**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC 1/SC 29/WG 11 N16146**

**San Diego, USA – February 2016**

|  |  |
| --- | --- |
| **Source:** | **Requirements**  |
| **Status:** | **Draft** |
| **Title:** | **Evaluation Procedure for the Draft Call for Proposals on Genomic Information Representation and Compression** |

[1 Introduction 1](#_Toc444284900)

[2 Encoded bitstreams 2](#_Toc444284901)

[3 Tools 2](#_Toc444284902)

[4 Evaluation Procedure 3](#_Toc444284903)

[4.1 Sequence reads extraction – compression of unmapped reads 3](#_Toc444284904)

[4.1.1 Coverage matrix 3](#_Toc444284905)

[4.2 Mapping and Alignment – Compression of mapped/aligned reads 4](#_Toc444284906)

[4.2.1 Coverage matrix 4](#_Toc444284907)

[4.3 Transport 4](#_Toc444284908)

[4.3.1 Coverage matrix 4](#_Toc444284909)

[5 Selection criteria 5](#_Toc444284910)

[6 References 5](#_Toc444284911)

[7 Annex A – Template for Self-Assessment of submissions 5](#_Toc444284912)

# Introduction

The Evaluation Procedure described in this document will be used to assess proposals in response to the MPEG Call for Proposals on Genomic Information Representation and Compression (WG11 document N16136). Some of the terms used in this document are defined in the “Terminology” section of the Call for Proposals.

This assessment will identify both the proposal which will become the General Model 0 (GM0) and other proposals from which a set of technologies will be possibly added to or integrated with the GM0. The addition and integration will be achieved through Core Experiments that will be specified by dedicated MPEG documents. A timetable and the procedure for the verification process and Core Experiments are specified in the Call for Proposals.

It should be noticed that the procedure described above implies that proposals will be accepted whether they fulfill (entirely or partially) all requirements but also in the case that they provide technology covering only a subset of the requirements.

This Evaluation Procedure will be used by:

• Proponents of Genomic Information Representation and Compression technology/solutions, both in answering to the Call for Proposals, and in self-assessing their proposal before submitting it;

• MPEG, to validate the self-assessment of the proponents and proceed to the evaluation of the proposed technologies.

The requirements defined in the Call for Proposals are divided into major needs (specified by a “shall” in the requirement text) and objectives (specified by a “should” in the requirement text). Major requirements shall be met by GM0 while objectives are less stringent even if highly desirable.

# Encoded bitstreams

Proponents are expected to supply encoded bitstreams obtained from a subset of the data listed in document N16145.

|  |  |
| --- | --- |
| **Input ID** | **Encoded bitstream name** |
| 01, 02, 03, 05, 07, 08, 09, 10,11,12 | MPEG\_HG\_ID\_mXXXXX.bit |
| 14, 15, 16, 18, 19 | MPEG\_BC\_ID\_mXXXXX.bit |
| 20 | MPEG\_MT\_ID\_mXXXXX.bit |
| 23 | MPEG\_CC\_ID\_mXXXXX.bit |
| 24,25 | MPEG\_PL\_ID\_mXXXXX.bit |
| 26 | MPEG\_SM\_ID\_mXXXXX.bit |

Table 1 - IDs of input data to be encoded and naming policy for encoded bitstreams

Notes:

1. FASTQ compression: In case the input item is composed by multiple files proponents are required to submit a single encoded bitstream.
2. mXXXXX is the registration number of the MPEG input contribution related to the bitstreams.
3. For input 01 only files ERR174324 to ERR174341 have to be considered (for a total of 36 FASTQ files).
4. For input 02 only file NA12878\_S1 has to be considered: ftp://ftp.sra.ebi.ac.uk/vol1/ERA172/ERA172924/bam/NA12878\_S1.bam
5. Input 14 shall be converted to FASTQ. The converted sample can be obtained by contacting Claudio Alberti: claudio.alberti@epfl.ch.

# Tools

The following tools will be used in order to assess the fulfillment of some of the requirements.

|  |  |
| --- | --- |
| **Tool** | **Purpose** |
| SFU Toolkit (to be published) | Verify shuffled decoded FASTQ, BAM |
|  |  |
|  |  |
|  |  |
|  |  |

# Evaluation Procedure

This section describes the test cases that will be performed to validate the fulfillment of the requirements listed in the Call for Proposals.

