IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 5, Issue 2, April - May, 2017
ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

Identification Of Disease Target Protein In Pancreatic And Prostate Cancer Using Big Data

Kangan Arora
Department of Computer Science and Engineering, Vellore Institute of Technology
Vellore, Tamilnadu-623014

Abstract— The genomic profile of the tissue causing cancer reveals both the possible treatments and underlying pathological mechanisms. For analysing the cancer part, need to create working model for diagnosis for cancer patients by analyzing the intermediate relationships of outcome-associated genes using the Hadoop-based concordance index tool kit which we implemented for this project. Our technique is based totally at the concordance index for information, which is non-parametric degree to quantify the energy of prediction rule. Specifically, we have implemented smoothed model of the concordance index toolkit. With the identified protein/genes, we will validate the genomic dataset involved in multiple parts of networks representing cancer marks and thereby visualizing Bayesian regulation network. Therefore, the main purpose is to identify target genes/protein using various genomic cancer types, algorithms and data types.

Keywords— MapReduce; Cancer; Treatment; Concordance Index; Network Analysis.

I. Introduction

Cancer has been considered as one of the most life-threatening diseases. Generally, it is being caused by abnormal genomic interactions and genetic mutation in the cells. [1] The abnormalities caused by cancer lead to uncontrollable growth and invasivess. Typical traits of cancer are proposed by Weinberg and Hanahan as Hallmarks of Cancer [2].

With “Big Data of cancer”, we are being able to inspect the relationships between gene expressions and the response to the treatment. Understanding their interrelationships may help us making treatment plan for complicated profiles.

Staggering amount 94 percent of pancreatic cancer patients die within five years of diagnosis and an approximate 74 percent of patients die just within the first year. Various factors do contribute to the low survival rate. It’s not only especially aggressive, but indeed also difficult to reveal the early and once detected, poorly susceptible to treatment. It’s a perfect mix of factors that make it one of the more deadly forms of cancer.

Nearly an estimate of about 40,000 people do die from pancreatic cancer every year. This disease is referred as twelfth common cancer with about 7.2% of people with disease surviving till five years after treatment and diagnosis, as said according to National Institutes of Health. It is also known as silent killer. One that has, in recent years, taken the lives of some of America’s most beloved celebrities including former Apple (AAPL) CEO Steve Jobs, and actor Patrick Swayze. It is a type of cancer which goes virtually undiscovered until it is in advanced stages, leaving cancer patient with less options to live. It is difficult to surpass this type of cancer as second most aggressive and deadly cancer in United States, as per analysis from Pancreatic Cancer Network, advertised in American Association for Cancer Research’s Cancer Research journal.

Prostate cancer (PC) serves as second most leading cause of deaths in Western world. Prostate Cancer can be diagnosed on basis of heightened levels of protein serum namely prostate specific antigen together with rectal digital examination and is accepted by prostate needle biopsies. However, biopsies and PSA often fail to differentiate between both aggressive forms and clinically indolent, thus leading to a treatment like irradiation and unnecessary prostactectomies which greatly degrades patient’s life quality.

One of the tools for biomedical data analysis that was not implemented in Apache Hadoop ecosystem, which is the family the concordance index. The concordance index is a non-parametric similarity metric for the censored data. It requires massive computation for counting the number of concordant pairs among all possible combinations between the two numbers in the input.

In this project, we will identify disease protein and genes associated with outcome of patients dealing with by creating toolkit based on Apache Hadoop ecosystem. We will further create Bayesian network of the genes and will validate the genes involved in the hallmarks of cancer with performance of model. We will implement web service for deep analysis of cancer and treatment suggestion based on assumption that assume optimal treatment strategy which can be identified by sorting outcomes among patients along with same genomic profiles.

