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 trait, ancestry, vulnerability of diseases etc.
- Exposes relatives' genome.
- Initiates genomic discrimination in insurance, employment etc.

Thesis Goal

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

Novelty:

We offer substantial improvement over cryptography-based methods[1,2,3] regarding queries for multiple diseases related to both the alleles of the same SNP.

Genomic Background

- ✓ Four nucleotides : A, C, G, T.
- ✓ **SNP:** Difference of a single nucleotide between
  - Members of same species
  - Paired chromosomes of an individual.
- ✓ Each SNP carries two alleles; one from each parent.
- ✓ Both alleles can contain risks of different diseases.

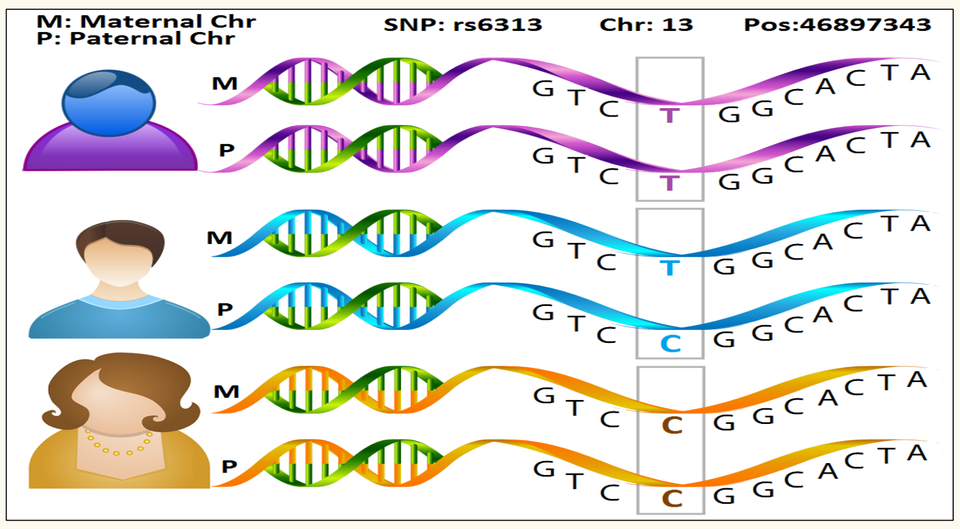

Fig 1: DNA fragments showing SNP rs6313

System Architecture

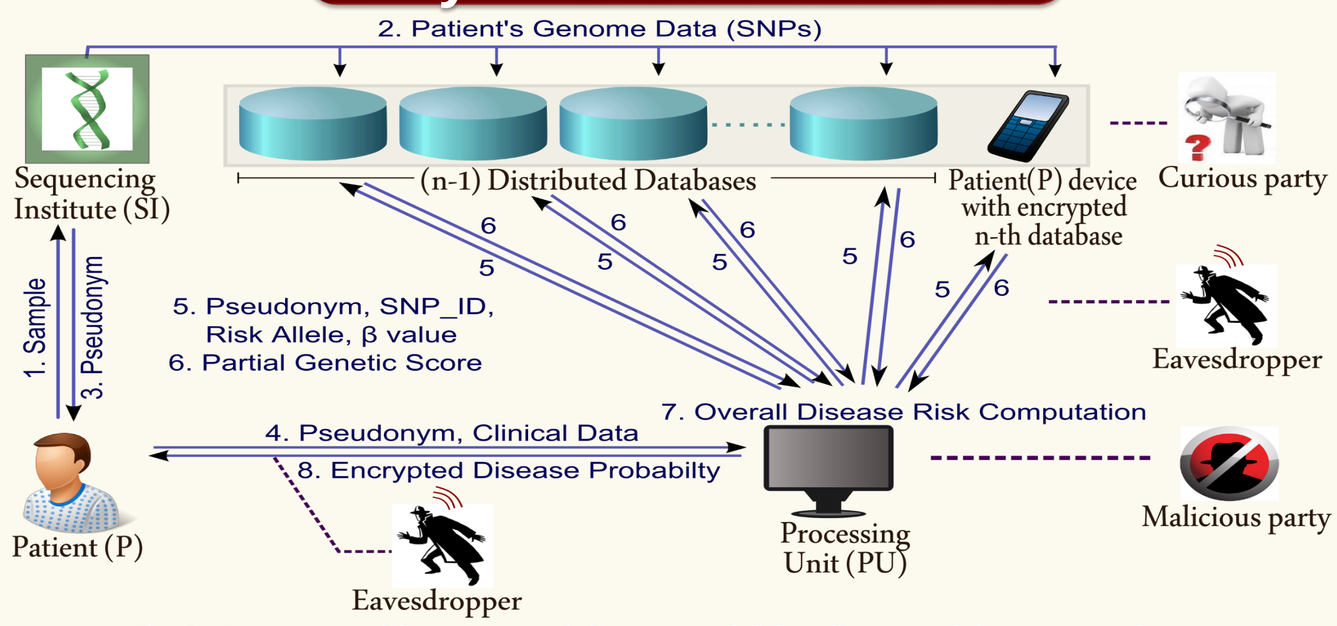

Fig 2: System architecture and threat model for disease risk computation

Our Approach

Genomic data is distributed in DBs such that only aggregated data reveals true contents. Dummy diseases are added in query using ℓ–diversity.

