HIV NOMOGRAM USING BIG DATA ANALYTICS

S.Avudaiselvi and P.Tamizhchelvi

Student Of Ayya Nadar Janaki Ammal College (Sivakasi)
Head Of The Department Of Computer Science, Ayya Nadar Janaki Ammal College (Sivakasi)

ABSTRACT

The developing countries like India with huge population face various problems in the field of healthcare with respect to the expenses, meeting the needs of the economically deprived people, access to the hospitals, and research in the field of medicine and especially in the time of spreading epidemics. The exponential growth of data over the last decade has introduced a new domain in the field of information technology called Big Data. Here, we have focused on ***Acquired Immune Deficiency Syndrome (HIV/AIDS)*** *which seems prevail in many women/children who come to Government Hospital in Sivakasi. HIV is a disease which is not easily identified. This involves certain tests to be carried out like ELISA,* ***OraQuick In-Home HIV Test*** *etc. Most of the people do not come forward for HIV test. So, we attempt to develop a NOMOGRAM which gives possibility percentage of AIDS in a person using big data analytics.*

KEYWORDS

HIV/AIDS, DNA, Hadoop, Map - Reduce framework.

1.INTRODUCTION

Big Data[1] is a high volume, high velocity and high variety information asset that demand cost-effective , innovative forums of information processing for enhanced insight and decision making. Big data, a buzzword in the business intelligence can handle petabytes or terabytes of data in a reasonable amount of time. Big data is distinct from large existing database which uses Hadoop framework for data intensive distributed applications.
The detection of DNA and its variation is critical for many fields, including clinical and veterinary diagnostics, industrial and environmental testing, agricultural researches and forensic science. Disease diagnosis and prognosis are based on effective detection of disease conditions (e.g. cancer), infectious organisms (e.g. HIV) and genetic markers. However, DNA analysis from original specimens is a complex process involving multiple chemical compositions as well as multistep reactions.

The **human genome**[2] is the complete set of nucleic acid sequence for humans (Homosapiens), encoded as DNA within the 23chromosome pairs in cell nuclei and in a small DNA molecule found within individual Mitochondria.DNA is the largest human chromosome, chromosome number 1, consists of approximately 220 million base pairs a would be 85 mm long if straightened.

If you had a perfect sequence of the human genome, then all you would need is the string of letters (A, C, G and T) that make up one strand of the human genome, and the answer would be about 700 megabytes. The volume of data DNA sequence is huge size. Rise of these data leads to a new technology such as big data that acts as a tool to process, manipulate and manage very large dataset along with the storage required.

The current commercially available HIV DNA PCR[3] assays have acceptable sensitivity for detection of most common group M HIV-1 subtypes including A, B, C, D, E, G and H. I use Needleman-wunsch algorithm to compare DNA sequences of HIV genes.

2.SOFTWARE FRAMEWORK

Hadoop [4] is a software framework that supports data-intensive distributed applications under a free license. It enables applications to work with thousands of computational independent computers and Hadoop was derived from Google's MapReduce and Google File System (GFS) papers. Hadoop is built up of mainly two parts:

1. HDFS
2. Map-Reduce Framework

2.1.HDFS

HDFS is the primary distributed storage used by Hadoop applications. A HDFS cluster primarily consists of a NameNode that manages the file system metadata and DataNodes that store the actual data. Hadoop is a software framework that supports data-intensive distributed applications under a free license. It enables applications to work with thousands of computational independent computers and Hadoop was derived from Google's MapReduce and Google File System (GFS) papers.

2.2.Map-Reduce

Hadoop Map-Reduce is a software framework for writing applications easily which processes vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner. A Map-Reduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks.

Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks. Typically the compute nodes and the storage nodes are the same, that is, the Map-Reduce framework and the Distributed File System are running on the same set of nodes.

