DRAFT INTERNATIONAL STANDARD

ISO/IEC DIS 23092-2

ISO/IEC JTC **1**/SC **29** Secretariat: **JISC**

Voting begins on:
2020-01-17

Voting terminates on:
2020-04-10

Information technology — Genomic information representation —

Part 2:
Coding of genomic information

Technologies de l'information — Représentation des informations génomiques —

Partie 2: Codage des informations génomiques

ICS: 35.040.99

This document is circulated as received from the committee secretariat.

Reference number
ISO/IEC DIS 23092-2:2020(E)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this specification and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This specification was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this specification may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see http://patents.iec.ch).

Any trade name used in this specification is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html.

This specification was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

A list of all parts in the ISO/IEC 23092 series can be found on the ISO website.

Any feedback or questions on this specification should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

The advent of high-throughput sequencing (HTS) technologies has the potential to boost the adoption of genomic information in everyday practice, ranging from biological research to personalized genomic medicine in the clinic. As a consequence, an extraordinary growth in volume of generated data has been recorded during the last few years, and an even more pronounced growth is expected in the near future.

At the moment genomic information is mostly exchanged through a variety of data formats, such as FASTA/FASTQ for unaligned sequencing reads and SAM/BAM/CRAM for aligned reads. With respect to such formats, the ISO/IEC 23092 series provides a new solution for the representation and compression of genome sequencing information by:

- specifying an abstract representation of the sequencing data rather than a specific format with its direct implementation
- being designed at a time point when technologies and use cases are more mature. This permits to address one limitation of the textual SAM format, for which incremental ad-hoc addition of features followed along the years, resulting in an overall redundant and suboptimal format which at the same time results not general and unnecessarily complicated
- normatively separating free-field user-defined information with no clear semantics from the normative genomic data representation. This allows a fully interoperable and automatic exchange of information between different data producers
- allowing multiplexing of relevant meta-data information with the data since data and meta-data are partitioned at different conceptual levels
- following a strict and supervised development process which has proven successful in the last 30 years in the domain of digital media for the transport format, the file format, the compressed representation and the application program interfaces

The ISO/IEC 23092 series provides the enabling technology that will allow the community to create an ecosystem of novel, interoperable, solutions in the field of genomic information processing. In particular it offers:

- Consistent, general and properly designed format definitions and data structures to store sequencing and alignment information. A robust framework which can be used as a foundation to implement different compression algorithms
- Speed and flexibility in the selective access to coded data, by means of newly designed data clustering and optimized storage methodologies
- Low latency in data transmission and consequent fast availability at remote locations, based on transmission protocols inspired by real-time application domains
- Built-in privacy and protection of sensitive information, thanks to a flexible framework which allows customizable secured access at all layers of the data hierarchy
- Reliability of the technology and interoperability among tools and systems, owing to the provision of a normative procedure to assess conformance to the standard on an exhaustive dataset
- Support to the implementation of a complete ecosystem of compliant devices and applications, through the availability of a normative reference implementation covering the totality of the specification.

The fundamental structure of the ISO/IEC 23092 series data representation is the *genomic record*. The genomic record is a data structure consisting of either a single sequencing read, or a paired sequencing read, and its associated sequencing and alignment information; it may contain detailed mapping and alignment data, a single or paired read identifier (read name) and quality values.

Without breaking traditional approaches, the genomic record introduced in the ISO/IEC 23092 series provides a more compact, simpler and manageable data structure grouping all the information related to a single DNA template, from simple sequencing data to sophisticated alignment information.

Although it is an appropriate logic data structure for interaction and manipulation of coded information, the genomic record is not a suitable atomic data structure for compression. To achieve high compression ratios, it is necessary to group genomic records into clusters and to transform the information of the same

type into sets of descriptors structured into homogeneous blocks. Furthermore, when dealing with selective data access, the genomic record unit is too small to allow effective and fast information retrieval.

For these reasons, this document introduces the concept of access unit, which is the fundamental structure for coding and access to information in the compressed domain.

The access unit is the smallest data structure that can be decoded by a decoder compliant with ISO/IEC23092-2. An access unit is composed of one block for each descriptor used to represent the information of its genomic records; therefore, a block payload is the coded representation of all the data of the same type (i.e. a descriptor) in a cluster.

In addition to clusters of genomic records compressed into access units, reads are further classified in six data classes: five classes are defined according to the result of their alignment against one or more reference sequences; the sixth class contains either reads that could not be mapped or raw sequencing data. The classification of sequencing reads into classes enables to develop powerful selective data access. In fact access units inherit a specific data characterization (e.g. perfect matches in class P, substitutions in class M, indels in class I, half-mapped reads in class HM) from the genomic records composing them, and thus constitute a data structure capable of providing powerful filtering capability for the efficient support of many different use cases.

