ICT for Health Science Research
A. Shabo (Shvo) et al. (Eds.)

doi:10.3233/978-1-61499-959-1-75

Adding Security and Privacy to Genomic Information Representation

Jaime DELGADO, Silvia LLORENTE and Daniel NARO
Information Modeling and Processing - Distributed Multimedia Applications Group (IMP - DMAG),
Computer Architecture Dept. (DAC),
Universitat Politècnica de Catalunya (UPC)

Abstract. Provision of security and privacy to genomic data is a key issue in current genomic information representation. Existing formats do not give a solution to these issues (or they provide a partial one), so new solutions are demanded. MPEG-G (ISO/IEC 23092, Genomic Information Representation) is an International Standard for the representation of genomic information being defined by the MPEG Committee (Moving Pictures Expert Group, ISO/IEC JTC1 SC29/WG11). We provide flexible protection to the information stored inside the MPEG-G format with a combination of security techniques and privacy rules.

Keywords. Security, Genomic information, Privacy rules, Protection, MPEG-G

1. Introduction

Sequencing complete genomes is becoming quicker and cheaper. The medical advantages of precision medicine are clear, and, as sequencing techniques have improved, many health professionals request genomic analysis for early detection of diseases and also to provide specific treatments to known (and future) medical conditions.

This situation implies that an increasing amount of genomic information needs to be stored with the consequence that the institutions doing so are starting to run out of space and budget. On the other hand, due to the special nature of the genomic information, i.e., it uniquely identifies a person and his/her blood relatives, protection of privacy of genomic information is a key issue when storing and giving access to it.

To do so, the use of cryptographic techniques, the management of genome owner consents and other protection techniques are a starting point to guarantee privacy and protection. These are good news for genome owners, but sometimes an obstruction to the researchers' task, as they have to re-ask for permission in order to perform new studies over some genomic information, or the genomic information has to be removed after the required research analysis has been done.

We have presented in [1] how security and privacy, and other features, can be applied to MPEG-G [2], a new genomic information representation (GIR) format being currently standardized inside the MPEG Committee [3].

This paper starts by presenting how security and privacy features are currently considered in different GIRs. Then, we describe the proposed techniques that have been included in MPEG-G [2], a standard that not only takes into account these aspects but also describes compression, how to perform metadata representation and defines a

hierarchy to represent genomic information. Inclusion of security and privacy features was considered in this format from the beginning of its design, based on our previous contributions [4] [5].

2. Methods: Security aspects in current genomic information representation

The representation of genomic information is currently done with formats like FASTA/FASTQ [6], which store the raw data coming from the sequencing machines, Sequence Alignment / Map format (SAM) [7] that stores aligned genomic information, and formats to list the known variants, such as the Variant Call Format (VCF) [8]. All of them are text-based formats, what gives room to compression techniques. Some of them already provide a compressed version, like Binary Alignment Map (BAM), that is a compressed version of SAM [7], or BCF that is a compressed VCF [8], but new and more complex requirements are being demanded by research centers when accessing and processing genomic information. These text-based formats share a common issue that is that security and privacy were not considered when they were designed, as their sole aim was to store the information coming from sequencing machines without losing information. This is an important point when storing genomic information, because, as the knowledge over genome regions meaning evolves, new analysis can be conducted in order to find previously unknown medical conditions.

On the other hand, security aspects are being under consideration also on different fronts. For instance, inside the Global Alliance for Genomics and Health (GA4GH) [9], there is an initiative to define a Security Technology Infrastructure [10], led by the Data Security Work Stream. This document describes how to implement a security infrastructure around genomic information. Given the fact that the GIR formats considered in GA4GH are SAM/BAM and Compressed Reference-oriented Alignment Map (CRAM) [11], it is expected that the Data Security Work Stream will present solutions to provide security for these formats.

Nevertheless, other specific privacy solutions exist, such as Selective retrieval on Encrypted and Compressed Reference-oriented Alignment Map (SECRAM) [12]. It defines encryption techniques for aligned data with the final aim of providing privacy with a granularity down to the nucleotide when accessing to genomic information. SECRAM has lower compression performance than CRAM due to the way data is represented.

Finally, the MPEG Standardization Committee (ISO/IEC JTC1 SC29/WG11) started working in genomic information representation in 2014. The result is MPEG-G [2], which defines a new GIR providing security and protection, privacy-by-design, compression and direct access, among other features, to genomic information. These features, to the best of our knowledge, are not provided by any other GIR nowadays. This format, and how security and privacy is applied, is described in detail in section 3.

3. Results: MPEG-G and its security and privacy

MPEG-G, ISO/IEC 23092 [2], currently has five parts, further described in [13]. They are Transport and Storage of Genomic Information, Coding of Genomic Information, Genomic Information Metadata and Application Programming Interfaces (APIs), Reference Software and Conformance.

