International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, June 2015
ISSN: 2395-3470
www.ijseas.com

MICROARRAY DATA ANALYSIS USING R PROGRAMMING

Nirusha P[1], Sabreen Taj B G[2], Thriveni T G[3], Gurumurthy H[4],

[1]Research Scholar, G M Institute of Technology, Davanagere, India

[2]Research Scholar, G M Institute of Technology, Davanagere, India

[3] Professor, Department of Biotechnology, G M Institute of Technology, Davanagere, India

[4]Professor & HOD of Biotechnology, G M Institute of Technology, Davanagere, India

ABSTRACT

Gene expression profiling provides unprecedented opportunities to study patterns of gene expression regulation, for example, in diseases or developmental processes. Bioinformatics analysis plays an important part of processing the information embedded in large-scale expression profiling studies and for laying the foundation for biological interpretation. Over the past years, numerous tools have emerged for microarray data analysis. One of the most popular platforms is R, an open source and open development software project for the analysis and comprehension of genomic data, based on the R programming language. In this work use R analysis packages to demonstrate the workflow of microarray data analysis for different cancer forms annotation, normalization, expression index calculation, and diagnostic plots to pathway analysis, leading to a meaningful visualization and interpretation of the data which will help in gene product identification which will further help in drug discovery to combat the cancer progressions.

1. INTRODUCTION

Microarray data analysis is becoming an increasingly integral part of biological research. Analysis of cell expression that would have previously taken months to perform can now be carried out in a matter of hours with the use of these miraculous chips. The analysis of gene

expression values is of key importance in bioinformatics. The technique makes it possible to give an initial answer to many important genetic type of questions.

1.1 MICRROARRAYS

A microarray is device that allows for fast and precise analysis of messenger ribonucleic acid (mRNA) directly from a cell. It consists of two parts: the chip and the optical reader. The chip is constructed from a plate of glass to which tens of thousands of cDNA genes are chemically attached in specific locations called *spots*.. The chip is then run through the optical reader which records the location and intensities of the fluorescent tags Hala .M. *et.al*, 2013[1]

1.2 ABOUT R:

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

1.3 CANCER

Cancer is one of the dreadful diseases, which causes a considerable death rate in humans. Cancer is featured by an irregular, unmanageable growth that may demolish and attack neighbouring healthy body tissues or somewhere else in the body. Milena B *et.al*, 2014[2].

2. LITERATURE REVIEW

According to the works of Michael R *et al*, 2009[3] the Gene Expression Omnibus developed by the National Centre for Bioinformatics (NCBI) at the National Institutes of Health is a repository of nearly 140 000 gene expression experiments. Establishes a bridge between GEO and bioconductor. Easy access to GEO data from bioconductor will likely lead to new analyses of GEO data using novel and rigorous statistical and bioinformatics tools.

Facilitating analyses and met analyses of microarray data will increase the efficiency with which biologically important conclusions can be drawn from published genomic data.

The work of Xiaosheng Wang 2012[4] on microarray analysis of ageing-related signatures and their expression in tumors based on a computational biology approach according to which ageing and cancer have been associated with genetic and genomic changes. The identification of common signatures between ageing and cancer can reveal shared molecular mechanisms underlying them. The convergent and divergent mechanisms between ageing and cancer were discussed. This approach provides insights into the biology of ageing and cancer, suggesting the possibility of potential interventions aimed at postponing ageing and preventing cancer.

The work of Jing Hua Zhao and Qihua Tan 2005 [5] described both the motivation and prospects for using R as an integrated environment for genetic data analysis while a formal presentation of R and comparison with R systems might have been given the description has been deliberately kept informal. First it provides the flexible, integrated environment for statistical computing using an object –oriented programming language.

3. AIMS AND OBJECTIVES

- Implementation of R/bioconductor in microarray analysis of different gene expression data sets.
- Analysis of different sets of microarray data of varied cancer types.
- To analyze the cancer type chosen by using two, three or multiple groups.

Determining the gene expression levels and comparing the experimental samples for control v/s the diseased datasets.

4. MATERIALS AND METHODOLOGY

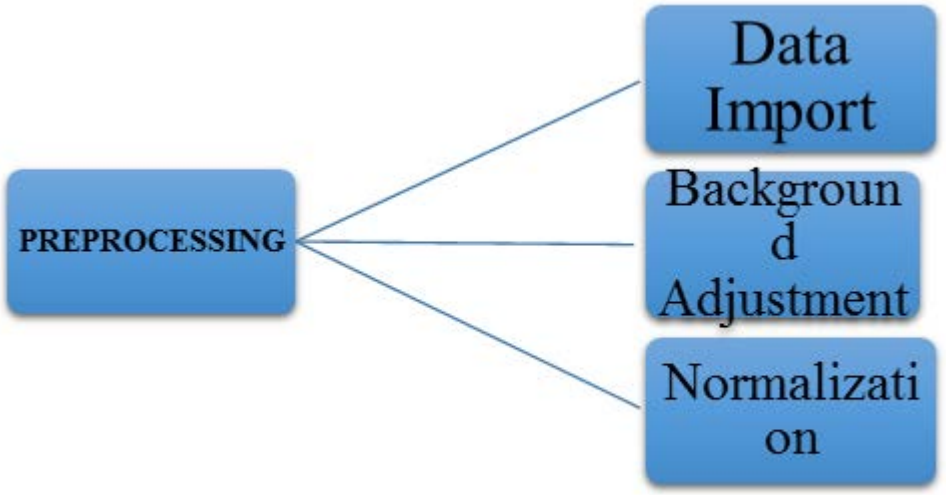


Fig. 1: Flow chart of the work

4.1 NATIONAL CENTRE FOR BIOTECHNOLOGY INFORMATION (NCBI):

The **National Center for Biotechnology Information** (**NCBI**) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health. The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Pepper. The NCBI houses a series of databases relevant to biotechnology and biomedicine.

Selected the different cancer types starting with NCBI which was used to access the data for different cancer types obtained accession numbers through the geo datasets and submitted the accession number to the geo accession viewer

4.2 GEO (Gene expression omnibus):

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

The geo dataset gives the accession number for different data sets the obtained accession number is submitted to GEO accession viewer.

4.3 GEO2R:

GEO2R is an interactive web tool that allows users to compare two or more groups of Samples in a GEO Series in order to identify genes that are differentially expressed across experimental conditions.

The data obtained from GEO accession viewer was analyzed by using geo2r tool .Using the geo2R defined the obtained GSM files into groups which may be 2 or multiple groups.

4.4 DAVID (Annotation tool):

The database for annotation, visualization and integrated discovery. This tool suite mainly provides typical batch annotation and gene-GO term enrichment analysis to highlight the most relevant GO terms associated with the given gene lists. From the tabulated data copy the top 250 gene ID's of particular type, was submitted to the DAVID functional annotation tool.

4.5 BLAST2GO:

Blast2GO is an ALL in ONE bioinformatics solution for functional annotation of (novel) sequences and the analysis of annotation data. Its main function is to assign information about the biological function of gene or protein sequences by making use of diverse public resources like comparison algorithms and databases. The software identifies already characterized similar sequences, and transfers its functional labels to the uncharacterized sequences.

