An Approach in Big Data Analytics Framework for Analysing Huge Gene Transcription Data

Rohit Roy, S. P. Syed Ibrahim

Abstract*: Gene Co-expression network analysis is increasingly used to explore system level functionality. In order to study the complexity within gene interactions and identify a target gene for clinical application the researchers apply co expression network based on correlation coefficient. Significance of weighted co expression network analysis is that it reduces the high dimension of the data and integrity of multi scale dataset and also identifies the hidden interactions among the genes. Construction of co expression network on a large sample size would improve the accuracy and robustness but statistical and computational methods applied for screening of multidimensional data are both space and time consuming. In present, the researchers are at the verge of acquiring a new methodology that analyzes huge data in short time period. However big data analytics method for analyzing gene co expression network is in infant stage. Our objective is to identify an approach using big data analytics framework that can enable scientific research community to process large scale data and support them in identifying clinically significant targets.

Index Terms*: Co-Expression, Correlation, Weighted Network, Network Analysis.

I. INTRODUCTION

Gene Co-expression Network is a directed graph with each node representing a gene and there is an edge, if there exists any relation between those genes. Significance of this is that it reduces the high dimension of the data and integrity of multi scale dataset. This identifies the hidden interactions among the genes. Researchers can focus on various gene modules as opposed to individual genes. In this paper we are going to work with cancer related dataset. Cancer is a vicious disease and therefore it is important to understand the gene patterns and interactions in its samples. So that significant medical inferences could be made out of the analysis of those patterns and interactions. Co expression is mathematical treatment which enables us to practice this analysis. Co expression has been executed in R, a free and open source factual programming dialect that is generally utilized, kept up and enhanced by a dynamic group. It was produced to all the more productively break down micro array datasets by measuring not just the relationships between individual sets of qualities, yet additionally the degree to which these qualities share similar neighbors. Network based strategies have been discovered valuable in numerous spaces. Co expression network analysis is important and significant because it reduces high dimensionality of the data. In this paper, we will aim for gene co-expression network based for transcriptions response of genes to changing challenges like increasing the data size along with a checking on accuracy and time efficiency.

Revised Manuscript Received on May 22, 2019.
Rohit Roy, SCSE, VIT University, Chennai, India.
Dr. S P Syed Ibrahim, SCSE, VIT University, Chennai, India.

II. LITERATURE SURVEY

After reading various research papers from various sources and journals, we have finally gathered some knowledge, by analyzing various problems related to this field of interest and how researchers have adapted different techniques and solutions to handle those problems. Few of the important papers among the survey made so far are mentioned below with the challenges and methodology of respective papers explained briefly.

Matthew V. DiLeo, Gary D. Strahan, Meghan den Bakker and Owen A. Hoekenga have proposed to perform analysis of co expression gene of tomato fruit metabolome, we learnt that the challenges they faced were dimensionality reduction in the process to cleanse the input dataset. To encounter this problem they used PCA and BL-SOM methods to overcome the challenges.

Ignacio Riquelme Medina, Zelmina Lubovac-Pilav have proposed that network based examination gives more elevated amount associations amongst cells and their contribution in various pathways, which is a decent beginning stage for exploring complex illnesses, for example, Type 1 Diabetes.

Hisham Abdel Latif Albukhaiti and Jiawei Luo proposed that in their investigation, it expects to dissect the distinction of non-straight quality co-articulation arrange between tumor related network and typical related network. We initially acquaint the TMM technique with standardize the quality articulation to decrease the predisposition of tests. At that point, the fundamentally differentially communicated qualities which are recognized with the DESeq technique are utilized to develop the co-articulation arrange. To secure the non-direct nature of the organic framework, we utilize the dCor, which is effective by and large, to gauge the reliance amongst qualities and supplant the first relationship measure in WGCNA to build the non-straight weighted quality co-articulation organize for both the gastric disease and typical specimens. To discover intrigued qualities in all the differentially communicated qualities, we recognize the modules in the two systems individually and find just a single module in each system. The qualities which are in the module of ordinary affiliation yet not in the module of disease affiliation arrange are chosen for facilitate investigation and call these qualities stray qualities for simple reference.

III. METHODOLOGY

Our objective is to pull out significant patterns in genes and understand system instead of reporting individual parts. Our focus should mainly be on clusters and not on individual genes.

