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: ???

Санкт-Петербург
2019

Геном

Геномом будем называть строку $\mathcal{S}$ над четырёхбуквенным алфавитом $\{A, T, G, C\}$.

1. Биологическая подоплёка этого понятия нас не интересует, поэтому в дальнейшем будем рассматривать его как самую обычную строку.
2. Однако, мы будем использовать некоторые стандартные для биоинформатики понятия.

Задача сборки

Рассмотрим некоторую строку $\mathcal{S}$.

Основные понятия

- **Рид** (или **прочтение**) — короткая подстрока $\mathcal{S}$.
- k-**мер** — подстрока $\mathcal{S}$, имеющая длину k.
- **Спектр** k-**меров** — множество всех k-меров, встречающихся в $\mathcal{S}$.

Предположим, имеется набор ридов строки $\mathcal{S}$ (обозначим его через $\mathfrak{R}$ и будем называть **библиотекой ридов**)

Задача: собрать из них как можно более длинные **контиги** — непрерывные подстроки исходной строки $\mathcal{S}$ (в идеале, конечно, всю строку целиком).

Граф де Брёйна строки $\mathcal{S}$

1. В качестве вершин графа берётся спектр k-меров строки $\mathcal{S}$.
2. Для каждого $(k+1)$-мера, содержащегося в $\mathcal{S}$, в граф добавляется ребро $v_1 \to v_2$, 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.

Свойства графа де Брёйна

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

Пусть G — граф де Брёйна строки $\mathcal{S}$. Тогда

1. В этом графе **существует** соответствующий исходной строке $\mathcal{S}$ эйлеров путь. Будем называть этот путь **геномным**.
2. Если в графе всего один эйлеров путь, то мы получаем возможность однозначно восстановить исходную строку.
3. Если путь не один, то можно попытаться восстановить хотя бы части исходной строки $\mathcal{S}$, но необходимо уметь выявлять отрезки геномного пути.

Сборка при помощи графа де Брёйна

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. *dense read model*).

Тогда можно извлечь из $\mathfrak{R}$ спектр её $(k+1)$-меров и построить граф де Брёйна, используя их.

Сборка при помощи графа де Брёйна

Однако,

1. Если предположение о плотности ридов нарушается, то качество сборки может оказаться неудовлетворительным. Этот случай мы не рассматриваем, считая, что предположение выполнено.
2. На качество сборки также существенно влияет и структура исходной строки $\mathcal{S}$ — повторы последовательностей (имеющие длину больше k) в $\mathcal{S}$ приводят к неединственности эйлерова пути.

Повторы

Простой повтор последовательности f: $\mathbf{e_1 f e_2 \ldots g_1 f g_2}$.

