ISSN:0975-9646

A Cryptographic Approach to Securely Share Genomic Sequences

G. Rajendra Kumar, V.Navya, M.Swetha, V.Sindhusha, T. Murali Kalyan
Department of Information Science and Technology, KL University, Guntur dt Andhra Pradesh, 522502, India

***Abstract*— Many basic tasks in computational biology involveoperations on individual DNA and genomic sequences. These sequences, even when anonymized, are vulnerable to re-identification attacks and may reveal highly sensitive information about individuals. To support large-scale biomedical research projects, organizations need to share person-specific genomic sequences without violating the privacy of their data subjects. We present a relatively efficient, privacy-preserving implementation of fundamental genomic computation without disclosing the raw genomic sequences. Organizations contribute encrypted genomic sequence records into a centralized repository, where the administrator can perform queries, without decrypting the data.**

***Keywords*—Databases, genomics, privacy, security.**

I. INTRODUCTION

Genomic data such as DNA and protein sequences are increasingly collected by government agencies for law enforcement and medical purposes, disseminated via public repositories for research and medical studies, and even stored in private databases of commercial enterprises. For example, deCODE Genetics aims to collect the complete genome sequences of the entire population of Iceland, while the non-profit HapMap Project is developing a public repository of representative genome sequences in order to help researchers to discover genes associated with specific diseases. The underlying genome records are typically collected from specific individuals, and thus contain a lot of sensitive personal information, including genetic markers for diseases, information that can be be used to establish paternity and maternity, and so on. Therefore, genomic records are usually stored in an anonymized form, that is, without explicit references to the identities of people from whom they were collected. However, to realize cost-effective specialized services, scientists need to characterize the influence of genomic variation over a wide array of health features, such as clinical diagnostics and treatment response. To facilitate data sharing, organizations in various countries, including Estonia, Iceland, Japan, Mexico, Norway, Sweden, the United Kingdom, and the United States are establishing data repositories that centralize person-specific biomedical records for research purposes. Despite the potential benefits to health care, person-specific genomic records must be shared in a manner that preserves the anonymity of the data subjects. This requirement is rooted in both social concerns and public policy. Many people fear that sensitive information learned from their medical and genomic records will be misused or abused. To mitigate such concerns, many countries have enacted policies that limit the sharing of a subject's genomic information in a personally identifiable form. In the United States, for instance, the National Institutes of Health (NIH) is drafting policy that will require scientists to share genomic data studied with NIH funding once "identifiable information" has been removed. Consider the following scenario. John is a principle investigator located at the University of Texas Southwestern Medical Center and Mike is a principle investigator located at the Vanderbilt University Medical Center. Both John and Mike are independently funded by the NIH to collect data from hospital patients and conduct genome wide association studies on Alzheimer's disease. To comply with the NIH policy, at the completion of their studies, John and Mike need to share their data collections to a centralized repository, so that researchers around the country, such as Charlie at the National Institute on Aging can perform scientific investigations on the integrated data, such as "How many records contain a diagnosis of juvenile Alzheimer's and gene variant X?" How can John and Mike share the biomedical records so that biomedical researchers can conduct their scientific investigations without revealing the identities of the data subjects? To summarize the problem, data collectors, such as John and Mike need to satisfy two goals when sharing genomic data:

1) Data utility: the data should be useful for scientific investigations;
2) Data privacy: the data should not reveal the subjects' identities.

Often, these goals are considered to be contradictory and existing privacy methods tend to favor one over the other. In this paper, however, we demonstrate that utility and privacy goals can be simultaneously satisfied for specific scientific endeavors.