## Sequence reads extraction – compression of unmapped reads

### Coverage matrix

|  |  |  |  |
| --- | --- | --- | --- |
| **Test case ID** | **Input data (ref. doc N16145)** | **Test case steps** | **Reqs** |
| 1.1 | All bitstreams listed in Table 1 except those obtained from encoding input data with IDs 11,12,26 | 1. Decompress encoded bitstream
2. Verify that the original
* Reads headers
* Nucleotides sequences
* Quality scores

are present in the decoded bitstreams 1. Verify that pairing information is preserved
 | 1.1, 1.3, 1.4, 1.5, 1.8 |
| 1.2 | Bitstreams obtained from encoding input data with IDs 05, 11 | 1. Decompress encoded bitstream
2. Verify that each read has a unique identifier

Note: *if pairing information is contained in the original headers, it has to be preserved in order to meet requirement 1.3* | 1.2 |
| 1.3 | Bitstreams obtained from encoding input data with IDs 11, 12, 26 | Perform the evaluation procedure using Sensitivity, Precision and F-Score as described in document N16135In addition to the evaluation procedure described in N16135, application-independent rate distortion metrics may be taken into account. Proponents are invited to document how metrics are defined in their supporting documentation and provide any required supporting tool. | 1.6 |
| 1.4 | N/A | The proponents are required to document to what extent the solution meets the requirements | 1.7.1, 1.7.2 |

## Mapping and Alignment – Compression of mapped/aligned reads

### Coverage matrix

|  |  |  |  |
| --- | --- | --- | --- |
| **Test case ID** | **Input data (ref. doc N16145)** | **Test case steps** | **Reqs** |
| 2.1 | Bitstreams obtained from encoding input data with IDs 05, 11 | Bitstream decoding will be performed on an offline device  | 2.1 |
| 2.2 | Bitstream obtained from encoding input data with ID 05 | 1. Decode the compressed BAM
2. Verify that the decompressed data contain
	1. Headers
	2. Sequences
	3. Quality scores
	4. Pairing information
	5. Reverse complementarity information
 | 2.2 |
| 2.3 | Bitstream obtained from encoding input data with ID 05 | The proponents shall describe how the tool enables access to specific regions of the aligned data. | 2.3 |
| 2.4 | Bitstream obtained from encoding input data with ID 05 | The proponents shall describe how the tool enables any other mode of selective access to the aligned data. | 2.4 |
| 2.5 | Bitstream obtained from encoding input data with ID 05 | 1. Decompress encoded bitstream
2. Verify that each read has a unique identifier
3. Pairing information has to be preserved
 | 2.5 |
| 2.6 | Bitstream obtained from encoding input data with ID 05 | 1. Decompress encoded bitstream
2. Verify that the original information listed in requirements 2.6, 2.7, 2.8, 2.9 is preserved.
 | 2.6, 2.7, 2.8, 2.9 |

## Transport

### Coverage matrix

|  |  |  |  |
| --- | --- | --- | --- |
| **Test case ID** | **Input data (ref. doc N16145)** | **Test case steps** | **Reqs** |
| 3.1 | Bitstream obtained from encoding input data with ID 05 | The proponents are required to document how and to what extent the solution meets the requirements | 3.1 |
| 3.2 | Bitstream obtained from encoding input data with ID 05 | The proponents are required to document how and to what extent the solution meets the requirements | 3.2 |
| 3.3 | Bitstream obtained from encoding input data with ID 05 | The proponents are required to document how and to what extent the solution meets the requirements | 3.3 |
| 3.4 | Bitstream obtained from encoding input data with ID 05 | The proponents are required to document how and to what extent the solution meets the requirements | 3.4 |
| 3.5 | Bitstream obtained from encoding input data with ID 05 | The proponents are required to document how and to what extent the solution meets the requirements | 3.5 |

# Selection criteria

Criteria to rank technologies will be:

1. requirements coverage,
2. size of the produced bitstream overall and per class of data:
	1. Headers
	2. Sequences
	3. Quality Scores
	4. Any other metadata (e.g. auxiliary data)
3. compression and decompression computational complexity, ....
	1. processing time
	2. memory requirements

# Annex A – Template for Self-Assessment of submissions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test case** | **Input item** | **Requirement met Y/N** | **Result** | **Comments** |
| 1.1 | 01 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 02 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 03 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 05 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 07 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 08 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 09 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 10 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 14 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 15 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 16 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 18 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 19 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 20 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 23 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 24 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 25 |  | *Compressed file size with size of the 4 classes of data: headers, sequences, QS, metadata* |  |
| 1.2 | 05 |  | *All reads have a unique identifier* |  |
| 11 |  | *All reads have a unique identifier* |  |
| 1.3 | 11 |  | *Compressed file size, Precision, Sensitivity, F-score* |  |
| 12 |  | *Compressed file size, Precision, Sensitivity, F-score* |  |
| 26 |  | *Compressed file size, Precision, Sensitivity, F-score* |  |
| 1.4 | N/A |  | *Reference to technical description on how the requirements are met* |  |
|  |
| 2.1 | 05, 11 |  |  |  |
| 2.2 | 05 |  |  |  |
| 2.3 | 05 |  | *Access block size*  |  |
| 2.4 | 05 |  |  |  |
| 2.5 | 05 |  |  |  |
| 2.6 | 05 |  |  |  |
|  |
| 3.1 | 05, 11 |  |  |  |
| 3.2 | 05, 11 |  |  |  |
| 3.3 | 05, 11 |  |  |  |
| 3.4 | 05, 11 |  |  |  |
| 3.5 | 05, 11 |  |  |  |