MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-979

Anonymizing Patient Genomic Data for Public Sharing Association Studies

Carlos Fernandez-Lozano[a], Guillermo Lopez-Campos[b], Jose A. Seoane[c], Victoria Lopez-Alonso[d], Julian Dorado[a], Fernando Martín-Sanchez[b], Alejandro Pazos[a]

[a] *Information and Communications Technologies Department, University of A Coruña, Spain*
[b] *Health & Biomedical Informatics Researh Unit, The University of Melbourne, VIC, Australia*
[c] *MRC Centre for Causal Analyses in Translational Epidemiology. University of Bristol, United Kingdom*
[d] *Bioinformatics Unit, Institute of Health Carlos III, Spain*

Abstract and Objective

The development of personalized medicine is tightly linked with the correct exploitation of molecular data, especially those associated with the genome sequence along with these use of genomic data there is an increasing demand to share these data for research purposes. Transition of clinical data to research is based in the anonymization of these data so the patient cannot be identified, the use of genomic data poses a great challenge because its nature of identifying data. In this work we have analyzed current methods for genome anonymization and propose a one way encryption method that may enable the process of genomic data sharing accessing only to certain regions of genomes for research purposes.

Keywords:

Anonymous, Genomics, Data sharing.

Introduction

There is a growing importance in the use of genomic data in clinics and different solutions have been developed to include these data in the clinical records ensuring their privacy and security. Nonetheless genomic data have certain particularities that make them special because it identifies univocally the individual, contains information about the patient like the risk for present and future health conditions but also provides the same information about past, present and futures relatives of the patient. In this context there is an increasing demand in research for the use of genomic data clinically annotated as for example GWAS (Genome Wide Association Studies), but these uses have arisen some privacy concerns since privacy decreases very quickly as the number of genetic variants from a single individual increases, so with a relatively small number of SNPs (Single Nucleotide Polymorphisms) it is possible to identify a patient.

Traditionally to share data extracted from the clinical record these data are anonymized splitting and removing all identifying elements. However, when genomics data are shared this approach based on the removal of patient identifying elements might not be enough since genomic data are an identifier themselves and are the data of interest for the sharing process. Therefore there is an increasing need of solutions enabling the use of this information using anonymization or de-identification techniques [1].

In this project we present an alternative to facilitate genomic data sharing using an encryption schema that allows access to small regions of the genome interested on accessing to particular regions of interest (i.e. genes, variants or coding regions).

Methods

Encrypting genomes into fragments of fixed size (tens of nucleotides) with a one-way encryption algorithm to facilitate their public sharing encrypted. In order to get clear information of a genome can only encrypt information that can match the encrypted, to check if it matches after applying the encryption algorithm. The more rare variants have a genome in a set of loci, the longer the process of decryption. Computational time could increase exponentially with the number of positions in the genome. Given a set of loci, genome fragments will be encrypted, using the possible variants of SNPs and mutations that may occur in these loci and then will compare it with all the available encrypted genomes in the system to look for coincidences. Genome fragmentation is carried out to ensure that trying to simultaneously get access to different areas of the genome with the aim of decrypt it would require a prohibitive time and computational power whereas the information can be accessible for queries such as those required i.e. for pharmacogenomic tests. The encryption tool mechanism is analogous to that one of UNIX passwords.

Genomic data anonymization requires the use of encryption methods with certain characteristics. The objective of the proposed method is not to hide the data and protect it with a shared password but to provide a method that allows decoding of small regions of the genome and makes the decoding process for larger or multiple regions of a genome intractable.

Conclusion

This approach could be used to make genomic data publicly available avoiding some of the inconvenient issues related with a symmetric or public-key encryption and simplifying password management.

References

[1] Greenbaum D, Sboner A, Mu XJ, Gerstein M (2011) Genomics and Privacy: Implications of the New Reality of Closed Data for the Field. PLoS Comput Biol 7(12): e1002278