International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Swift Retrieval of DNA Databases by Aggregating Queries

S. Dhanalakshmi[1], R.S. Bhavya[2], P.M. Dharini priya[3], M. Sangeetha[4]

[1,2,3]B.E .Students, Department of Computer Science And Engineering, Panimalar Engineering College, Poonamallee, Chennai, Tamil Nadu, India.
[4]Associate Proffesor, Department of Computer Science and Engineering, Panimalar Engineering College, Poonamallee, Chennai, Tamil Nadu, India.

Abstract: *This paper addresses the problem of sharing person-specific genomic sequences without contravening the privacy of their data subjects to support large-scale biomedical research projects. The proposed method builds on the framework but extends the results in a number of ways. One improvement is that our scheme is deterministic, with zero probability of imprecise results. We also deployed a new operating point in the space-time substitution, by offering a scheme that is twice as fast as theirs but uses twice the storage space. This point is prompted 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 manage a richer set of queries than exact matching between the query and each sequence of the database, including: (i) Counting the number of rivals between the 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 a "not equal to" requirement for a query symbol (iii) Support for the extended alphabet of nucleotide base codes that encompasses ambiguities in DNA sequences (iv) Queries that indicate the number of appearances of each kind of symbol in the specified sequence positions (v) A threshold query whose answer is 'yes' if the number of matches exceeds a query-specified threshold. (vi) For all query types we can conceal the answers from the deciphering server, so that only the client learns the answer. (vii) In all cases, the client conclusively learns only the query's answer, except for query type (viii) where we fathomed the (very small) statistical leakage to the client of the actual count.*

Keywords: DNA, Cloud Computing, Fast Access, Database, Sequences.

1. INTRODUCTION

Cloud computing facilitates the storage and management of huge amounts of data. It is the on-demand delivery of compute power, database storage, applications, and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing. Cloud storage is a paradigm of computer data storage in which the digital data is stored in logical pools. The physical storage spans multiple servers in multiple locations and the physical environment is typically owned and managed by a hosting company. Cloud computing offers significant research and economic benefits to healthcare organizations. It provides a safe place for storing and managing large amounts of such sensitive data. Under conventional flow of gene information, gene sequence laboratories send out raw and deduced information via Internet to several sequence libraries. DNA sequencing storage costs will be minimized by use of cloud service. By DNA analysis now becoming economical, more swiftly than data computation or data storage, the genome informatics is migrating to the cloud. The knack of cloud computing is to search for common patterns and to generalize results and to accelerate the development of treatments and diagnostic tools.

Genomics deals with the study of genomes which involves the sequencing, mapping and analysis of genomes. Cloud computing in genomics is a scalable service, where genetic sequence information is stored and processed virtually usually via networked, large-scale data centers accessible remotely through various clients and platforms over the Internet. With the use of this system, the stored DNA sequences can be accessed by the healthcare organizations with high speeds and genome sequences are used for prediction of diseases and drug designs.

The existing system is human DNA data (DNA sequences within the 23 chromosome pairs) are private and sensitive personal information. However, such 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 encouraging research in this field is mainly confronted by privacy concerns. Today, the ample 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, exact and approximate sequence search and various tests (diagnosis, identity, ancestry and paternity). 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. Whereas unidentified techniques such as de-identification, data augmentation, or database partitioning solve this problem partially, they are not sufficient because in many cases, re-identification of persons is possible.

There is no universal method to create a protocol for secure multi-party computation and handling aggregate queries on encrypted data. Several holomorphic systems only support a subset of mathematical operations, like addition, or exclusive-or. From a security perspective, only the additive and the multiplicative are classified to be IND-CPA (stands for in distinguishability under chosen plaintext attack). Partially holomorphic cryptosystems are more desirable from a performance point of view than somewhat holomorphic cryptosystems, which support a limited operation depth. Fully holomorphic systems have a huge cost and cannot be deployed in practice.

