International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Blockchain enabled DNA banking and comparative analysis using Hyperledger Fabric

Preethi V., Saurav Surve

Abstract: ***In the year 2001, the human genome was sequenced for the first time costing about $3 billion. Today, it is far cheaper to get the genome sequenced. Thus, this would lead to wider adoption of personal genome sequencing as it enables better diagnosis, disease prevention, and personalized therapies. If data is shared with researchers, the causes of many diseases will be identified and new drugs developed. Opportunities worth billions of dollars are being created and envisioned. There are generally steps being undertaken to protect our data and genetic privacy by the labs processing our genetic information. But they also could be used or misused for a variety of reasons. Cases of identity theft have been proved worrisome for various experts. Using the same databases, marketers could also target individuals with particular diseases; or in a more unnerving scenario, those whose genetic information suggests they are likely to develop certain illnesses in the future. The individual would have little to no control on the genetic data being stored and thus, security breach on such types of data could lead to revelation of an individual's family history. Thus, the project aims to enhance data protection, enable buyers to efficiently acquire data and address the challenges of big data. These goals will be accomplished through decentralization, cryptography, and its utilization.***

Index Terms: ***Blockchain, Database Security, DNA Banking, Hyperledger Fabric***

I. INTRODUCTION

Genomics is an area under genetics which studies the sequencing of an organism's genome. In fact, the whole set of genetic material of an organism is being studied in the field of genomics. Sequencing is a method used to determine the correct order of the four nucleotide bases – adenine (A), guanine (G), cytosine (C) and thymine (T). These four nucleotide bases make up a strand of DNA. These bases provide the underlying genetic basis, also called the genotype for telling a cell what to do. The outcome of these instructions result in the phenotype of an organism. About 3 billion pairs of Deoxyribonucleic Acid (DNA) bases are being distributed on 23 pairs of chromosomes. An important expansion of human DNA sampling and data collecting in order to exploit and study the genetic information collected. The strategic importance of this activity for genetic research and its applications is enormous in various fiels related to genetics. A DNA database is a database of DNA profiles which can be used in the analysis of hereditary diseases and other profiles related to the field of geneaology. It is also called DNA bank. DNA storage databases may be public or private, the largest ones generally collected by the government's DNA databases. DNA databases require more storage when compared to other type of databases. This is due to the enormous size of each DNA sequence. This has given rise to a major challenge to the storage, data transfer, retrieval and search of these databases. Companies like 23andMe, Ancestry.com, etc. provide consumer genomics services for individuals who want to conduct DNA testing on their samples. These companies also provide DNA sequencing data to healthcare and pharmaceutical companies who need such data for medicinal research regarding drug discovery and production, patents and commercialization of drugs invented. However, the individual consumer's consent is taken for supplying his/her DNA data to these enterprises. Companies like 23andMe provide facilities for storage of DNA databases. However, these databases are centralized in storage and prone to malicious attacks by a bad actor. Hence, the project aims to provide a blockchain based solution for storing these DNA databases using Hyperledger Fabric platform currently developed mainly by IBM.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Preethi V*, Department, of Computer Science and Engineering SRM Institute of Science and Technology (Ramapuram Campus) Chennai, India
Saurav Surve, Department, of Computer Science and Engineering SRM Institute of Science and Technology (Ramapuram Campus) Chennai, India

II. EXISTING SYSTEMS

Various private companies like 23andMe offer direct-to-consumer DNA banking system through its web-based platform. This web-based business model is summarized in the following steps:

1. Getting consent from the individual online and ordering the kit
2. Shipment of the DNA sample to the company
3. DNA sequencing and banking
4. DNA sample provided to the healthcare companies for genetic testing
5. Providing the testing results to the consumer through online platform

The individual consumer has to pay certain amount for conducting their DNA testing. On the other hand, as mentioned above, healthcare institutions and pharmaceutical enterprises also require lots of human genetic data for conducting research on drug discovery and production, patents issuance, research publication and commercialization of the drug invented. These companies pay huge amount of fee to such consumer genomics companies as these companies provide the storage of DNA databases.

Of course, these healthcare and pharmaceutical enterprises cannot collect or use an individual DNA data without requiring the individual's voluntary consent. Various countries around the world have formulated regulations regarding genetic privacy violations of these would lead to serious penalties for the enterprises. Fig.1 shows the traditional DNA banking business model. Thus, 23andMe company may considerably contains a two-sided platform, with two kinds of consumers: individuals who want information about their own genes for various reasons.

