ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AHG on Requirements on Genome Compression and Storage (m39627)

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 (N16527).
3. To carry out the core experiments described in document N16526 based on the results of the answers to the Call for Proposals.
4. To analyze genome information applications and identify new requirements related to file format definition and compression technology.
5. To further validate and possibly extend the benchmark framework for quasi-lossless compression of sequencing quality values defined in N16525 and the identified metrics.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AHG Activity

- Mainly focused around running the 4 Core Experiments
  - Develop the SW tools according to the CE descriptions
  - Cross checking the results according to the test database
- Coordination activities with TC276 and other ISO committees
- More than 100 emails on the reflector and several hundreds among CE participants
- 27 Input documents

Dissemination and coordination with other ISO TCs

- Sharing Information about relevant Genome Sequencing Data Projects, such as genome compression, NGS data quality evaluation, whole genome sequencing for typing and characterization…
  - Different information sharing platforms are tested.
  - A Wiki-Style Website will be finally established by mid-March
  - An information request form will be sent out for collecting information.
  - Leaders may have full-access rights to maintain the contents.
  - Recommend to share the documents to public (under certain rules, and owners/groups can make their own decision)
- Action Point Plan:
  - Website will be launched by the end of March
  - Information will be collected by leaders by the mid of March

Summary of CE 1 “Unified Representation for Reads”

- 9 participants submitting technologies
- 1 input documents summarizing results for all different classes of data (M39845):
  - Ranking of cross-checked results is available for all classes of data
  - 1 candidate is better ranked in almost all data sets and categories aligned data (no support for PacBio reads)
  - 1 candidate is better ranked for raw data category
  - 1 candidate is better ranked for metadata in almost all data sets
  - Initial definition of a common representation based on a single set of descriptors
- Conclusion/Consensus: complete the definition of a common set of descriptors supporting the coding modes of best ranked tools

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Summary of CE 2 “Quality Values”

- 8 Participants
- 7 QV lossless tools submitted, 5 have been cross-checked (3 partial cross-checks)
- 6 QV quasi-lossless tools, 4 cross checked (none completed)
- 1 input document summarizing cross-check results:
  - all lossless tools present similar performance
  - 3 quasi lossless tools provide comparable “best” results, but measurement method must improve to increase dynamic range to provide a meaningful rank
- Conclusion/Consensus:
  - Lossless: work to defined a unified approach to compression
  - Quasi-lossless: complete the cross-checks with a single analysis pipeline
  - Work during the week to define a set of descriptors specifying decoding syntax
  - Work during the week to define a more sensitive measurement method

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Summary of CE 3 “Read Identifier Compression”

- Validate and evaluate 8 submitted technologies from 5 submitters the CfP
- 1 Candidate for WD (three technologies ranked)
- 1 input document (M40047): summary of CE results
- Conclusion/Consensus for WD: extract the decoding syntax for inclusion in the WD
- Recommendations: cross check results with reads compressor actual ordering, block size and more tech providers PacBio and Oxford Nanopore

Summary of CE 4 “Genomic Access Abstract Layer”

- 4 proposers expressed interest to participate to the CE
- 2 technologies submitted to the CE, 4 participants to the CE
  - 2 technologies cross-checked
  - Steps to integrate the two technologies in a single specification including all functionality
- Summary of results
- Joint submission presenting the integration of the two technologies (M39945)
- Definition of use cases (M39828)

- Conclusion/Consensus:
  - Solve some issues pending on file format
  - Verify streaming capabilities when the file format will include all functionality of the two technologies

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Input documents

No.	Title	Authors
m39663	Core Experiment 2 summary	Jan Voges
m39664	Core Experiment 2 results LUH-Stanford/UIUC	Jan Voges, Mikel Hernaez
m39665	Core Experiment 1 results LUH	Jan Voges
m39666	Core Experiment 3 results LUH	Jan Voges
m39748	Core Experiment 1 cross-check SFU	Jan Voges
m39749	Core Experiment 2 cross-check SFU	Jan Voges
m39750	Core Experiment 3 cross-check SFU	Jan Voges
m39828	Use cases Cases for an efficient genomic information representation	Nicolas Guex, Christian Iseli, Thierry Schuepbach, Ioannis Xenarios, Claudio Alberti, Marco Mattavelli
m39829	Data structures used in CRAM v3	James K. Bonfield, Claudio Alberti
m39845	Report on Core Experiment 1 on Genome Compression	Claudio Alberti
m39870	Tools, Technology and Results for CE1 on Genomic Information Representation	Giorgio Zoia, Daniele Renzi,
m39871	Tools, Technology and Results for CE4 on Genomic Access Abstract Layer	Giorgio Zoia, Daniele Renzi
m40109	Cross check of CE3 of the submission by Wellcome Trust Inst.	Stanford University

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Input Documents

No.	Title	Authors
m39939	Genomic Information Compression and Storage CE4: DMAG-UPC, GENIFF v.2 implementation	Jaime Delgado, Silvia Llorente, Daniel Naro, Lukasz Roguski, Dmitry Repchevsky
m39940	GENIFF (GENomic Information File Format) v2	Jaime Delgado, Silvia Llorente, Daniel Naro, Lukasz Roguski, Dmitry Repchevsky
m39941	Users Guide for DMAG-UPCs GENIFF v.2 software used for Genome information CEs	Jaime Delgado, Silvia Llorente, Daniel Naro, Lukasz Roguski, Dmitry Repchevsky
m39943	Genomic Information Compression and Storage CE4: Cross-check of HEGIF	Daniel Naro, Jaime Delgado, Silvia Llorente
m39944	GENIFF (GENomic Information File Format) v2: Security and signature issues	Daniel Naro, Jaime Delgado, Silvia Llorente
m39951	Genomic Information Compression and Storage CE1, CE2 & CE3: Cross-checkings	Dmitry Repchevsky, Jordi Portell, Jaime Delgado, Lukasz Roguski
m39945	Towards a WD for a format on genomic information compression and storage	Jaime Delgado, Silvia Llorente, Daniel Naro, Giorgio Zoia
m39948	FAPEC v.2 for Core Experiments on Genomic Information Compression and Storage	Jordi Portell, Jaime Delgado
m39950	Description of FAPEC v.2 Core Experiments execution and results on Genomic Information Compression and Storage	Jordi Portell, Jaime Delgado
m40047	Summary of CE3	Mikel Hernaez
m39875	CE1: Cross-check of the PirBright Genomic Data Compression proposal	Tom Paridaens, Glenn Van Wallendael, Wesley De Neve, Peter Lambert
m39876	CE2: Cross-check of the PirBright Genomic Data Compression proposal	Tom Paridaens, Glenn Van Wallendael, Wesley De Neve, Peter Lambert
m39881	CE1 & CE2: Proposal for a genomic data file compression framework, based on existing MPEG practices and technologies (update)	Tom Paridaens, Glenn Van Wallendael, Wesley De Neve, Peter Lambert
m40109	Context Model Compression of Genomic aligned data - CE1	Idoia Ochoa, Mikel Hernaez, Reggy Long

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Recommendations

- Establish the dissemination and information exchange platform with the support of TC276 to include information on other ISO TCs genomic related activities
- Continue with the 4 CE extending the scope according to current results
- Produce a WD with the current results of Ces and reached consensus
- Establish a new CE on entropy coding
- Establish a liaison with GA4GH