**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

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

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2014/N15047**

**October 2014, Strasbourg, France**

|  |  |
| --- | --- |
| **Source** | **Requirements** |
| **Status** | **Draft** |
| **Title** | **White paper on genomic information compression and storage** |
| **Authors** | Claudio Alberti, Marco Mattavelli (EPFL), Leonardo Chiariglione (CEDEO), Ioannis Xenarios, Nicolas Guex, Heinz Stockinger, Thierry Schuepbach, Pascal Kahlem, Christian Iseli, Daniel Zerzion, Dmitry Kuznetsov (SIB), Yann Thoma, Enrico Petraglio (HEIG-VD). |

Table of Contents

[**1** **Executive Summary** 3](#_Toc403718787)

[**2** **Genomic information generation and manipulation** 4](#_Toc403718788)

[2.1 Genome versus sequence data compression 4](#_Toc403718789)

[2.2 DNA sequencing 5](#_Toc403718790)

[2.2.1 The NGS Industry Landscape 6](#_Toc403718791)

[2.3 Bioinformatics and genomics 6](#_Toc403718792)

[2.4 Applications 7](#_Toc403718793)

[2.4.1 Diagnostics and Personalized Medicine 7](#_Toc403718794)

[2.4.2 Drug Discovery 7](#_Toc403718795)

[2.4.3 Biomarker Discovery 7](#_Toc403718796)

[**3 Some relevant initiatives** 7](#_Toc403718797)

[3.1 ISO Technical Committees 7](#_Toc403718798)

[3.1.1 ISO TC 215 - Health Informatics 7](#_Toc403718799)

[3.1.2 ISO TC 276 – Biotechnology 8](#_Toc403718800)

[3.2 Non-ISO initiatives and groups 8](#_Toc403718801)

[3.2.1 Pistoia Alliance Inc. 8](#_Toc403718802)

[3.2.2 Expert Committee on Biological Standardization of the WHO 8](#_Toc403718803)

[**4** **File Formats** 9](#_Toc403718804)

[***4.1*** ***Sequencing output FastA/FastQ*** 9](#_Toc403718805)

[***4.2*** ***The IUPAC ambiguity codes*** 10](#_Toc403718806)

[**5 Genome compression** 10](#_Toc403718807)

[5.1 Methods 10](#_Toc403718808)

[5.1.1 Naive bit encoding 10](#_Toc403718809)

[5.1.2 Dictionary based 10](#_Toc403718810)

[5.1.3 Statistical methods 10](#_Toc403718811)

[5.1.4 Referential methods 11](#_Toc403718812)

[5.2 Tools 11](#_Toc403718813)

[6 Available sequence data 12](#_Toc403718814)

[7 Requirements from identified applications 12](#_Toc403718815)

[8 Existing tools versus requirements 13](#_Toc403718816)

[9 Beyond storage 14](#_Toc403718817)

[10 Conclusions 14](#_Toc403718818)

[**11** **References** 14](#_Toc403718819)

1. **Executive Summary**

The sequencing of the genetic information of human genome has become affordable due to high-throughput sequencing technology [1], [2]. This opens new perspectives for the diagnosis and successful treatment of cancer and other genetic illnesses. However, there remain challenges, scientific as well as computational, that need to be addressed for this technology to find its way into everyday practice in healthcare and medicine. The first challenge is to cope with the flood of sequencing data. For instance, a database covering the inhabitants of a small country like Switzerland would need to store a staggering amount of data, about 2’335’740 Terabytes. The second challenge is the ability to process such a deluge of data in order to 1) increase the scientific knowledge of genome sequence information and 2) search genome databases for diagnosis and therapy purposes. High-performance compression of genomic data is required to reduce the storage size, increase transmission speed and reduce the cost of I/O bandwidth connecting the database and the processing facilities.

The current trends in sequencing data generation show clearly that the storage and transfer (bandwidth) costs will soon become comparable to the costs of sequencing. This means that IT costs may soon become a major obstacle to such genome analysis applications as personalized medicine, early diagnostics and drugs discovery, unless genetic data compression reduces IT costs on par with sequencing costs.

This document has been drafted with the goal to help MPEG to assess the opportunity to start a standardization effort in genetic information processing, particularly compression, and provides

1. An overview of the current status of tools and technology supporting genomic information compression and storage
2. An analysis of related challenges for the stakeholders
3. A first list of essential requirements checked against a set of tools recently presented in literature and used by bioinformaticians.
4. **Genomic information generation and manipulation**

Figure 1 shows the main stages of genomic information manipulation in existing bioinformatics applications. The steps depicted include:

1. Sequencing: expression of genomic information as strings (a.k.a. sequences or reads) of nucleotides identifiers.
2. Alignment/mapping: sequences arrangement to identify regions of similarity among them (*de-novo* assembly) or with respect to an external reference (a pre-constructed genome). Sequences are encoded in the form of SAM files and its binary dual named BAM [1].
3. Compression: data encoding to use less bits.
4. Storage: compressed data is stored and made available via database interfaces or files.
5. Decompression/access: access to data to perform analysis.
6. Update: previously sequenced genomic information might be updated by means of new alignment techniques or new sequencing (a.k.a. re-sequencing).

