GABAC: AN ARITHMETIC CODING SOLUTION FOR GENOMIC DATA

TOM PARIDAENS1, JAN VOGES2, MIKEL HERNAEZ3, JAN FOSTIER1, AND JÖRN OSTERMANN2

1 IDLAB, GENT UNIVERSITY – IMEC, GENT, BELGIUM
2 INSTITUT FÜR INFORMATIONSVERARBEITUNG (TNT), LEIBNIZ UNIVERSITY, HANNOVER, GERMANY
3 CARL R. WOESE INSTITUTE FOR GENOMIC BIOLOGY, UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN, URBANA, ILLINOIS, USA

Introduction: In an effort to provide a response to the ever-expanding generation of genomic data, MPEG (Moving Picture Experts Group), under the auspices of ISO (International Organization for Standardization), is designing a new solution for the representation, compression and management of genomic sequencing data: the MPEG-G standard. Part 2 of the MPEG-G standard focuses on specifying the coding of the sequencing data. This work discusses the first implementation of an MPEG-G compliant entropy encoder/decoder: GABAC.

Methods: GABAC combines proven coding technologies, such as context-adaptive binary arithmetic coding (CABAC), binarization schemes, and transformations into one straightforward solution for the compression of the sequencing data.

Results

- To analyze the performance of the GABAC encoder, a test set of 206 descriptor stream files has been selected. Files were extracted from the MPEG-G Genomic Information Database (total uncompressed size: 12.97 GiB).
- Adding GABAC to the CRAM codec set offers a significant performance gain, both in compression ratio and in encoding speed.

Compressed Size | Encoding Time | Decoding Time
--- | --- | ---
gzip | 3,524 MiB | 3h 25m 18s | 06m 02s
bzip2 | 3,088 MiB | 33m 55s | 20m 00s
xz | 2,944 MiB | 4h 47m 38s | 09m 25s
rANS-0 | 4,143 MiB | 06m 01s | 07m 08s
rANS-1 | 3,400 MiB | 06m 54s | 08m 20s
GABAC | 2,877 MiB | 45m 25s | 20m 18s
CRAM | 2,879 MiB | 2h 25m 58s | 09m 32s
CRAM + GABAC | 2,800 MiB | 1h 01m 17s | 20m 08s

Conclusions

- We present the first implementation of an MPEG-G Part 2 compliant entropy codec.
- Our implementation already outperforms well established codecs and can serve as a reference for future implementations. The performance of new implementations of the specification is expected to improve over time.
- Adding GABAC to the CRAM codec set offers significant performance gains, both in compression ratio and in encoding speed.