Access units are the fundamental, finest grain data structure in terms of content protection and in terms of metadata association. In other words each access unit can be protected individually and independently. **Figure 1** shows how access units, blocks and genomic records relate to each other in the ISO/IEC 23092 series data structure.

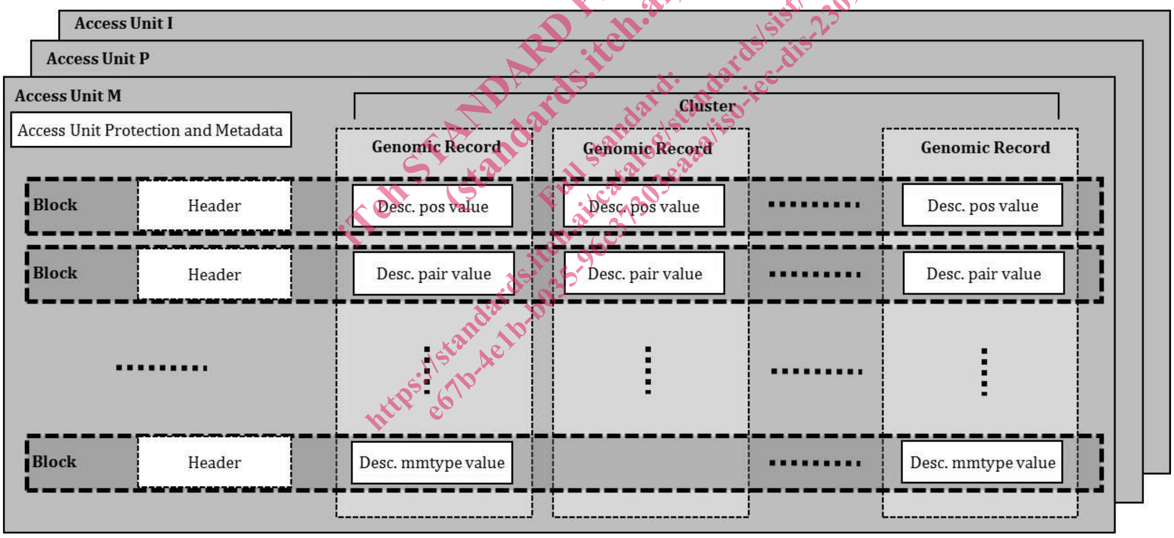

Figure 1 –Access units, blocks and genomic records

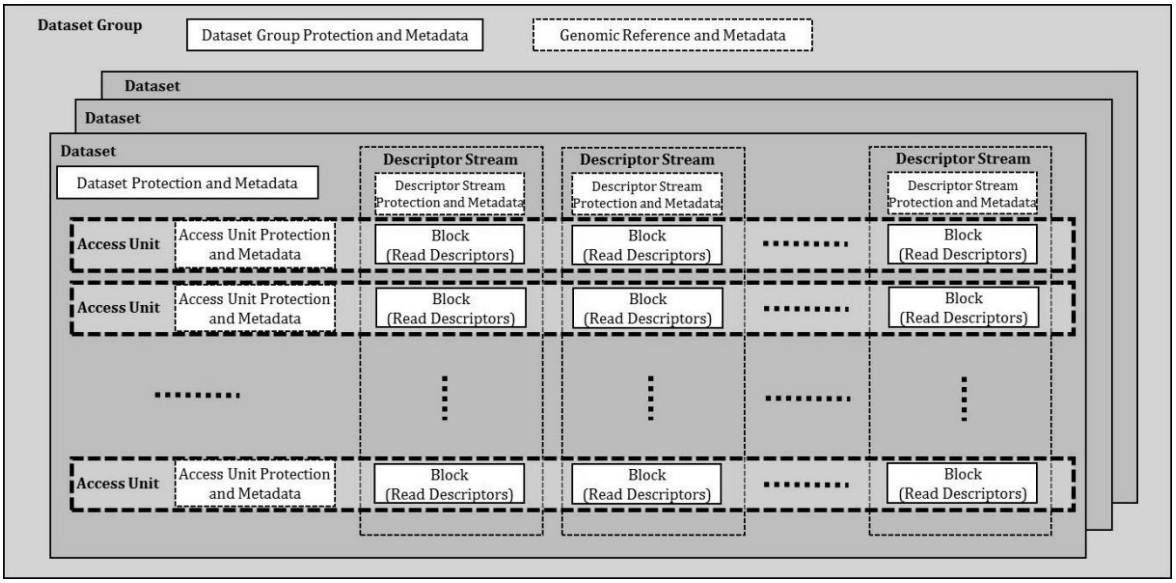

Figure 2 – The high-level data structure: datasets and dataset group

A dataset is a coded data structure containing headers and one or more access units. Typical datasets could for example contain the complete sequencing of an individual, or a portion of it. Other datasets could contain for example a reference genome or a subset of its chromosomes. Datasets are grouped in dataset groups, as shown in **Figure 2**.

