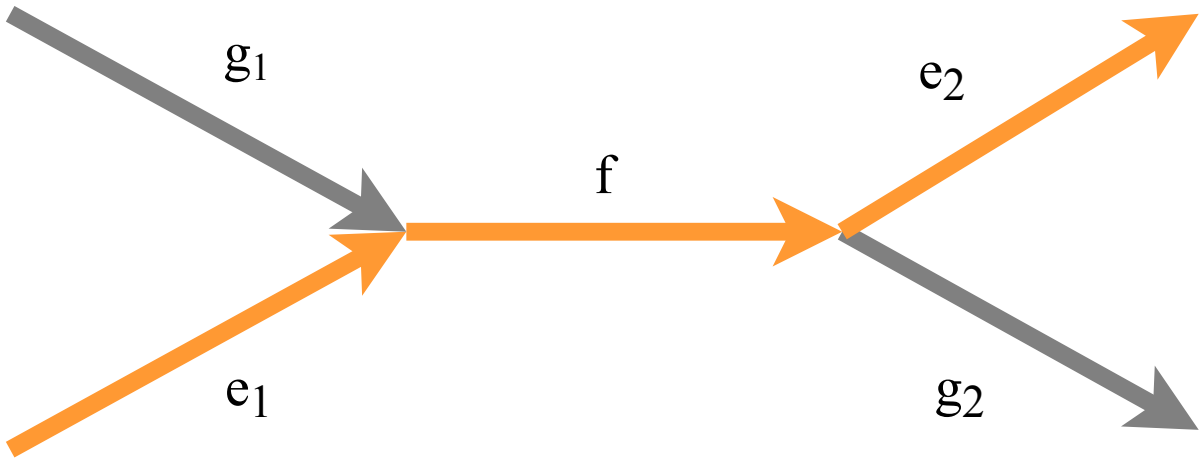


Рис. 1: Простой повтор в графе

Как должен проходить эйлеров путь,

- $e_1 \to f \to e_2$ или $e_1 \to f \to g_2$?
- $g_1 \to f \to g_2$ или $g_1 \to f \to e_2$?

Повторы

- Следовательно, повторы жизненно необходимо каким-то образом разрешать.
- Для разрешения повторов в графе сборки обычно используются специальные структуры, несущие дополнительную информацию о связи между последовательностями на рёбрах графа.
- Одной из таких структур являются так называемые **парные риды**.

Вероятностная модель парных ридов

Пусть

- ξ – дискретная случайная величина с носителем $\{1, \ldots, |\mathcal{S}|\}$, имеющая смысл координаты в геноме,
- η — независимая от ξ неотрицательная целочисленная случайная величина (т. н. **длина вставки**),
- 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 $\mathcal{S}[\xi + \eta - \ell, \xi + \eta]$.

Разрешение повторов

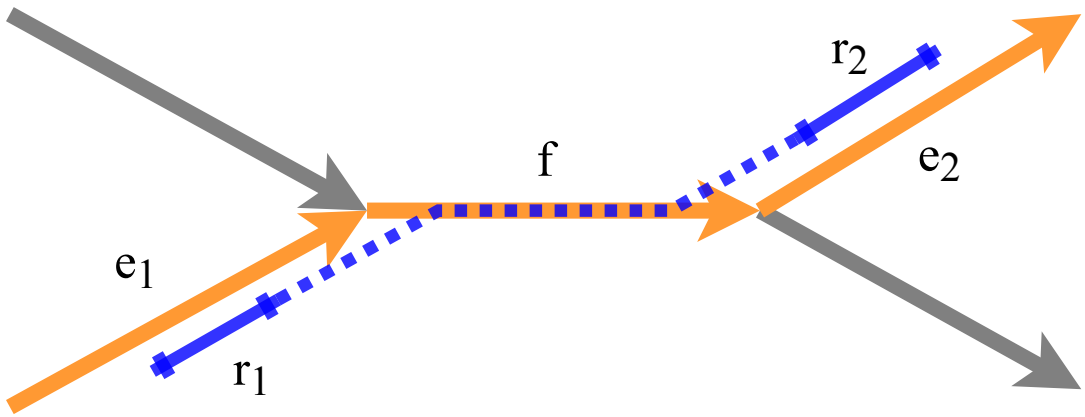


Рис. 2: Простой повтор

Графовое расстояние между r_1 и r_2 вдоль $\mathbf{p} = (e_1, f, e_2)$:

$$d_{graph}(r_1, r_2) = d(e_1, e_2) - r_1^{(s)} + r_2^{(s)},$$

где $d(e_1, e_2) = |\mathbf{p}| - |e_2|$ — расстояние между e_1 и e_2 вдоль $\mathbf{p}$, $r_i^{(s)}$ — координата начала r_i на e_i.