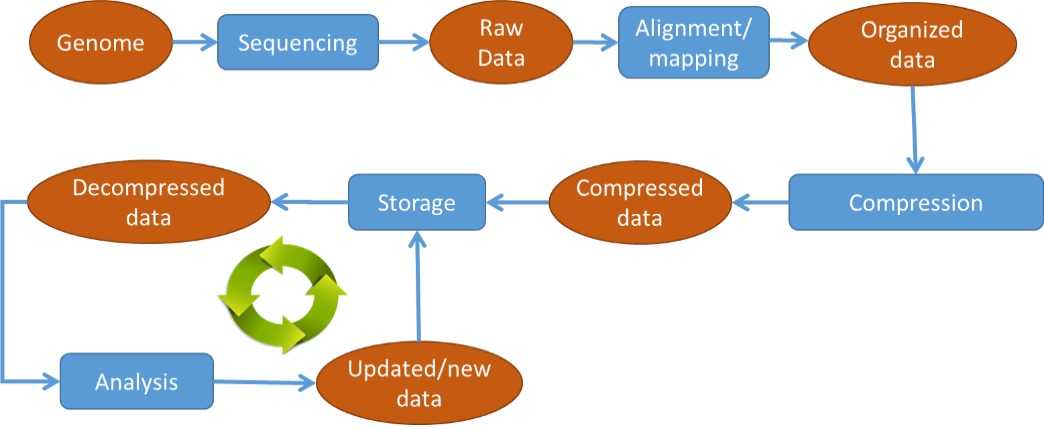


Figure 1 – Genomic information generation and manipulation stages

## Genome versus sequence data compression

One important distinction that is worth stressing here is the difference between entire *genome compression* and *sequence data compression*.

*Genome compression* tools aim at encoding the genetic information of a living organism expressed as a sequence of symbols representing the nucleotides. This string is about 3.2 billion symbols long for the human being (organized in 23 chromosomes) and can be up to 100+ billion symbols long for other organisms. The encoding of an entire genome is the result of a long (error prone) process of analysis that today can only provide a close approximation to the real genetic sequence.

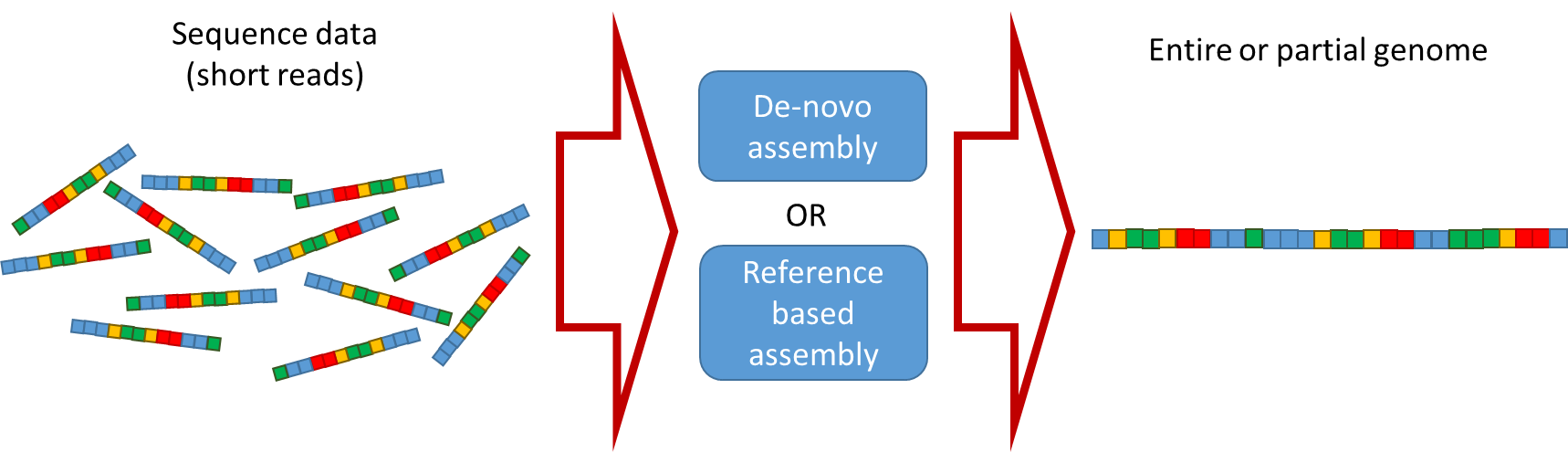


Figure 2 – From short reads to genome

On the other hand the *compression of sequence reads* focuses on encoding the output of new generation sequencing machines which are able to extract large amounts of short (from 35 to over 1,000) nucleotides sequences (“reads”). This is the type of information that is nowadays in need of efficient compression in order to enable the wide range of applications made possible by recent advances and discoveries in genomics. This document will focus on compression of sequence data (short reads).

