Privacy-Preserving Processing of Filtered DNA Reads

Maria Fernandes*, Jérémie Decouchant*, Marcus Völp*, Francisco M. Couto† and Paulo Esteves-Verissimo*
* SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg
†LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa

Abstract

The rapid evolution of sequencing technologies promoted a variety of opportunities for genomic data, which have introduced significant performance and privacy challenges. In this manuscript, we introduce the approach our research group has been following, in order to protect users' privacy at low performance cost, with a special focus on the early stages of genomic data processing, which are routinely executed in public clouds. The keystone of our approach is a 'privacy filter' [1], which detects sensitive information (i.e., variants) contained in unaligned sequencing data (reads). When compared to previous works, this filter includes several design innovations that allow it to be more sensitive, accurate, and more widely usable, since it can be applied to the long reads that the current sequencing technologies produce. A direct application of this filter — the excision, or masking out, of all the known sensitive nucleotides of a genome — inspired the development of a novel privacy-preserving alignment algorithm [2], which harmlessly performs alignment of the genome in public clouds, once stripped of the sensitive reads, with just 2% less accuracy than the best known plaintext algorithm (BWA). More recently, we designed a stratification method to classify sensitive information into levels [3], which would replace the classical boolean sensitivity classification.

I. Introduction

Genomic sequencing evolved at an unprecedented rate with the advances of sequencing technologies. To follow the sequencing data production growth, traditional genomic data processing has been routinely executed in public clouds. Read alignment, which determines the position of each unaligned read in the genome, is one key step of this processing. However, privacy concerns stemming from unauthorized genomic data accesses have been highlighted by several privacy attacks [4]–[7]. These attacks stressed the weaknesses of non-cryptographic privacy protection methods [8] when used alone. In addition, few works considered protecting the information contained in raw reads. Cryptographic approaches [9], [10] are often of limited applicability due to their slower performance and the restricted application range.

In order to design a high performance and fully privacy-preserving genomic data processing workflow, we developed a filtering approach [1] that not only identifies the privacy-sensitive information automatically, i.e., possible variants in reads, but does so at the very beginning of the workflow, even before alignment starts. The output of this filter contains reads where sensitive nucleotides are tagged, and can therefore be excised from the payload file and replaced by a neutral placeholder that reveals absolutely no information about the initial content. Consequently, by masking out the sensitive information in reads, the resulting files do not reveal any personal information. However, they can still be aligned with high success, as we showed in [2].

II. Per-Nucleotide Read Classification

The read filter [1] we developed improves on a previous work that detects privacy-sensitive reads [11] in the sense that it is more sensitive and accurate, in particular when filtering long reads, because it relies on a per-nucleotide sensitivity detection, while the previous approach used read classification, which resulted in a high proportion of the human genome being detected sensitive. Ayday et al. [12] detailed a method that protects manually declared variants in aligned reads. Our filter treats raw reads in an automatic way, albeit at the price of generating false positives.

Our filter relies on several dictionaries of DNA sequences that are created based on public databases (e.g., the 1000 Genomes Project). Each dictionary contains genomic sequences where a nucleotide at a precise location is part of a variant (e.g., Dic_0 contains sequences where the first nucleotide is part of a variant). The filtering process can then be described as a sliding window iteration on each read. When a subsequence in a read is detected in a dictionary (using a Bloom filter), we classify as sensitive the corresponding nucleotide in the subsequence. Compared to the previous approach, the results showed that our filter produces 10% of false positives, instead of 60%, and misses less privacy-sensitive nucleotides (i.e., less than 10 nucleotides are not detected per human genome). In addition, our approach is more efficient when dealing with sequencing errors, since it correctly detects 86% of the nucleotides instead of 56%.

III. MaskAl: Privacy-preserving alignment

The read filter we presented can be used to produce masked reads, which are reads where the sensitive parts have been masked out. We therefore studied the impact of masking reads on the alignment step, and designed MaskAl [2], a privacy-preserving alignment approach that relies on masked reads. MaskAl achieves high performance by doing most of the alignment process based on masked reads in public clouds, using plain-text alignment algorithms, and relies on Intel's software guard extensions (SGX), which is a trusted execution environment, to further refine the alignment results based on the full reads' information. Our performance evaluation included accuracy comparison with state-of-the-art alignment algorithms (e.g., BWA, LAST, Balaur), and memory and network bandwidth comparison with existing privacy-preserving alignment approaches.

