**ISO/IEC JTC 1/SC 29/AG 03 N**Graphical user interface, text

Description automatically generated**0112**

**ISO/IEC JTC 1/SC 29/AG 03  
MPEG Liaison and Communication   
Convenorship: KATS (Korea, Republic of)**

**Document type:** Output Document

**Title:** White Paper on Coding of Genomic Annotations

**Status:** Approved

**Date of document:** 2023-04-28

**Source:** ISO/IEC JTC 1/SC 29/AG 03

**No. of pages:** 9 (with cover page)

**Email of Convenor:** kyuheonkim@khu.ac.kr

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29ag3>

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/** **AG 03 MPEG Liaison and Communication**

**ISO/IEC JTC 1/SC 29/AG 03 N0112**

**April 2023, Antalya**

|  |  |
| --- | --- |
| **Title** | **White Paper on Coding of Genomic Annotations** |
| **Authors** | **Marco Mattavelli, Yeremia Gunawan Adhisantoso, Patrick Y.H. Cheung, Ünsal Öztürk, Raymond Krasinski, Jaime Delgado, Simone Casale-Brunet, Paolo Ribeca, Mikel Hernaez, Itaru Kaneko, Jan Voges** |
| **Source** | **AG 03, MPEG Liaison and Communication** |
| **Status** | **Approved** |
| **Serial Number** | **22805** |

White Paper: Coding of Genomic Annotations

1. Introduction
   1. Overview and History

The introduction of high-throughput DNA sequencing has led to the generation of large quantities of genomic sequencing data that must be stored, transferred and analyzed. The ISO/IEC 23092 family of standards, Part 1 to 5, have addressed the problem of an efficient representation, compression and transport of genome sequencing data. Once the sequencing data is available, an important usage of the data is the association of the data with the results of the analysis that are generated by genomic processing pipelines and by the information added by analysts. Analysis results and additional information are referred here as “genomic annotations”. The newest ISO/IEC 23092 standard, Part 6, addresses the need of providing compressed representations of genomic annotations linked to the compressed representation of raw sequencing data and metadata.

By doing this ISO/IEC 23092, Part 6 is extending the MPEG Genomics standard so as to incorporate not only the primary (raw sequencing data) and secondary (aligned sequencing data), but also tertiary genomic data, including variant calls, gene expressions, mapping statistics. contact matrices (e.g. Hi-C), genomic tracks information and functional annotations, which are collectively called Annotation Data in the ISO/IEC 23092 standard, with efficient compression, indexing and searching capabilities. The extended format also includes advanced features including selective encryption and signing of the data, auditing support, data provenance information, traceability and support for direct linkage to external clinical data repositories expressed in common standard formats.

* 1. Introduction to Genomic Annotations

Primary data analysis (i.e., the analysis of the data derived from sequencing), performed at the level of the read, is usually followed by secondary analysis stages, which are performed at the level of the genomic interval. The table below shows some of the typical steps of a genomic variant-calling analysis workflow:

|  |  |  |
| --- | --- | --- |
|  | Step | Description |
| 1 | Sequence reads extraction | The process of extraction of fragments of DNA/RNA in the form of sequences of nucleotides from a biological sample. Sequences of nucleotides are commonly referred to as “reads”. |
| 2 | Mapping and Alignment | Sequence alignment refers to the process of arranging sequence reads by finding regions of similarity that may be a consequence of functional, structural, or evolutionary relationships among the sequences. When the alignment is performed with reference to an existing DNA sequence the process is called “mapping”. |
| 3 | Variant detection | Variant detection (a.k.a. variant calling) is the process of translating the output of DNA sequencing machines, (reads mentioned in step 1 and aligned in step 2), to a summary of the unique characteristics of the organism being sequenced. These characteristics are called “variants” because they are expressed as differences between the sequenced genome and a reference genome. |
| 4 | Variant annotation | Variant annotation is the process of assigning functional information to the DNA variants identified in step 3. This implies the classification of variants according to their relationship to coding sequences in the genome and according to their impact on the coding sequence and the gene product. |
| 5 | Functional & Structural Analysis | Analysis of DNA (variants, CNV = copy number variation, methylation etc,) strands to define their relationship with genes (and proteins) functions and structure. |

Table 1 – The main stages of a typical genomic variant calling pipeline

While the representation of the data and metadata covering the steps 1 and 2 are fully supported by the published Parts 1 to 5 of the ISO/IEC 23092 standard, the new part 6 provides an extension for the support of Annotation Data thus covering also the subsequent steps 3 to 5. Figure 1 depicts a functional diagram of the typical genomic information life cycle expressed as different processing steps and the intermediate file formats.



