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.

De novo сборка генома

- Геном невозможно прочитать как одну большую строку.
- *Риды* (или *прочтения*) — короткие подстроки генома, получаемые в процессе секвенирования.
- Геном при секвенировании покрывается большим числом перекрывающихся ридов.

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

Граф де Брёйна

k-мер — подстрока длины k.

Конструкция графа де Брёйна:

1. Строится спектр $(k+1)$-меров генома.
2. В качестве вершин графа берутся все k-меры.
3. Для каждого $(k+1)$-мера из спектра добавляется ребро $v_1 \to v_2$, где v_1 и v_2 — его левый и правый k-меры соответственно.
4. Количество таких рёбер равно количеству вхождений соответствующего $(k+1)$-мера в спектр.

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

Сборка при помощи графа де Брёйна

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.
2. Если существует единственный такой 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.

Paired-End риды

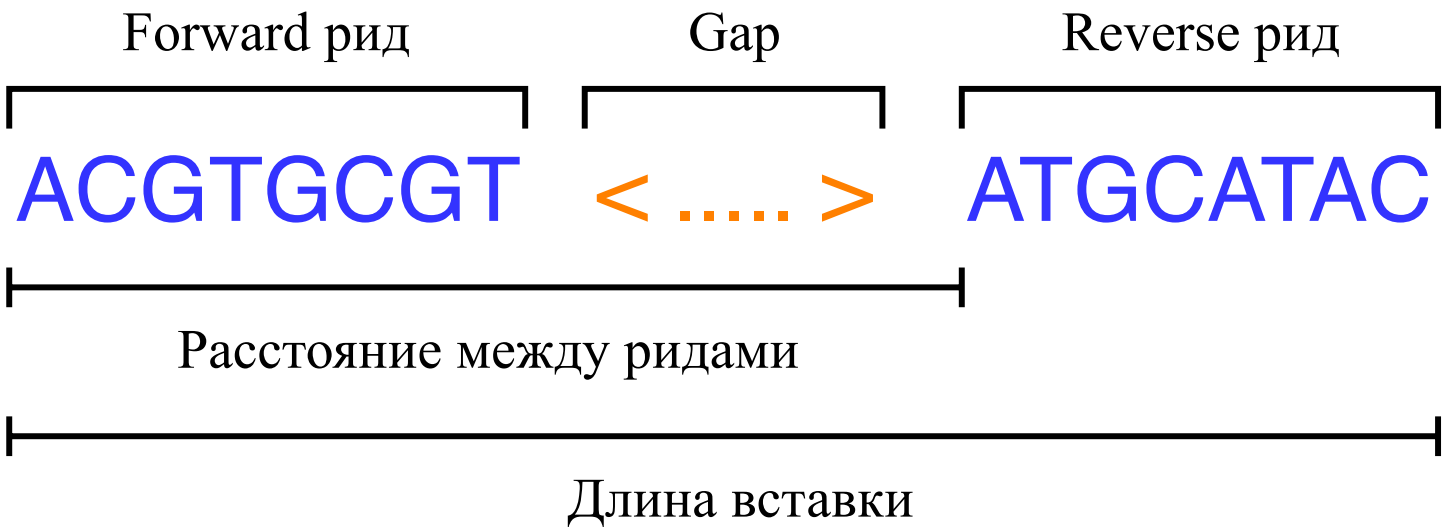

Рис. 1: Paired-End рид

- Длина вставки — случайная величина, распределение которой является характеристикой библиотеки ридов.
- Длины самих ридов также не фиксированы.