IV. Sensitivity levels for DNA reads

To further improve performance, we extended the boolean classification of nucleotides (i.e. sensitive or non-sensitive) into levels of sensitivity that we computed based on quantitative and qualitative features of DNA [3]. Simply put, a more sensitive level contains SNPs that reveal more information about their donor (e.g., because they are rarer). We relied on a genomic variant imputation tool to make sure that the sensitivity levels are disconnected, and that an adversary located in a cloud could not infer more sensitive data than the one it could access should an attack be successful. To classify the reads into sensitivity levels, we relied on the filtering approach we developed. In case a nucleotide matches in different sensitivity levels, we classify it with the highest sensitivity level, to ensure adequate protection. Considering the allele frequencies of the 1000 Genomes Project population and the linkage disequilibrium relationships, we obtained that 5% of the 100-bases reads of a genome have very high sensitivity, 23% have high sensitivity and the remaining 72% have low sensitivity.

In summary, whilst we give an innovative mechanism, it is not rigid and can be re-applied and shaped accordingly to the user needs. For example in an Alzheimer's disease study (qualitative feature), we could set as highly sensitive the genes causing the disease, moderately sensitive the ones involved in drugs response and little sensitive the remaining genes. Additionally, different levels of sensitivity require different needs of protection. Finally, the classification into sensitivity levels allows the adaptation of the privacy protection to the criticality of the information while improving performance.

V. Conclusion

We believe to have advanced the s.o.t.a. significantly, by addressing the conundrum 'privacy vs. sharing' in genomic data processing. In fact, the increased speed and price decay of sequencing data production, put high pressure for using high throughput alignment algorithms in cheap but unprotected environments (e.g., clouds) . However, privacy risks due to the manipulation of data in plain-text in the cloud, cast enormous shadows on this approach, and the recent adoption of GDPR measures will make this option untenable. However, the other extreme, like crytographically strong alignment algorithms, provides high privacy protection, but has very limited performance, yielding delays that are unacceptable in a real-world production cycle. Thus, the current challenge, addressed by our results, consists in providing data privacy protection while taking advantage of the storage and computational power of clouds.

We developed a filtering approach [1] to improve the privacy and the performance of alignment using existing algorithms. Our approach reduces false positives to 10%, which should be compared to the 60% of the previous approach. Additionally, we also improve accuracy in presence of errors in the sequences (we detect 86% of the reads instead of the previous 56%). Furthermore, we studied how masked reads could be used in the alignment step to design fast and privacy-preserving algorithms. The alignment scheme we proposed, MaskAl [2], has an accuracy of more than 96% of aligned masked reads, which is close to the 98% of aligned plaintext reads achieved by the BWA algorithm. Regarding computation time, MaskAl is 87% faster than existing privacy-preserving algorithms while it also shows memory and network bandwidth cost improvements. However we showed that masked reads do not affect alignment accuracy while improving privacy. To further improve performance, we designed a method to classify genomic information into sensitivity levels [3], which can be used to apply more efficient and less privacy-preserving algorithms to the reads with lower sensitivity. Future work will include the design of methods based on masked reads in downstream steps of the workflow, such as variant calling.

References

[1] J. Decouchant, M. Fernandes, M. Völp *et al.*, "Accurate filtering of privacy-sensitive information in raw genomic data," *Journal of Biomedical Informatics*, vol. 82, pp. 1–12, 2018.

[2] C. Lambert, M. Fernandes, J. Decouchant *et al.*, "Maskal: Privacy preserving alignment of masked reads using intel sgx," in *SRDS*, 2018.

[3] M. Fernandes, J. Decouchant, M. Volp *et al.*, "Dna-seal: Sensitivity levels to optimize the performance of privacy-preserving dna alignment," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2019.

[4] B. Malin, *Protecting dna sequence anonymity with generalization lattices*. Carnegie Mellon University, School of Computer Science [Institute for Software Research International], 2004.

[5] N. Homer, S. Szelinger, M. Redman *et al.*, "Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays," *PLoS Genet*, vol. 4, no. 8, p. e1000167, 2008.

[6] D. R. Nyholt, C.-E. Yu, and P. M. Visscher, "On jim watsons apoe status: genetic information is hard to hide," *European Journal of Human Genetics*, vol. 17, pp. 147–149, 2009.

[7] M. Gymrek, A. L. McGuire, D. Golan *et al.*, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

[8] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews*, vol. 15, pp. 409–421, 2014.

[9] J. Baron, K. El Defrawy, K. Minkovich *et al.*, "5pm: Secure pattern matching," *SCN*, pp. 222–240, 2012.

[10] Y. Chen, B. Peng, X. Wang *et al.*, "Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds." in *NDSS*, 2012.

[11] V. V. Cogo, A. Bessani, F. M. Couto *et al.*, "A high-throughput method to detect privacy-sensitive human genomic data," in *WPES*, 2015, pp. 101–110.

[12] E. Ayday, J. L. Raisaro, U. Hengartner *et al.*, "Privacy-preserving processing of raw genomic data," in *Data Privacy Management and Autonomous Spontaneous Security*. Springer, 2014, pp. 133–147.