II. GENOMIC DATA PRIVACY TECHNIQUES

To date, various privacy protection strategies have been designed to remove identifying information prior to

sharing genomic data. For the most part, existing genomic data privacy techniques. can be roughly grouped into two approaches with distinct benefits and drawbacks: 1) data deidentification and 2) data augmentation. Privacy protections based on "deidentification" advocate the removal, or encryption, of person-specific identifiers, such as name and social security number, initially associated with genomic records. Deidentification enables data collectors to disclose all genomic information that has been collected, but it is an ad hoc process and provides no guarantees of privacy protection. In fact, it was recently shown that in many cases, knowledge gleaned from deidentified genomic data can be exploited to "reidentify" records to named subjects in publicly accessible resources through simple automated methods. Data augmentation techniques provide exact guarantees of privacy protection. As an example, consider that a prime factor in reidentification is that a subject's DNA is uniquely distinguishable. In particular, experimental evidence indicates that less than 75 single nucleotide polymorphisms (SNPs), features common to genomic studies, are sufficient to uniquely distinguish a subject's DNA record in a population. A formal model of privacy protection that addresses uniqueness is the generalization of a subject's DNA sequence so that the resulting record is indistinguishable from other shared records. For instance, the DNA sequences AACTAA and AAGTAC can be generalized to the common AA[C or G]TA[A or C]. Privacy protection based on generalization is controlled by varying the number of records that are rendered indistinguishable. Though generalization formally prevents data reidentification, it changes the genomic records in ways that may limit their scientific usefulness.

III. CRYPTOGRAPHY ARCHITECTURE

In this paper, we propose an alternative approach to genomic data privacy protection that is based on cryptography. Our model ensures that: 1) the data utility of protected records is equivalent to that achieved by deidentification and 2) the data privacy is equivalent to that achieved by data augmentation schemes. As an overview, our model works as follows. Data holders John and Mike transmit encrypted versions of their records to a third party's data repository. The repository administrator executes queries on behalf of Charlie the researcher without decrypting any of the records. The results of the query are then sent to a third party who decrypts the aggregation of the result and sends the answer to the scientist. This architecture incorporates two different third parties for security-related benefits. There is no opportunity to decrypt the data unless both third parties collaborate. As a result, the use of multiple third parties ensures that there is no single point of data compromise. Thus, if a hacker breaks into one of the third party's computer systems, the hacker cannot learn the sensitive information in the encrypted records. Recognize that though the data remains encrypted at all times, the results of queries themselves can violate privacy requirements. For instance, if the answer to Charlie's query is such that there is only one record with DNA sequence "AATCAATGAA" and juvenile Alzheimer's disease, then Charlie has uniquely pinpointed an individual's record. Thus, it is necessary for the third party to ensure that query results, or the combination of a series of query results issued by a researcher, do not permit the triangulation of an individual's record. This process, known as query restriction, is necessary to ensure that our framework achieves identity protection; however, this topic has been studied extensively in the database security community, and thus, we neglect the presentation of query restriction in this paper. The main contribution of our model is in the analysis of encrypted genomic data. To the best of our knowledge, there is no off-the-shelf product or literature that can be applied to satisfy this component of the framework. As such, this paper focuses on the cryptographic protocols that are necessary to build and query encrypted genomic databases. In addition, we provide experimental validation so that in our framework, queries can be answered efficiently for real world biomedical applications.

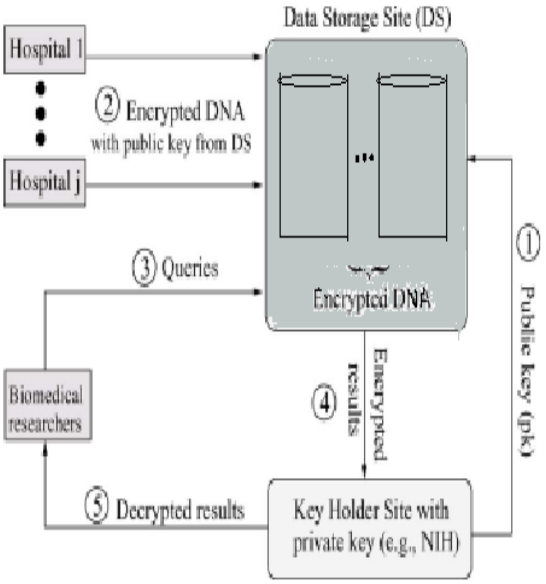


Fig:1. Architecture of Proposed System

IV. METHODS

The goal of our research is to create a system that simultaneously: 1) stores DNA sequences in a database securely, 2) supports querying tasks that would be performed on the original sequences, 3) facilitates the DNA data holders to submit their records to our system