Разрешение повторов

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

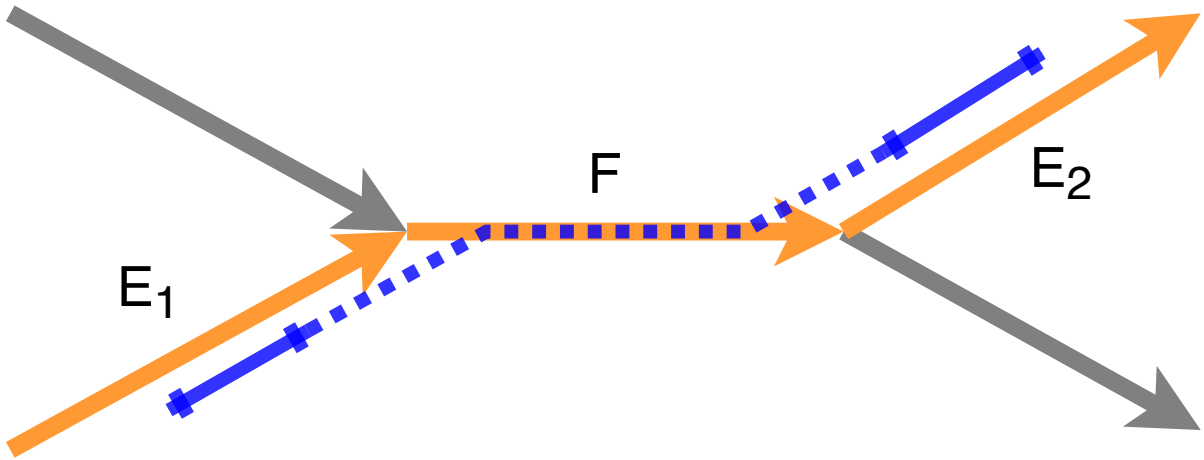

Рис. 2: Простой повтор и выровненный парный рид

Тогда если расстояние между ридами не совпадает с графовым расстоянием, то можно утверждать, что данный путь не является геномным.

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:

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

где $r_i^{(s)}$ — координаты начала r_i при выравнивании на e_i, ξ — длина вставки, $|\ .\ |$ обозначает длину строки.

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

Постановка задачи

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.

Модель смеси распределений

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

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.

Модель смеси нормальных распределений

Предположим, что $\mathcal{P}^{(i)} = \mathrm{N}(d_i, \sigma_i^2)$, 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 $\mathrm{N}(\mu, \sigma^2)$.

- 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 $\mathbb{X}$.

Оценка параметров

Для оценки параметров модели воспользуемся методом максимального правдоподобия. Пусть

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

Запишем логарифм правдоподобия:

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

EM-алгоритм для нормальной смеси

Введём в дополнение к выборке $\mathbb{X}$ «скрытые» случайные векторы $\Delta_i = \left(\Delta_i^{(1)}, \ldots, \Delta_i^{(k)}\right)$ $(i \in 1 : N)$:

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

$$\mathbb{P}[\Delta_j^{(i)} = 1] = \pi_i.$$

Плюс: Если выборка $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_N)$ наблюдается, то логарифм правдоподобия принимает более простую форму:

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

Минус: Выборка $\boldsymbol{\Delta}$ не наблюдается.

EM-алгоритм для нормальной смеси

Так как значения величины Δ_i не наблюдаются, то вместо них будем рассматривать векторы γ_i, которые определяют «уверенность» в том, что X_i было порождено j-й компонентой:

$$\gamma_i = \left(\gamma_i^{(1)}, \ldots, \gamma_i^{(k)}\right), \quad \sum_{j=1}^{k} \gamma_i^{(j)} = 1.$$

Зафиксируем начальное приближение θ_0.

EM-алгоритм для нормальной смеси

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

$$\gamma_i = \mathbb{E}\left[\Delta_i | \theta_0, \mathbb{X}\right].$$

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

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

$$\hat{\theta} = (\hat{\pi}, \hat{d}, \hat{v}).$$

Пары **E**- и **M**-шагов повторяются до сходимости. Можно доказать, что каждая итерация алгоритма не уменьшает правдоподобие.

Результаты

TBA

Дальнейшие планы

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.