Consider the example analysis of Esophagus Cancer

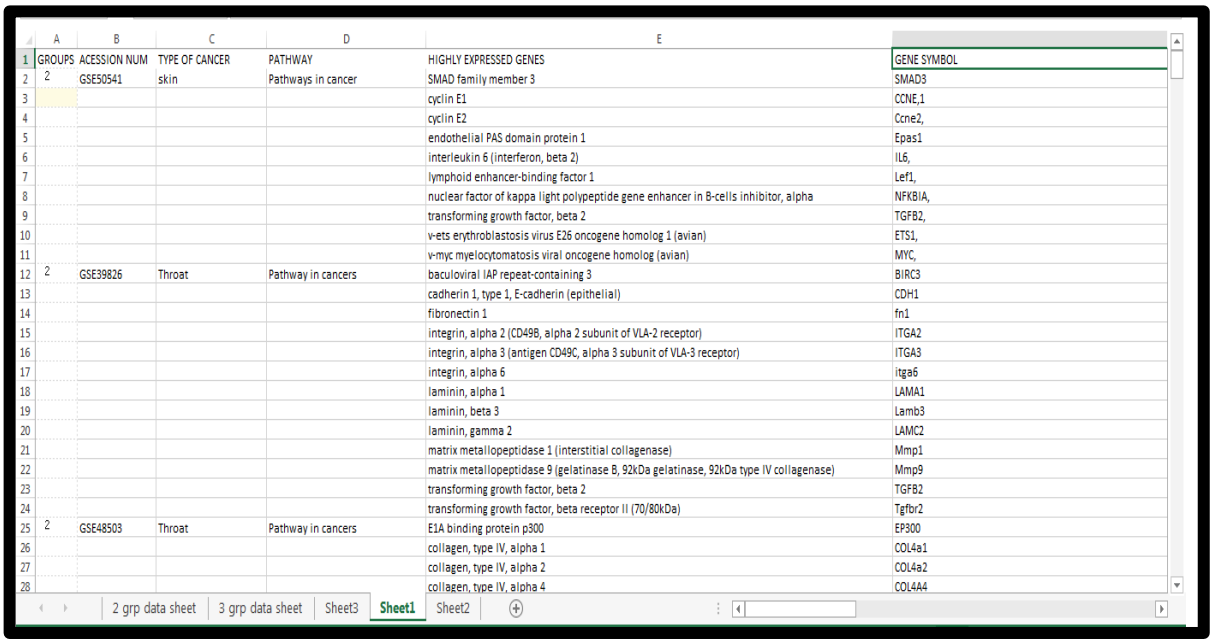
Selecting the particular cancer type to obtain the acession number using NCBI and GEO.Accession Number**: GSE19472** retrieved from GEO accession viewer and obtained complete summary of the cancer type. Analysis with GEO2R and obtaining the GSM files.Submit the obtained accession number in GEO2R tool for GSM files grouping. Defining the groups into 2.GSM files grouping has carried to make two or three data sets. Similarly the

data's for different cancer types with different accession numbers being defined into groups of 2 or more, and the data is characterized. Data for top 250 genes.Analyzed with GEO2R and got the list of top 250 highly expressed genes, created the excel work sheets for the top 250 genes. Copying the gene ID's from the created excel work sheet. Copying of highly expressed top 250 gene ID's from the created excel sheet is carried out in order to submit to the DAVID functional annotation tool. Uploading the gene ID's to the DAVID functional annotation tool. From the tabulated data copied the top 250 gene ID's of particular type, submit the gene ID's to the DAVID functional annotation tool and annotation summary results are obtained.

Pathway responsible for cause of cancer identified. There are different pathways involved in the cancer but we selected only the P53 signalling pathway and pathways in cancer. From the annotation summary results we have selected only P53 signalling pathway by referring the KEGG chart from which highly expressed genes with their official gene symbols were identified. The official gene symbols were submitted to the NCBI tool to get the gene sequences in FASTA format. From the NCBI mRNA and protein sequences were downloaded in the FASTA format for the obtained gene symbols. In this project we concentrating on Homo sapiens so that we have selected Homo sapiens mRNA sequence and protein, and the sequence summary for particular organism.FASTA format sequences for all the genes involved in the pathway are loaded as total sequence into the Blast2GO pro tool. After the sequences are loaded into the Blast2GO pro tool run the blast and interpro, mapping and annotation. The combined graphs for biological process, molecular function and cellular components for the given genes were identified by using graph tool.

GROUPS	ACESSSION NUMBER	ORGANISM	SUBMISSIOM DATE	COUNTRY
2	GSE44327	HOMOSAPIENS	FEB,14 2013	GERMANY
2	GSE36776	HOMOSAPIENS	MAR,23 2012	UK
2	GSE50541	HOMOSAPIENS	SEP,03 2013	AUSTRIA
2	GSE48281	HOMOSAPIENS	JUN,25 2013	TIAWAN
2	GSE37552	HOMOSAPIENS	APR,24 2012	USA
2	GSE31185	HOMOSAPIENS	NOV,03 2011	USA
2	GSE63571	HOMOSAPIENS	NOV,23 2014	USA

Table 1: Collected data for the different cancer types with two group



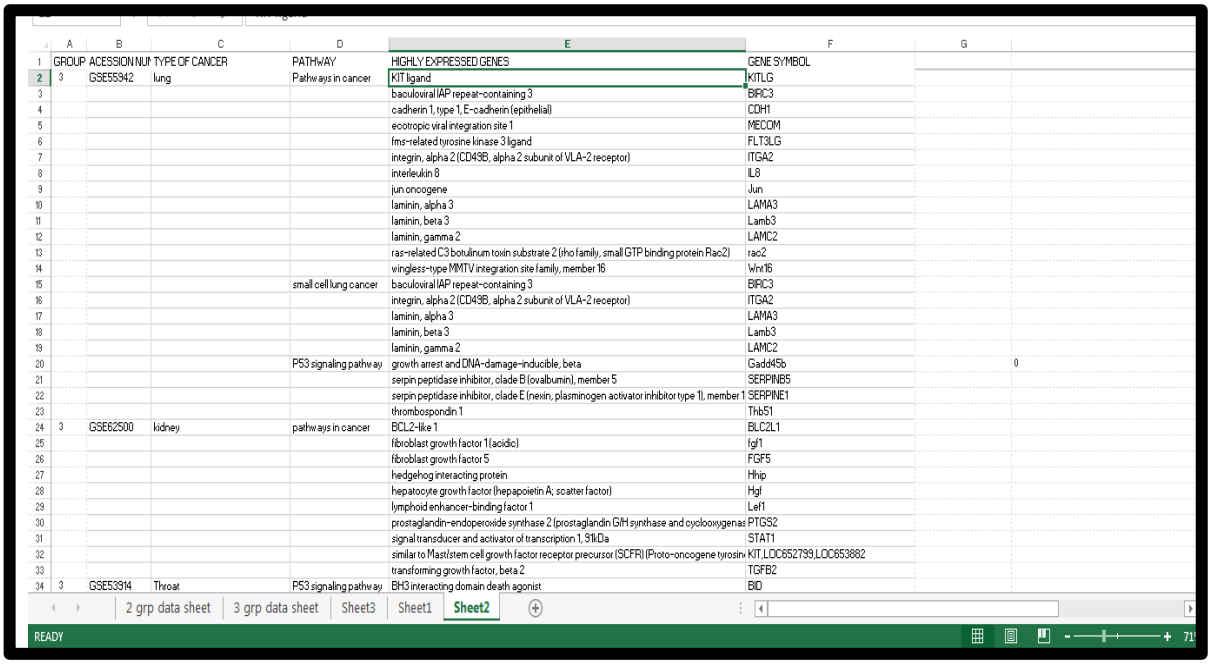
	A	B	C	D	E	
1	GROUPS	ACESSION NUM	TYPE OF CANCER	PATHWAY	HIGHLY EXPRESSED GENES	GENE SYMBOL
2	2	GSE50541	skin	Pathways in cancer	SMAD family member 3	SMAD3
3					cyclin E1	CCNE,1
4					cyclin E2	Ccne2,
5					endothelial PAS domain protein 1	Epas1
6					interleukin 6 (interferon, beta 2)	IL6,
7					lymphoid enhancer-binding factor 1	Lef1,
8					nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	NFKBIA,
9					transforming growth factor, beta 2	TGFB2,
10					v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)	ETS1,
11					v-myc myelocytomatosis viral oncogene homolog (avian)	MYC,
12	2	GSE39826	Throat	Pathway in cancers	baculoviral IAP repeat-containing 3	BIRC3
13					cadherin 1, type 1, E-cadherin (epithelial)	CDH1
14					fibronectin 1	fn1
15					integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)	ITGA2
16					integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)	ITGA3
17					integrin, alpha 6	itga6
18					laminin, alpha 1	LAMA1
19					laminin, beta 3	Lamb3
20					laminin, gamma 2	LAMC2
21					matrix metallopeptidase 1 (interstitial collagenase)	Mmp1
22					matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)	Mmp9
23					transforming growth factor, beta 2	TGFB2
24					transforming growth factor, beta receptor II (70/80kDa)	Tgfbr2
25	2	GSE48503	Throat	Pathway in cancers	E1A binding protein p300	EP300
26					collagen, type IV, alpha 1	COL4a1
27					collagen, type IV, alpha 2	COL4a2
28					collagen, type IV, alpha 4	COL4A4

2 grp data sheet | 3 grp data sheet | Sheet3 | Sheet1 | Sheet2

Fig. 2: Data sets of various types of cancers with two groups with the pathways involved and the genes involved in the cancer pathways with their gene symbols