## DNA sequencing

The expression “Next Generation Sequencing” (NGS) designates the process of fast extraction of large amounts of genomic information from samples of organic material belonging to a living organism. Modern sequencing devices are able to produce several hundred million “reads” (genomic information related to pieces of the whole genome) per day. According to the specific technology employed, each read can contain from a few dozens to several thousand bases (the atomic unit of genomic information) together with optional metadata. This rate of information generation dramatically outpaces any progress in digital information storage and transmission. In this context the last decade has witnessed several attempts of finding suitable solutions to compress genomic information efficiently and robustly. These efforts have been produced by research institutions, universities, industries with a wide range of diverse priorities and drivers. The result is a proliferation of tools and formats able to address only the specific needs of their authors, without any perspective to be flexible enough to meet the various needs of the scientific and industrial communities.

As a consequence, nowadays most of the players of the “genomic revolution” are open to initiatives aiming at making the growing amount of genomic information more manageable and rapidly “consumable” by the tools used for analysis.

For instance the human genome is composed by a sequence of about 3 billion nucleotide bases. Research projects can produce with just one sequencing analysis, a volume of data (in the form of relatively small fragments of the genome) that reaches up to 400-500 times the size of the complete human genome. Faster sequencing technology produce a much higher volume of data with a much higher redundancy which requires much more efficient (in terms of both size and processing speed) compression than the current simple and non-standard methods available today.

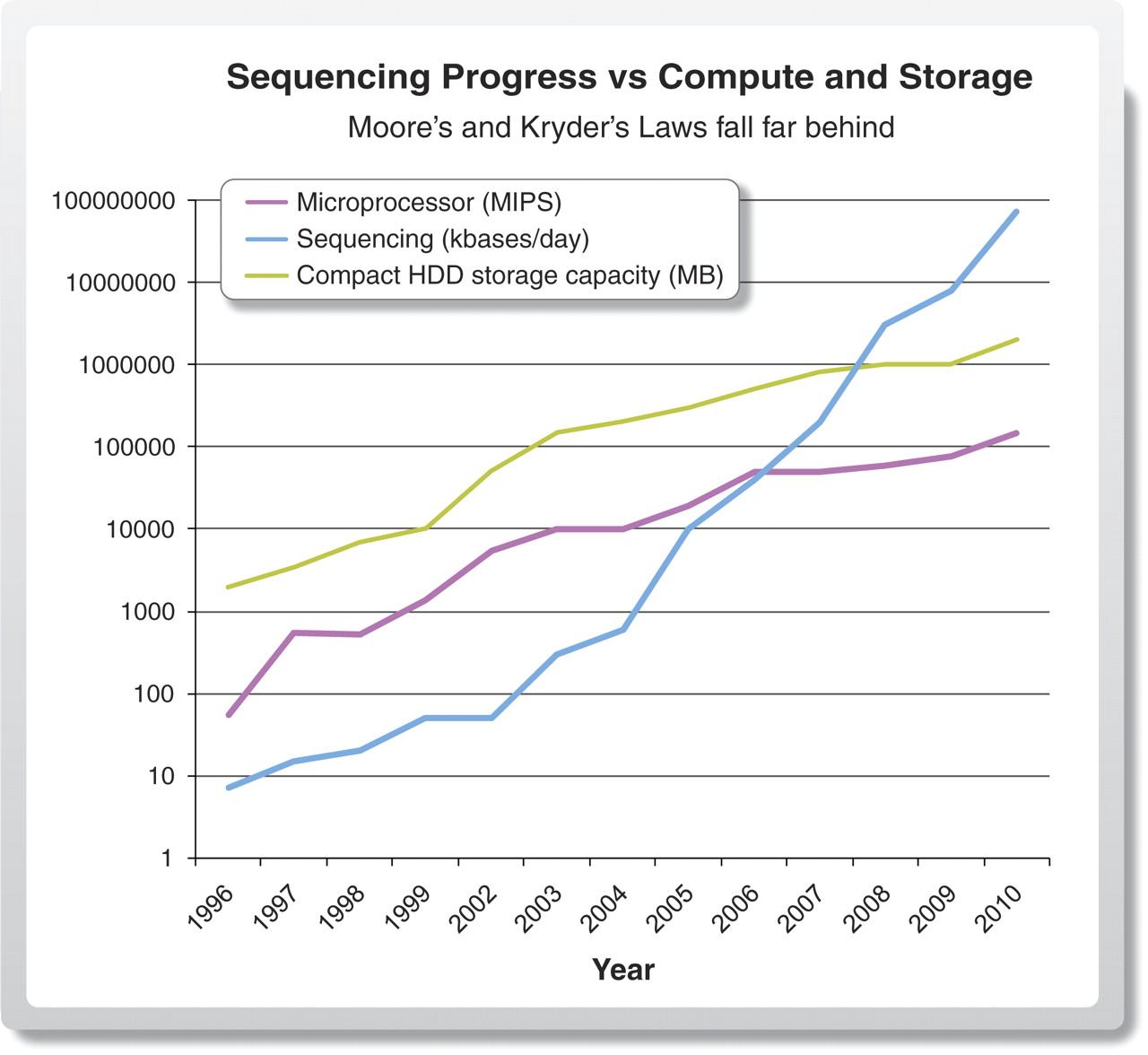


Figure 3 Moore’s law versus Sequencing, from [2]

What is important to be remarked is that within such huge amount of data, even if most of the fragments can be considered redundant versus the theoretical size of the genome, these cannot be simply discarded, because on one side it is the statistical indication of the correctness of the reads and on the other side small differences in some fragments might indicate pathologies that might be appropriately taken care of.