The working of Gene Co expression analysis is that it will construct a network using the patterns and interactions among the genes. After that it will identify those modules, so that module based analysis is possible, then it is going to relate that module with some micro dataset which is done to find the biologically interesting modules. After completing that, we study the module preservation of our data. This is to check the strength (strong or weak) of the module. For example, if we have a lung cancer module in a dataset, we will find out which module could be found in another lung cancer dataset. After finding the interested modules, we find the key drivers in those modules. This means, to check whether these modules contain certain transcript, or any particular identifying gene in these modules.

A. Existing System

The data set after being taken as input undergoes through a process to calculate the similarity among the genes. This is done by calculating Pearson correlation coefficient among those genes. This shows the level of consistency between the gene expressions. After that this correlation matrix thus formed is transformed into a adjacency matrix with an aim to create network modules. Finally when the network modules are identified, we apply big data technologies to analysis those network. In order to implement the above environment, we use R scripting. Since, WGCNA is already having R packages, it becomes easier to implement. Also, R provides with various libraries which could be used to analysis various clinical significances. Formula for person correlation is;

$$a_{ij} = | cor(x_i, x_j) |^{\beta}$$

Where i and j are two different nodes, a is the similarity factor and x and j are weights of the respective genes. Beta is often taken as 6 as a constant. The diagrammatic view of the existing methodology is given in **fig (a)**.

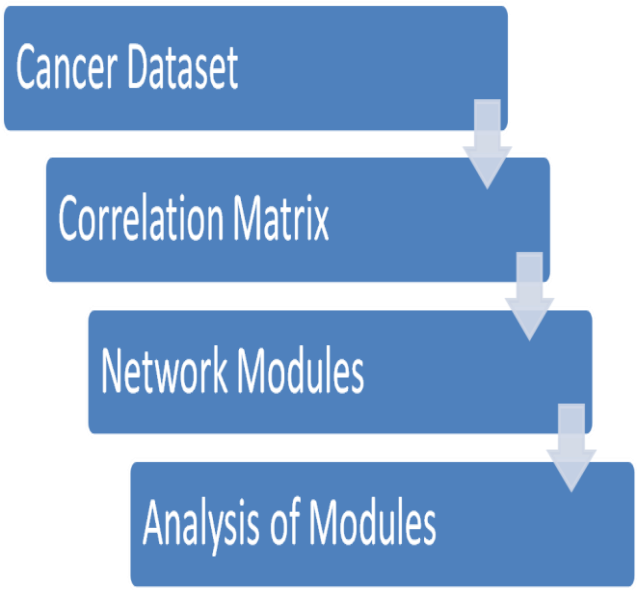

Fig (a)

B. Challenges

This method can process a limited amount of data only of 500 to 1000 rows and columns of dataset at maximum. Also, Statistical and computational methods applied for screening of multidimensional data are both space and time consuming.

C. Proposed System

To identify a method using big data analytics framework that can enable scientific research community to process large scale data and support them in identifying clinically significant targets. In order to increase the efficiency and accuracy of the system, we are going to increase the input data size so that better and accurate inference can be made. After taking huge cancer data set, we need to clean the data and perform dimensionality reduction techniques on that data set. After the data is clean, we need to construct graph based network, so that our analysis can be made faster and less time consuming. Further to improve the efficiency, we will optimize those graphs as well. . The diagrammatic view of the existing methodology is given in **fig (b)**.

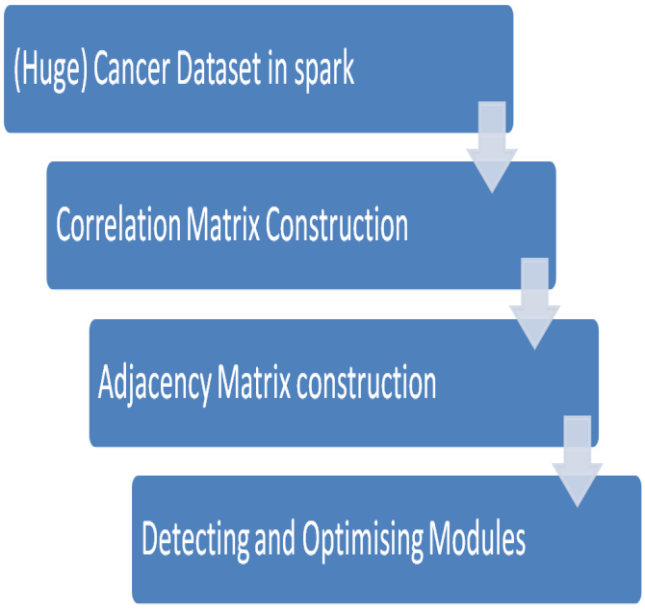

Fig (b)