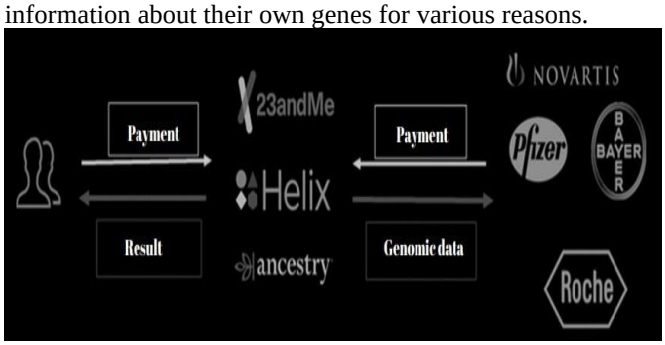

Fig.1 Traditional DNA banking business model

The other class of consumers are those researchers and large institutions and healthcare companies who want access to genetic, web behaviour and self-reported information for a large number of people[1].

III. ISSUES IN THE EXISTING SYSTEM

The traditional DNA banking system has raised certain problematic issues regarding ethics of genetic privacy and monetization of an individual's genetic data. Firstly, the issues of genetic discrimination has been controversially raised since these DNA banking facilities can map out important portions of the genome. Genetic discrimination takes place when people are are provided differently by their employer or insurance company only due to a presence of a gene mutation that causes or increases the risk of an inherited disorder. This is a common fear among people who are considering genetic testing. There has been an uncertain and risk-based regulatory environment for these gentic companies. Individuals also share a fear of identity theft since these companies could do that only due to lack of positive regulations.The company collects various types of data from their consumers. This includes mainly genetic and personal information from customers who order DNA tests. The data about users' web behaviour is also captured through cookies from their website as well as the company's smartphone application. Such data is useful not only for people belonging to the pharmaceutical field but also from those belonging to law enforcement and legal team. Consumers are generally not aware of their information being shared through such techniques. During data breaches this could get even worse. The centralized data repositories of such DNA banking services could be hacked and used for wrong intentions as provided above. The file size of a human genome for complete sequencing could take upto Gigabits of data storage which could make it for such companies to maintain data of billions of people in the future.

IV. PROPOSED SYSTEM

The DNA data of individuals should be stored on blockchain for protecting its integrity. Through this, the data will be stored on a growing list of records, which would be secured through a cryptographic hash, which would make it practically unhackable. The traditional business model encourages centralization of data storage and functionality towards DNA banking platforms like 23andMe. The blockchain based decentralized platform would eliminate these companies acting as middlemen for providing DNA banking services, thereby making individual consumer as the sole owner of their genetic data and cost-effective genetic testing. The blockchain based DNA sharing platform would enable Direct-To-Consumer DNA banking service, empowering them to earn money for sharing their genetic data. The project aims to utilise the Hyperledger Fabric platform, which is a permissioned blockchain infrastructure, originally contributed by IBM. The business model will be executed based on Smart Contract between the consumers and genetic testing laboratories. A smart contract is a self-executing contract used to validate, verify, or enforce the negotiation or performance of a contract. Smart contracts are to provide security that is superior to traditional contract law. This also helps in reducing other transaction costs associated with contracting. Blockchain is a distributed and immutable database, shared and automatically automated among all participants. This distributed database technology was first developed to be used as a public ledger in the popular decentralized cryptocurrency, Bitcoin. Today, blockchain is seen as a viable option for conducting business and operational transaction through the usage of the Smart Contracts. The most important features of the blockchain technology are decentralization, immutability and security. Decentralization means that no single entity or some few entities control the organization or the database in our project's aspects. Immutability doesn't allow for the past records to be changed. The security aspect of blockchain provides data protection through encrypted cryptographic hash function. The Bitcoin blockchain is encrypted through SHA-256 hah function. However, Bitcoin is just a blockchain platform for enabling financial transactions. We can't utilise the Bitcoin network for storing large amounts of data or build interactive applications. Thus, the concept of smart contracts emerged in the year 2013 when the Ethereum blockchain network's whitepaper was released. Ethereum is an open-sourced decentralized software platform based on blockchain technology that enables developers to build and deploy decentralized applications. Similar to Bitcoin, Ethereum is a distributed public blockchain network. The concept of Ethereum was devised by Russian-Canadian programmer Vitalik Buterin There are no significant technical differences between both Bitcoin and Ethereum but their purpose and scope are significantly varied. A peer to peer electronic cash system that enables online Bitcoin payments is being offered by Bitcoin. Bitcoin is used for transferring financial assets in a peer-to-peer fashion (bitcoin) but the Ethereum blockchain focuses on running the programming code of any decentralized application. However, Ethereum blockchain transaction speed is just 15-17 Tx/sec which makes it unable to build large-scale decentralized applications on its network. Moreover, all the transactions on the Ethereum blockchain is public and since we require a certain level of privacy in maintaining a genetic databank, the Ethereum platform could not be used for the same purpose.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

