International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 87-92
e-ISSN:2278-621X

BIG DATA IN BIOINFORMATICS- A REVIEW ON PROBLEMS AND SOLUTIONS

Jithin Mathew[1], Rohan D'souza[2] *and* Hemalatha N[3]

Abstract- One of the greatest challenges in Bioinformatics is the big data. The data is piling up from heterogeneoussources every minute, from all over the world. Next generation sequencing(NGS), a high throughput sequencing method is the major reason for this voluminous data generation is growing at an exponential phase.This data is of high importance because,speaking from the point of a protein function to an evolutionary relationship, the answer to every question lies in the genetic code of that organism, where understanding this code can solve any errors or even understand the reason for existence of life on earth. Bioinformatics is the tool to this study and big data is a huge vault of treasure where the person with right key can unlock this information. Everywhere and everyone are trying to tackle with the problem of Big Data using their own methods and solutions, either it is by creating a new and sophisticatedalgorithm or creating a new way of computing by developing the hardware that can coup up with the data load piling up daily. Some approaches even look at this in an entirely different way like DNA data storage or a development of quantum computing which could change the face of data storage and data retrieval of Big Data for ever , but its far away from reality for the time being. But the rate at which we are closing by is also promising to the solutions to Big Data. This paper aims at providing the insight of present problems of big data in Bioinformatics and the solutions to this. This review will help an individual in understanding the problems and the solutions to the big data in Bioinformatics.

Keywords- Big data, DNA data storage, Bioinformatics, Cloud computing, Parallel computing.

I.INTRODUCTION

In today's world Big Data is a very important term. Big Data typically includes masses of unstructured data that need more real-time analysis[1]. Now this data can be of any type, form and size. Today data is generated almost by everything,and it is accumulated and stored in such a way as it to be easily retrieved during a query, but this happens at a smaller level. When it comes to big data, if the data is stored in an unstructured manner the retrieval of the necessary file from the massive database will be difficult than finding a needle in the haystack, which is why a proper tool is needed to find exactly what we are looking for. These is one of the many problems that Big Data is facing today.

To those who are not very familiar with the term 'Big Data' here is the easiest and simplest way of understanding it. Let us take the example of a local hard disk or the storage system of a smart phone where one day we run short of storage capacity, and for time being we may manage to upgrade the storage unit, but eventually it reaches the end of up gradation capacity, now with the further addition of data, we will notice that system gets slower and less efficient which in fact consumes more energy, time and money. And this is exactly what is happening in the case of big data except for the fact that it is in a very large scale. With the advancement in the Information technology, the growth of such massive unstructured data in a database is exponential. Thus the voluminous increase in data, has in turn placed a high demand in the development of compatible hardware and software for storage, processing and retrieval of information from huge databases where the data can be redundant or unstructured. This has ledto a transformation of the world we live in today, In January

[1] *Aloysius Institute of Management and Information technology, Mangalore,Karnataka , India.*
[2] *Aloysius Institute of Management and Information technology, Mangalore,Karnataka , India.*
[3] *Aloysius Institute of Management and Information technology, Mangalore,Karnataka , India.*

2007, Jim Gray, a pioneer of database software, called such transformation “The Fourth Paradigm”[2]. Now this has evoked the interest of giant databases across the world to come up with every possibility to overcome this problems,and as we go on we are going to see different possibilities that researchers, data analysts and data scientists, have come up to deal with the problem of big data.

A. Big Data in Bioinformatics

The day man realised that the solutions to fundamental problems of biological organism is buried deep inside the molecular level of the body itself , a quest to find out the genetic information and tools to manipulate it became highly valuable. With the advancement of biological knowledge and its tools, sequencing of genome entered into a new level, “Next Generation Sequencing”. This led to the new era of ‘Omics’.when the sequencing cost came down to an affordable price which is depicted in figure 1, more volume of omic data of thousands of organisms began do decode in biological research laboratories. Next Generation Sequencing (NGS) platform that use semiconductors or nanotechnology have exponentially increased the rate of biological data generation in the last few years[3]. Today we have many number of databases for Bioinformatics which store and retrieve variety ofbiological data using different storage and retrieval methods. NCBI, EMBL,DDBJ are the most prominent biological databse in the world. To store and analyse these data we requires massive amounts of computational power of databases, data formats, software packages and pipelines[3].And the information derived out of big data using all these so called powerful tools must be valid and true,because no researcher wants to end up with a wrong conclusion, this puts a pressure in developing tools that are accurate and efficient as well. The retrieval of information from the database can be divided into three stages and all the stages has it’s own problems due to traditional database management methods[4].

Big data is not entirely a problem but a new era of raw information in science. The International Data Corporation defines big data in terms of four ‘V’s , which include volume, velocity, variety and value[1]. Now this is very important definition in terms of ‘Biological’ big data because the information is of higher value in case of biology with comparison to other areas where big data is generated. There is a large volume of high standard information buried inside the huge unstructured big data, this pose a great challenge to the Bioinformaticians to selectively retrieve and deliver the right information in the shortest of the time possible when each query is made.

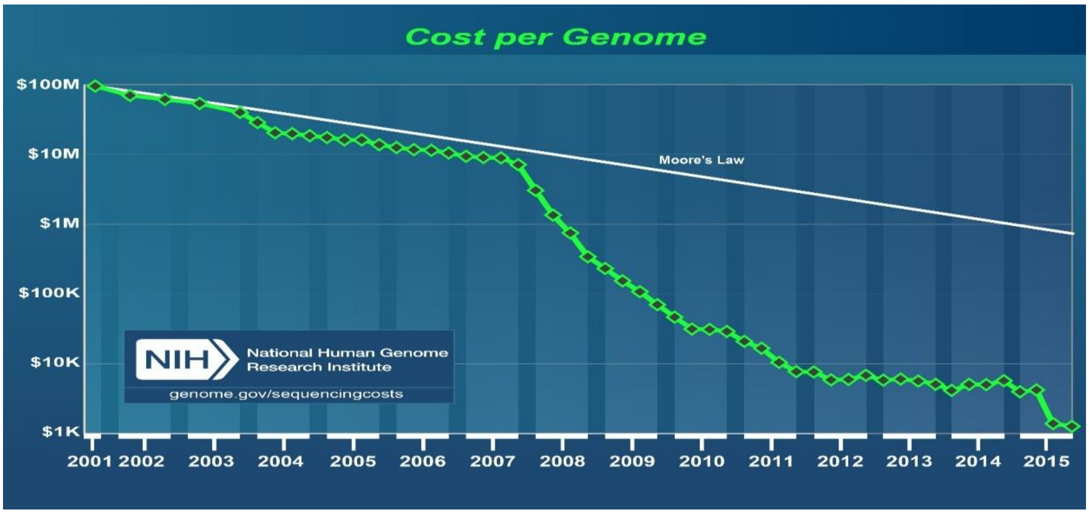

Figure 1.

II. PROBLEMS ASSOCIATED WITH BIG DATA IN BIOINFRMATICS

With the increase in the volume of data, there is a steep rise in the need for higher computational speed, better software packages and pipelines to process, retrieve or submit the genomic and Proteomics data[3](which is referred to as velocity). One of the major challenge that Bioinformatics face today is that "Biological data is generated at a faster pace that the Moore's law of computational power does not meet the requirements of molecular biology[5]. The tools that are being used currently is not very efficient in dealing the complexity of every biological aspects. Most of the tools are powerful but they still possess some flaws that is not encouraged by biologists.

A. Storage

The storage and maintenance of Petabytes of data in a data centre consumes energy, space and money which in addition require proper cooling facility with labourers[6].The storage of such large volume of data might be possible at the current scenario by governmental organisations like NCBI and large private sectors like, google, amazon etc, but data scientists currently predict by analysing the statistical records that soon the exponential growth of data will move on from Petabyte to Exabyte.This is mainly because the generation of data arises from multiple sources. The storage of data cannot be avoided because of its high value and importance in this data driven world, the only remedy is to contain it in a useful manner.When the data is collected from multiple sources and this data being heterogeneous, it is very important to filter and select the useful data to avoid redundancy. When several researchers from all over the world might sequence the genome of a particular organism and submit their data to a database, if the difference in the data is small or negligible, where if the difference do not hold any biological significance then the data can cause redundancy if stored.Database design is an art today, and is carefully executed in the enterprise context by highly-paid professionals[7]. There are quite number of hypothetical theories and successful experiments conducted by data scientists to overcome the storage problem of big data in Bioinformatics. The solutions to this will be explained in section three.