This paper provides a new method that addresses a huge set of problems and provides a faster query response time than the technique introduced. Our stratagem is based on the fact that, given current pricing plans at many cloud services providers, storage is cheaper than computing. Therefore, we favor storage over computing resources to reduce cost. Moreover, from a user experience perspective, response time is the most tangible indicator of performance; hence it is natural to aim at reducing it. Moreover, our ciphering 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. My sql is the database used in this project and for connectivity JDBC is employed. Linear and logistic regression algorithms are used. Linear regression is a common Statistical Data **Analysis** process. In **this technique** a single independent variable is used to predict the value of a dependent variable. Logistic regression is a statistical method for examining a dataset in which there are one or more independent variables that determine an outcome. The result is measured with a dichotomous variable (in which there are only two possible outcomes). In logistic regression, the dependent variable is a binary value or a dichotomous value i.e. it only contains data coded as 1 TRUE, success, pregnant or 0 FALSE, failure, non-pregnant.

2. LITERATURE SURVEY

Wei Jiang, Ying Liu and other team members of the paper, “A Securely Share and Query Genomic Sequences” have apprised that many organisations and institutions grant encrypted genomic sequence records into a incorporate repository, where the administrator can perform queries, such as frequency counts, etc., without decrypting the data. Here the competence of the structure and framework is assessed with the actual and existing database of SNP i.e., Single Nucleotide Polymorphism and manifest that the era required to outright count queries is expedient for the applications of the real world. Here consider this experiment which illustrate that count query above SNP’s of 40 in a database of 5000 records could be finished more or less relatively in half an hour along with the off-the-shelf mechanics. The dissemination of the aspect of the third party afford further prosperity. Precisely consider that the maximum of data administration is bumped onto DS, whereas KHS serves as a last point of control in the structure. With the status of the KHS this project endorse that the authentic third party i.e., the NIH plays the part of KHS. Now that the participants are mapped from the original sequence of events to the roles in the secure framework and the question is left away, which could be accomplished by identifying that DS is curbed by many circumstances. At first the DS must be a trusted entity with no data of its own at stake, and so there is no strife of enthusiasm. And the second is that the DS must have enough and proper storage and a bandwidth capability in administrating the large databases with access at the same time. This role could be expected by particular and specialized Information management corporation that is juridically bound to DS for liability aspirations.

In IEEE International Conference the paper “Transforming Semi-Honest Protocols to Ensure Accountability” have implied that Secure multi-party computation (SMC) stabilises the benefit and confidentiality i.e., secrecy of the distributed data. This is notably substantial and valuable for privacy-preserving data mining (PPDM). Most secure multi-party computation protocols are only proven secure under the semi-honest model, providing insufficient security for many PPDM applications. SMC protocols under the mischievous antagonist model customarily have impractically high intricacy for PPDM. Here an accountable computing (AC) framework is proposed that empowers liability for privacy compromise to be entrusted to the responsible party without the complexity and cost of an SMC-protocol concealed by the malicious model. We show how to transform a circuit-based semi-honest two-party protocol into a simple and efficient protocol satisfying the AC-framework.

In “Protocols for secure communications” the author explores the problem of n people desires to gauge the value of a function $f(x_1, x_2, x_3, ..., x_n)$, which is an integer-valued function of n integer variables xi of bounded range. Assumption is that, at first person Pi knows the value of xi and no other x's. Is it possible for them to calculate the value of f, by interacting among themselves, without immensely giving away any information about the values of their own variables? The author gives a precise formulation of this general problem and describe three ways of solving it by use of one-way functions (i.e., functions which are easy to evaluate but hard to invert). These results have applications to secret voting, private querying of database, oblivious negotiation, playing mental poker, etc.. He also discusses the complexity question "How many bits need to be exchanged for the computation," and describes methods to prevent participants from cheating. Finally, he studies the question "What cannot be accomplished with one-way functions."

According to Andrew Chi-Chih Yao a new gizmo for regulating the knowledge transfer procedure in cryptographic protocol layout is introduced. It is applied to solve a prevailing league of problems which comprises most of the two-party cryptographic problems in the literature. Explicitly, it is shown how two parties A and B can bilaterally engender a random integer N = p·q such that its secret, i.e., the prime factors (p, q), is concealed from either party individually but is rectifiable jointly if chosen. This can be resort to to give a protocol for two parties with private values i and j to compute any polynomically computable functions f(i,j) and g(i,j) with minimal knowledge transfer and a firm fairness idiosyncrasy.