Разрешение повторов

Предположим, что длины ридов и длина вставки известны точно.

В этом случае известно **геномное расстояние** между r_1 и r_2 (то есть расстояние между ними как подстроками генома):

$$d_{genome}(r_1, r_2) = L - |r_2|,$$

где L — точное значение длины вставки.

Тогда если

$$d_{graph}(r_1, r_2) \neq d_{genome}(r_1, r_2),$$

то можно утверждать, что путь $\mathbf{p} = (e_1, f, e_2)$ не является частью геномного пути.

Постановка задачи

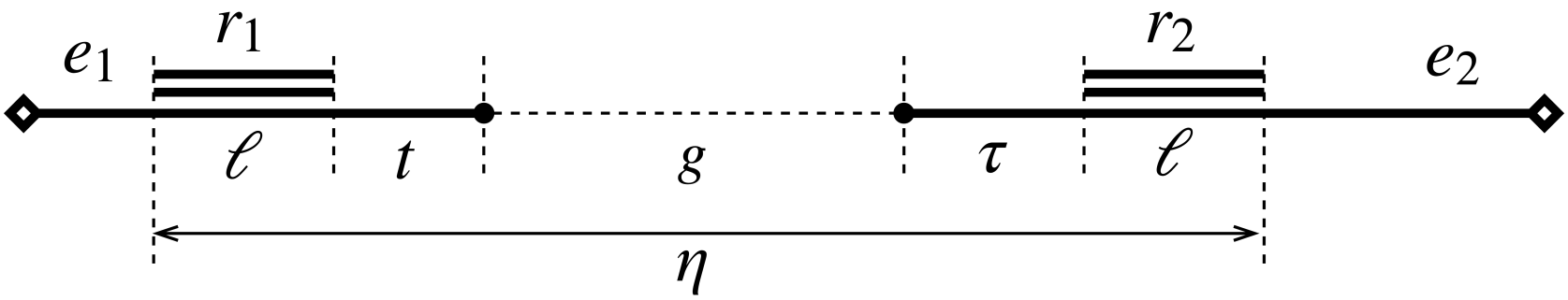


Рис. 3: Расположение ридов на рёбрах графа

Зафиксируем пару рёбер e_1, e_2. Пусть известны координаты ридов r_i на рёбрах e_i. Введём обозначения:

1. g — гэп 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.

Постановка задачи

r_i — случайные подстроки генома в терминах вероятностной модели парных ридов. Будем считать, что рид r_1 приложен к ребру e_1.

Введём событие

$$A_{e_2}(r_2) = \{\text{рид } r_2 \text{ приложен к } e_2\}.$$

Если $A_{e_2}(r_2)$ произошло, то мы получаем наблюдение (t, τ).

Рассмотрим $\mathbb{T} = \Big((t_1, \tau_1), \ldots, (t_n, \tau_n)\Big)$ — повторную независимую выборку из совместного распределения t и τ при условии $A_{e_2}(r_2)$. Пусть гэп g имеет неизвестное распределение $\mathcal{P}_g$, а распределение η известно в точности.

Задача: статистический вывод для $\mathcal{P}_g$ по выборке $\mathbb{T}$.

Замечание: кроме выборки $\mathbb{T}$ мы также располагаем и некоторой *априорной* информацией — набором

$$\mathbb{D} = \{g^{(1)}, \ldots, g^{(k)}\}$$

расстояний между рёбрами e_1 и e_2 в графе. Логично попытаться ей воспользоваться. Сделать это позволяет *байесовский вывод*.

Введём *априорное распределение* g:

$$g \sim \begin{pmatrix} g^{(1)} & \ldots & g^{(k)} \\ 1/k & \ldots & 1/k \end{pmatrix},$$

и найдём $p(g|\mathbb{T})$.

Апостериорное распределение длины гэпа

Утверждение

Пусть $\eta = \lfloor \tilde{\eta} \rfloor$, где $\tilde{\eta}$ имеет распределение $\mathrm{N}(\mu, \sigma^2)$ с известными средним μ и дисперсией σ^2.

