Задача оценивания кратности повторных последовательностей в геномах

Карасов Николай Дмитриевич, гр. 15Б-04мм

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доцент Коробейников А.И.
Рецензент: м.н.с. Шлемов А.Ю.

СтатМод

Санкт-Петербург
2019г.

Одна из задач биоинформатики — восстановление (сборка) генома по множеству его фрагментов.

- Геном — строка над конечным алфавитом $\mathfrak{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.
- Целью сборки генома является получение как можно более длинных участков генома без потери точности.

Для решения задачи сборки генома используют различные структуры данных, одна из них — граф де Брюйна.

Исходному геному в графе де Брюйна соответствует Эйлеров путь с учетом кратностей ребер.

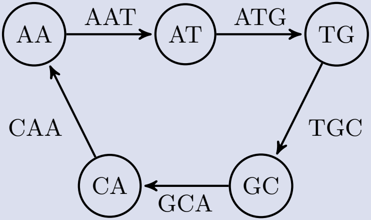
Проблема

Наличие повторяющихся подпоследовательностей в геноме приводит к тому, что нельзя однозначно построить Эйлеров путь.

- Граф де Брюйна — ориентированный граф, в котором в качестве вершин выступают строки длины k, называемая k-мером, и две вершины соединены ребром тогда и только тогда, когда существует строка длины $k+1$, префиксом и суффиксом которой они являются.

Пример

Пусть имеется строка AATGCAA*. Построим по набору ее подстрок длины* 3 *граф де Брюйна:*



Часто в наборе подстрок генома имеются одинаковые последовательности, приводящие к появлению повторов в графе.

- Для некоторых видов повторов можно оценить их кратность локально.

Пример

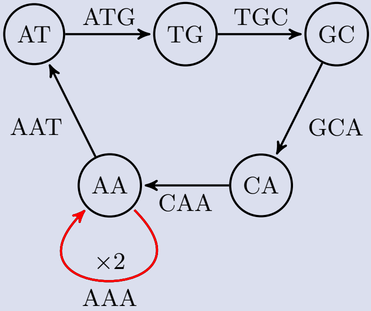


Рис.: *Граф де Брюйна строки* AATGCAAAA, $k = 3$.

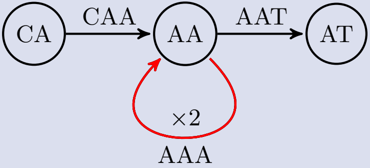


Рис.: *Подграф графа де Брюйна строки* AATGCAAAA.

Виды повторов

Пусть Q, B, D, E — некоторые строки над алфавитом $\mathfrak{A}$, $n \geq 0$.
Рассмотрим повторы двух видов:

- последовательный повтор вида $QB\overbrace{B\ldots B}^{n\text{ раз}}E$

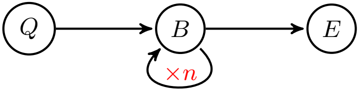


Рис.: Схематичное изображение последовательного повтора.

- тандемный повтор вида $QB\overbrace{DB\ldots DB}^{n\text{ раз}}E$

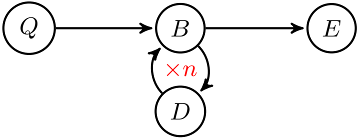


Рис.: Схематичное изображение тандемного повтора.

Задача

Оценить кратность последовательного и тандемного повторов.

Для получения оценки кратности воспользуемся набором парных чтений.

- Фрагмент — подстрока случайной длины η вида $S[\xi, \xi + \eta]$ генома S, где ξ — случайная координата левого конца фрагмента, случайную величину η называют длиной вставки.
- Случайная величина ξ имеет распределение $\mathcal{P}_\xi$, а случайная величина η имеет распределение $\mathcal{P}_\eta$.
- Левое чтение — префикс длины d фрагмента $S[\xi, \xi + \eta]$.
- Правое чтение — суффикс длины d фрагмента $S[\xi, \xi + \eta]$.
- Парные чтения — рассмотренные вместе левое и правое чтения.

Аналитическая формула длины вставки

Пусть имеется набор парных чтений, считаем, что N — максимальное возможное значение кратности повтора.

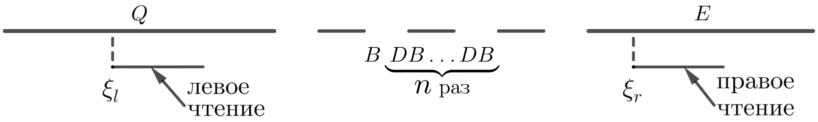


Рис.: Пример расположения пары чтений относительно строк Q и E.

Обозначим $\xi_{i,l}$ и $\xi_{i,r}$ — крайние левые позиции в Q и E левого и правого чтений i-ой пары чтений соответственно.