D. Huge Cancer Dataset in Spark

In this paper, cancer related data set is being used. The dataset is in tabular format of a csv/excel file, with the rows depicting the various number of genes and the column depicting various number of samples of those genes. Sample data and R software tutorials: http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork. We overcome the challenge of limited dataset by storing the dataset into big data framework such as spark, and after that apply operations over the dataset by integrating spark with R. SparkR provides the combined features of R packages and parallel processing environment of spark.

E. Correlation Matrix

After reading the data into **spark**, we perform dimensionality reduction by first taking the data into a spark data frame and calculating the similarity among the data that consists of co expressions values of cancer affected genes. While calculating, we find the correlation between one observations with all the remaining observations. A sample of the correlation calculated is shown in **fig (c)**.

```
File  Edit  View  Search  Terminal  Help
V3     -1.875712e-01 -6.798927e-02  1.588922e-02 -9.284842e-02  7.752864e-02
V4      2.192527e-01  3.374738e-01  3.620092e-01  2.232779e-01  2.563157e-01
              V23311        V23312        V23313        V23314        V23315
V1      1.463688e-01 -7.510026e-02  1.399422e-01  1.614030e-02  1.274163e-01
V2     -1.047545e-01 -3.340400e-02 -3.244113e-01 -7.153019e-02  9.919312e-03
V3     -1.459670e-01  5.749140e-03 -9.487075e-02 -2.515648e-02 -6.404322e-02
V4      3.613919e-01  7.892729e-03  2.779376e-01  2.977277e-01  2.746770e-01
              V23316        V23317        V23318        V23319        V23320
V1      9.353486e-03 -2.184617e-01  1.925046e-01 -1.493895e-02  1.518988e-01
V2     -9.857384e-03 -2.203987e-01 -4.867989e-02  6.527380e-02  2.151412e-01
V3     -1.051718e-01  2.758670e-01 -1.163739e-01 -5.561289e-02 -5.661252e-02
V4      2.400366e-01 -1.425345e-01  2.878351e-01 -1.131369e-01  1.759234e-01
              V23321        V23322        V23323        V23324        V23325
V1      2.692682e-02 -6.066075e-02  2.745229e-02  7.900262e-03  1.169924e-01
V2     -1.321950e-01  4.726857e-02 -4.918346e-02  3.972292e-02 -1.446467e-01
V3     -1.577759e-01 -1.989502e-02 -9.673080e-02 -1.127494e-01 -1.030722e-01
V4      1.036298e-02  1.799221e-01  1.092421e-01 -3.102178e-02  3.598932e-01
              V23326        V23327        V23328        V23329        V23330
V1      1.656060e-02  5.208272e-02  7.383889e-02 -1.343666e-01  4.728268e-02
V2     -2.879049e-02 -6.506106e-02 -6.450225e-02  4.457644e-01  3.698610e-02
V3     -9.583561e-02 -1.394464e-02 -5.208711e-02 -3.499301e-03 -4.797418e-02
V4      3.220678e-01  3.873926e-01  3.926545e-01 -1.741927e-01  2.115196e-01
              V23331        V23332        V23333        V23334        V23335
V1      1.337502e-01 -9.317904e-02  1.389655e-02  4.652749e-02  8.273692e-02
V2     -3.221314e-02  1.043063e-01 -1.129558e-01 -3.520621e-02 -1.735514e-02
V3     -6.543391e-03  7.321116e-02  4.998407e-02 -1.682864e-01 -7.136723e-04
V4      3.907544e-01 -2.105053e-01  2.966151e-01  1.134333e-02  3.170987e-01
              V23336        V23337        V23338        V23339        V23340
V1      1.453126e-01  8.869397e-03  1.699555e-01 -1.520103e-02  3.189069e-01
V2     -7.135148e-02  3.328800e-01 -1.446395e-01  3.177488e-01 -1.102095e-01
V3     -3.413037e-01  7.233810e-02 -9.401686e-02 -2.144151e-02 -4.509861e-03
V4      1.857067e-01  1.915620e-01  4.285959e-01  2.822068e-01  1.021978e-01
              V23341        V23342        V23343        V23344        V23345
V1     -1.844389e-01 -9.818421e-02  1.985488e-01  1.027754e-01 -8.808000e-03
V2      2.909885e-01  4.159044e-01  1.903357e-01  1.871286e-01  1.648929e-01
V3      3.159634e-02 -5.051099e-02 -8.229840e-03 -1.041483e-01  2.327395e-02
V4     -1.857053e-01 -1.499758e-01  1.231285e-01  1.194860e-01 -2.323867e-02
              V23346        V23347        V23348
V1      9.163144e-02  6.949091e-03  9.992102e-02
V2      3.980625e-01  2.828974e-01 -2.679859e-01
V3     -2.802056e-02 -5.604564e-02 -9.125713e-02
V4      5.891530e-02  7.818843e-02  2.025166e-02
 [ reached getOption("max.print") -- omitted 23344 rows ]
```

