ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AHG on Genomic Information Representation (m42285)

M. Golebiewsky (HITS), J. Delgado (UPC),
M. Mattavelli (EPFL)

Joint AHG with ISO/IEC TC276

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Mandates

1. To disseminate the information on the activities jointly carried out by ISO/IEC JTC 1/SC 29/WG 11 and ISO TC 276/WG 5 to other ISO TC and other organizations.
2. To make available on line the MPEG genome data to be used for core experiments (N16726).
3. To carry out the core experiments described in document N17324.
4. To contribute to the editing and to the revision of the CD of Part 1 and Part 2, Part 3 and for the WD of Part 4.
5. To finalize the collection and definition of test item descriptions and binary streams for Conformance testing
6. To identify work items for MPEG-G Version 2
7. To finalize the organization of the WS on MPEG-G and genomic data processing on the 18th of April in San Diego

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

AHG Activity

- Mainly focused around:
  - Editing the DIS text for Part 1 (Transport and Storage of Genomic Information) and DIS text of Part 2 (Compression of Genomic Information)
  - Running the active Core Experiment:
    - CE5 entropy coding
  - Preparing the study of CD for Part 3
- Organization of the WS on MPEG-G in San Diego
- Coordination activities with TC276 and other ISO committees:
  - No activity since Gwangju meeting.
- 18 Input documents at this meeting

Summary of CE 5 “Entropy Coding”

- 3 input document on CE 5 (M42645, M42646, M42550):
  - **Multiple alignments:** CE confirms that the current syntax and semantic of Part 2 is capable of supporting all cases found in the data set and potentially a wider set of cases.
  - **CABAC configuration:** a proposed set of “default” values for entropy coding of read names and sequence descriptors. Need of a similar table also for QV. TBD during the week.
  - **Current entropy coding specification:** validation of current entropy coding and possible alternatives and simplifications. Simplifications are considered for Part 2 editing.
- Conclusion/Consensus:
  - Close CE5

Other input documents review

- Most of inputs reviewed during the Sat. and Sun. AHG meeting
- Comments from NB suggesting changes to Part 2:
  - Review of the editorial and a few technical changes completed for about 90%, almost all NB comments approved for inclusion in the next version.
- Suggested changes to Part 1 for inclusion in the DIS text document:
  - Essentially removing inconsistencies after modifications asked by NBs.
- Part 3 is under ballot and some proposed changes have been already discussed:
  - The existing profile has been improved by using the extension mechanism to include the relevant EGA metadata.
  - Privacy rules specification has been revised so that it fully matches only operations by means of APIs.
  - APIs specification has been completed and included in the study document.
- Part 5: first test items for conformity tests were generated.

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Other input documents review

- Validation of functional equivalence of MPEG-G (M42306)
- New use case (Efficient storage of RNAseq data) (M42659) needs to be discussed.
- First inputs of work items to be considered for version 2:
  - Proposal for classification of chimeric pairs to support efficient access and identification of “chimeras” in MPEG-G files (M42364)
  - Efficient filtering of “duplicates” in MPEG-G files (M42364)
  - Inclusion of Variant Calling information and other statistical information in MPEG-G files (M42591)
  - New coding mode preserving coding order of reads (M42658)
- Addition to new test material (M42657)

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

WS on MPEG-G on Wednesday 18th

A workshop on applications of genomic information processing will be held on 18th April 2018 co-located with the 122nd MPEG meeting in San Diego:

Specifically the workshop will addresses:

- The perspectives of genomic information in medicine and public health
- The vision of interdisciplinary approaches to the analysis of genome sequencing data
- The challenges for the management of very large volumes of genome sequencing data
- The progresses of sequencing technology and data generation
- The reasons for supporting availability and exchange of genome sequencing data for improving scientific progress
- A status report on MPEG-G its new features and performance

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

WS on MPEG-G on Wednesday 18th

Start	End	What	Who
12:30	13:00	Registration	
13:00	13:15	Welcome & workshop goals	
13:15	13:40	“Genome and medical information portability, retrieval and analysis”	Amalio Telenti (Scripps Research Institute, USA)
13:40	14:05	“From womb to tomb sequencing: on the advantages on bringing multidisciplinary R&D to develop standards and analytics”	Ioannis Xenarios, (SIB Switzerland)
14:05	14:30	“Future of Genomics and Big Data”	Dawn Barry (Luna DNA, USA)
14:30	14:55	“Generation and Management of Large Sequence Files: Perspectives from the DNA Sequencing Core”	Alvaro G. Hernandez (UIUC DNA Services, USA)
14:55	15:10	Presentation of demonstrations	
15:10	16:00	Demo session and Coffee Break	
16:00	16:25	“The role of compression in the genomics data life cycle”	Come Raczy (Illumina Inc., USA)
16:25	16:50	"An overview of the MPEG-G standard for the compression and processing of genomic sequencing data"	Marco Mattavelli (EPFL, Switzerland)
16:50	17:30	Panel discussion, Q&A and concluding remarks	All speakers
17:30	18:00	Demo session resumes	

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Recommendations

- Continue technical and editorial work on the DIS text of Part 1 and of Part 2 during the week
- Close CE5
- Produce a study document of CD of Part 3
- Continue the integration work on Part 4 (Reference SW for both Part 1 and Part 2) before promoting it to CD level (July meeting)
- Continue the work on identifying and generating conformance test items
- Continue new use cases validation and the identification of new work items and corresponding solutions for version 2.