II. RELATED WORKS

Due to the inherent complexity of the reason, that's ordinary interactions between genes, currently cancers are labeled into multiple subtypes [3] and numerous treatment strategies had been developed [2]. Though the goods for precise cancer kinds have been permitted [4, 5], a way to connect the cancer subtypes to the surest remedy remains tough [6].

The PhD researchers have amassed data associated with molecules that has inspired the activity of genes. Authors have construct a database of genetic facts from blood samples, thus predicts how gene is being tormented by cancer remedies and have advanced computing models to apprehend why a couple of cancers occur together. They help us in figuring out most cancers, fit the remedy to right man and recognize how most cancers cells broaden, delivering our method of better prognosis, higher remedy and better prevention. The predominant purpose is of unlocking energy of microRNAs to extricate combative cancers from among low-danger ones and also tell us which one is likely to be taken into consideration exceptional for man's most cancers.

The researchers have taken into consideration a gross amount of knowledge available in entirety of genetic cloth. They help in locating out which genes are on or off in respective cellular kind at given time, what's controlling those genes whether or not they are active or now not which bits of DNA mess with bits of DNA and protein, whether or no longer to listing the mutations of suitable genes that have an effect on how they work and consequently list how the mobile behaves. The main goal is to workout with what genetic changes appear as most cancers develops. A big statistics is being accumulated, helping in operating out on which patients can be suitable for which remedy and attempting it clinically based totally on the facts genetically to be had specially most cancers. The database made built by means of authors may be useful in designing prostate most cancers drugs due to size of database accordingly making it feasible to have a look at a populace level.

Prostate cancer has been considered as 2d most most cancers not unusual among guys and is fourth maximum commonplace usual. It is the maximum recognized most cancers in men which accounts for one-zone of new instances diagnosed in line with annum. Radiotherapy has been typical as first-line treatment attributing to high prices of controlling it regionally. Authors have studied about techniques which offer multiplied conformity with overall dose at the same time as destructive regular structures. They have studied superior treatment equipment which generate extra quantities of information than modern opposite numbers. They are speaking in phrases of evaluation which can render modeling of plans.

Prostate most cancers is referred to as heterogeneous disorder at several tiers and in spite of technical improvements, still there is a threat of most cancers recurrence after therapy.

Pancreatic most cancers has baffled researchers. Most drug treatments thrown at the disorder have supplied marginal enhancements at quality. New studies from the University of Glasgow published inside the journal Nature gives new desire to pancreatic most cancers patients.

The researchers say they've located that there are four exceptional subtypes of pancreatic cancer: squamous, pancreatic progenitor, immunogenic, and aberrantly differentiated endocrine exocrine, or ADEX. These subtypes describe the differentiating factors among them, and being able to understand them offers a risk to make remedies more centered to the weaknesses of each patient's subtype. In short, the potential to apprehend those characteristics of a most cancers makes the sickness extra treatable.

However, the step forward of excessive-throughput genomic profiling generation makes big cancer genome records profiling possible. One of the biggest Pan-Cancer information set is curated and maintained by way of The Cancer Genome Atlas (TCGA) [7]. The statistics set incorporates extra than 18,000 genome profiles with complete scientific records consisting of survival facts, treatment file, and medical phenotypes.

III. SYSTEM OVERVIEW

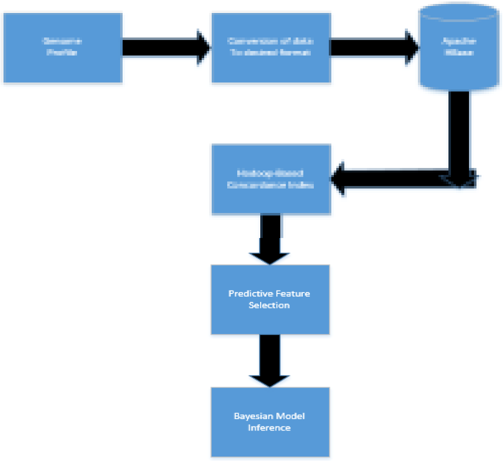
In the following sections, we will elaborate on the data set we used and the methods we implemented for this project.

