ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 4, Issue 4, April 2017

Big Data Analysis On Gene Mutation Detection For Bacterial Diseases

[1] Sharath Kumar S, [2] Sanil A Jain, [3] Veeresh Shenoy S, [4] Vinith G, [5] M Ravi Krishna
[1] Assistant Professor [2] [3] [4] [5] UG Scholar
Department of Information Science and Engineering
Siddaganga Institute of Technology, B H Road, Tumakuru, Karnataka, India 572103

***Abstract:*-- Big data and analytics is a method of examining a lot of datasets to uncover hidden patterns, market trends, and other useful business information. In a human body, the gene will generate a significant amount of data; the real challenge here is to find, collect, analyze and manage information so that we can avoid the occurrence of the illness and be healthier. Nowadays many people are losing their lives because of bacterial diseases like MRSA, CRE, etc. So we are trying to collect the data sets of a particular disease and analyze it by detecting gene mutation, which helps people to find which gene is the actual cause of the particular disease. Here we will be finding which mutation is responsible for that drug resistance. This drug resistant gene will be found out in a large population. A mutation is the permanent change of the nucleotide sequence of an organism, genetic elements, virus or extrachromosomal DNA. Errors during DNA replication results in the mutation which further undergoes error-prone repair. Through genetic recombination, mutations can involve the duplication of vast sections of DNA.**

***Index Terms*—NumPy, Mutation, Gene, MRSA (Methicillin resistant Staphylococcus Aureus), DNA.**

I. INTRODUCTION

Big data analysis (BDA) on gene mutation detection for bacterial diseases is a method of detecting the gene that is resistant to a particular drug. Big data and analytics is a process of examining a lot of datasets to uncover hidden patterns, market trends, and other useful business information. A mutation is the permanent change of the nucleotide sequence of an organism, genetic elements, virus or extrachromosomal DNA. Bacterial diseases like MRSA, CRE, mic, etc. are fatal nowadays. We are working on MRSA (Methicillin-Resistant Staphylococcus Aureus) is a bacterium that is resistant to many antibiotics.

Genomic data production now has the velocity, volume, variety and veracity to be considered ‘Big Data’. As a result, the processing of genomic data can be improved by applying lessons and techniques from other industries that have worked successfully with Big Data. Pathogenic bacteria are bacteria that can cause infection. Although most bacteria are harmless or often beneficial, some are pathogenic. One of the bacterial diseases with the highest disease burden is Tuberculosis, caused by the bacterium Mycobacterium Tuberculosis, which kills about 2 million people a year, mostly in Africa.

Pathogenic bacteria contribute to other globally important diseases, such as pneumonia, which can be caused by bacteria such as Streptococcus. The preeminent goal of architecting big data solutions is to create reliable, scalable and capable infrastructure. At the same time, the analytics, algorithms, tools and user interfaces will need to facilitate user with users, specifically those in execute-level.

Big data architecture is similar to any other architecture that originates or has a foundation from reference architecture. Understanding the complex hierarchal structure of reference architecture provides a good background for understanding big data and how it complements existing analytics, business intelligence, databases and other systems.

Methicillin-Resistant Staphylococcus Aureus (MRSA) is a disease prone to human beings usually working in hospitals and clinical laboratories. Genomic data production now has the velocity, volume, and variety to be considered ‘Big Data.' For our work, we are considering only volume and variety. As a result, there will be the improvement in the processing of genomic data by applying lessons and techniques from other industries that have worked successfully with Big Data.