### The NGS Industry Landscape

A proliferation of new sequencing technology is rapidly spreading across the market of NGS machines. Illumina, Life Technologies (Thermo Fisher Scientific), 454 Lifesciences (Roche) and Pacific Biosciences are companies that commercialize equipment relying on different sequencing methods briefly described below.

The main difference among the sequencing devices commercialized by these companies is the technology employed to process the organic material to determine the precise order of nucleotides within DNA strands. The different methods result in extremely different segments lengths and performance in terms of speed, accuracy, cost and throughput.

Table 1 is taken from [Wikipedia](http://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods) and compares the methods employed by the 4 players mentioned above.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Read length** | **Accuracy** | **Reads per run** | **Time per run** | **Cost per 1 million bases (in US$)** | **Advantages** | **Disadvantages** |
| **Single-molecule real-time sequencing (Pacific Biosciences)** | 5,500 bp to 8,500 bp avg (10,000 bp [N50](http://en.wikipedia.org/wiki/N50_statistic)); maximum read length >30,000 bases | 99.999% consensus accuracy; 87% single-read accuracy | 50,000 per SMRT cell, or ~400 megabases | 30 minutes to 2 hours | $0.33–$1.00 | Longest read length. Fast. Detects 4mC, 5mC, 6mA | Moderate throughput. Equipment can be very expensive. |
| **Ion semiconductor (Life Technologies)** | up to 400 bp | 98% | up to 80 million | 2 hours | $1 | Less expensive equipment. Fast. | Homopolymer errors. |
| **Pyrosequencing (454 Lifesciences)** | 700 bp | 99.9% | 1 million | 24 hours | $10 | Long read size. Fast. | Runs are expensive. Homopolymer errors. |
| **Sequencing by synthesis (Illumina)** | 50 to 300 bp | 98% | up to 3 billion | 1 to 10 days | $0.05 to $0.15 | Potential for high sequence yield, depending upon sequencer model and desired application. | Equipment can be very expensive. Requires high concentrations of DNA. |

Table 1 - Comparison of next-generation sequencing methods

## Bioinformatics and genomics

The term *bioinformatics* designates the scientific domain that uses computing infrastructures to analyse biological data. The term is very broad and includes several fields and disciplines. Among them *genomics* applies DNA and RNA sequencing methods and computers to analyse the genomes of human beings and other organisms.

The main goals of genomics include:

1. The identification of the complete genome of organisms
2. The comparison among genomes of different organisms
3. The study of mutations in time of the genome of a given organism
4. The identification of genes (portions of a genome) functions
5. The definition of the spatial structure of genes within a genome

All applications of genomics such as genomic medicine, synthetic biology and bioengineering share the need of accessing and transporting genomic data rapidly and efficiently. The rapid evolution in genomic information generation is requiring dramatic advancements in databases technology, computational platforms, mathematical and statistical methods and theory to meet the requirements of the management and analysis of large biological datasets.

Nowadays tools used by bioinformaticians range from simple scripts to large commercial products. A large literature of open source software is available in various forms from development projects followed by communities of developers to simple reference software accompanying scientific publications. In some cases such tools implement very sophisticated compression schemes using entropy and arithmetic coding, but none of them meets all the requirements of the applications mentioned above. In particular, support for random access to data according to criteria expressed in a formal way is an important feature that the scientific community in looking for.

## Applications

### Diagnostics and Personalized Medicine

Genome-based diagnostic tests have been recently developed to make personalized treatment possible thanks to discovered links between specific genetic variants and diseases. Such tests have the potential to predict risk and drive preliminary therapeutic interventions, to detect onset of disease, or detect residual disease. Although clinicians and patients are still far from being educated in how best to apply genetic knowledge in better targeting (that is, in whom to intervene) and tailoring (how best to intervene) preventive efforts, improved health is a major goal of genomic research.

### Drug Discovery

Complete knowledge of the functions of all human genes might dramatically change drug discovery development processes and drug research as a whole. The application of genomic technologies to the clinical development of new and existing drugs is known as pharmacogenomics. Thanks to the recent development in genomics and pharmacogenomics in clinical research and clinical medicine, diseases could be treated in a close future according to genetic and specific individual markers, so that medications and dosages could be optimized according to the genetic profile of individual patients.

### Biomarker Discovery

In medicine a biomarker is an indicator of the presence of a disease state or any other physiological state. More generally anything that is measurable and related to the state of an organism can be considered a biomarker. The pharmaceutical industry is increasingly interested in biomarker discovery because biomarkers could represent early signals of disease in clinical trials, and possible drug targets.

# Some relevant initiatives

## ISO Technical Committees

### ISO TC 215 - Health Informatics

The scope of this ISO Technical Committee is defined as: *“Standardization in the field of information for health, and Health Information and Communications Technology (ICT) to promote interoperability between independent systems, to enable compatibility and consistency for health information and data, as well as to reduce duplication of effort and redundancies.”*

Among the produced standard documents the most interesting is **ISO 25720:2009 - Genomic Sequence Variation Markup Language (GSVML)**.