B. Data transfer

A singlehuman DNA comprises around 3 billion base pairs (bp) representing approximately 100 Gigabytes (GB) of data[8].The transfer of such huge data from a database to local sever or a personal computer is time consuming and inefficient process. When a study is conducted which involves the comparison of quite large number of organisms and its omic data, the transfer and analysis of each set of data is a difficult and time consuming process. In addition to that, the individual or the research institute should possess advanced and updated tools to analyse and handle the raw data especially when the data is very huge, this is again a time consuming and not very economical method to every biologist.

C. Security

With the vast amount of information flowing in and out of the databases, it is an extremely challenging task to maintain the confidentiality and security of the entire system. Without a proper encryption of the data that is being transferred,it pose a great risk of loss or damage of the data due to various reasons. Even in a cloud computing system the privacy and security of data depends on the quality of service provider depending on their terms and conditions. Personal information in the field of Bioinformatics (Especially in the field of personalised medicine) is very sensitive and must be archived in secure and safe Database.

D. Computational power--

Computational power is a need that grows parallel to the data volume in Big Data. However promising, parallel computing still requires new paradigms in order to harness the additional processing power for Bioinformatics[9].Even though there are various sophisticated algorithms developed to manage the storage and processing of biological big data, it still do not meet the requirement which the big data in omics place in front of the computer science. As long as sequencing takes place and more comparative and analytical research are carried out in the field of biology, a need for the new and updated tools to compute the statistical data in biology is never ending.

III. SOLUTIONS TO BIG DATA ASSOCIATED PROBLEMS IN BIOINFORMATICS.

There are so many different approaches the Bioinformaticians have put forward and still working on to overcome the problems of big data. To solve this problem it is necessary to look at the possibilities in a broader way by a better understanding of biology and computer science.

A. *Cloud computing*

Since we are facing the problems of data transfer and storage in personal storage system, Cloud computing is gaining a wider acceptance these days. Cloud computing have a number of advantages due to its scalability, extensibility and provision capability[10].

Cloud computing works in a way where individual can transfer large volume of data from databases like NCBI and create an environment within the cloud application interface or software frameworks, thereby the user can analyse the raw data and derive the information out of it without downloading and computing the raw data in a personal computer with necessary analytical tool[8].Cloud computing provides a promising solution to the problems of data transfer because of its higher data transfer speed and computational power in comparison to peer to peer data transfer or server to client data transfer (Figure 2).

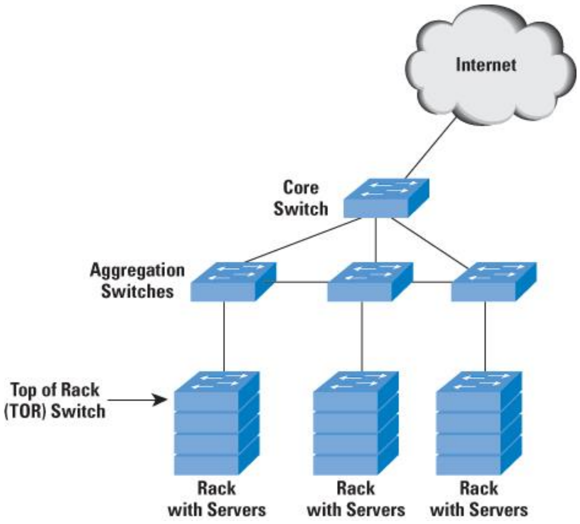

Figure 2. Architecture of Cloud computing

B. *Storage of data in DNA*

With the amount of biological data that is being produced every year, the datacentres are running out of finances to maintain big data. The present technological advance in data storage is not adequate to meet the challenges that big data storage will face in the future. Therefore it is necessary to think out of the box to attain a better solution to contain this big data. In the past decade, there has been a great development in the field of molecular biology which at its current rate could result in a molecular revolution. This is clearly evident by the positive results that are generated out of several researches carried out all over the world. The entire process of Data storage and retrieval in DNA is depicted in figure 3. Digital data storage in DNA is progressing at a rate of 10 times every year in comparison to Moore’s law on computational power and electronics[11]. 1 gram of dry DNA has a storage potential of 455 Exabytes of information, but at the current situation the cost of storage and retrieval of one Megabyte(MB) of data in DNA is 12,500 USD and 220 USD respectively. But the high durability of DNA molecule (for several million years) in addition to their confidentiality and security in archival storage of data is driving the interest of researchers to focus their attention to this field (Figure 3).

