The Use of Aggregate Queries to Create a Flexible and Secure DNA Database

D.Padma, G. Chandrika
Assistant Professor, Dadi Institute of Engineering & Technology
dpadma@diet.edu.in, chandrika@diet.edu.in

Abstract:
In order to facilitate extensive biomedical research projects, the proposed research activity focuses on the issue of exchanging person-specific genetic sequences without invading the privacy of its data subjects. We also propose a new operational point in the space-time tradeoff by proposing a faster but larger storage-intensive approach than theirs. The fact that storage is less expensive than computation in the present cloud computing pricing models serves as the foundation for this argument. We can also handle a wider range of queries thanks to the data's encoding, such as (i) counting the number of matches between a query's symbols and a sequence, (ii) logical OR matches where a query's symbols match a sequence, and (iii) logical AND matches where a query's symbols match a sequence.
Key words : DNA Databases, Cloud Security, Secure Outsourcing.

Introduction

The cloud computing paradigm has changed how the information technology infrastructure is used and managed. Cloud computing features on-demand self-services, ubiquitous network connectivity, resource pooling, elasticity, and quantifiable services. Cloud computing is a clear candidate for adoption by businesses, organizations, and individual users due to the aforementioned characteristics. The benefits of low cost, minimum management (from the users' point of view), and greater flexibility come with an increase in security risks. Security is one of the main things holding back the mainstream use of cloud computing. Cloud features (data recovery flaws, VM escapes, session riding, etc.) and cloud service offerings (structured query language injection, loose authentication, etc.) can all lead to security problems in the cloud. a person's likelihood of contracting a specific disease, the finding of a drug allergy, or the prediction of a treatment's likelihood of success are examples of research and investigations. Concerns over privacy stand in the way of developing a freely accessible DNA database for

this kind of research. The enormous computation and storage capability of cloud services today allows for the practical hosting and sharing of DNA databases as well as efficient processing of genomic sequences.

Literature Survey

A cryptographic strategy was put out by M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin to securely share and query genomic sequences [1]. The authors of this study introduce a novel cryptographic framework that enables businesses to facilitate genomic data mining without revealing the unprocessed genomic sequences. Encrypted genomic sequence records are contributed by organizations to a central repository, where the administrator may run queries like frequency counts without having to decode the data. They use existing databases of single nucleotide polymorphism (SNP) sequences to test the effectiveness of their architecture and show that the time required to finish count queries is reasonable for use in practical applications. They also demonstrate how approximation techniques can be used to dramatically shorten query execution times while maintaining high accuracy. The framework can be used in biomedical environments on top of already-existing information and network technologies.

When evaluating and creating anonymity protection systems, B. Malin and L. Sweeney described how (not) to secure genetic data privacy in a dispersed network [2]. They used trail re-identification. In this article, writers investigate how privacy is compromised when genetic data, whether pseudonymous or data regarded as anonymous, are made available in a dispersed healthcare setting. We describe a set of methods, collectively known as RE-Identification of Data in Trails (REIDIT), that use particular characteristics in patient-location visit patterns to connect genetic data to identified persons in publicly accessible records. Using experiments on real-world data, we provide algorithmic proofs of re-identification and show that vulnerability to re-identification is not trivial nor the product of strange single occurrences. The authors suggest using these strategies to test a system's privacy protection capabilities. An Overview of Issues and Recent Developments in Cloud Computing and Storage Security was proposed by E. Aguiar, Y. Zhang, and M. Blanton [3].With cloud services' recent explosive development in popularity and accessibility, convenient remote storage and computing are now possible. But among the main barriers preventing a wider use of cloud technologies are worries about security and privacy. The lack of direct control over one's data or

