Protecting Genomic Privacy in Medical Tests using Distributed Storage

Maitraye Das[1], Sharmin Afrose[2], Tanzima Hashem[3]
[1,2,3]Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
[1]0905052.md@ugrad.cse.buet.ac.bd, [2]0905028.sa@ugrad.cse.buet.ac.bd,
[3]tanzimahashem@cse.buet.ac.bd

ABSTRACT

With the increasing utilization of genomic data in various medical tests, privacy of an individual is currently undergoing potential risks like genetic discrimination etc. In this work, we propose a secure system using distributed databases for storing genomic data to calculate disease risks of a patient. Our scheme reduces computational overhead compared to present cryptography-based approaches. Furthermore, we offer substantial development over existing methods regarding multiple disease risk queries.

1. INTRODUCTION

Genomics is one of the most emerging research fields in current world. Personal genetic variation is largely associated with an individual's predisposition to several diseases. As such, with the radical development in genomic research, significant progress in diagnosis and treatment of diseases is expected. However, with this growing utilization of genomic data, privacy of an individual is going through potential risks due to following reasons: (i) human genome carries valuable information regarding a person's trait, health condition and susceptibility to diseases like Alzheimer's, (ii) even if a person publishes his/her own DNA sequence publicly, his relatives' genetic information can also be extracted without their consent [14], (iii) leakage of genetic information can cause severe disasters resulting in genetic discrimination in health insurance, employment, education etc.[3] Hence, privacy enhancing technologies need to be implemented in medical tests so that sensitive genomic data of a person cannot be inferred by adversaries.

For disease risk tests and personalized medicines of an individual, medical companies and hospitals need to sequence their patients and store genomic data in their databases. Though heavy layer of access control and legislation is applied, keeping sensitive genomic information in different hospital's storage is not safe, because (i) the threat from a potential hacker or any resentful employee cannot be ignored in this case, (ii) if a person undergoes medical tests in two different hospitals, his genomic data is stored in two different places exposing it to more probable attacks. Identity anonymization is also ineffective for storing genomic data [13, 9], as DNA sequence is the ultimate identifier of a person [10]. Furthermore, a tradeoff between privacy of genomic data and accuracy of medical test results needs to be done, if techniques like obfuscation is used. Due to huge size of genomic data, encryption also adds substantial overhead and complexity in computation of disease risks. Hence, we design a new architecture using distributed databases for privacy ensured storage of genomic information. In addition, our proposed scheme offers significant development over existing techniques in answering queries about multiple disease risks.

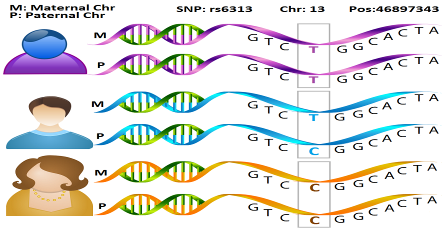

Figure 1: DNA fragments showing SNP

2. PRELIMINARIES

DNA, the main genomic material of human, is a double stranded molecule consisting of four nucleotides, such as Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Between any two given individuals, around 99.9% of the entire genome is same [11]. The remaining 0.1% part is responsible for many of our distinguishable characteristics. Single Nucleotide Polymorphism (SNP) is the most common form of human genetic variations in which a single nucleotide in the genome differs between members of the same species or paired chromosomes of an individual [15]. For example, Figure 1 shows three sequenced DNA fragments from three different persons including SNP rs6313.

So far, dbSNP has listed 112,743,739 SNPs in human population [1]. Each individual carries two alleles (i.e. two nucleotides) at each SNP position; one inherited from mother and one from father. SNP is called homozygous and het-

