ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

Various Approaches for Secured and Fast transfer Data Minimization Approach of Big Data using Secure Encoding: A Survey

Dabhade Jyoti Goraksh[1], Prof. Kore Kunal Sidramappa.[2]

P.G. Student, Department of Comp Engineering, Sharadchandra Pawar College of Engineering, Otur, Pune, India

Assistant Professor, Department of Comp Engineering, Sharadchandra Pawar College of Engineering, Otur, Pune, India

ABSTRACT: In the age of Big Genomics Data, institutes such as the National Human Genome Research Institute (NHGRI), 1000-Genomes project, and the international cancer sequencing consortium are faced with the challenge of sharing large volumes of data between internationally dispersed sample collectors, data analyzers, and researchers, a process that up until now has been plagued by unreliable transfers and slow connection speeds. These occur due to the inherent throughput bottlenecks of traditional transfer technologies. One suggested solution is using the cloud as an infrastructure to solve the store and analysis challenges. However, the transfer and share of the genomics datasets between biological laboratories and from/to the cloud represents an ongoing bottleneck because of the amount of data, as well as the limitations of the network bandwidth. Therefore, transfer challenges can be solved by either increasing the bandwidth or minimizing the data size during the transfer phase. One way to increase the efficiency of data transmission is to increase the bandwidth, which might not always be possible due to resource limitations. Another way to maximize channel capacity utilization is by decreasing the bits that need to be transmitted for a given dataset. Traditionally, transmission of big genomics datasets between two geographical locations is commonly done using general-purpose protocols, such as hypertext transfer protocol (HTTP) and file transfer protocol (FTP). In this dissertation, a novel deep learning-based data minimization algorithm is presented and aims to: 1) minimize the datasets during transfer over the carrierchannels; 2) protect the data from the man-in-the-middle (MITM) and other attacks by changing the binary representation (codewords) several times for the same dataset. This innovative data minimization strategy exploits the alphabet limitation of DNA sequences and modifies the binary representation (codewords) of dataset characters by usingdeep learning-based random sampling that utilizes the convolutional neural network (CNN) and Fourier transform theory. This algorithm ensures transmission of big genomics datasets with minimal bits and latency, thereby lending to a more efficient and expedient process. To evaluate this approach, extensive actual and simulated tests on various genomics datasets were conducted. Results indicate that the proposed data minimization algorithm is up to 99-fold faster and more secure than the current use of the HTTP data-encoding scheme and 96-fold faster than FTP on tested datasets.

KEYWORDS : Data encoding, decoding, data Minimization, data transfer, CNN, RNN, Deep Learning, TCP, network protocol

I. INTRODUCTION

DNA sequencing is needed in the most critical areas such as criminal investigations, genotyping and determination of disease-relevant genes or agents causing diseases, mutation analysis, screening of single nucleotide polymorphisms (SNPs), detection of chromosome abnormalities [1], global determination of post-translational modification [2], and to identify disease- and/or drug-associated genetic variants to advance precision medicine [3] [4]. Also, the use of the nextgeneration sequencing (NGS) technologies such as whole-genome sequencing (WGS) and wholeexome sequencing (WES), are significantly decrease the sequencing costs and enable the genomic datasets to join to the big data club.

