Představení programu Bowtie 2

Principy zpracování sekvenačních dat a datové formáty

Co je Bowtie?

- program na zarovnání DNA sekvencí
- optimalizovaný pro krátké sekvence 30-50bp
- algoritmus Burrows-Wheeler
- podpora většiny OS

Autor: Ben Langmead, University of Maryland

Vlastnosti Bowtie

+ velice rychlý pro krátké sekvence
+ nízké paměťové nároky
+ podpora standardních formátů (FASTA...)
+ nastavení citlivosti vs. dopad na rychlost
- není tak obecný (př. BLAST)
- v době uvedení podpora zarovnání pouze jednoho vlákna (Paired-End alignment)

Datové formáty

- Vstupní
  - FASTA
  - FASTQ
  - raw...
- Výstupní
  - SAM/BAM (podpora SAMtools)
  - konverzní programy pro .map (Maq)
  - (index pro BWT)

FASTA a FASTQ

- textové formáty pro zápis sekvencí
  - nukleotidů
  - proteinů
- FASTA

```
>hlavička
ATGCGGTGTAGGATGAGCCA...
```

- FASTQ
  - FASTA + hodnocení kvality sekvence
  - rozdíly v interpretaci, kódováno v ASCII pod sekvencí

Výstupní datový formát SAM

- SAM = Sequence Alignment / Map
- standardizovaný textový formát pro zápis zarovnání sekvencí

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
name flag refname pos mapq cigar rnext pnext tlen seq quality
```

- Kompletní specifikace
  - http://samtools.sourceforge.net/SAM1.pdf

Burrows-Wheelerova transformace

- algoritmus původně pro kompresi dat
- založena na výběru prvků ze všech rotací řetězce
- výhoda - nízké paměťové nároky

agat$

$agat
agat$
at$ag
gat$a
t$aga

t$gaa

Burrows-Wheelerova transformace

- rekonstrukce řetězce inverzní transformací
- Last-First mapování

F L

$agat

agat$

at$ag

gat$a

t$aga

$ -> t

t -> a

a -> g

g -> a

agat

BWT pro hledání prefixu

- efektivní hledání řetězců s odpovídajícím prefixem
- výsledek: N po sobě jdoucích řádků
- Bowtie rozšiřuje o "non-exact match"

gat	gat	gat
$agat	$agat	$agat
agat$	agat$	agat$
at$ag	at$ag	at$ag
gat$a	gat$a	gat$a
t$aga	t$aga	t$aga

Bowtie 2

- opět založen na BWT
- optimalizován pro delší sekvence 100-1000bp
  - moderní metody sekvenování
- podporuje lokální zarovnání
- podpora překryvu neurčitých bází (N)
- přepracován algoritmus hodnocení

Bowtie 2 - Lokální zarovnání

- Globální zarovnání (úplné konce)

```
GACTGGGCGATCTCGACTTCG
|||||  ||||||||||| |||
GACTG--CGATCTCGACATCG
```

- Lokální zarovnání

```
ACGGTTGCGTTAA-TCCGCCACG
     |||||||||  ||||||
TAACTTGCGTTAAATCCGCCTGG
```

Systém hodnocení zarovnání

- možnost nastavení konkrétních hodnocení
  - bonus shody
  - malus neshody
  - malus mezery
  - malus neurčité báze (N)
- podobné hodnocení u algoritmů Needleman-Wunsch a Smith-Waterman

Praktické věci

Možnost vyzkoušet i předkompilovanými indexy (E.coli 5MB .. člověk 2.7GB)

- http://bowtie-bio.sourceforge.net
- http://bowtie-bio.sourceforge.net/bowtie2
- Pro Bowtie2 je jich méně

- Dostupnost pro Win/Linux/OS X
  - Nebo zdrojové soubory

Odkazy

- Bowtie
  - http://bowtie-bio.sourceforge.net
- Bowtie 2
  - http://bowtie-bio.sourceforge.net/bowtie2
- Bowtie paper
  - http://genomebiology.com/2009/10/3/R25
- SAMtools
  - http://samtools.sourceforge.net/
- BWT v Bowtie
  - http://www.biomedcentral.com/content/pdf/gb-2009-10-3-r25.pdf