Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-318

Protecting Privacy of Genomic Information

Jaime DELGADO[1], Silvia LLORENTE and Daniel NARO
Distributed Multimedia Applications Group (DMAG), Computer Architecture Dept. (DAC), Universitat Politècnica de Catalunya (UPC)

Abstract. The ISO/IEC committee in charge of standardizing the well-known MPEG audiovisual standards has launched, in cooperation with the ISO committee on Biotechnology, a new activity for efficient compressed storage and transmission of genomic information. The paper presents proposals for adding privacy and security to such in-progress standards.

Keywords. Privacy, Security, Genomic Information

1. Introduction

Nowadays, there is an increasing amount of genomic information being generated for different purposes: research, genetic analysis, precision medicine, search of relatives, forensics, etc. This means that the information generated will probably need to be stored, transferred and, ideally, protected against attacks.

On the one hand, several attacks have already been identified, usually to find out if an individual's genome is inside a set of genomes [1], but also to infer regions of the genome which have to remain private [2].

On the other hand, there is an initiative of the Moving Picture Experts Group (MPEG) Standardization Committee [3] for the representation of genomic information in a compressed way, including security and privacy aspects.

The authors have presented several proposals (such as [4]) as answer to the MPEG Committee call for proposals [5], but we concentrate on the one that includes the provision of security and privacy mechanisms to support the storage and transmission of genomic information.

In the next sections, we describe the MPEG Committee activities on genomic information, our proposal for security and privacy and some conclusions and future work.

2. Methods: The Moving Pictures Experts Group (MPEG) Work on Genomics

The Moving Picture Experts Group (MPEG) [3] is a working group of ISO/IEC (ISO/IEC JTC 1/SC 29/WG 11). Since 1988, the group has produced standards for coded representation of digital audio and video and related data.

[1] Corresponding author, Jaime Delgado, Universitat Politècnica de Catalunya, Jordi Girona, 1-3, 08034 Barcelona, Spain; E-mail: jaime.delgado@ac.upc.edu.

Based on their successful previous experience in audiovisual content compression, a new initiative inside MPEG to provide compression mechanisms for genomic information was started two years ago. After a detailed process of obtaining requirements, a call for proposals was launched in July 2016 [5] in order to provide solutions to the genomic information compression and representation problem. The 15 answers to this call have been discussed in the 116th meeting held in Chengdu in October 2016 [6].

The paper focuses on the proposal presented by a group of organizations and companies led by the authors [4]. The contribution responds to the transport requirements [7], defining a format for storage and transport of genomic information, including aspects like metadata, definition of security mechanisms and application of privacy rules.

3. Results: Privacy and Security in Genomics contributed to MPEG

In [4], we propose a format (that we call GENIFF, for GENomic Information File Format) based on ISOBMFF [8] to support the inclusion of compressed and uncompressed genomic information. Moreover, we also provide mechanisms to include metadata associated to the information stored inside the format. Metadata elements may apply to a study, an individual, partial or complete genomic information, etc.

Apart from generic metadata describing things like the genetic study done, the machine used for generating the genomic information and other kind of metadata defined by European Bioinformatics Institute (EBI) [9], we have focused on providing placeholders to apply security and privacy to the genomic information. In this way, only authorized users will be able to access to the information contained in the file. Security and privacy elements may apply to different levels, like the complete study or only to specific information, for example, a SAM / BAM file [10] or a Variant Call File (VCF) [11]. In this way, we can provide a high level of flexibility.

3.1. Privacy Provision, Use of XACML to Authorize Access to Genomic Information

In order to provide privacy when accessing genomic information, we use XACML [12] rules. First ideas on how to use XACML where already presented to the MPEG Committee in [13] and are reflected in our proposal for representing genomic information [4]. We are also using XACML rules to provide privacy protection when accessing medical information, as described in [14]. The rules for controlling access to genomic information may include, among other, the following information:

- who is able to access to the genomic information (user roles or individuals);
- what information can be accessed (the complete file, a chromosome, etc.);
- when it can be accessed;
- with which purpose (genetic analysis, anonymized study, etc.);
- if the data provider has to be informed when information is accessed;
- which permission is given (`viewFile`, `viewChromosome`, etc.).

The rules can be included inside our file format (GENIFF), or even inside the existing formats for genomic information, like SAM or BAM [10].