According to the ISO/IEC 23092 series, the compressed sequencing data can be multiplexed into a normative bitstream suitable to packetization for real-time transport over typical network protocols. In storage use cases coded data can be encapsulated into a file format with the possibility to organize blocks per descriptor stream or per access unit, to further optimize the selective access performance to the type of data access required by the different application scenarios. The ISO/IEC 23092 series further provides a reference process to convert a normative transport stream into a normative file format and vice versa.

The ISO/IEC 23092-2 series define the syntax and semantics of the compressed genome sequencing data representation and the deterministic decoding process that reconstructs the contents of datasets. The decoding process is fully specified such that all decoders that conform to ISO/IEC 23092-2 will produce identical decoded output. A simplified diagram of the decoding process is shown in **Figure 3**.

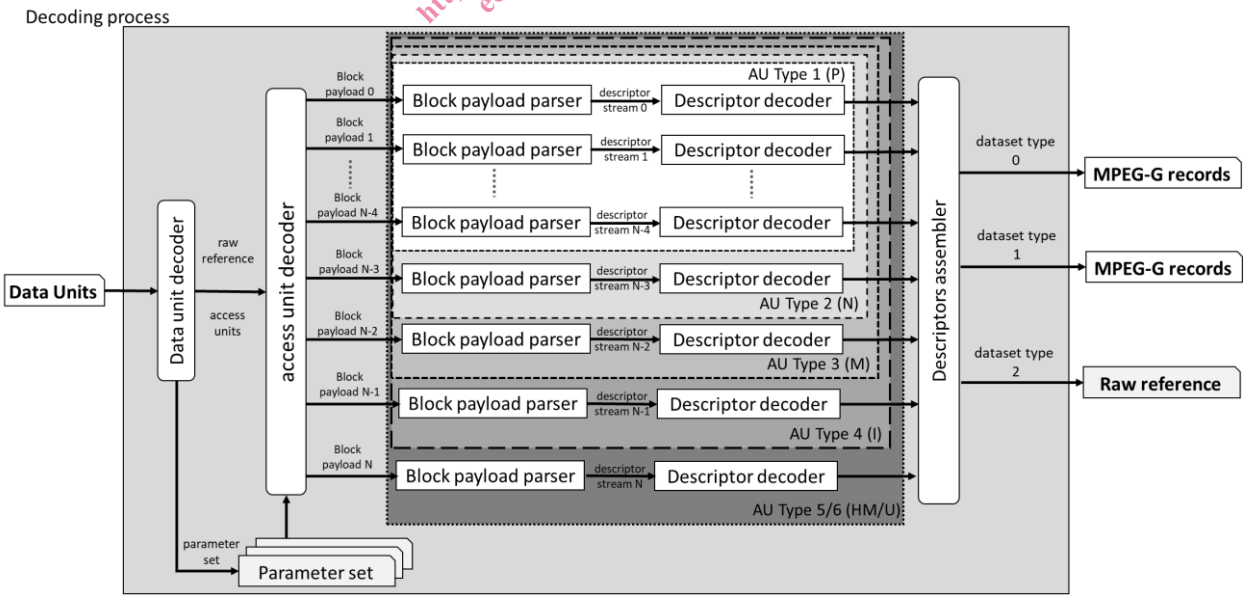

Figure 3 – The decoding process

Information Technology — ISO/IEC 23092 — Part 2: Coding of Genomic Information

1 Scope

This document provides specifications for the normative representation of the following types of genomic information:

- unaligned sequencing reads including read identifiers and quality values
- aligned sequencing reads including read identifiers and quality values
- reference sequences

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this specification. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 10646, *Information technology — Universal Coded Character Set (UCS)*

ISO/IEC 23092-1, *Information technology — Genomic information representation — Part 1: Transport and storage of genomic information*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 23092-1 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
alignment
information describing the similarity between a sequence [typically a *sequencing read* (3.28)] and a reference sequence (for instance, a reference genome)

Note 1 to entry: An alignment is described in terms of a position within the reference, the strand of the reference, and a set of edit operations (matches, mismatches, insertions and deletions, clipping of the sequence ends and splicing information) needed to turn the first sequence into the second.

