Genie: an MPEG-G conformant software to compress genomic data.

B. Bliss
University of Illinois at Urbana-Champaign
Urbana, IL, USA

J. Allen
University of Illinois at Urbana-Champaign
Urbana, IL, USA

S. Baheti
Mayo Clinic
Rochester, MN, USA

M.A. Bockol
Mayo CLinic
Rochester, MN, USA

S. Chandak
Stanford University
Stanford, CA, USA

J. Delgado
Universitat Politecnica de Catalunya
Barcelona, Spain

J. Fostier
IDLab, Ghent University
Ghent, Belgium

J.L. Gelpi
Universitat de Barcelona
Barcelona, Spain

S.N. Hart
Mayo Clinic
Rochester, MN, USA

M. Hernaez Arrazola
University of Illinois at Urbana-Champaign
Urbana, IL, USA

M. Hudson
University of Illinois at Urbana-Champaign
Urbana, IL, USA

M.T. Kalmbach
Mayo Clinic
Rochester, MN, USA

E.W. Klee
Mayo Clinic
Rochester, MN, USA

L.S. Mainzer
University of Illinois at Urbana-Champaign
Urbana, IL, USA

F. Müntefering
Institut für Informationsverarbeitung (TNT), Leibniz University
Hannover, Germany

D. Naro
Barcelona Supercomputing Center
Barcelona, Spain

I. Ochoa-Alvarez
University of Illinois at Urbana-Champaign
Urbana, IL, USA

J. Ostermann
Institut fur Informationsverarbeitung (TNT), Leibniz University
Hannover, Germany

T. Paridaens
IDLab, Ghent University
Ghent, Belgium

C.A. Ross
Mayo Clinic
Rochester, MN, USA

J. Voges
Institut fur Informationsverarbeitung (TNT), Leibniz University
Hannover, Germany

E.D. Wieben
Mayo Clinic
Rochester, MN, USA

M. Yang
University of Illinois at Urbana-Champaign
Urbana, IL, USA

T. Weissman
Stanford University
Stanford, CA, USA

M. Wiepert
Mayo Clinic
Rochester, MN, USA

Supercomputing '19, November 17-22, 2019, Colorado

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

ABSTRACT

Problem statement. The cost of sequencing a whole human genome has dropped from was around $20 million in 2004 to a million in 2008, and a mere $1,000 in 2015. This decrease in cost, together with the advancements in sequencing technology, has allowed the field of medical genomics to rapidly develop, enabling the design of individualized drugs and diagnoses, helping mitigate risks, prevent diseases and treat them effectively when they occur. Currently the

amount of genomic sequencing data is doubling approximately every seven months. At this rate more than an exabyte of sequencing data will be produced per year, approaching zettabytes by 2025 [10]. Often, these data are unique: the samples are not available for re-sequencing. The tools used to process these data also improve over time. It is beneficial to regularly revisit and re-analyze data, which requires their long-term storage. Consequently, data storage, transmission, visualization and scalable processing have become major challenges in the advancement of biological and medical science research. This situation calls for state-of-the-art, efficient and secured compressed representations of massive biological datasets, that can not only alleviate the storage requirements, but also facilitate the exchange, dissemination and accession of these data. This undertaking is of paramount importance, as data storage and acquisition are becoming the major bottleneck, evidenced by the recent flourishing of solutions enabling processing the data directly in the cloud.

Defining the solution. Motivated by these facts, the Moving Picture Experts Group (MPEG)—a joint working group of the International Standardization Organization (ISO) and the International Electrotechnical Commission (IEC)—has developed MPEG-G, a new open standard [1] to compress, store, transmit and process genomic sequencing data (https://mpeg.chiariglione.org/standards/mpeg-g). A detailed specification embeds mechanisms to resolve the above technical difficulties, while ISO backing provides the assurance of long-term support. The next crucial step is developing software that delivers the benefits of MPEG-G.