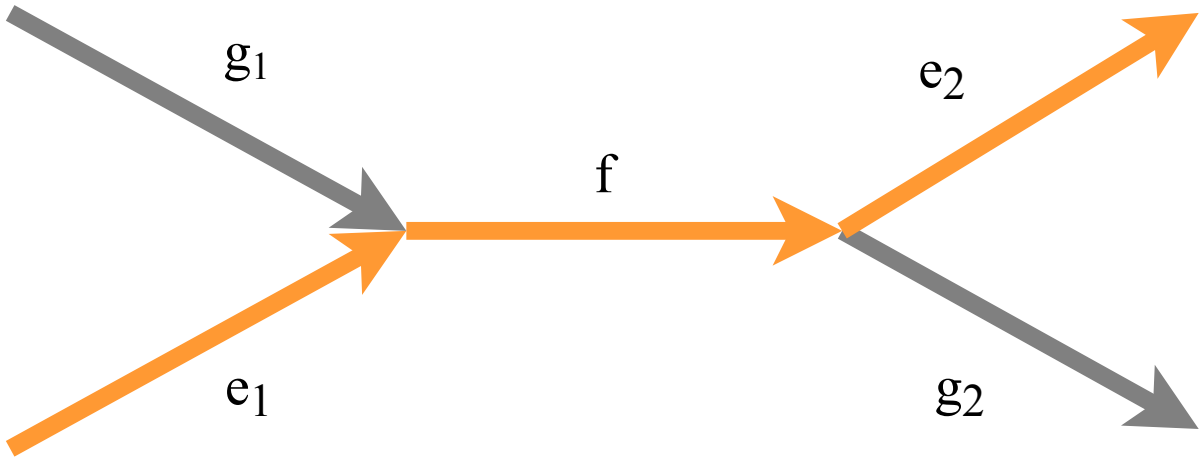

Рис. 1: Простой повтор в графе

Имея такую топологию графа, невозможно без дополнительной информации определить, как должен проходить эйлеров путь:

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

Повторы

- 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**.

