International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

Next Generation Sequencing in Big Data

Chinmayee C, Amrita Nischal, C R Manjunath, Soumya K N
Department of Information Science & Engineering,
Jain University – School of Engineering & Technology, Karnataka, India

ABSTRACT

A huge revolution has taken place in the area of Genomic science. Sequencing of millions of DNA strands in parallel and also getting a higher throughput reduces the need to implement fragment cloning methods, where extra copies of genes are produced. The methodology of sequencing a large number of DNA strands in parallel is known as Next Generation Sequencing technique.

An overview of how different sequencing methods work is described. Selection of two sequencing methods, Sanger Sequencing method and Next generation sequencing method and analysis of the parameters used in both these techniques. A Comparative study of these two methods is carried out accordingly. An overview of when to use Sanger sequencing and when to use Next generation sequencing is described. Increase in the amount of genomic data has given rise to challenges like sharing, integrating and analyzing the genetic data. Therefore, application of one of the big data techniques known as Map Reduce model is used to sequence the genetic data. A flow chart of how genetic is processed using MapReduce model is also present. Next Generation Sequencing for analysis of huge amount of genetic data is very useful but it has few limitations such as scaling and efficiency. Fortunately recent researches have proven that these demerits of Next Generation Sequencing can be easily overcome by implementing big data methodologies.

Keywords: *Next Generation Sequencing, DNA Strands, Bi data analytics, MapReduce model, Sanger Sequencing*

I. INTRODUCTION

Here are lots of efforts going on to gather and understand genetic data. Industries are gathering and holding onto larger amount of samples, sequencing the samples more rapidly and mining the data for useful medical patterns, information, and trends. The amount of data collection will continue to increase tremendously in future. The big data terms variety, velocity and volume come into the picture. Volume of genetic data can be measured based on the number of interesting samples of information for each occurrence. A multinational pharmaceutical corporation would probably consider 500 terabytes as big data. Data variety can be structured, unstructured and semi-structured data. Data velocity is all about the speed at which the information is created, gathered and processed. But the actual problem arises when we have to secure the information collected. Privacy is the biggest issue. There may be personally identifiable information or sensitive information that has to be protected and prevented from unauthorized access.

Bioinformatics is a field that develops software tools and methods that helps in analyzing and interpreting biological data such as genomic data. Computer programming is an integral part of bioinformatics which helps in identifying the candidate genes. DNA sequences can also be identified and analyzed in bioinformatics. But before the sequences are analyzed, it has to be obtained first. And obtaining raw data is still a problem due to the noise or weak signals. The data obtained might be very huge and analysis might be very difficult.

Next Generation Sequencing is a young field, and the first few machines were introduced in the year 2005.

Within 10 years Next generation sequencing has become the foundation of molecular biology and genetic studies. Being familiar with the technical aspects will help us in understanding and analyzing the available surveys or researches better.

Overall Next Generation sequencing has become a popular methodology that can help in various scientific fields and has transformed bioinformatics field. Millions of DNA sequences can immediately be processed at very lower rates. We can say that this is only the beginning of a new scientific era.

Another famous definition of Next Generation Sequencing, it is a sequencing method where millions of sequencing reactions are carried out in parallel which automatically increases the sequencing throughput.

To understand next generation sequencing, first we need a clear image of what is DNA sequencing and few biological terms as well. As we all know DNA is the outline of life which consists of chemical building blocks known as nucleotides. Nucleotides are generally made of phosphate, sugar groups and one of the nitrogen bases. The nitrogen bases can be any of these four mentioned: Adenine(A), Thymine(T), Guanine(G), Cytosine(C).

There are strands of DNA that needs to be analyzed in sequencing. Therefore, to form a strand of DNA, nucleotides are linked into chains. The biological instructions can be determined through the analysis of order of the bases mentioned previously. We can take a simple example, sequence ATCGTT may refer to blue eyes, whereas ATCGCT might be for brown eyes. What is gene? Gene is made up of DNA or RNA, that can act as instructions to make molecules known as proteins. Generally, in humans the size of the gene can range from 1000 bases to 2300 kilo bases. DNA has a double helical structure.

The story or the era of DNA sequencing began centuries before, when Sanger studied the insulin and showed how important sequencing is in biological studies. Two of the famous methods of sequencing were developed during the same period. The two methods were Sanger's Chain termination sequencing method and Maxam-Gilbert sequencing method. But the common method of choice in these two methods was Sanger's method because of its cost effectiveness and simplicity factor. As the technology advanced several new next generation sequencing methods have surfaced upon, which provides higher throughput and amazing opportunities in bioinformatics field. Generally, the DNA sequencing is referred for techniques that determine the order of nucleotides bases such as Adenine, Guanine, Cytosine and Thymine in a DNA molecule.