With the current rate of development in the field of Bioinformatics, the possibility of storage of digital data in Biological molecules such as DNA is inevitable with respect to time[11].

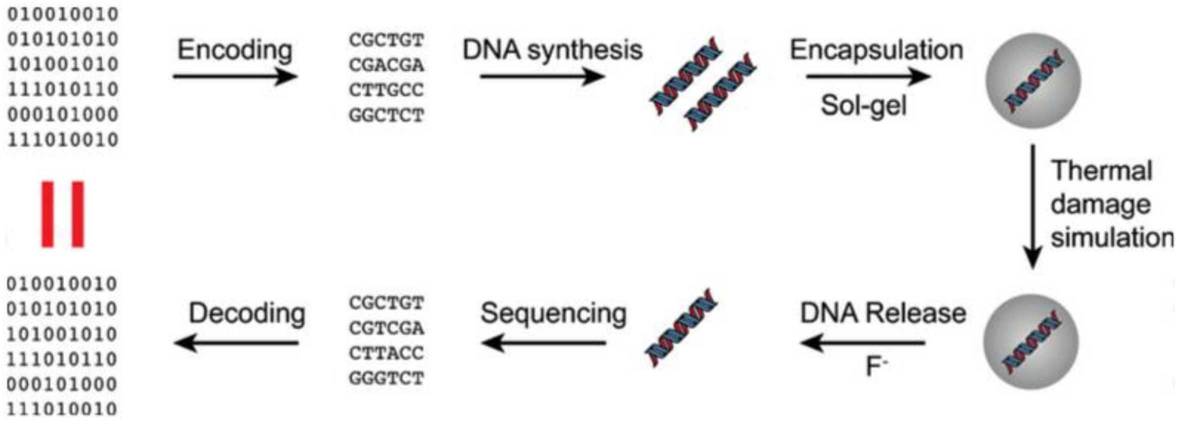

Figure 3. Data flow in storage of data in DNA

C. *Parallel computing*

A parallel computer uses a set of processors that are able to cooperate in solving computational problems[9]. Parallel computing is an example for the use of high computational power in data analysis. Here, a large problem is broken down into small tasks and distributed to multiple processors, then each bit of solved data is collected and submitted as a single solution. Two or more microprocessors can be used simultaneously, in parallel processing, to divide and conquer tasks that would overwhelm a single, sequential processor.Database searching (DBsrch) is the most heavily used Bioinformatic application[9]. From Biological data point of view, the data is highly diverse and consist of different set of information encoded as labels,some of the basic queries are to find the structural, functional and evolutionary analogies between the sequences. Therefore to process these queries there is a need for dividing the task into multiple processes where each process will analyse and compute a set of data, put it together and give out the result in the shortest possible time.

D. *Hadoop open source framework*

Hadoop, a software framework, is said to be the most efficient tool in parallel commode computing for handling biological data[12]. There are several tools already existing in Hadoop that deal with Bioinformatics problems. Parallel computing using software frameworks like Hadoop can get the result out with less cost in comparison with the use of supercomputers. Hadoop was developed by Google's MapReduce that is a software framework where an application breaks down into various parts.

The HDFS (Hadoop Distributed File System) architecture of Hadoop allows the data storage without the loss or damage of data. The HDFS system stores the data in multiple servers where the whole information is broken into pieces for storage. Unlike other Database systems, HADOOP HDFS architecture protects the information even if one memory location is damaged. This is very important because biological data holds high value and high priority must be given for its existence for a long duration.