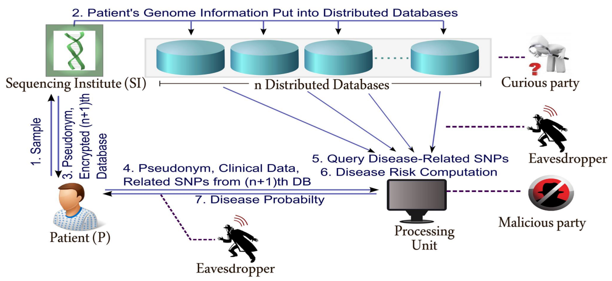
2. Patient's Genome Information Put into Distributed Databases
Sequencing Institute (SI)
n Distributed Databases
Curious party
1. Sample
3. Pseudonym, Encrypted (n+1)th Database
Eavesdropper
4. Pseudonym, Clinical Data, Related SNPs from (n+1)th DB
5. Query Disease-Related SNPs
6. Disease Risk Computation
7. Disease Probabilty
Patient (P)
Processing Unit
Malicious party
Eavesdropper

Figure 2: System architecture for disease risk computation using distributed databases

erozygous respectively, depending on both the alleles being similar or not. Generally, for a SNP associated with a particular disease, one of the alleles carries the risk and other does not. Furthermore, it is possible that both the alleles of a particular SNP carries risk for two different diseases [2].

3. RELATED WORK

String matching and comparison algorithms with encryption has been used widely for protecting security of genomic data [16, 8, 12, 7]. In [4, 5], Ayday *et al.* proposed a privacy-preserving disease susceptibility test using modified Paillier cryptosystem and proxy re-encryption. They also included clinical and environmental data of the patients in addition to genomic data in [6]. These works involve secure computation of weighted average similar to our proposal. However, the main difference is in the architecture, where we used distributed storage mechanism for protecting genomic privacy instead of cryptographic methods used in these works. Furthermore, their proposal focuses on a particular disease, whereas in our system, we incorporate the option to answer multiple disease risk queries by different patients.

4. APPROACH AND UNIQUENESS

In this work, our motivation is to find a secure medical test procedure that calculates disease risks using genomic data and various clinical factors. The detailed system architecture is shown in Figure 2.

4.1 Gene Sequencing

We propose that a sequencing institute (SI) performs the sequencing of genomic data of a patient (P). SI sequences the sample, e.g. saliva, hair etc. provided by P and extracts the SNPs from the raw genomic data.

4.2 Storing Data in Distributed Databases

SI stores the SNPs in (n+1) distributed databases where n is even. We assign four two-bit strings 00, 01, 10 and 11 to represent four bases A, C, G or T respectively. Each database contains one bit corresponding to each SNP. First n/2 databases contains entries such that their exclusive-or (XOR) gives the first bit of the two-bit string. And rest n/2 database entries give the second bit after XOR operation. The $(n+1)^{th}$ database is encrypted using public key of the patient and stored in his personal device like mobile or computer. After retrieving content of one allele of a particular SNP, we need to know whether it is homozygous or heterozygous to find out another allele. This information is kept in n^{th} database where, 0 denotes homozygous and 1 denotes heterozygous. Figure 3 shows storage and retrieval process of SNP in detail. Here, n equals to 4.

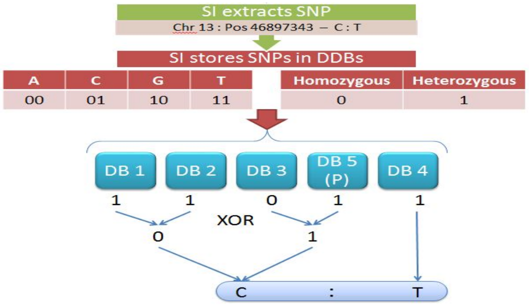

Figure 3: Storage and Retrieval of SNP content

4.3 Computation of Disease Risk

For computation of disease risk, we propose a Processing Unit (PU) which contains the names of particular SNPs and clinical factors responsible for various diseases. These information are not rare, as SNPs associated with specific diseases are now-a-days available from GWAS (Genome-Wide-Association-Studies). P provides his clinical data like age, sugar level etc. and pseudnym to the PU and asks for the chances of a particular disease. PU conducts queries to all the distributed databases for those SNPs of P related to the corresponding disease. For each SNP query, P also provides content from his personal database. PU then computes disease risk by weighted averaging contributions of all SNPs and other clinical factors via logistic regression model. As we are storing the actual content of the SNP in our storage, the probability of P to grow any disease can be computed precisely. The final result is encrypted using a public key of P and returned to him.