without ever knowing the secret keys that can be used to decrypt the encrypted data, and 4) prevents a single point of failure to ensure that if a hacker breaks into any single site, he/she will not be able to learn the confidential DNA data. To achieve these goals, we designed an architecture that incorporates four types of participants: data holders, data users, a DS, and a KHS. In Fig. 1, we depict the relationship of these participants and a broad overview of the architecture. Imagine that the set of data holders are a set of hospitals and that the set of data users are biomedical researchers. For this research, we assume each hospital maintains one or more DNA records and that all hospitals collect records on the same set of attributes (i.e., the same regions of the genome). Recall, in the earlier scenario, the data holders need to share the data with a third party for public dissemination purposes, which in the context of genome wide association studies in the United States will be the NIH. Yet, notice that in our framework, we incorporate two third parties: a data storage site (DS) and a key holder site (KHS). The additional third party is crucial to the security of the framework. The DS is where encrypted DNA is stored and processed, whereas KHS manages the cryptographic keys that are used for encryption and decryption of the genomic records stored in a database at DS, as well as the query results to biomedical researchers. Thus, if one of the third parties is , the decrypted DNA records are not revealed.

V. CRYPTOGRAPHIC BASICS

A)Secure Hash Algorithm

For the implementation of architecture we are using the SHA-1 algorithm. The Secure Hash Algorithm (SHA) was developed by National Institute Of Standards and Technology (NIST) and published as a federal information processing standard (FIPS 180) in 1993; and is generally referred as SHA-1. SHA is based on the MD4 algorithm and its design closely models MD4.

SHA-1 Logic

The algorithm takes as input a message with a maximum length of less than 2^{64} bits and produces as output a 160-bit message digest. The input is processed in 512-bit blocks. The overall processing of a message follows the structure as MD5 with a block length of 512 bits and a hash algorithm and chaining variable length of 160 bits. The processing consists of the following steps:

Step 1: Append Padding bits: The message is padded so that its length is congruent to 448 module 512 (length = 448 mod 512). Padding is always added, even if the message is already of 1 to 512. The padding consist of a single 1-bit followed by necessary number of 0-bits.

Step 2: Append Length: A block of 64 bits is appended to the message. This block is treated as an unsigned 64-bit integer and contains the length of the original message (before the padding).

Step 3: Initialize MD buffer: A 160-bit buffer is used to hold intermediate and final results of the hash function. The buffer can be represented as five 32-bit registers (A, B, C, D, E). These registers are initialized to the following 32-bit integers(hexadecimal values):

A=67452301
B=EFCDAB89
C=98BADCFE
D=C3D2E1F0

Note that the first four values are the same as those used in MD5. However, in the case of SHA-1, these values are stored in big-endian format, which is the most significant byte of word in the low-address byte position. As 32-bit strings, the initialization values appear as follws.

Word A: 67 45 23 01
Word B: EF CD AB 89
Word C: 98 BA DC FE
Word D: 10 32 54 76
Word E: C3 D2 E1 F0

Step4: Process message in 512-bit (16-word) blocks: The heart of the algorithm is a module that consists of four rounds of processing of 20 steps each. The logic is illustrated in the Figure below. The Four rounds have a similar structure, but each uses a different primitive logical function, which we refer to as f_1, f_2, f_3, and f_4.

Each round takes as input the current 512-bit block being processed (Y_q) and the 160-bit buffer value ABCDE and updates the contents of the buffer. Each round also makes use of an additive constant k_t, where $0<=t<=79$ indicates one of the 80 steps across five rounds. In fact, only four distinct constants are used. The output of the fourth round (eightieth step) is added to the first round (CV_q) to produce CV_{q+1}.

The addition is done independently for each of the five words in the buffer with each of the corresponding words in CV_q, using addition module 2^{32}.

Step 5: Output: After all L 512-bit blocks have been processed, the output from the Lth stage is the 160-bit message digest.

We can summarize the behavior of SHA-1 as follows:

CV_0=IV
CV_{q+1}=SUM_{32} (CV_q, $ABCDE_q$)
MD=CV_L

Where

IV = initialize value of the ABCDE buffer.

