DNA – Compression Algorithms

Martin Všetička

Lempel–Ziv–Welch (LZW)

- Family of lossless compression techniques known as *dictionary coders*.
- The basic idea is to replace repetitions by (shorter!) references to a “dictionary”.
- *Adaptive dictionary* is used
  - Dictionary is being build in a single pass, while at the same time also encoding the data.

Algorithms

- Biocompress (1993) [LZW]
  - First DNA-specific compression algo.
- Biocompress-2 (1994) [LZW]
- Cfact (1996)
- GenCompress (1999) [LZW]
- DNACompress (2002) [LZW]
- DNAPack (2005)

DNACompress

- Uses LZW compression scheme.
- Typically much faster than GenComp.
- Two phases:
  - – Find all approximate repeats including complementary palindromes
    - **PatternHunter** – approximate repeat search engine; at the same sensitivity levels is over two orders of magnitudes faster than Blastn!
  - – Encode approximate repeat regions and non-repeat regions.

[Paper]

DNACompress (2) - Algorithm

1. Run *PatternHunter* and output all approximate repeats (and approximate reverse complements) into a list *A* in the order of descending scores;
   - The selection of which repeats are more optimal for sequence compression can be deferred at the end of PatternHunter homology search!
2. Extract a repeat *r* with highest score from list *A* and add *r* into another repeat list *B*;
3. Process each repeat in *A* so that there’s no overlap with the extracted repeat *r*;
4. Goto step 2 if the highest score of repeats in *A* is still higher than a pre-defined threshold; otherwise exit.

DNACompress (3) - Algorithm

- *DNACompress* checks each repeat to see whether it saves bits to encode. If not, it will be discarded. At the end, all the remaining regions other than repeats are concatenated together and then sent as input to a two-order arithmetic coder
- DNACompress uses almost the same encoding as GenCompress.

DNACompress (4)

Sequence	Size	Biocompress-2	GenCompress	CTW+LZ	DNACompress	Encoding time
CHMPXX	121024	1,6848	1,673	1,669	**1,6716**	6.21s
CHNTXX	155844	1,6172	1,6146	1,6129	**1,6127**	5.58s
HEHCMVCG	229354	1,848	1,847	**1,8414**	1,8492	5.41s
HUMDYSTROP	38770	1,9262	1,9226	1,9175	**1,9116**	3.21s
HUMGHCSA	66495	1,307	1,1048	1,0972	**1,0272**	7.45s
HUMHBB	73323	1,88	1,8204	1,8082	**1,7897**	4.04s
HUMDABCD	58864	1,877	1,8192	1,8218	**1,7951**	6.13s
HUMHPRTB	56737	1,9066	1,8466	1,8433	**1,8165**	5.08s
MPOMTCG	186608	1,9378	1,9058	1,9	**1,892**	5.84s
PANMTPACGA	100314	1,8752	1,8624	**1,8555**	1,8556	4.22s
VACCG	191737	1,7614	1,7614	1,7616	**1,758**	6.60s
average	---	1,7837	1,7434	1,7389	**1,7254**	---*

[Source]