Among other things, MPEG-G defines a file format that represents genomic information hierarchically, as can be seen in the left side of Figure 1. The decision of having this file structure came from several proposals, done by different research organizations. In particular, we proposed in [4] [5], a possible hierarchy for genomic information, based on existing formats combined with the requirements coming from research institutions, universities and other actors in the genomic information value chain.

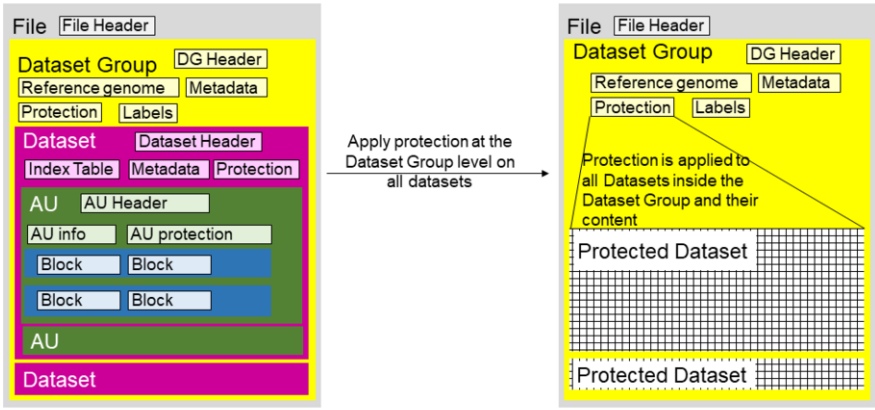

Figure 1. MPEG-G File Format.

To construct the hierarchy, several levels are considered: Genomic Study (Dataset Group in MPEG-G), Genomic Dataset (Dataset in MPEG-G), Genomic Records (Access Unit in MPEG-G) and Genomic Data (Block in MPEG-G). At each level, protection elements convey the encryption and authentication strategies. They specify the encryption and signature parameters (e.g. cipher used or key reference) for the other boxes of the same level, and the protection boxes of the layer below. The protection element at the genomic record level also specifies if and how the encoding units (i.e. Block in Figure 1) are encrypted. The right side of Figure 1 shows a case where the protection is applied to all Datasets contained inside the Dataset Group. By combining these features, it is possible to specify with high granularity the data to protect: in one dataset one can encrypt metadata whilst a different region encrypts genomic data. Figure 2 shows the structure of the Protection Box.

The security strategies in MPEG-G provide 1) confidentiality and integrity to user-specified elements in the file and 2) access control for specific methods of the API. They are defined in Part 3 of the Standard. The confidentiality and integrity methods are based on usual cryptographic strategies such as AES [14], RSA [15] or Elliptic curves [16]. At the different hierarchy levels of the file (Dataset Group, Dataset, or Access Unit) the protection element is used to provide a list of the security strategies employed to protect the selected subset of elements within that layer instance (header information, metadata, data blocks), or the protection elements of a layer below. The list is represented in XML form and provides the parameters needed to decrypt and/or check the signature of the elements, which are identified through URLs. The keys used in the security strategies are identified through names. There must be an external key transport channel, or the value of the key can be derived from other information: either applying pbkdf2 [17] (using a shared knowledge as password), or unwrapping an encrypted key. These options can be leveraged to reduce the number of keys to transport. For example, in a Dataset group with three datasets, if the recipient is allowed to read only one dataset, only the

key for that dataset is transmitted, but if the authorization is for the 3 datasets, only a general key could be shared from which the 3 specific keys could be derived.

```
<protection_box>
        <Encryptions>
                <EncryptedData>...</EncryptedData>
                <EncryptedData>...</EncryptedData>
        </Encryptions>
        <Privacy_rules>...</Privacy_rules>
        <Signatures>
                <Signature>...</Signature>
                <Signature>...</Signature>
        </Signatures>
</protection_box>
```

Figure 2. Protection Box Structure.

The privacy rules, which act as access control rules, are based on the API: the rules indicate under which conditions each action can be executed for accessing genomic regions, datasets or dataset groups that have been previously protected. The rules are conveyed using the eXtensible Access Control Markup Language (XACML) [18] specification. The party distributing the file can thus place such conditions as to limit the access to a specific region on a specific chromosome to users having a specific role, for example researchers. The access control rules do not have any means to enforce that they are being followed, thus access control rules should be coupled to confidentiality strategies. When access to one of these protected information structures is requested, the corresponding rule has to be evaluated in order to decide if access is granted or not. Some initial thoughts on this mechanism were presented in [19], where we demonstrated how privacy rules were able to provide access control to parts of a SAM file according to different known parameters. We have applied the same concepts to MPEG-G file format, defining access to different parts of the file according to roles, actions and conditions.