The first DNA sequencing was obtained by researchers in the early 1970's. Dye based sequencers have made DNA sequencing simple, easy and fast. DNA sequencing of genes has become an integral part of researches, biological processes, diagnostics and forensic researches as well.

We can list down some of the limitations or difficulties that are common in DNA sequencing:

1. DNS sequence reaction may fail.
2. The quality of data maybe poor.
3. Few sequencing methods can sequence very small amount of DNA maybe about 300 to 1000 bases only.
4. The quality after sequencing few pieces of DNA may degrade. They maybe good in the beginning few bases (15 to 40), but later the quality may decrease.
5. Some of the sequencing methods maybe time consuming as well
6. Parallel sequencing is not possible in all the DNA sequencing methods except in Next Generation sequencing.

NGS & big data:

The NGS tends to deliver crude information or data that is of size as vast as 1TB for every example frequently postures trouble for the data mining and sensible understanding. Presently the sequencing limit for DNA bases per year is measured to be 13 quadrillion DNA bases.

There is a huge amount of biological data that is generated through Next generation sequencing in research laboratories all over the world. Big data often comes with five key words which is: Variety, Velocity, Volume, Veracity and Value. Therefore, there is an immediate requirement for new databases that has the capacity to store large amount of varied data sets. Extracting and transferring the huge amount of genomic data can prove to be a difficult work. There are researches going on where researchers are trying to shrink the large volume of NGS data by finding out several patterns in the data and mapping and reducing them.

II PROBLEM DEFINITION

Genome sequencing is similar to decoding, but a sequence is a code. We can say that genome sequence is a long string of letters in a mysterious language. The explosion in the amount of genomic data has given rise to challenges like sharing, archiving, integrating and analysing the genetic data. Researchers have often found interest in genome sequencing. There has been drastic changes and advances in this area of research. Sequencing methods were discovered that gave high throughput. The data obtained might be very huge and analysis might be difficult. The variety and volume of the data will obviously increase. Therefore, Big data models are useful tools to analyse genetic data.

We will give an overview of sequencing methods and also a comparative study of Next generation sequencing and Sanger sequencing method. We will also describe how MapReduce model can be used for genetic data.

III LITERATURE SURVEY

Recently in an article in The International weekly journal of Science, it was said that storing and processing of genome data would exceed the computing challenges of running YouTube and Twitter[1]. The computing resources needed to handle the genome data will increase. A team of biologists and computer scientists say that the needs of computing in genomics will be huge as sequencing costs drop and even more genomes are analysed.

In a report published in the journal PloS Biology experts say that by 2025 between 100 million and 2 billion human genomes could be have been sequenced. The data storage requirements for this genome data could run up to 2 to 40 exabytes. This is due to the fact that the amount of data that needs to be stored for a single genome is 30times larger than the genome size itself. There may be errors that would have occurred while sequencing or preliminary analysis of genetic data. Therefore, big data analytics plays an important role in analysing the genomic data.

Following are the few researches that have been carried out in the area of analysis of genetic data, sequencing methods and Big Data:

1. In a survey that was carried out by Rashmi Tripathi and Pawan Sharma, from IIT-Allahabad, they have given how the Next generation sequencing has evolved through the Big data analytics[2]. They have said that the massive amount of genomic data has been on the rise and the new challenges such as sharing, integrating and archiving have arisen. Analysis of genetic data has become more difficult due to scale and efficiency challenges of Next generation sequencing. In the paper they have analysed few big data tools and algorithms that can overcome the limitations of NGS. The APACHE based Hadoop framework models that they have discussed in the paper are MapReduce, CloudBurst, Myrna and DistMap. They have also applied big data for NGS and tried to reveal hidden patterns in sequencing and analysis. This survey paper gives a summary, limitations and usage of all the current applications of Hadoop technology in NGS point of view.

2. In a survey paper "Next-Generation Big Data Analytics: State of the Art, Challenge that was published by researchers ZhihanLv, Houbing Song, Pablo Basanta-Val, Anthony Steed, they have given a summary of big data problems and limitations[3]. They have given an overview of the different data types present, storage models, privacy issues, Data security and applications in the area of Big data. Concluding, they have summarized the challenges and development in the area of big data so that the current and future trends can be predicted. They have also given a gist of analysis methods such as MapReduce and Hadoop based Distributed File System.