3.WORKING PRINCIPLE OF HIV NOMOGRAM USING BIG DATA ANALYTICS

Three main steps of the project are:

1. Collecting symptoms of HIV
2. HIV DNA gene sequence alignment using Mapper /Reducer
3. Final score for HIV affected percentage using inputs from 1 and 2

3.1.Symptoms Of HIV

"In the early stages of HIV infection, the most common symptoms are none," says Michael Horberg, MD, director of HIV/AIDS for Kaiser Permanente, in Oakland [5]. This tool is designed based on that symptoms and the values are assigned based on the advise of doctors. The details of HIV symptoms are collected from sivakasi and virdhunagar government hospitals. Within a

month or two of HIV entering the body, 40% to 90% of people experience the following symptoms known as acute retroviral syndrome (ARS).
The Symptoms ranges are as follows:

- A mild fever, up to about 102 degrees F.
- The fever, if it occurs at all, is often accompanied by other usually mild symptoms, such as fatigue, swollen lymph glands, and a sore throat.
- Skin rashes can occur early or late in the course of HIV/AIDS.
- Anywhere from 30% to 60% of people have short-term nausea, vomiting, or diarrhea in the early stages of HIV.
- Once called "AIDS wasting," weight loss is a sign of more advanced illness and could be due in part to severe diarrhea.
- A person is considered to have wasting syndrome if they lose 10% or more of their body weight and have had diarrhea or weakness and fever for more than 30 days.
- Dry cough is an "insidious cough that could be going on for weeks that doesn't seem to resolve.

3.2.Needleman-Wunsch Algorithm To Align Sequences

A scoring function (σ): defines the score to give to a substitution mutation eg. -1 for a match, -1 for mismatch.

A gap penalty: defines the score to give to an insertion or deletion mutation, eg. -1
A recurrence relation: defines what actions we repeat at each iteration3 (step) of the algorithm; for N-W this is:

1. T (i-1, j-1) + σ (S1 (i), S2 (j))
2. T (i , j) = max T (i-1, j) + gap penalty
3. This will be T (i , j-1) + gap penalty

There are 2 parts to computing the best alignment using the N-W algorithm:
1) Fill up a matrix (table) T using the recurrence relation
2) The traceback step: use the filled-in matrix T to work out the best alignment

3.3.DNA Gene Sequence Alignment Using Mapper / Reducer

The Needleman-wunsch algorithm is implemented on hadoop framework:

1. Read the sequence files as input
2. Build score matrix
3. Traceback method

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases adenine, guanine, cytosine, and thymine—in a strand of DNA. The two sequence files such as query file(HIV affected sequence) and the non-affected DNA sequence [2-3] are given as input to the mapper where we pass offset as key and value is line of file and file tag is also passed to verify the file. Mapper reads each line as input. The reducer() reads aline and align the

sequences using Needleman-wunsch algorithm. Finally, the aligned sequences are obtained from reducer.

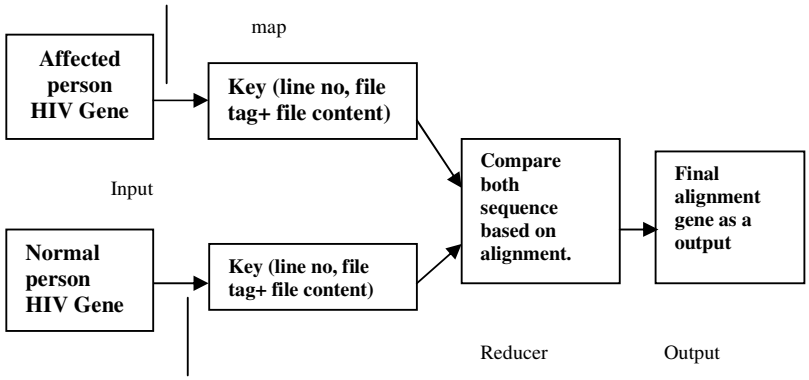

Figure1.HIV gene sequence alignment

The symptoms percentage calculated from user data and the similarity percentage of gene sequences are added and the result is the percentage of a person affected by HIV.

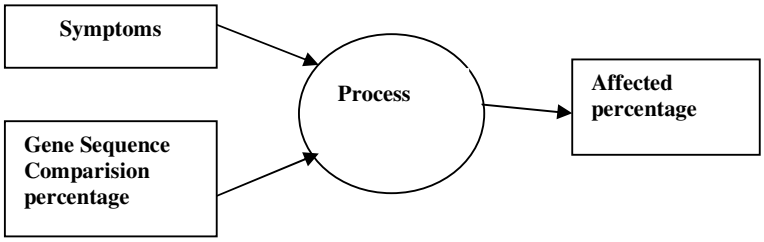

Figure 2. HIV Nomogram workflow

This system uses the symptoms and HIV DNA sequence comparison result to identify the possibility percentage of AIDS. Here, most common symptoms already mentioned are taken as input. After the alignment received from reducer, it finds the similarity percentage of two sequences. Finally, add the similarity percentage of two sequences [6-7] and result of symptoms percentage to predict the possibility percentage of AIDS in a person.

