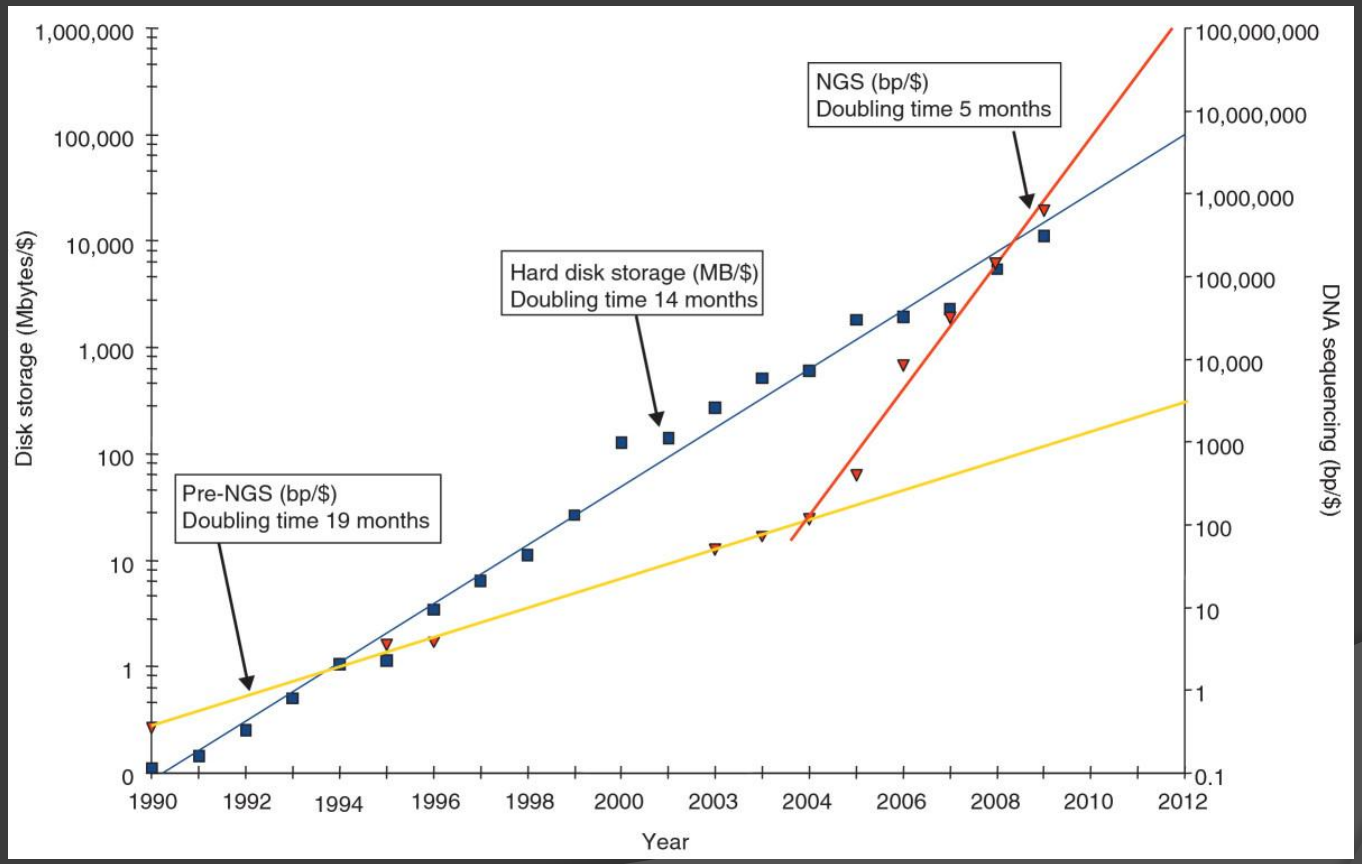
Kompresia príbuzných DNA sekvencií

Tomáš Šurín

Zima 2011/2012

Motivácia

- cena sekvencovania DNA klesá rýchlejšie ako cena za MB na úložných médiách



Lincoln D. Stein (2010)