A. Methods

Summarizing the methodology as follows:

Analysis and modeling have been performed on the pancreatic and prostate data set using the programs created for Hbase, Hadoop, and R. Also analysis in network and modelling on genomics dataset. The delineaton of these components (above) gives us more stable and refined picture of major driving events in elucidation of cancer development and progression.

Further, we will implement the modules and algorithms using Java, Hadoop, R, Pig and JavaScript evaluating the performance on virtual machine.


Genomic Profile
Conversion of data To desired format
Apache HBase
Hadoop-Based Concordance Index
Predictive Feature Selection
Bayesian Model Inference

Architecture design- figure 1

The genomic dataset has been first converted into desired particular data format i.e after running “ci_preprocess.R” , we have compiled the java code and converted the CSV files into sequential files. The second step is to upload the generated sequential files to HDFS engine. Then we will compute the concordance index of single gene using the latter algorithm. After that computation of multiple genes is being carried out by using the Map-reduce concordance index algorithm. Then, from here a specific set of features have been selected from the above carried out steps. The last step is to identify genes associated outcome using Bayesian Inference Model by installing “bnlearn packages” in R thus resulting in a graph/model on the console window.

B. Algorithm

Map-Reduce Concordance Index Algorithm

Map-reduce concordance index function C has been implemented via mapping combinations of the predictions into pairs and checking if they're true/valid pairs and concordant with the response: If one pair’s order is concordant to the particular response when patient with the shorter survival time is deceased, add 1 to n_c. Otherwise if the order isn't concordant to the response, add 1 to n_d. The subject statuses are definitive as defined in Kendall’s τ. Then in reducing part, the two pairs with groups are gathered and the concordance index is computed as formula (1).

$$C = \frac{n_c}{n_c+n_d} \qquad (1)$$

Because the survival time is always integer and the reputation is binary (deceased), we hashed the survival facts into a floating quantity, then we looked after the survival information with regarded to the cost of the characteristic-of-interest. Therefore, for each pair of numbers of the taken care of input, the concordance checking subroutine calls for most effective one iteration, which reduces no longer best the complexity on area but on time.

Varieties of MapReduce–primarily based concordance index program have been applied for validation. The algorithm we described above is described as part A. We have carried out part B for validation. In Algorithm B, the concordance index assessment was accomplished with the aid of mapper in place of reducer. The reducer will collect only the last concordance values of index.

Concordance index is basically defined as global index for validating predictive ability of various survival models. It is actually the pairs of fraction in dataset taken, in which the observation with high survival time has more probability of survival as being predicted by working model of the project.

Using the concordance index algorithm, thus we demonstrated the most survival-associated genes out of large data set of features on the GENOME data set.

C. Inference, Validation and Visualization of the Networks

The elucidation above thereby infers the Bayesian community of genes in opposition to using the bnlearn package deal in R. Default parameters have been used. In addition to this, we have visualized the inferred Bayesian Network the usage of D3.Js.

Probabilistic models are based on directed acyclic graphs (DAG) having rich and long tradition of beginning with the work of geneticist Sewall Wright. Many variants have been appeared in various fields. But amongst within statistics, such models are called as directed graphical models; and within artificial intelligence and cognitive science, these models are called as Bayesian networks.

These networks are considered as type of Probabilistic Graphical Model which are used to construct models from data and/or opinion of experts. They are commonly called as **Bayes nets**, **Belief networks** and **Causal networks**.

Bayesian networks are more general that dealing stricty with Inputs and Outputs. This is because any variable in the graph can be an input or output or even both. We could even predict the joint probability of an Output and a missing Input.

The subnetworks of genes enriched have proven the usage of the publicly gene ontology evaluation tool Gene Ontology Consortium Enrichment Analysis.

