RepeatScout

Сермухамедов Денис

Руководитель: Комиссаров Алексей
Центр геномной биоинформатики им. Ф.Г.Добржанского

RepeatScout

De novo поиск повторов

C, Perl

Price A.L., Jones N.C. and Pevzner 2005

Письмо Певзнера

Repeat SCout was developed by my former postdoc Alkes Price who is now professor at Harvard working on human genetics - here is not supporting RepeatScout anymore.

I cc to Steve 'O'Brien and Alla Lapidus - do you know who is working on RepeatScout at Dobzhansky?

Thanks!

RepeatScout состоит из двух частей

- Подсчет к-меров с указанием последней позиции к-мера в геноме
- Поиск повторяющихся семейств согласно найденным к-мерам

Идея

Используем к-меры для расширения

TAGCACCTTAGGGCGTCTCGCAACGTCTGCCCACGAACGTTAATCAGTAA

GATTATCATGAAGCGCTTCGCAACGTCTGCAGCTGTCCAGACCGCTGTCA

TATATCCGGTAATCGCCCCGCAACGTCTGCTAACGGGCGTACGGTCGAAT

TGACCTGCTCAGGAGCCTTGCAACGTCTGCTCGCGGATGTGTATGCACGC

ATCCATGCTCGGTATGAATCCAACGTCTGCTCATGGACATCTCATGACGT

CGATCCTCTGAGGCACCTCACAACGTCTGCTCACTGACGCACGGTTGCTG

Идея

GGGCGTCTCGCAACGTCTGCCCACGAACGT

AAGCGCTTCGCAACGTCTGC

AATCGCCCCGCAACGTCTGCTAACGGGCGT

AGGAGCCTTGCAACGTCTGCTCGCGGATGT

CAACGTCTGCTCATGGACAT

AGGCACCTCACAACGTCTGCTCACTGACGC

Consensus: AGGCGCCTCGCAACGTCTGCTCACGGACGT

Greedily extend right/left to optimize $A(Q, S_1, \ldots, S_n)$

Проблема

- Не работает на больших геномах
  - изначально проблема с переполнением int везде по коду
- Ест много памяти
  - 1Gb геном 16Gb на расчет кмеров
  - 1Gb геном >200Gb на поиск повторов (не дождались)
- Работает очень долго
  - подсчет кмеров ~10Mb/min
  - поиск повторов ~1Mb/min

Задачи

1. Нужно минимум починить RepeatScout.
2. Заменить алгоритм подсчета к-меров внутри RepeatScout на что-нибудь более современное и быстрое.
3. Максимальная задача: улучшить алгоритм расчета границ повторов.

Результаты. Починка запуска

- Расширены типы для работы с большими геномами
- Уменьшен расход памяти

Результаты. Замена алгоритма подсчета к-меров

Внутренняя “считалка” очень медленная и прожорливая по памяти, а также снабжала выход дополнительной информацией о положении к-меров. Поэтому был произведен переход на входные данные с jellyfish, что потребовало модификации кода

ссылки

Исходный код:

http://bix.ucsd.edu/repeatscout/

Репозиторий с нашим проектом:

https://github.com/DobzhanskyCenter/RepeatScout2

Спасибо