PatternHunter

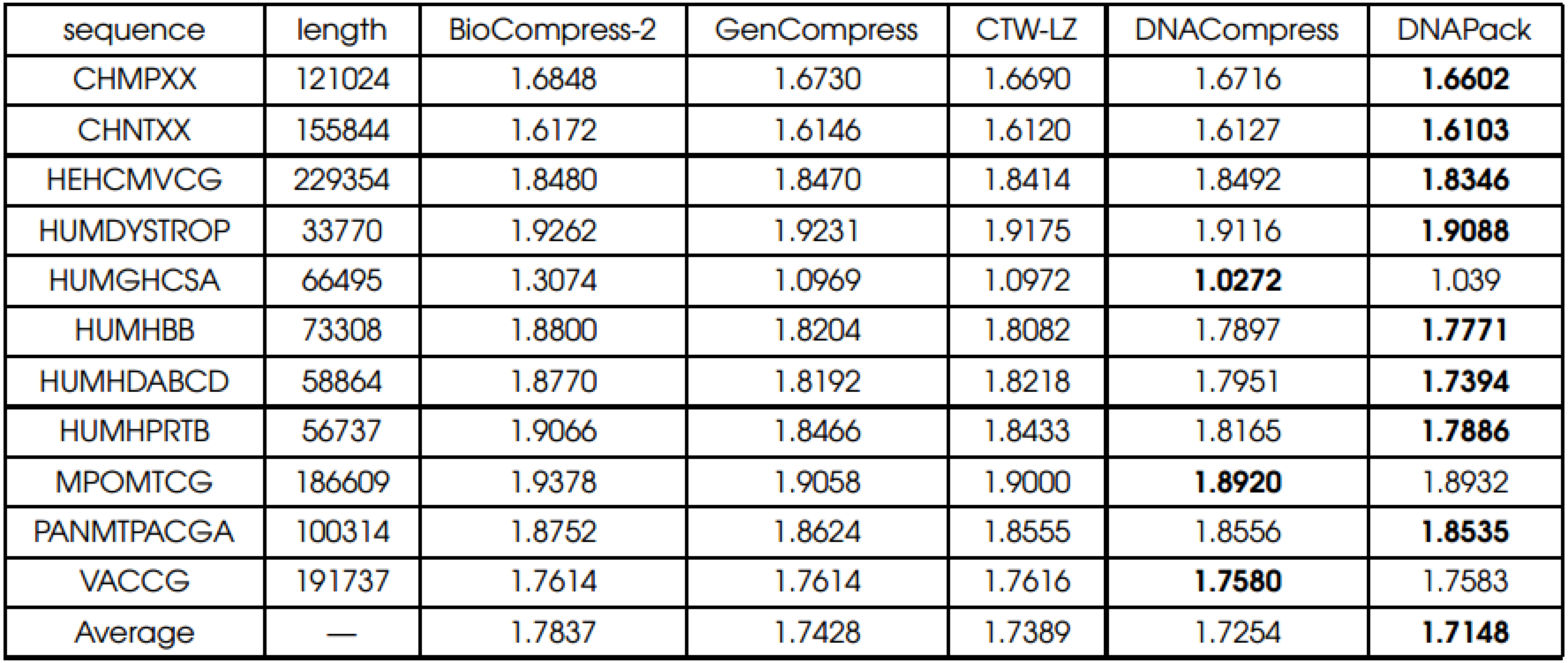
- Commercial program
- Blast finds short exact ‘seed’ matches (hits), which are then extended into longer alignments.
- Blast looks for matches of *k* (default *k = 11* in Blastn) **consecutive** letters as seeds. PatternHunter looks for **nonconsecutive** *k* letters as seeds. This seemingly simple change has a surprisingly large effect on sensitivity.

[Paper]

DNAPack

- http://fabrice.lefessant.net/src/dnapack
- Source codes: **No**
- Binary: **No**
- Paper: Available