3. A paper on "Genomic Data Integration", by Emanuel Weitschek, Fabio Cumbo, Eleonora Cappelli and Giovanni Felici(2016)" Genomic Data integration [4] was a research paper that was written by Emanuel Weitschek and Fabio Cumbo, in the year 2016. in this paper the authors have done a research on the advances of NGS technologies in Bioinformatics and how they faced the challenges of larger amounts of genetic and clinical data which are massively growing. In this paper they have focused mainly on analysis and integration of Next generation sequencing data that has been extracted from "The Cancer Genome Atlas"(TCGA).

4. "A Survey Paper on Map Reduce in Big Data"by P. Sudha and Dr. R. Gunavathi who ave carried out a survey on big data and its techniques [5]. The authors

have analysed some of the issues in big data, such as the variety of data types (structures/semi-structured/unstructured) and also the complexity of data. They have analysed the MapReduce model for processing these huge amounts of data that has been distributed on a large cluster. In this paper the authors have highlighted the different applications using Map reduce programming model.

5. In a research paper on "What is Next Generation Sequencing, authors Sam Behjati and Patrick S, have given a clear idea of what exactly is a Next generation sequencing[6]. They have given how the next generation sequencing produces higher throughput parallely. They have also described about how DNA sequencing has helped in genetics research. They have also given a description about the Sanger sequencing methodology. The whole objective of this paper is to review the applications of NGS in paediatrics.

IV SEQUENCING METHODS

A) Sequencing by Hybridization:

In 1990 Ed Southern introduced the method of sequencing by Hybridization, a procedure that depends on the identification of a particular DNA sequence by utilizing hybridization of complementary probes. Back then this was much faster and less expensive than the other methods available. Sequencing by hybridization is one of the strategies for sequencing a DNA sequence. It comprises of two stages: the biochemical and computational ones.

In the biochemical phase a DNA chip, which contains a library of a few distinctive oligonucleotides, is presented to the solution containing numerous duplicates of target DNA. Hybridization happens between the DNA arrangement and oligonucleotides in their corresponding spots. After the response, one acquires an arrangement of short subfragments of the analyzed grouping by perusing a fluorescent or radioactive picture of the DNA chip. This range is called spectrum.

Reconstruction of the target DNA from the oligonucleotides is don't in the second stage, i.e., the computational phase. Sequencing by hybridization makes utilization of an all-inclusive DNA microarray, which harbors all nucleotides of length. These oligonucleotides are hybridized to an obscure DNA section, whose arrangement one might want to decide.

In actual Hybridization experiments there is a possibility of encountering errors. There are two types of errors which is explained below:

A. Positive errors: These are extra oligonucleotides that show up in the range, however don't fit to the target DNA.
B. Negative errors: These are the missing oligonucleotides in the range, contrasting with the perfect range. An extraordinary case are negative mistakesoriginating from reiterations, that are rehashed subsequences in the first succession however display just in one duplicate in the range.

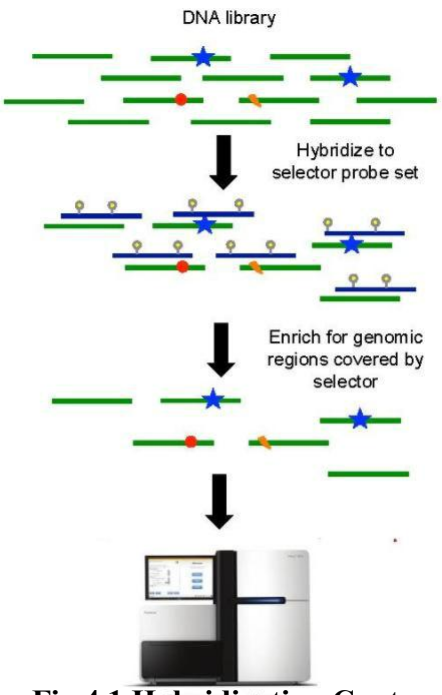


Fig 4.1 Hybridization Capture

With the occurrence of these errors DNA sequencing becomes a harder task. With the expanse of technology there are much cheaper and more accurate methods of DNA sequencing, such as sequencing by synthesis, which came after Hybridization and replaced it. Although Hybridization is still used in some sequencing schemes such as reverse hybridization.

B) Sanger Sequencing

Sanger's method of DNA sequencing, also known as dideoxy chain termination method was the method ever to sequence the genetic code and was devised by Fredrick Sanger in the year 1977. The strategy has been widely used to propel the field of practical and functional genomics, transformative hereditary qualities and disease research. Prominently, the

dideoxy technique was utilized in sequencing the main human genome in 2002. On account of its appropriateness for routine approval of cloning examinations and PCR sections, Sanger sequencing remains a mainstream method in numerous research centers over the world.

In this method, first the DNA is divided into two strands. With the help of chemically altered bases the first strand to be sequenced is copied. These altered bases act as the identification markers and stop the copying process each time one of these is encountered. This same procedure is repeated for all the four bases and then the pieces are put together, finally revealing the original sequence of the DNA.

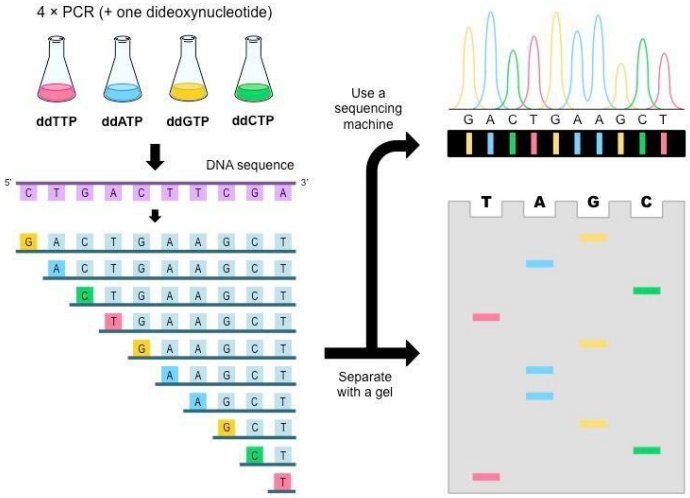
Sanger sequencing consists of three steps:

i. Template Preparation: The quality of the template prepared is one of the most important things in Sanger method. Copies of template strands are prepared at each 3' end. This is done with the help of a DNA primer.

ii. Generation of nested set of labelled fragments: Various replication reactions are performed using copies of each template by dividing them into four parts. The standard primer is copied and used in all four batches along with polymerase I. To blend parts that ends at A, ddATP is included to the response blend cluster I alongside dATP, dTTP,dCTP and dGTP, standard preliminary and DNA polymerase I. Thus, to produce, all sections that ends at C, G and T, the separate ddNTPsieddCTP,ddGTP and ddTTP are added individually to various response blend on various clump alongside normal dNTPs.

iii. Electrophoresis and Gel Reading: The response blend from four clusters are stacked into four distinctive well on polyacrylamide gel and electrophoresed.The autoradiogram of the gel is perused to decide the request of bases of correlative strand to that of format strand.The band ofbriefest pieces are at the base of autoradiogram, so the arrangements of correlative strand is perused from base to top.

There are various advantages of Sanger sequencing such as focusing on littler genomic districts in a bigger number of tests, validating the results of Next Generation Sequencing and many others but Next Generation Sequencing is still a faster, cheaper and better option.



4.2 Generation of nested set of labelled fragments: Sanger Sequencing

C) Maxam Gilbert's Chemical Sequence method

Before the Sanger sequencing method, there were two DNA sequencing techniques presented by Alan Maxam and Walter Gilbert in 1973 and 1976. The techniques were:

i. Wandering-spot Analysis: Made use of 24 base pairs.

ii. Chemical Sequencing: Chemical processes were used to terminate DNA strands and then run through gel to determine the sequence.

