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

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.

Граф де Брёйна

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 $v_1 \to v_2$, 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 k- 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 каждому ребру столько раз, какова его кратность. 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.

Парные риды

- **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*).

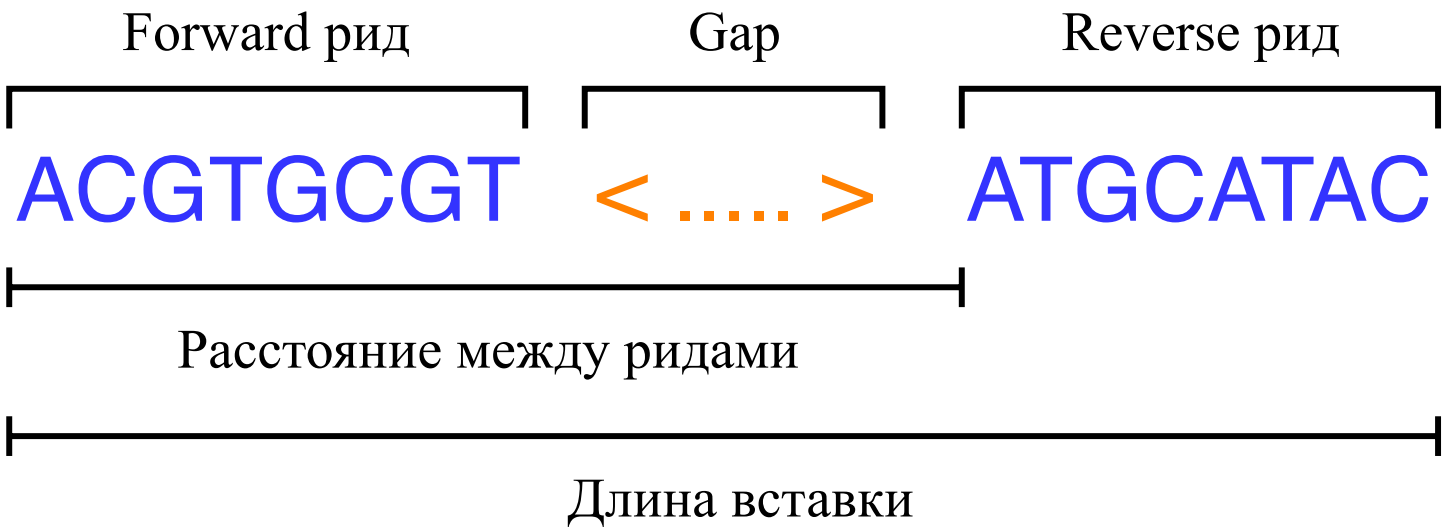

Рис. 1: Структура парного рида

Разрешение повторов

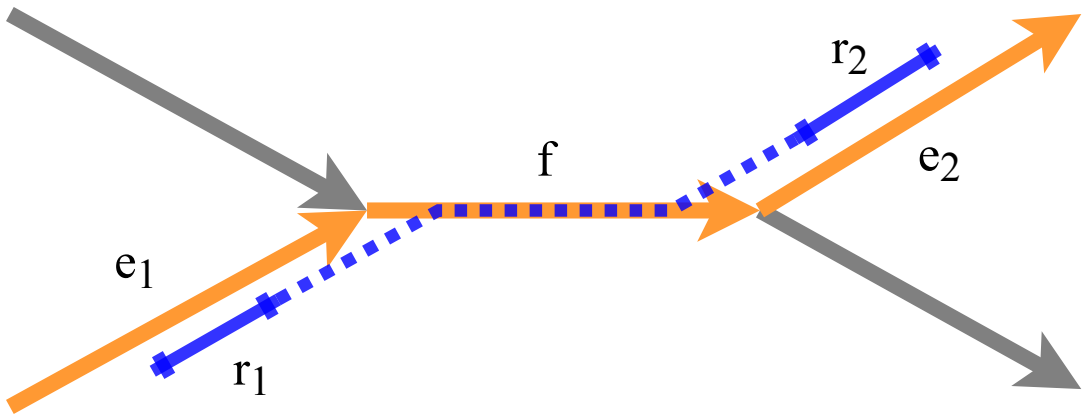

Рис. 2: Простой повтор

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:

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

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 координаты r_1 и r_2 на рёбрах 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 на e_i, а η — длина вставки.

$\mathbb{X}_{e_1,e_2} = \{d(r_1, r_2) \mid r_i$ является подстрокой $e_i\}$ — 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:

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

Оптимизировать эту конструкцию по θ напрямую не представляется возможным.

EM-алгоритм для нормальной смеси

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

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

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

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

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

Пары **E**- и **M**-шагов повторяются до сходимости.

Программная реализация

Пакет mclust реализует множество инструментов для работы со смесями нормальных распределений.

- Оценка параметров производится при помощи EM-алгоритма.
- Оптимальное число компонент смеси выбирается автоматически на основании *байесовского информационного критерия (BIC)*:

$$BIC = k \log N - 2 \log L^*,$$

где $k = 3n - 1$ — число оцениваемых параметров, n — количество компонент смеси, N — объем выборки, L^* — максимальное значение правдоподобия.

Решающее правило

Оценки расстояний между рёбрами графа получены. Теперь требуется определить правило, по которому мы сможем отличать геномные пути от негеномных.

Воспользуемся для этого классификацией на основе полученной нами модели.

Решающее правило
Будем классифицировать путь длины d как геномный, если найдётся такое i, что

$$d_i - \sigma_i \leq d \leq d_i + \sigma_i,$$

где d_i и σ_i — оценки параметров смеси, полученные при помощи EM-алгоритма.

Данные и результаты

Организм: *E. coli*, штамм *MG1655*.
Данные: Референсный геном и библиотека парных ридов с длиной вставки 298 ± 17.
Объём генома: 4.7 Mbp.

Качество классификации:
Точность (доля правильно классифицированных): 0.59,
Чувствительность (доля правильно классифицированных геномных): 0.17,
Специфичность (доля правильно классифицированных негеномных): 0.75.

Вывод: качество классификации *в целом* оставляет желать лучшего, но классификатору удаётся довольно удачно отсеивать негеномные пути.

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.