In the paper "Geonomic research and Human Subject privacy", Public genetic sequence databases are a detracting part of our academic biomedical research framework and infrastructure. However, human genetic data should only be made public if we can competently insulate the privacy of research prone subjects. Diacritic genomic sequence data (such as SNPs) are quite identifiable with the help of common definitions, while our intention to figure out disease susceptibility or therapeutic opportunity require approach to large genomic data sets. The authors of this policy forum argue that surprisingly minuscule amounts of genomic sequence data are identifiable. Therefore, the peculiar privacy challenges posed by genomic data is required to be addressed with new policies or innovative technical advances and approaches.

3. EXISTING SYSTEMS

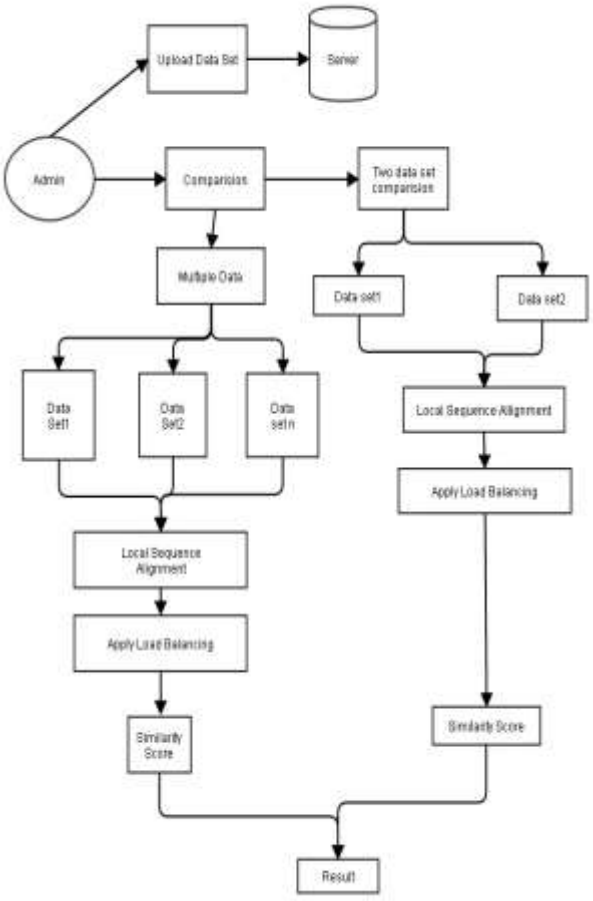
DNA data which consists of DNA sequences within the 23 chromosome pairs belonging to human are private and sensitive personal information. However, conducting biomedical research and studies on such type of DNA data is critical, for example, detection of pre-disposition to develop a specific disease, drug allergy, or prediction of success rate in response to a specific treatment. Providing a publicly available DNA database for espousing research in this field is mainly confronted by privacy concerns. Today, the profuse 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, exact and approximate sequence search and various tests (diagnosis, identity, ancestry and paternity). 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. This problem can be solved partially by using anonymization techniques such as de-identification, data augmentation, or database partitioning, they are not sufficient because in many cases, re-identification of persons is viable.

Disadvantage:

- In the authors address the longest common subsequence as a private search problem.
- In our model, hospitals that have DNA sequences do not have the computing and processing capabilities to process researchers requests, so they all store their DNA sequences at a server.
- We have presented two new operating points in the space-time tradeoff of the private query problem.

4. PROPOSED SYSTEM

This paper proposes an innovative method for addressing an ample set of problems and providing a faster query response time than existing systems. Our approach is based on the element that, storage is cheaper than computing, given current pricing plans at many cloud services providers. Therefore, favoring storage over computing resources reduces cost. Moreover, from a user experience perspective, response time is the most tangible indicator of performance. Hence it is natural to aim at reducing response time. Our method enhances the state of the art at both the conceptual level and the implementation level. Encoding of the data makes it possible to handle a richer set of queries than exact matching between the query and each sequence of the database.