The scope of this standard is the specification of a common format for the exchange of genomic sequence variation data among existing databases. The aim is to define a standard envelop able to carry all the major existing formats for human genomic data.

While this is the most interesting effort for the standardization of a file format for the encoding of genomic information, the design of ISO 25720:2009 does not specifically address issues around efficient data compression and support of next generation sequencing technologies. Some of the main weaknesses are listed below.

* It has not been revised in the light of the new sequencing technologies that produce both human, virus and bacteria sequences in shot.
* It is not conceived to improve storage efficiency as it’s entirely XML based and annotation-based (variant annotation to be precise)
* It only meets the requirements of those applications interested in the variation at a single position in a gene. Since 2010 the field has evolved tremendously and this is not sufficient any more for a broad range of applications.

### ISO TC 276 – Biotechnology

Among the mandates of this Technical Committee, the standardization of “Computing tools, bioinformatics for international comparability and integrability of data” is mentioned.

The most recent activity has been the organization of a [Workshop in October 2011 on “International Standards for Biotechnology”](http://www.iso.org/sites/biotechnology2011/index.html). The goal of the workshop was to create an opportunity *“to promote a dialogue among the organizations most active in standardization for biotechnology, to foster better understanding among the key players and to capture input, recommendations on relevant matters and possible priority action items which will be channeled for consideration to the existing ISO technical and governance bodies”*.

The workshop outcome has been a set of recommendations on how ISO work in the biotechnology field should be structured. In particular it is interesting that one of the recommendations includes the need to ***standardize data structuring and processing for genomic applications*.**

## Non-ISO initiatives and groups

### Pistoia Alliance Inc.

The Pistoia Alliance Inc. [3] is a private consortium of pharmaceutical industries, universities and research centres which aims at supporting collaboration in the development of tools and technology for the manipulation of biological data. Its mission is to “*lower barriers to innovation by improving the interoperability of R&D business processes through precompetitive collaboration*”. While the scope of the Alliance is very broad, it is worth mentioning an initiative promoted in 2012 for the comparison of the most popular and efficient tools for DNA information compression: the SequenceSqueeze contest [4].

### Expert Committee on Biological Standardization of the WHO

The World Health Organization website [5] states that the “*Expert Committee on Biological Standardization is commissioned by WHO to establish detailed recommendations and guidelines for the manufacturing, licensing, and control of blood products, cell regulators, vaccines and related in vitro diagnostic tests. Members of the Expert Committee are scientists from national control agencies, academia, research institutes, public health bodies and the pharmaceutical industry acting as individual experts and not as representatives of their respective organizations or employers. The decisions and recommendations of the Committee are based entirely on scientific principles and considerations of public health*”.

As of today the committee had no specific activity on the standardization of compression or file format for genomic information storage, but when contacted they have shown interest in following any activity in this sense.

1. **File Formats**

Nowadays the largest majority of public repositories of sequence data provide data formatted in two - very similar - textual file formats named FastA and FastQ. FastQ exists in a few different flavors [6] defined by different sequencing machine vendors.

FastA and FastQ have been adopted in the recent past when the amount of generated information was not so important to create any issue of storage space. In addition, text files can easily be parsed and analyzed using scripting languages (e.g. bash, Perl, python) very popular on the UNIX platforms commonly used in this domain.

The explosion of the throughput of NGS machines pushed the adoption of popular file compression tools such as zip, tar and all the related flavors. These generic approaches to compression can anyway save between 50% and 75% of the original utilized space, which is currently becoming inadequate of at least one order of magnitude with respect to the requirements of faster and faster sequencing technology. The main drawback of this approach is the total lack of support for random access to portions of the compressed information.

FastA and FastQ are described and compared in Section 4.1 while Section 4.2 introduces a new standardized notation for nucleotides sequences aiming at merging the characteristics of FastA and FastQ towards a single file format.

* 1. ***Sequencing output FastA/FastQ***

FastA and FastQ are very similar textual formats that are used for genomic information generated by NGS machines. Table 2 compares the two formats.

|  |  |  |
| --- | --- | --- |
| FASTQ | Field | FASTA |
| @HWUSI-EAS100R:6:73:941:1973#0/1 | *Header (Unique ID plus other information). Only the first character is standard.* | **>HWUSI-EAS100R:6:73:941:1973#0/1** |
| GATTTGGGGT….. | *Nucleotides sequence* | **GATTTGGGGT……** |
| +SRR001666.1 071112\_SLXA-EAS1\_s\_7 | *Optional description. Only the first character is standard.* | Not present |
| !''\*((((\*\*\*+) | *Quality scores* | Not present |

Table 2 - Comparison between FastQ and FastA

Both file formats start with a header field where only the first character (“@“ for FastQ and “>” for FastA) is standardized to signal the start of a new read. The remaining text in the header usually identifies the originating experiment, the type of sequencing machine or technology adopted and other information aiming at identifying the source of the data.

The second field contains the symbols used to represent nucleotides in both FastA and FastQ. They are usually 5 types of symbols:

* A, C , G, T (T is replaced by U in case of RNA sequencing)
* A fifth symbol “N” used when the sequencing machine cannot take any decision.