ABCDEq= the output of the last round of processing of the qth message block

L = the number of blocks in the message (including padding and length fields)

SUM_{32} = Addition modulo 2^{32} performed separately on each word of the pair of inputs

MD =final message digest value

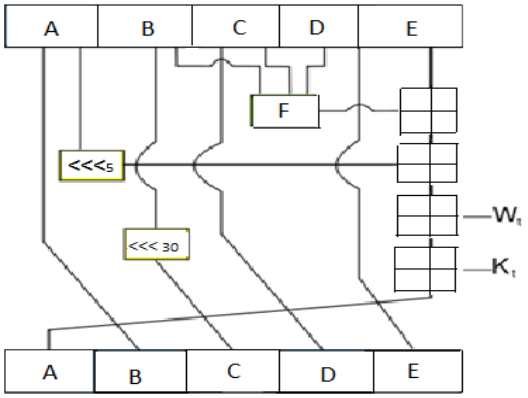
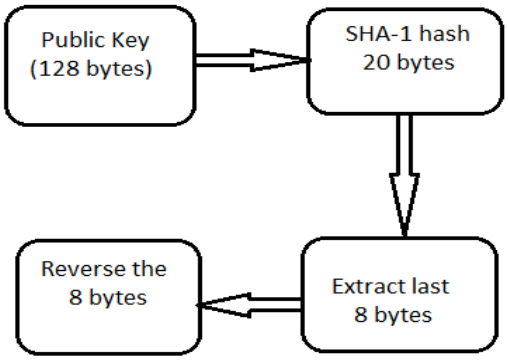


Fig:2 SHA Operation(single step)

B. MAC(Message authentication code)

After applying the SHA-1 on the message we get the message digest, then we apply the MAC for the purpose of double encryption, so that to enhance the security of the message.



VI .Conclusion And Future Scope

Even though cryptography is a ground-breaking technology, de-identification of data is the bottle neck for the security of genomic sequences, so we providing security by using the sha algorithm.

Our futures work aims at further expanding the security as well as speed up the query execution times with minimal loss in accuracy.

Acknowledgement

We would like to extend our sincere thanks to Prof. Dr. V. Srikanth HOD, Department of Information Science and Technology, K.L. University, for the untiring help and encouragement. It is our extreme pleasure to acknowledge our thanks to the authors of previous journals and papers which aided us to put a new perspective in the field of Cryptography and Security.

References

[1] M. West, G. Ginsburg, A. Huang, and J. Nevins, "Embracing the complexity of genomic data for personalized medicine," Genome Res., vol. 16, pp. 559–566, May 2006.

[2] W. Evans and M. Relling, "Pharmacogenomics: Translating functionalgenomics into rational therapeutics," Science, vol. 286, pp. 487–491, 1999.

[3] A. Roses, "Pharmacogenetics and pharmacogenomics in the discovery and development of medicines," Nature, vol. 38, pp. 815–818, 2000.

[4] U. Sax and S. Schmidt, "Integration of genomic data in electronic health records—Opportunities and dilemmas," Methods Inf. Med., vol. 44, pp. 546–550, 2005.

[5] D. Gurwitz, J. Lunshof, and R. Altman, "A call for the creation of personalized medicine databases," Nature Rev. Drug Discov., vol. 5, no. 1,pp. 23–26, 2006.

[6] Anonymous, "Medicine's new central bankers," The Economist, vol. 377,no. 8456, pp. 28–30, Dec. 2005.

[7] V. Barbour, "UK Biobank: A project in search of a protocol?," Lancet, vol. 361, pp. 1734–1738, 2003.

[8] E. Clayton, "Ethical, legal, and social implications of genomic medicine," New England J. Med., vol. 349, pp. 562–569, 2003.

[9] M. Rothstein and P. Epps, "Ethical and legal implications of pharmacogenomics," Nature Rev. Genetics, vol. 2, pp. 228–231, 2001.

[10] B. Malin, "An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future," J. Amer. Med. Inf. Assoc., vol. 12, no. 1, pp. 28–34, 2005.

[11] Z. Lin, A. Owen, and R. Altman, "Genomic research and human subject privacy," Science, vol. 305, no. 5681, p. 183, 2004.