Linear and logistic regressions:

Linear and logistic regressions are the algorithms are used in this method. Linear regression is a linear approach to simulating the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to elucidate the data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

5. 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 based on additive homomorphism encryption, and two servers: one holding the keys and one storing the encrypted records. 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. Big data analytics over genetic data is a superior future work direction. There are rapid recent advancements that address performance limitations of holomorphic encryption techniques. We hope that these evolutions will lead to more practical solutions in the future that can handle larger-scale genetics data. It is worth mentioning that our approach is not restricted to a fixed holomorphic encryption technique and therefore, it would be possible to use and inherit the advantages of newly developed ones.

REFERENCES

1) T. Hara, V. I. Zadorozhny, and E. Buchmann, Wireless Sensor Network Technologies for the Information Explosion Era, Stud. Comput. Intell. Springer-Verlag, 2010, vol. 278.

2) Y. Wang, G. Attebury, and B. Ramamurthy, "A Survey of Security Issues in Wireless Sensor Networks," IEEE Commun. Surveys Tuts., vol. 8, no. 2, pp. 2–23, 2006.

3) Abbasi and M. Younis, "A survey on clustering algorithms forwireless sensor networks," Comput. Commun., vol. 30, no. 14-15, pp. 2826–2841, 2007.

4) W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "AnApplication-Specific Protocol Architecture for Wireless MicrosensorNetworks," IEEE Trans. Wireless Commun., vol. 1, no. 4, pp. 660–670, 2002.

5) Manjeshwar, Q.-A.Zeng, and D. P. Agrawal, "An Analytical Modelfor Information Retrieval in Wireless Sensor Networks Using Enhanced APTEEN Protocol," IEEE Trans. Parallel Distrib. Syst., vol. 13, pp.1290–1302, 2002.

6) S. Yi, J. Heo, Y. Cho et al., "PEACH: Power-efficient and adaptiveclustering hierarchy protocol for wireless sensor networks," Comput. Commun., vol. 30, no. 14-15, pp. 2842–2852, 2007.

7) K. Pradeepa, W. R. Anne, and S. Duraisamy, "Design and Implementation Issues of Clustering in Wireless Sensor Networks," Int. J. Comput. Applications, vol. 47, no. 11, pp. 23–28, 2012.

8) L. B. Oliveira, A. Ferreira, M. A. Vilac¸a et al., "SecLEACH-On the security of clustered sensor networks," Signal Process., vol. 87, pp. 2882–2895, 2007.

9) P. Banerjee, D. Jacobson, and S. Lahiri, "Security and performance analysis of a secure clustering protocol for sensor networks," in Proc. IEEE NCA, 2007, pp. 145–152.

10) K. Zhang, C. Wang, and C. Wang, "A Secure Routing Protocol for Cluster-Based Wireless Sensor Networks Using Group Key Management,"

11) M. Blanton, M. M. J. Atallah, K. B. K. Frikken, and Q. Malluhi, "Secure and Efficient Outsourcing of Sequence Comparisons," Comput. Secur. 2012, pp. 505–522, 2012.

12) M. Franklin, M. Gondree, and P. Mohassel, "Communication-efficient private protocols for longest common subsequence," in Topics in Cryptology--CT-RSA 2009, Springer, 2009, pp. 265– 278.

13) M. Gondree and P. Mohassel, "Longest common subsequence as private search," in Proceedings of the 8th ACM workshop on Privacy in the electronic society, 2009, pp. 81–90.

14) D. Szajda, M. Pohl, J. Owen, B. Lawson, and V. Richmond, "Toward a practical data privacy scheme for a distributed implementation of the SmithWaterman genome sequence comparison algorithm," in Proceedings of the 12th Annual Network and Distributed System Security Symposium (NDSS 06), 2006.

15) 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.

16) J. R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik, "Privacy preserving error resilient dna searching through oblivious automata," in Proceedings of the 14th ACM conference on Computer and communications security, 2007, pp. 519–528.

17) K. B. Frikken, "Practical private DNA string searching and matching through efficient oblivious automata evaluation," in Data and Applications Security XXIII, Springer, 2009, pp. 81–94.

18) K. Kozl and C. Listy, "Biochemical nomenclature and related documents," Chem. List., vol. 72, pp. 288–305, 1978.

19) P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in Proceedings of the 17th international conference on Theory and application of cryptographic techniques (EUROCRYPT'99) , 1999, pp. 223–238.

20) D. Chaum, "Blind signatures for untraceable payments," in Advances in cryptology, 1983, pp. 199– 203.