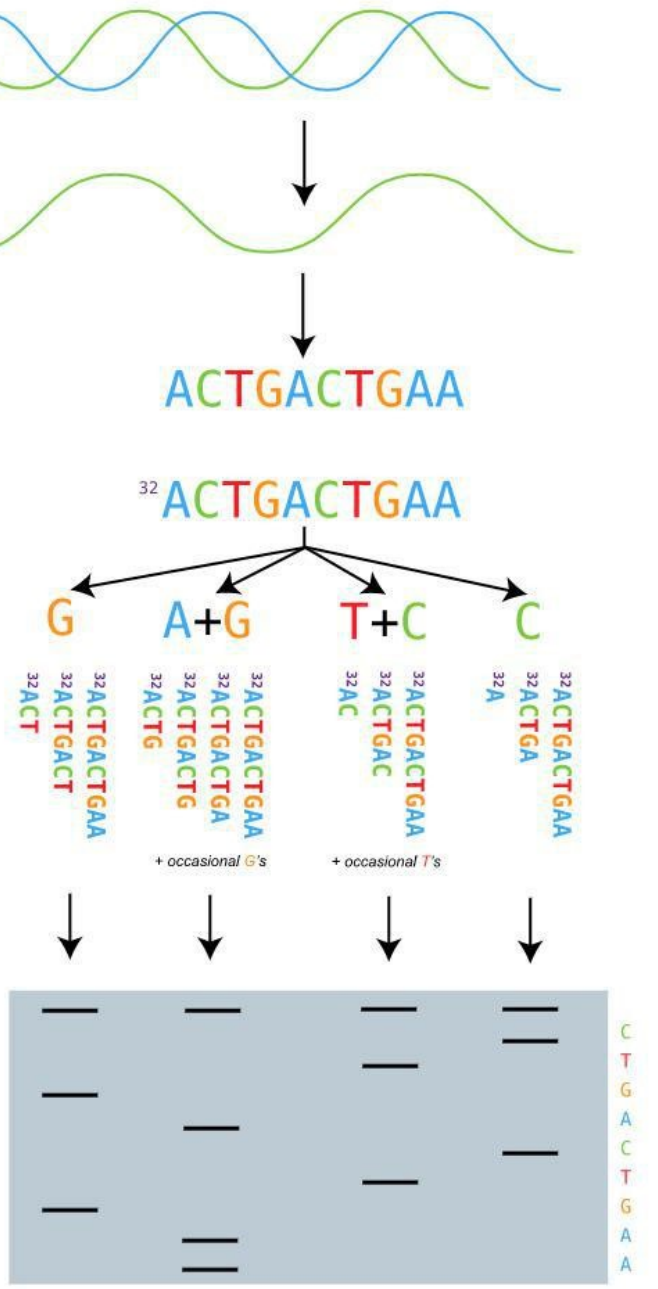
Maxam Gilbert's method consists of the following steps:

i. The double stranded DNA is broken into single stranded DNA's by increasing the temperature.

ii. Then the DNA is labelled at 5' using P^{32}

iii. The following stage separates the DNA. Furthermore, this is the place the Maxam-Gilbert sequencing gets extremely intriguing. With the help of Piperidine, the DNA is partitioned with different focuses. The grouping is done such as, wherever there is a C, or there is a C or T. If this sample is partitioned into four tubes, depending on the chemicals, four different fragments are obtained.

iv. These reactions are then loaded on to a high percentage polyacrylamide gel, to differentiate fragment sizes. The fragments are visualized via the radioactive tag.



4.3 Steps in Maxam Gilbert's Method This method was a start to the idea of DNA sequencing but had many flaws in it. It could only sequence 200-300 bases of DNA in a couple of days. Usage of a large amount of radioactive material was involved and thus it was considered hazardous. If any one step would go wrong there were chances of the entire process going wrong.

D) Automatic DNA Sequencer

After Sanger, in 1986 Lorey Hood invented the Automatic DNA Sequencer. Lorey solved the problem that came with the radioactive markers by changing them to fluorescent markers. With the use of four different colors, namely, Red, Green, Blue or Yellow, the four reactions were combined into one and increasing the throughput. Even better, the outline utilized a laser to examine the examples in the gel and a PC to peruse the outcomes. That first framework, showcased by Applied Biosystems of Foster City, Calif., could create then-noteworthy 4,800 bases of arrangement for every day. Suddenly, institutional center sequencing offices were a pragmatic probability.

The fluorescent dye has no environmental hazard and came with no special handling instructions. Shifting degrees of robotization are likewise accessible. For full mechanization, all that is required is to stack an example plate with layout DNA; the hardware plays out the marking and investigation. The other alternative is to play out the marking responses with fluorescent colors, stack the examples onto a gel, and place the gel into the DNA sequencer. The hardware plays out the detachment and examination. The framework consequently distinguishes the nucleotide succession and recoveries the data on the PC. In this way, just a survey of the information is important to guarantee no irregularities were misidentified by the computer. The greatest obstacle back then with this technique was the insufficient knowledge to interpret the computer-generated results. Today we have better, and more efficient computer-generated techniques and people have a better knowledge of the workings of a computer.

V NEXT GENERATION SEQQUENCING AND SANGER SEQUENCING TECHNIQUES

Researchers these days have many options to choose among the different sequencing technologies for their genetic researches. The biggest confusion here is "Which would be the best sequencing method for our application?". Sanger sequencing method or next generation sequencing?

Back in the 1990s, sequencing of genetic data was done in a very old-fashioned manner, where they used gels and manually called the bases. After some years, Sanger sequencing method was introduced in this bio informatics field. This sanger sequencing method was introduced by a British biochemist known as DrFrederick Sanger in the year 1970. In this method, a single stranded DNA.

