Original Research Paper Computer Science

Bio Analytics - A Big Data Analytics on Disease Identification Using Pathology

R.N.V.Jagan Mohan	R & D Cell Incharge & Professor, Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapuram-534 280. A.P., India.
Y.Vamsidhar	Head & Professor, Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapuram-534 280. A.P., India.
Thota Mohana Lakshmi Tulasi	Research Scholar, Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapuram-534 280. A.P., India.

ABSTRACT

Genome based Disease Identification data is in volumes and also increased exponentially, unrelatedly of whether it is phenotype data namely genetic data. It's derived from various sources and has gigantic diversities is lively and heterogeneous. Therefore, it can be characterized as Big Data. Then, all the aspects of Big Data Analytics can be applied in the field of Bio-Analytics. The author interests with the initiation of pathology laboratory techniques such as high throughput data sequencing and software capable of genome-wide analysis, sequence gaining has become increasingly less expensive and time-consuming. Therefore, it's providing an important helps to science in the form of well-organized in the disease gene identification technique. The experimental results, bio analytics is a big data analytics on clinical pathology of identification of disease detection.

KEYWORDS : : Bio-analytics, Big Data Analytics, Disease Identification, Genome-wide analysis, Pathology etc.

Introduction: In the modern living habits and life styles of the human being, people are suffering with variety of diseases like for which they are habituated to consult the medical practioners. Nowadays, medical practitioners are purely dependents upon variety of clinical tests for detection and identification of diseases. Clinical pathology is a significant component of the causal study of disease and major fields in modern medicine and diagnosis. There are different varieties of pathologies. One can observe them as general medical pathology, anatomical pathology, dermoto pathology, cytopathology, forensic pathology and neuro pathology etc. Because pathology is an incredibly vital part of the medical field, which will grow continuously in the upcoming future. Due to emergence of the new diseases day-by-day, some new improvements in the disease detection, treatment and classification are very much essential. In this direction, we proposed pathologist want to ensure the genome based laboratory testing and disease detection.

In this approach, there are mainly two types gene forecast methods namely **Ab initio and Evidence based.** The first one is **Ab initio** approach rest on signals within the DNA order. It is an automatic process whereby a computer is given directions for discovery the genes order and is then left to find them. The system appearances for communal sequences known to be found at the start and end of genes such as organizer sequences does not known to the gene product. The other is **Evidence-based** approach depend on indication out there the DNA sequence. It includes get-together various pieces of genetic material from the transcript order (mRNA), and known protein orders of the genome. With these fragments of indication it is then possible to get an idea of the original DNA order by working back to front through transcription and translation. For Instance, when you have the protein sequence it is possible to get ready the family of likely DNA sequences it could be derived from by get ready which amino acids to create the protein in which mixture of codons could code for those amino acids and so on, until you get to the DNA order.

Measurement is one of the most useful process in science and in our daily life. When a person become ill then the doctor measures condition. Doctors measurement is depends on several health tests and treatment will be continue to the patient. Therefore health tests plays vital role in patients treatment. The process of health tests should be validated and they should be verified with good reliability. Verification and Validation, is very important for any identification systems. Both are two important topographies of testing process. However, these two expressions appear to have similar meanings, but there exists enormous difference between them, like verification is a process to ensure that the given human trait or phenotype satisfies almost all the identifications which were placed previous to its development, while, validation is a process to ensure that the given detection of human trait or phenotype satisfies the patient requirements. Hence, in order to ensure that the given recognition of diseases in human trait is true in all aspects, and then it has to satisfy the verification and validation process successfully. During the creation of genome based testing algorithm, a development team can employ various Verification and Validation practices to improve the quality of the disease detection. The behavior of data analyst plays a crucial role in the scope of disease detection**.** The developed system should renovate the unstructured data into a structure format and generate the processed data sets, based on these data sets the testing report that will be turned out. The information contains with flags and labels are represented on the huge amounts of data classification system for different time series.

In this paper, the up-to-date Section deals with the introduction in section. The details of proposed work and discussion of unstructured based Patient Diseased Data, which affect its performance, are discussed in section 1. Section 2 deals with the basic algorithm. The section 3 deals with the MapReduce Procedure. The experimental results are highlighted in section 4. Section 5 deals with the conclusion.