DOI: 10.15680/IJIRCCE.2019. 0706018

Currently, the major big data generators are Astronomy, YouTube, and Twitter and are expected to demonstrate continued dramatic growth in the volume of data to be acquired. For example, the Australian Square Kilometer Array Pathfinder (ASKAP) project currently acquires 7.5 terabytes/second of sample image data, a rate projected to increase 100-fold to 750 terabytes/second (~25 zettabytes per year) by 2025 [5] [6]. YouTube currently has 300 hours of video being uploaded every minute, and this could expand to 1,000 - 1,700 hours per minute (1 - 2 exabytes of video data per year) by 2025 if we extrapolate from current trends. Today, Twitter generates 500 million tweets/day, each about 3 kilobytes including metadata. While this figure is beginning to plateau, a projected logarithmic growth rate would suggest a 2.4-fold growth by 2025, to 1.2 billion tweets per day, 1.36 petabytes/year. A big data generator will appear soon that will exceed 35 petabases per year [7]. For example, the cost of sequencing genomes reduced by a factor of 1 million in less than 10 years to reach to a few hundred dollars to sequence and map genomes faster than ever before. However, growing the genomic datasets brought challenges such as storing, handling, analyzing, visualizing, sharing, and transferring the genomic information generated by NGS technologies that need to be addressed. For instance, sequencing a single whole genome generates more than 250 gigabytes of data since there are over three billion base pairs (sites) on a human genome[8]. In fact, the growth rate of DNA sequencing over the last 10 years has generated a massive amount of data to produce a double amount approximately every 7 months. According to [9] to date, there are more than 2,500 high-throughput sequencing instruments distributed over 55 countries placed in about 1,000 sequencing centers. The United States National Institutes of Health National Center for Biotechnology Information (NIH/NCBI) maintains the most archived sequencing reads. For example, NIH/NCBI currently maintains more than 3.6 petabases of sequence reads, distributed into approximately 32,000 microbial genomes, 5,000 animal and plant genomes, and 250,000 human genomes [10].