Парные риды

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

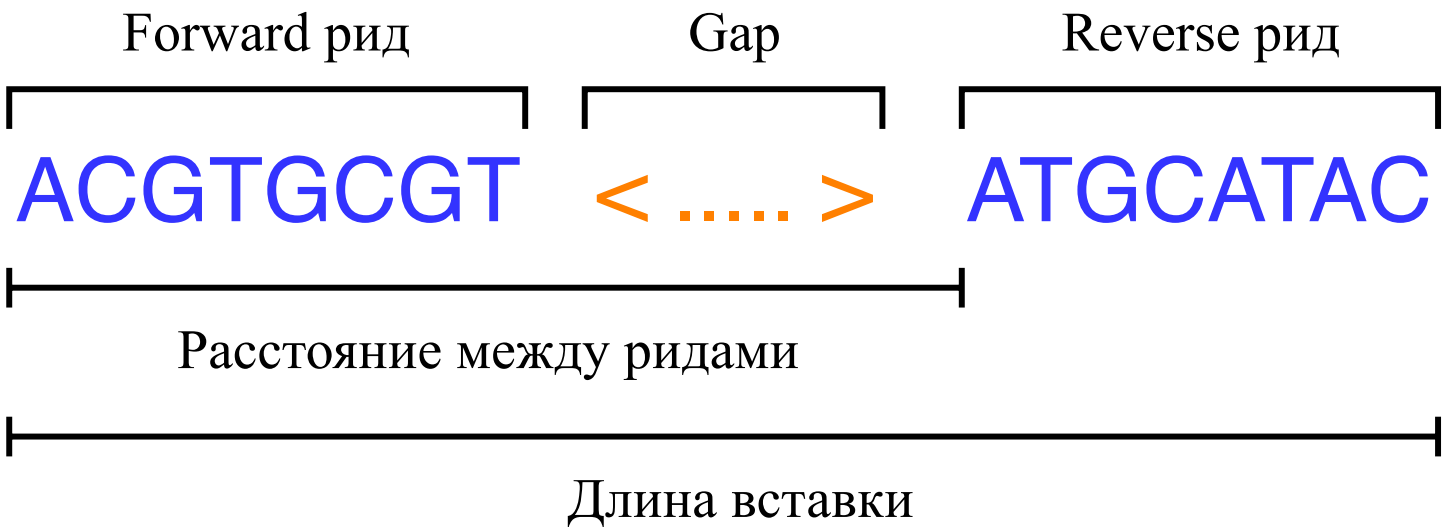

Рис. 2: Структура парного рида

Разрешение повторов

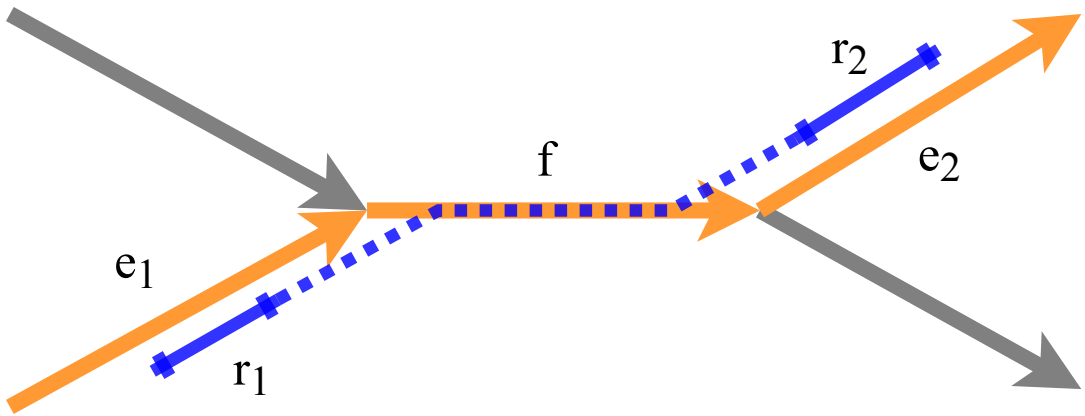

Рис. 3: Простой повтор

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.

Существующее решение

В геномном сборщике **SPAdes** реализована следующая процедура оценки расстояний:

1. Строится упорядоченный по возрастанию набор графовых расстояний между e_1 и e_2.
2. Выбрасываются все графовые расстояния, которые отстоят от границ выборки $\mathbb{X}$ более чем на $\alpha\sigma_\xi$, где α — настраиваемый коэффициент.
3. Для каждого элемента выборки находится ближайшее графовое расстояние, и к его весу прибавляется 1 (если их два, то добавляется по $1/2$ каждому).
4. Далее над получившимся взвешенным набором расстояний производится иерархическая кластеризация.
5. Оценка расстояний — центроиды кластеров.

Существующее решение

- Иногда происходят ошибки сборки — на гистограмме наблюдаемых расстояний наблюдается «пик», но соответствующего ему пути в графе нет. Информация об этом полностью теряется.
- Хотелось бы получить модель, которая бы позволила избежать потери этой информации.

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. Обозначим

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

Логарифмическая функция правдоподобия для нашей модели имеет вид

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

Оптимизировать эту конструкцию по $\boldsymbol{\theta}$ напрямую не представляется возможным аналитически и весьма сложно численно.

EM-алгоритм для нормальной смеси

Рассмотрим «скрытые» случайные векторы $\boldsymbol{\Delta}_j$ $(j \in 1:N)$:

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

Шаг E(xpectation) Считая $\boldsymbol{\theta}$ известным и равным $\boldsymbol{\theta}_0$, вычислим

$$\boldsymbol{\gamma}_j = \mathbb{E}\left[\boldsymbol{\Delta}_j | \boldsymbol{\theta}_0, \mathbb{X}\right], i \in 1:N.$$

Шаг M(aximization) Используя $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_N)$, вычислим оценку $\boldsymbol{\theta}$:

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

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

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 классификации

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

Качество классификации:
Accuracy (доля правильно классифицированных): 0.59,
Точность (доля правильно классифицированных как негеномные): 0.75,
Полнота (доля выявленных негеномных): 0.71.

Вывод: классификатору удаётся довольно удачно отсеивать негеномные пути, хотя качество *в целом* — не лучшее.

Данные и результаты классификации

Классификация здесь оказывается плохой сразу по нескольким причинам:

- Несбалансированность классов — негеномных путей значительно больше, чем геномных.
- Большое количество ложно-отрицательных срабатываний объясняется тем, что многие истинные расстояния на самом деле **не могут** наблюдаться из-за недостаточной длины вставки.

Дальнейшие планы

- Сравнить подход, реализованный в **SPAdes**, с **GMM** на уровне оценённых расстояний. Это осложняется тем, что как правило эти подходы выдают «разное» количество расстояний.
- Попытаться улучшить текущую модель.
- Опробовать иные модели.