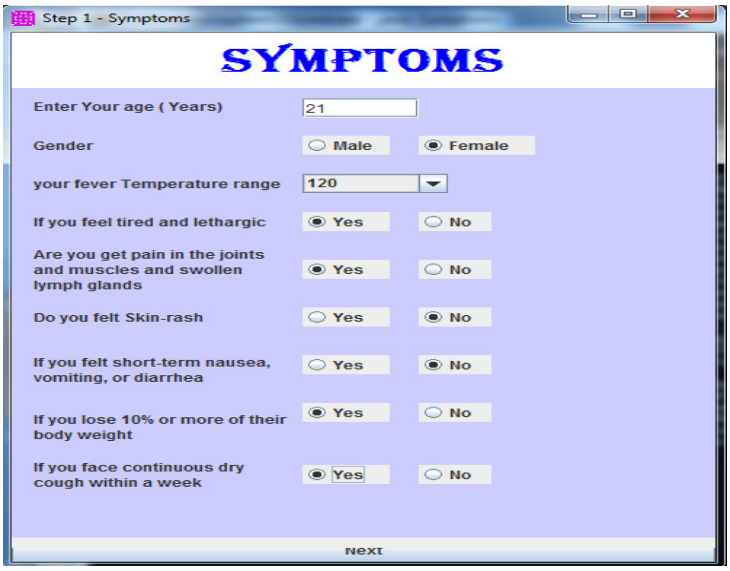

Figure 3.HIV NOMOGRAM Symptoms

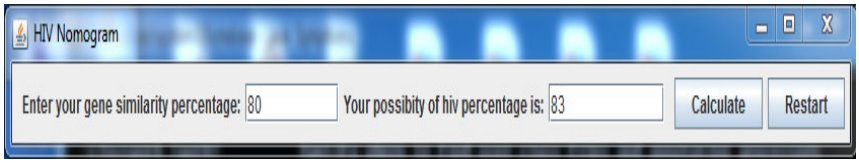

Figure 4.HIV NOMOGRAM Result prediction

4.CONCLUSION

HIV NOMOGRAM USING BIG DATA ANALTICS" helps to find the possibility percentage of HIV in a person. This system uses MapReduce technology in hadoop for DNA sequence alignment. MapReduce framework processes vast amount of data in parallel on large cluster of commodity hardware in a reliable, fault-tolerant manner. This tool is used for detecting HIV/AIDS disease in very effective manner than early approaches. So, it is faster than other existing systems.

5.REFERENCES

[1] Amir H. Payberah,' Introduction to Big Data - SICS', April-8, 2014. [online]:https://www.sics.se/~amir/files/download/dic/introduction.pdf.

[2] Sandrine Dudoit and Robert Gentleman, 'Introduction to Genome Biology', 2003. [online]:cs.mcgill.ca/~blanchem/561/GenBio.pdf.

[3] World Health Organization, ' Early detection of HIV infection in infants and children'. [online]:http://www.who.int/hiv/paediatric/EarlydiagnostictestingforHIVVer_Final_May07.pdf.

[4] A.Hammad, A.Garcia,'Hadoop tutorial', September7, 2011. [online]:https://gridkaschool.scc.kit.edu/.../Hadoop_tutorial-1-Introduction.pdf.

[5] www.health.com,'16 Signs You May Have HIV'. [Online]:http://www.health.com/health/gallery/0,20539037, 00.html.

[6] http://thedownloadandshutup.us,'Hiv dna sequence download', [Dec 16, 2014]. [Online]: http://thedownloadandshutup.us/graphic-design-software/hiv-dna-sequence-download.html.

[7] www.ncbi.nlm.nih.gov, 'Hiv dna sequence download'. [online].http://www.ncbi.nlm.nih.gov/gene/155971.

Authors

S.AVUDAISELVI studying in Ayya Nadar Janaki Ammal College specializing in Computers.

P.TAMIZHCHELVI, Head Department of Computer Science in Ayya Nadar Janaki Ammal College, Sivakasi.