Figure 1 – Typical structure of a genomic variant-calling pipeline from sequencing to analysis.

The following data types are supported by ISO/IEC 23092 Part 6:

|  |  |  |
| --- | --- | --- |
| Data type | Description | Indicative file format(s) |
| Mapping statistics | Mapping statistics may be pre-computed with custom defined granularity, for instance at the level of thousands of nucleotides or larger. Summary statistics can be computed for sub intervals, at different scales as needed. | There is no indicative file format, but for a description of relevant statistics refer to ISO/IEC 23092-3 |
| Quantitative browser tracks | Mappings between nucleotides on the genome and numerical values with formats such as bigWig containing snapshots at different scales for ease of visualization. Can be used to visualize read coverage (and hence ChIP-seq or methylation/epigenetics experiments, variant frequency, etc.). | wig, bigWig, bedGraph |
| Variants | Genomic variants (genomic or transcriptomic) can be simple (SNPs or short indels) or complex (large-scale rearrangements). | VCF |
| Genome functional annotation | Functional annotations can include localization of gene models on the genome, such as lists of UTRs, exons and coding intervals, and other biological features such as repeats, ontology annotations, gene names. In the case of ChIP-seq experiments, features could be a list of called peaks. | BED, GTF, GFF(3), GenBank |
| Expression values | Expression values from NGS describes the number of reads, or related metrics (FPKM, etc.) associated to a list of genes or transcripts. Additional metadata for the sample(s) might also be present. | Non-sparse and sparse matrix formats (e.g. MatrixMarket) |
| Chromosome conformation capture (3C)  data | 3C data expresses information such as presence/absence of contact, or intensity/frequency of contact between a position in the genome and another position, thus specifying a matrix of values. As with other data, different resolutions are possible. Depending on the experiments, matrices could be sparse. | This information is typically coded in .hic files. |

Table 2 - The main stages of a typical genomic variant calling pipeline

1. MPEG Genomics Part 6: Support for Genomic Annotations
   1. Compression of Genomic Annotations

In ISO/IEC 23092-6, genomic annotation data fields are encoded either as descriptors or attributes. Descriptors refer to data fields that are known a priori and thus come with a pre-defined coding process; whereas attributes refer to ad hoc data fields whose coding processes can be individually determined. To allow for such flexibility, the definition of each attribute and its encoding process, consisting of a sequence of transformation and compression steps, are specified in attribute and compressor parameter sets stored as part of the data file. By parsing these parameter sets, a decoder can obtain the instructions for the proper decoding of each attribute. This attribute coding mechanism is particularly useful for handling genomic data that are volatile and rapidly evolving.

For certain descriptors that represent a huge portion of the data, specific coding technologies are used to exploit the properties of the data and maximize the compression ratio. For example, the genotype and the likelihood account for more than 80% of the Variants data and contact matrix accounts for more than 95% of the 3C experiments data.

Variants data is represented as a tabular structure where each row represents the variant at a specific genomic position and the columns represent a sample in an experiment. Genome in a population tends to be similar due to meiotic recombination. Additionally, the variants at different genomic positions tend to correlate that can be described as Linkage Disequilibrium. To exploit such properties in both Genotype and Likelihood, the codec uses a series of transformations consisting of row-column sorting and binarization. The sorting allows similar data to be near each other, increasing the efficiency of run-length encoding and context-based encoding. The number of alternate alleles varies from one position to another.

The contact matrix is a two-dimensional array, representing all interactions between all chromosomes, therefore the matrix is symmetrical. Each row and column of the contact matrix corresponds to a region in a certain chromosome. Each value in the contact matrix represents the number of contacts between a pair of regions or loci. Many of the values are zero, therefore the contact matrix is sparse.

Contact matrix is further split and classified into two categories based on the chromosome pair: intra-chromosomal for contact between a chromosome with itself and inter-chromosomal for contact between two different chromosomes (see Figure 2). Intra-chromosomal contact matrix is diagonal-dominant, symmetrical, and sparse. Furthermore, the diagonal entries are magnitudes greater than the entries in both upper and lower triangle of the matrix.