Now when the technology has evolved, next generation sequencing has revolutionized the way the

sequencing is performed. The biological sciences has been revolutionized through a parallel sequencing technology which is called as next generation sequencing. This method is a parallel way of sequencing the genetic data. As we know that NGS has high throughput, scalability and speed which is best for sequencing genes. Before it was not possible to study the biological systems or genetic systems easily, but now NGS allows researchers to perform and analyse the biological data. NGS is almost similar to Sanger sequencing method. In both the methods we sequence DNA fragments. But in NGS, we can sequence millions of DNA fragments in a single run. Whereas in Sanger sequencing, it produces one forward and reverse read.

These days complex genomic research needs a in-depth information other than the traditional DNA sequencing techniques. In this case Next generation sequencing helps and has become a daily research tool to address genomic survey questions for researchers. NGS supports variety of applications such as gene expression profiling, chromosome counting, molecular analysis etc.

Although sanger sequencing gives us 99.99% accuracy for researches, the newer methodologies like NGS are also becoming popular because of the higher throughput it offers and it is also cost effective. That is sequencing can be done at a lower cost.But which one should we use? Sanger or NGS? For this we will be doing a comparative study, to find out which methods suits best for what kind of input.

5.1 Comparative study of Sanger sequencing method and
Next generation sequencing:

Let us take an example where we have to sequence a cancer sample from a patient. Here the cancer cells are 10% of the total cells present in a human body. Let us also assume that we know that cancer cells contain a single nucleotide mutation. What will the result of Sanger sequencing vs NGS be?

If Sanger sequencing method is used, we would generate a single sequence and also there would be a combination of all the DNA fragments and there is no other way to isolate the individual signals that originate from each of the DNA molecule. Therefore, at nucleotide of interest, a mixed signal originates from 90% normal cells and 10% cancer cells.
In case Next generation sequencing, we prepare a target DNA molecule library and load it to generate DNA clusters which are then parallelly sequenced. Here we have to note that, each cluster is usually generated from a single molecule of DNA targeted. At the end, we get 90% of clusters from normal cells and 10 % from cancer cells. This allows us to efficiently isolate, detect and quantify signals. We can finally say that, Sanger sequencing individual data channel is analogical whereas Next generation sequencing is Digital since it allows us to separate each

Table 5.1.1 Sanger Sequencing Method vs Next Generation Sequencing:

SL No	SANGER SEQUENCING METHOD	NEXT GENERATION SEQUENCING
1	In Sanger method, for sequencing a single gene, there is a separate reaction.	In next generation sequencing methods, for simultaneous analysis of various genes there is one single reaction. NGS has facilitated collection of large amounts of nucleotide information in sequence read length from 30 to1500 nucleotides for 100's of 1000's to millions of DNA molecules simultaneously.
2	This method has a higher precision	This is highly cost effective and also very efficient due to the faster analysis. Higher throughput is produced here
3	This method is expensive and also time consuming and labour intensive.	Here the interpretation of abundant data is quiet a challenge.
4	This Sanger method does not offer a fast turnaroundtime but it is more economical technology.	**Next generation sequencing** offers fast turnaround time and takes only about 4 hours to complete a run.

Table 5.1.2 When to choose Sanger sequencing method and when to choose Next Generation sequencing method?

SL No	**Choose Sanger sequencing method when:**	**Choose Next generation sequencing method when:**
1	Sequencing single genes	Sequencing more than 100 genes at a time and cost effectively
2	Sequencing 1-100 amplicon targets at the lowest cost	Finding novel variants by expanding the number of targets sequenced in a single run.
3	Sequencing up to 96 samples at a time.	Sequencing more than 96 samples for multiple targets
4	Microbial Identification	Sequencing samples that have low inputamounts of starting material,
5	Fragment analysis is done.	

5.2 Map Reduce and Sequencing:

In simple terms, MapReduce can be defined as a list of values which is mapped into another list of values, which further gets reduced into a single value. MapReduce can also be defined as a programming model which is used for processing and analysing large data sets in parallel. distributed method on a cluster. Hadoop Distributed File System is a system that is having the capability to store large data sets and also to stream those data sets at high bandwidth.

To handle big data, MapReduce can be coupled with HDFS. The combination of HDFS and MapReduce can be referred to as Hadoop.in MapReduce data is treated as Key and value pair. Before the data or information is fed to the MapReduce model, the data types in the information, whether it is structured or unstructured, it must be translated first. MapReduce model has two functions which is known a Map-function and Reduce function.

