ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Accessing Securely DNA Database using Aggregate Queries

Shweta Mokal[1], Vedanti Ghate[1], Shubham Sawake[1], Ankit Mugali[1], Prf. Pallavi Ahire[2]

Student, Department IT, Sinhgad Institute of Technology, Savitribai Phule Pune University, Pune, India[1]

Professor, Department IT, Sinhgad Institute of Technology, Savitribai Phule Pune University, Pune, India [2]

ABSTRACT: DNA or Deoxyribonucleic Acid is the medium of long-term storage and transmission of genetic information for all modern living organisms. DNA information of specific person is sensitive therefore DNA dataset must be secure or needs security on the cloud. High security systems are required to protect data within the cloud.This paper addresses the problem of sharing person-specific genomic sequences without violating the privacy of their data subjects to support large-scale biomedical research projects. One improvement is that our scheme is deterministic, with zero probability of a wrong answer (as opposed to a low probability). We also provide a new operating point in the space-time tradeoff, by offering a scheme that is twice as fast as theirs but uses twice the storage space. This point is motivated by the fact that storage is cheaper than computation in current cloud computing pricing plans.Moreover, our encoding of the data makes it possible for us to handle a richer set of queries than exact matching between the query and each sequence of the database. Also we use aggregate queries at the time of DNA data searching on the cloud. There are four main modules namely clients i.e researchers, hospital, key holder and cloud server. This system is useful for clients i.e. researchers. Our system gives best performance as compare to other state of art systems.

KEYWORDS: DNA Databases, Cloud Security, Secure Outsourcing, Aggregate Query

I. INTRODUCTION

The cloud computing paradigm has reformed the usage and management of the information technology infrastructure. Cloud computing is characterized by on-demand self-services, ubiquitous network accesses, resource pooling, elasticity, and measured services. The aforementioned characteristics of cloud computing make it a striking candidate for businesses, organizations, and individual users for adoption. However, the benefits of low-cost, negligible management (from a users perspective), and greater flexibility come with increased security concerns. Security is one of the most crucial aspects among those prohibiting the wide-spread adoption of cloud computing. Cloud security issues may stem due to the core technologies implementation (virtual machine (VM) escape, session riding, etc.), cloud service offerings (structured query language injection, weak authentication schemes, etc.), and arising from cloud characteristics (data recovery vulnerability, Internet protocol vulnerability, etc.). For a cloud to be secure, all of the participating entities must be secure. For a cloud to be secure, all of the participating entities must be secure. High security systems are required to protect data within the cloud. DNA or Deoxyribonucleic Acid is the medium of long-term storage and transmission of genetic information for all modern living organisms. DNA information of specific person is sensitive therefore DNA dataset must be secure or needs security on the cloud. DNA data is critical for conducting biomedical research and studies, for example, diagnosis of pre-disposition to develop a specific disease, drug allergy, or prediction of success rate in response to a specific treatment. Providing a publicly available DNA database for fostering research in this field is mainly confronted by privacy concerns. Today, the abundant computation and storage capacity of cloud services enables practical hosting and sharing of DNA databases and efficient processing of genomic sequences. What is missing is an efficient security layer that preserves the privacy of individuals' records and assigns the burden of query processing to the cloud. Existing systems are not sufficient because in many cases. Therefore we propose Securing Aggregate Queries for DNA Databases. In this paper, we consider the framework proposed where the DNA records coming from several hospitals are encrypted and stored at a data storage site, and

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

biomedical researchers are able to submit aggregate counting queries to this site. Counting queries are particularly interesting for statistical analysis. This paper provides a new method that addresses a larger set of problems and provides a faster query response time.

II. LITERATURE SURVEY

M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin proposed a cryptographic approach to securely share and query genomic sequences [1]. In this paper, authors present a novel cryptographic framework that enables organizations to support genomic data mining without disclosing the raw genomic sequences. Organizations contribute encrypted genomic sequence records into a centralized repository, where the administrator can perform queries, such as frequency counts, without decrypting the data. They evaluate the efficiency of their framework with existing databases of single nucleotide polymorphism (SNP) sequences and demonstrate that the time needed to complete count queries is feasible for real world applications. They further show that approximation strategies can be applied to significantly speed up query execution times with minimal loss in accuracy. The framework can be implemented on top of existing information and network technologies in biomedical environments.

