ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AHG on Genomic Information Representation (m41383)

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

Joint AHG with ISO/IEC TC276

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 N17078.
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
6. To collect preliminary test item descriptions and bytestreams for Conformance testing

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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
  - Intermediate ad-hoc meeting in Woking (27th, 28th, 29th September)
- Collecting feedback on metadata definition
- Coordination activities with TC276 and other ISO committees:
  - No meetings after MPEG in Torino, next TC276/WG5 meeting on 29th-30th Nov.
- Initial specification of list items for conformance
- 21 Input documents at this meeting

Summary of CE 5 “Entropy Coding”

- 4 input documents
- Summary of results:
  - Descriptors streams have been generated and distributed including QV descriptors and read name descriptors
  - CABAC-based solutions have been cross checked and can provide better results than standard general purpose entropy coding solutions (LZMA, PPMd, ….) in the range of 5-10 %
  - Significantly faster in speed (at least a factor 10)
- Conclusion/Consensus:
  - Continue CE5 with read descriptors, read name descriptors and QVs:
    - for new test items with the new descriptors of multiple alignments
    - for achieving higher compression or lower computational complexity

Summary of CE 6 “Multiple alignments”

- Working session at the intermediate ad-hoc group meeting in Woking
- 2 Input documents (M41477, M41478)
- Solution provided in terms of:
  - Definition of global parameters in current syntax
  - Definition of new descriptors
  - Definition of an extended CIGAR syntax
- The solution also supports new features:
  - Splices strandedness
  - Secondary Alignments with different CIGAR string than the primary
- Conclusion
  - Proposed syntax extensions goes to CD of Part 2
  - New syntax elements go to CE5 Entropy Coding
  - Performance results need to be validated with other data within CE5 : a) larger sets of paired-end sequences, b) variable length reads (e.g. PacBio)
  - CE 6 can be closed

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Input document review

- All inputs reviewed during the Sat. and Sun. AHG meeting
- All proposed changes to Part 1 (Syntax improvements):
  - Approved for inclusion in the WD
- Syntax changes to Part 2 approved for inclusion in WD:
  - Decoding process for references and QVs
  - Alignment with Part 1
  - Multiple alignments
- Other proposed changes to Part 2 (M41641, M41771):
  - New descriptors and coding modes still under discussion
- Part 3 related inputs revised according to major user feedbacks approved for Part 3 WD
- Input on Reference SW status reviewed

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Recommendations

- Promote Part 1 and Part 2 to CD level
- Continue technical and editorial work on the WD of Part 1 and part 2 during the week
- Continue CE5 to provide technical input to Part 2.
- Continue the work on metadata and APIs definition:
  - Possibly collect feedback also from other identified major users.
- Promote Part 3 (Metadata and APIs) to WD level
- Promote Part 4 (Reference SW for both Part 1 and Part 2) to WD level
- Continue the work on identifying conformance test items for preparing the issue of Part 5 WD at next meeting