Сборка скэффолдов геномной последовательности на основе принципа максимального правдоподобия

Андрей Васин
научный руководитель:
к.т.н., доцент кафедры КТ, Ф.Н.Царев

Университет ИТМО

18 июня 2014 года

Введение

- **Геном** — совокупность наследственного материала, заключенного в клетке организма
- **Парные чтения** (mate-pairs) — концы короткого участка генома, полученные в результате его секвенирования
- **Сборка генома** — процесс восстановления исходного генома по имеющемуся набору чтений

Сборка генома

- Процесс сборки генома обычно разделяется на три этапа:
  1. Исправление ошибок в парных чтениях
  2. Сборка **контигов** — длинных последовательностей исходного генома
  3. Сборка **скэффолдов** — наборов упорядоченных и ориентированных контигов с оценками расстояний между соседними контигами

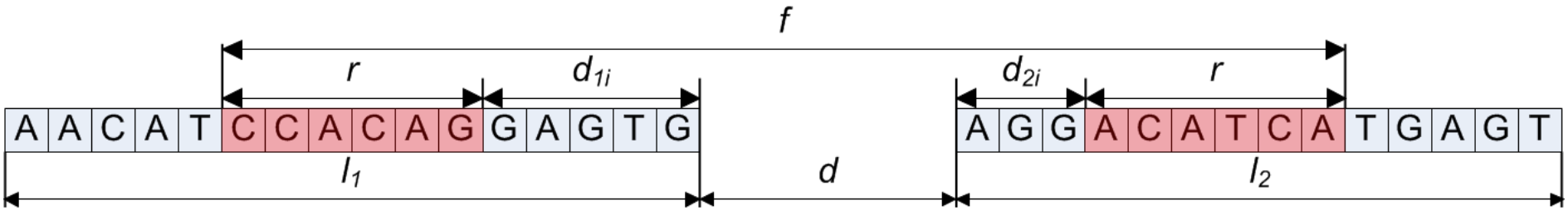
Сборка скэффолдов

- В данной работе сборка скэффолдов подразделяется на четыре этапа:
  1. Первичная оценка расстояний между контигами
  2. Определение взаимного порядка контигов в скэффолдах
  3. Определение взаимной ориентации контигов
  4. Уточнение оценки расстояний между контигами в скэффолдах

Определения

- **Связывающее чтение** — парное чтение, обе части которого картируются на разные контиги
- **Несвязывающее чтение** — любое парное чтение, не являющееся связывающим

Вклад связывающих чтений в формулу правдоподобия



- Вероятность получения фрагмента заданной длины

$$p(fragment\_size_i|d) = \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(fragment\_size_i - \mu)^2}{2\sigma^2})$$

$$fragment\_size_i = d_{1i} + d_{2i} + d + 2r$$

- Вероятность получения фрагмента в данной позиции

$$p(fragment\_position_i|d) = \frac{1}{L}$$

- Вклад связывающих чтений

$$p(connecting|d) = \prod_{i=1}^{n} p(fragment\_size_i|d)p(fragment\_position_i|d)$$

Вклад несвязывающих чтений в формулу правдоподобия

- Вероятность получения случайного связывающего чтения

$$p(random\_connect|d) = \sum_f \frac{w_d(f)}{\sqrt{2\pi}\sigma L} exp(-\frac{(f-\mu)^2}{2\sigma^2})$$

- Вероятность получения несвязывающего чтения

$$1 - p(random\_connect|d)$$

- Вклад несвязывающих чтений

$$p(non - connecting|d) = (1 - p(random\_connect|d))^{R-n}$$

Формула для первичной оценки расстояний

- Итоговая формула

$$p(connecting|d)p(non-connecting|d)$$

- Недостатки:
  - Учитывает наличие только двух контигов
  - Возможна более точная оценка расстояний

Распространение формулы

- Вместо расстояния d вводится вектор расстояний между соседними контигами $\overline{d}$
- Вклад расстояния для связывающего чтения

$$p(fragment\_size_i|\overline{d}) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(fragment\_size_i - \mu)^2}{2\sigma^2})$$

$$fragment\_size_i = d_{1i} + d_{2i} + 2r + \sum_{k=n}^{m-1} \overline{d}_k + \sum_{t=n+1}^{m-1} l_t$$

Распространение формулы

- Вероятность получения случайного связывающего чтения

$$p(random\_connect|\overline{d}) = \sum_{f} \frac{\tilde{w}_d(f)}{\sqrt{2\pi}\sigma L} exp(-\frac{(f-\mu)^2}{2\sigma^2})$$

- Итоговая формула

$$(\prod_{i=1}^{n} p(fragment\_size_i|\overline{d})p(fragment\_position_i|\overline{d}))\times$$

$$\times(1 - p(random\_connect|\overline{d})^{R-n}$$

Результаты

Набор	SOPRA	GAPEST	OPERA	Улучшенный метод
ITMO 1	217 ± 13	211 ± 16	207 ± 14	155 ± 11
ITMO 2	206 ± 14	195 ± 15	212 ± 13	130 ± 11
ABySS	189 ± 13	173 ± 13	163 ± 13	143 ± 10
SPAdes	195 ± 15	175 ± 14	177 ± 13	117 ± 13

Результаты

Набор	Улучшенный метод vs SOPRA	Улучшенный метод vs GAPEST	Улучшенный метод vs OPERA
ITMO 1	71%/6%	70%/6%	65%/5%
ITMO 2	73%/8%	72%/4%	60%/10%
ABySS	75%/8%	71%/3%	64%/6%
SPAdes	73%/5%	75%/7%	63%/8%

Результаты

- Разработана формула правдоподобия для множества контигов в скэффолде
- Усовершенствован существующий сборщик геномной последовательности
- Разработан метод оценки расстояний между контигами в скэффолде на основе принципа максимального правдоподобия, применимый для любого набора скэффолдов

Спасибо за внимание