The inclusion of the rules inside the genomic files allows us to extract them and authorize access to the file according to the permissions defined in the rules. The workflow of operations required to authorize the access to genomic information is shown in Figure 1. It works as follows:

1. A user requests, to a repository, access to some specific genomic information. To do so, an access request is sent including, among other, the following context information: user or role requesting access, the operation requested, the time when the request is done, which information wants to be accessed and the level of granularity (i.e. complete file, chromosome, etc.).
2. The genomic information repository extracts the rule(s) from the requested genomic information and asks, to the Authorization point, for user authorization according to the rule(s) and the request received.
3. The authorization result could be Permit, Deny or Not Applicable.
4. In the case of Permit, the access is permitted, so the requested genomic information should be decrypted and given to the user from the repository.

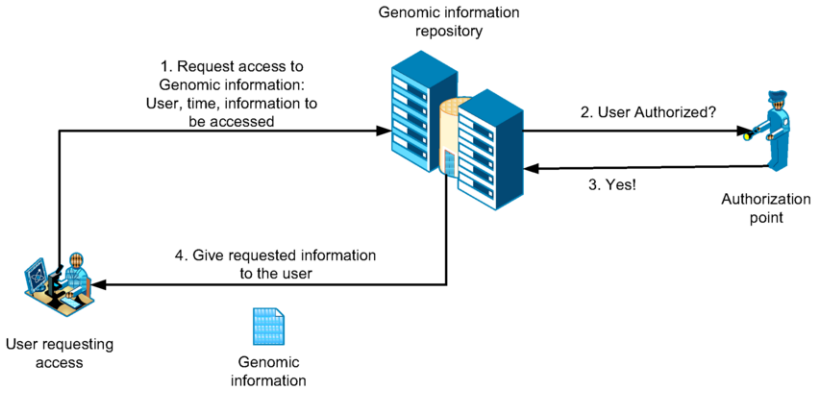

Figure 1. Genomic data authorization flow.

Figure 2 shows a snippet of an XACML rule giving VIEW permission to a physician for the file `genomic-file.sam`.

3.2. Encrypting Portions of the Data

When securing genomic information, one of the preferred strategies is to aggregate results of multiple individuals and to only respond to queries concerning the whole pool of data. Such queries might be, for example, the frequency of a certain mutation in the studied population. This is intended to ensure the privacy at the individual level, although this approach was proven to have some flaws [1].

The result of MPEG's work has to be a format for genomic information, where the encryption of the data is used as a way to enforce the privacy rules. Certain regions of the DNA can be considered as safe to make them public, while others require a key to decrypt the content. By encrypting those regions, we ensure that the file can be transmitted, and the access to the content can vary depending on the access the user asks for. The actual policy to decide which regions should be protected is left to the user creating the file, but one approach might be, as in [2], encrypting the regions to remain private, but also include those allowing inference attempts.

The encryption is done on well-defined regions of the DNA, for example for a given chromosome, using symmetric encryption. Homomorphic encryption might add the benefits to enable secure computations on encrypted content (e.g. [15]); however, this encryption methodology appears to be tightly bounded to the intended usage, thus limiting broad applicability. Therefore, our proposed format uses symmetric encryption, leveraging on the employed codec for better granularity (e.g. BAM's blocks).

```
<Rule RuleId="urn:oasis:names:tc:xacml:3.0:RuleSAM" Effect="Permit">
  <Description> A physician may view the genomic information file for which he or
    she is the designated primary care physician
  </Description>
  <Target>
    <AnyOf>
      <AllOf>
        <!-- Which kind of user: physician -->
        <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
          <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
            physician
          </AttributeValue>
          <AttributeDesignator MustBePresent="false"
            Category="urn:oasis:names:tc:xacml:3.0:role" AttributeId="role"
            DataType="http://www.w3.org/2001/XMLSchema#string"/>
        </Match>
        <!-- Which resource -->
        <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:regexp-string-match">
          <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
            genomic-file.sam
          </AttributeValue>
          <AttributeDesignator MustBePresent="false"
            Category="urn:oasis:names:tc:xacml:3.0:attribute-category:resource"
            AttributeId="urn:oasis:names:tc:xacml:1.0:resource:resource-id"
            DataType="http://www.w3.org/2001/XMLSchema#string"/>
        </Match>
        <!-- Which action  -->
        <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
          <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
            VIEW
          </AttributeValue>
          <AttributeDesignator MustBePresent="false"
            Category="urn:oasis:names:tc:xacml:3.0:attribute-category:action"
            AttributeId="urn:oasis:names:tc:xacml:1.0:action:action-id"
            DataType="http://www.w3.org/2001/XMLSchema#string"/>
        </Match>
      </AllOf>
    </AnyOf>
  </Target>
</Rule>
```

Rule effect,

Authorized role:

Resource: genomic-

Permission

Figure 2. XACML rule giving VIEW permission to a SAM file.