D. HBase-Hadoop Interface

Because PIG and the Hbase shell do not provide the characteristic for *SCHEMA FOR BIG MATRIX, SO WE CARRIED OUT java class for IMPORT THE GENOMIC DATA WITH FIXED NUMBER OF COLUMNS.* Apache HBase is known as an open source resource as a NoSQL database which provides a high real-time read/write access to particular large datasets.

Apache Hbase is known for providing random, a real time access to one’s data in Hadoop. It was being created for hosting huge tables thus making itself a great choice for storing sparse or multi-structured. Hbase can be queried by users for a particular point of time.

HBase linearly scales data to have primary key by having all tables. The space has been divided into sequential blocks which are allotted to region. Hbase can further subdivide the space automatically if keys are accessed frequently within region, thus making data sharding unnecessary for manual works. Thus in this paper, PIG function has been applied for accessing Hbase formatted output and storage.

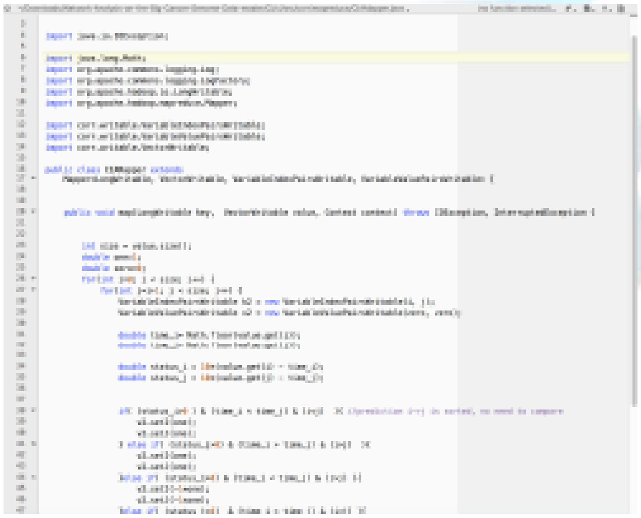
IV. SOFTWARE PACKAGE DESCRIPTION

The package incorporates the toolkit for evaluating two forms of MapReduce-based totally concordance index algorithms in Hadoop environment and the HBase equipment for information uploading and exporting data, as well as the R script files for feature analysis and modelling , preprocessing and Bayesian inference.

V. EXPERIMENTAL RESULTS

A. The Toolkit for Data Analysis

Map reduce algorithm has been implemented that consists of concordance index function and their respective importing, exporting and preprocessing tools. As described above, by hashing and sorting input, we have decreased the size of input since the input of the algorithm is taken as a floating array instead of three-column based matrix. Using traditional method, we have evaluated features using the gene values.



```
import java.io.IOException;

import java.lang.Math;
import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.mapreduce.Mapper;

import [illegible]
import [illegible]
import [illegible]

public class CIMapper extends
    Mapper<LongWritable, VectorWritable, VariableIndexPairsWritable, VariableValuePairsWritable> {

    public void map(LongWritable key, VectorWritable value, Context context) throws IOException, InterruptedException {

        int size = value.size();
        double one=1;
        double zero=0;
        for(int i=0; i < size; i++) {
            for(int j=i+1; j < size; j++) {
                VariableIndexPairsWritable k2 = new VariableIndexPairsWritable(i, j);
                VariableValuePairsWritable v2 = new VariableValuePairsWritable(zero, zero);

                double time_i= Math.floor(value.get(i));
                double time_j= Math.floor(value.get(j));

                double status_i = 10*(value.get(i) - time_i);
                double status_j = 10*(value.get(j) - time_j);

                if( (status_i==1 ) & (time_i < time_j) & (i<j) ) // [illegible]
                    v2.set1(one);
                    v2.set2(one);
                } else if( (status_j==1) & (time_j < time_i) & (i<j) ) {
                    v2.set1(one);
                    v2.set2(one);
                }else if( (status_i==1) & (time_i < time_j) & (i>j) ){
                    v2.set1(-1*one);
                    v2.set2(-1*one);
                }else if( (status_j==1) & (time_j < time_i) & (i>j) ){
```

B. Bayesian Inference, Validation and Visualization of the networks

The Bayesian network for genes is reconstructed and the three phenotypes were visualized using D3.js.