FastQ has two additional fields:

* An optional container of additional metadata starting with “+”
* Quality scores expressing the level of confidence for each nucleotide encoded in the second field. The value and meaning of each symbol vary with the sequencing machine adopted.
  1. ***The IUPAC ambiguity codes***

The International Union for Pure and Applied Chemistry (IUPAC) has recently standardized a [larger set of symbols](http://www.bioinformatics.org/sms/iupac.html) (16, including the 5 currently used) that includes the information related to the incertitude of the read between one or more basis. This standard set of codes better represents and can replace the non-standard metadata previously used to indicate the "quality" of the base read. Therefore, supporting this slightly larger set of symbols can replace the support of the current non-standard metadata (the latter including quality scores that are usually machine dependent).

While the IUPAC standard is starting being supported by the latest Next Generation Sequencing machines, its use is still limited in the scientific community. Nonetheless the trend seems to points towards a progressive wider adoption with the gradual substitution of quality scores by IUPAC ambiguity codes. A FastA file format extended to 16 (or more) symbols seems to be on the (not so far) horizon.

# Genome compression

Several genomic data compression tools have been developed by researchers and developers with different interests in terms of requirements to be met (compression ratio, speed, memory footprint). Comparing such tools is impaired by their very diverse approaches to the problem in terms of test dataset, output format, actual availability of implementations and the related specifications. This section aims at providing a summary of the main approaches and tools currently in use; it is nevertheless incomplete due to the rapid proliferation of new methodologies and implementations.

## Methods

### Naive bit encoding

These highly inefficient methods are worth mentioning here only because they were among the first to be used and are still used in some circumstances. They simply encode several nucleotides within the same byte using fixed-length encoding [8]. For example the 4 nucleotides (A, C, G, T) can be encoded using a 2 bit alphabet so that 1 byte can encode 4 nucleotides (compression ratio 4:1 with respect to textual encoding).

### Dictionary based

A dictionary of repeated substrings is built at runtime or offline and then compression is performed by replacing each substring with a reference to the dictionary [9].

### Statistical methods

Also referred to as entropy encoding algorithms, they derive a probabilistic model from the input data. When appropriately defined, the model is supposed to use the available information to predict next symbols of the sequence. When a reliable model is built, these methods result in very high compression rates.

### Referential methods

Also known as reference based approaches, these methods encode substrings by means of references to an external genome. For each substring of the input data that can be mapped to the reference, only the position and possible deviations with respect to the reference are encoded. With respect to dictionary-based approaches here the reference is static, while dictionaries are usually dynamic and can be updated during compression.

## Tools

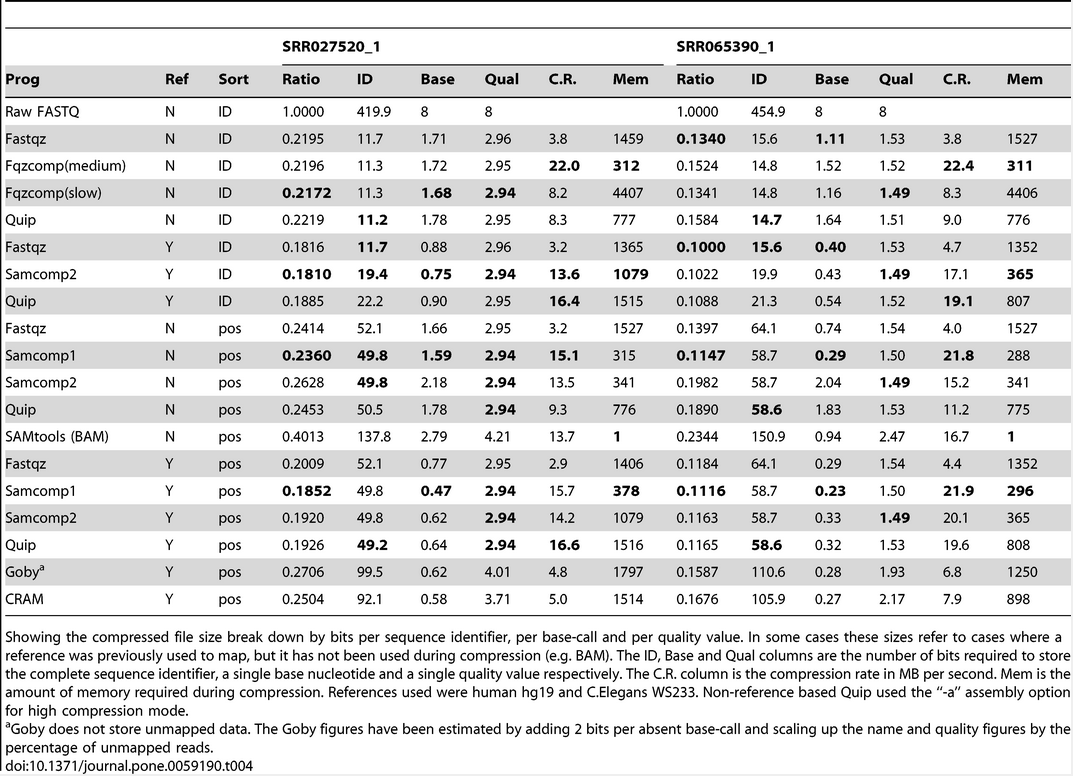
The landscape of existing compression tools can be partitioned in two major classes according to the choice of using an external reference or not. When ordering reads with respect to an external reference, only the relative positions and the differences are encoded in the compressed output, therefore generating highest compression ratios. The limitation of such an approach is that the reference shall be available both at the encoding and decoding side and if not available it shall be transferred with the compressed data (affecting the efficacy of compression).