Fig (c)

F. Adjacency Matrix

Taking the correlation as the input to Topological Overlap Module function of R, efforts are made to calculate the adjacency among the highly correlated genes. Calculation of the correlation and the adjacency in the existing method takes ample amount of time to throw the result, whereas in the approach of this paper the computational time taken to calculate the above steps was very less. A sample of the adjacency calculated is shown in **fig (d)**.

```
PAX8                                0.056810289 0.0062104066 0.0712822667
                                          ZBED9 LOC102724782      KBTBD12
DDR1                                0.014275749 0.0365010933 0.018325550
RFC2                                0.022834290 0.0503014622 0.006673490
HSPA6                               0.011084608 0.0110143955 0.005576812
PAX8                                0.007622556 0.0413137127 0.016622049
                                    CTD-2118P12.1          TEX36       MTCP1
DDR1                                  0.010733834 0.018381884 0.016380433
RFC2                                  0.020614293 0.011545171 0.019739027
HSPA6                                 0.013850186 0.008486395 0.007622507
PAX8                                  0.042163825 0.029105702 0.012933144
                                        LINC01159 RP11-664D1.1 LOC101928779
DDR1                                0.0303473841 0.0172432711 0.0211894339
RFC2                                0.0061191808 0.0131128649 0.0104356499
HSPA6                               0.0081351378 0.0085369830 0.0143624265
PAX8                                0.0988209233 0.0834338846 0.1114336437
                                          HP08942          NAF1 RP11-945A11.2
DDR1                                0.0239582331 0.0065739157   0.020615959
RFC2                                0.0104731301 0.1426938251   0.019429322
HSPA6                               0.0113352369 0.0152997971   0.011633553
PAX8                                0.1139935606 0.0049555572   0.049408629
                                          CCDC63     AF086126  RP11-495P10.6
DDR1                                0.0331836593 0.0086885902  0.0169739097
RFC2                                0.0128380291 0.0283372257  0.0076118737
HSPA6                               0.0150215050 0.0238743253  0.0209370890
PAX8                                0.1130635666 0.0037836889  0.0742483039
                                    RP11-337C18.10  LOC101929475 RP11-1260E13
DDR1                                  0.020526924  0.025174317   0.03526665
RFC2                                  0.012601639  0.014066937   0.01002139
HSPA6                                 0.005172051  0.015558212   0.00127623
PAX8                                  0.016719055  0.081569895   0.04341938
                                       LOC389641    LOC388210         GALR3
DDR1                                0.0164751629 0.0400715425 0.0142529754
RFC2                                0.0876125545 0.0061194909 0.0818117883
HSPA6                               0.0237580314 0.0086405370 0.0137195782
PAX8                                0.0447218571 0.1328239487 0.0694334131
                                    MINOS1-NBL1        NUS1P3   MROH7-TTC4
DDR1                                0.0822441206 0.0045978958 0.0084047932
RFC2                                0.0077543732 0.0723359572 0.1259473922
HSPA6                               0.0152069389 0.0188312345 0.0114487921
PAX8                                0.0280141443 0.0045552243 0.0058939675
                                          LIME1  LOC102725263       C1orf50
DDR1                                0.0463184911   0.028102343 0.0148172209
RFC2                                0.0444464287   0.043732733 0.0390423577
HSPA6                               0.0148691536   0.008076754 0.0179378971
PAX8                                0.0313614921   0.030756151 0.0135690916
                                       FAM86B1       HNRNPUL2  LOC100505915
DDR1                                0.0264408997 0.0162879007 0.0276687182
RFC2                                0.1166754563 0.0696580958 0.0024040115
HSPA6                               0.0131753601 0.0110541668 0.0087996646
PAX8                                0.0220286238 0.0245464127 0.0176223527
 [ reached getOption("max.print") -- omitted 23344 rows ]
```

