genomeweb

Will Bioinformatics Professionals Embrace MPEG-G Data Compression Standard?

Nov 13, 2019 | Neil Versel

CHICAGO – The genomic informatics world got a new standard for data compression over the summer when the International Organization for Standardization (ISO) approved MPEG-G.

That is the same MPEG — which stands for Motion Picture Expert Group — that brought the world the MP3 format for digital audio and MPEG-4 specification for compressing video files.

MPEG and ISO claim that the new standard addresses the issue of "efficient and cost-effective handling of genomic data by providing not only new compression and transport technologies, but also a family of standard specifications associating relevant information in the form of metadata." The collection also includes application programming interfaces to enable third parties to create interoperable applications and services for handling large genomic data files.

Among other things, MPEG-G promises to give individuals access to genomic data and other aspects of personalized medicine on mobile devices, according to Claudio Alberti, cofounder and CTO of genomic software analysis company GenomSys. Alberti was involved in the creation of MPEG-G when he was a project manager and software developer at the Swiss Federal Institute of Technology.

In 2014, Alberti was approached by bioinformatics professionals in Switzerland who wanted to know if the standards set by the MPEG group within ISO for multimedia data compression could be used to compress genomic data.

"They were struggling with existing formats. They couldn't do what they would like to do," which is compress, manipulate, and stream files, Alberti recalled.

"We spent two years of exploration trying to involve as many people as possible," Alberti said.

In 2016, ISO, upon request of MPEG and this working group of informaticians from universities, research centers, and technology companies from all over the world, approved the MPEG-G project. MPEG then issued an open call for technology, with a list of requirements collected from those who participated in the exploratory phase.

That led to a standard ISO "core experiment." For MPEG, this was the comparison of several technologies that seem to address the same requirements. These technologies eventually became the

MPEG-G standard.

Alberti is the lead author on a prepublication paper about MPEG-G listed on BioRxiv since September 2018.

The paper's authors note that MPEG-G represents the "largest coordinated international effort to specify a compressed data format that enables large scale genomic data processing, transport, and sharing." The authors claim that MPEG supports compression representing "more than 10x improvement over the BAM format."

At the same time, the authors said, this standard supports conversion to and from FASTQ, SAM, and BAM files to ensure full interoperability with current genomic data pipelines.

MPEG-G compresses raw data into blocks that can be indexed. These blocks are then encoded into the MPEG file format, which allows people to find the information they need and then extract blocks without having to decompress the entire file.

MPEG standardizes interfaces for applications for querying data.

"If you are interested in some specific patterns included in the reads, you can build clusters of reads around these patterns," Alberti said. "Then you compress them into separate blocks."

MPEG-G also can serve as an archival format because of the querying capability. It does not rely on a single input format, while BAM and CRAM require SAM files as input. "SAM is a textual representation with constraints," Alberti said.

SAM has 11 columns, containing alignment information on sequence reads, while an optional 12th column the optional one is for auxiliary information, including metadata and user-defined fields. Anything user-defined is a potential barrier to interoperability, according to Alberti.

But not everyone in the bioinformatics community is happy about the introduction of MPEG-G.

For its part, the Global Alliance for Genomics and Health (GA4GH) last month introduced five new standards as part of its GA4GH Connect five-year strategic plan, meant to address issues in data security, cloud computing, phenotype and variant data exchange, and the ethical implications of personal data use. Those are some of the same areas MPEG-G is focusing on.

At the Intelligent Systems for Molecular Biology and European Conference on Computational Biology (ISMB/ECCB) conference in Basel, Switzerland in July, several attendees of a session on data compression standards questioned why MPEG-G was even necessary.

Those against MPEG-G generally support file specifications adopted by GA4GH.

"It's very silly because the genomics community have already adopted BAM and FASTQ, for example, as standard formats, and these are GA4GH standards," said Dan Greenfield, cofounder and CEO of PetaGene, a British company that builds compression technology meant to enhance GA4GH-endorsed file formats. "The last thing the genomics community needs is yet another genomic standard."

Greenfield contended that competition risks fragmentation, undermining the point of standards. "There may be some justification for competing standards bodies if there's a real practical benefit, but we don't see one in this case," he said.

"We take great pains to make sure that our compression encryption works transparently when we're mixing tools and pipelines which see ordinary BAM and FASTQ files. We don't believe there should be a new file format standard for genomic data," Greenfield added.