Motivácia

- projekty sekvencovania DNA
  - 1000 Genomes project
  - 1000 Plant genomes project
  - Genome 10K vertebrate genome project
  - a veľa ďalších
- výstup týchto projektov väčšinou obsahuje viacero sekvencií od rovnakých alebo podobných organizmov
- vysoká redundancia v dátach (u ľudí sú genómy podobné na 99.9%)

Tradičné metódy kompresie DNA

- všeobecne zamerané algoritmy – gzip, compress
- kanonické kódovanie – 2bit/báza
- kompresia opakujúcich sa podreťazcov/palindromov
  - Offline
- kompresia opakujúcich sa podreťazcov používajúca nepresné zhody
  - GenCompress
- modernejšie
  - XM

Kompresia príbuzných sekvencií

- využitie univerzálnej metódy kompresie
- využitie self-indexu
- využitie referenčnej sekvencie

Využitie univerzálnej metódy kompresie

- zip
  - slovníková metóda (LZ77)
  - malé náhľadové okno
- 7zip
  - slovníková metóda (LZ77)
  - veľké náhľadové okno (~2GB)
  - žiadny náhodný prístup
- gramatická kompresia (re-pair, comrad)
  - využitie bezkontextovej gramatiky
  - umožňuje náhodný prístup
  - zachytí opakovania aj ďaleko od seba

Self-index

- **index** = štruktúra umožňujúca nasledujúce operácie nad textom T
  - `count(p)` – počet výskytov vzoru p v T
  - `locate(p)` – miesta výskytov vzoru p v T
  - `display(i,j)` – vráti podreťazec T[i,j]
- **self-index** = index, ktorý nepotrebuje text T pre implementáciu vyššie uvedených operácií a jeho pamäťová zložitosť je menšia ako veľkosť textu T
- na webe Pizza&Chilli [1] je dostupných viacero implementácií self-indexov

Implementácie

- RLCSA
  - podľa autorov najlepší self-index pre opakujúce sa sekvencie
  - všeobecné zameranie – SVN, wiki, DNA, proteíny
- LZ77 based self index [2]
  - podľa autorov je lepší než RLCSA (väčšinou 2x lepší kompresný pomer a rýchlejšie operácie)
  - stále horší kompresný pomer než 7zip

Využitie referenčnej sekvencie

- **Princíp:** pri kódovaní využiť kontext referenčnej sekvencie
- napr. kódovanie rozdielov voči referencii
  - zmena bázy
  - vloženie/vymazanie bázy
- **Problémy:**
  - voľba referenčnej sekvencie
  - ako nájsť zmeny v dátach?

Implementácie

- DNACompress
  - na nájdenie zmien používa komerčný program PatternHunter
- BioZip (human genomes as email attachments)
  - kompresia genómu Jamesa Watsona z približne 3GB na 4 MB
  - využíva referenčnú sekvenciu a SNP mapu (spolu 4.2 GB)
  - problém s distribúciou referenčnej sekvencie a SNP mapy
- RLZ (relative Lempel-Ziv factorization)
  - slovníková metóda
  - rozdelí kódovanú sekvenciu na najdlhšie fráze, ktoré sa vyskytujú v referenčnej sekvencii
  - kóduje len miesta a dĺžky výskytov frází
  - umožňuje aj implementáciu rýchleho náhodného prístupu

Testované nástroje

- general-purpose
  - 7zip (v9.20)
  - zip (zip v2.32 / unzip v6.0)
  - Re-Pair (v1.0.1)
- špeciálne určené pre DNA kompresiu
  - RLZ (verzia z februára 2011)
  - XM (v2.0)
- self-index
  - RLCSA (verzia z augusta 2011)

Testované sekvencie

- chrípka (influenza) ~ 78000 sekvencií
- kvasinky (S. paradoxus) ~ 800 sekvencií
- hemoglobín ~ 15200 sekvencií
- človek chromozóm Y (chrY)– hg15, hg16, hg17, hg18, hg19

Predspracovanie dát

- kvôli niektorým testovaným nástrojom bolo potrebné predspracovanie sekvencií
  - odstránenie metainformácií o sekvenciách
  - zmena na lowercase
  - odstránenie znakov iných než acgtn