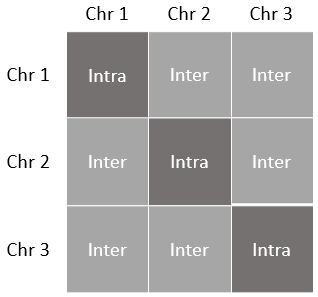


Figure 2: Intra- and Inter-Chromosonal Regions in a Conact Matrix

ISO/IEC 23092, in the latest edition of ISO/IEC 23092-2, allows the compression of different descriptor types with different entropy coding algorithms. Currently, the standard employs five different entropy coding modes, namely CABAC, LZMA, ZSTD, BSC, and Procrustes.

CABAC (Context-Adaptive Binary Arithmetic Coding) is a codec already widely used in previous AVC and HEVC MPEG standard in the field of video coding, is in general the most efficient and is suitable for achieving high compression rates at the expense of relatively high processing costs.

LZMA (Lempel-Ziv Markov Chain Algorithm) makes use of probability-based range encoding and a dictionary matching and compression scheme similar to that of LZ77 based on Markov chain-based dictionary searches.

ZSTD is based on a LZ77-like dictionary matching stage with a large search window, entropy encoding through tANS, a table-based version of ANS and Huffman coding.

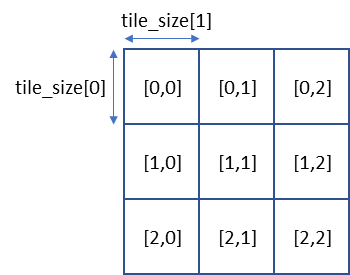
BSC (Block-Sorting Compression), based on the Burrows-Wheeler transform (BWT), reversibly rearranges data based on its suffix array, in a way such that the transformed data contains runs of repeated symbols.

Procrustes is a self-indexing codec that provides at the same time compression and fast exact string matching capabilities in the compressed domain.

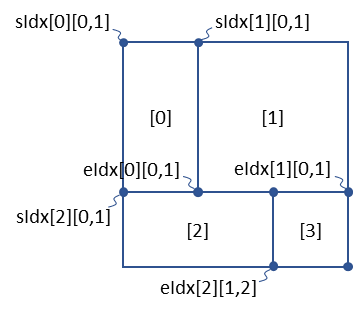
Each of the different entropy coding algorithm can be selected to compress and index the “descriptors” or “attributes” of annotation data.

* 1. Indexing of Genomic Annotations

To facilitate random access, descriptor/attribute data in an annotation table is split into tiles (ranges of rows and columns), each being compressed individually. Depending on the characteristics of the data, tiles can be of a uniform or variable size. For uniform-size tiles, only the fixed numbers of rows and columns per tile are specified; whereas for variable-size tiles, the specific ranges of row and column indexes associated with each tile are specified. Figure 3 shows examples of uniform and variable tile sizes. To randomly access data in selected rows and columns of an annotation table, the tiles overlapping with the selected region can be identified through the tile configuration and decoded.



(a)



(b)

**Figure 3 – Examples of (a) uniform and (b) variable tile sizes. The number(s) at the center of each tile is the tile index(es). For uniform tile size, *tile\_size*[*i*] specifies the number of rows (*i = 0*) and columns (*i = 1*) per tile. For variable tile size, *sIdx*[*i*][*j*] and *eIdx*[*i*][*j*] specify respectively the starting and ending indexes of the rows (*i = 0*) and columns (*i = 1*) defining the boundary of each tile.**

If the annotation data is sorted by genomic coordinates in one or two dimensions, the genomic coordinates defining the boundary of each tile can be specified in a similar fashion as the row and column indexes.. This extension supports random access and searches, such as substring and numeric range search, on descriptor and attribute values by storing pre-built and optionally compressed indexing data on selected descriptors and attributes as needed.

* 1. Metadata

While ISO/IEC 23092, Part 6 extends the MPEG Genomics standard to incorporate tertiary genomic data, this is not sufficient to fully support typical annotation data use cases. Rich metadata is also needed to fully make use of the data in context, from information of the originating source of the data to the detailed pipeline specification used to produce the data as well as how the data was organized and changed over time. For this reason, ISO/IEC 23092, Part 3, which defines metadata for sequencing data, has been extended to support annotation data.

Metadata providing information on the origin of the data, including details of the lab workflow used to produce the underlying data, has been added to the standard in order to record the provenance of the originating study or experiment. This provenance metadata is necessary to support proper analysis and organization of the data,

Analytics metadata containing detailed specifications of the software pipelines for generating the data is necessary to allow for verification of data reproducibility by re-running the analysis using exactly the same input data, computational environment, software and pipeline settings.