E. *Grid computing*

Unlike parallel computing, Grid computing holds an advantage of being cost effective method of analysing large problems. It has the overwhelming potential to apply supercomputing power to address a vast range of Bioinformatics problems[13]. Grid computing is a technology that enables the division of a large problems (like the genome analysis) into smaller tasks and distributing them over a vast network of computing environment over different geographical locations (Figure 4). This is made possible by sophisticated algorithms that divide the Bioinformatics problems efficiently into executable smaller processes which is computed over different computing environment and finally collected and organised accurately. Grid computing appears to be a promising trend for its ability to make more cost effective use of a given amount of computer resources[14].This can be justified as one of the major reason for the development of grid computing.

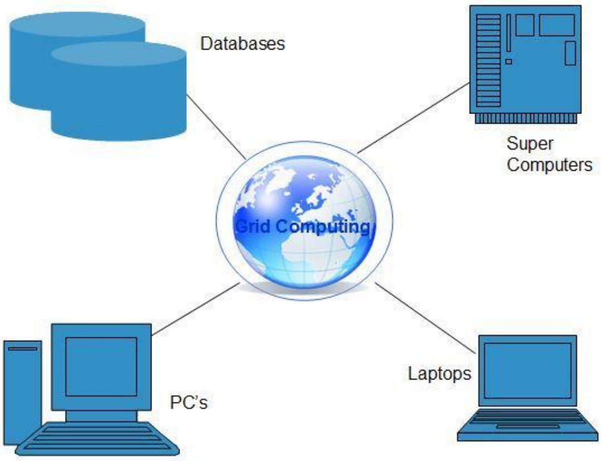

Figure 4. Grid Computing frame work

IV. CONCLUSION

The rate of genomic sequence is almost doubling every year and volume of biological data is definitely growing parallel to the sequencing rate. When the need for a long lasting solution is very high, several projects are initiated at 'big' level to deal with big data. The solutions which currently exist are going to be incompatible in dealing the big data any day. In spite of all this hurdles, the rate at which the computer science is innovating to build a long lasting solution is promising. From past few years there are researches carried out to effectively utilise the value out of Bioinformatics big data.New algorithms are created every day to meet the advancement in biology. Big data is not considered as a problem anymore, but a high potential solution to the existing problems.

REFERENCES

[1] M. Chen, S. Mao, and Y. Liu, “Big data: a survey,” Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.

[2] T. Hey, S. Tansley, K. M. Tolle et al., “The fourth paradigm: dataintensive scientific discovery. Microsoft research Redmond”, WA, vol. 1, 2009.

[3] P. Nemade and H. Kharche,“Big data in bioinformatics & the era of cloud computing,” IOSR-JCE, vol. 14, pp. 53–56, 2013.

[4] S. Ceri, A. Kaitoua, M. Masseroli, P. Pinoli, and F. Venco, “Data management for next generation genomic computing,” 2016.

[5] F. F. Costa, “Big data in genomics: challenges and solutions,” GIT Lab J, vol. 11, pp. 1–4, 2012.

[6] M. C. Schatz, B. Langmead, and S. L. Salzberg, “Cloud computing and the dna data race,” Nature biotechnology, vol. 28, no. 7, pp. 691, 2010.

[7] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han et al., “Challenges and opportunities with big data 2011-1,” 2011.

[8] P. Singh, “Big genomic data in bioinformatics cloud,” Appli Microbio Open Access, vol. 2, no. 1000113, pp. 2, 2016.

[9] O. Trelles, “On the parallelisation of bioinformatics applications,” Briefings in Bioinformatics, vol. 2, no. 2, pp. 181–194, 2001. [10] Y. Han and H. Kim, “A scalable computing framework for large-scale bioinformatics analysis,” 2013.

[11] S. Shrivastava and R. Badlani, “Data storage in dna,” International Journal of Electrical Energy, vol. 2, no. 2, pp. 119–124, 2014.

[12] M. Shahzad and K. Ahsan, “Comparison of big data analytics tools: A bioinformatics case study,” FUUAST Journal of Biology, vol. 4, no. 1, pp. 113, 2014.

[13] Y.-M. Teo, X. Wang, and Y.-K. Ng, “Glad: a system for developing and deploying large-scale bioinformatics grid,” Bioinformatics, vol. 21, no. 6, pp. 794–802, 2005.

[14] M. KA and G. Raju, “A study on applications of grid computing in bioinformatics.”