

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

- **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 .
- **Спектр k -меров** — множество всех k -меров, встречающихся в S .

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

Задача сборки генома:

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

Граф де Брёйна строки S :

1. В качестве вершин графа берётся спектр 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.
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.

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

Пусть G — граф де Брёйна строки S . Тогда

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

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.

Плохое качество сборки может быть следствием

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

Предположим, что геном имеет вид $e_1 f e_2 \dots g_1 f g_2$, где e_i , g_i и f — некоторые строки.

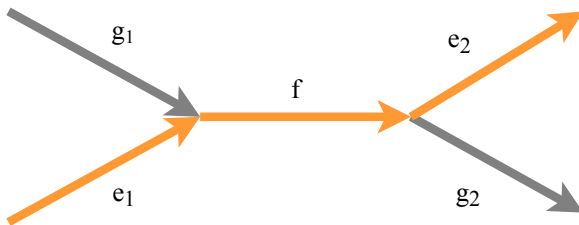


Рис. 1: Простой повтор в графе

Как должен проходить эйлеров путь,

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

- 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**.

Пусть

- ξ — дискретная случайная величина с носителем $\{1, \dots, |\mathcal{S}|\}$, 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 $\mathcal{S}[\xi + \eta - \ell, \xi + \eta]$.

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 .

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.

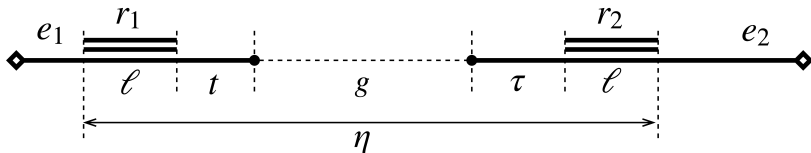


Рис. 2: Расположение ридов на рёбрах графа

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

1. g — гэд 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

$$\mathbb{T}' = \left((t_1, \tau_1, g_1), \dots, (t_n, \tau_n, g_n) \right)$$

из совместного распределения t , τ и g .

Будем считать, что рид r_1 приложен к ребру e_1 . Введём событие

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

На самом деле, мы наблюдаем реализации только при условии A_{e_2} , а g_i не наблюдаем вовсе.

Получаем набор реализаций

$$\mathbb{T} = \left((t_1, \tau_1), \dots, (t_n, \tau_n) \right).$$

Пусть

$$\mathbf{D}_{graph} = \{g^{(1)}, \dots, g^{(k)}\}$$

— набор гэпов между рёбрами e_1 и e_2 в графе сборки, а

\mathbf{D}_{genome} — набор гэпов между ними в геноме.

Положим

$$\mathbf{D} = \mathbf{D}_{genome} \cap \mathbf{D}_{graph}.$$

Совместное распределение вектора (t_i, τ_i) зависит от g_i как от параметра. При этом t_i , τ_i и g_i связаны соотношением

$$\tau_i = \eta_i - t_i - g_i - 2\ell,$$

где $g_i \in \mathbf{D}$ — один из графовых гэпов, который одновременно является и геномным.

ЗАДАЧА: определить, какие из $g^{(i)} \in \mathbf{D}_{graph}$ являются геномными, при помощи выборки \mathbb{T} .

Здесь и далее в формулах для упрощения будем опускать условие A_{e_2} .

Предложение

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

Тогда

$$p(g|\tau, t) = \frac{q(\tau, g, t)}{\sum_{j=1}^k q(\tau, g^{(j)}, t)},$$

где

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

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

- Пусть имеется библиотека парных ридов \mathcal{R} .
- На практике для каждого рида $(r_1, r_2) \in \mathcal{R}$ реализуется собственный гэп $g^{(i)} \in \mathbf{D}_{genome}$ для некоторого i .
- Поэтому нельзя напрямую использовать статистический вывод по повторной независимой выборке из распределения (τ, t) .

- Формально $(\tau, t, g) \sim \mathcal{P}_{\tau, t, g}$, где на g накладывается априорное распределение:

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

- На практике g — скрытая переменная, распределение которой мы хотим оценить.

Предложение

В тех же условиях

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

где

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

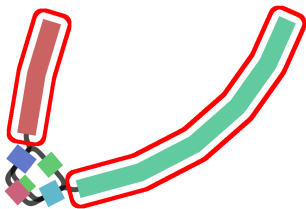
а Φ — функция распределения закона $N(\mu, \sigma^2)$.

За основу были взяты первые 400 тысяч нуклеотидов генома *E.coli*. При помощи пакета **art** были промоделированы парные риды с длиной вставки, имеющей распределение $N(1000, 30)$.

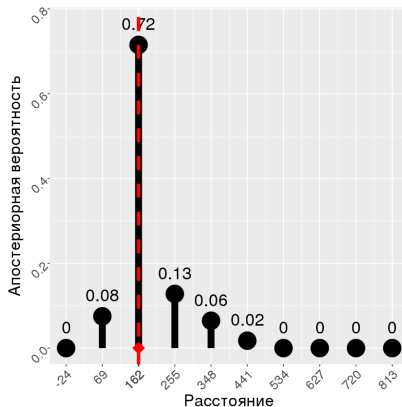
По получившимся ридам при помощи геномного ассемблера **SPAdes** был построен граф де Брёйна.

Для выравнивания рёбер получившегося графа на исходный геном и выравнивания ридов на рёбра использовался пакет **bwa**.

Пример апостериорных вероятностей



(a) Граф



(b) Распределение