V. HYPERLEDGER PLATFORM AND PROJECT MODULES

On its website, Hyperledger is defined as "Hyperledger is an open source collaborative effort created to advance cross-industry blockchain technologies. It is a global collaboration, hosted by The Linux Foundation, including leaders in finance, banking, Internet of Things, supply chains, manufacturing, and Technology". A lot of projects are being developed in the Hyperledger, one of them being the Fabric. It was intended for developing solutions with a modular architecture. Hyperledger platform allows the components to be plug-n-play. It is a private as well as a permissioned Blockchain system. It is different from the public network systems where the unknown identities are allowed to participate in the network. In Fabric, the members enrol through Membership Service Provider (MSP). It also offers the ability to create channels, allowing a group of participants to create a separate ledger of transactions. The benefits of the Hyperledger Fabric are data protection and privacy through use permissions for membership and access rights, confidential transactions providing flexibility and security for conducting business operations and enabling smart contract to create applications catering to various business needs. The main advantage of Hyperledger Fabric platform is that it doesn't require the mandatory usage of cryptocurrencies as required in various other blockchain networks.

The architecture diagram for the project is given the Fig.2. The project can be divided into the following modules:

1. **Data Owner Nodes -**Data owner nodes belong to individuals who want to share their personal phenotypic and genomic data or organizations that own genomic databanks. These nodes utilize Hyperledger Fabric network to privately store their data.
2. **Data Buyer Nodes** - Data buyer nodes typically belong to pharma and biotech companies. These nodes purchase genomic and phenotypic data from data owner nodes with a financial transaction and analyze the data on secure compute nodes. Furthermore, data buyer nodes can subsidize sequencing costs of selected data owners and send out survey questions to generate phenotypic data.

3. Human Genome Sequencing - The DNA samples will be sequenced using next-generation DNA sequencing at the DNA sequencing facilities. Next-generation sequencing of a human genome generates billions of short reads of up to about 250 letters in length. The process is quite error-prone, thus the letters are sequenced multiple times.. A typical personal genome sequencing file is about about 1 billion sequencing reads and is approximately 150 to 200 gigabytes in size.

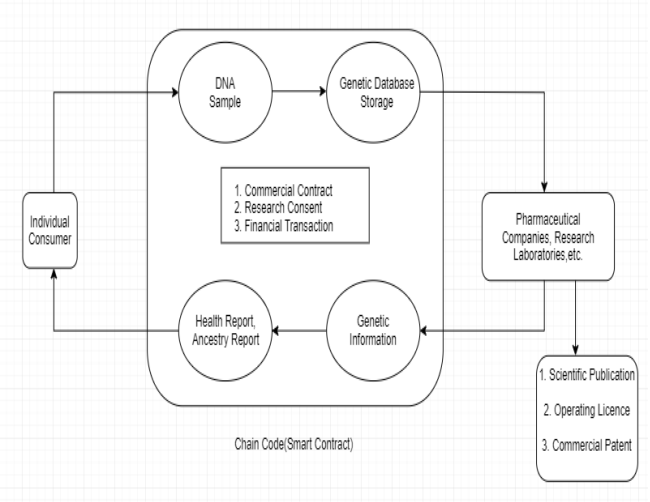

Fig. 2 Architecture Diagram for implementing project modules

4. **Phenotypic Data Generation -** It is necessary to generate phenotypic data and thus, the questions related to the same are being sent to the consumers. The survey questions are designed as such that the answers for most of the questions are interdependednt on the previous questions. This helps in elimination of fake answers given by the users. Additionally, it will be possible to evaluate the correctness of survey responses on a population level since the prevalence of common medical conditions is well known.

