Vol.2018-DPS-176 No.10
Vol.2018-EIP-81 No.10
2018/9/13

遺伝子符号化表現の標準化の現状と個人情報保護の検討

金子 格†

ISO/IEC JTC 1/SC 29/WG 11 では 5 パートからなる遺伝子情報符号化ファイルフォーマットの標準化を進めている。同作業グループは MPEG 動画像符号化で有名であり、遺伝子情報の符号化はこれまでの同グループの標準化対象と大きく変わっているが、セキュリティや大量なデータの符号化、構造化において共通点がある。
本発表では、特にプライバシー保護について検討を加える。

Current status of the protection of personal information and in the standardization of MPEG Coded Representation of Genomic Information

ISO/IEC JTC 1/SC 29/WG11 is currently drafting five parts series of international standard for genomic code compression. While the area of standardization is different to the past standardization of the working group, which accomplished standardization of MPEG video coding, there is common technology in both standardization areas, e.g. security, compression of large data, structure of data and privacy management.
This report will review, privacy related specifications within the current openly reviewable specification of the standard.

ITARU KANEKO†

1. Introduction

ISO/IEC JTC 1/SC 29/WG 11 aka MPEG is about publishing its new standard, Genomic Information Representation. Even the standardization of Genomic Information was not included in past MPEG standard, it is within the working area of WG11 because working area is defined as coding representation of multimedia. In same time, in fact, many considerations, problems and technical elements are shared between Genomic Information Representation with representation of other medium.

In this report, we first describe standardization process of Genomic Information Processing, explain structure of the standard and describe possible consideration of the protection of privacy data.

2. Standardization process

ISO/IEC JTC 1/SC 29/WG 11 is the working group under ISO/IEC which work on the standardization of coding representation of multimedia information. It is also known as MPEG which developed dozens of the standards for audiovisual data representation. For example, MP3(ISO/IEC 11172.3), MPEG-2(ISO/IEC 13818.2), MPEG-4(ISO/IEC 14496-x), MPEG AAC(ISO/IEC 13818.7 and ISO/IEC 14496-3), MPEG-HEVC(ISO/IEC 23008.x)

To start new standardization project, such project can be proposed to ISO/IEC upper committee by member body of ISO/IEC or working group such as WG11. Coding of genomic information representation is first considered by ISO TC 276 which is technical committee which are standardization group for bio-technology. ISO TC276/WG5 which work on Data processing and integration, considered standardization of Genomic information processing. And then, since ISOIEC JTC 1/SC 29/WG 11had experience of standardizing wide spectrum of multimedia, TC 276 and WG11 established liaison. Discussion and drafting of

† 東京工芸大学
Tokyo Polytechnic University

the proposal of new project took place among ISO TC276, ISO/IEC JTC 1/SC 29/WG 11 and all relevant body, standardization process had been started in 2017.

3. The structure of the standard

Table 1 shows the structure of ISO/IEC 23092 series which consists of 4 parts. ISO/IEC 23092-1 Transport and Storage of Genomic Information specifies basic and fundamental structure of the genomic information in this standard. ISO/IEC 23092-2 Coding of Genomic Information specifies the coding representation of genomic information itself. ISO/IEC 23092-3 Genomic Information Metadata and APIs specifies metadata which can be embedded in the genomic information in the standard as well as APIs to operate genomic information. ISO/IEC 23092-4 Reference Software specifies the reference software which is most of MPEG standards requires to be standardized to insure the clearness of the specification by providing example software implementation of the specification.

Table 1 The structure of ISO/IEC 23092 series, Genomic Information Representation

ISO/IEC 23092-1 Transport and Storage of Genomic Information
ISO/IEC 23092-2 Coding of Genomic Information
ISO/IEC 23092-3 Genomic Information Metadata and APIs
ISO/IEC 23092-4 Reference Software

4. Coding of genomic data

4.1. Committee Draft and IS

While drafts of those standard is under ballot to be promoted to the next stage, those will be also sent to liaison standardization body e.g. ISO/TC276. And these can be also made to the public for the reviewing, WG11 is usually do make them public. Therefore, everyone interested in can review those specifications. However, the final specification can be only available from ISO web site and various ISO member organization in each Country.

Referring CD 23092-2 (which means committee draft of 23092-2 Coding of genomic information), data format is defined as following.

Secton 1 to 6 of the standard includes various information to describe specification e.g. terms and definitions, representations etc. Section 7 is describing coding form of the genomic information.

4.2. Basic element of gnomic information

It is essential to understand basic element of the genomic information to be stored. If the problem is how to describe “genomic element” which is four code ATGC as well known, but it is not the case. We need to understand, "sequence read", "pairs" and "sequences". Section 3 of the standard defines those terms and according to committee draft those are defined as following.

“sequence read The readout, by a specific technology more or less prone to errors, of a continuous part of a segment of nucleotides extracted from an organic sample.”

“mate pairs Two reads from the same (long) DNA strand extracted by sequencing machines. The orientation is the opposite of paired ends. “

“paired ends A couple of reads produced from the same (short) DNA fragment by sequencing both ends. The orientation is the opposite of mate pairs.”

Since today DNA analysis is processing DNA by first split them to many small pieces then analyze sequence nucleotides of those splited DNA. Those massive amount of sequence is called "sequence read".

To determine original single sequence, reads are then compared to the other reads and reference sequence. By finding those matching original complete sequence will be determined. But there are always some ambiguity and possibility of errors. Therefore there is needs to store and transmit all data derived form the acquisition process, as it is just as they had derived from the acquisition process.