computation, in other words, necessitates new ways for service providers' openness and accountability. This is in addition to the new security dangers that arise with the adoption of new cloud technology. This chapter aims to give a thorough overview of recent research on many facets of cloud security. The authors discuss recent assaults on cloud service providers, their defenses, and protection techniques designed to enhance the confidentiality and integrity of client data and computations. Authentication, virtualization, availability, accountability, privacy, and the integrity of remote storage and computing are among the themes covered in this survey. The authors of cryptographic Approaches to Privacy in Forensic DNA Databases are P. Bohannon, M. Jakobsson, and S. Srikwan [4]. Authors investigate access control for one category of these databases, forensic DNA databases, which are used to compare groups of prospective suspects—typically convicted criminals—against unidentified culprits. Our main finding is that for legal forensic searches, the querying agent already has access to the sensitive information belonging to the target person in the form of a blood or tissue sample from a crime scene. They demonstrate how forensic DNA databases can be set up so that only valid searches are practical. In instance, unless the required genetic information for each individual is already known, a person with unfettered access to the database will not be able to extract information on any individual. They create a framework for a general solution and demonstrate how to use databases to address specific scenarios including incomplete or inaccurate DNA samples. The security of this framework and its procedures is based on accepted cryptographic presumptions, and they are applicable to the wider issue of encrypting data using keys that are only partially known or right. Matching of DNA profiles while protecting privacy was provided by F. Bruekers, S. Katzenbeisser, K. Kursawe, and P. Tuyls [5].The authors of this work introduce cryptographic privacy-enhancing methods that enable the most popular DNA-based ancestry, paternity, and identity tests, enabling the implementation of privacy-enhanced online genealogical services or research initiatives. The techniques are resilient against typical types of measurement mistakes that might occur during DNA sequencing and ensure that no important information about the relevant DNA is disclosed under the model of a semi-honest attacker. In terms of communication and computing complexity, the protocols are useful and effective. Finite automata were suggested by M. Blanton and M. Aliasgari [6] for the secure outsourcing of DNA searches. Through the use of oblivious evaluation of finite automata, this study addresses the issue of error-resilient DNA searching in situations where a customer has a DNA

sequence and a service provider has a pattern that relates to a genetic test. When the privacy of both the pattern and the DNA sequence must be respected, error-resilient searching is accomplished by modeling the pattern as a finite automata and evaluating it on the DNA sequence (which is treated as the input). Existing interactive solutions to this issue can be taxing on the parties involved. In order to prevent computational servers from learning any information, authors in this work present methods for secure outsourcing of oblivious evaluation of finite automata to computing servers. Any type of finite automaton can be solved using these methods, but DNA searching is the context for the optimizations. The idea of Secure outsourcing of sequence comparisons was put up by M. J. Atallah and J. Li [7].A poor computational device connected to such a grid can be less constrained by its insufficient local computational, storage, and bandwidth resources thanks to internet computing technologies like grid computing. However, because its data are sensitive, such a poor computational device (PDA, smartcard, sensor, etc.) frequently cannot take advantage of the numerous resources on the network. This drives the development of systems for computational outsourcing that protect privacy, i.e., without disclosing one's data or the result of the calculation to the remote agents whose computational power is being employed. This study examines such secure outsourcing for broadly applicable sequence comparison problems and provides a practical technique for a customer to safely outsource sequence comparisons to two remote agents. The customer's local computations are linear in terms of the length of the sequences, while the external agents' computation and communication costs are roughly equivalent to the temporal complexity of the most effective known solution for addressing the problem on a single computer.

This motivates the creation of computational outsourcing systems that protect privacy, i.e., without disclosing one's data or the calculation's outcome to the remote agents whose computational power is being utilized. This study analyzes such secure outsourcing for widely applicable sequence comparison problems and offers a workable method for a customer to confide safely on two remote agents to perform sequence comparisons. While the external actors' computation and communication costs are generally similar to the temporal complexity of the most efficient known method for handling the problem on a single computer, the customer's local computations are linear in terms of the length of the sequences. This study demonstrates secure data insertion, deletion, and update techniques.

The main concepts are that sensitive information of deleted tuples is left behind to protect privacy of both deleted and undeleted individuals, and that data is added or modified into an encrypted temporary table until enough data is available to properly decrypt it. Identification of Participants in the Personal Genome Project by Name was provided by L. Sweeney, A. Abu, and J. Winn [9].Authors connected names and contact details to Personal Genome Project profiles that were accessible to the general public. In addition to demographic data like date of birth, gender, and zip code, these profiles also include medical and genomic information, including specifics regarding treatments, procedures, and disorders. They properly identified between 84 and 97 percent of the profiles for which we provided names by tying demographics to public information like voter lists and mining for identities hidden in connected papers. Their capacity to remember their names is determined by their demography, not by their DNA, resurfacing an old weakness that might be readily closed with little loss of study value. Hard-to-reverse negative databases were suggested as a method of protecting data privacy by F. Esponda, E. S. Ackley, P. Helman, H. Jia, and S. Forrest [10].The idea of using negative representations of information to increase privacy is expanded upon in this study. Simply put, the complement set of a set of data components can be used to represent the set. That is, all things that are not kept in the database are shown, and the database itself is not expressly stored. In contrast to the single NDB technique used in earlier work, the authors review the negative database (NDB) representation scheme for compactly storing a negative picture and suggest a design for representing a multiple record DB using a set of NDBs. Finally, they provide a way for modifying a method for producing 3-SAT formula in order to create negative databases that are difficult to reverse in practice, i.e., from which it is difficult to obtain DB.

