Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 3 – 9

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Survey on Molecular Cryptographic Network DNA (MCND) Using Big Data

R.Kannadasan[1], M.S.Saleembasha[2], I.ArnoldEmerson[3]

[1,3]*School of Computing Science and Engineering-School Bio Sciences and Technology, VIT University, Vellore, India.*

[1]Kannadasan.r@vit.ac.in

[3]iarnoldemerson@vit.ac.in

[2]*Department of Computer Science, Mazoon University College, Sultanate of Oman.*

[3]m.s.saleembasha@gmail.com

Abstract

The term Big Data commonly refers to enormous bulk of data that is bring out by human (genomes) from highly used digital devices like cameras, internet, mobile phones, sensors etc. One can predict many possible outcomes such as human nature, his interest etc. By developing advanced analytics on the top of big data. Methods like encryption and decryption helps to address many issues related to Molecular DNA big data before making the use of it. Analysing and finalizing better technical solution for specific application is all time astonishing issue from the developers and analysts perspective. Xampp server tries to give the base for big data storage by the help of MySQL, PHP and Perl. Along with mentioned technologies there are many more data storage and processing technologies evolved related to Hadoop. This paper contains the survey on recent Molecular DNA big data technologies which gives the performance characteristic of each technique.

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Keywords: Big Data, DNA (MCND), Encryption, Decryption.

1. Introduction

By the time data is growing tremendously, so there is need of more and more advance way of storing those data. One evolving idea to store this data is making the use of DNA. Let by George at Harvard University's *Wyss* institute[1], a team of researchers have recently figured out a way of a whopping 5.5 petabytes (= 700 terabytes) within a single gram of DNA. Rather than enciphering binary data on magnetic drives, scientists are leveraging strands of DNA to microcode data .DNA, capable of storing 96 bits, is synthesized with each of the ATGC bases representing a binary value fig.1(a): (T&C representing 0 and A&G representing 1).

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.003

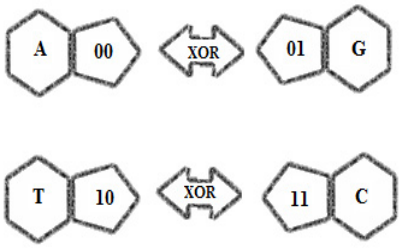

Fig .1(a). DNA double stranded molecule

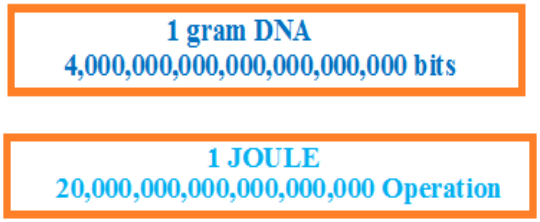

Fig.1 (b). DNA measurements

It is found that DNA is very compact to store one bit base with each base only a few atoms large. DNA is also volumetrically meaningful to store in a beaker or other incarnation rather than a Hard disk fig.1 (b). Finally while some advanced storage system need to be kept in sub-zero vacuums "DNA can be stored in a box in your house". Poignantly, big data can also help to advance our understanding of DNA with the ability to manage incredibly large data sets, were able to better understand our own genetic makeup. Big data research projects such as charge of R&D seem to be one of the most gripping use causes for research. It might still be some way off, but with the help of companies like DNA relationship.

2. Related Works

2.1 DNA based cryptography

Many of the research work has been carried out on DNA computing in which they made the use of test tubes some simulating changes of DNA using computers Gehani ET. Substitution method which make the use of distinct one time pads libraries introduced by the al[2] at its first trial of DNA based cryptography, which mainly states a specific, auto generated, combination mapping whereas an XOR scheme utilizes molecular calculations and ranking, random key strings are used for encryption.[2] DNA sequence combination operation which is based on an image encryption algorithm is presented by[3] . In this original image is encoded in order to get DNA sequence matrix and later it is classified into equal blocks and two logistic maps, in which addition has been carried out by DNA complementarity and DNA sequence addition operation. Encrypted image can be obtain by DNA sequence matrix. Leier et al in his paper mainly focused on two different cryptographic approaches based on DNA binary strands with the idea that a potential interceptor cannot distinguish between dummies and message strand[4].where information can be hide in DNA binary strands in first approach and the another approach used to designed molecular checksum. PCR (Polymers Chain Reaction) and subsequent gel electrophoresis helps in the process of Decryption.