The analysis in network enrichment shows that the genes which we identified are being involved in networks relating to hallmarks of cancer, which are interleukin-4 secretion, embryonic morphogenesis and tissue development. The result obtained hereby validates the outcome-associated genes which are involved in pathways of hallmarks of cancer.

C. HBase-Hadoop Interface

The Base-Hadoop interface has been implemented using Pig for exporting data and Java for importing. However, importing into Hbase, one gene takes about few less seconds than that of exporting data into Hbase. Difference in running times may indicate certainly Pig as not perfect choice for manipulating big genome data on distributed system, although it may perform better in operations with HBase.

The main aim is to implement various genomic profiles along with platforms into analysis pipeline so to improve their accuracy.

On a more technical front, we have attempted to implement Apache Thrift in order to connect it with Hadoop, Hbase.

VI. CONCLUSION

Here in this paper, by implementing the Hadoop based concordance index device kit, associated genes have been recognized in the results, by using information set gathered from over more than one of most cancers types.

Also using Bayesian network, as a graphical model that incorporates probabilistic relationships amongst variables of interest.

When it is used in conjunction with Hadoop Hbase, the model has more advantages for analysing data. Some advantages amongst many are:

Among all variables, the model encodes dependencies. It handles situations wherein some data entries/values are missing/skipped. Secondly, Bayesian network can be used in learning causal relationships and hence it can be made to use in gaining understanding about particular problem domain and in predicting the consequences of given problem set.

Apache HBase provides its high availability in many interesting forms such as listed below:

Firstly, High amount of available cluster topology information through production deployments with multiple instances.

Secondly, Data distribution across various number of nodes meaning that loss of single node will affect only data stored on that particular node.

Thirdly, HBase HA allows data storage, thus ensuring loss of single node so as not resulting in loss of data availability. Format of HFile stores data directly in HDFS. HFile can be read or written onto by Apache Pig, Apache Hive and MapReduce.

Therefore, Set of genes have been demonstrated , contained in multiple parts of networks thus considering the cancer hallmarks thereby inferring and visualizing the Bayesian networks community.

VII. References

[1] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, "The cancer genome atlas pan-cancer analysis project," Nature genetics, Sept. 2013, Vol. 45, pp. 1113-1120, doi:10.1038/ng.2764.

[2] L. J. van't Veer, H. Dai, M. J. Van De Vijver , Y. D. He, A. A. Hart, M. Mao, "Gene expression profiling predicts clinical outcome of breast cancer," Nature, Jan. 2002, Vol. 415, pp. 530-536, doi:10.1038/415530a.

[3] N. I. Simonds, M. J. Khoury, S. D. Schully, K. Armstrong, W. F. Cohn, D. A. Fenstermacher, "Comparative effectiveness research in cancer genomics and precision medicine: current landscape and future prospects," J. Nat. Cancer Insti., Mar. 2013, djt108, doi: 10.1093/jnci/djt108

[4] M. J. Van De Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, "A gene-expression signature as a predictor of survival in breast cancer," New Eng. J. Med., Dec. 2002, Vol. 347, pp. 1999-2009, doi: 10.1056/NEJMoa021967.

[5] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," Cell, Mar. 2011, Vol. 144, pp. 646-674, doi:10.1016/j.cell.2011.02.013.

[6] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, and J. P. Mesirov, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, Vol. 286, Oct 1999, pp. 531-537, doi: 10.1126/science.286.5439.531.

[7] American Cancer Society, "Cancer Facts & Figures 2014," Atlanta: American Cancer Society, 2014, pp. 1-2.