Software development. We have developed GENIE, the first open-source implementation of an encoder-decoder pair compliant with the MPEG-G specifications (https://github.com/mitogen/genie/tree/develop). GENIE is now focused on compression (up to 3-fold reduction in disk footprint with respect to the current *de facto* standard *gzip*), but also supports development of efficient data transfer and APIs for operating directly on the compressed data.

Results. GENIE is a package constructed from several codes that we integrated into a single application [2–5, 7–9, 11], implementing parallelization using the same OpenMP paradigm across all modules. The unaligned input data are split into streams to separate out read IDs and sequences from the quality scores. These streams are directed into SPRING and CALQ, respectively, for initial processing and conversion into descriptor streams that are maximally compressible by GABAC. GABAC is our rendition of the popular CABAC (Context-Adaptive Binary Arithmetic Code [6]) that is specifically designed for genomic sequences.

Given an input stream, the compression process of GABAC consists of a five stage pipeline: (1) input parsing, (2) (optional) 3-step transformation, (3) symbol binarization, (4) context selection, and (5) CABAC. First the input descriptor stream is parsed into a stream of symbols. These symbols are processed by the 3-step transformation stage that converts the symbol stream into transformed sub-streams. For each transformed sub-stream, a binarization algorithm converts each symbol into a bit string. It is chosen together with a context selection algorithm. Finally, each bit of the binarization is combined with a context and both are processed using CABAC. GENIE packages the compressed data into a single output file in a format that follows the MPEG-G specification.

A performance comparison was performed across several codecs, including GABAC, gzip, bzip2, xz, rANS order-0 or rANS order-1. Each codec was run on human whole genome sequencing chromosome 11 data from items 01 and 02 of the MPEG-G Genomic Information Database (https://mpeg-g.org). The corresponding BAM files are 6.9 GB and 4.2 GB in size, respectively. To bring all codecs to the same denominator and make them comparable for this analysis, we modified the compression tools CRAM and DeeZ to enable access to their internal data representations. These data were used as descriptor streams, each encoded with the entropy codecs used in CRAM (gzip, bzip2, xz, rANS order-0 or rANS order-1), plus GABAC. This resulted in a test set of 129 descriptor streams. To further emulate block-wise compression (random access capabilities), all streams were limited to 200 MiB. This approach allows for a more extensive test set in a random access environment, while preserving a reliable representation of the coding performance for each of the compared codecs. Measurements of compression ratio and speed on each descriptor stream were ranked, and the rank compared across the different input datasets. GABAC yields the best compression ratios on average, and is faster than gzip and xz in its optimal configuration. In its default configuration, freshly downloaded from the repository, Genie gives a consistent 6.5X compression ratio on Illumina FASTQ data (GIAB NA24694) in 1.5 hrs - 5.5 hrs (25X and 100X sequencing depth, respectively). The genomic sequence stream itself is compressible up to 40X on large datasets (>35X sequencing depth).

Conclusion. The GENIE framework delivers all the benefits of MPEG-G data standard and will create a step-change in the field of medical genomics by making genomic data storage 3-fold cheaper and (re-)analysis 2-fold faster. Data sharing and annotation, which is so important for research purposes, will be unburdened through built-in security mechanisms. GENIE supports the vital tradition of maintaining an open source ecosystem by fostering open data format: it is already based on the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. Likewise, MPEG-G is designed to continue the tradition of community-driven data infrastructure that has been established in bioinformatics via the widely used FASTQ and SAM formats.

CCS CONCEPTS

• **Applied computing → Computational genomics**.