II. LITERATURE SURVEY

In paper [1], the molecular processes MRSA infection are poorly understood. Although a major role has been attributed to the acquisition of virulence determinants by horizontal gene transfer, there are insufficient epidemiological and functional data supporting that concept. We here report the spread of clones containing a previously extremely rare mobile genetic element–encoded gene, sasX. Our study identifies sasX as a quickly spreading crucial determinant of MRSA pathogenic success and a promising target for therapeutic interference. Our results provide proof of principle that horizontal gene transfer of key virulence determinants drives MRSA epidemic waves. In paper [2], it provides an overview of recent developments in big data in the context of biomedical and health informatics. It outlines the key characteristics of big data and how medical and health informatics, translational bioinformatics, sensor informatics, and imaging informatics will benefit from an integrated approach of piecing together different aspects of personalized information from a diverse range of data sources, both structured and unstructured, covering genomics, proteomics, metabolomics, as well as imaging, clinical diagnosis, and long-term continuous physiological sensing of an individual. It is expected that recent advances in big data will expand our knowledge for testing new hypotheses about disease management from diagnosis to prevention to personalized treatment. The rise of big data, however, also raises challenges in terms of privacy, security, data ownership, data stewardship, and governance. This paper discusses some of the existing activities and future opportunities related to big data for health, outlining some of the key underlying issues that need to be tackled. In paper [3], One of the most pressing issues facing the pharmaceutical and biotechnology industry is the tremendous dropout rate of lead drug candidates. Over the last two decades, several new genomic technologies have been developed in hopes of addressing the issues of target identification and lead candidate optimization. Gene expression microarray is one of these technologies and this review describes the four main formats, which are currently available: (a) cDNA; (b) oligonucleotide; (c) electro kinetic; and (d) fiber optic. Many of these formats have been developed with the goal of screening large numbers of genes. Recently, a high-throughput array format has been developed where a large number of samples can be assayed using arrays in parallel. In addition, focusing on gene expression may be only one avenue in preventing lead candidate failure. Proteomics or the study of protein expression may also play a role. Examples of several gene and protein expression studies as they apply to drug discovery and development are reviewed. These studies often result in large data sets. These newer genomic and proteomic technologies and their analysis and visualization methods have the potential to make the drug discovery and development process less costly and more efficient by aiding to select better target and lead candidates.

III. SYSTEM DESIGN

The system design consists of six major steps. The general block diagrams of these steps are shown in the Fig.1.

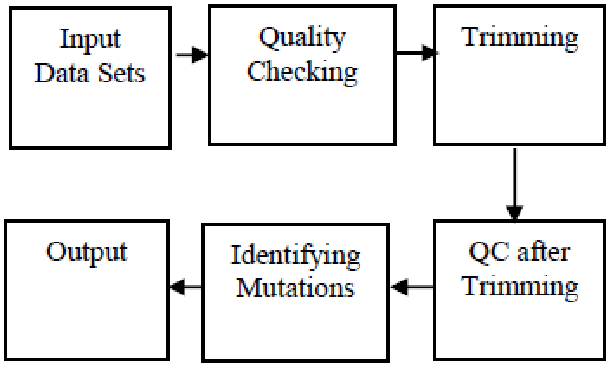
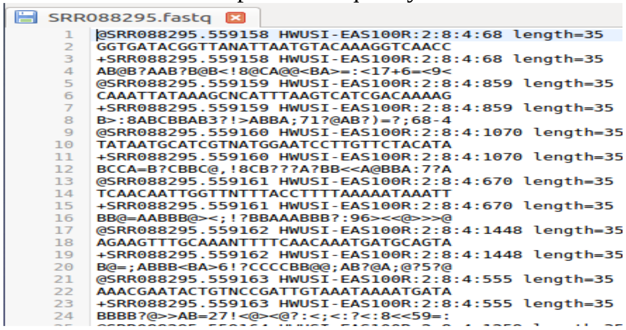