Problems for MapReduce:

i. Parallel problems
ii. Problems that are having larger data sets
iii. Problem data is complex in structure.
iv. Ability to perform common operations on all data
v. Problems that can be converted into key value pair.

Table 5.2.1 Simple Example of Map reduce

One Rat	**Map()**	**One-1**	**Rat-1**
Two Rat	**Map()**	**Two-1**	**Rat-1**
Three Rat	**Map()**	**Three-1**	**Rat-1**
Blue Rat	**Map()**	**Blue-1**	**Rat-1**
White Rat	**Map()**	**White-1**	**Rat-1**
One-1	**Reduce()**	**One-1**	
Two-1	**Reduce()**	**Two-1**	
Three-1	**Reduce()**	**Three-1**	
Blue-1	**Reduce()**	**Blue-1**	
White-1	**Reduce()**	**White-1**	
Rat-1-1-1-1-1	**Reduce()**	**Rat- 5**	

The same example shown above can be done for nucleotides as well:

Table 5.2.2 MapReduce for Nucleotides

ddATP	**Map()**	**d-1-1-1-1-1-1-1-1**
ddGTP	**Map()**	**A-1**
ddCTP	**Map()**	**T-1-1-1-1-1**
ddTTP	**Map()**	**P-1-1-1-1**
	Map()	**G-1**
	Map()	**C-1**
d-1-1-1-1-1-1-1-1	**Reduce()**	**d-8**
A-1	**Reduce()**	**A-1**
T-1-1-1-1-1	**Reduce()**	**T-5**
P-1-1-1-1	**Reduce()**	**P-4**
G-1	**Reduce()**	**G-1**
C-1	**Reduce()**	**C-1**

The following flow chart depicts an overall view of how genetic information can be mapped and reduced:

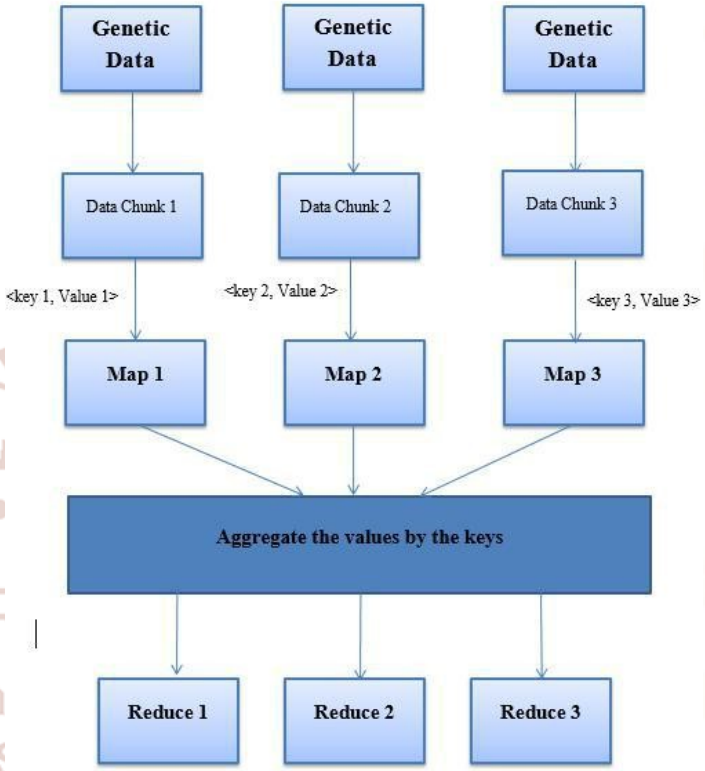


Figure 5.1 Flow chart for MapReduce

5.3 The Past, Present and Future of DNA sequencing:

Fredrick Sanger believed that:

"The order of nucleic acids in polynucleotide chains ultimately contains the information for the hereditary and biochemical properties of terrestrial life. Therefore the ability to measure or infer such sequences is imperative to biological research."

DNA sequencing has seen many phases and it started back in 1953 when Watson and Crick could figure out the three dimensional structure of the DNA. Initial research only focused on the structure of easily available RNA's of bacteria. There was a still a long way to go for the scientists to be able to sequence the DNA. In 1968, primer reduction methods were introduced. In 1973, Maxam and Gilbert were able to come up with a sequencing method. In 1977, Fredrick Sanger came up with the still widely used Sanger method of sequencing. By 1987, automated Fluorescence based Sanger sequencing machines were available. These could generate around 1000 bases per

day. With the exponential growth of Genomic data it was getting more and more difficult to sequence it. Int the year 1986, over 10 million bases were deposited in the GenBank.