KEYWORDS

genomics, petascale storage, data compression, individualized pedicine

ACM Reference Format:
B. Bliss, J. Allen, S. Baheti, M.A. Bockol, S. Chandak, J. Delgado, J. Fostier, J.L. Gelpi, S.N. Hart, M. Hernaez Arrazola, M. Hudson, M.T. Kalmbach, E.W. Klee, L.S. Mainzer, F. Müntefering, D. Naro, I. Ochoa-Alvarez, J. Ostermann, T. Paridaens, C.A. Ross, J. Voges, E.D. Wieben, M. Yang, T. Weissman, and M. Wiepert. 2018. Genie: an MPEG-G conformant software to compress genomic data.. In *Proceedings of Supercomputing '19: November 17âĂŞ22, 2019 (Supercomputing '19).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

ACKNOWLEDGMENTS

This work was a product of the Mayo Clinic and Illinois Strategic Alliance for Technology-Based Healthcare. Major funding was provided by the Mayo Clinic Center for Individualized Medicine and the Todd and Karen Wanek Program for Hypoplastic Left Heart Syndrome. We thank the Interdisciplinary Health Sciences Institute, UIUC Institute for Genomic Biology and the National Center for Supercomputing Applications for their generous support and access to resources. We particularly acknowledge the support of Keith Stewart, M.B., Ch.B., Mayo Clinic/Illinois Grand Challenge Sponsor and Director of the Mayo Clinic Center for Individualized Medicine. Special gratitude to Amy Weckle, Katherine Kendig and Gay Reed for managing the project. We are also grateful for the support of the Chan Zuckerberg Initiative DAF (2018-182798 and 2018-182799).

REFERENCES

[1] Claudio Alberti, Tom Paridaens, Jan Voges, Daniel Naro, Junaid J Ahmad, Massimo Ravasi, Daniele Renzi, Paolo Ribeca, Giorgio Zoia, Idoia Ochoa, et al. 2018. An introduction to MPEG-G, the new ISO standard for genomic information representation. *bioRxiv* (2018), 426353.

[2] Shubham Chandak, Kedar Tatwawadi, Idoia Ochoa, Mikel Hernaez, and Tsachy Weissman. 2018. SPRING: A next-generation compressor for FASTQ data. https://github.com/shubhamchandak94/Spring. [Online; accessed 2018].

[3] Mikel Hernaez, Idoia Ochoa, and Tsachy Weissman. 2016. A cluster-based approach to compression of quality scores. In *Proceedings. Data Compression Conference*, Vol. 2016. NIH Public Access, 261.

[4] Reggy Long, Mikel Hernaez, Idoia Ochoa, and Tsachy Weissman. 2017. GeneComp, a new reference-based compressor for SAM files. In *Data Compression Conference (DCC), 2017*. IEEE, 330–339.

[5] Greg Malysa, Mikel Hernaez, Idoia Ochoa, Milind Rao, Karthik Ganesan, and Tsachy Weissman. 2015. QVZ: lossy compression of quality values. *Bioinformatics* (2015), btv330.

[6] Detlev Marpe, Heiko Schwarz, Gabi Blattermann, Guido Heising, and Thomas Wiegand. 2002. Context-based adaptive binary arithmetic coding in JVT/H. 26L. In *Proceedings. International Conference on Image Processing*, Vol. 2. IEEE, II–II.

[7] Idoia Ochoa, Himanshu Asnani, Dinesh Bharadia, Mainak Chowdhury, Tsachy Weissman, and Golan Yona. 2013. QualComp: a new lossy compressor for quality scores based on rate distortion theory. *BMC bioinformatics* 14, 1 (2013), 1.

[8] Idoia Ochoa, Hongyi Li, Florian Baumgarte, Charles Hergenrother, Jan Voges, and Mikel Hernaez. 2019. AliCo: a new efficient representation for SAM files. In *2019 Data Compression Conference (DCC)*. IEEE, 93–102.

[9] Łukasz Roguski, Idoia Ochoa, Mikel Hernaez, and Sebastian Deorowicz. 2018. FaStore–a space-saving solution for raw sequencing data. *Bioinformatics* 1 (2018), 9.

[10] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. 2015. Big data: astronomical or genomical? *PLoS biology* 13, 7 (2015), e1002195.

[11] Jan Voges, Jörn Ostermann, and Mikel Hernaez. 2017. CALQ: compression of quality values of aligned sequencing data. *Bioinformatics* 34, 10 (2017), 1650–1658.