Тогда

$$p(g|\mathbb{T}, A_{e_2}) = \frac{\prod_{i=1}^{n} q_i(\tau_i, g_i^*, t_i)}{\sum_{j=1}^{k} \prod_{i=1}^{n} q_i(\tau_i, g^{(j)}, t_i)},$$

где

$$q_i(x, y, z) = \frac{\Phi\big(x + y + z + 2\ell + 1\big) - \Phi\big(x + y + z + 2\ell\big)}{1 - \Phi\big(y + z + 2\ell\big)},$$

а Φ — функция распределения закона $\mathrm{N}(\mu, \sigma^2)$.

Моделирование (прямое)

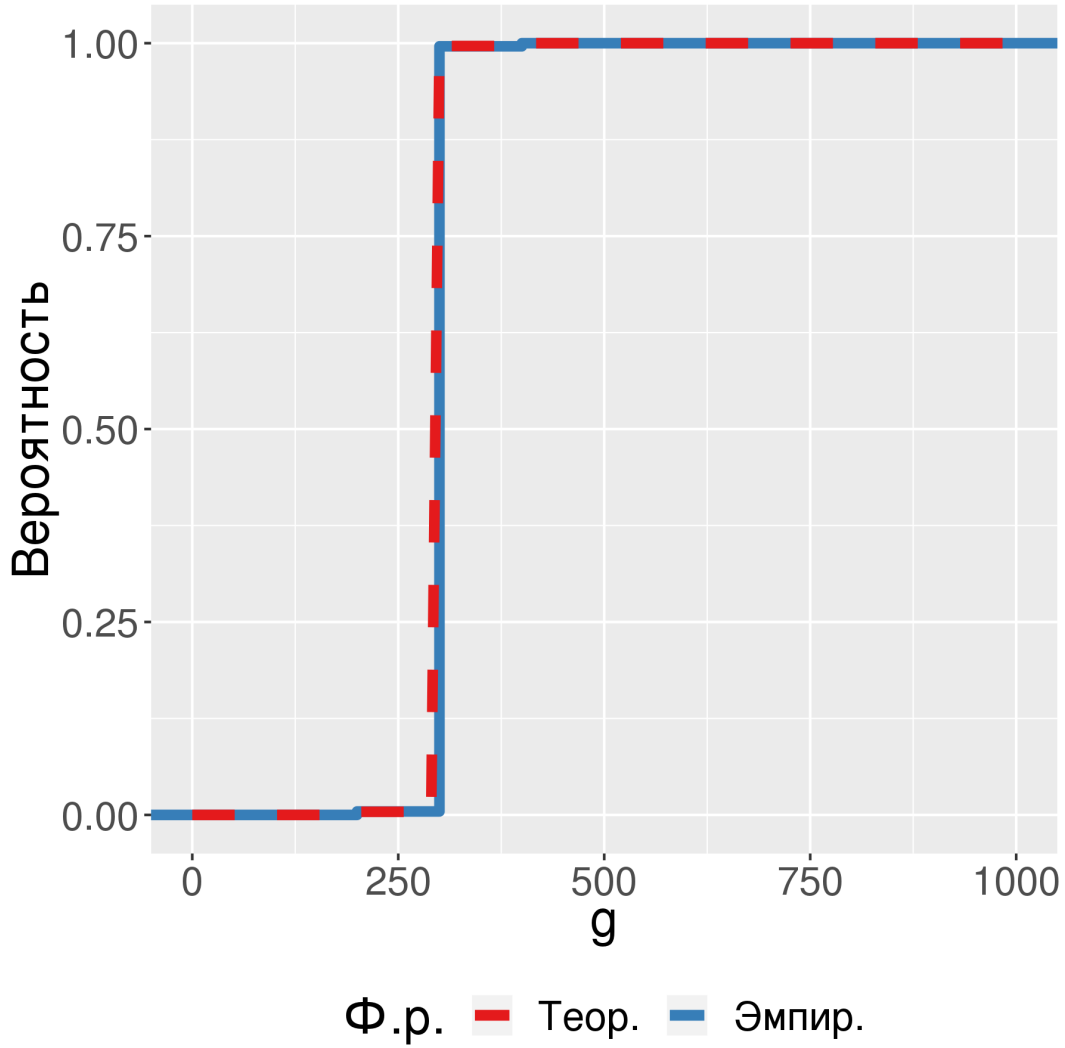
Пусть $\mu = 1000$, $\sigma = 30$, $\ell = 100$. В качестве априорного распределения g выберем равномерное дискретное со значениями из $\{100, 200, \ldots, 800\}$.

