

# ISO/IEC 23092-1:2025-01 (E)

## Information technology - Genomic information representation - Part 1: Transport and storage of genomic information

---

### Contents

Page

Foreword	v
Introduction	vii
<b>1 Scope</b>	<b>1</b>
<b>2 Normative references</b>	<b>1</b>
<b>3 Terms and definitions</b>	<b>1</b>
<b>4 Conventions</b>	<b>4</b>
4.1 Operators and functions	4
4.1.1 Arithmetic operators	4
4.1.2 Logical operators	4
4.1.3 Relational operators	4
4.1.4 Bitwise operators	4
4.1.5 Assignment operators	5
4.1.6 String/Character functions and operator	5
4.1.7 Data structure function and operator	5
4.1.8 Mathematical functions	5
4.1.9 Array operation functions	5
4.2 Syntax and semantics	6
4.2.1 Method of specifying syntax in tabular form	6
4.2.2 Bit ordering	6
4.2.3 Specification of syntax functions	6
4.2.4 Processes	7
<b>5 Structure of coded genomic data</b>	<b>7</b>
5.1 Genomic sequencing data record	7
5.2 Genomic annotation data records	8
5.3 Data classes	9
5.4 Access units	10
5.5 Datasets	10
5.6 Annotation data tile	11
5.7 Annotation tables	11
5.8 Annotation access units	11
5.9 Selective access	12
<b>6 Data format</b>	<b>12</b>
6.1 Format structure	12
6.1.1 General	12
6.1.2 Box order	17
6.2 Syntax for representation	18
6.3 Output data unit	19
6.4 Data structures common to file format and transport format	20
6.4.1 File header	20
6.4.2 Dataset group	20
6.4.3 Dataset	29
6.4.4 Access unit	40
6.4.5 Block	46
6.4.6 Annotation Table	47
6.4.7 Attribute Group	57
6.4.8 Annotation access unit	59
6.4.9 AAU block	63

6.5	Data structures specific to file format .....	64
6.5.1	General .....	64
6.5.2	Indexing .....	64
6.5.3	Descriptor stream .....	74
6.5.4	Offset .....	76
6.6	Data structures specific to transport format .....	77
6.6.1	General .....	77
6.6.2	Data streams .....	77
6.6.3	Dataset mapping table list .....	77
6.6.4	Dataset mapping table .....	78
6.6.5	Packet .....	80
6.7	Reference procedures to convert transport format to file format .....	81
6.7.1	Procedure for genomic sequencing data .....	81
6.7.2	Procedure for genomic annotation data .....	83
<b>7</b>	<b>String indexing technologies .....</b>	<b>87</b>
7.1	Master string index .....	87
7.1.1	General .....	87
7.1.2	Syntax .....	87
7.1.3	Master String Index Header .....	87
7.1.4	String index .....	88
7.1.5	Compressed string index .....	90
7.2	Decoding and querying processes .....	96
7.2.1	String index payload .....	96
7.2.2	Helper functions .....	97
7.2.3	Substring decoding process .....	98
7.2.4	Suffix array lookup process .....	99
7.2.5	Inverse suffix array process .....	99
7.2.6	Character decoding process .....	100
7.2.7	LF-mapping process .....	101
7.2.8	Extended LF-mapping process .....	101
7.2.9	Substring position search process .....	102
7.2.10	Searching for substring positions with the string index .....	103
7.2.11	Decoding a subset of the string index .....	104
7.2.12	Decoding all the strings of a specific annotation data tile .....	104
7.2.13	Retrieving whole strings with the string index .....	106
7.2.14	Retrieving data tile index(es) associated with a position and record indexes .....	107
<b>8</b>	<b>Indexing for numeric range searches .....</b>	<b>110</b>
8.1	B-Tree indexing .....	110
8.1.1	General .....	110
8.1.2	Syntax .....	110
8.1.3	Semantics .....	111
	<b>Annex A (informative) IETF RFC 3986 specification summary .....</b>	<b>112</b>
	<b>Annex B (informative) Selective access strategies for genomic sequencing data .....</b>	<b>113</b>
	<b>Annex C (informative) Selective access strategies for genomic annotation data .....</b>	<b>116</b>
	<b>Annex D (informative) Depacketization process .....</b>	<b>132</b>
	<b>Annex E (informative) Efficient handling of symmetric annotation data .....</b>	<b>135</b>
	<b>Bibliography .....</b>	<b>137</b>