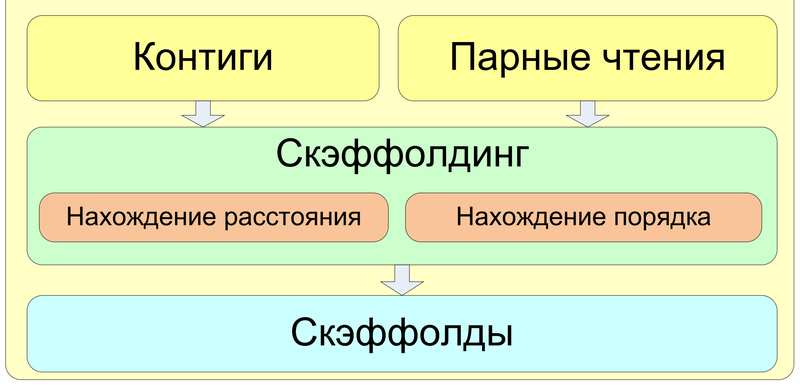
Метод оценки расстояния между контигами на основе принципа максимального правдоподобия

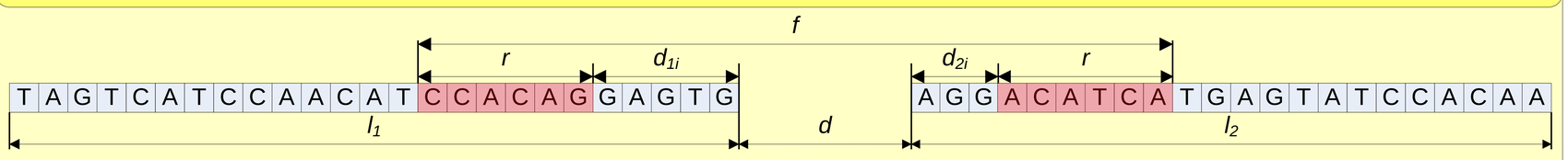
Антон Ахи, Нияз Нигматуллин, Алексей Сергушичев, Федор Царев
Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики
Лаборатория «Алгоритмы сборки геномных последовательностей»

Постановка задачи

- Входные данные: контиги и наборы парных чтений (mate pairs).
- Необходимо: найти расстояния между контигами.
- Задача является частью задачи построения скэффолдов – упорядоченных наборов контигов с известными расстояниями между ними.



Учет связывающих чтений



Вероятность получения фиксированного чтения:

$$p(d, d_{1i}, d_{2i}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\mu - (d_{1i} + d_{2i} + d + 2r))^2}{2\sigma^2} \right) \frac{1}{L}$$

Вероятность фиксированной длины фрагмента

Вероятность фиксированного местоположения фрагмента

$$\ln p(d) = -n \ln\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mu - (d_{1i} + d_{2i} + d + 2r))^2 - n \ln L$$

- Для нахождения правдоподобия вероятности перемножаются.
- Для повышения точности вычисляется логарифм вероятности.
- Выражение – квадратный трехчлен от *d*. После предподсчета вычисляется за *O*(1).
- Для картирования чтений на контиги используется сторонняя программа bowtie.

Учет несвязывающих чтений

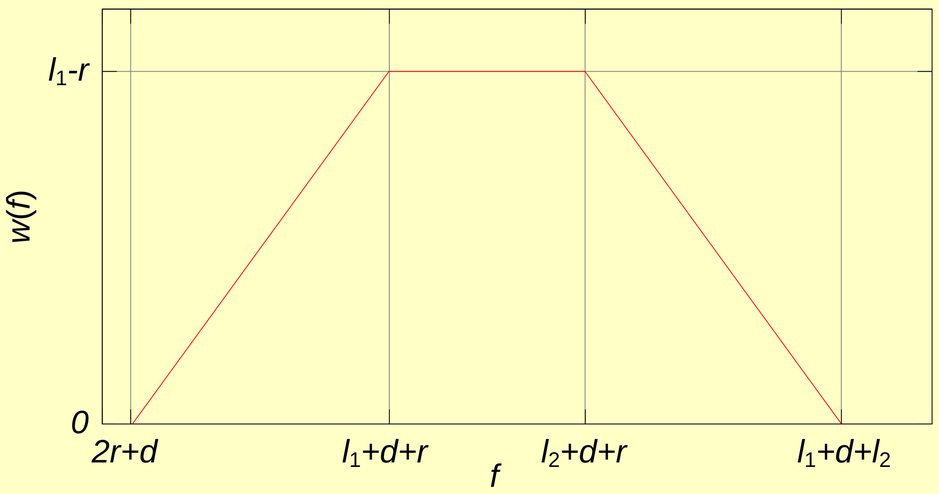
- Остальные чтения не должны связывать контиги.
- *q*(*d*) – вероятность случайного чтения связать пару контигов.

$$q(d) = \sum_{f=2r+d}^{l_1+l_2+d} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\mu - f)^2}{2\sigma^2} \right) \frac{w_d(f)}{L}$$

Вероятность фиксированной длины фрагмента

Вероятность «хорошего» местоположения фрагмента

$w_d(f)$ – число способов разместить чтение с длиной фрагмента *f*, чтобы оно связывало контиги.



q(*d*) можно аппроксимировать, заменив сумму на интеграл, который вычисляется с помощью табличных значений. Это позволяет вычислять *q*(*d*) за *O*(1).