B. Malin and L. Sweeney presented how (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems [2]. In this paper, authors study the erosion of privacy when genomic data, either pseudonymous or data believed to be anonymous, are released into a distributed healthcare environment. Several algorithms are introduced, collectively called RE-Identification of Data in Trails (REIDIT), which link genomic data to named individuals in publicly available records by leveraging unique features in patient-location visit patterns. Algorithmic proofs of re-identification are developed and we demonstrate, with experiments on real world data, that susceptibility to re-identification is neither trivial nor the result of bizarre isolated occurrences. Authors propose that such techniques can be applied as system tests of privacy protection capabilities.

E. Aguiar, Y. Zhang, and M. Blanton proposed an Overview of Issues and Recent Developments in Cloud Computing and Storage Security [3].The recent rapid growth in the availability and popularity of cloud services allows for convenient on demand remote storage and computation. Security and privacy concerns, however, are among the top impediments standing in the way of wider adoption of cloud technologies. That is, in addition to the new security threats that emerge with the adoption of new cloud technology, a lack of direct control over one's data or computation demands new techniques for service provider's transparency and accountability. The goal of this chapter is to provide a broad overview of recent literature covering various aspects of cloud security. Authors describe recently discovered attacks on cloud providers and their countermeasures, as well as protection mechanisms that aim at improving privacy and integrity of client's data and computations. The topics covered in this survey include authentication, virtualization, availability, accountability, and privacy and integrity of remote storage and computation.

P. Bohannon, M. Jakobsson, and S. Srikwan, presented cryptographic Approaches to Privacy in Forensic DNA Databases [4]. Authors study access control for one class of such databases, forensic DNA databases, used to match unknown perpetrators against groups of potential suspects – usually convicted criminals. Our key observation is that for legitimate forensic queries, the sensitive information belonging to the target individual is already available to the querying agent in the form of a blood or tissue sample from a crime scene. They show how forensic DNA databases may be implemented so that only legitimate queries are feasible. In particular, a person with unlimited access to the database will be unable to extract information about any individual unless the necessary genetic information for that individual is already known. They develop a general solution framework, and show how to implement databases which handle certain cases of missing or incorrect DNA tests. This framework and techniques are applicable to the general problem of encrypting information based on partially known or partially correct keys, and its security is based on standard cryptographic assumptions.

F. Bruekers, S. Katzenbeisser, K. Kursawe, and P. Tuyls, presented privacy-preserving matching of DNA profiles [5].In this paper, authors introduce cryptographic privacy enhancing protocols that allow performing the most common DNA based identity, paternity and ancestry tests and thus implementing privacy-enhanced online genealogy services or research projects. In the semi-honest attacker model, the protocols guarantee that no sensitive information about the involved DNA is exposed, and are resilient against common forms of measurement errors during DNA sequencing. The protocols are practical and efficient, both in terms of communication and computation complexity.

M. Blanton and M. Aliasgari, proposed secure outsourcing of DNA searching via finite automata [6]. This work treats the problem of error-resilient DNA searching via oblivious evaluation of finite automata, where a client has

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

a DNA sequence, and a service provider has a pattern that corresponds to a genetic test. Error-resilient searching is achieved by representing the pattern as a finite automaton and evaluating it on the DNA sequence (which is treated as the input), where privacy of both the pattern and the DNA sequence must be preserved. Interactive solutions to this problem already exist, but can be a burden on the participating parties. In this work authors propose techniques for secure outsourcing of oblivious evaluation of finite automata to computational servers, such that the servers do not learn any information. Our techniques are applicable to any type of finite automata, but the optimizations are tailored to the setting of DNA searching.

M. J. Atallah and J. Li, proposed Secure outsourcing of sequence comparisons [7].Internet computing technologies, like grid computing, enable a weak computational device connected to such a grid to be less limited by its inadequate local computational, storage, and bandwidth resources. However, such a weak computational device (PDA, smartcard, sensor, etc.) often cannot avail itself of the abundant resources available on the network because its data are sensitive. This motivates the design of techniques for computational outsourcing in a privacy-preserving manner, i.e., without revealing to the remote agents whose computational power is being used either one's data or the outcome of the computation. This paper investigates such secure outsourcing for widely applicable sequence comparison problems and gives an efficient protocol for a customer to securely outsource sequence comparisons to two remote agents. The local computations done by the customer are linear in the size of the sequences, and the computational cost and amount of communication done by the external agents are close to the time complexity of the best known algorithm for solving the problem on a single machine.

