ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AHG on Genomic Information Representation (m40761)

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 N16759 based on the results of the answers to the Call for Proposals.
4. To contribute to WD editing for Part 1 and Part 2.
5. To collect requirements for specifying the format of metadata to be included in Part 3 : Genomic Information Representation APIs

AHG Activity

- Mainly focused around:
  - Editing the 2 WDs for Part 1 (Transport and Storage of Genomic Information) and Part 2 (Compression of Genomic Information)
  - Running the 2 active Core Experiments:
    - CE5 entropy coding
    - CE6 multiple alignments
- Coordination activities with TC276 and other ISO committees:
  - WD Part1 and WD Part 2 have been presented to the plenary TC276 meeting in Seoul in May
- 24 Input documents to this meeting

Summary of CE 5 “Entropy Coding”

- 3 participants reported results on CE 5 (M40860, M40861, M40804, M40994)
- Summary of results:
  - Descriptors streams have been generated and distributed including QV descriptors and read name descriptors
  - CABAC solutions have been cross checked and can provide comparable solutions to best performing solutions (LZMA +/-2%)
  - Significantly faster in speed (at least a factor 10)
- Conclusion/Consensus:
  - Continue experiments with read descriptors from more test items

Summary of CE 6 “Multiple alignments”

- 1 Input document (M40893)
- Solution provided in terms of:
  - Definition of global parameters in current syntax
  - Definition of 4 new descriptors
- The solution also supports:
  - Multiple Alignments on different sequences
  - Secondary Alignments with different CIGAR string than the primary
  - Multiple Primary Alignments
- Conclusion
  - Proposed syntax goes to WD
  - New syntax elements go to CE 5 Entropy Coding
  - Performance results need to be validated with other data: a) larger sets of paired-end sequences, b) variable length reads (e.g. PacBio)
  - Study extensions necessary to support indexing
  - CE 6 continues

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Input document review

- All inputs reviewed during the Sat. and Sun. AHG meeting
- One new technology proposed for indexing unmapped reads (M40894):
  - Approved for inclusion in the WD
- Extensions, improvements of identified limitations of current WD technologies:
  - Proposed solutions approved for inclusion in WD
- Proposals of 3 new “use cases” supports:
  - Several relevant objections raised
  - Three proposal rejected
- Information about a related NIH activity

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Recommendations

- Define by the end of the meeting: Specification of reference SW components for the different Parts and a workplan for each component development.
- Schedule editing sessions on:
  - identified limitations, issues and extensions discussed and approved for Part 1 and Part 2 WDs.
  - on new technologies approved to be included in Part 1 and Part 2 WDs
- Continue CE5 and CE6 with modified/extended mandates.
- Continue the work on metadata definition:
  - action point to collect feedback from major users.