Annotation linkage metadata is used for specifying relationships that exist between annotation data and other data, stored either within or outside the current file archive, to facilitate cross-referencing capabilities for purposes such as data exploration, navigation, visualization and complex queries.

To support the exploration of different analytical hypotheses, formatting metadata extensions have been included. Formatting supports filtering and sorting of the underlying data and the ability to store this formatting with the annotation data allows for the formatting rules to be shared among users.

Protection metadata has been extended to support selective encryption, digital signature and access control to ensure data security, privacy and integrity in the part of the annotation data that is considered sensitive.

Finally, history metadata that records all data access and modifications has been added to support data tracking, traceability and, ultimately, the ability to audit data usage.

1. Benefits of ISO/IEC 23092, Part 6

ISO/IEC 23092 Part 6, by means of the extensions to Annotation data, increases the applications supported by the MPEG-G standard series. By including Annotation Data within the ISO/IEC 23092 unified format, advanced visualization and analysis applications can be supported by a structured, compressed and indexed file format representation. This includes the ability to perform random access on the compressed data, access by indexed information, and perform complex searches on the compressed data, as well as navigate across the various data modalities.

ISO/IEC 23092 Part 6 supports five encoding mode, with each specific sets of encoder-specific encoding parameters, which can be utilized to handle different use cases and scenarios in optimal manners. ISO/IEC 23092 Part 6 also allows for the use of different coding modes for different descriptors, which in turn lets users choose which types of data are to be self-indexed with the coding mode described in this section, and which types of data to be coded with other coding modes that are more specialized in compression-decompression rates or speeds, or are closer to the Pareto frontier in the space of rates and speeds.

ISO/IEC 23092 provides support for efficient exact string matching of appropriate data types in the compressed domain at descriptor level. The “Procrustes” encoding mode defined in ISO/IEC 23092 arranges the data into compressed data structures indexing. It is based on an efficient FM-index implementation associated to a binary wavelet tree decomposition. The bitvectors at each node are further compressed using succinct data structures supporting efficient character rank queries. Strings can be matched without the decompression of the payload. Assuming a constant alphabet size, the algorithm can match queries in linear time in the length of the query.

The coding mode and the API used, referred to and defined in ISO/IEC 23092-1, 23092-2, and 23092-6 creates several venues for optimizations and tradeoffs in compression rates and times, query efficiency and complexity, and decompression times. Descriptors or input data may be compressed independently. Such payloads need to be independently queried, resulting in the overall query matching time complexity to have a factor of the number of independent payloads. Since independently compressed payloads do not share common data structures related to the coding mode, querying, compression, and decompression may be parallelized on a payload-level, which introduces yet another dimension of optimization and tradeoffs. Having smaller payloads in the presence of sufficiently abundant computational resources allows higher degrees of parallelization of encoding-decoding and querying, hence decreasing compression-decompression times while maintaining hardware-related time bounds on query times. Another advantage provided by smaller payload sizes is increased spatial locality: the traversal that takes place during a query does not necessarily exhibit spatial locality, and the algorithm may traverse anywhere in the payload for the next step in the compression; however, it is guaranteed to stay within the payload.

1. Conclusion/What’s next?

The development of ISO/IEC 23092 Part 6 Coding of genomic Annotations is a first attempt to solve the main drawbacks introduced by the current practices of representing, handling and processing genomic annotations in a variety of loosely specified and structured textual formats disconnected from the sequencing data they have been generated or they are linked to. The main novelty is that genomic annotations can be part of a stand-alone file or included, and linked, to the genome sequencing data they are associated to, in the same standard ISO/IEC 23092 file. Both type of genomic data information are compressed. ISO/IEC 23092 provides a unified compressed representation of a variety of annotation file formats by employing a combination of state of the art and newly developed compression technologies, avoiding current file syntax and semantic ambiguities, thus simplifying the tasks of the users that can rely on optimized compression approaches for the various type of data and use cases. State of the art and newly developed indexing technologies are also natively supported for both string and range value type of searches. String searches are very efficient and available directly in the compressed domain. In addition, when sequencing data and annotation data are compressed in the same file, the results of string and range searches of annotation data, can provide a direct selective access to the sequencing data they are related to.

While the publication of this new Part of the ISO/IEC 23092 standard series constitutes an important achievement and step forward in the performance, and functionality, of genomic annotation data processing, it is not the conclusion of the work. In the best MPEG tradition, limitations and potential new functionality are currently under study and, if effective, they will be considered for future extensions and developments of the standard family.