Section 8 of the standard describe such format which describe reads of sequence as an basic unit of the representation and then has flexibility to describe reads as sequence of nucleotides, read id (of same sequence) and pointer to the reference sequence.

Level of confidence and precision of those information to determine actual sequence is jus a matter of performance of the sequencing device. If more reads will be analyzed, more precise result will be given.

5. Considering need for anonymization

MPEG-G is valuable new challenge for both WG11 activity and medical information processing. And MPEG-G version 1 had made great achievement to bring benefit of state the

art media information processing technology to the area of genomic information processing.

To consider further progress of the standard, secondary use of genomic information for the medical research and protection of privacy information is the important matter.

Privacy regulation is well established and sophisticated rules were agreed and shared. Secondary use of medical information is difficult matter. Owner's medical care, which is recognized primary use of medical information usually do not have a problem. Accumulation of genomic information and use of statistical analysis is recognized as secondary use and will not be allowed without complying related regulations.

One of the emerging but already established technique is called anonymization. This is a method to convert medical information to another data form, and with that data, none of personal data is identified, while preserving all useful medical statistical attributes. Support of anonymization is valuable additional function for the MPEG-G.

To study usefulness and technical problem to support anonymization, this contribution explain several related information.

The objective of this contribution is to share better understanding of the purpose and technical challenge of anonymization in case of MPEG-G.

5.1. Basic mechanism of anonymization

k-anonymity of release of data is set of data such as information for each person contained in the release cannot be distinguished from at least (k – 1) individuals. In other words, for an example, if one data chosen in 10000 anonymity data set, we can say it is one of the 10000 people but any further identification is not possible.

5.2. Case study of the clinical study of large medical data, ALLSTAR project

This is one successful case of anonymization of big medical data. In this case, large size medical database of 24-hr ambulatory electrocardiogram was built. The project called ALLSTAR. a 24-hr ambulatory electrocardiogram database of 81,615 males and 103,038 females (≥20 yr) from all over Japan. With this database, we examined if regional differences in heart rate (HR) and HR variability (HRV) are associated with the inter-prefecture rankings of healthy life expectancy (HALE) and of average life expectancy (ALE) in Japan. According to reports by the Japanese Ministry of Health, Labor and Welfare (2013), subjects in each sex were grouped into short, middle, and long HALE and ALE tertiles by their living prefectures. Standard deviation of 24-h normal-to-normal R-R intervals (SDNN) increased progressively with increasing HALE tertiles in both sexes (Ps < 0.001), while it showed no consistent associations with ALE. Conversely, HR decreased progressively with increasing ALE tertiles in females (P < 0.001), while it showed no consistent association with HALE.

These study shows practicability and benefit of anonymization of large medical data.

5.3. National EHR

EHR, stands for electronic health record, includes medical information and health information. National EHRs are already in service in several countries e.g. Canada, USA, UK, Finland, Estonia, Singapore, Australia, New Zealand and Russia (Moscow). National EHR also started in Japan. In case of Japanese version of National EHR, privacy of medical information is very important and strictly protected. However, secondary use of medical information, such as clinical study, is also recognized important issue and thus included in the scope of this system. And for that purpose, anonymization is included in the basic system structure.

5.4. Considerations of requirements

We currently do not have specific use case nor requirements. However, this document intends to explain generic benefit of having support of anonymization. And also provide hint for architecture to support them.

Obviously MPEG-G can be great platform for the big data usage of genomic information. We hope MPEG-G can develop useful standard for such benefit.

If the consideration shared by the group, we would like to continue working on the use cases and requirements.

6. Further work

Use case of the anonymization of genomic information representation is the next step in the standardization process. Use case is can be described as following.

Genomic information is sensitive privacy information. Therefore, we expect more concern on the protection of genomic information.

Currently when people would like to provide genomic information there is only two choices which is give it all or give nothing. Which is difficult decision if people will have more consideration in the privacy exposure by doing it. Ironically, expansion of the application of genomic information. People need to aware of sickness, various personal attributes even evidence of past crime by relatives. Then combination of several level of anonymization of genomic data will ease people to use this technology.

7. Conclusion

New standard for Genomic Information Representation is about to be published. The standard is designed based on the technological characteristics of the genomic analysis. In same time, author expect more privacy issues on the use and collection of genomic information in the future. Contribution to the standardization will be useful to make the standard to be applicable for the more applications. Further consideration of the anonymization of genomic information representation is beneficial to expand application of the new standard.

Acknowledgement

Emi Yuda , kindly provided the information of example case of anonymization which is described in section 5.2.

References:

[1] MPEG-G ISO/IEC 23092, Genomic Information Representation, https://mpeg-g.org/ (2018)
[2] Itaru Kaneo,Emi Yuda, “Consideration of anonymizations in MPEG-G and brief survey of personal medical data in Japan.” MPEG m43376, (2018)
[3] Yuda E, Furukawa Y, Yoshida Y, Hayano J & ALLSTAR Research Group, Association between Regional Difference in Heart Rate Variability and Inter-prefecture Ranking of Healthy Life Expectancy: ALLSTAR Big Data Project in Japan, Proceedings of the 7th EAI International Conference on Big Data Technologies and Applications (BDTA), Chung-ang University, Seoul, South Korea, November 17-18 (2016) P004
[4] Samarati, Pierangela; Sweeney, Latanya (1998). "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression" (PDF). Harvard Data Privacy Lab. Retrieved April 12, 2017. https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf
[5] YOSHIHARA Hiroyuki, gEHR Project: Nation - wide EHR Implementation in JAPAN, Kyoto Smart city Expo, https://expo.smartcity.kyoto/2016/doc/ksce2016_doc_yoshihara.pdf (2016)