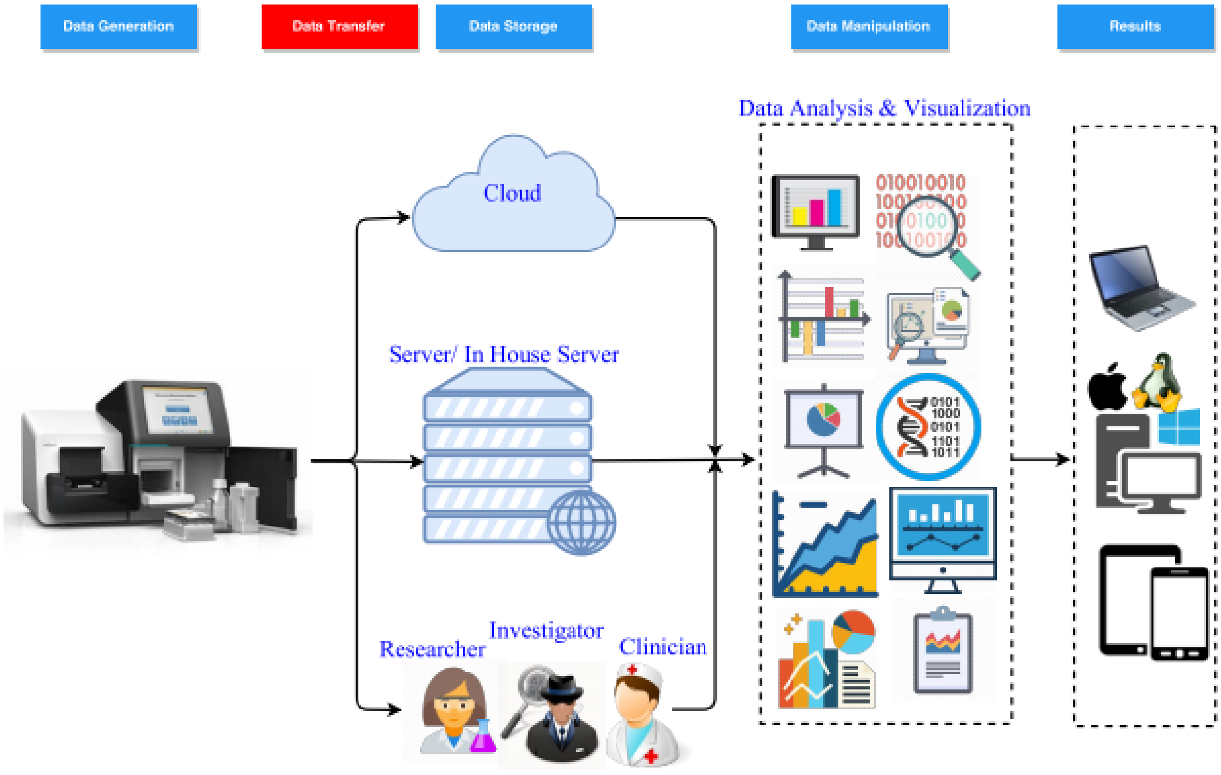

Figure 1: Genomic lifecycle

However, the current estimate of the global sequence reads is more than 35 petabases annually [11]. The tentative expectation of DNA sequencing for the next 8 years (2025) is one zettabase of annual sequencing and expanding to double per 7 months, as shown in Table 1.1 on page 10. There two more estimates of DNA sequencing that are doubling every 12 months according to Illumina's estimate [10] and every 18 months based on Moore's law. Also, biologists anticipate they need to sequence the most known species of plants and animals that are comprised of approximately 1.2 million genomes [1]. American and Chinese researchers plan to sequence about 1 million genomes in the next few years [12] [13]. For all listed information, it is necessary to prepare for the greatest challenge of big genomic datasets: data transfer. Recently, biologists ascertained that the bottleneck in the advent of the big genomics revolution is an inability to share and transfer large datasets in a timely manner. Therefore, some projects have begun to navigate the possible solutions to access big data and share them with researchers worldwide. The possible solutions are either to minimize the data volumes during the transfer over the networks or expand the network bandwidth. For example, the Human Genome Project [14] and the HapMap project

Motivation

The revolution of the new technology solved the data generating challenges and resulted in the creation of big datasets. The generated data led to other challenges such as data storage, process, and share. Many efforts have been made to solve the challenges of big data storage, and manipulation, including data analysis and visualization. However, the challenges of big data sharing still constitute a major challenge that must be addressed and resolved. Also, designing and implementing transfer protocols equipped with data minimization techniques that relied on neural network techniques and that aimed to transfer big data in shorter times with added security against attacks, did not garner the attention of many researchers. Healthcare instruments and biology laboratories became big data generators that required singular methodologies in terms of transfer time, accuracy, speed, and security. Big genomic datasets are part of the big data club that require special handling from generating and processing to transferring between two or more biology laboratories. Many solutions have been developed to address the challenges of big data generating and analysis. However, transmission challenges have not been addressed at the same level, mainly due to compatibility issues.

II. LITERATURE SURVEY

A Review of Related Network Transfer Protocols

Many protocols have been implemented to transfer different data volumes, such as HTTP and FTP, but these protocols operate in a data-oblivious manner. HTTP is a request/response protocol that resides in the first layer of the Open Systems Interconnection (OSI) conceptual model (application) and transfers data among web applications i.e. client(s)-server. HTTP communicates by sending a request from the client (requestor) to the server, followed by a response from the server to the client. Requests and responses are present in a simple ASCII format (8 bits). HTTP requests contain many elements: a method such as GET, PUT, POST and a uniform resource locater (URL). Also, HTTP requests include message headers and content-encoding, along with all data needed by the client. The server handles the request, and then responds according to the specified method. After that, the server sends a response to the client, including the status code, indicating if the request succeeded or failed and the reason why.

A Review of Data Encoding Schemes

To the best of our knowledge, this work is the first network-based data minimization solution for big genomic datasets that utilizes data-encoding as a mechanism. A file transfer protocol which uses the BitTorrent technique to transfer genomic datasets, and which was originally designed to support distributed peer-to-peer (P2P) file transfer applications. In other words, GeneTorrent distributes the same file(s) on different machines settled in different locations and configures those machines to transfer certain part(s) of that file(s) to a requester. Although higher throughput can be achieved by using multiple machines for transferring data, the underlying data are still transferred using general-purpose protocols. This protocol is no longer in use, and there is a need to create a data-aware network transfer protocol for the DNA genomic datasets that use minimum resources of the network to deliver data efficiently.