Fig (d)

G. Detecting and Optimizing Analyzing Module

The adjacency matrix calculated in the previous step will be saved as a R data file and will be taken to a R IDE for better visualization purpose. In IDE, on the adjacency matrix, agglomerative hierarchical clustering will be applied and by using a cut tree function the required modules will be detected. These modules would still contain certain noise factor which needs to be removed. For this purpose we will apply power transformation on the adjacency matrix and repeat the above process of finding the modules. This time, the modules are observed to have less noise than earlier. The module detected is shown in **fig (e)**.

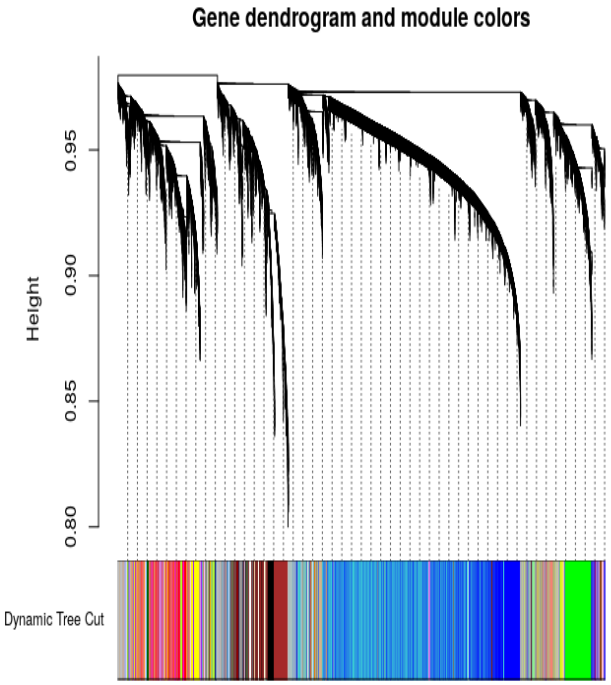

Fig (e)

IV. CONCLUSION

So far WGCNA can process a limited amount of data only. This is because Statistical and computational methods applied for screening of multidimensional data are both space and time consuming. Therefore, our objective is to identify a novel computational method using big data analytics framework that can enable scientific research community to process large scale data and support them in identifying clinically significant targets, keeping in mind the efficiency and accuracy of the system to be increased. We have so far calculated correlation in 3:48 min and adjacency in 1:32 min in spark environment, where as existing system took 10:02 min and 7:59 min respectively.

REFERENCES

1. Hisham Abdel Latif Albukhaiti and Jiawei Luo, “Using differential nonlinear gene co-expression network analysis for identification gastric cancer related genes” PLoS ONE, 2017, 6(10), p.e26683.
2. Sun, C. Yuan, Q. WU, D., Meng, X. And Wang, B.,) Identification of core genes and outcome in gastric cancer using bioinformatics analysis. PLoS ONE. 2017, Sun et al., 2017.
3. Riquelme Medina, I. and Lubovac-Pilav, Z., Gene Co-Expression Network Analysis for Identifying Modules and Functionally Enriched PLOS ONE, 2016, 11(6), p.e0156006.
4. Ruan J, Dean AK, Zhang W., “A general co- expression network-based approach to gene expression analysis. comparison and applications.,” BMC Sys Biol 2010; 4: 8, 2010.
5. Sipko VD, Urmo V, Adriaan van der G, Lude F, Joao PM, “Gene co-expression analysis for functional classification and gene-disease predictions” Bioinform 2017; 1-18.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

6. Lgnacio RM, Zelmina LP. "): Co-expression network analysis for identifying modules and functionally enriched pathways in type 1 diabetes." PLoS One 2016; 1.
7. Langfelder P, Horvath, WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

AUTHORS PROFILE

Rohit Roy Mtech Computer Science with specialization in Big Data Analytics, SCSE, VIT University, Chennai, India

Dr. S P Syed Ibrahim Professor in SCSE, VIT University, Chennai, India