	Pôvodná veľkosť	Veľkosť po spracovaní
hemoglobin	9.6 MB	7.0 MB
influenza	117.4 MB	107.4 MB
sparadoxus	11.4 MB	11.3 MB
chrY	268.5 MB	263.3 MB

Výsledky - Hemoglobin

hemoglobin			
	Veľkosť	**Pamäť**	**Čas**
7zip	0.63 MB / 8.96%	**200 MB**	**0:04/0:01**
zip	0.80 MB / 11.31%	**3.5 MB**	**0:04/0:01**
repair	**0.96 MB / 13.58%**	130 MB	**0:04/0:01**
xm	**0.53 MB / 7.56%**	190 MB	**9:55/10:24**
RLCSA	0.94 MB / 13.42%	70 MB	0:06/0:02

Výsledky - Influenza

influenza			
	Veľkosť	**Pamäť**	**Čas**
7zip	1.87 MB / 1.75%	200 MB	0:56/0:05
zip	7.52 MB / 7.00%	**3.5 MB**	**0:46**/0:06
repair	3.08 MB / 2.87%	**1.5 GB**	1:13/**0:03**
xm	**1.69 MB / 1.57%**	900 MB	**3:51:59/-**
RLCSA	**13.40 MB / 12.48%**	350 MB	12:59/6:20

Výsledky – S. paradoxus

sparadoxus			
	Veľkosť	**Pamäť**	**Čas**
7zip	2.84MB / 25.12%	200 MB	0:17/0:02
zip	3.15MB / 27.82%	**3.5 MB**	0:19/**0:01**
repair	4.86MB / 34.12%	160 MB	**0:11/0:01**
xm	**2.63MB / 23.24%**	**250 MB**	**5:48/7:19**
RLCSA	**7.62MB / 67.43%**	100 MB	0:16/0:04

Výsledky – ChrY

chrY			
	Veľkosť	**Pamäť**	**Čas**
7zip	**22.02 MB / 8.36%**	200 MB	3:47/**0:05**
zip	30.82 MB / 11.71%	**3.5 MB**	**3:14**/0:06
repair	-	-	-
xm	-	-	-
RLCSA	**52.87 MB / 20.08%**	**400 MB**	**34:20/14:25**

Výsledky RLZ

RLZ kompresia				
Referenčná sekvencia	**Veľkosť referencie**	**Veľkosť sekvencií bez referencie po kompresii**	**Velkosť po 2bit kódovaní referencie**	**Čas**
hg15	48.5MB	3.84MB	**15.97MB / 6.06%**	3:27/0:06
hg16	47.9MB	1.1MB	**13.08MB / 4.97%**	3:00/0:05
hg17	55MB	0.922MB	14.67MB / 5.57%	3:00/0:06
hg18	55MB	0.476MB	14.23MB / 5.40%	2:57/0:05
hg19	56.6MB	0.474MB	14.62MB / 5.55%	3:06/0:06

využitie pamäte: 250 MB

Záver

- RLZ
  - najlepší pre chrY
  - referenčnú sekvenciu možno uchovávať mimo komprimovaných dát
    - problém so správou verzií referencií
- XM
  - veľmi dobré výsledky
  - avšak pomalý
- gramatická kompresia
  - doháňa 7z
  - navyše ponúka náhodný prístup
  - vytvoriť optimálnu gramatiku je NPC
  - optimálna gramatika je však horšia než LZ77
- self-index
  - nie je vhodný ak
    - potrebujeme len ukladať dáta
    - potrebujeme prístup len k niektorým pozíciam (napr. začiatkom sekvencií)
  - príliš veľky overhead

Zdroje

[1] *Pizza&Chilli.* http://pizzachili.dcc.uchile.cl

[2] Sebastian Kreft, Gonzalo Navarro. *Self-Indexing Based on LZ77*. 2011.

[3] Shanika Kuruppu. *Comrad, RLZ*. http://ww2.cs.mu.oz.au/~kuruppu/

[4] *XM.* ftp://ftp.infotech.monash.edu.au/software/DNAcompress-XM/README.html