ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Securing Aggregate Queries for DNA Databases in Cloud Server

T.Hema Sree
Department of Computer Science and Engineering,
MJR College of Engineering & Technology,
Piler, A.P - 517214, India.

G.Varaprasad
Department of Computer Science and Engineering,
MJR College of Engineering & Technology,
Piler, A.P - 517214, India.

Karamala Suresh
Department of Computer Science and Engineering,
MJR College of Engineering & Technology,
Piler, A.P - 517214, India.

ABSTRACT:

The matter of sharing individual person genomic sequence arrangements without providing the security of their information to support large scale biomedical research projects. In any case, expands the outcomes in various ways. One change is that our plan is deterministic, with zero likelihood of a wrong answer. This approach is proven effective in maintaining the privacy constraint against an adversarial server. We introduce cryptographic privacy for queries that allow to performing the most common DNA based identity. The capacity is very less expensive than calculation in current distributed computing evaluating plans. 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, including: (i) calculate the number of matches between query symbols and a sequence; (ii) logical OR matches where a query symbol is allowed to match a subset of the alphabet thereby making it possible to handle (as a special case) a "not equal to" requirement for a query symbol (e.g., "not a G"); (iii) support for the extended alphabet of nucleotide base codes that encompasses ambiguities in DNA sequences (this happens on the DNA sequence side instead of the query side); (iv) queries that specify the number of occurrences of each kind of symbol in the specified sequence positions (e.g., two 'A' and four 'C' and one 'G' and three 'T', occurring in any order in the query-specified sequence positions); (v) a start query whose answer is 'yes' if the number of matches exceeds a query-specified threshold. (vi) For all query types we can hide the answers from the decrypting server, so that only the client learns the answer. (vii) In all cases, the client deterministically learns only the query's answer, except for query type (v) where we quantify the (very small) statistical leakage to the client of the actual count.

Keywords: ***DNA Databases, Cloud Security, Secure Outsourcing.***

I. INTRODUCTION

DNA or Deoxyribonucleic Acid is the medium of life-long storage and transmission of genetic information for all contemporary living organisms. Human DNA data (DNA sequences within the 23 chromosome pairs) are private and sensitive personal data. However, such information is critical for organize biomedical research and studies, for example, diagnosis of pre-disposition to implement 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 heavy computation and storage capacity of cloud services enables practical hosting and sharing of DNA databases and efficient processing of genomic sequences, such as performing sequence comparison, perfect and approximate sequence search and numerous tests (diagnosis, identity, ancestry and paternity). What is missing is an efficient security layer that preserves the privacy of individuals' documents and assigns the load

Cite this article as: T.Hema Sree, G.Varaprasad & Karamala Suresh, "Securing Aggregate Queries for DNA Databases in Cloud Server", International Journal & Magazine of Engineering, Technology, Management and Research, Volume 5 Issue 7, 2018, Page 46-50.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

of query processing to the cloud. Whereas anonymization techniques such as de-identification [2], data augmentation [3], or database partitioning [4] solve this issue partially, they are not sufficient because in many cases, re-identification of persons is possible [5]. It follows that the DNA data must be protected, not just unlinked from the corresponding persons.

In this paper, we deal with the framework proposed in [1] where the DNA documents coming from few hospitals are encrypted and keep the data at a data storage site, and biomedical researchers are able to submit aggregate counting queries to this site. Counting queries are particularly interesting for statistical analysis.

This proposed schema provides a new method that addresses a larger set of problems and provides a faster query reply time than the technique introduced in [1]. Our approach is based on the fact that, given current pricing plans at many cloud services vendors, storage is cheaper than computing. Therefore, we favour storage through computing resources to optimize cost. Moreover, from a customer experience point of view, reply time is the most tangible indicator of performance; hence it is natural to aim at reducing it. Our method increases the state of the art at both the conceptual level and the implementation level.

II. EXISTING SYSTEM

To support large-scale biomedical research projects, management 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 de-identified genomic data can be "re-identified" to named individuals using simple automated methods. What is missing is an efficient security layer that preserves the privacy of individuals' documents and assigns the load of query processing to the cloud. Existing systems are not good because in many cases. It follows that the DNA data must be protected, not just unlinked from the corresponding persons. Several works focus on protecting biometric computations over genomic sequence records in the context of secure multi-party computations (SMC). Secure outsourcing is a particular case of SMC where a client with low resources (energy, memory, CPU) requests the service of one or more outsourcing agents with abundant resources. Secure outsourcing finds a real projection in the current business models thanks to the proliferation of cloud-based services. Cloud computing and storage security problems have been subject to ostensive research in the past decades [6]. Areas of interest include client authentication, hardware virtualization hazard, flooding and denial of service attacks as well as issues of accountability, storage protection and computation protection. In the context of DNA data protection, related works can be classified into five groups depending on the function or the query being addressed: forensic databases, profile matching, sequence comparison, testing by finite automata and aggregate queries.

A. Forensic databases: In a forensic database, a doubtful record has to be tested across an entire database. A document of the database can be decrypted only if it matches the doubtful record. This protects the other records from being unveiled [7].