Утверждение

Если хотя бы один k-мер каждого из чтений лежит в Q и E, то длина вставки i-ой пары чтений равна

$$\eta_i = |Q| - \xi_{i,l} + (n+1)\cdot|B| + n\cdot|D| + \xi_{i,r} + d - 2\cdot(n+1)\cdot k,\ n \in \{0,\dots,N\}.$$

В противном случае длина вставки не наблюдается.

Аналитическая формула кратности повтора

Рассмотрим $\mathbf{X} = \{x_1, \ldots, x_M\}$ — выборка размера M тех парных чтений, у которых хотя бы один k-мер каждого из чтений лежит в Q и E.

- Для i-ой пары чтений из выборки $\mathbf{X}$ длину вставки можно найти по полученной формуле:

$$\eta_i = |Q| - \xi_{i,l} + (n+1) \cdot |B| + n \cdot |D| + \xi_{i,r} + d - 2 \cdot (n+1) \cdot k, \ n \in \{0, \ldots, N\}.$$

- Рассмотрим функцию правдоподобия:

$$\mathrm{L}(n \mid \mathbf{X}) = \prod_{i=1}^{M} \mathrm{P}(\eta = |Q| - \xi_{i,l} + (n+1) \cdot |B| + n \cdot |D| + \xi_{i,r} + d - 2 \cdot (n+1) \cdot k).$$

Оценка кратности повтора

$$\hat{n} = \operatorname*{argmax}_{0 \le j \le N} \log(\mathrm{L}(j \mid \mathbf{X})).$$

Оценка кратности последовательного повтора

- Для получения парных чтений и их позиций в строках Q и E воспользуемся программным пакетом art [Weichun Huang et al., 2011].

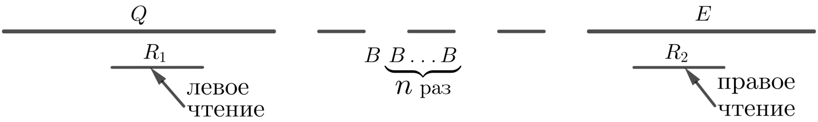


Рис.: Пример расположения пары чтений относительно строк Q и E.

Проведем эксперименты со следующими параметрами:

- $|Q| = |E| = 10000,\ |B| = 100,\ |R_1| = |R_2| = 75$. Длина вставки имеет распределение со средним 1000 и стандартным отклонением 10.
- $|Q| = |E| = 255,\ |B| = 50,\ |R_1| = |R_2| = 75$. Длина вставки имеет распределение со средним 500 и стандартным отклонением 50.

Для каждого случая было проведено 100 моделирований при разных значениях $n \in \{1, \ldots, 7\}$. Все полученные оценки совпали с истинным значением кратности петли.

- Тандемный повтор имеет вид $QB\overbrace{DB\ldots DB}^{n \text{ раз}}E$, где Q, B, D, E — некоторые строки над алфавитом $\mathfrak{A}$, $n \geq 0$.
- Рассмотрим S_0 — геном кишечной палочки $E.\ coli\ K$-$12\ MG1655$. Смоделируем набор парных чтений. Длина вставки имеет нормальное распределение со средним 500 и стандартным отклонением 100.

i	Кратность повтора, n	Оценка кратности, $\hat{n}$
1	1	—
2	2	1
3	1	1
4	1	1
5	1	1

Таблица: Оценки кратностей повторов.

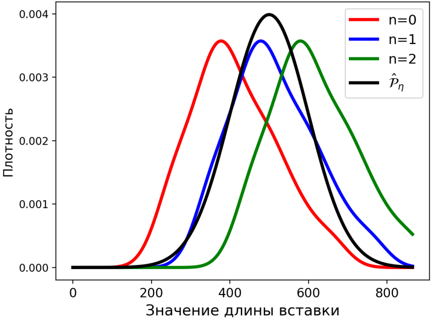


Рис.: Распределение длины вставки.

- Рассмотрим S_0 — геном кишечной палочки $E.\ coli\ K\text{-}12\ MG1655$. Воспользуемся теперь реальным набором парных чтений для S_0 с распределением $\hat{\mathcal{P}}_\eta$ со средним $\hat{\mu} = 503.14$ и стандартным отклонением $\hat{\sigma} = 34.74$.

Изобразим информацию о тандемных повторах и оценки кратностей в виде таблицы.

i	Кратность повтора, n	Оценка кратности, $\hat{n}$
1	1	1
2	1	1
3	1	2
4	1	1
5	1	1

Таблица: Оценки кратностей повторов.