Fig.1 Work Flow of System

A. Input Data Sets

The storage capacity of the data sets of genomic sequences of MRSA is around 140 GB. In the first process, we give the input data sets of MRSA which we have downloaded from DDBJ (DNA Data Bank of Japan). These datasets contain the sequences (forward and reverse sequence) which are formed of four major nucleotides namely A (Adenine), T (Thymine), G (Glycine) and C (Cytosine) respectively. The fig.2 represents the file in the FASTQ format. This file contains the information about the dataset i.e. filename with fastq extension, the length of the sequence or read and the nucleotides with their particular quality score.



```
@SRR088295.559158 HWUSI-EAS100R:2:8:4:68 length=35
GGTGATACGGTTANATTAATGTACAAAGGTCAACC
+SRR088295.559158 HWUSI-EAS100R:2:8:4:68 length=35
AB@B?AAB?B@B<!8@CA@@<BA>=:<17+6=<9<
@SRR088295.559159 HWUSI-EAS100R:2:8:4:859 length=35
CAAATTATAAAGCNCATTTAAGTCATCGACAAAAG
+SRR088295.559159 HWUSI-EAS100R:2:8:4:859 length=35
B>:8ABCBBAB3?!>ABBA;71?@AB?)=?;68-4
@SRR088295.559160 HWUSI-EAS100R:2:8:4:1070 length=35
TATAATGCATCGTNATGGAATCCTTGTTCTACATA
+SRR088295.559160 HWUSI-EAS100R:2:8:4:1070 length=35
BCCA=B?CBBC@,!8CB???A?BB<<A@BBA:7?A
@SRR088295.559161 HWUSI-EAS100R:2:8:4:670 length=35
TCAACAATTGGTTNTTTACCTTTTAAAAATAAATT
+SRR088295.559161 HWUSI-EAS100R:2:8:4:670 length=35
BB@=AABBB@><;!?BBAAABBB?:96><<@>>>@
@SRR088295.559162 HWUSI-EAS100R:2:8:4:1448 length=35
AGAAGTTTGCAAANTTTTCAACAAATGATGCAGTA
+SRR088295.559162 HWUSI-EAS100R:2:8:4:1448 length=35
B@=;ABBB<BA>6!?CCCCBB@@;AB?@A;@?5?@
@SRR088295.559163 HWUSI-EAS100R:2:8:4:555 length=35
AAACGAATACTGTNCCGATTGTAAATAAAATGATA
+SRR088295.559163 HWUSI-EAS100R:2:8:4:555 length=35
BBBB?@>>AB=27!<@><@?:<;<:?<:8<<59=:
```

Fig.2 FASTQ file format

B. Quality Checking

Quality checking tool aims to check the quality of data sets that are in the FASTQ format. The tool is developed using NumPy – a Python library meant for faster processing of huge amount of data. The parameters that are mainly considered are as follows:

(i) %GC= No. of G's + No. of C's
Total No. of Nucleotides
Where 'G' and 'C' are Glycine and Cytosine nucleotides.

(ii) Sequence Length: Total number of nucleotides in a single read.

(iii) Total number of sequences: Total number of reads in a data set file.

The graphical report is generated after quality checking as shown in fig.3. Here the X-axis represents nucleotides in a sequence and Y-axis represents its corresponding ASCII value.

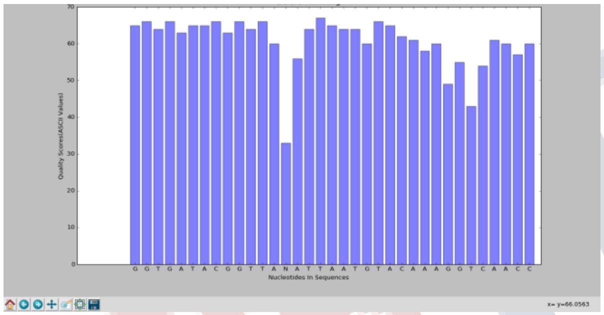


Fig.3 Graphical Report of Quality Checking Tool

C. Trimming

Trimming tool is used to trim and crop the unwanted nucleotides in the FASTQ data sets. The unwanted nucleotides are represented by 'N' and the respective quality scores. The type of trimming done here is pair ended mode since we have two input files (for forward and reverse reads). There are four output files (for forward paired, forward unpaired, reverse paired and reverse unpaired reads). Paired reads have good quality of data and unpaired reads contains unwanted trimmed data. Hence, paired reads are considered for further data processing.

D. Assembling of data

For the assembly of data, we would only consider only paired reads which are obtained after trimming and the unpaired reads are ignored. A tool named Velvet is used for assembling forward read and backward read file into a single file. Different K-mer values are considered for assembling data where 'K' is the length of the sequence. The value of 'K' considered is odd since the sequence length of MRSA is 35 i.e., odd. The value of 'K' lies between 19 (half of the sequence length) to 35. The graphs are generated as shown in fig. 4. The X-axis represents the number of matches occurred for that particular K-mer value throughout the file and the Y-axis represents the nodes value.