A. E. Nergiz, C. Clifton, and Q. M. Malluhi, presented updating outsourced anatomized private databases [8].Authors introduce operations to safely update an anatomized database. The result is a database where the view of the server satisfies standards such as k-anonymity or l-diversity, but the client is able to query and modify the original data. By exposing data where possible, the server can perform value added services such as data analysis not possible with fully encrypted data, while still being unable to violate privacy constraints. Update is a key challenge with this model; native application of insertion and deletion operations reveals the actual data to the server. This paper shows how data can be safely inserted, deleted, and updated. The key ideas are that data is inserted or updated into an encrypted temporary table until enough data is available to safely decrypt, and that sensitive information of deleted tuples is left behind to ensure privacy of both deleted and undeleted individuals.

L. Sweeney, A. Abu, and J. Winn, presented identifying Participants in the Personal Genome Project by Name [9].Authors linked names and contact information to publicly available profiles in the Personal Genome Project. These profiles contain medical and genomic information, including details about medications, procedures and diseases, and demographic information, such as date of birth, gender, and postal code. By linking demographics to public records such as voter lists, and mining for names hidden in attached documents, they correctly identified 84 to 97 percent of the profiles for which we provided names. Their ability is to learn their names is based on their demographics, not their DNA, thereby revisiting an old vulnerability that could be easily thwarted with minimal loss of research value. So, they propose technical remedies for people to learn about their demographics to make better decisions.

F. Esponda, E. S. Ackley, P. Helman, H. Jia, and S. Forrest, proposed protecting data privacy through hard-to-reverse negative databases [10].The paper extends the idea of negative representations of information for enhancing privacy. Simply put, a set DB of data elements can be represented in terms of its complement set. That is, all the elements not in DB are depicted and DB itself is not explicitly stored. Authors review the negative database (NDB) representation scheme for storing a negative image compactly and propose a design for depicting a multiple record DB using a collection of NDBs—in contrast to the single NDB approach of previous work. Finally, they present a method for creating negative databases that are hard to reverse in practice, i.e., from which it is hard to obtain DB, by adapting a technique for generating 3-SAT formula

III. EXISTING SYSTEM APPROACH

To support large-scale biomedical research projects, organizations need to share person-specific genomic sequences without violating the privacy of their data subjects. In the past, organizations protected subjects' identities by removing identifiers, such as name and social security number; however, recent investigations illustrate that deidentified genomic data can be "reidentified" to named individuals using simple automated methods. What is missing is an efficient security layer that preserves the privacy of individuals' records and assigns the burden of query processing to the cloud.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Existing systems are not sufficient because in many cases. It follows that the DNA data must be protected, not just unlinked from the corresponding persons.

IV. PROPOSED SYSTEM APPROACH

We address the problem of sharing person-specific genomic sequences without violating the privacy of their data subjects to support large-scale biomedical research projects. One improvement is that our scheme is deterministic, with zero probability of a wrong answer (as opposed to a low probability). We also provide a new operating point in the space-time tradeoff, by offering a scheme that is twice as fast as theirs but uses twice the storage space. This point is motivated by the fact that storage is cheaper than computation in current cloud computing pricing plans.Moreover, our encoding of the data makes it possible for us to handle a richer set of queries than exact matching between the query and each sequence of the database. Also we use aggregate queries at the time of DNA data searching on the cloud.