Bit Representation of Nucleotides

A = 00	C = 01	G = 10	T= 11

SNP Division by SI

S_1: Rs1799950 (A/G) Patient Alleles: AA

	a_1	w_1	a_2	w_2
DB1	00	42	10	45
DB2	00	18	10	-12
DB3	10	0	00	100
DB4	00	40	10	-33

S_4: Rs6311 (C:T) Patient Alleles: CC

	a_1	w_1	a_2	w_2
DB1	01	10	01	85
DB2	11	-42	11	99
DB3	01	90	11	-99
DB4	11	42	01	15

SNP Set Partition by PU

Target Disease , X
$P(X) = \{S_1, S_2, S_3, S_4\}$

Dummy Diseases, $Y_{1, \ldots, Y_{\ell-1}}$
e.g. $P(Y_1) = \{S_5, S_6, S_7\}$

Partition → $Q_1 = \{S_1, S_4\}$, $Q_2 = \{S_3, S_2\}$

Partition → $Q_3 = \{S_5\}$, $Q_4 = \{S_6, S_7\}$

β_i Value Division by PU

Let, β_i = 0.55 → Division → 0.12, 0.43, 0.1, 0.3

e_{1i}, e_{2i} — Original; e_{3i}, e_{4i} — Dummy

(a)

Query Message Sent by PU

S_1, 10, e_{21},e_{41},e_{31},e_{11} : S_4, 01, e_{24},e_{44},e_{34},e_{14} |
S_5, 01, e_{15},e_{35},e_{25},e_{45} |
S_3, 00, e_{23},e_{43},e_{33},e_{13} : S_2, 10, e_{22},e_{42},e_{32},e_{12} |
S_6, 01, e_{46}, e_{26},e_{16},e_{36} : S_7, 11, e_{47},e_{27},e_{17},e_{37} |

Marker Array of PU

T	F	F	T	F	F	F	F	T	F	F	T	F	F	F	F

Partial Calculation at DB1

For S_1, weight of risk allele, 10 = 45
For S_4, weight of risk allele, 01 = 10 + 85
Partial genetic score, $S_{1,1}$= 0.45 e_{21} + 0.95 e_{24}
Return message, $R_1 = S_{1,1} : S_{2,1} : \ldots \ldots \ldots : S_{16,1}$

Total Genetic Score, $\mathbb{S}$ Calc. at PU

Sum all $S_{i,j}$ corresponding to marker entry 'T'

$$\mathbb{S} = \sum_{i=1}^{\lambda} \beta_i \times p_i(X)$$

Total Disease Risk Calc. at PU

Patient provides clinical data, $n_i(X)$, e.g. age

$$\beta_f = \ln\left(\frac{Pr}{1-pr}\right) = \mathbb{S} + \sum_{i=1}^{\varphi} \beta^-_i \times n_{i(X)}$$

Total disease risk, $\Pr = \frac{e^{\beta_f}}{1+e^{\beta_f}}$

(b)

Fig 3: Flowchart of (a) query message generation (b) total disease risk computation

Implementation and Evaluation

Privacy Analysis

- ✓ SNP retrieval attack by
  - Semi-honest DDB
  - Semi-honest PU
  - Dishonest-but-covert PU
- ✓ Test inference attack by semi-honest DDB
- ✓ β value inference attack by semi-honest DDB

Dataset Size

- ✓ Over 0.3 million.

Assumption

- ✓ DDBs maintain protocols.