B. Profile matching: In [9] the authors address a multitude of tests such as identity, ancestry and paternity tests based on Short Tandem Repeat (STR) profiles. The authors distribute each test into an algebraic expression and provide a homomorphic encryption scheme allowing two semi-honest persons to compare their stored profiles in a semantically secure manner. The proposed approach allows specific answers or small error tolerance as practically required by the tests.

C. Sequence comparison: The edit distance is the optimal cost of insertion, deletion and substitution of characters to go from a sequence λ to a sequence μ. The edit script is the chart of the steps leading to the optimal edit distance. Atallah et al. [10] offers a solution for securely outsource a dynamic programming solution for

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

finding the edit distance and the edit script for two given sequences. In [11] their scheme has been improved for performance and requires space only linear in the input size. The work in [12] addresses a similar dynamic programming solution for finding the longest common subsequence.

D. Sequence testing by finite automata: Sometimes the queries on DNA require taking into account various errors such as irrelevant mutations, incomplete specifications and sequencing errors. Therefore, the pattern of the query should be expressed using regular expressions. Many works address practical and privacy-preserving outsourcing of this regular expression type of queries, implemented as oblivious evaluation of automata.

E. Aggregate queries: For biomedical researchers, essential queries have often the form "How many documents contain a diagnosis of Alzheimer disease and gene variant X?" Secure outsourcing of the database and allowing such type of queries without requiring the server to decrypt the data has been addressed in [1]. The paper presents very practical results. For example, a count query over 40 records in a database of 5000 records takes 30 minutes.

III PROPOSED SYSTEM

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 trade off, 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.

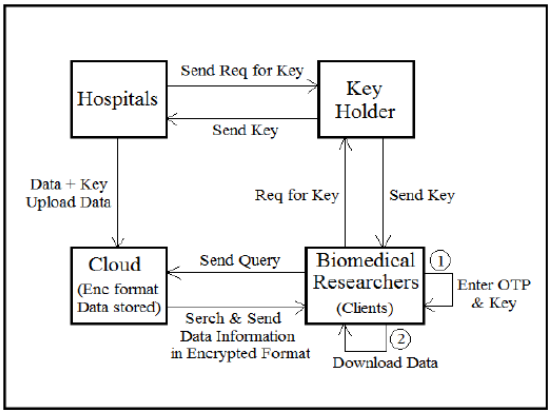

Figure 1: System Architecture

The architecture of our proposed 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 information storage where all the encrypted DNA records are stored and is responsible of processing the queries. Key Holder is a trusted party that generates and keeps the private and public keys of the homomorphic encryption scheme. The hospitals send the request for key to Key Holder and earn 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 through the encrypted records and sends the results which are in encrypted format to Clients in order to be decrypted. Cloud is needed to permute the results for individual records before sending them out. The permutation secures the records if in any case the order of the records can be linked to some protected data. 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 very safely.

IV CONCLUSION

In this paper, we have revisited the challenge of sharing person-specific genomic sequences without violating the

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

privacy of their data subjects in order to support large-scale biomedical research projects. We have used the framework proposed by Kantarciogluet 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 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 trade off 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.

V 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] Z. Lin, A. B. Owen, and R. B. Altman, "Genomic research and human subject privacy," Science (80-.)., vol. 305, no. 5681, p. 183, 2004.

[4] 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.

[5] L. Sweeney, A. Abu, and J. Winn, "Identifying Participants in the Personal Genome Project by Name," Available SSRN 2257732, 2013.

[6] 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.

[7] 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.

[8] 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.

[9] F. Bruekers, S. Katzenbeisser, K. Kursawe, and P. Tuyls, "Privacy-preserving matching of dna profiles," IACR Cryptol. ePrint Arch., vol. 2008, p. 203, 2008.

Author Details

T.Hema Sree Pursuing M.Tech at MJR College of Engineering & Technology, Department of CSE, Piler, Chittoor Dist.

Varaprasad Gajjala Working as a Assistant Professor in MJR College Of Engineering And Technology, Piler, A.P. and he obtained Graduation Degree in Computer Science and Engineering from JNTUA university in 2012, and obtained his Master's degree in 2014 specialized in Artificial Intelligence under Department of Computer Science & Engineering JNTUA college of Engineering, Anantapuramu, India. He made robot based on Artificial Intelligence while pursuing Master's. His research interest includes Web Designing, Information Security, Image Processing, Natural Language

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Processing, Robotics, and Artificial Intelligence, **Email:**gajjala.varaprasad@gmail.com.

Karamala Suresh Working as a Head of the Department in MJR College Of Engineering And Technology, Department Of Cse, Piler, Chittoor dist. He is having 14 years of teaching experience in engineering colleges, he received B.Tech(CSE) from JNTU Hyderabad in 2002,Received M.E(CSE) from satyabama university Chennai in 2006, and Received M.Tech(CSE) in 2015 from JNTU Anantapuramu, he published several research papers in various national and international journals. He is interested in Computer Networks.