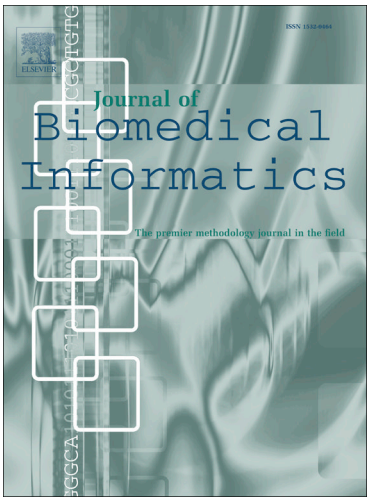

Secure Count Query on Encrypted Genomic Data

Mohammad Zahidul Hasan, Md Safiur Rahman Mahdi, Md Nazmus Sadat, Noman Mohammed

PII: S1532-0464(18)30045-5
DOI: https://doi.org/10.1016/j.jbi.2018.03.003
Reference: YJBIN 2944

To appear in: *Journal of Biomedical Informatics*

Received Date: 4 August 2017
Revised Date: 5 December 2017
Accepted Date: 12 March 2018

Please cite this article as: Hasan, M.Z., Rahman Mahdi, M.S., Sadat, M.N., Mohammed, N., Secure Count Query on Encrypted Genomic Data, *Journal of Biomedical Informatics* (2018), doi: https://doi.org/10.1016/j.jbi.2018.03.003

11.5 Genomic data, Cloud computing, Data sharing

Secure Count Query on Encrypted Genomic Data

Mohammad Zahidul Hasan, Md Safiur Rahman Mahdi, Md Nazmus Sadat*, Noman Mohammed

Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada

Abstract

Human genomic information can yield more effective healthcare by guiding medical decisions. Therefore, genomics research is gaining popularity as it can identify potential correlations between a disease and a certain gene, which improves the safety and efficacy of drug treatment and can also develop more effective prevention strategies [1]. To reduce the sampling error and to increase the statistical accuracy of this type of research projects, data from different sources need to be brought together since a single organization does not necessarily possess required amount of data. In this case, data sharing among multiple organizations must satisfy strict policies (for instance, HIPAA and PIPEDA) that have been enforced to regulate privacy-sensitive data sharing. Storage and computation on the shared data can be outsourced to a third party cloud service provider, equipped with enormous storage and computation resources. However, outsourcing data to a third party is associated with a potential risk of privacy violation of the participants, whose genomic sequence or clinical profile is used in these studies. In this article, we propose a method for secure sharing and computation on genomic data in a semi-honest cloud server. In particular, there are two main contributions. Firstly, the proposed method can handle biomedical data containing both genotype and phenotype. Secondly, our proposed index tree scheme reduces the computational overhead significantly for executing secure count query operation. In our proposed method, the confidentiality of shared data is ensured through encryption, while making the entire computation process efficient and scalable for cutting-edge biomedical applications. We evaluated our proposed method in terms of efficiency on a database of Single-Nucleotide Polymorphism (SNP) sequences, and experimental results demonstrate that the execution time for a query of 50 SNPs in a database of 50000 records is approximately 5 seconds, where each record contains 500 SNPs. And, it requires 69.7 seconds to execute the query on the same database that also includes phenotypes.

1. Introduction

Analysis of human genome can reveal essential information about an individual, like predisposition to a specific disease such as breast cancer, diabetes, and Alzheimer's [2]. This kind of analysis is usually done by querying an individual's genome against a list of known variations and then predicting the disease susceptibility [2]. Most of these analysis rely on genome-wide association study (GWAS). GWAS helps to understand and identify the genetic variations that are associated with a particular disease [3]. To guarantee significant accuracy in this type of analysis, a large number of genomic sequences are required, the collection of which are sometimes beyond the capability of a sole organization [4]. Allowing the access of the genomic data surpassing the premise of the organization responsible for initial collection is a viable solution. But, delegating the access of the data, be it owned by a government organization or a private research institution, is not always very straightforward because of the nature of the genomic data.

Genomic data cannot be treated as any other data; it has some distinctive features. Naveed *et al.* [5] identified six special features of genomic data. Genomic data does not change considerably over time and it is unique – two

*Corresponding author.

Email addresses: zahidul@cs.umanitoba.ca (Mohammad Zahidul Hasan), mahdi@cs.umanitoba.ca (Md Safiur Rahman Mahdi), sadat@cs.umanitoba.ca (Md Nazmus Sadat), noman@cs.umanitoba.ca (Noman Mohammed)