1. Промоделируем выборку с g из априорного распределения,
2. Промоделируем выборку с g из распределения, отличного от априорного.

В обоих случаях сравним получаемое условное распределение с вычисленным теоретически.

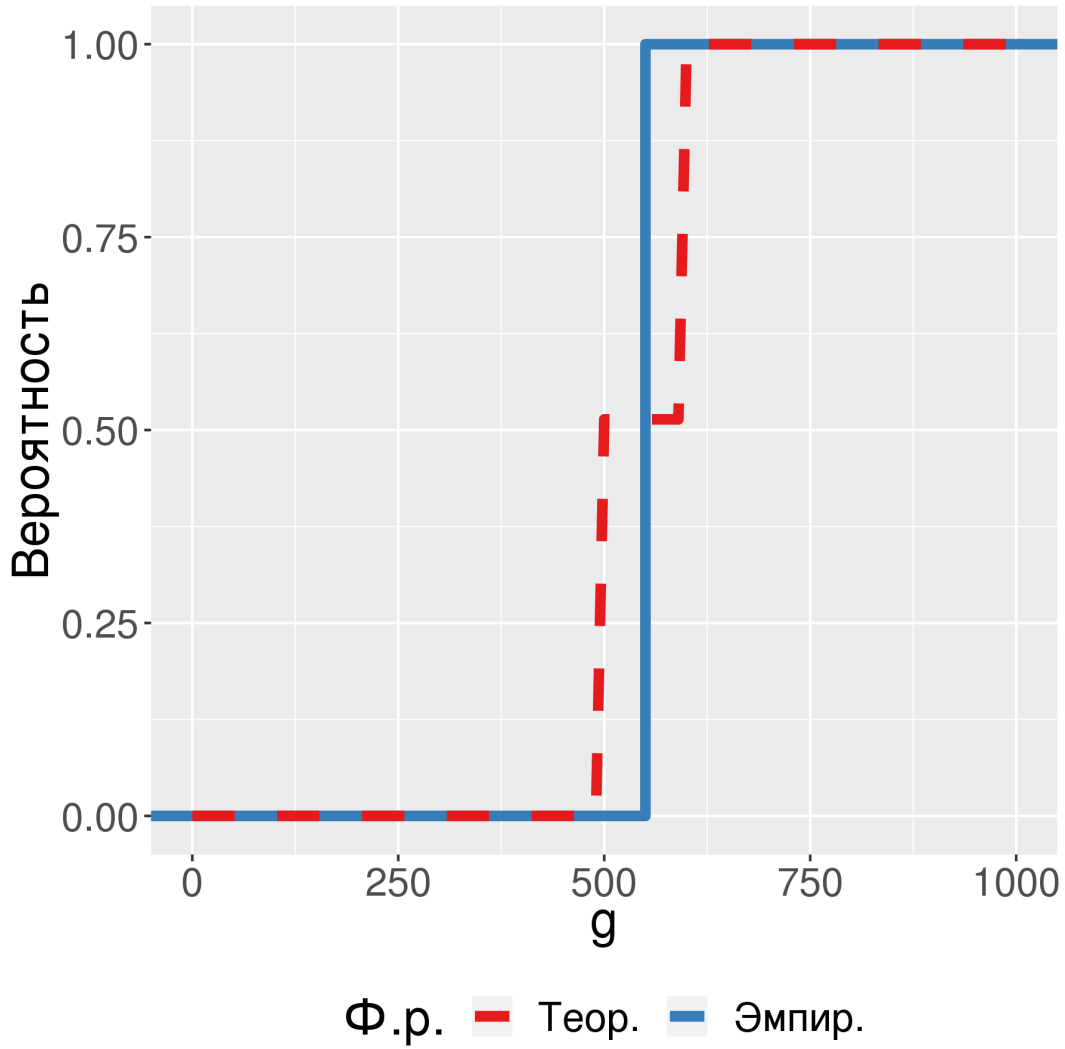
Моделирование (прямое)

Моделируем g из априорного распределения.



Моделирование (прямое)

Теперь пусть g всегда равно 550, а априорное распределение по-прежнему сосредоточено на $\{100, \ldots, 800\}$.



Дальнейшие планы

Проверить полноценно на реальном геноме — с построением графа, выравниванием ридов на рёбер:

1. На симулированных ридах,
2. На реальных ридах.