1. Unstructured Patient Diseased Data: Disease detection is sample collections of gene i.e., unstructured data. Gene expression, in the classical sense, refers to the production of the concerned trait or phenotype by a gene. The phenotype production, but is based on production of a specific protein encoded by this gene and the subsequent participation of this protein in a metabolic pathway. The sequence of events in phenotype production may, therefore, be written as follows. However, in terms of gene expression has considered to mean transcription that is production of RNA transcript of the gene. In this regard, the term Reliability and Optimization in which plays a

crucial role in the area of disease detection testing. These are derived from the statistical assumptions. It is important for operational, transactional and reliability processes. Reliability is the probability of a human trait is to perform its functions in a specific time. Optimization is to minimize the Number of deceases while executing the process [1].

2. Basic Algorithm: The main objective of work is to develop an algorithm for the precise disease detection using genome-based data.

The following steps in procedure of decease detection is as follows

- Collect the diseases data from social medical field based on patient medical report.
- Perform genome-based test to obtain genome data in the form of cluster.
- Identification of gene, and to compare with existing gene.
- Finally, prepare the relational gene data.

3. Map Reduce Practice: A Genome based patient datasets are prepared by collecting the information from various patients across the country including the social-media networks. This is an un-labeled(unstructured) dataset in which involves each and every person or entity means of applying disease detection technique, we are classifying the patient according to a particular disease to find out the gene dataset. The process of categorizing the set of patients or entities of same type in one place i.e., the process of similar patients for the purpose of cluster having same characteristics from the data information. The analysis follows same procedure to map the data into a category on the basis of input dataset [2, 3, 4, 5, 6, 7, 8 and 9] and compare with structured gene dataset. Basing on the similarities one can go for a correct diagnosis for variety of pathological problems. The following sequence of events in the production of the phenotype by a gene:-

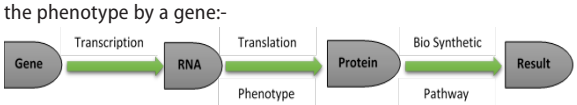

Fig 1: The production of the phenotype by a gene.

The mechanism by which the expression of different gene os controlled in different tissues and/or at different times. Since, we have suppose handle with large data sets with different types of information's the concept and tools of Big Data Analytics is used in the study. In this regard, the mapreduce procedure two methods are involved namely mapper and reducer. Initially, Unstructed based Patient Diseased Data has taken and data sequentially mapping the mapper from several patients and assigns the index. Once taken the data to transform into the Reducer. The reducer identify the patient data as unstructured to compare with original gene then transform the result. The result information store into the HDFS is a distributed file system.

4. Experimental Results: The experimental results is conducted on Unstructed based Big Data processing using Map Reduce technique like disease gene and compare with original gene. Initially, phenotype gene collected from several patients who are believed to have the same genetic disease. Then, their phenotype gene samples are analyzed and partitioned to determine likely regions where the mutation could potentially reside. These techniques are mentioned in above.

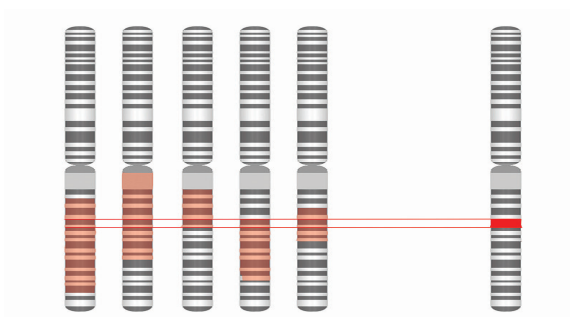

Fig 2: Disease Identification Data Set with Original Data Set

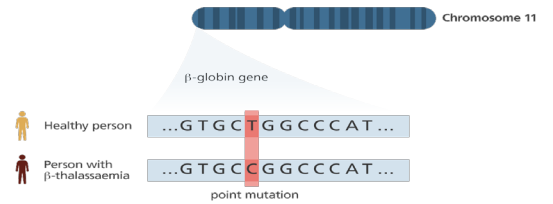

Fig 3: Disease Identification Data Set with Original Data Set

In this regard, phenotype gene on the left show possible disease gene locations as identified by any of the above methods for affected individuals. Red area in the 'composite gene' on the right signifies the overlap of these regions, and this the most probable location of the disease gene. This information can then be used to investigate similarities in the phenotypes of the humans.