One special thing about Fabric is the endorsement process. Any transaction, initiated by a client application, must be endorsed. A client application generates a transaction proposal and send it for validation to endorsement peer(s) of its choice.The endorsing peers verify (1) that the transaction proposal is well formed, (2) it has not been submitted already in the past, (3) the signature is valid, and (4) that the submitter is properly authorized to perform the proposed operation on that channel (e.g. writing into the channel). The inputs as arguments to the invoked chain code's function are being taken by the endorsing peers. The chain code is processed and executed against the current ledger state to produce transaction results, including a response value, read and write sets. However, the updates to the ledger are not being made at this point of time. The set of these values, along with the endorsing peer's signature and a yes/no endorsement statement is passed back as a "proposal response" to the client application. The client application check whether the validation was successful, check the signature of endorsing peer(s) and only then a transaction goes for execution.

VI. RESULTS AND DISCUSSION

A proof-of-concept of our proposal is currently under way of prototyping. A dedicated blockchain platform on IBM Fabric Hyperledger has already been implemented, and smart contracts have been developed for sequencing data acquisition and management procedures.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Cloud off-chain files can be stored on IBM systems and the pipeline (in cloud) that uses raw DNA data for reporting to users has been completed. In addition, the first reports were created, highlighting the first somatic traits and predisposition to certain diseases and neurological conditions.

VII. CONCLUSION

The new frontier of medicine is personalized medicine, where doctors will be able to recommend the most effective medicines based on our DNA. More and more advanced technologies will make our DNA a priceless mine of information. We can now analyse our DNA in an affordable cost, and soon we will be able to discover more and more about our past, present, and future. The more DNA will be analysed, the faster the development will be. We will not only help ourselves, but we will do our part by contributing to the development of science. Blockchain is an upcoming technology which might be used to help solve some of the problems we encounter in genomics. It offers various new and exciting solutions to all these problems in the modern field of DNA testing. However, it should be noted that field of blockchain is not yet mature, and there is still room for further development, especially to ensure cryptographic security. In this perspective, we tried to speculate on how blockchain can be an integral part of solving several problems in genomics through Hyperledger Fabric.

ACKNOWLEDGMENT

We thank our colleagues from SRM Institute of Science and Technology who provided their insight and expertise that greatly assisted the research. We would also like to show our gratitude to the faculties of the Dept. Of Computer and Science and Engineering for sharing their pearls of wisdom with us during the course of this research, and we thank the project reviewers for their so-called insights.

REFERENCES

1. Henri-Corto Stoeklé, Marie-France Mamzer-Bruneel, Guillaume Vogt and Christian Hervé (2016) 23andMe: a new two-sided data-banking market model. BMC Medical Ethics. DOI 10.1186/s12910-016-0101-9
2. Vitalik Buterin (2014) Ethereum: A Next-Generation Generalized Smart Contract and Decentralized Application Platform. https://web.archive.org/web/20140111180823/http://ethereum.org/ethereum.html
3. Ian Grigg (2017) EOS- An Introduction.
4. https://eos.io/documents/EOS_An_Introduction.pdf
5. Cecile Monteill (2018) Blockchain and Health. Digital Medicine pp.41-47 https://doi.org/10.1007/978-3-319-98216-8_4
6. Halil Ibrahim Ozercan, Atalay Mert Ileri, Erman Ayday and Can Alkan (2018) Realizing the potential of blockchain technologies in genomics. http://www.genome.org/cgi/doi/10.1101/gr.207464.116. Accessed 3 August,2018.
7. IBM (2018) Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains. https://arxiv.org/pdf/1801.10228.pdf. Accessed 17 April 2018

AUTHORS PROFILE

Preethi V. Assistant Professor (O.G) Area or Subject: Network Security, cloud computing, wireless sensor network Department of Computer Science and Engineering, Ramapuram Campus, SRM Institute of Science and Technology (formerly known as SRM University)

Saurav Surve Undergraduate Student, Department of Computer Science and Engineering, Ramapuram Campus, SRM Institute of Science and Technology (formerly known as SRM University)