3.2
CIGAR string
CIGAR
textual way of representing an *alignment* (3.1)

Note 1 to entry: Several definitions have been used by different programs; the one referred to here is the one used in the SAM format. It encodes a set of edit operations (matches, mismatches, insertions and deletions, clipping of the sequence ends and splicing information) needed to turn the sequencing read into the reference.

3.3
dataset
compression unit containing one or more of: reference sequences; *sequencing reads* (3.28); and *alignment* (3.1) information

Note 1 to entry: Datasets shall be as specified in ISO/IEC 23092-1.

3.4
deletion
contiguous removal of one or more bases from a genomic sequence

3.5
E-CIGAR
extended CIGAR syntax specified as a superset of the CIGAR syntax

Note 1 to entry: Among other things, E-CIGAR enables the unambiguous representation of substitutions, spliced reads and splice strandedness.

3.6
edit operation
modification of a sequence of *nucleotides* (3.20) by means of a substitution, *deletion* (3.4), *insertion* (3.18) or clip

3.7
FASTA
GIR that includes a name and a *nucleotide* (3.20) sequence for each *sequencing read* (3.28)

Note 1 to entry: Additional information is usually encoded in the read identifier by bioinformatics tools (such as database information, and base calling information).

3.8
FASTQ
GIR that includes *FASTA* (3.7) and *quality values* (3.22)

3.9
first end
end 1
read 1
first segment of a paired-end *template* (3.33)

Note 1 to entry: Illumina platforms usually store first and second ends in two separate files and in the same order — i.e. the n-th read of the first FASTQ file and the n-th read of the second FASTQ file belong to the same template.

3.10
genomic descriptor
descriptor
element of the syntax used to represent a feature of a genomic *sequencing read* (3.28) or associated information such as *alignment* (3.1) information or *quality values* (3.22)

3.11
genomic information representation
way to describe a sequence and some information associated with it

Note 1 to entry: Which information is represented varies depending on the GIR.

3.12
genomic record
record
data structure representing a *tuple* (3.34) optionally associated with *alignment* (3.1) information, *read identifier* (3.24) and *quality values* (3.22)

3.13
genomic record index
position of a genomic record in the sequence of *genomic records* (3.12) encoded in an access unit

3.14
genomic record position
0-based position of the leftmost mapped base on the reference genome of the first *alignment* (3.1) contained in a *genomic record* (3.12)

Note 1 to entry: A base present in the aligned read and not present in the reference sequence (insertion) and bases preserved by the alignment process but not mapped on the reference sequence (soft clips) do not have mapping positions.

3.15
genomic reference
reference
collection of reference sequences

Note 1 to entry: Typical examples are a reference genome or a reference transcriptome.

3.16
hard clip
base or set of bases originally present at either side of a read, and removed from it following *alignment* (3.1)

Note 1 to entry: The bases are no longer present in the sequence of the read.

3.17
indel
contiguous stretch of *nucleotides* (3.20) that, when aligning two sequences, are inserted into one sequence, or alternatively deleted from the other, in order to make the two sequences the same

Note 1 to entry: From "insertion or deletion".

3.18
insertion
contiguous addition of one or more bases into a genomic sequence

3.19
leftmost read end
leftmost read
sequencing read (3.28) generated by a paired-end sequencing run and mapped at a position on the reference sequence which is smaller than the mapping position of the other read in the pair

3.20
nucleotide
base
base pair
monomer of a nucleic acid polymer such as DNA or RNA

Note 1 to entry: Nucleotides are denoted as letters ('A' for adenine; 'C' for cytosine; 'G' for guanine; 'T' for thymine which only occurs in DNA; and 'U' for uracil which only occurs in RNA). The chemical formula for a specific DNA or RNA molecule is given by the sequence of its nucleotides, which can be represented as a string over the alphabet ('A', 'C', 'G', 'T') in the case of DNA, and a string over the alphabet ('A', 'C', 'G', 'U') in the case of RNA. Bases with unknown molecular composition are denoted with 'N'.

3.21
paired-end read
paired-end template
tuple (3.34) made of two segments

Note 1 to entry: Typically the segments correspond to the beginning and the end of the same nucleic acid molecule.

3.22
quality value
quality score
number assigned to each *nucleotide* (3.20) base call in automated sequencing processes

Note 1 to entry: Quality values express the base-call accuracy, i.e. the probability (or a related measure) for a nucleotide in the sequence to have been incorrectly determined.