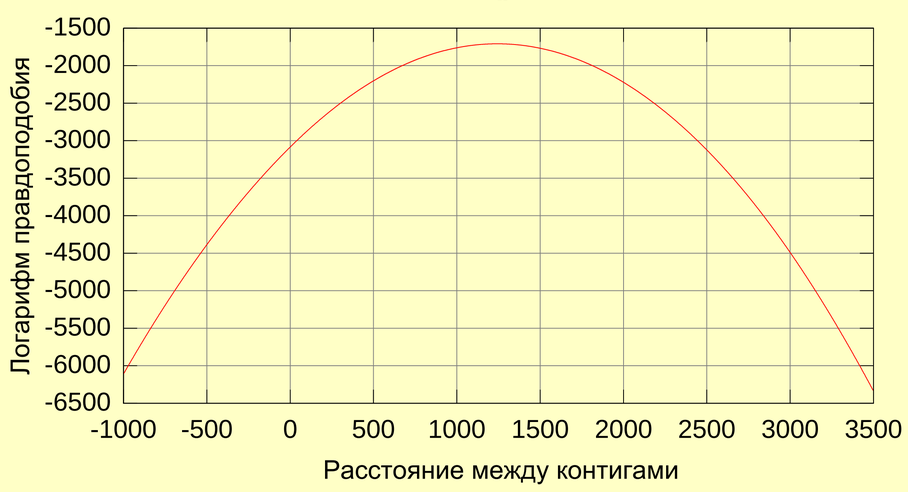
Нахождение расстояния

Функция правдоподобия получается комбинированием вероятностей чтений, связывающих контиги, и чтений, не связывающих контиги.

$$L(d) = p(d)(1 - q(d))^{(reads - n)}$$

- Логарифм функции правдоподобия выпуклый.
- Находим максимум тернарным поиском.

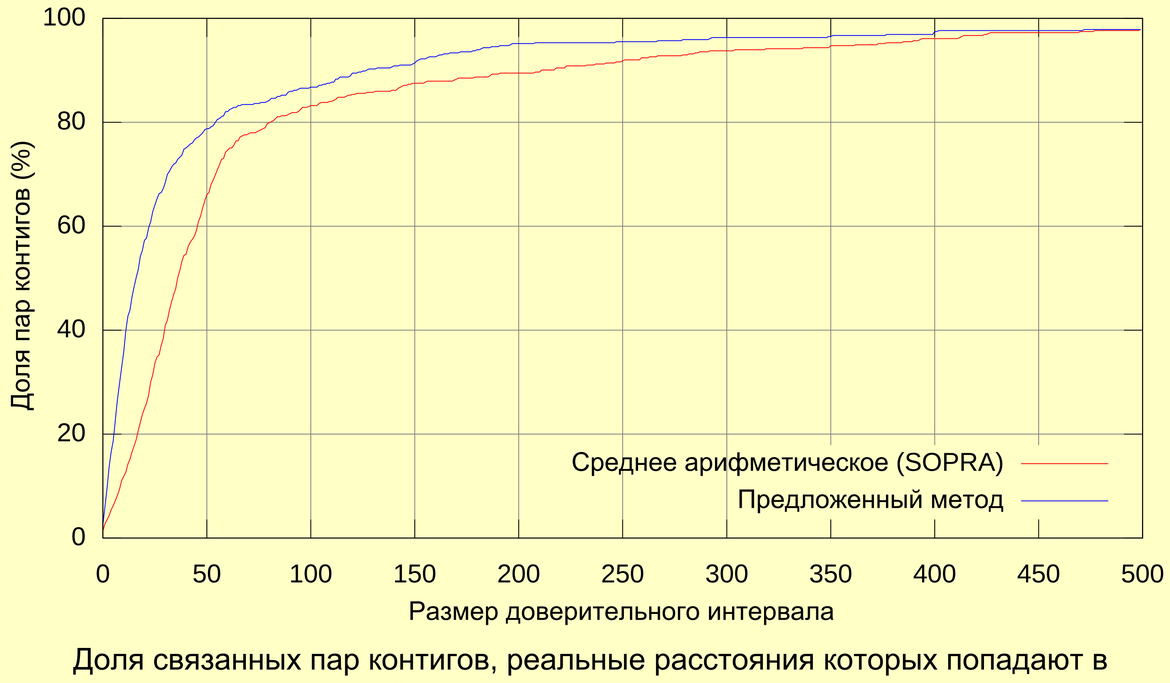
$$d_{opt} = \arg\max_d \ln L(d)$$



- Для фиксированного расстояния вычисление функции правдоподобия осуществляется за *O*(1).
- Нахождение расстояния осуществляется за *O*(*n* + log *L*).

Эксперименты

- Данные – чтения и контиги генома бактерии *E.Coli.*
- 502 контига: N50 = 18047, минимальная длина 235, максимальная длина 73908, средняя длина 9126.
- 300000 пар чтений: длина чтений 36, средний размер фрагмента 3000, стандартное отклонение 300.



Доля связанных пар контигов, реальные расстояния которых попадают в доверительный интервал, в зависимости от его размера.

Результаты

Сравнение с методом среднего арифметического (SOPRA, http://www.biomedcentral.com/1471-2105/11/345):

- В 66% случаев – расстояние оценивается точнее.
- В 10% случаев – оценки совпадают.
- Выигрыш – в среднем на 48%.
- Проигрыш – в среднем на 39%.

Поддержка

Исследование поддержано в рамках Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России на 2009-2013 годы» (Государственный контракт №16.740.11.0495, соглашение №14.B37.21.0562)

anton.akhi@gmail.com http://genome.ifmo.ru/ru/