Existing System Approach

Organizations must disclose individualized genetic sequences without invading the privacy of their data subjects in order to enable extensive scientific research programs. Historically, organizations safeguarded people' identities by deleting identifiers like names and social security numbers; however, new research show that deidentified genetic data may be "reidentified" to named persons using straightforward automated approaches. What is required is a strong security layer that protects the confidentiality of personal data while shifting the responsibility for query processing to the cloud.

Existing technologies are insufficient since DNA data needs to be preserved as well as separated from the corresponding individuals.

Proposed System Approach

In order to facilitate extensive scientific research projects, we address the issue of exchanging person-specific genetic sequences without invading the privacy of its data subjects. Our system's determinism, which results in a low likelihood of getting the answer incorrect rather than zero, is one improvement. By giving a system that is twice as quick as theirs but takes up twice as much storage space, we also offer a new operational point in the space-time tradeoff. This argument is supported by the fact that, according to current cloud computing pricing models, storage is less expensive than compute. In addition, our data encoding enables us to handle a wider range of queries than only those that require a precise match between the query and every database sequence. Additionally, we employ aggregate queries when looking up DNA information online.

Figure 1: Architecture for systems

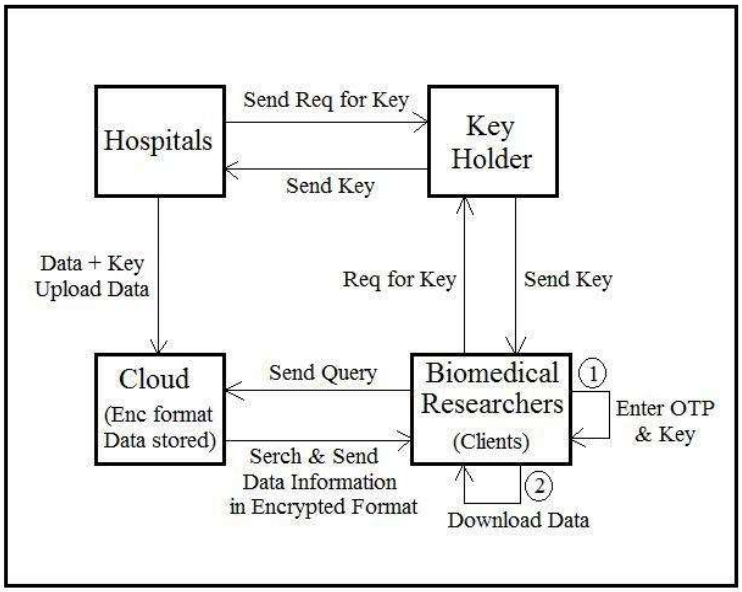

Figure 1 up top depicts the architecture of our system. The key modules are Hospitals, Key Holder, Cloud, and Biomedical Researchers, or Clients. The cloud stands for the data repository that houses all of the encrypted DNA records and is in charge of handling queries. The private and public keys for the homomorphic encryption technique are generated and kept by the Key Holder, a reliable third party. In order to encrypt their DNA records and upload them to the

cloud, the hospitals send a request for a key to the key holder and receive one in return. Biomedical researchers posing as clients send Cloud a question. In order to decrypt the results, clients must get them from the Cloud once it has processed the query over the encrypted records. Before sending out the findings on individual records, Cloud is necessary. If the order of the records may ever be related to any protected information, the permutation safeguards the records. Our system received a request from the client for an OTP (One Time Password). The system then produces and sends the OTP to the client. Additionally, the client sends the key holder a request. Key Holder produces and sends key to Client. The client then inserts the key and OTP, and the data or file will then download.

Conclusion

In order to facilitate extensive biomedical research projects, we have reviewed the problem of sharing person-specific genomic sequences without invading the privacy of their data subjects in this study. We used the additive homomorphic encryption-based framework Kantarciogluet al. [1] suggested.

We employed two cloud servers in our work; one is referred to as the key holder and is used to store the keys, while the other is referred to as the cloud server itself and is used to store the encrypted records. We create this project with biomedical researchers in mind because they are our clients. Additionally, we have a hospital module that uses our technology to store DNA database in the cloud. The suggested solution covers new types of queries that were not supported by earlier work and provides two new operating points in the space-time tradeoff. A crucial and useful prerequisite for biological researchers, the approach also supports an extended alphabet of nucleotides.

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. Security.*, vol. 6, no. 6, pp. 403–415, 2007.