

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.



Санкт-Петербург
2018г.

Основные понятия

- **Геном** — строка над четырёхбуквенным алфавитом $\{A, T, G, C\}$.
- **Рид (или прочтение)** — короткая подстрока генома, получающаяся в результате секвенирования.
- **k -мер** — подстрока генома, имеющая длину k .
- **Спектр k -меров** — множество всех k -меров, встречающихся в геноме.

Геном при секвенировании покрывается большим числом перекрывающихся ридов.

Задача: собрать из них как можно более длинные *контиги* — непрерывные подстроки исходного генома.

Конструкция графа де Брёйна:

1. В качестве вершин графа берётся спектр k -меров.
2. Для каждого $(k + 1)$ -мера из спектра $(k + 1)$ -меров добавляется ребро $v_1 \rightarrow v_2$, где v_1 и v_2 — его левый и правый k -меры 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.

Сборка при помощи графа де Брёйна

Построим граф де Брёйна, используя спектр k -меры и $(k + 1)$ -меры ридов.

Предположение: риды содержат все $(k + 1)$ -меры, имеющиеся в геноме (т. н. *dense read model*).

Тогда

1. В этом графе существует соответствующий исходному геному эйлеров путь — путь, проходящий по всем рёбрам ровно один раз. Будем называть этот путь **геномным**.
2. Если в графе существует единственный эйлеров путь, то получаем собранный геном.

Проблема: повторы последовательностей в геноме приводят к неединственности эйлерова пути — в графе появляются пути, не имеющие отношения к истинному геному.

- **Фрагмент** — подстрока генома, имеющая вид $S[\xi, \xi + \eta]$, где ξ — случайная координата начала фрагмента, η — случайная длина фрагмента (т. н. **длина вставки**).
- **Парный рид** — пара (r_1, r_2) , где r_1 — случайный префикс фрагмента (*forward-рид*), r_2 — случайный суффикс фрагмента (*reverse-рид*).

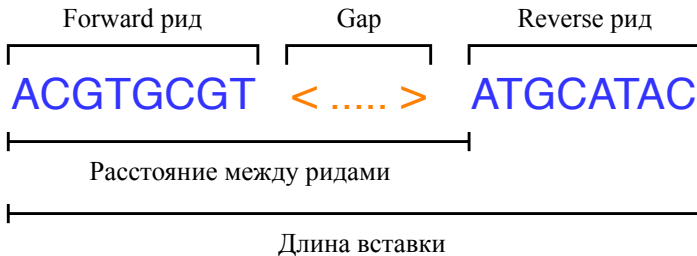


Рис. 1: Структура парного рида

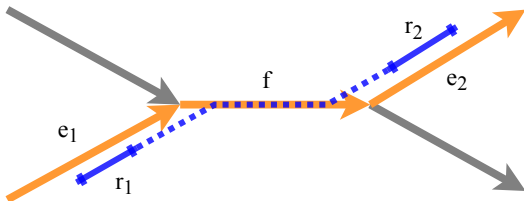


Рис. 2: Простой повтор

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

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

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 при выравнивании на e_i .

Предположим, что длины ридов и длина вставки известны точно, выравнивание ридов является точным.

В этом случае известно геномное расстояние между r_1 и r_2 :

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

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.

Выборка расстояний между рёбрами

- Предположим, имеется парный рид (r_1, r_2) и выравнивание r_1 и r_2 на рёбра графа e_1 и e_2 соответственно.
- Расстояние между рёбрами:

$$d(r_1, r_2) = \eta - |r_2| + r_1^{(s)} - r_2^{(s)},$$

где $r_i^{(s)}$ — координаты начала r_i при выравнивании на e_i , а η — длина вставки.

$\mathbb{X}_{e_1, e_2} = \{d(r_1, r_2) \mid r_1, r_2 \text{ выровнялись на рёбра } e_1, e_2\}$ — выборка расстояний между e_1 и e_2 .

Зафиксируем e_1 и e_2 .

- Пусть $\mathcal{P} = \mathcal{P}_{e_1, e_2}$ — распределение расстояний между e_1 и e_2 .
- Так как оба ребра могут встречаться в геноме несколько раз, то и расстояний между ними может быть несколько.

Входные данные:

1. $\mathbb{X} = \mathbb{X}_{e_1, e_2}$ — выборка расстояний между e_1 и e_2 ,
2. Графовые пути между e_1 и e_2 .

Задача: построить модель, которая по выборке \mathbb{X} позволит оценивать геномные расстояния между рёбрами e_1 и e_2 , а также отличать потенциально геномные пути между ними от негеномных.

$$\mathcal{P} = \sum_{i=1}^n \pi_i \mathcal{P}^{(i)},$$

где

1. n — количество геномных путей из e_1 в e_2 ;
2. π_i — веса, то есть $\pi_i > 0$ и $\sum_{i=1}^n \pi_i = 1$;
3. $\mathcal{P}^{(i)}$ — абсолютно непрерывное распределение, математическое ожидание которого равно одному из геномных расстояний.

Предположим, что $\mathcal{P}^{(i)} = N(d_i, \sigma_i^2)$, где d_i — длина одного из геномных путей. Тогда плотность распределения расстояния имеет вид

$$\varphi(t) = \sum_{i=1}^n \pi_i \varphi_{d_i, \sigma_i^2}(t),$$

где φ_{μ, σ^2} — плотность распределения $N(\mu, \sigma^2)$.

- Модель содержит $3n - 1$ параметр: π_j и d_i, σ_i^2 ($i \in 1 : n, j \in 1 : n - 1$).
- Параметры можно оценить по выборке \mathbb{X} .

Для оценки параметров модели воспользуемся методом максимального правдоподобия. Пусть

$$\begin{aligned}\pi &= (\pi_1, \dots, \pi_{n-1}), \quad d = (d_1, \dots, d_n), \quad v = (\sigma_1^2, \dots, \sigma_n^2), \\ \theta &= (\pi, d, v), \\ \mathbb{X} &= (X_1, \dots, X_N).\end{aligned}$$

Запишем логарифм правдоподобия:

$$\ell(\theta; \mathbb{X}) = \sum_{j=1}^N \log \left(\sum_{i=1}^n \pi_i \varphi_{d_i, \sigma_i^2}(X_j) \right).$$

Оптимизировать эту конструкцию по θ напрямую не представляется возможным.

Рассмотрим «скрытые» случайные векторы Δ_j ($j \in 1 : N$):

$$\Delta_j^{(i)} = [X_j \text{ порождено } i\text{-й компонентой смеси}].$$

Шаг E(xpectation) Считая θ известным и равным θ_0 ,
вычислим

$$\gamma_j = \mathbb{E}[\Delta_j | \theta_0, \mathbb{X}], i \in 1 : N.$$

Шаг M(aximization) Используя $\Gamma = (\gamma_1, \dots, \gamma_N)$, вычислим
оценку θ :

$$\hat{\pi}_i = \sum_{j=1}^N \gamma_i^{(j)}, \quad (\hat{d}, \hat{v}) = \arg \max_{d, v} \ell(\hat{\pi}, d, v; \mathbb{X}),$$
$$\hat{\theta} = (\hat{\pi}, \hat{d}, \hat{v}).$$

Пары E- и M-шагов повторяются до сходимости.

ТВА

1. Часть геномных расстояний может в принципе не наблюдаться — это нужно учесть в модели.
2. Геномные пути могут пропадать.
3. Распределение длины вставки на самом деле наблюдается не полностью, а с неким цензурированием, что тоже нужно учесть.