GROUPS	ACESSSION NUMBER	ORGANISM	SUBMISSION DATE	COUNTRY
3	GSE60519	HOMOSAPIENS	AUG,19 2014	USA
3	GSE55942	HOMOSAPIENS	MAY,16 2014	USA
3	GSE36150	HOMOSAPIENS	FEB,29 2012	USA
3	GSE44660	HOMOSAPIENS	FEB,25 2013	USA
3	GSE56280	HOMOSAPIENS	MAR,27 2014	USA
3	GSE62500	HOMOSAPIENS	OCT,20 2014	USA

Table 2: Collected data for the different cancer types with three groups



	A	B	C	D	E	F	G
1	GROUP	ACESSION NUI	TYPE OF CANCER	PATHWAY	HIGHLY EXPRESSED GENES	GENE SYMBOL	
2	3	GSE55942	lung	Pathways in cancer	KIT ligand	KITLG	
3					baculoviral IAP repeat-containing 3	BIRC3	
4					cadherin 1, type 1, E-cadherin (epithelial)	CDH1	
5					ecotropic viral integration site 1	MECOM	
6					fms-related tyrosine kinase 3 ligand	FLT3LG	
7					integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)	ITGA2	
8					interleukin 8	IL8	
9					jun oncogene	Jun	
10					laminin, alpha 3	LAMA3	
11					laminin, beta 3	Lamb3	
12					laminin, gamma 2	LAMC2	
13					ras-related C3 botulinum toxin substrate 2 (rho family, small GTP binding protein Rac2)	rac2	
14					wingless-type MMTV integration site family, member 16	Wnt16	
15				small cell lung cancer	baculoviral IAP repeat-containing 3	BIRC3	
16					integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)	ITGA2	
17					laminin, alpha 3	LAMA3	
18					laminin, beta 3	Lamb3	
19					laminin, gamma 2	LAMC2	
20				PS3 signaling pathway	growth arrest and DNA-damage-inducible, beta	Gadd45b	0
21					serpin peptidase inhibitor, clade B (ovalbumin), member 5	SERPINB5	
22					serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1	SERPINE1	
23					thrombospondin 1	Thb51	
24	3	GSE62500	kidney	pathways in cancer	BCL2-like 1	BLC2L1	
25					fibroblast growth factor 1(acidic)	fgf1	
26					fibroblast growth factor 5	FGF5	
27					hedgehog interacting protein	Hhip	
28					hepatocyte growth factor (hepapoietin A; scatter factor)	Hgf	
29					lymphoid enhancer-binding factor 1	Lef1	
30					prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenas	PTGS2	
31					signal transducer and activator of transcription 1, 91kDa	STAT1	
32					similar to Mast/stem cell growth factor receptor precursor (SCFR) (Proto-oncogene tyrosin	KIT,LOC652739,LOC653882	
33					transforming growth factor, beta 2	TGFB2	
34	3	GSE53914	Throat	PS3 signaling pathway	BH3 interacting domain death agonist	BID	

Fig. 3: Data sets of various types of cancers with two groups with the pathways involved and the genes involved in the cancer pathways with their gene symbols

5. RESULTS AND DISCUSSION

Analysis of the microarray data of different cancer types with two or multiple group GSM files in the GEO datasets using the GEO2R analysis tool where normalization of the data and the identification of top 250 highly expressed genes is done and the gene ID's are uploaded in the DAVID functional annotation tool to obtain the annotation summary results and identify the pathways involved in the cancer expression and the genes responsible for the cancer cause is identified with their gene symbols and submitted to the Blast2GO pro tool and sequences run with blast where DNA sequences are converted into proteins and with the interpro scan similar matches for the proteins are identified followingly annotation statistics are studied for

gene ontology distribution and direct gene ontology count for molecular function is made . A key goal of cancer studies is to systematically characterize the cellular molecular mechanisms involved in disease progression.

5.1 ANALYSIS OF TWO GROUP DATASET USING GEO2R:

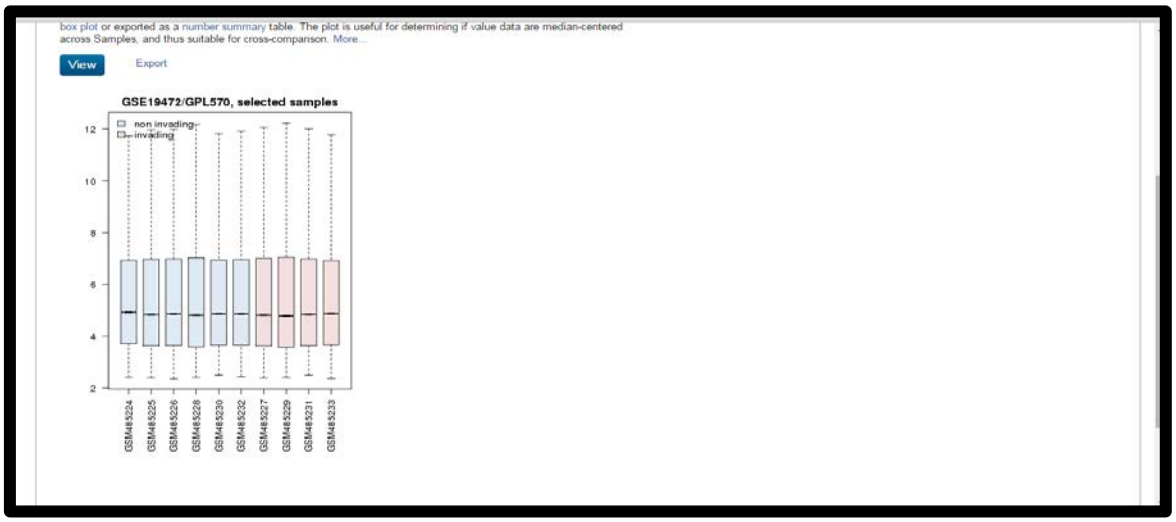
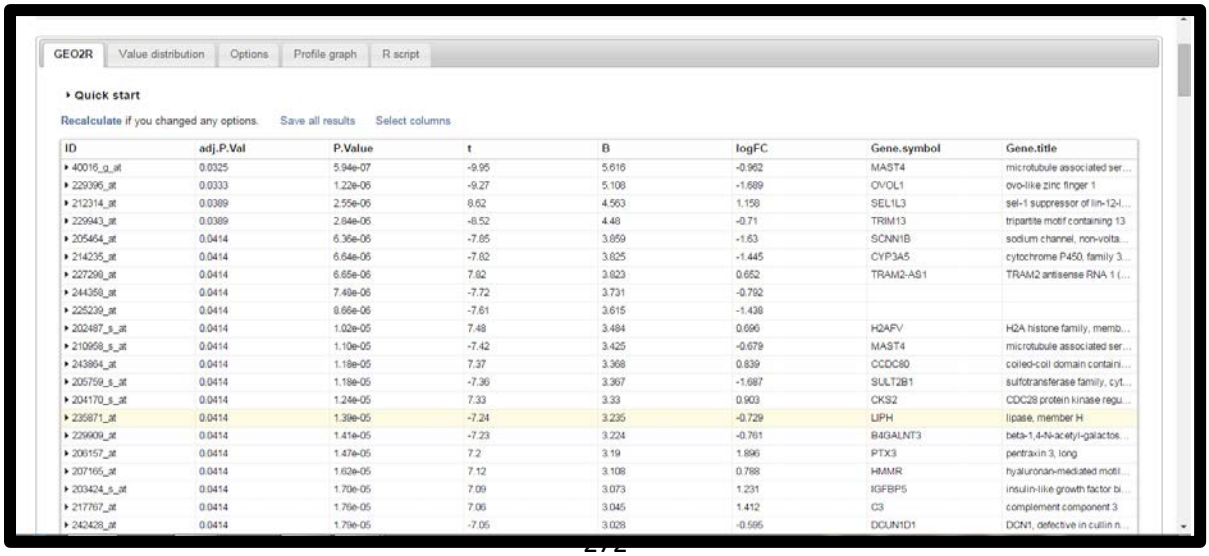


Fig. 4: Result of value distribution in the form of box plot for the esophagus cancer with two group GSM files with accession number GSE1947



ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
40016_g_at	0.0325	5.94e-07	-9.95	5.616	-0.962	MAST4	microtubule associated ser...
229396_at	0.0333	1.22e-06	-9.27	5.108	-1.589	OVOL1	ovo-like zinc finger 1
212314_at	0.0389	2.55e-06	8.62	4.563	1.158	SEL1L3	sel-1 suppressor of lin-12-l...
229943_at	0.0389	2.84e-06	-8.52	4.48	-0.71	TRIM13	tripartite motif containing 13
205464_at	0.0414	6.36e-06	-7.65	3.659	-1.63	SCNN1B	sodium channel, non-volta...
214235_at	0.0414	6.64e-06	-7.62	3.625	-1.445	CYP3A5	cytochrome P450, family 3...
227298_at	0.0414	6.65e-06	7.62	3.623	0.652	TRAM2-AS1	TRAM2 antisense RNA 1 (...
244358_at	0.0414	7.49e-06	-7.72	3.731	-0.792		
225239_at	0.0414	8.66e-06	-7.61	3.615	-1.439		
202487_s_at	0.0414	1.02e-05	7.48	3.484	0.699	H2AFV	H2A histone family, memb...
210958_s_at	0.0414	1.10e-05	-7.42	3.425	-0.679	MAST4	microtubule associated ser...
243864_at	0.0414	1.18e-05	7.37	3.368	0.839	CCDC80	coiled-coil domain containi...
205759_s_at	0.0414	1.18e-05	-7.36	3.367	-1.687	SULT2B1	sulfotransferase family, cyt...
204170_s_at	0.0414	1.24e-05	7.33	3.33	0.903	CKS2	CDC28 protein kinase regu...
235871_at	0.0414	1.39e-05	-7.24	3.235	-0.729	LIPH	lipase, member H
229909_at	0.0414	1.41e-05	-7.23	3.224	-0.761	B4GALNT3	beta-1,4-N-acetyl-galactos...
206157_at	0.0414	1.47e-05	7.2	3.19	1.896	PTX3	pentraxin 3, long
207165_at	0.0414	1.62e-05	7.12	3.108	0.788	HMMR	hyaluronan-mediated motil...
203424_s_at	0.0414	1.70e-05	7.09	3.073	1.231	IGFBP5	insulin-like growth factor bi...
217767_at	0.0414	1.76e-05	7.06	3.045	1.412	C3	complement component 3
242428_at	0.0414	1.79e-05	-7.05	3.028	-0.595	DCUN1D1	DCN1, defective in cullin n...

Fig. 5: List of top 250 highly expressed genes with their gene ID's

5.1.1 DAVID Annotation Tool Results.

Gene ID's were submitted to the DAVID annotation tool and for top 250 gene ID's the annotation summary results were obtained as shown in the fig.6. followingly the pathways involved in the expression of a cancer was identified as shown in fig.7, P53 signaling pathway is analyzed as shown in fig 8, and the genes involved in the pathway of cancer were identified and their official gene symbols were recorded as shown in the fig 9.

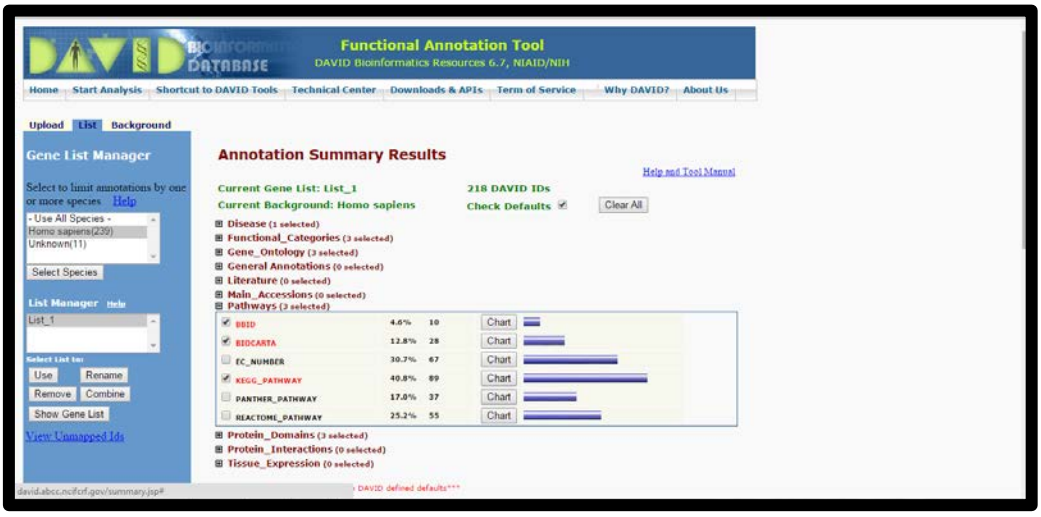


Fig. 6: Annotation summary results after submitting the gene ID's to the DAVID functional annotation tool.

Annotation summary results from DAVID annotation tool, from which the pathways involved in cancer were selected for further analysis as shown in fig.6.

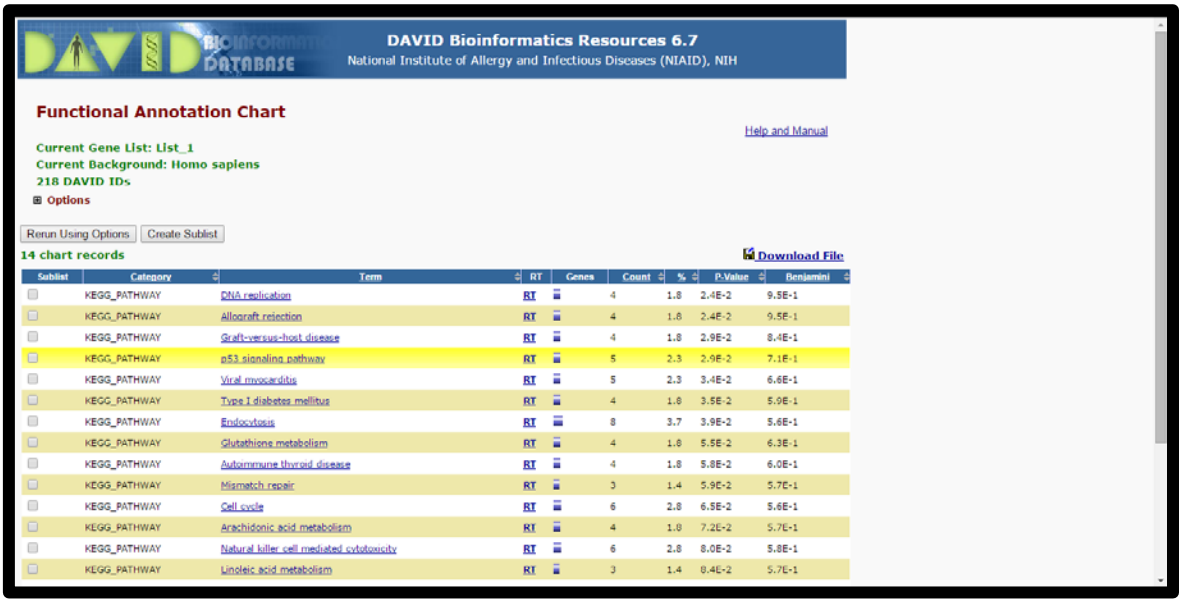


Fig. 7: Selection of pathway involved in cancer

Through the functional annotation chart, P53 signalling pathway is found ,responsible for the cause of cancer and is analysed further to know the genes responsible for the pathway as shown in fig.7.

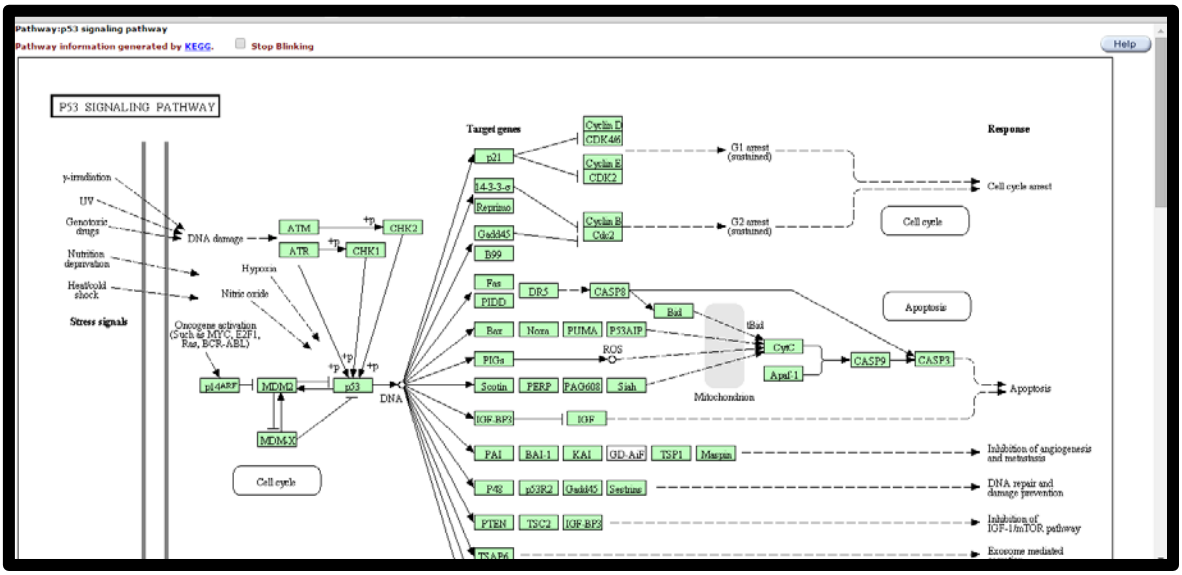


Fig. 8: P53 signalling pathway is analysed

Analysis of P53 signalling pathway is done to know which genes are responsible for the pathway of cancer causing the Esophagus cancer as shown in fig.8.

