

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

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

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

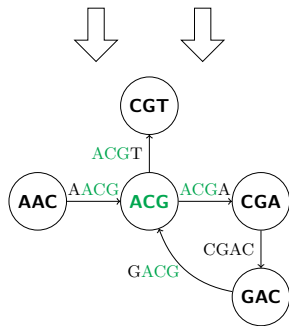
k -мер строки S — это её подстрока длины k .

Граф де Брёйна G , $k \in \mathbb{N}$:

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

Неформально говоря, граф де Брёйна состоит из всех уникальных подстрок длины k генома S , которые соединены в том порядке, в котором они встречаются в S .

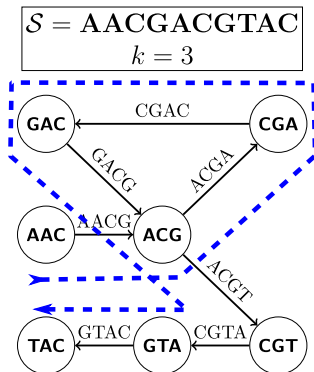
$S = \text{AACGACGT}$
AAC
ACG
CGA
GAC
ACG
CGT



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*, так как \mathcal{S} неизвестна.
- Если эйлеровых путей в полученном графе несколько, то неизвестно, какой из них соответствует \mathcal{S} .



Сборка генома \iff поиск \mathbf{p}^* среди всех эйлеровых путей в графе де Брёйна G , построенном по ридам.

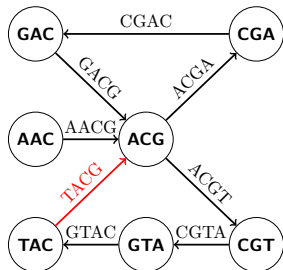
Проблема: повторы последовательностей (длины $\geq k$) приводят к образованию циклов.

Добавим к строке 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?

$S = \text{AACGACGTAC}\textbf{G}$
 $k = 3$



По графу ответить на этот вопрос невозможно!

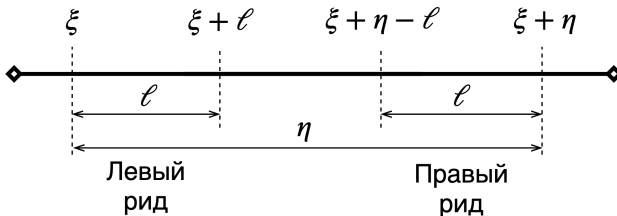
Для разрешения повторов требуется внешняя информация.

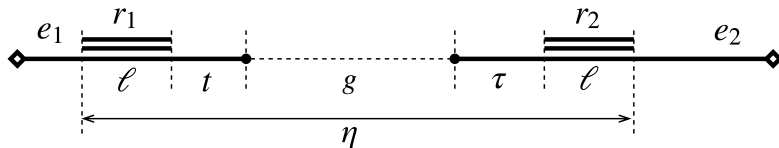
Рассмотрим независимые случайные величины

- $\xi \sim U(\{1, \dots, |\mathcal{S}|\})$, играющую роль координаты в геноме;
- $\eta > 0$ с произвольным распределением \mathcal{P}_η , которую мы далее будем называть **длиной вставки**.

Рассмотрим случайную подстроку генома, имеющую вид $\mathcal{S}[\xi, \xi + \eta]$.

Парный ряд — это пара из её префикса $\mathcal{S}[\xi, \xi + \ell]$ (*левый ряд*) и суффикса $\mathcal{S}[\xi + \eta - \ell, \xi + \eta]$ (*правый ряд*) фрагмента.





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 .

Тогда

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

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

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

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

Приходим к **модели смеси**:

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

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

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 ридов с длиной вставки $\eta \sim N(\mu, \sigma^2)$:

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

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

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 рядам.

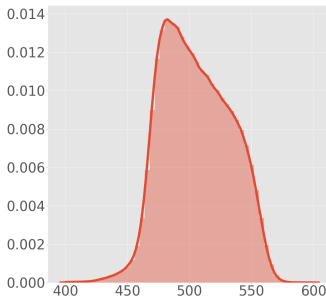
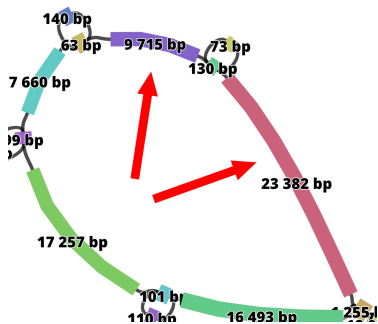
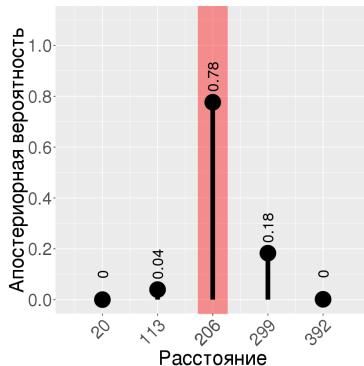


Рис. 1: Распределение длины вставки η

Пример: рёбра без повторов



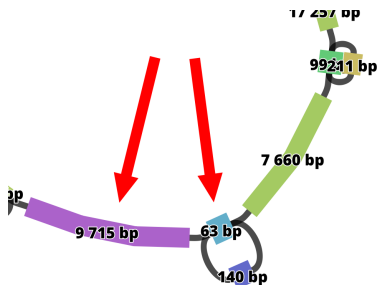
(a) Фрагмент графа



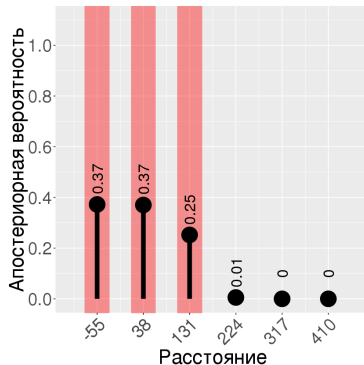
(b) Апостериорное распределение расстояния g

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

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 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.