4. Discussion, Conclusions and Future Work

We have developed software to demonstrate the feasibility of the format and its granular access control. We have implemented some policy rules with XACML and used real SAM/BAM files to evaluate the encryption of content for supporting the access control. The results are satisfactory for the moment.

Our next step is aligned with the standardization process. Some of the presented ideas have been initially accepted for consideration by the MPEG Committee. Now it is time to perform some formal tests to check their feasibility. For this purpose, a set of experiments has been defined [16] and the results have to be presented for discussion at the next MPEG meeting, in January 2017.

In particular, we will further analyze and implement the presented approach in experiment 4, which will check that the genomic information representation format defines the proper tools for allowing selective access to the content for both data storage and transfer. In addition, these tools should support access control (privacy and security) and high-level metadata applying to different genomic information contained in the format. We will particularly focus on providing tools for privacy, security and metadata definition inside the genomic information.

Acknowledgements

The work presented in this paper has been partially supported by the Spanish Government under the project: Secure Genomic Information Compression (GenCom, TEC2015-67774-C2-1-R).

References

[1] Nils Homer et al., *Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays*. PLoS genetics, 4(8):e1000167, 2008.
[2] Nyholt, Dale R and Yu, Chang-En and Visscher, Peter M. On Jim Watson's APOE status: genetic information is hard to hide. European journal of human genetics : EJHG march 2009.
[3] ISO/IEC JTC 1/SC 29/WG 11, *Moving Picture Experts Group (MPEG),* http://mpeg.chiariglione.org/.
[4] MPEG2016/M39175, *GENIFF (GENomic Information File Format), a proposal for a Secure Genomic Information Transport Layer (GITL) based on the ISO Base Media File Format*, Chengdu, October 2016.
[5] ISO/IEC JTC 1/SC 29/WG 11 - ISO/TC 276/WG 5 MPEG2016/N16320, *Joint Call for Proposals for Genomic Information Compression and Storage*, Geneva, June 2016.
[6] MPEG2016/M38890, *AHG on Requirements on Genome Compression and Storage*, Chengdu, October 2016.
[7] ISO/IEC JTC 1/SC 29/WG 11 - ISO/TC 276/WG 5 MPEG2016/N16323, *Requirements on Genomic Information Compression and Storage*, Geneva, June 2016.
[8] ISO/IEC IS 14496-12, *Information technology - Coding of audiovisual objects - Part 12: ISO base media file format*, Fifth edition, December 2015.
[9] European Bioinformatics Institute (EBI), http://www.ebi.ac.uk/, 2016.
[10] *Official Sequence Alignment/Map (SAM) Format Specification*, https://samtools.github.io/hts-specs/, 2016.
[11] *Official Variant Call Format (VCF) Format Specification*, https://samtools.github.io/hts-specs/, 2016.
[12] OASIS Standard, *eXtensible Access Control Markup Language (XACML) Version 3.0*, http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html, January 2013.
[13] Jaime Delgado, Silvia Llorente, MPEG2015/M36405 - *Some application scenarios for privacy and security requirements on genome usage, compression, transmission and storage*, Warsaw, June 2015.
[14] Jaime Delgado, Silvia Llorente, Martí Pàmies and Josep Vilalta, *Security and Privacy in a DACS*, Exploring Complexity in Health: An Interdisciplinary Systems Approach, doi: 10.3233/978-1-61499-678-1-122, IOS Press, 122-126, 2016.
[15] Jung Hee Cheon, Miran Kim, and Kristin Lauter, *Homomorphic computation of Edit Distance*, Pages 194-212 Springer Berlin Heidelberg 2015.
[16] ISO/IEC JTC 1/SC 29/WG 11- ISO/TC 276/WG 5 MPEG2016/ N16526, *Core Experiments on Genomic Information Representation*, Chengdu, October 2016.