V. PROPOSED ARCHITECTURE

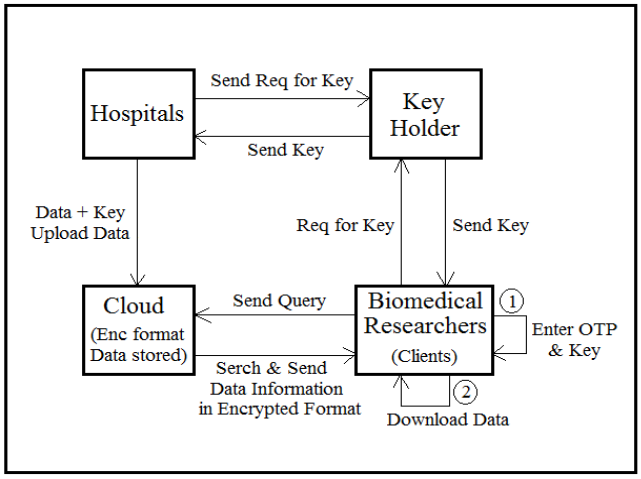

Figure 1: System Architechture

The architecture of our system is shown in above Figure 1. There are four main modules: Hospitals, Key Holder, Cloud and Biomedical Researchers i.e. Clients. Cloud represents the data store where all the encrypted DNA records are stored and is responsible of processing the queries. Key Holder is a trusted party that generates and holds the private and public keys of the homomorphic encryption scheme. The hospitals send the request for key to Key Holder and obtain the key in order to encrypt their DNA records and upload them to Cloud. Biomedical researchers representing clients submits a query to Cloud. The Cloud processes the query over the encrypted records and sends the results which are in encrypted format to Clients in order to be decrypted. Cloud is required to permute the results for individual records before sending them out. The permutation protects the records if in any case the order of the records can be linked to some protected information. Client send request for OTP (One Time Password) to our system. Then system generates OTP and send to Client. Client also send request for key to Key Holder. Key Holder generates key and send to Client. Finally Client enters OTP and key and after that data/file will download.

VI. CONCLUSION

In this paper, we have revisited the challenge of sharing person-specific genomic sequences without violating the privacy of their data subjects in order to support large-scale biomedical research projects. We have used the framework proposed by Kantarcioglu*et al.* [1] based on additive homomorphic encryption. In our project we used two cloud servers; in that one is called as key holder which is used to holding the keys and other is called as cloud server itself which is used to storing the encrypted records. We develop this project for biomedical researchers which act as client in

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

our project. Also we have hospital module which stores DNA database on the cloud with the help of our system. The proposed method offers two new operating points in the space-time tradeoff and handles new types of queries that are not supported in earlier work. Furthermore, the method provides support for extended alphabet of nucleotides which is a practical and critical requirement for biomedical researchers.

REFERENCES

[1] M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin, "A cryptographic approach to securely share and query genomic sequences," *Inf. Technol. Biomed. IEEE Trans.*, vol. 12, no. 5, pp. 606–617, 2008.
[2] B. Malin and L. Sweeney, "How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems," *J. Biomed. Inform* vol. 37, no. 3, pp. 179–192, 2004.
[3] E. Aguiar, Y. Zhang, and M. Blanton, "An Overview of Issues and Recent Developments in Cloud Computing and Storage Security," in *High Performance Cloud Auditing and Applications*, 2014, pp. 3–33.
[4] P. Bohannon, M. Jakobsson, and S. Srikwan, "Cryptographic Approaches to Privacy in Forensic DNA Databases," in *Public Key Cryptography*, vol. 1751, H. Imai and Y. Zheng, Eds. Springer Berlin Heidelberg, 2000, pp. 373–390.
[5] F. Bruekers, S. Katzenbeisser, K. Kursawe, and P. Tuyls, "Privacy-preserving matching of DNA profiles," *IACR Cryptol. ePrint Arch.*, vol. 2008, p. 203, 2008.
[6] M. Blanton and M. Aliasgari, "Secure outsourcing of DNA searching via finite automata," in *Data and Applications Security and Privacy XXIV*, Springer, 2010, pp. 49–64.
[7] M. J. Atallah and J. Li, "Secure outsourcing of sequence comparisons," *Int. J. Inf. Secur.*, vol. 4, no. 4, pp. 277–287, Mar. 2005.
[8] A. E. Nergiz, C. Clifton, and Q. M. Malluhi, "Updating outsourced anatomized private databases," in *Proceedings of the 16th International Conference on Extending Database Technology*, 2013, pp. 179–190.
[9] L. Sweeney, A. Abu, and J. Winn, "Identifying Participants in the Personal Genome Project by Name," *Available SSRN 2257732*, 2013.
[10] F. Esponda, E. S. Ackley, P. Helman, H. Jia, and S. Forrest, "Protecting data privacy through hard-to-reverse negative databases," *Int. J. Inf. Secur.*, vol. 6, no. 6, pp. 403–415, 2007.