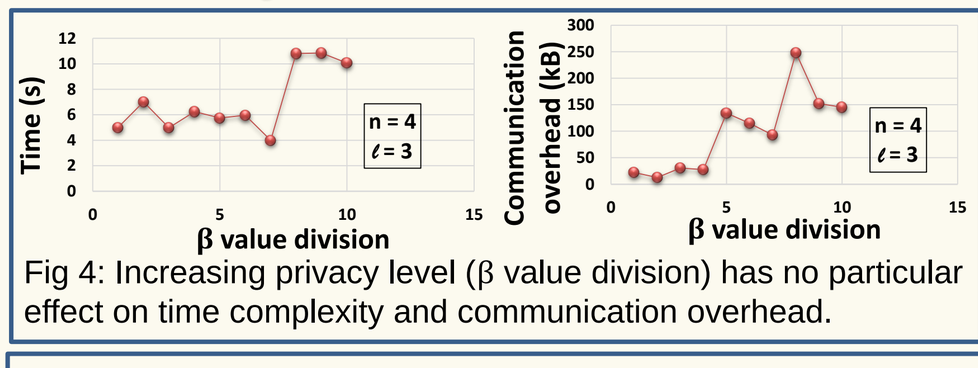

Fig 4: Increasing privacy level (β value division) has no particular effect on time complexity and communication overhead.

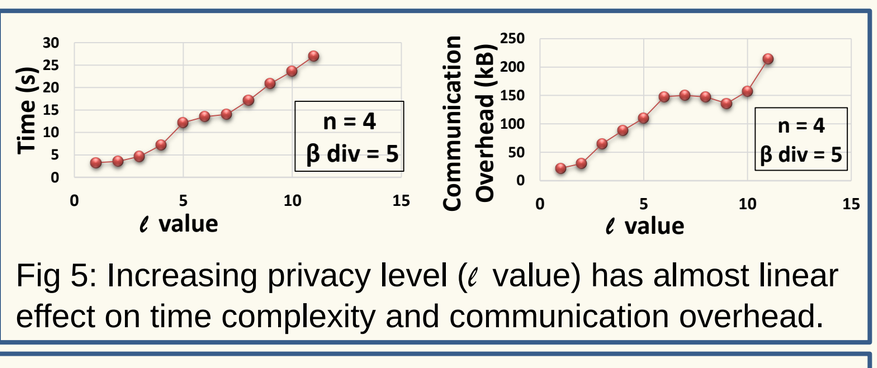

Fig 5: Increasing privacy level (ℓ value) has almost linear effect on time complexity and communication overhead.

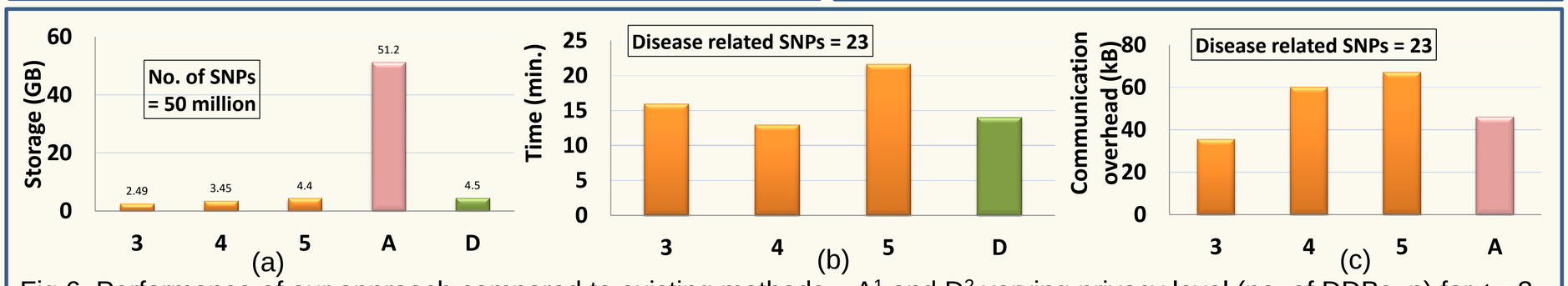

Fig 6. Performance of our approach compared to existing methods – A[1] and D[2] varying privacy level (no. of DDBs, n) for ℓ = 3 and β division = 5 with respect to (a) storage (b) time complexity (c) communication overhead

Conclusion

Our proposed system preserves genomic privacy in medical tests using distributed storage and precisely computes risks for multiple diseases in all possible scenarios. Estimation of performance compared to existing methods over real dataset shows its practicality in real-life implementation.

References

1. E. Ayday, J. L. Raisaro, P. J. McLaren, J. P. Hubaux, and J. Rougemont, "Protecting and evaluating genomic privacy in medical tests and personalized medicine", in *WPES,* 2013.
2. G. Danezis and E. D. Cristofaro, "Fast and private genomic testing for disease susceptibility", in *WPES,* 2014.
3. L. Barman, M.-T. Elgraini, J. L. Raisaro, J.-P. Hubaux, and E. Ayday, "Privacy threats and practical solutions for genetic risk tests", in *GenoPri*, 2015.
4. http://edition.cnn.com/2013/08/07/health/henrietta-lacks-genetic-destiny/

Department of Computer Science and Engineering (CSE), BUET