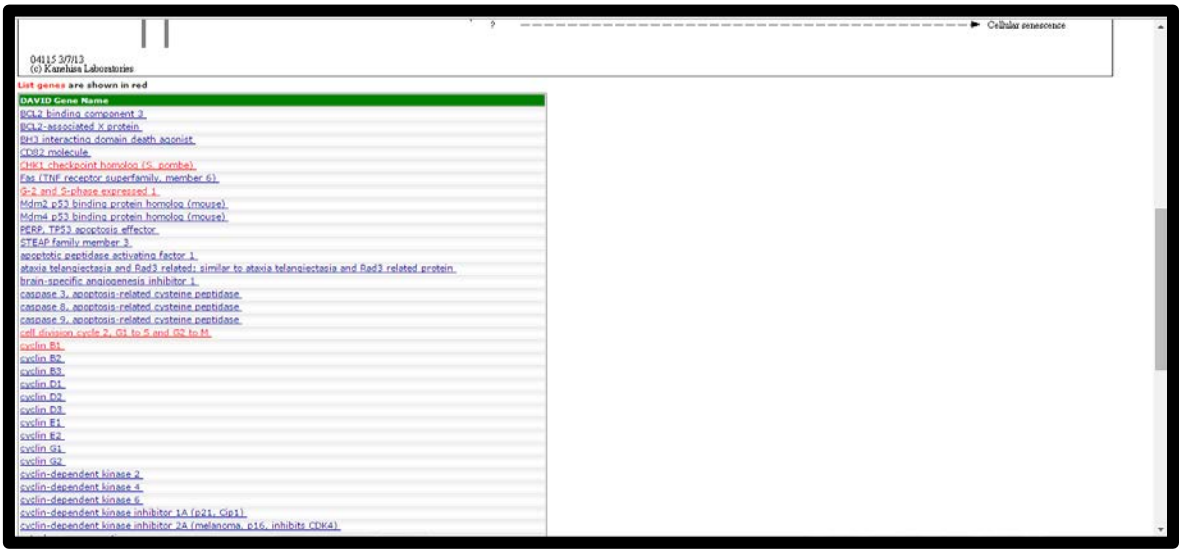


Fig. 9: Genes involved in the pathway are identified and their official gene symbols are recorded.

5.1.2 BLAST2GO Tool Results:

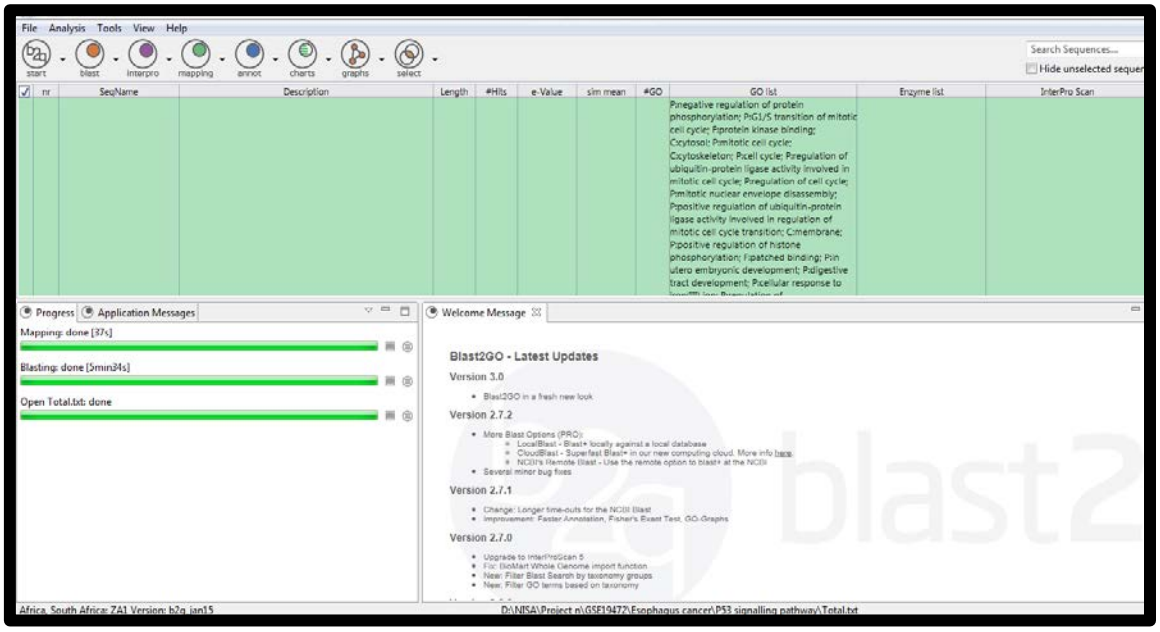


FIG. 10: Results for the blast and proceed for the inter pro scan and mapping.

All the FASTA format sequence of the genes responsible for the cancer are uploaded to the Blast2GO PRO tool to analyse through various functions of it such as visualization, management, and statistical analysis of annotation results as shown in fig.10.

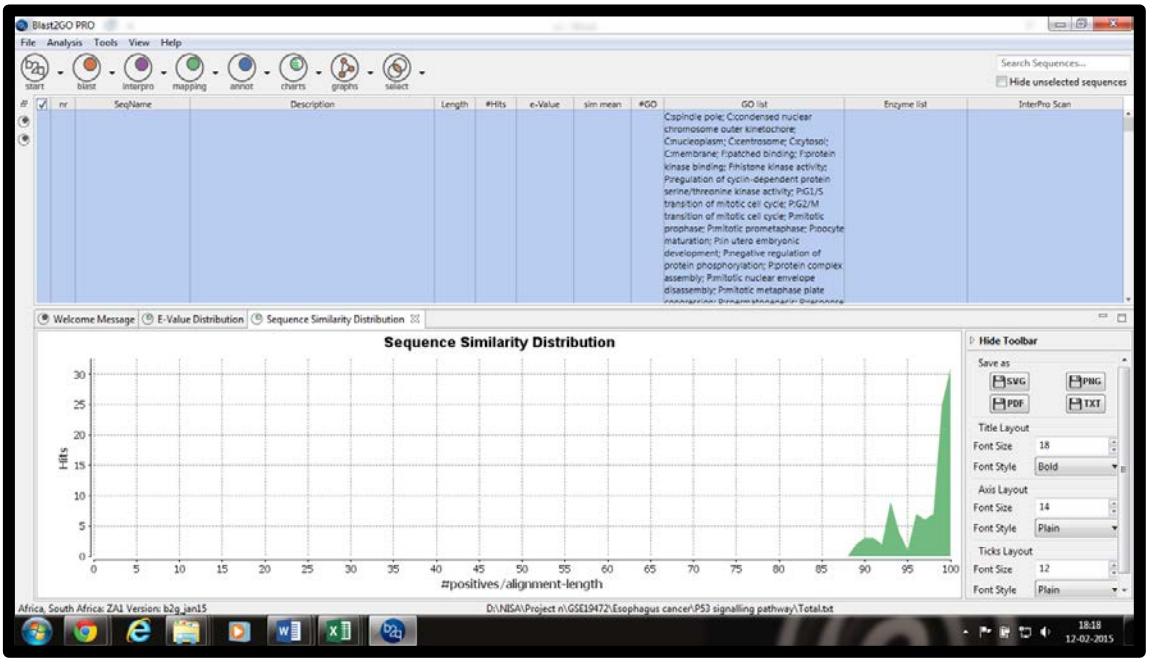


Fig. 11: The sequence similarity distribution and E value distribution

Through this chart, sequence similarity distribution can be analysed which denotes the distribution of all calculated sequence similarities (percentages), as shown in fig. 11.