A recent call for technology has been issued to compare the performance of different approaches to FastQ compression (applicable to FastA as well). This initiative was prompted by the Pistoia Alliance [3] by means of the *SequenceSqueeze* competition.

The results of the comparison have been published in [4], are reported in Table 2, and are accessible at this URL:

<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0059190>.

The list of submitted tools includes both FastQ and SAM compressors. The latter are therefore not exactly compressor of raw sequence data as a pre-processing stage to transform raw sequences to SAM is needed.



*Table 3 – Results of the SequenceSqueeze competition.*

The results of the *SequenceSqueeze* competition show that the best compression ratios are in the order of 20% of the original size for one dataset and 10% for the other.

# Available sequence data

Sequence data for research purposes are published by several organizations around the globe. Among the richest dataset we can mention:

* The 1000 genome project [9] with more than 2000 sequence data from human genomes
* The Gene Expression Omnibus (GEO) repository [10] from the US National Center for Biotechnology Information (NCBI) [11]
* The European Nucleotide Archive (ENA) [12] of the European Bioinformatics Institute (EBI) [13]

The vast majority of sequence data available on these repositories exists in the form of zipped FastQ files with sizes in the order of several GB. All publications describing the principles and functioning of compression tools usually refer to some of the data available from these repositories when measuring performance. Comparison among different works is usually difficult as the chosen dataset are different both in terms of biological characteristic (originating organisms) and sequence technology.

One important step towards a coherent comparison among different compression tools and approaches would be the definition of a shared reference dataset that covers the widest possible range of organisms and sequencing technology.

# Requirements from identified applications

A survey among industrial and academic players in the domain of bioinformatics who manipulate daily large amounts of sequence data files has produced this initial list of requirements for compression of genomic information.

|  |  |  |
| --- | --- | --- |
|  | **Requirement** | **Rationale** |
| R1 | 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 would allow applying newly discovered alignment and mapping approaches to original data. |
| R2 | Compressed data should be structured into data access units. | Access to data shall work in terms of Access Units that would be accessed without the need to decompress the whole genomic information. I.e. selective access and manipulation shall be supported. |
| R3 | 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 released to the market. While Access Units can have a fixed length for a specific application, the standard shall support variable reads lengths. |
| R4 | 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 occurrences of a sample” * “go to a specific location and dump the content”   This implies the ability to support random access to specific positions within a whole genome without the need to read it from the beginning. |
| R5 | 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. |
| R6 | Compressed data formats should be able to integrate annotations – e.g. features on a given region, **non-contiguous** regions of the data structure. | Annotations are metadata that are used for example to indicate relations between reads located at different spatial locations. |
| R7 | The nucleotide information shall support at least the 16 IUPAC ambiguity 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. |
| R8 | Conformance testing shall cover the widest possible range of meaningful biological samples and sequencing technology. | The conformance to the norm shall be defined against a predefined dataset that represents the widest possible range of input data. Compression ratios may heavily vary with respect to the nature of input data. Therefore the conformance dataset shall be selected so as to cover the widest range of meaningful biological samples and sequencing technology. |

# Existing tools versus requirements

This section compares the most popular genomic data compression tools - including the most interesting among those submitted to the SequenceSqueeze contest - against the requirements listed in Section 7. The assessment summarized in Table 4 is built only according to related documents such as the papers accompanying each tool, the SequenceSqueeze contest report [4] or the comprehensive overview compiled by Deorowicz and Grabowski in 2013 [14].

It is worth noting that CRAM and Goby are toolkits containing several tools while BAM is a file formats used by the SAMTools [1] framework.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tool name | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | Reference based | Notes |
| **Raw sequencing data compressors** | | | | | | | | | | |
| Fastqz | Y | N | N | N | N | N | N | N | Y | Optionally accepts a reference as input. Input shall be in Sanger FASTQ format. |
| Fqzcomp | Y | N | N | N | N | N | N | N | N | Needs adaptation to support a wider range of FASTQ |
| Quip | Y | N | N | N | N | N | N | Y | Y | Weak on deep coverage, implements internal alignment. |
| DSRC-1/2 | Y | Y | Y | Y | N | N | Y | Y | N |  |
| **Aligned reads compressors (reference based)** | | | | | | | | | | |
| Quip | Y | N | N/A\* | N | N | N | N | Y | Y | Weak on deep coverage, implements internal alignment. |
| CRAM | Y | Y | Y | Y | N | N/A\* | Y | Y | Y | Toolkit and file format |
| BAM | Y | Y | Y | Y | N | Y | Y | Y | Y | Binary version of SAM. Poor compression ratios. |
| Goby | Y | Y | Y | Y | N | Y | N | Y | Y | Toolkit and file format |

*\*Explicit mention of support of the concerned functionality could not be found*

Table 4 – Summary of the supported features of popular compression tools

# Beyond storage

Efficient genomic information compression can help the scientific community not only by saving transfer time and storage space but also in improving the performance of another critical stage of genomics that is *de-novo* assembly. *De-novo* assembly tries to build (parts of) a genome from raw sequence reads without the help of an external reference. This is usually implemented using de Brujin graphs [15] that might require hundreds of GB of memory during processing. Studies have shown that efficient compression can help in reducing the memory usage of *de-novo* assembly by one order of magnitude [16].