Sheriff ET. Al[5]. In his paper proposed The YAEA DNA encryption algorithm which make the use of search technique which helps in locating and returning the location of quadruple DNA nucleotide sequence which play important role on representing binary octets of plain text characters. Where input to algorithm will be Plain text character and a random binary file which produces the PTR as a pointer to the location of the found quadruple DNA. In which they made the use of images for the encryption process in order to show randomness of the selection of DNA octet's locations[6]. Pseudo encryption methodology has been explained by the Kang[7]. Which converts plain text to DNA sequence later converted into the spliced and protein form of data by cutting them with specified pattern and again converted into mRNA form and which later transform to protein form of data[8] .fig 2.

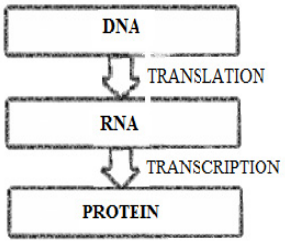

Fig. 2. Seclusion of RNA, DNA and proteins

Secure channel is used to send the protein form of data. This method uses the mechanism of DNA function not DNA sequences therefore, it is known as pseudo DNA cryptography method[7] because it simulates the transcription, splicing, and translation process of the central dogma. Authors Borda and Tornea in their paper invented a secret writing method in which they used DNA with the concept of one time pad[9] XOR OTP tiles and chromosome indexing the message is encrypted. As digital information continues to accumulate, higher density and some longer term storage solutions. DNA has many benefits as a medium for immutable like, DNA storage is very dense. At theoretical maximum, DNA can encode two bits per nucleotide (nt) or 455 Exabyte's per gram of DNA[10], where most digital storage medias lags.

2.2 Digital Information Storage in DNA

As day by day digital Medias are increasing exponentially in order to handle such huge data higher density and longer term storage solutions are necessary[11].The advantage of DNA storage is that it is not restricted to any kind of limit, and is often readable despite degradation in non-ideal conditions over millennia (*4*, *5*). Finally, DNA's essential biological role helps to access natural reading and writing enzymes and grantees that DNA there will not be more changes in its original form which keeps it in readable form which will be useful for the foreseeable future. The concept of DNA storage was initially experimented in 1988 and the largest project to date encoded 7920 bits[10]with very small work which was done previously they succeeded in forming writable and readable long well-formed DNA sequences, and which has few amount of broader applications, we developed a mechanism which helpful in encoding the digital data by the help of novel encoding scheme which uses the very advanced DNA synthesis and sequencing technologies. As a part of experimented we transformed an html coded e-book which consist 53,426 words, 11 JPG images and 1 JavaScript pro-gram into a 5.27 megabit bit stream. Later we translated it on 54,898 159nt oligonucleotides (oligos) each encoding a 96-bit data block (96nt), where 19-bit address denotes the exact position of the data block in the bit stream (19nt), and flanking 22nt similar sequences for amplification and sequencing. Highly synthesized ink-jet printer and high-fidelity DNA microchips are used by oligo library.in order to make encoded book readable the library was amplified by limited-cycle PCR and then sequenced on a single lane of an Illumina[12] We combined neighbor paired-end which were ending at each other 100nt reads in order to decrease the effect of sequencing error. At the end by making the use of reads which produced 115-nt length as well as perfect barcode sequence, we produced consensus at each base of each data block at an average of ~3000-fold coverage. All data blocks were recovered with a total of 10 bit errors out of 5.27 million, which were predominantly located within homo-polymer runs at the end of the oligo where we only had single sequence coverage. Proposed mechanism has approximately 5 benefits over past DNA storage approaches. Proposed mechanism encodes one bit per base (A or C for zero, G or T for one), instead of two. This allows us to achieve messages many ways which helps to avoid sequences which are not in proper readable and writable format like extreme GC content, repeats, or secondary structure. By splitting the bit stream into addressed data blocks, we succeed in eliminating the need for long DNA constructs which were difficult to combine at this scale. We synthesized, stored, and ranked many copies of of individual oligo[13] in order to clone and ranking verification constructs. Since other copies of synthesis and sequencing contains rare errors, vitro approach has been used in order to avoid cloning and stability issues of in vivo approaches[14] At last we succeeded in implementing the advanced mechanism in DNA synthesis as well as sequencing which allows us to encode and decode of huge amount of data for ~100,000-fold less cost which was much lesser than first

generation encodings. Archival storage is the best application which suitable for DNA. Density, stability, and energy efficiency are the main benefits of DNA storage, while costs and times for writing and reading are currently impractical for all but century-scale archives comparison to other measured by the log10 of bits encoded in the report or commercial technologies. We plotted information density (log n=10 of bits/mm3) versus current scalability as unit fig 3.

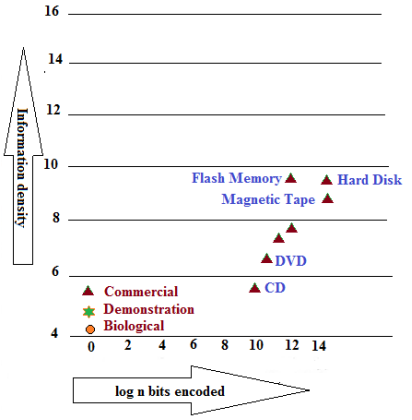

Fig. 3. Information density in DNA

However, per year, the cost of DNA synthesis and sequencing have been decreasing exponentially of 5- and 12-fold, which much faster than digital devices at 1.6-fold per year. Vastly simplify reading DNA-encoded information is available in the form of Hand-held, single-molecule DNA sequences. The proposed approach of using addressed data blocks mixed with library synthesis and consensus sequencing should be suitable with future DNA sequencing and synthesis technologies. Reciprocally, large-scale use of DNA such as for information storage could accelerate development of synthesis and sequencing technologies. Future work could use compression, redundant encodings, parity checks, and error correction to improve density, error rate, and safety. Other polymers or DNA modifications can also be considered to maximize reading, writing, and storage capabilities.

3. DNA Big Data Storing Technology

3.1. Big data Next generation

As data continues to amass current method of storing it will soon be mothballed. News from Harvard that bacteria in the average human stomach can store more data than the most advanced computer hard drives available today should be a harbinger of things to come[15].

Fig.4. DNA the Next Big Data Storage Technology

The supposition that digital data could be warehoused in DNA is no longer; it is fact. The vaunting task of storing vast amount of data is suddenly a foregone conclusion. The content of the entire internet will soon be stored in 75grams, about 75 paper clips worth, of DNA material; the cardinal rules of data management are being rewritten.

4. The Characteristic of Big Data

The main sources of big data are data information within the organization, sensory information in the Internet of things and interactive information in the Internet world. It is difficult to form a unified concept of big data. Big data evolving new shape to the data considering its main properties which are volume, variety and velocity.

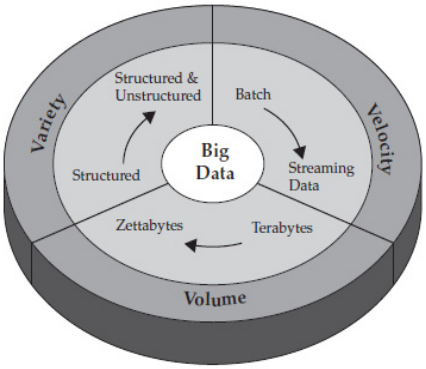

Fig.5.Characteristic of Big Data

The Big Data mainly classified in three dimensions: Volume, Velocity and Variety.

- *Volume* – The size of data is very large and in Terabytes and petabytes and continuously increasing.

- *Velocity* – It should be used when streaming in to the enterprise in order to maximize its value to the business. The role of time is very critical here.
- *Variety* – It extends beyond the structured data, including unstructured data of all varieties: text, audio, video, posts, and log files.

5. Proposed Method

In the proposed method for large number of storing the text and image we are using the technology as big data. If user can send image or text to receiver that data will be converted into DNA sequence, using encryption for converting original text or image to DNA (A, T, G, and C) sequence. If any third party or unauthorized person can access the database it will show only the DNA (A, T, G, and C) sequence not an original data. If receiver wants to view the original data with the help of secret key, that key will be created by user and send a copy key along with converted data to the receiver. All the converted data's are stored in the databases. Suppose if you need to get particular records, you can retrieve from the database. So that you can get all the information in the database.

6. Encryption and Decryption using Big Data

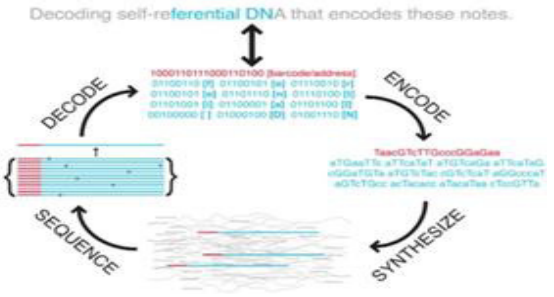

Fig. 6. Encryption and Decryption using big data

Encryption: The process of converting plain text into an unintelligible format (cipher text) is called Encryption.

Decryption: Process of converting cipher text into a plain text is called Decryption.

7. DNA Encryption Decryption Using PHP

Big data is used to large number of storing data; here we have used DNA to encrypt and decrypt the data by using Xampp tool below fig.7.1.1&2 shows that user has to login with username and password to enter.

WELCOME:
E-Mail:kannadasan.r@vit.ac.in
Password:******************
SIGN IN

Fig.7 (a) O/P of DNA Encryption Decryption Using PHP

WELCOME:
Enter the String: HELLO
Encrypt String : cgcggctgctgcgga
Decrypt String : HELLO

Fig.7 (b) O/P of DNA Encryption Decryption Using PHP

The above show that, after user login the user has to enter the message, so that it will encrypt the message in the form of A, T, G, C, and it will store in to the database.

7.2 DNA Table for Encoding Method

Fig.8 (a) Database admin

a-cga	l-tgc	w-ccg	3-gac
b-cca	m-tcc	x-cta	4-gag
c-gtt	n-tct	y-aaa	5-aga
d-ttg	o-gga	z-ctt	6-tta
e-ggc	p-gtg	_-ata	7-aca
f-ggt	q-aac	,-tcg	8-agg
g-ttt	r-tca	.-gat	9-gcg
h-cgc	s-acg	:-get	Space-ccc
i-atg	t-ttc	0-act	
j-agt	u-ctg	1-acc	
k-aag	v-cct	2-tag	

Fig.8 (b) DNA Table for Encoding Method

8. Conclusion

Big data has numerous number of technologies for data storage as well as processing big data, which faces many challenges to meet user's expectations when we consider storage, processing and performance. There are full-time implemented solutions for many of the file storage and data processing techniques. This paper mainly focus on three file systems along with four data processing languages which is widely using in Hadoop ecosystem and gives further study outlook.

Reference

1. Birney E, Stamatoyannopoulos JA, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. 2007;447(7146):799-816.
2. Gehani A, LaBean T, Reif J. DNA-based cryptography. In: *Aspects of Molecular Computing*. Springer; 2004:167-188.
3. Zhang Q, Guo L, Xue X, Wei X. An image encryption algorithm based on DNA sequence addition operation. In: *Bio-Inspired Computing, 2009. BIC-TA'09. Fourth International Conference on*.; 2009:1-5.
4. Shiu HJ, Ng K-L, Fang J-F, Lee RCT, Huang C-H. Data hiding methods based upon DNA sequences. *Inf Sci (Ny)*. 2010;180(11):2196-2208.
5. Jacob G, Murugan A. DNA based Cryptography: An Overview and Analysis. *Int J Emerg Sci*. 2013;3(1):27-36.
6. Jacob G, Murugan A. A Hybrid Encryption Scheme using DNA Technology. 2013.
7. Ning K. A pseudo DNA cryptography method. *arXiv Prepr arXiv09032693*. 2009.
8. Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*. 1997;18(3-4):533-537.
9. Borda M, Tornea O. DNA secret writing Techniques. In: *IEEE Conferences*.; 2010.
10. Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. *Science (80-)*. 2012;337(6102):1628.
11. Conway P. *Preservation in the Digital World*. Commission on Preservation and Access Washington, DC; 1996.
12. Caporaso JG, Lauber CL, Walters WA, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J*. 2012;6(8):1621-1624.
13. Ellis T, Adie T, Baldwin GS. DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr Biol*. 2011;3(2):109-118.
14. Saxena P, Singh A, Lalwani S. Use of DNA for Computation, Storage and Cryptography of Information.
15. Baharvand H, Aghdami N. *Advances in Stem Cell Research*. Springer; 2012