Data-Encoding Approaches

Data minimization can be divided into two main forms: data encoding and data compression. Data minimization using data encoding assigns the lowest possible bits to each alphabet's symbol using content-encoding schemes without complex computations, whereas the data minimization using data compression assigns the lowest possible bits to the entire dataset: this process involves complex computations and an extended period of time to compress and decompress operations. In general, binary representation can be divided into two categories: Fixed-Length Binary Encoding (FLBE) and Variable-Length Binary Encoding (VLBE). FLBE scheme, also called singular encoding, converts symbols into a fixed number of output bits, such as in an ASCII code which consists of an 8-bit long for each codeword

Naive Bit Encoding

This approach works by assigning fixed-length codeword/binary representation to each alphabet symbol in a way that represents more than a single symbol in a single byte, such as 2-bit length to genomic symbols.

Hybrid Encoding

This approach works by combining two or more encoding methods. For example, The Burrows-Wheeler transform (BWT) [60][55] and [61], is one of the hybrid encoding methods, especially popular in bioinformatics, used for data minimization. The BWT method works by permuting the input sequence in a way that symbols are grouped by their neighborhood. Our proposed data minimization algorithm can be classified as a hybrid encoding method by incorporating elements of the standard (8-bit) and the statistical encoding methods (variation of 1 - 3 bits.)

Statistical/Entropy Encoding

This approach works by statistics, prediction, and a probabilistic model from the input [57] and [58], such as Huffman's coding. Huffman's coding, introduced in 1952, is a statistical method that assigns a fixed-length codeword/binary representation to alphabet symbols, such as 2-bit, 3-bit, 8-bit, etc. The codewords will have different lengths, and the lowest frequency symbols will be assigned with the longest codewords and vice versa. This research utilizes this type of encoding with CNN deep learning algorithm to ensure the assignment of the lowest possible codewords to the more frequent dataset characters, and to undertake this process several times during the data transfer phase.

Dictionary-based/Substitutional Encoding

This approach stores different patterns of the input symbols in a dictionary or a database, along with their codewords, and then replaces the new input parts with predefined portions algorithm works by replacing the repeated occurrences of symbols with their references that indicate length and location of that string, which occurred before, and which can be presented in the tuple (offset, length, symbol).

III. PROBLEM ANALYSIS

Although low-cost, high-throughput instruments and cloud-based services have solved big data generating and processing challenges to certain extents, they do not solve the data transfer speed and security problems very efficiently when it comes to big genomic datasets. That is, transferring big data between two or more places (e.g. between two biology laboratories or between lab-cloud-lab) still results in a bottleneck due to the use of traditional transfer protocols such as HTTP [19] and FTP [20]. Recently, biologists ascertained that the bottleneck in the advent of the big genomics revolution results in an inability to share and transfer large datasets in a timely manner.

We designed and implemented a novel data minimization algorithm to transfer big genomic datasets in a more expedient and secure way and to allow scientists to easily share their data and analyses. We used the HTTP as a baseline protocol to compare and assess our implementation results of transferring big genomic datasets. The goals of our new data minimization algorithm are as follows: 1) reduce the size of data that need to be transferred between a server and a client; 2) secure and protect the privacy of the data from unauthorized access due to attacks or data breach, such as MITM attack. Our heuristic model, simulation, and implementation results proved that our data minimization algorithm reduces significant amounts of data and makes more efficient use of network bandwidth, while also protecting the data by preventing unauthorized individuals from accessing them, should a breach occur.