As time passed by scientists came to realize the importance of understanding the DNA to go further in biomedical research. Then came the Human Genome Project. The approach shotgun sequencing was developed during this phase. In his approach the long DNA fragments could be broken down into small sequences which could be analyzed easily. The complete sequence could then be analyzed by the reassembly of these results. This is what made the sequencing of the entire Human Genome possible.

The only problem remaining now was the time it took for the DNA sequencing. This is what compelled the scientists to come up with the science of Next Generation Sequencing with the help of which large number of fragments could be sequenced in parallel. This is the reason NGS is also known as massively parallel sequencing.

5.4 Applications of Next Generation Sequencing:

SL No	Field	Application
1	Forensics	To help identify individuals as each individual has a different genetic sequence
2	Agriculture	The mapping and sequencing of a genome of microorganisms has helped make them useful for crops and food plants.
3	Medicine	i. Used to help detect the genes which are linked to various genetic disorders. ii. Prediction of response to immunotherapy in patients with cancer.

Table 5.4.1 Applications of Next Generation Sequencing

VI CONCLUSION

It is difficult to exaggerate the significance of DNA sequencing to biomedical research, since this is the property by which all living things can be characterized and separated from each other. This is the reason why scientists all over the world have spent their lifetimes researching on faster and better ways to sequence the DNA.

Humans have predicted the relationship between heredity and diseases for a long time. Only in the beginning of the last century, scientists begin to discover the connotations between different genes and disease phenotypes. Recent trends in next-generation sequencing (NGS) technologies have brought a great momentum in biomedical research that in turn has remarkably augmented our basic understanding of human biology and its associated diseases. State-of-the-art next generation biotechnologies have started making huge strides in our current understanding of mechanisms of various chronic illnesses like cancers, metabolic disorders, neurodegenerative anomalies, etc. We are experiencing a renaissance in biomedical research primarily driven by next generation biotechnologies like genomics, transcriptomics, proteomics, metabolomics, lipidomics etc.

We are currently in the era of "big data," in which big data technology is being rapidly applied to biomedical and health-care fields. Big Data analytics is a blessing since it helps deal with the massive volume and complexity of bioinformatics data. Combining the knowledge of both the fields(Big Data and Next Generation Sequencing) humans finally have hope to cure various terminal illnesses like Cancer and HIV.

REFERENCES

1) Erika Check Hayden, "The International Weekly Journl of Science, an article "Genome Researchers raise an alarm over big data" July 7th 2015

2) Rashmi Tripathi, Pawan Sharma, Pavan Chakraborty & Pritish Kumar Varadwaj , A survey paper on "Next-generation sequencing revolution through big data analytics", (2016).

3) Zhihan Lv, Houbing Song, Pablo Basanta-Val, Anthony Steed, Minho JoA paper on "Next-Generation Big Data Analytics: State of the Art, Challenge",2017

4) Emanuel Weitschek, Fabio Cumbo, Eleonora Cappelli and Giovanni Felici , A paper on "Genomic Data Integration",2016 .

5) P. Sudha , Dr. R. Gunavathi ,"A Survey Paper on Map Reduce in Big Data ",(2015)

6) Sam Behjati and Patrick S Tarpey, A paper on "What is next generation sequencing?" (2015)

7) Schuster, S.C. et al. Method of the year, next-generation DNA Sequencing, Functional genomics and medical applications, 2015

8) CA Cummings, E Peters, L Lacroix, F Andre, MR Lackner, The Role of Next-Generation Sequencing in Enabling Personalized Oncology 2016

9) Reva Kakkar Basho, Agda Karina Eterovic, Funda Meric-Bernstam, Clinical Applications and Limitations of Next-Generation Sequencing. 2016

10) Jake Luo, Min Wu, Deepika Gopukumar and Yiquing

11) Zhao,Big Data Application in Biomedical Research and Health Care: A Literature Review. 2016

12) Jay Shendure, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss8 & Robert h. Waterston , DNA sequencing at 40: past, present and future. 2017

13) Sinchita Roy-Chowdhuri, Somak Roy, Sara E. Monaco, Mark J. Routbort and Liron Pantanowitz Big data from small samples: Informatics of next-generation sequencing in cytopathology, 2016

14) Karen Y. He, Dongliang Ge and Max M. "Big Data Analytics for Genomic Medicine", 2017