Threat Model & Security Analysis

In our model, we assumed that SI is a trusted entity. This assumption is inevitable in the sense that sequencing has to be done at an institution to obtain the genomic profile of a person. Government or any kind of central medical institute can play the role of SI.

Here, we considered three types of potential adversaries. One, hacker, eavesdropper or curious party at distributed databases; two, disgruntled employee or malicious party at PU; three, eavesdropper at the connection line between patient and PU. First adversary is resisted using the distributive storage mechanism. It is ensured that even if one or all of the databases are hacked or eavesdropped, the adversaries cannot retrieve true content of an SNP. Second adversary is controlled by the fact that true genomic data is not revealed unless all distributed databases including patient's personal one colludes. Attack from third adversary is protected using encryption of patient's personal database and final disease risk result. Threat modeling is highlighted in Figure 2 along with the system architecture.

5. CONTRIBUTIONS

The main contributions of our work are summarized in the following.

- We develop a new architecture where a central sequencing institute and processing unit communicate with the patients to satisfy their query regarding disease risk issues.
- The chances of developing multiple diseases can be computed correctly in our proposed system which is a significant development over the existing methods.
- Our approach shows notable reduction in computational overhead compared to current cryptography based approaches.
- In our proposal, we managed to securely protect sensitive genetic information from third parties without compromising with the accuracy of the result which is the most important aspect for a medical test.

6. REFERENCES

[1] National Center for Biotechnology Information, United States National Library of Medicine (NCBI) dbSNP build 142 for human. summary page. Visited on 27/Feb/2015.

[2] Rs6313, SNPedia. Visited on 27/Feb/2015.

[3] E. Ayday, E. D. Cristofaro, G. Tsudik, and J. P. Hubaux. The chills and thrills of whole genome sequencing. *arXiv:1306.1264*, 2013.

[4] E. Ayday, J. L. Raisaro, and J.-P. Hubaux. Personal use of the genomic data: Privacy vs. storage cost. In *IEEE Global Communications Conference, Exhibition and Industry Forum (GLOBECOM)*, pages 2723–2729, December 2013.

[5] E. Ayday, J. L. Raisaro, J.-P. Hubaux, and J. Rougemont. Protecting and evaluating genomic privacy in medical tests and personalized medicine. In *WPES'13: Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society*, page 95âĂŞ106, Berlin, Germany, November 2013.

[6] E. Ayday, J. L. Raisaro, P. J. McLaren, J. Fellay, and J.-P. Hubaux. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. In *HealthTech'13: Proceedings of USENIX Security Workshop on Health Information Technologies*, 2013.

[7] M. Canim, M. Kantarcioglu, and B. Malin. Secure management of biomedical data with cryptographic hardware. *IEEE Transactions on Information Technology in Biomedicine*, 16(1):166–175, 2012.

[8] E. D. Cristofaro, S. Faber, and G. Tsudik. Secure genomic testing with size- and position-hiding private substring matching. In *WPES '13: Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society*, pages 107–118, Berlin, Germany, November 2013.

[9] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science*, 339:321–324, January 2013.

[10] N. Homer, S. Szelinger, M. Redman, D. Duggan, and W. Tembe. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4, August 2008.

[11] L. B. Jorde and S. P. Wooding. Genetic variation, classification and race. *Nature Genetics*, 36, 2004.

[12] M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin. A cryptographic approach to securely share and query genomic sequences. *IEEE Transactions on Information Technology in Biomedicine*, 12(5):606–617, 2008.

[13] B. A. Malin. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *Journal of the American Medical Informatics Association*, 12:28–34, 2005.

[14] F. Stajano, L. Bianchi, P. Lio, and D. Korff. Forensic genomics: kin privacy, driftnets and other open questions. In *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society*, 2008.

[15] P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S. Thomas, and D. N. Cooper. The human gene mutation database: 2008 update. *Genome Medicine 2009*, 1(1), January 2009.

[16] J. R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik. Privacy preserving error resilient dna searching through oblivious automata. In *CCS '07: Proceedings of the 14th ACM Conference on Computer and Communications Security*, October-November 2007.