IV. CONCLUSION

Our proposed data minimization mechanism for the transfer protocols enables them to be smart protocols via using standard codewords for dataset headers, while using our data minimization mechanism for the dataset body. We test our data minimization algorithm by using three different transfer protocols: HTTP, FTP and BitTorrent and by considering such variables as versatility, security and flexibility. These are commonly used protocols that transfer different data types in a variety of browsers, such as Google Chrome. Also, these protocols have certain security features in the transport layer because they run on top of TCP. These protocols are flexible because they are equipped with the ability to modify one or more components, such as content-encoding schemes, compression algorithms, and message headers. This dissertation is an extended version of our works existing system. We extend this previous work by employing the convolutional neural network to update the encoding codewords periodically and to ensure the assignment of the minimum binary representation to the most repetitive characters in the file

V. FUTURE WORK

To implement a system with various data encoding approaches on large scale data and evaluate the comparative analysis on different network environments.

REFERENCES

[1] C. Mora, D. P. Tittensor, S. Adl, A. G. Simpson, and B. Worm, “How many species are there on earth and in the ocean?” PLoS biology, vol. 9, no. 8, e1001127, 2011.
[2] J.-Y. Li, J. Wang, and R. S. Zeigler, “The 3,000 rice genomes project: New opportunities and challenges for future rice research,” GigaScience, vol. 3, no. 1, p. 8, 2014.
[3] F. S. Collins and H. Varmus, “A new initiative on precision medicine,” New England Journal of Medicine, vol. 372, no. 9, pp. 793–795, 2015.
[4] T. C. Carter and M. M. He, “Challenges of identifying clinically actionable genetic variants for precision medicine,” Journal of healthcare engineering, vol. 2016, 2016.
[5] N. Drake et al., “Cloud computing beckons scientists.,” Nature, vol. 509, no. 7502, pp. 543– 544, 2014.
[6] R. Spencer, “The square kilometre array: The ultimate challenge for processing big data,” in Data Analytics 2013: Deriving Intelligence and Value from Big Data, IET Seminar on, IET, 2013, pp. 1–26.
[7] M. C. Schatz, “The next 20 years of genome research,” Simons Center for Quantitative Biology, 2015.
[8] M. Aledhari, M. Di Pierro, T. Barkman, M. Hefeida, and F. Saeed, “Deep learning-based data minimization algorithm to fast and securely transfer of big genomic datasets,” IEEE TRANSACTIONS ON BIG DATA, 2017.
[9] J. Hadfield and N. Loman, Next generation genomics: World map of high-throughput sequencers, 2014.
[10] A. Regalado, “Emtech: Illumina says 228,000 human genomes will be sequenced this year,” Technology Review, vol. 24, 2014.
[11] S. C. Baker, “Next-generation sequencing challenges,” 2017. [12] A. Manzoor, “Emerging role of big data in public sector,” Managing Big Data Integration in the Public Sector, pp. 268–88, 2015.
[13] J. Zhu, “A year of great leaps in genome research,” Genome medicine, vol. 4, no. 1, p. 4, 2012.
[14] N. I. of Health et al., “An overview of the human genome project,” 2005.
[15] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch’ang, W. Huang, B. Liu, Y. Shen, et al., “The international hapmap project,” 2003.
[16] W. S. Bush and J. H. Moore, “Genome-wide association studies,” PLoS computational biology, vol. 8, no. 12, e1002822, 2012.
[17] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, et al., “The ncbi dbgap database of genotypes and phenotypes,” Nature genetics, vol. 39, no. 10, pp. 1181–1186, 2007.
[18] J. Kaye, C. Heeney, N. Hawkins, J. De Vries, and P. Boddington, “Data sharing in genomics– re-shaping scientific practice,” Nature reviews. Genetics, vol. 10, no. 5, p. 331, 2009.
[19] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, “Hypertext transfer protocol–http/1.1,” Internet Engineering Task Force (IETF), Tech. Rep., 1999.
[20] S. Deorowicz and S. Grabowski, “Compression of dna sequence reads in fastq format,” Bioinformatics, vol. 27, no. 6, pp. 860–862, 2011.