ISO is a standards body for many industries, but Greenfield noted that some industries have their own standards bodies, such as the Internet Engineering Task Force for internet protocols. "GA4GH is the standards body for genomics," he said. He dismissed the notion of an "outside organization" like ISO coming into bioinformatics.

Further, according to Greenfield, MPEG might eventually require licensing fees or royalty payments. "I don't think the genomics community would be happy for [a licensing authority] to oversee the use of our genomic data," Greenfield said.

Proponents of MPEG-G said that GA4GH is not an official standards body in the sense that ISO is, and that CRAM, BAM, and derivatives are more file formats than properly developed standards.

Greenfield said that GA4GH is in fact a formal standards body because each specification goes through approval and improvement processes. "Just because MPEG-G don't want to recognize them doesn't mean that the genomics community at large doesn't recognize them," he said.

However, Alberti said that there is a difference between a standard and a widely used implementation, and just because something is widely used doesn't make it a standard. MPEG is a technical specification of something that people can implement in different ways, according to different use cases.

While BAM and CRAM are compressors of SAM files, MPEG-G is more than a data compressor. "MPEG-G is a framework," Alberti said. This multipart standard offers "specification of interfaces and algorithms to systems to interoperate through standard interfaces," he said.

For example, part three of the MPEG-G specification standardizes how non-MPEG-compliant systems can interact with MPEG files. "Even if you are not an entirely MPEG-compliant system, you can still query through a standard interface MPEG data and you can expect the output of the query in a standardized format," he said.

There is a standard application programming interface. Users of this standard API can expect a consistent output with each query, according to Alberti.

An MPEG-G file is like a database in that it provides data aggregation and segmentation features that BAM lacks. He said that MPEG has a different indexing mechanism for compressing and organizing data.

"You can classify your data according to other features that genomic data can present," to allow people to query compressed data by feature, not just by genomic interval, Alberti said. "This is about indexing."

James Bonfield, principal software developer at the Wellcome Sanger Institute and author of the CRAM file format's implementation in the C programming language, cochaired the session on data compression at ISMB/ECCB. He also is a longtime GA4GH contributor.

Bonfield declined to comment on MPEG-G, but said that GA4GH has a standard mantra for all its activities, which is to collaborate on the interfaces and compete on the implementations.

"We're quite happy with competitive work, but we ideally want to be trying to collaborate on what I call interfaces, which in this case means file formats," Bonfeld said. "And the idea there clearly is to try and reduce the number of different file formats because if we can all agree and collaborate on a single format, that's great."

He said a new format would be justified only if it produces significant improvements over previous formats. "I haven't actually seen much in the way of justifications yet," Bonfield said.

Bonfield noted that there is always a cost in changing formats, in terms of money, time, and even culture. "You have to do that cost benefit analysis to determine how much am I saving by changing in terms of disk space," he said.

He added that it is "helpful" to have a formal standards process to go through and said that GA4GH was not very formal in its early days. But Bonfield now believes that the alliance is now a formal standards body with clear processes for proposing or changing standards, as well as rigorous review processes.

Sanger designed the SAM and BAM formats and the European Molecular Biology Laboratory's European Bioinformatics Institute created the CRAM format. The Broad Institute is deeply involved in BAM development as well.

"This is where GA4GH fits in, to try to make sure that the formats are not owned by any one institute, but are governed by a committee with various interested parties from all over the world, both academic and commercial," Bonfield said. "They weren't developed by GA4GH, but GA4GH is essentially now the maintainer in governance of the formats."

Alberti, the MPEG-G booster, said that FASTQ is widely used only because "nobody has ever taken the time to find anything better." BAM and CRAM are specific implementations for specific functionalities, according to Alberti.

But it might be possible for the two camps to coexist peacefully.

Ewan Birney, director of GA4GH, director of EMBL-EBI, and nonexecutive director of Genomics England, said that GA4GH actually is working closely with ISO Technical Committee 215 — the one tasked with improving interoperability of health information — and another standards body, Health Level Seven International. "I do think that BAM, CRAM, and VCF will be three of the more community-accepted standards which will go through this extra level of commercial or industry-level standardization," he said.

He said that GA4GH has been discussing standards harmonization with HL7 and ISO for about two years. HL7 is more focused on electronic health records and other clinical information, while GA4GH concentrates on genomics. The groups are trying to improve interoperability at the nexus of phenotype and genotype data, according to Birney.

Filed Under Informatics Editor's Pick Europe North America technology development GA4GH data sharing genomics