**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 N14553**

**July 2014, Sapporo, Japan**

|  |  |
| --- | --- |
| **Source** | **Requirement Subgroup** |
| **Status** | **Approved** |
| **Title** | **Issues in genome compression and storage** |
| **Authors** | **Marco Mattavelli (EPFL)**  **Claudio Alberti (EPFL)**  **Yann Thoma (HEIG-VD)**  **Heinz Stockinger (SIB)**  **Jaime Delgado (UPC)**  **Ioannis Xenarios (SIB)** |

# Problem definition

Recent technological advances in sequencing genomes (including humans) have enabled a faster and more cost-efficient approach to sequence individual genomes. In a not so far future it will be possible to sequence and analyze the entire genome of any patient as frequently as we currently can resort to blood tests and standard clinical analysis. However, large scale sequencing and analysis efforts still have many scientific and technological challenges that currently prevent a readily usable genome in the clinical practice being in hospitals or health care institutions.

Creation, storage and analysis of sequence data has many technological and engineering challenges such as the amount of data, the structure of data, the computer hardware limitations (memory, storage space, processing speed, I/O bandwidth, network bandwidth etc.). In addition, there is the inherent necessity to rerun analysis over time due to the change of the underlying knowledge.

**Storage and analysis of the human genome**

Latest generation DNA sequencing methods are increasingly applied to address the genetic diversity of humans and different organisms affecting health and well-being of human populations. Such sequencing experiments typically yield insufficient data to cover multiple times (as of 2012 roughly 50 times coverage of a given position) the complete human genome with a length of ~3 billion base pairs. Applying such approaches to a large number of different genomic samples exemplified by the “1000 personal genomes project” [alt10] produces a considerable amount of data which is still frequently stored in human readable text files. Such text data formats could and are currently compressed using standard text compression algorithms (i.e. Lempel-Ziv), but such option is by far not the best option if we consider the special format of the genome data for which the information content of a nucleotide sequence with 4 bases has a specific structure that admits more appropriate coding models. Moreover, since the majority of computer programs depend on a textual input, current practice typically requires repeated cycles of compression and decompression which results inefficiencies. In practice, such an approach is only feasible for some processing/analysis cases. A solution denoted as “compressive genomics” [loh12] suggests to compute specific analysis directly on compressed genomic representations, which means that faster access to data since no decompression is required. Similar approaches are typical in the database world (e.g. access to compressed bitmap indices [wu10]), but are not yet common practice in the domain of genomics.

**Current status of genome data collection**

Over the last years, several different reference genomes have been painfully established for humans, some animals and plants. However, since each individual has genetic differences (Single Nucleotide Variant-SNV) with respect to a reference genome, there are currently many different genome analysis projects world-wide (1000 genomes [alt10] www.1000genomes.org, ENCODE [mah12]) with the aim of sequencing individuals. Most of these sequencing efforts use a reference genome as a “golden standard” to assemble and map these individual genomes. Unfortunately, reference genomes represent an “average genome” (e.g. for specific parts of the genome no sequence information is available) and newly sequenced individual genomes, that are mapped on the reference genome, usually suffer from the same loss of information.