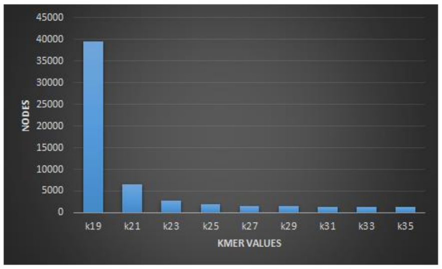


Fig.4 No. of Nodes v/s K-mer values

E. Output-Identifying Mutations

Mutation in gene is identified by 'contig' file generated by the tool Velvet. This contig file would be compared with reference genomic sequence of MRSA. We are considering the genomic sequences of mecA and gyrA. These mecA and gyrA are the two basic genes in MRSA. The genome considered is mecA and gyrA whose sequences are divided into three nucleotides each and coded along Genetic code table. This is implemented using dictionary phenomenon of Python. These divided nucleotides would be compared with standard genomic sequences of mecA and gyrA. If the divided nucleotides and reference genomic sequences of MRSA doesn't match, then there is a mutation in the gene else there exists no mutation. This report on genetic mutation will be to identify the drug that is helpful in curing the disease. There is no mutation in the gene as shown in fig.5 since the upper bound representing the part of reference genomic sequence of MRSA and the lower bound representing the divided nucleotides matches.

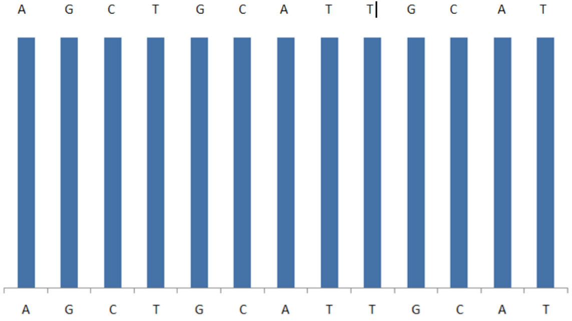


Fig.5 Graphical Representation of unoccured mutation

There is mutation in gene as shown in fig.6 since the upper bound representing the part of reference genomic sequence of MRSA and the lower bound representing the divided nucleotides doesn't match each other respectively.

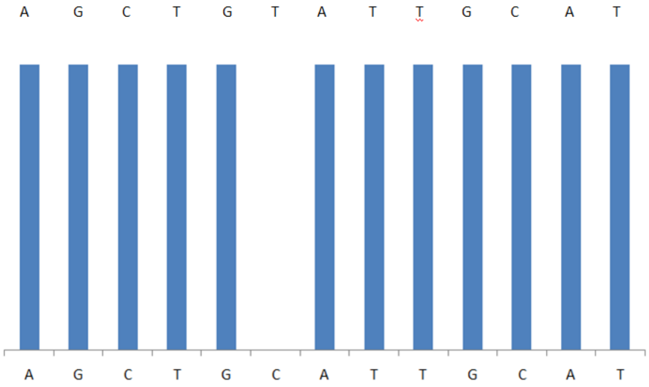


Fig.6 Graphical Representation of occurred mutation

IV. APPLICATIONS

The method replaces the traditional approach of testing the DNA samples which saves a lot of time. The lab technician can upload the data file of genomic sequences into the designed system and further data processing takes place in a pre-defined manner and finally the gene responsible for drug resistance is identified and the conclusions are drawn.

V. CONCLUSION AND FUTURE WORK

The lab technician cannot depend on the research person every time to find the mutation in the genomic sequences. Similarly, using the approach of Big Data and Analytics a system can be designed which helps the human beings to get rid of various deadly bacterial and viral diseases.

REFERENCES

[1] Min Li, XinDu and et.al, "MRSA epidemic linked to a quickly spreading colonization and virulence determinant," NatureMedicine18,816–819(2012)doi:10.1038/nm.2692, 22 April 2012.

[2] IEEE Journal of Biomedical and Health Informatics (Volume: 19, Issue: 4, July 2015).

[3] M.J. Cunningham and et. genomics and proteomics: "The new millennium of drug discovery and development". Journal of Pharmacological and Toxicological Methods, Volume 45, Issue 1, January–February 2001.

[4] Aisling O'Driscoll, Jurate Daugelaite, Roy D. Sleator: "Big data, Hadoop and cloud computing in genomics".

[5] Matti Niemenmaa, Aleksi Kallio, André Schumacher, Petri Klemela, Eija Korpelainen and Keijo Heljanko, "Hadoop-BAM: directly manipulating next generation Sequencing data in the cloud". Bioinformatics 2012, 28(6):876-877, doi:10.1093/bioinformatics/bts054