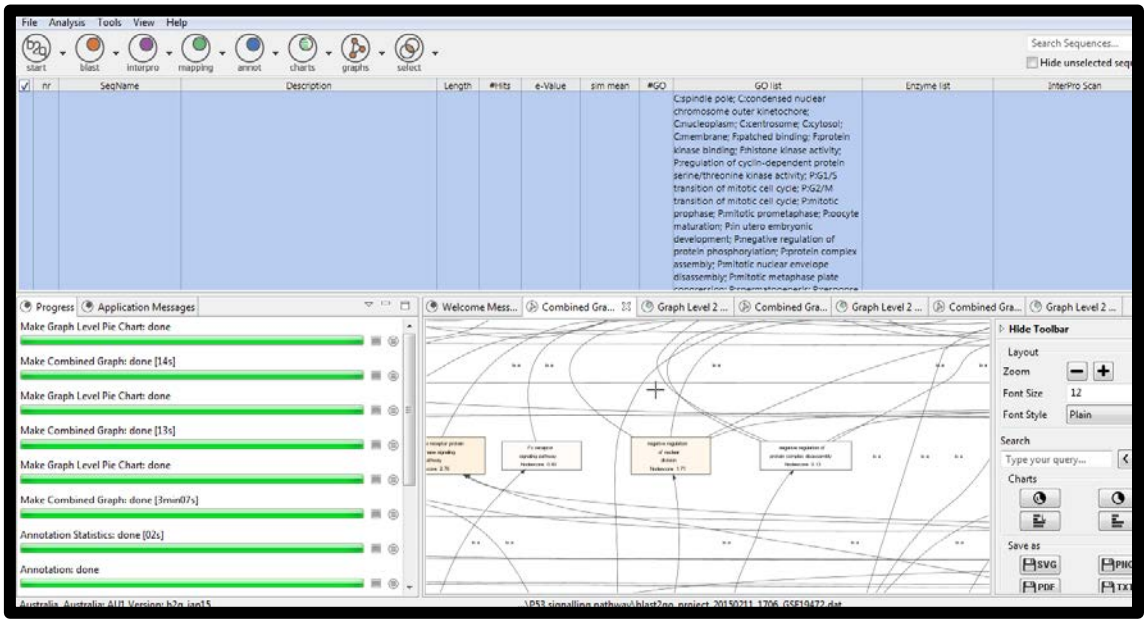
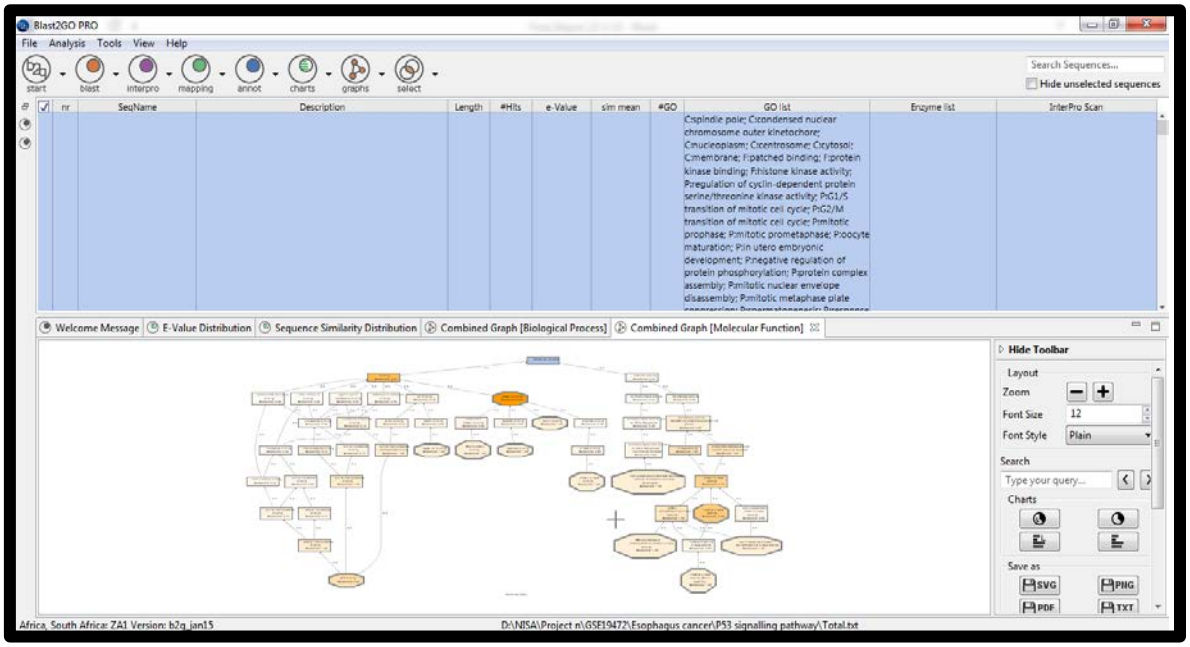


Fig. 12: Combined graph for the biological process.



International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, May 2015
ISSN: 2395-3470
www.ijseas.com

Fig. 13: The combined graph for the molecular function.

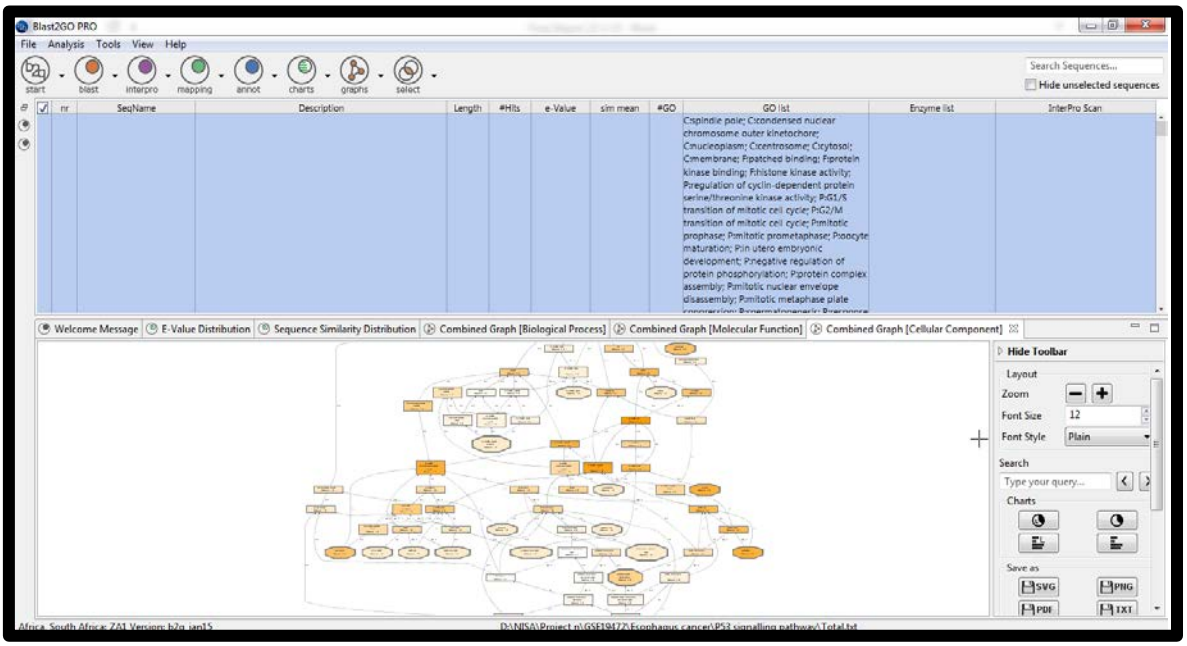


Fig. 14: Combined graph for cellular component

This confluence score takes into account the number of sequences converging at one GO term and at the same time penalizes by the distance to the term where each sequence was actually annotated. Assigned sequences and scores can be displayed at the terms level as shown in fig. 12, 13, 14.