# Draft requirements in genome compression and storage

**General requirements**

|  |  |  |
| --- | --- | --- |
|  | **Requirement** | **Rationale** |
| 1 | Maintain the data integrity produced by the sequencing technology. | It shall be possible to revert to the original representation of data (or equivalent) that was produced by the sequencing devices. This is due to leave the possibility to apply newly discovered alignment and mapping approaches to old data. |
| 2 | Compressed data should be structured into data access units. | Access to data shall work in terms of Access Units that be accessed without the need to decompress the whole genomic information. I.e. selective access and manipulation shall be supported. |
| 3 | Access units should be able to contain variable size sequence lengths. | Sequences lengths heavily vary with the sequencing device and will keep changing with new sequencing technologies going to the market. While Access Units can have a fixed for a specific application, the standard shall support variable reads lengths. |
| 4 | Data should be “queryable”. | The data set structured according to the standard shall be accessible using queries such as:   * “search for the first match for a given sample” * “search all occurrencies of a sample” * “go to a specific location and dump the content” |
| 5 | Stored sequence data should be extensible allowing adding or modifying sequencing information. Such modifications should be traceable, reversible and incremental. | An existing dataset shall support update, revert, history, diff operations.  For each operation support for integrity and correctness checks is recommended. |
| 6 | Compressed data formats should be able to integrate annotations - e.g. features on a given region, **non-contiguous** (this is the most important of all) regions of the data structure. | Annotations are metadata that are used for example to indicate relations between reads located at different spatial locations. |
| 7 | The nucleotide information shall support at least the 16 IUPAC codes but shall be able to encode more in the future | This is required to support the coding of other genomic data such as amino acids. For example basic “profiles” could support only the 16 IUPAC codes while extended ones would be able to additionally encode the 23 amino acids codes. |

**Data security-related requirements**

Genome data normally comes either from (anonymous) people who provide their data for general analysis or from patients whose health status is specifically analyzed. This implies that security (in general terms) is key when formatting, processing, storing and interchanging genome data.

The security-related requirements could be summarized as follow:

|  |  |  |
| --- | --- | --- |
|  | **Requirement** | **Rationale** |
| 8 | Data protection. | Ability to prevent unauthorized access. |
| 9 | Support for anonymization, when requested. | Data shall be made anonymous in such a way that there is no way to retrieve the identity of the originating organism (especially human) while preserving all the genomic information for analysis. |
| 10 | Privacy and confidentiality. | Knowing, by non-authorized agents, sensible data about the health of a person has to be avoided. Furthermore, individuals should be allowed to decide their required level of privacy. |
| 11 | Transparency | How and for which purpose the information is used should be known. Usage restriction shall be applicable to the data. |
| 12 | Accountability and traceability. | Data access and manipulation shall be traceable together with the identity of parties having access to data. |

**Use of existing MPEG (and non-MPEG) standards**

Where possible, MPEG standards, or even non-MPEG ones, should be used to provide the identified requirements.

Initial examples (although further discussion and consideration is needed) could include MPEG-21 Event Reporting for traceability and transparency; or MPEG-21 IPMP, REL/MVCO and CEL/MCO for data protection and privacy.

# Conclusion

The vast amount of genome data currently generated by the different sequencing projects raises the issue of efficient storage and processing of genome data satisfying several specific requirements. Such challenges imply the use/existence of an appropriate **open standard** for both compression and storage, that might have common properties or in part use existing technology applied so far to audio/video and 3DG signals, but that currently does not exist.

A solution that meets the requirements listed in this document would enable the scientific and industrial community working on genomic information to address the challenges of a domain where the variety amongst individuals is higher than they were expecting only a few years ago.

The two driving elements of the design process should be on one side the reuse of existing well-established technologies for compression, storage, privacy and security, etc. (e.g. MPEG compression schemes, MPEG-21), and on the other the flexibility to incrementally address current and future needs without being bound to specific application constraints.

**References**

[Ka11] S. D. Kahn. On the Future of Genomic Data. Science, 331(6018): 728-729, 2011.

[alt10] Altshuler D, Durbin R, et al. A map of human genome variation from population-scale sequencing, Nature. 467(7319):1061-73, 2010.

[loh12] Loh P.-R, Baym M, and Berger B. Compressive genomics. Nat. Biotechnol. 30:627–630, 2012.

[wu10] Wu K, Shoshani A, Stockinger K. Analyses of multi-level and multi-component compressed bitmap indexes. ACM Trans. Database Syst. 35(1), 2010.

[mah12] Maher B, ENCODE: The human encyclopaedia. Nature. 489(7414):46-8, 2012.