Another application of efficient genome compression is the training of expert models on one specific sequence to be able to use the acquired knowledge to align another similar sequence. The resulting aligners are shown to have a higher quality despite a lower speed [17].

# Conclusions

In a period of investigation for improved sequencing data representations this document aims at providing a rough (even though incomplete) overview of existing tools and approaches for data compression. While performance in terms of compression and speed might be acceptable in some cases, what appears to be missing is a solution that meets at least the requirements listed in section 7. Such a solution 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 what was expected only a few years ago.

The two driving elements of the design process should be on one hand the reuse of existing well-established technologies for representation, compression, storage, access, etc., and on the other hand the flexibility to incrementally address current and future needs without being bound to specific application constraints.

1. **References**

|  |  |
| --- | --- |
| [1] | S. D. Kahn, "On the Future of Genomic Data," *Science,* vol. 331, pp. 728-729, 2011. |
| [2] | S. Wandelt, M. Bux and U. Leser, "Trends in Genome Compression," *Journal of Current Bioinformatics,* 2013. |
| [3] | H. Li, "SAM/BAM and related specifications," [Online]. Available: http://samtools.github.io/hts-specs/. |
| [4] | "Pistoia Alliance," [Online]. Available: http://www.pistoiaalliance.org/. |
| [5] | J. K. Bonfield and M. V. Mahoney, "Compression of FASTQ and SAM Format Sequencing Data," 2013. [Online]. Available: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0059190. |
| [6] | WHO, "WHO Expert Committee on Biological Standardization," [Online]. Available: http://www.who.int/biologicals/expert\_committee/en/. |
| [7] | P. Cock, C. Fields, N. Goto, M. Heuer and P. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acid Research,* vol. 38, no. 6, pp. 1767-1771, 2009. |
| [8] | L. Chen, S. Lu and J. Ram, "Compressed pattern matching in dna sequences," in *Proceedings of IEEE Computational Systems Bioinformatics Conference*, 2004. |
| [9] | A. Moffat and J. Larsson, "Off-line Dictionary Based Compression," in *Proceedings of the 1999 Conference on Data Compression*, 2000. |
| [10] | "1000 Genomes," [Online]. Available: http://www.1000genomes.org/. |
| [11] | "Gene Expression Omnibus," [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/. |
| [12] | "The National Center for Biotechnology Information," [Online]. Available: http://www.ncbi.nlm.nih.gov/. |
| [13] | "European Nucleotide Archive," [Online]. Available: https://www.ebi.ac.uk/ena. |
| [14] | "European Bioinformatics Institute," [Online]. Available: https://www.ebi.ac.uk/. |
| [15] | S. Deorowicz and S. Grabowski, "Data compression for sequencing data," *Algorithms for Molecular Biology,* vol. 8, no. 25, 2013. |
| [16] | P. Compeau, P. Pevzner and G. Tesler, *How to apply de Bruijn graphs to genome assembly,* vol. 29, 2011, pp. 987-991. |
| [17] | T. Conway and A. Bromage, "Succinct data structures for assembling," *Bioinformatics,* vol. 27, no. 4, pp. 479-486, 2011. |
| [18] | M. Cao, T. Dix and L. Allison, "A genome alignment algorithm based on compression," *BMC Bioinformatics,* vol. 11, no. 1, p. 599, 2010. |
| [19] | "SamTools," [Online]. Available: http://samtools.sourceforge.net/. |
| [20] | M. H.-Y. Fritz, R. Leinonen, G. Cochrane and E. Birney, "Genome Research," 2011. [Online]. Available: http://genome.cshlp.org/content/21/5/734.long. |
| [21] | "Cram Toolkit," ENA European Nucleotide Archive, [Online]. Available: http://www.ebi.ac.uk/ena/software/cram-toolkit. |
| [22] | J. Bonfield, "sam\_comp," [Online]. Available: http://sourceforge.net/projects/samcomp/. |
| [23] | F. Campagne, K. C. Dorff, N. Chambwe, J. T. Robinson, J. P. Mesirov and T. D. Wu, "Compression of structured high-throughput sequencing data," 2013. [Online]. Available: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0079871. |
| [24] | P. Li, X. Jiang, S. Wang, J. Kim, H. Xiong and L. Ohno-Machado2, "HUGO: Hierarchical mUlti-reference Genome cOmpression for aligned reads," *Journal of the American Medical Informatics Association,* 2014. |
| [25] | B. G. Chern, I. Ochoa, A. Manolakos, A. No, K. Venkat and T. Weissman, "Reference Based Genome Compression," in *IEEE ITW - IEEE Information Theory Workshop*, Lausanne, 2012. |
| [26] | S. Grabowski, S. Deorowicz and Ł. Roguski, "Disk-based genome sequencing data compression," *Bioinformatics,* 2014. |