Про ДНК, исправление ошибок и Echo

Дмитрий Грошев



СПбГУ

26.10.2012

Общие понятия

ДНК

- ...ACGGGCACACTTACG...
- 4 нуклеотида: аденин, тимин, цитозин, гуанин (но это неважно)
- ▶ ДНК \rightarrow РНК \rightarrow белки \rightarrow жизнь
- ▶ 4 миллиарда нуклеотидов у человека, 4Gb данных

Секвенирование

- чтение последовательности ДНК
- текущая технология прочтение случайных участков по 100
- ▶ хорошее покрытие 30+ (4Gb→120Gb)
- очень сложно собирать
- неидеальное прочтение

Ошибка прочтения

Ошибка прочтения

```
GCGC____
__GCTG__
___CTAC__ <---
___TGCA_
___GCAC
```

Ошибка прочтения

```
GCGCTGCAC
GCGC
__GCTG__
__CTAC__ <---
__TGCA_
__GCAC
```

Терминология

- ▶ k-мер
- замена
- ▶ индел (indel, insertion/deletion)

Об ошибках

- ▶ Геном достаточно неравномерен, чтобы риды в большинстве случаев сильно отличались друг от друга
- Покрытие каждого нуклеотида значительно больше 1
- ► Секвенаторы Illumina редко делают инделы, чаще замены
- Секвенаторы Illumina делают больше ошибок ближе к концу рида
- ▶ Ошибки обычно независимы друг от друга и зависят от исходного нуклеотида $(P(A \rightarrow T) \neq P(A \rightarrow C))$

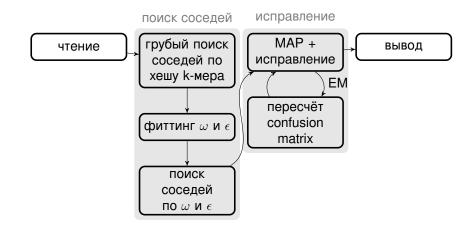
Один из достаточно успешных тулов — ECHO (Wei-Chun Kao, Andrew H. Chan, and Yun S. Song, University of California, Berkeley)

Как работает ЕСНО

2 стадии

- поиск пересекающихся ридов (соседей)
- ▶ исправление ошибок

ECHO



Точные соседи

GCGCTGCACAGTTCGAG

Неточные соседи

GCGCTTCACAGTTCGAG
ATATGAGCTGCACAG

Неточные соседи

GCGCTTCACAGTTCGAG ATATGAGCTGCACAG ^ ^^ ^^^ 12345678901

- $\sim \omega = 11$
- $\epsilon = 2$

Поиск соседей

Для каждого рида:

- поиск точно совпадающего k-мера hashmap'ом (k подобран полуэмпирически)
- подбор ω и ϵ (покрытие конкретного нуклеотида должно иметь распределение Пуассона)
- ightharpoonup поиск неточных соседей по ω и ϵ

Исправление

- Для каждой позиции есть набор значений, нужно наиболее вероятное
- Вероятное в данном случае = Maximum A Posteriori (MAP)
- ▶ Вводится понятие confusion matrix
- Ищутся наиболее вероятные нуклеотиды в перекрывающихся областях

Maximum Likelihood

$$P(A_i) = P_{observed}(A_i)$$
 $\underset{X=A,T,G,C}{\operatorname{arg max}} P(X)$

Maximum A Posteriori

$$P(A_i) = P(A_i | A_{obvserved}) = rac{P(A_{observed} | A_i) P(A_i)}{P(A_{observed})}$$
 $rg \max_{X=A,T,G,C} P(X) pprox P(A_{observed} | P_i) P(A_i)$

Confusion matrix

$$\Phi_{b,b'}^{(m)} = P(r_m = b'|H_m = b)$$

 r_m — нуклеотид в риде в позиции т H_m — нуклеотид в истинном сиквенсе в позиции т

MAP, revisited

Maximum A Posteriori

$$P(A_i) = \prod_{j=1}^{N} \Phi_{READS_{j,i},A}$$
 $rg \max_{X=A,T,G,C} P(X)$

Исправление

- ► Берётся prior confusion matrix (из общих рассуждений или предыдущих экспериментов)
- В группе соседей делается MAP по каждому нуклеотиду с помощью существуеющей confusion matrix и строятся исправленные риды
- На основе исправленных ридов пересчитывается confusion matrix
- Повторять до схождения (исправленный рид не меняется)

Это ЕМ-алгоритм



Вопросы?

Слайды: https://github.com/si14/uni-kse-2012-10/

Использовались картинки:

- http://en.wikipedia.org/wiki/File:Thomas_Bayes.gif
- http://mlp.wikia.com/wiki/File:Happy_Rainbow_Dash_S1E1.png