DNAPack (2) - Compression results



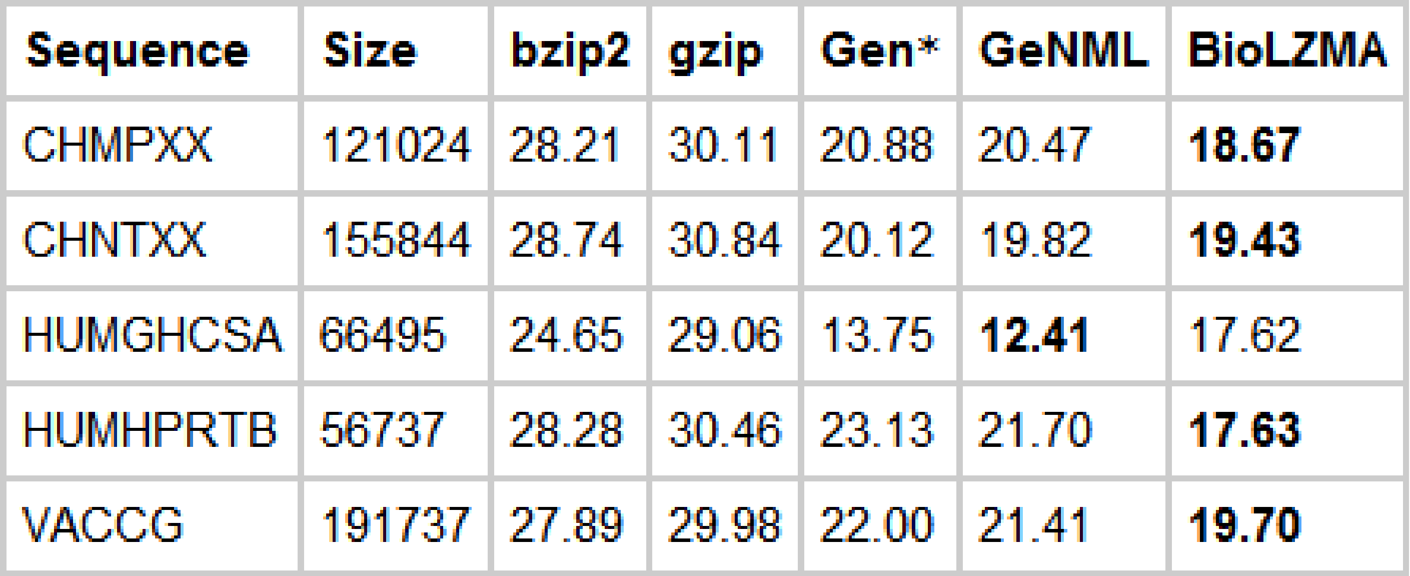
sequence	length	BioCompress-2	GenCompress	CTW-LZ	DNACompress	DNAPack
CHMPXX	121024	1.6848	1.6730	1.6690	1.6716	**1.6602**
CHNTXX	155844	1.6172	1.6146	1.6120	1.6127	**1.6103**
HEHCMVCG	229354	1.8480	1.8470	1.8414	1.8492	**1.8346**
HUMDYSTROP	33770	1.9262	1.9231	1.9175	1.9116	**1.9088**
HUMGHCSA	66495	1.3074	1.0969	1.0972	**1.0272**	1.039
HUMHBB	73308	1.8800	1.8204	1.8082	1.7897	**1.7771**
HUMHDABCD	58864	1.8770	1.8192	1.8218	1.7951	**1.7394**
HUMHPRTB	56737	1.9066	1.8466	1.8433	1.8165	**1.7886**
MPOMTCG	186609	1.9378	1.9058	1.9000	**1.8920**	1.8932
PANMTPACGA	100314	1.8752	1.8624	1.8555	1.8556	**1.8535**
VACCG	191737	1.7614	1.7614	1.7616	**1.7580**	1.7583
Average	—	1.7837	1.7428	1.7389	1.7254	**1.7148**

BioLZMA

- BioLZMA is a user-friendly cross-platform DNA data compression software developed by Shenzhen University - Texas Instruments DSPs Laboratory.
- Binary: **Yes, not stable though!**
- Paper: **No**
- Source codes: **Yes**
- License: **GNU GPL v3**
- http://code.google.com/p/biolzma

BioLZMA (2)

- Comparison:



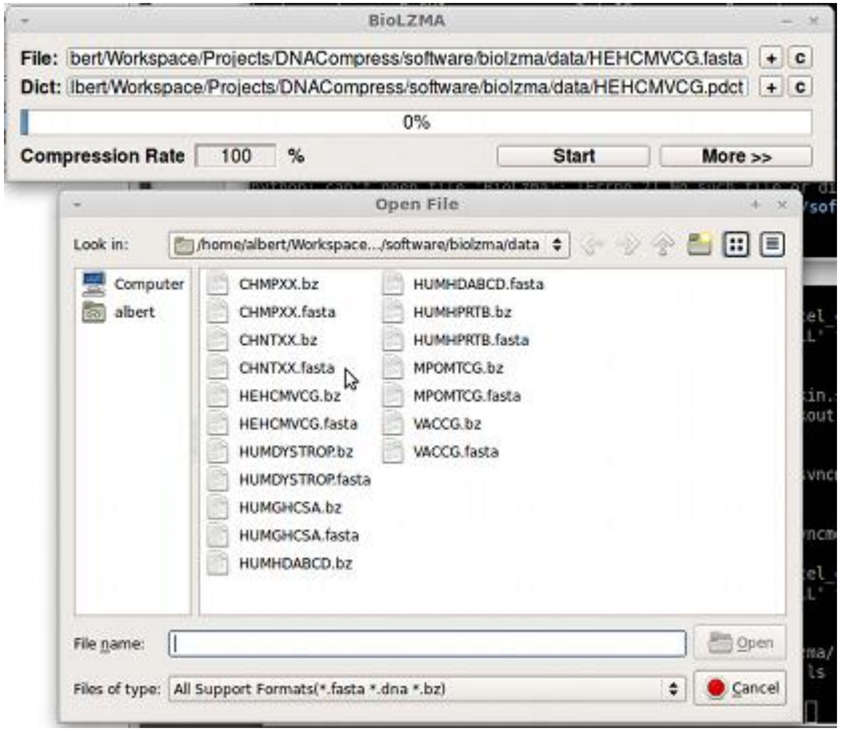
Sequence	Size	bzip2	gzip	Gen*	GeNML	BioLZMA
CHMPXX	121024	28.21	30.11	20.88	20.47	**18.67**
CHNTXX	155844	28.74	30.84	20.12	19.82	**19.43**
HUMGHCSA	66495	24.65	29.06	13.75	**12.41**	17.62
HUMHPRTB	56737	28.28	30.46	23.13	21.70	**17.63**
VACCG	191737	27.89	29.98	22.00	21.41	**19.70**

  - Gen* is the abbreviation of GenCompress

BioLZMA (3) - Advantages

- **Simple**: BioLZMA based on existing compression techniques like Huffman coding and LZMA compression. It's easy to implement.
- **Modularity**: BioLZMA consists of several encoding sub-procedures. These procedures can be replaced or reconfiguration for each compression process in order to achieve better performance.
- **Bioinformatics Meanings**: In BioLZMA, the DNA base symbols ('A', 'T', 'C' and 'G') will be translated into (one or several) amino acid symbols before compression. Experimental results show that by doing this, the compression rate can be significantly improved. It shows that the fragments similarities in amino acid sequences is higher than that in DNA symbol sequences.
- **High Performance**

BioLZMA (4)



