

Про ДНК, исправление ошибок и Echo

Дмитрий Грошев



СПбГУ

26.10.2012

Общие понятия

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

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

Ошибка прочтения

___GCTG___ <- рид
GCGC___
___GCAC
___CTAC___
___TGCA_

Ошибка прочтения

GCGC____
__GCTG__
__CTAC__ < - - -
____TGCA_
____GCAC

Ошибка прочтения

GCGCTGCAC

GCGC_____

___GCTG___

___CTAC___ < - - -

___TGCA_

___GCAC

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

- ▶ Геном достаточно неравномерен, чтобы риды в большинстве случаев сильно отличались друг от друга
- ▶ Покрывание каждого нуклеотида значительно больше 1
- ▶ Секвенаторы Illumina редко делают инделов, 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 ECHO

2 стадии

- ▶ поиск пересекающихся ридов (соседей)
- ▶ исправление ошибок



Точные соседи

____GCGCTGCACAGTTCGAG
ATATGCGCTGCACAG____
^^^^^^^^^^

Неточные соседи

____GCGCTTCACAGTTCGAG
ATATGAGCTGCACAG____
 ^ ^^^ ^^^^^

Неточные соседи

_____GCGCTTCACAGTTCGAG
ATATGAGCTGCACAG_____

^ ^^^ ^^^^^

12345678901

- ▶ $\omega = 11$
- ▶ $\epsilon = 2$

Для каждого ряда:

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

- ▶ 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

$$P(A_i) = P_{observed}(A_i)$$

$$\arg \max_{X=A,T,G,C} P(X)$$

Maximum A Posteriori

$$P(A_i) = P(A_i|A_{observed}) = \frac{P(A_{observed}|A_i)P(A_i)}{P(A_{observed})}$$

$$\arg \max_{X=A,T,G,C} P(X)$$

$$P(X) \approx P(A_{observed}|P_i)P(A_i)$$

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

r_m — нуклеотид в ряде в позиции m

H_m — нуклеотид в истинном сиквенсе в позиции m

Maximum A Posteriori

$$P(A_i) = \prod_{j=1}^N \Phi_{READS_{j,i}, A}$$

$$\arg \max_{X=A,T,G,C} P(X)$$

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

Это EM-алгоритм



Вопросы?

Слайды: <https://github.com/sil4/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