Protecting Genomic Privacy in Medical Tests using Distributed Storage

Sharmin Afrose (0905028), Maitraye Das (0905052)

Introduction

Individual's chances of diseases are largely associated with personal genetic variations. Hence, Genomic data is significantly used in disease susceptibility tests and personalized medicine.

Privacy Threats

- Reveals traits, ancestry, ethnicity, vulnerability of diseases etc.
- Exposes relatives' genomic data[3]
- Genomic discrimination in health insurance, employment, education etc.

Thesis Goal

Privacy-preserved and precise computation of multiple disease risks using genomic and clinical data

Novelty:

Existing cryptography-based methods[1,2] support only single disease risk test, whereas our approach offers substantial improvement regarding multiple disease risk queries in a privacy-preserving manner.

Genomic Background

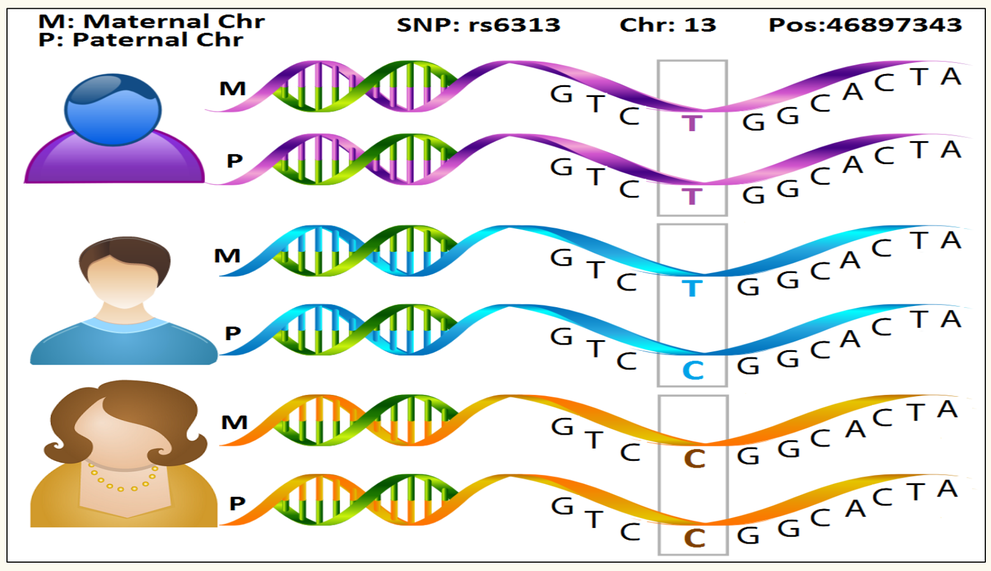

Fig 1: DNA fragments showing SNP rs6313

Genome

- ✓ Four nucleotides: A, C, G and T
- ✓ 3 billion (approx.) base pairs
- ✓ 99.9% of entire genome is same between two persons
- ✓ 112,743,739 enlisted SNPs by dbSNP

SNP

- ✓ Difference of a single nucleotide between
  - Members of same species
  - Paired chromosomes of an individual
- ✓ Each SNP carries two alleles - one from father, one from mother
- ✓ Both alleles can contain risks for two different diseases

System Architecture & Threat Model

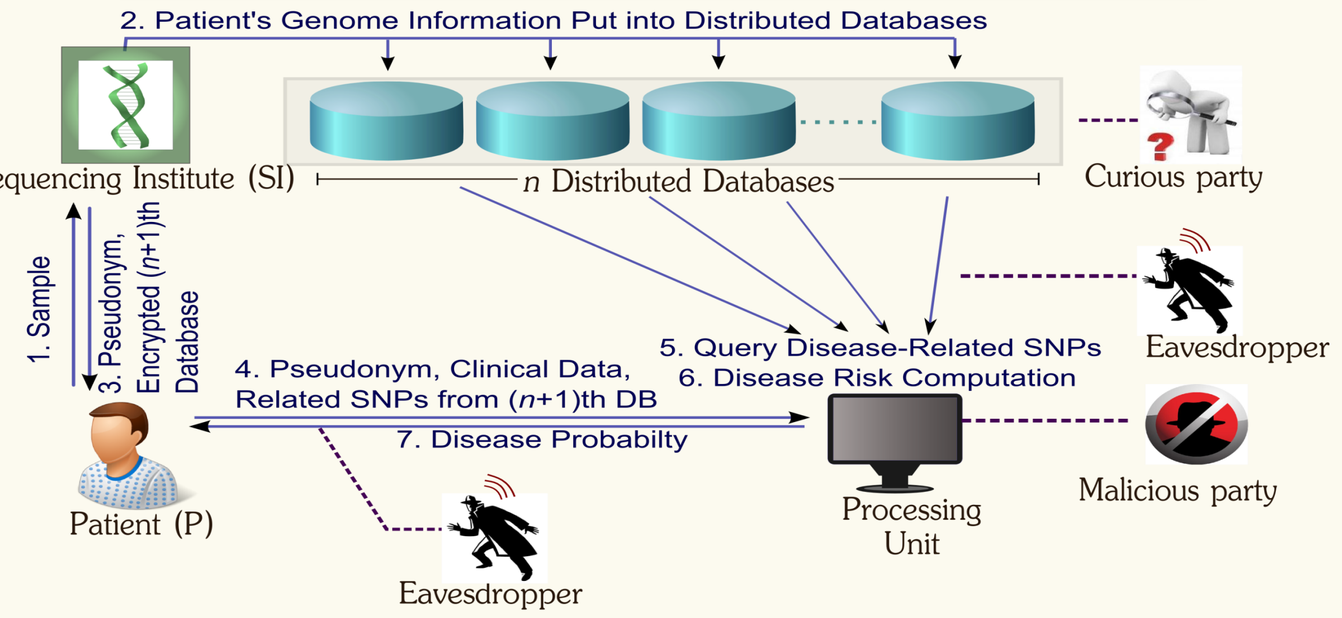

Fig 2: System architecture for disease risk computation

Our Approach

Assumptions:

- Total (n+1) databases used
- (n+1)th database encrypted with patient's public key and stored in her personal device
- DDBs maintain protocols appropriately

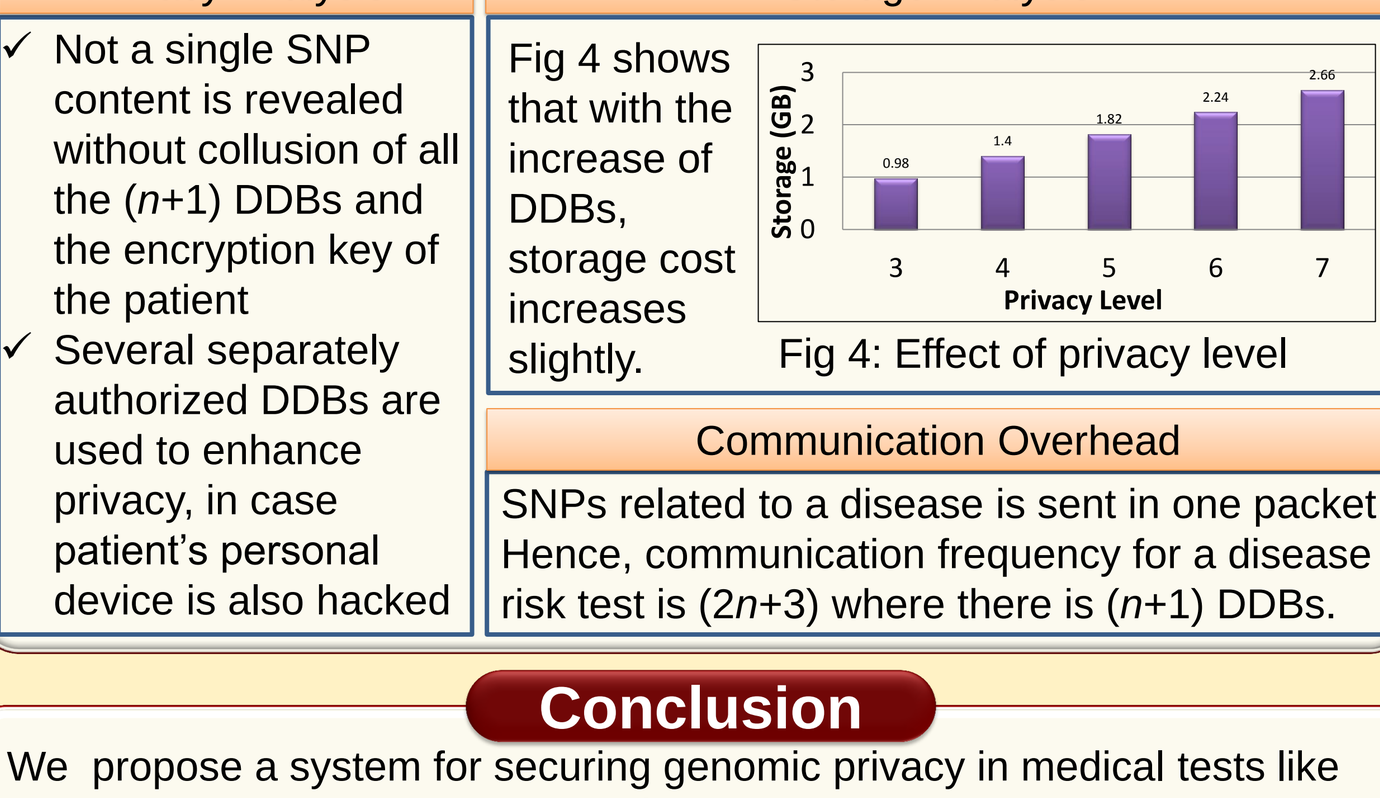

Fig 3: Flowchart of disease risk test (a) Sequencing and SNP extraction (b) Retrieval of SNP and computation

Evaluation

Privacy Analysis

- ✓ Not a single SNP content is revealed without collusion of all the (n+1) DDBs and the encryption key of the patient
- ✓ Several separately authorized DDBs are used to enhance privacy, in case patient's personal device is also hacked

Storage Analysis

Fig 4 shows that with the increase of DDBs, storage cost increases slightly.

Fig 4: Effect of privacy level

Communication Overhead

SNPs related to a disease is sent in one packet. Hence, communication frequency for a disease risk test is (2n+3) where there is (n+1) DDBs.

Conclusion

We propose a system for securing genomic privacy in medical tests like disease risk computation, personalized medicine using distributed storage. Currently, we are working on implementation and complexity analysis of the system in detail.

References

1. E. Ayday, J. L. Raisaro, J. P. Hubaux, and J. Rougemont, *Protecting and evaluating genomic privacy in medical tests and personalized medicine*, in the proceedings of WPES'13, p. 95-106.
2. 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 the proceedings of HealthTech'13.
3. http://edition.cnn.com/2013/08/07/health/henrietta-lacks-genetic-destiny/

Department of Computer Science and Engineering (CSE), BUET