5. Conclusion: The Big Data Analytics can be applied in the field of Bio-Analytics. The initiation of pathology laboratory techniques such as high throughput data sequencing and software capable of genome-wide analysis, sequence gaining has become increasingly less expensive and time-consuming. It's providing an important helps to science in the form of well-organized in the disease gene identification technique. The bio-analytics is a big data analytics on clinical pathology of identification of disease detection.

6. References:

1. Baker SJ, Fearon ER, Nigro JM, Hamilton SR, Preisinger AC, Jessup JM, vanTuinen P, Ledbetter DH, Barker DF, Nakamura Y, White R, Vogelstein B (April 1989). "Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas". Science. 244(4901): 217–21. doi:10.1126/science.2649981. PMID 2649981.
2. Jump up, Lee AS, Seo YC, Chang A, Tohari S, Eu KW, Seow-Choen F, McGee JO (September 2000). "Detailed deletion mapping at chromosome 11q23 in colorectal carcinoma". Br. J. Cancer. 83 (6): 750–5. doi:10.1054/bjoc.2000.1366. PMC 2363538 .PMID 10952779.
3. Jump up, Bell R, Herring SM, Gokul N, Monita M, Grove ML, Boerwinkle E, Doris PA (June 2011). "High-resolution identity by descent mapping uncovers the genetic basis for blood pressure differences between spontaneously hypertensive rat lines". Circ Cardiovasc Genet. 4 (3): 223–31. doi:10.1161/CIRCGENETICS.110.958934. PMID 21406686.
4. Jump up, Sherman EA, Strauss KA, Tortorelli S, Bennett MJ, Knerr I, Morton DH, Puffenberger EG (November 2008). "Genetic mapping of glutaric aciduria, type 3, to chromosome 7 and identification of mutations in c7orf10". Am. J. Hum. Genet. 83 (5): 604–9. doi:10.1016/j.ajhg.2008.09.018. PMC 2668038. PMID 18926513.
5. Jump up to: a b Broman KW, Weber JL (December 1999). "Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain". Am. J. Hum. Genet. 65 (6): 1493–500. doi:10.1086/302661. PMC 1288359 .PMID 10577902.
6. Jump up Zeliha Görmez; Burcu Bakir-Gungor; Mahmut Şamil Sağıroğlu (Feb 2014). "HomSI: a homozygous stretch identifier from next-generation sequencing data". Bioinformatics. 30 (3): 445–447. doi:10.1093/bioinformatics/btt686. PMID 24307702.
7. Jump up Luo J, Emanuele MJ, Li D, Creighton CJ, Schlabach MR, Westbrook TF, Wong KK, Elledge SJ (May 2009). "A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene". Cell. 137 (5): 835–48.doi:10.1016/j.cell.2009.05.006. PMC 2768667. PMID 19490893.
8. Jump up to: a b Fortier S, Bilodeau M, Macrae T, Laverdure JP, Azcoitia V, Girard S, Chagraoui J, Ringuette N, Hébert J, Krosl J, Mayotte N, Sauvageau G (2010). "Genome-wide interrogation of Mammalian stem cell fate determinants by nested chromosome deletions". PLoS Genet. 6 (12): e1001241. doi:10.1371/journal.pgen.1001241. PMC 3000362. PMID 21170304.
9. Jump up, Chen WJ, Lin Y, Xiong ZQ, Wei W, Ni W, Tan GH, Guo SL, He J, Chen YF, Zhang QJ, Li HF, Lin Y, Murong SX, Xu J, Wang N, Wu ZY (December 2011). "Exome sequencing identifies truncating mutations in PRRT2 that cause paroxysmal kinesigenic dyskinesia". Nat. Genet. 43 (12): 1252–5. doi:10.1038/ng.1008. PMID 22101681.