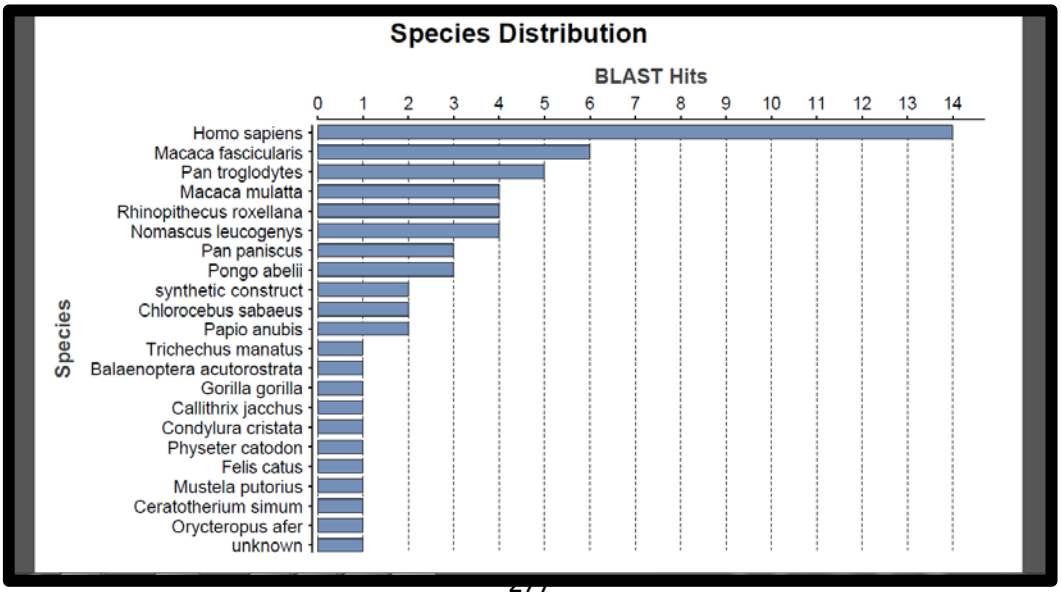
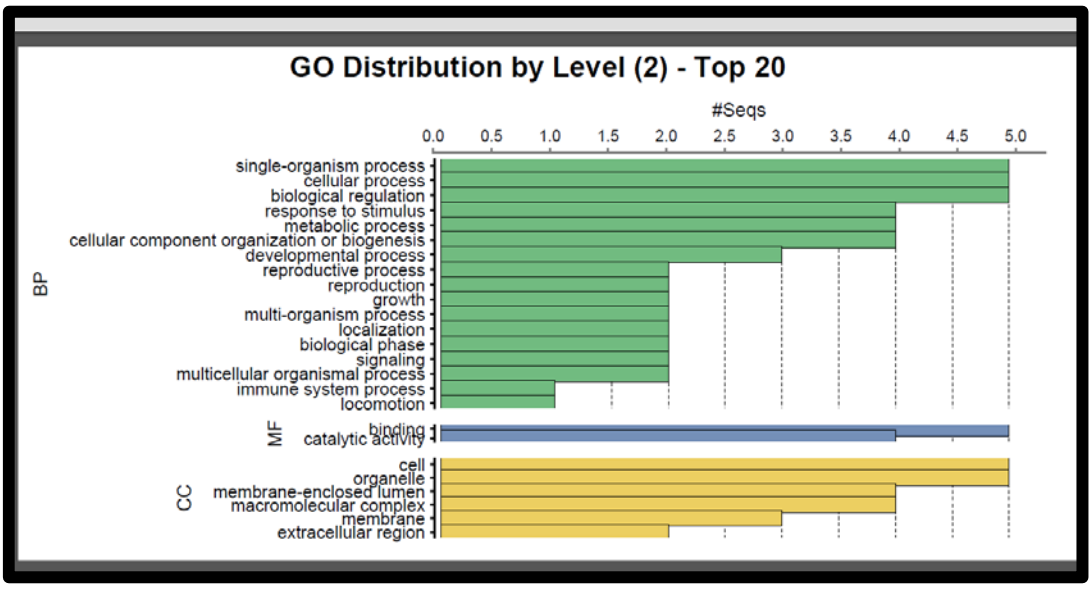


Fig. 15: Blast statistics result

When the recorded gene symbols are submitted to NCBI and the sequences in FASTA format are downloaded and loaded in Blast2GO tool. By using Blast2GO tool species distribution was analysed and this chart gives the distribution of different species to which more sequences were aligned during the Blast run, as we have selected Homosapien that shows maximum Blast Hit represented in the fig.15.

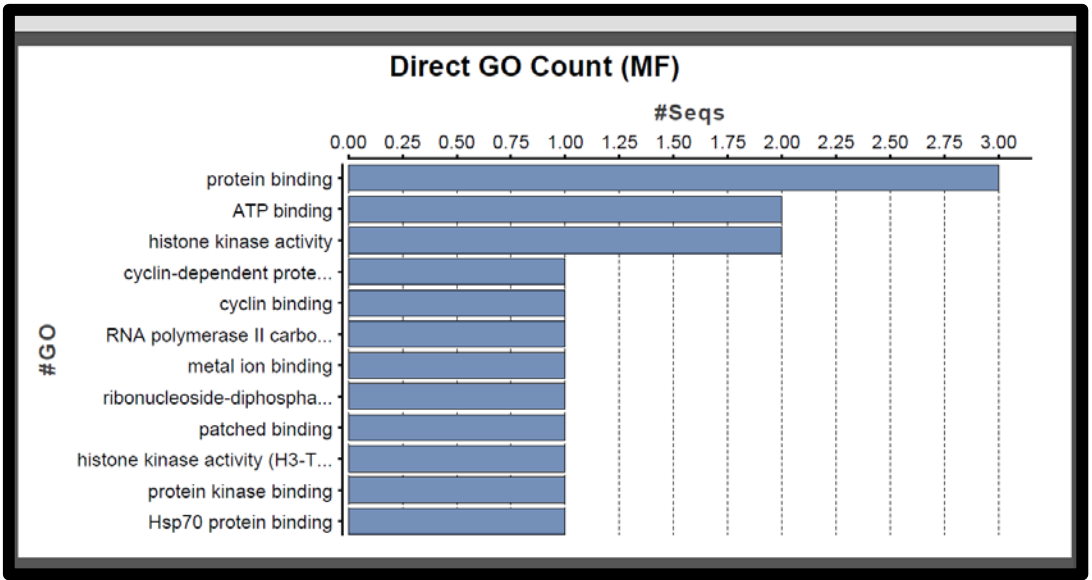
Fig. 16:



Results of annotation statistics for biological process, molecular function, cellular component

Cellular component consists of cell organelle, membrane enclosed lumen, macromolecular complex, extracellular region and the molecular function consists of depiction of binding and catalytic activity shown in fig.4.

Fig. 17:



Results of annotation statistics for biological process, molecular function, cellular component.

By using Blast2GO tool the direct GO count was analysed and by using this chart ,which gives total number of protein binding, ATP binding, as shown in fig. 17.

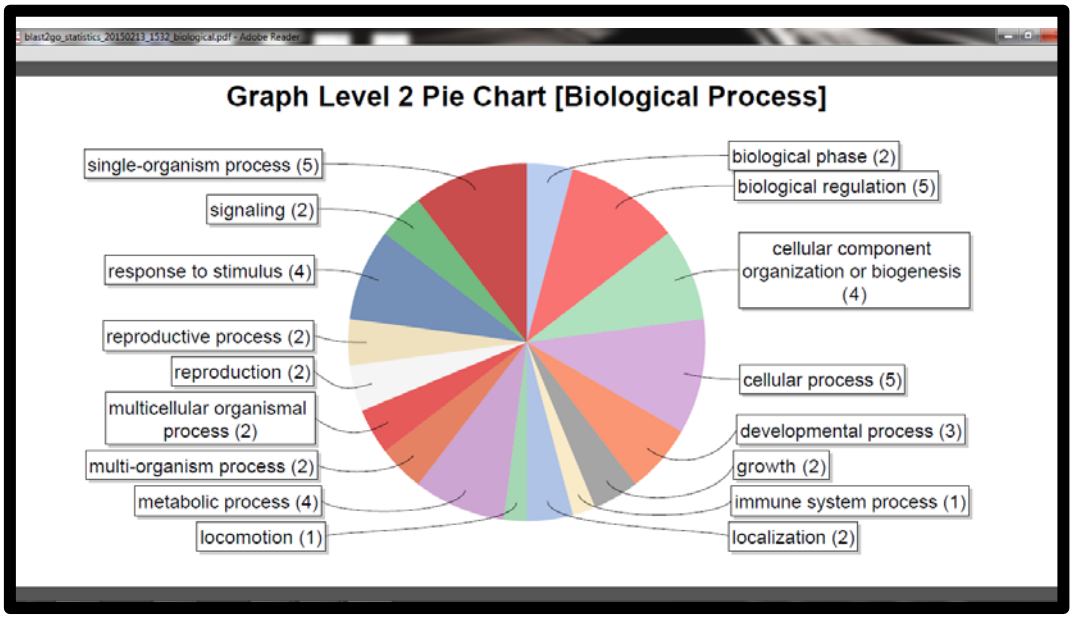


Fig. 18: Results of Gene ontology for biological process with pie chart

By using Blast2GO tool the pie chart for the biological process was obtained and this chart represents a series of events such as growth, immune system process, localization, biological regulation which are effected by the cancer gene activity on them as shown in fig.18.