4. Discussion

It is worth noting that MPEG-G is being defined using privacy-by-design principles from its beginning. The information stored inside MPEG-G files needs to be protected in order to prevent unauthorized access due to its sensitive nature. Moreover, MPEG-G goes a step further than current state of the art GIR, as it conveys information on the consent of the user, alongside rich and protected metadata.

On the other hand, the security and privacy approach considered in MPEG-G could be applied to other existing GIR, as presented in [19]. In this way, one could take the advantage of protecting genomic information whilst maintaining its current data format. Nevertheless, this option is not recommended, as the other extra features from MPEG-G would be lost, although it could be a transitory situation before moving to the new representation. Finally, MPEG-G also defines a transport format, in order to allow efficient streaming of genomic information.

5. Conclusions and future work

Security and privacy are key issues inside genomic information representation. The sensitive nature of the stored data, which not only affects to the individual whose

information is represented but also to her relatives, needs specific solutions in this direction.

In this paper, we have presented how MPEG-G, a new GIR being standardized by the MPEG Standardization Committee, implements some of our research results on privacy and security. However, MPEG-G not only provides security and privacy techniques but also other interesting features for the storage of genomic information, being able to achieve higher compression ratios than other existing formats, and direct access to genomic information regions.

The security and privacy approach based on boxes describing encryption, signature and privacy rules provides a very flexible solution to selective encryption of genomic information.

Acknowledgements

This work is partly supported by the Spanish Government (GenCom, TEC2015-67774-C2-1-R) and by the Generalitat de Catalunya (2017 SGR 1749).

References

[1] Jaime Delgado, Privacy, metadata and APIs in compressed genomic information: The MPEG-G case, GA4GH & MPEG Genome Compression Workshop, https://drive.google.com/file/d/14Y7qK5TmRM5b_5_F5x8UZfwJX3tDS11d/view?usp=sharing, 2018.
[2] ISO/IEC, ISO/IEC 23092, *MPEG-G, Genomic Information Representation*, https://mpeg.chiariglione.org/standards/mpeg-g, 2018.
[3] ISO/IEC JTC 1/SC 29/WG 11, *Moving Picture Experts Group (MPEG),* http://mpeg.chiariglione.org.
[4] M39175, GENIFF (GENomic Information File Format), a proposal for a Secure Genomic Information Transport Layer (GITL) based on the ISO Base Media File Format, http://dmag.ac.upc.edu/downloads/mpegg/GENIFF.pdf, 2016.
[5] M39940, *GENIFF v2*, http://dmag.ac.upc.edu/downloads/mpegg/m39940-GENIFFv2.pdf, 2017.
[6] Peter J. A. Cock, et al., *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants*. Nucleic Acids Res. 2010 Apr; 38(6): 1767–1771.
[7] *Sequence Alignment / Map (SAM) Format Specification*, https://samtools.github.io/hts-specs/, 2018.
[8] *Variant Call Format (VCF) Format Specification*, https://samtools.github.io/hts-specs/, 2018.
[9] Global Alliance for Genomics and Health, https://www.ga4gh.org/, 2018.
[10] GA4GH, Security Technology Infrastructure, https://www.ga4gh.org/wp-content/uploads/2016May10_REV_SecInfrastructure.pdf, 2016.
[11] CRAM format specification (version 3.0), https://github.com/samtools/hts-specs/blob/master/CRAMv3.pdf, 2018
[12] Zhicong Huang et al., A privacy-preserving solution for compressed storage and selective retrieval of genomic data, Genome Res. 2016. 26: 1687-1696, doi: 10.1101/gr.206870.116, https://genome.cshlp.org/content/26/12/1687.full, 2016.
[13] Claudio Alberti et al., An introduction to MPEG-G, the new ISO standard for genomic information representation, https://www.biorxiv.org/content/early/2018/10/08/426353, 2018.
[14] ISO/IEC 18033-3:2010 Information technology - Security techniques - Encryption algorithms - Part 3: Block ciphers, 2010.
[15] Public-Key Cryptography Standards (PKCS) #1: RSA Cryptography Specifications Version 2.1, 2003.
[16] Elliptic curve SEC 1: Elliptic Curve Cryptography Certicom Research, 2009.
[17] PBKDF2 PKCS #5: Password-Based Cryptography Specification Version 2.0, 2000.
[18] OASIS, eXtensible Access Control Markup Language (XACML) V 3.0 Errata 01, http://docs.oasis-open.org/xacml/3.0/errata01/os/xacml-3.0-core-spec-errata01-os-complete.html, 2017.
[19] Jaime Delgado et al., Protecting Privacy of Genomic Information, Studies in Health Technology and Informatics, 2017. Volume 235 (318-322).