BioLZMA

File: bert/Workspace/Projects/DNACompress/software/biolzma/data/HEHCMVCG.fasta

Dict: lbert/Workspace/Projects/DNACompress/software/biolzma/data/HEHCMVCG.pdct

0%

Compression Rate 100 %

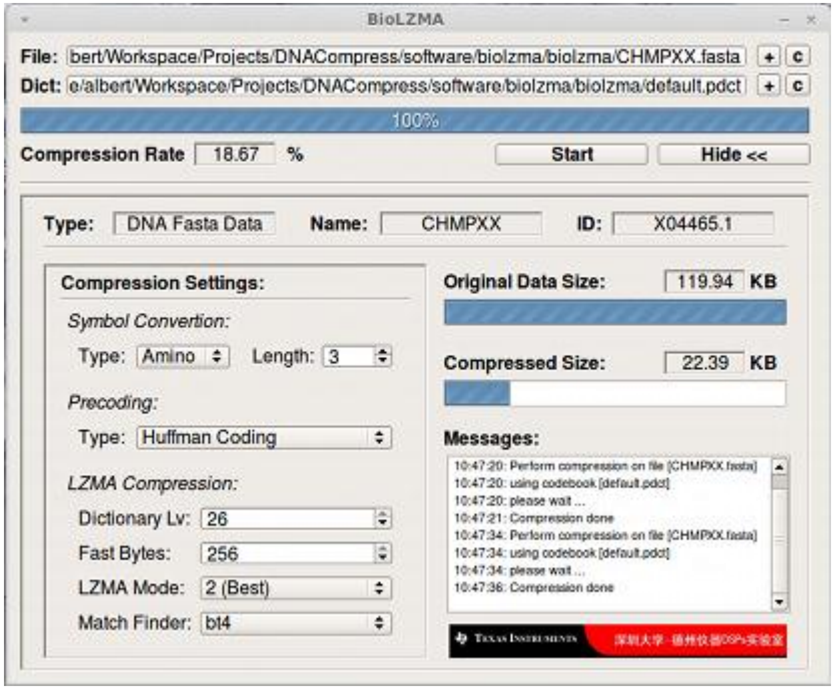
Open File

Look in: /home/albert/Workspace.../software/biolzma/data

```
CHMPXX.bz          HUMHDABCD.fasta
CHMPXX.fasta       HUMHPRTB.bz
CHNTXX.bz          HUMHPRTB.fasta
CHNTXX.fasta       MPOMTCG.bz
HEHCMVCG.bz        MPOMTCG.fasta
HEHCMVCG.fasta     VACCG.bz
HUMDYSTROP.bz      VACCG.fasta
HUMDYSTROP.fasta
HUMGHCSA.bz
HUMGHCSA.fasta
HUMHDABCD.bz
```

File name:

Files of type: All Support Formats(*.fasta *.dna *.bz)



BioLZMA

File: bert/Workspace/Projects/DNACompress/software/biolzma/biolzma/CHMPXX.fasta

Dict: e/albert/Workspace/Projects/DNACompress/software/biolzma/biolzma/default.pdct

100%

Compression Rate 18.67 %

Type: DNA Fasta Data   Name: CHMPXX   ID: X04465.1

Compression Settings:

Symbol Convertion:

Type: Amino   Length: 3

Precoding:

Type: Huffman Coding

LZMA Compression:

Dictionary Lv: 26

Fast Bytes: 256

LZMA Mode: 2 (Best)

Match Finder: bt4

Original Data Size: 119.94 KB

Compressed Size: 22.39 KB

Messages:

```
10:47:20: Perform compression on file [CHMPXX.fasta]
10:47:20: using codebook [default.pdct]
10:47:20: please wait ...
10:47:21: Compression done
10:47:34: Perform compression on file [CHMPXX.fasta]
10:47:34: using codebook [default.pdct]
10:47:34: please wait ...
10:47:36: Compression done
```

Other algorithms

- DNABIT Compress
  - Genome compression algorithm
  - Article
  - Supplementary material
- ReCoil
  - An algorithm for compression of extremely large datasets of DNA data
  - Article
- GRS
  - A novel compression tool for efficient storage of Genome Re-Sequencing data
  - Source code

Other algorithms (2)

- G-SQZ
  - Genomic Squeeze (G-SQueeZ™) is a technique to encode genomic sequence-quality data into an indexed, compact binary format, and that can result in substantial savings in storage and processing over conventional plain text formats (such as FASTQ, CSFASTA/QUAL formats).
  - Website
  - Paper
  - Example data
- DSCR
  - DNA Sequence Reads Compression
  - http://sun.aei.polsl.pl/dsrc/

Sequence Squeeze Competition

- The Pistoia Alliance, in the interests of promoting pre-competitive collaboration, is putting forward a prize fund of **US$15,000** to the best novel open-source NGS compression algorithm submitted before the closing date of 15 March 2012.

Thank you!

DNA Sequence Reads Compression (DSRC)

- DSRC is able to:
  - compress files from DNA sequencing in FASTQ format,
  - decompress whole file,
  - decompress only a single record without decompressing the complete file.
- Compression factor - DSRC is usually:
  - 35–55% better than gzip,
  - 15–25% better than bzip2