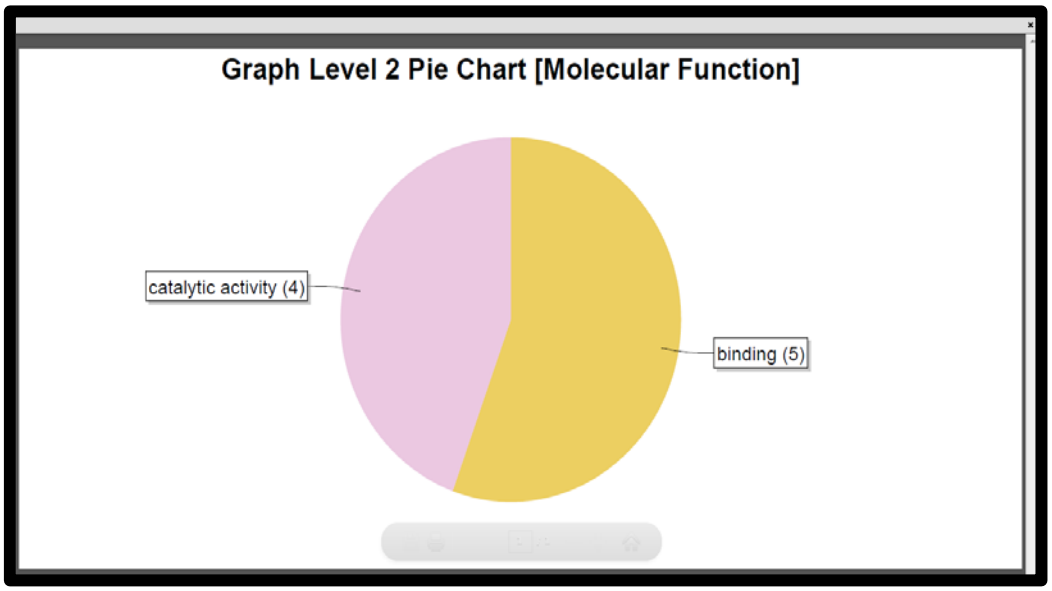


Fig. 19**:** Result of Gene ontology for molecular function with pie chart

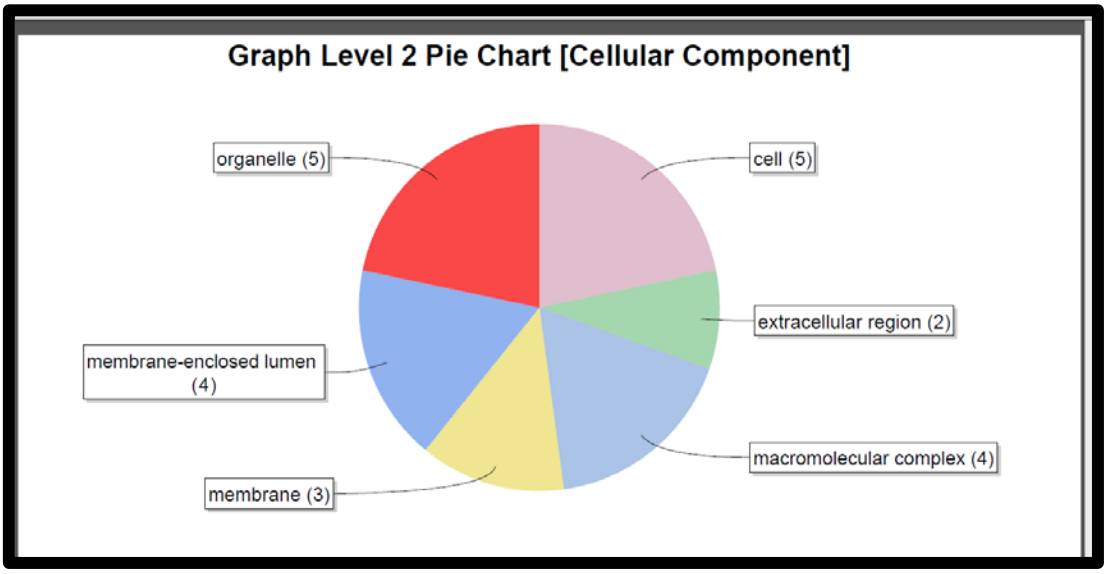


Fig. 20**:** Result of Gene ontology for cellular component with pie chart

Through Blast2GO tool the pie chart for the cellular component was analysed and this chart represents a part of a cell lumen, extracellular region, membrane, macromolecular complex as shown in fig.20.

Similar result analysis for 3 group data set is carried out.

CONCLUSION

Cancer is one of the dreadful diseases, which causes a considerable death rate in humans. Cancer is featured by an irregular, unmanageable growth that may demolish and attack neighboring healthy body tissues or somewhere else in the body. Microarray based gene expression profiling has been emerged as an efficient technique for cancer classification, as well as for diagnosis, prognosis, and treatment purposes.

Microarray based gene expression profiling using R has become an important and promising approach that can be used for cancer classification. This project mainly gives an overview of the genes involved in the expression of the various cancer types with two or multiple group data's as an example of the results mentioned above of the esophagus cancer with GEO accession number GSE19742 results which clearly depict the highly expressed genes with the

gene symbols CHEK1,GTSE1, CDK1,CCB1,RRM2 from which their functions, count of highly expressed genes, annotations, blast statistics, pathways involved in the gene expressions, gene ontology of biological process, gene ontology of molecular function, gene ontology of cellular components results were obtained, similarly it is analysed for 20 varied cancer types. This is an important step for diagnosis and prognosis purposes.

SCOPE OF FUTURE WORK

Microarray data analysis using R programming cancer studies can be carried out systematically and characterize the cellular, molecular mechanisms involved in disease progression and highly expressed genes causing cancer has been identified.

In future work for the highly expressed genes, their gene products will be identified which will be further help in the drug discovery to combat the cancer progressions.

REFERENCES

1. Hala .M, Ghada .H, Alshamlan, Badr, Jorge Andrade, & Bao Riyue “A Study of Cancer Microarray Gene Expression Profile: Objectives and Approaches” in London, U.K. Proceedings of the World Congress on Engineering, 2013 vol 2.
2. Milena B. Furtado, Hieu T. Nima, Jodee A. Gould, Mauro W, Costa, Nadia A. Rosenthal , Sarah E. Boyd. “Microarray profiling to analyse adult cardiac fibroblast identity”. 2014 Genomics Data, vol 2, pp.345–350.
3. Michael R. Stratton, Peter J. Campbell& P. Andrew Futreal.“The cancer genome”. 2009, Nature, vol 458, pp.719-724
4. Shizhu Zang , Ruifang Guo , Rui Xing , Liang Zhang , Wenmei Li ,Min Zhao , Jingyuan Fang , Fulian Hu , Bin Kang , Yonghong Ren , Yonglong Zhuang, Siqi Liu, Rong Wang, Xianghong Li, Yingyan Yu,Jing Cheng, Youyong Lu. “Identification of Differentially-expressed Genesin Intestinal Gastric Cancer by Microarray Analysis” 2014 Genomics Proteomics Bioinformatics, vol 12, pp.276–283.
5. Jing Hua Zhao and Qihua Tan “Integrated Analysis of Genetic Data with R” 2006 Human Genomics, vol 2, pp.258-265.

6. Audrey Kauffmann, Tim F. Rayner, Helen Parkinson, Misha Kapushesky, Marg us Lukk, Alvin Brazma1 and Wolfgang Huber. “Importing Array Express datasets into R/Bioconductor” 2009 CB vol 2.
7. Sandrine Dudoit1, Robert C. Gentleman, and John Quackenbush, **“**Open Source Software for the Analysis of Microarray Data”.2003 Bio Techniques, vol 34, pp.45-51
8. Sean Davis and Paul S. “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and Bio Conductor”.Bioinformatics,2007 vol 23, pp.1846–1847.