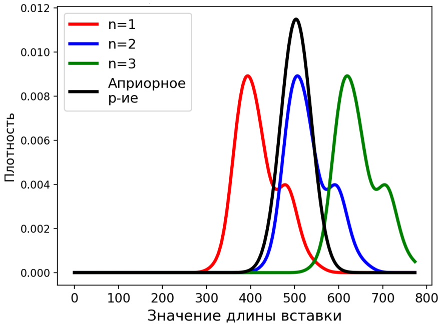


Рис.: Распределение длины вставки.

Оценивание распределения длины вставки

Последовательный повтор имеет вид $QB\overbrace{B\ldots B}^{n \text{ раз}}E$.

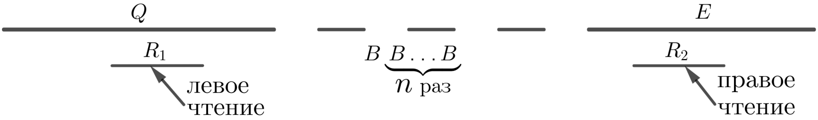


Рис.: Пример расположения пары чтений относительно строк Q и E.

- Для получения оценки кратности повтора использовались чтения, у которых хотя бы один k-мер каждого из чтений лежит в Q и E.
- Таким образом, оцененное распределение длины вставки может отличаться от реального распределения, внося неточности в оценки.

Задача

Оценить распределение длины вставки тех парных чтений, у которых хотя бы один k-мер каждого из чтений лежит в Q и E. Сравнить его с распределением всех парных чтений.

- Пусть $R_1, \ R_2$ — левое и правое чтения соответственно.
- На практике строки Q и E конечной длины и не могут быть короче длины одного k-мера.

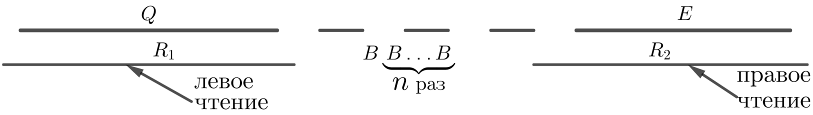


Рис.: Пример расположения пары чтений относительно строк Q и E.

- Если длина чтения больше длины Q или E, то в этом случае считаем, что чтение лежит в строке, если совпали хотя бы k символов.

Именно поэтому считаем, что длина вставки наблюдается, если хотя бы один k-мер каждого из чтений лежит в Q и E.

Рассмотрим случайную величину ζ, которая принимает значения

$$\zeta = \eta, \text{ при условии, что } R_1 \subset Q, R_2 \subset E.$$

- Условие на то, что длина вставки наблюдается, будет следующим:

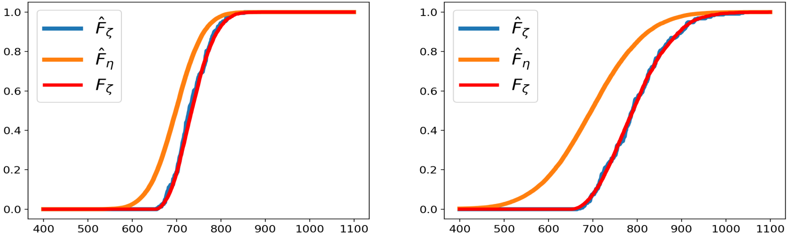
$$|Q| - \xi + n \cdot |B| < \eta - k \text{ и } |Q| - \xi \geq k.$$

Утверждение

$$\mathrm{F}_\zeta(t) = \mathrm{P}(\eta < t \mid |Q| - \xi + n \cdot |B| + k < \eta,\ \xi \leq |Q| - k) =$$
$$= \frac{\sum\limits_{i=0}^{t-1} (\max\{0, \mathrm{F}_\xi(|Q| - k) - \mathrm{F}_\xi(|Q| + n \cdot |B| + k - i)\}) \cdot \mathrm{P}(\eta = i)}{\sum\limits_{i=0}^{|Q|-k} (1 - \mathrm{F}_\eta(|Q| + n \cdot |B| + k - i)) \cdot \mathrm{P}(\xi = i)}.$$

Проверим полученную формулу моделированием и сравним с распределением длины вставки всех парных чтений $\hat{\mathcal{P}}_\eta$ со средним $\hat{\mu}$ и стандартным отклонением $\hat{\sigma}$. Рассмотрим повтор $QB\underbrace{B\ldots B}_{n\text{ раз}}E$. Возьмем следующие параметры:

$$|Q| = |E| = 3000,\ |B| = 100,\ d = 75,\ n = 5.$$



а) $\hat{\mu} = 700,\ \hat{\sigma} = 50.$ б) $\hat{\mu} = 700,\ \hat{\sigma} = 100.$

Рис.: Функция распределения длины вставки парных чтений.

Результаты

- Получены оценки значения кратности последовательного и тандемного повторов.
- Получены аналитические формулы для функции распределения длины вставки.
- Полученные теоретические результаты проверены как с использованием смоделированных чтений, так и с реальными данными.