Privacy-Preserving Similar Patient Queries for Combined Biomedical Data

1 Bibliographic Reference

@article{ SBHB19,
author = ”Ahmed Salem and Pascal Berrang and Mathias Humbert and Michael Backes”,
title = ”Privacy-Preserving Similar Patient Queries for Combined Biomedical Data”,
journal = ”Proceedings on Privacy Enhancing Technologies”,
year = ”2019”,
volume = ”2019”,
number = ”1”,
pages: 47 - 67”,
url = ”https://content.sciendo.com/view/journals/popets/2019/1/article-p47.xml” }

2 Abstract

The decreasing costs of molecular profiling have fueled the biomedical research community with a plethora of new types of biomedical data, enabling a breakthrough towards more precise and personalized medicine. Naturally, the increasing availability of data also enables physicians to compare patients' data and treatments easily and to find similar patients in order to propose the optimal therapy. Such similar patient queries (SPQs) are of utmost importance to medical practice and will be relied upon in future health information exchange systems. While privacy-preserving solutions have been previously studied, those are limited to genomic data, ignoring the different newly available types of biomedical data.

In this paper, we propose new cryptographic techniques for finding similar patients in a privacy-preserving manner with various types of biomedical data, including genomic, epigenomic and transcriptomic data as well as their combination. We design protocols for two of the most common similarity metrics in biomedicine: the Euclidean distance and Pearson correlation coefficient. Moreover, unlike previous approaches, we account for the fact that certain locations contribute differently to a given disease or phenotype by allowing to limit the query to the relevant locations and to assign them different weights. Our protocols are specifically designed to be highly efficient in terms of communication and bandwidth, requiring only one or two rounds of communication and thus enabling scalable parallel queries. We rigorously prove our protocols to be secure based on cryptographic games and instantiate our technique with three of the most important types of biomedical data namely DNA, microRNA expression, and DNA methylation. Our experimental results show that our protocols can compute a similarity query over a typical number of positions against a database of 1,000 patients in a few seconds. Finally, we propose and formalize strategies to mitigate the threat of malicious users or hospitals.

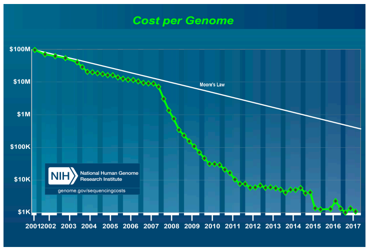
3 Link

https://content.sciendo.com/view/journals/popets/2019/1/article-p47.xml

Privacy–Preserving Similar Patient Queries for Combined Biomedical Data

Ahmed Salem[1], Pascal Berrang[1], Mathias Humbert[2], and Michael Backes[1]
[1]CISPA – Helmholtz Center for Information Security [2]Swiss Data Science Center

Motivation

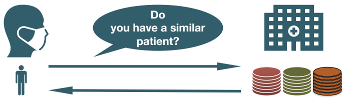


> "8 out of 10 people with a known cancer risk gene mutation don't know they have it, despite them frequently engaging with the healthcare system"

2018 Study by Geisinger's MyCode Community Health Initiative

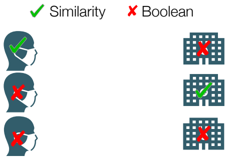
- Cost of data sequencing
- GA4GH (MatchMaker Exchange)
- The effect of data sharing
- The usage of DNA data in different industries like sports

Requirements



- From genomics to omics data
- Different metrics to measure the similarity
- Data privacy
- Specific positions
- Efficient
- Low communication cost

Modes and Metrics



- Euclidean Distance $\sqrt{\sum_{i\in I}(x_i - y_i)^2}$
- Pearson's correlation coefficient

$$\frac{n\sum_{i\in I} x_i y_i - \sum_{i\in I} x_i \sum_{i\in I} y_i}{\sqrt{n\sum_{i\in I} x_i^2 - (\sum_{i\in I} x_i)^2}\sqrt{n\sum_{i\in I} y_i^2 - (\sum_{i\in I} y_i)^2}}$$

Ciphertext Squaring

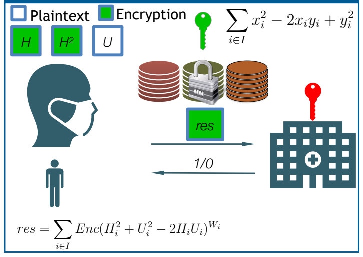
Encryption

1. Let *Enc(m) = c*
2. Pick a random number *a* ← *M*
3. Calculate *β = Enc(m-a)*
4. Calculate α = *Enc(-a² +2am)*
5. Return *(α, β)*

Decryption

Dec(α)+Dec(β)².

SPQ Using Pearson's Correlation Coefficient

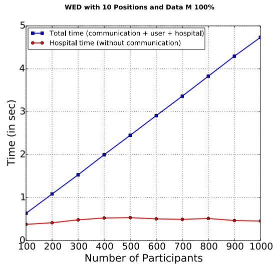


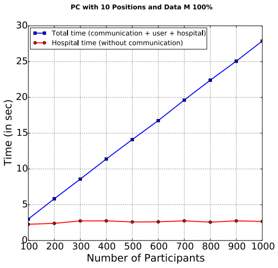
$$\sum_{i\in I} x_i^2 - 2x_i y_i + y_i^2$$

$$res = \sum_{i\in I} Enc(H_i^2 + U_i^2 - 2H_iU_i)^{W_i}$$

Protocol	Data sent by the user	Data sent by the hospital
ED & WED	802 Bytes	47 Bytes
PC	3594 Bytes	849 Bytes

Performance





- All Schemes are proved secure against honest but curious adversaries, with extensions to be secure against fully malicious adversaries
- A typical number of positions of interest can be queried for 1,000 patients in less than 5 seconds for the weighted Euclidean distance and in less than 30 seconds for the Pearson correlation coefficient
- ability to query a specific subset of data and positions with different weights.

ahmed.salem@cispa.saarland
https://ahmedsalem2.github.io/

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY