ISSN 2664-8075 (Print) & ISSN 2706-5774 (Online)
South Asian Research Journal of Applied Medical Sciences
Abbreviated Key Title: South Asian Res J App Med Sci

| Volume-5 | Issue-5 | Sep-Oct- 2023 | **DOI: 10.36346/sarjams.2023.v05i05.005**

Original Research Article

Text Mining Identifies Key Pattern Analysis and Process in Prostate Cancer

Boluwatife J. Ope[1,2*], Gabriel S. Akinbami[1], Oghenetejeri Ukochowera[2], Eloho Ukochovwera Ologan[3,4]

[1]Biochemistry Department, Faculty of Science, Lagos State University, Nigeria
[2]Bioengineering Department, Faculty of Engineering, Cyprus International University, Mersin 10, Turkey
[3]Rural and Northern Health, Kinesiology and Health Sciences, Laurentian University, Sudbury, Ontario, Canada
[4]Dr Gill Arcand Centre for Health Equity formerly the Centre for Social Accountability, Northern Ontario School of Medicine University (NOSMU) Sudbury, Ontario, Canada

***Corresponding Author:** Boluwatife J. Ope
Biochemistry Department, Faculty of Science, Lagos State University, Nigeria

Article History
Received: 16.08.2023
Accepted: 19.09.2023
Published: 23.09.2023

Abstract: One of the most typical malignancies in males is prostate cancer, and its global burden is increasing. Using text-mining technology, this study seeks to pinpoint important genes and biochemical processes related to prostate cancer. Using certain terms related to gene expression, PUBMED abstracts of interest were found. The extracted abstracts included gene pairings and functional connections. On the genes identified from the function interactions, biological procedures enrichment, network analysis, and gene prioritizing utilizing edge centrality of betweenness were carried out. For the modules containing at least five genes, which were retrieved from the network analysis, gene clustering and pathway enrichment analyses were built. The biological functions of the newly identified genes showed that they were involved in positive transcriptional regulation from the RNA polymerase II promoter, positive regulation of cell proliferation, and drug responsiveness. The prostate cancer enrichment analysis processes revealed that the NF signalling pathway, PI3k-Art signalling pathway, thyroid hormone signalling, and ErbB signalling pathways were enriched. According to the network analysis results, which were further sorted by their values for degree of between-ness, it was discovered that AKT1, AR, and KDM3A were the important genes. In conclusion, by concentrating on the discovered hub genes, prostate cancer can be medically treated.

Keywords: Prostate cancer, Text mining, Pubmed, Gene expression, signaling pathway.

1. INTRODUCTION

One of the most dangerous and severe diseases to affect people is cancer, which can be fatal. Cancer develops when a cell's DNA (deoxyribonucleic acid) is damaged (mutated), causing the cell to lose its natural activity and instead acquire the capacity to proliferate endlessly until normal tissue functionalities are compromised (Jurca *et al.*, 2016a). Cell division genes typically contain cancerous DNA mutations, which can arise from an intricate combination of genetic and environmental variables. Prostate cancer, the cancer diagnosed most frequently and the sixth most prevalent cause of cancer mortality in males, is the malignancy most inclined to affect men in the Western world (Jemal *et al.*, 2011). But only about 20% of these guys will be found to have lethal, aggressive diseases, and the rest of the patients eventually pass away for other reasons (Greene *et al.*, 2005; Ploussard & de la Taille, 2010). It is still a significant public health issue in every Western nation. According to (Siegel *et al.*, 2014), there were 233,000 new instances of PSA diagnosed in the US in 2014, and 29,480 men died. Unfortunately, no effective medication has yet been found that provides a surefire cure for cancer. As a result, researchers from a wide range of disciplines continue to work hard to identify substances (often genes or protein molecules) that might potentially be utilized as cancer-related biomarkers.

Medical researchers seek to recognize and characterize biomarkers that are indicative of each form of cancer to deliver the most precise diagnosis to patients and to personalize therapeutic methods for cancer patients (van 't Veer & Bernards, 2008). According to (Mishra & Verma, 2010), a cancer biomarker is a material or activity that acts as a sign of

Citation: Boluwatife J. Ope, Gabriel S. Akinbami, Oghenetejeri Ukochowera, Eloho Ukochovwera Ologan (2023). Text Mining Identifies Key Pattern Analysis and Process in Prostate Cancer. *South Asian Res J App Med Sci*, *5*(5), 100-112.

cancer in the body. Genetics is a typical instance of a cancer biomarker. Genetic biomarkers' essential building block is the gene. An area of DNA called a gene typically contains the instructions needed to build proteins. One unit of DNA that frequently carries the information required to generate proteins is the gene, which serves as the fundamental building block of genetic biomarkers. This theory's central assumption is that genes undergo translation into the intermediary molecule of RNA and that RNA eventually transforms into protein molecules, which carry out the fundamental tasks of life (Jurca *et al.*, 2016b). If a protein-producing gene that fights cancer is damaged or inhibited, the cell may develop cancer. Similar to this, if a gene that produces a protein that encourages cancer activity grows, a cell may likewise develop cancer. It is vital to identify the many genes and circumstances likely to result in malignancies, should gene expression be up- or down-regulated, to evaluate whether it is prostate cancer. The issue is important because several internal and environmental factors could damage the cells and cause cancer. Different people conduct and display different behaviours (Jurca *et al.*, 2016b).

Since these genes or molecules could be utilized as cancer biomarkers, researchers from various fields continue to devote a lot of work to finding them. Numerous techniques have been created. Numerous procedures are used in the research, including computer scientists' computational techniques and biologists' wet lab tests. Multidisciplinary techniques used are text mining, which comprises retrieval of information, text evaluation, extraction of information, clustering, categorization, visualization, the use of databases, machine learning, and data mining (Vardakas *et al.*, 2015). It is an effective tool for swiftly identifying the most important information from voluminous biological literature. However, to fully utilise this potential, the researcher needs to be well-versed in the accessibility, applicability, adaptability, interoperability, and comparative accuracy of the available text-mining resources. The researcher can quickly extract pertinent data from massive amounts of literature using the text-mining (Feldman & Dagan, 1995). As a result of the ensuing research, far fewer molecules need to be considered as potential markers, which is encouraging. Therefore, genetic biomarkers suggestive of the illness are being sought after by biomedical researchers. The existing literature can be used to determine new biomarkers. However, this is a challenging undertaking given the enormous number of research articles on prostate cancer. This study offers a framework that looks into existing literature data in search of instructive findings. It mixes text mining and social network analysis to pinpoint important pattern analysis and processes in prostate cancer. This study shows the significant benefits of text mining in locating important patterns and processes in the study of prostate cancer. Large volumes of scientific literature and databases can be mined for insightful information by using cutting- edge tools and approaches.

This study emphasizes the crucial function of text mining as a potent instrument that supports conventional research techniques and offers a practical way to make use of the enormous amount of textual data available in prostate cancer research. Researchers and healthcare practitioners can discover hidden patterns, find novel biomarkers, and acquire a deeper knowledge of the molecular pathways underlying prostate cancer growth by utilizing the capabilities of artificial intelligence and natural language processing tools. In addition to potential future paths and research opportunities in this rapidly emerging discipline, the difficulties and restrictions connected with text mining were explored.

2. MATERIALS AND METHODS

2.1 Methodology

Our objective is to offer fresh concepts for knowledge discovery (KD) and hypothesis creation about genes involved in prostate cancer. The initial stage of our approach was the information retrieval (IR) step, where we aimed to find all pertinent publications associated with our area of interest: cancer. We chose to look at abstracts since they have the most significant and concise keywords, even though full-text analysis has more information. Cancer. Even though full-text analysis provides more extensive information, we decided to focus on abstracts because they contain the most crucial and concise keywords. Additionally, evaluating abstracts is faster, allowing for a more thorough text analysis. We also anticipated that full texts might mention irrelevant genes or genes related to other malignancies, which could introduce noise. In simpler terms, full-text mining may yield more results, while abstract-based mining may yield more precise findings. As a result, our initial step was to search for as many academic abstracts on cancer as possible. We accessed the MedLine database through the PubMed API and obtained each abstract used in our study. Because it is known for being the most widely used search engine for scientific literature (Faro *et al.*, 2012; Zhu *et al.*, 2013), we decided to utilize PubMed.

2.1.1 Gene Curation and the Gathering of Literature

The Ensemble database was searched for several gene groupings that are typically associated with or cause cancer. For our query search, these gene categories were pre-set as standard in our Pubmed_extractor programme. We searched using the terms "Mesh term" and "Neoplasm," as well as terms like "Overexpressing," "Downregulating," "Enhancement," etc. There were 614,163 abstracts in total that were obtained from PubMed (Ono & Kuhara, 2014). To manually examine and further curate all 614,163 PubMed abstracts, we downloaded them all in Medline format. The publications were then further screened to ensure that the remaining sample only contained cancers that we were specifically interested in researching, such as prostate, cervical, pancreatic, glioblastoma, and breast cancer. Only abstracts from articles with genes relevant to the aforementioned cancer types were included in the set of papers. As a result, 12,566 papers were used in the analysis from our final paper set. 2,522 abstracts of articles that addressed prostate cancer. The abstracts that were disregarded after the NER stage might have been related to other aspects of prostate cancer, such as those that have to do with health care or psychology, rather than the genetic part that we are interested in. 373 Entrez human genes were ultimately discovered from 2,522 PubMed abstracts. The human gene database GeneCard, an integrated database that offers thorough information on all predicted and annotated human genes, validated their conventional names. It incorporates gene-centric information from over 150 websites' worth of genomic, transcriptomic, proteomic, biological, and functional data (Safran *et al.*, 2010). After deleting the genes that appeared twice or more, there were only 296 left. With additional research being done on these remaining genes.

2.1.2 Biological Annotation

Functional annotations for each gene were gathered to better understand the biological function of the genes involved in prostate development. This was done using the David database, which was used for annotation, visualisation, and Integrated Discovery. It gives researchers a complete set of gene function annotations so they can comprehend the biological significance of a long list of genes. From the KEGG pathway, an embedded database in David, associated biological pathways for certain genes were retrieved. A set of databases known as KEGG (Encyclopaedia of Genes and Genomes) deals with genomes, biological pathways, illnesses, medications, and chemical compounds. In addition, KEGG Disease database reports of connections with additional disorders were included (Kanehisa *et al.*, 2010). Gene Ontology was also derived for functional annotation, clustering, and gene functional classification using the David database. We applied KEGG and GOstats in the present study.

2.1.3 Protein-Protein Interaction and Network Reconstruction

We examined how often genes were found together in the abstracts. Since the genes may serve as the "actors" and the number of abstract instances may represent the "activity" that occurs among two genes, network analysis was an appropriate technique for investigating gene-gene relationships. To expand our knowledge of prostate cancer, we compiled all the interactions between 296 prostate-involved genes using a unique pathway-human interactome. This interactome, based on the modules formed by these genes, offers a comprehensive view. Using a module-searching technique, the pathway-based interactome was mapped to the glioblastoma-implicated genes (Zhao *et al.,* 2013). We used the R programming language to map all the genes related to prostate cancer onto the integrated interactome. Then, by joining these genes together along their shortest pathways, we created a subnetwork. Unlike examining a single gene, biological networks are frequently too complex to determine the function of each component. However, because a few simple topological principles relate to network function, the topological properties of a network are usually used to define its overall function. For the social network analysis in this case, the R programming language was used. For each gene in the system, we calculated the degree, or the number of connections between nodes in a network, as well as the short path, or the separation between two genes (Barabási & Oltvai, 2004).

2.1.4 Analysis of Functional Enrichment

Functional enrichment analysis is a powerful method that combines biology and mathematics to handle large gene chip data. In our study, we used the GO-stats and KEGG-db toolbox in the David database for this analysis. We selected GO entries with a Count value of 2 or more and a p-value less than 0.05. For KEGG pathways, we considered those with a p-value below 0.05. To determine expression correlation scores and associated P values, we utilized R (version 2.14.0) and accounted for multiple tests using a false discovery rate (FDR) adjustment. David helped identify statistically representative KEGG and GO-term pathways for each gene set, with P values adjusted using Benjamini-Hochberg methods. We used human protein-coding genes as background in these analyses, and pathways with a corrected P value less than 0.01 were considered significant (Zhang *et al.*, 2019;Ferreira, 2007).

3. RESULTS

Table 1a: Shows how functional annotations for genes in modules 9, 22, and 66 of the Gene Ontology (GO) biological process overlap. With the help of a modified Fisher's exact test (EASE score), the significance of gene-term enrichment was investigated. Higher importance is indicated by smaller P-values. The expected percentage of false discoveries is controlled by the false discovery rate (FDR), which was set at 0.05.

TERM	NUMBER OF GENE	P-Value	FDR
negative regulation of apoptotic process	20	3.48E-12	5.83E-09
positive regulation of transcription, DNA-templated	20	2.98E-11	5.00E-08
cellular response to mechanical stimulus	10	1.52E-10	2.55E-07
Positive transcriptional control by the RNA polymerase II promoter	25	2.64E-10	4.42E-07
extrinsic apoptotic signaling pathway in absence of ligand	8	5.18E-10	8.69E-07
apoptotic process	19	1.19E-09	1.99E-06

TERM	NUMBER OF GENE	P-Value	FDR
regulation of apoptotic process	12	2.06E-08	3.45E-05
Cytochrome c-mediated stimulation of cysteine-type endopeptidase activity implicated in the apoptotic pathway	5	9.96E-08	1.67E-04
reaction of the intrinsic apoptotic signaling system to DNA damage	7	1.93E-07	3.23E-04
positive regulation of cell proliferation	15	2.00E-07	3.36E-04
peptidyl-serine phosphorylation	9	3.81E-07	6.39E-04
liver regeneration	6	4.48E-07	7.51E-04
response to estradiol	8	6.08E-07	0.001018471
regulation of Golgi inheritance	4	6.13E-07	0.001028328
response to drug	12	7.47E-07	0.001252407
inverse control of gene expression	9	7.69E-07	0.001288794
Endoplasmic reticulum stress triggers the intrinsic apoptotic signalling pathway.	6	8.81E-07	0.001476213
protein phosphorylation	14	1.03E-06	0.001722708
epithelial to mesenchymal transition	6	1.03E-06	0.001723357
execution phase of apoptosis	5	1.40E-06	0.002342463
Transcription is negatively regulated by the RNA polymerase II promoter.	17	1.43E-06	0.002390621
positive regulation of gene expression	11	1.53E-06	0.002557269
peptidyl-threonine phosphorylation	6	1.82E-06	0.003056088
cellular response to DNA damage stimulus	10	1.88E-06	0.0031532
apoptotic signaling pathway	7	2.31E-06	0.003878376
macrophage differentiation	5	2.33E-06	0.003905861
trachea formation	4	3.04E-06	0.00510121
positive regulation of apoptotic process	11	5.11E-06	0.008566214
positive regulation of protein phosphorylation	8	5.73E-06	0.009603371
canonical Wnt signaling pathway	7	5.79E-06	0.009702681
response to toxic substance	7	6.65E-06	0.011145501
positive regulation of cell migration	9	6.99E-06	0.011708872
ERK1 and ERK2 cascade	5	7.89E-06	0.013229489
stress-activated MAPK cascade regulation	4	8.45E-06	0.01417079
regulating the transfer of early endosomes to late endosomes	4	8.45E-06	0.01417079
Bergmann glial cell differentiation	4	1.26E-05	0.021172017
cellular response to hypoxia	7	1.34E-05	0.022523752
Proteolysis	13	1.59E-05	0.0265858
positive control of growth of smooth muscle cells	6	1.81E-05	0.030270891
germ cell development	5	1.99E-05	0.033278779
response to gamma radiation	5	2.27E-05	0.038050228
response to wounding	6	2.29E-05	0.03846175
phosphatidylinositol-mediated signaling	7	2.37E-05	0.039686656
epithelial to mesenchymal conversion is positively regulated.	5	2.93E-05	0.049073464

Table 1b: Shows the functional enrichment statistics for the KEGG pathways for modules 9, 22, and 66 for glioblastoma. These pathways were identified as unique and significant based on an FDR value set at < 0.05.

TERM	NUMBER OF GENE	PVALUE	FDR
Prostate cancer	20	3.78E-21	4.54E-18
Hepatitis B	23	6.45E-21	7.74E-18
Pathways in cancer	31	6.54E-20	7.84E-17
Colorectal cancer	17	3.10E-19	3.71E-16
Endometrial cancer	15	3.35E-17	4.01E-14
Acute myeloid leukemia	14	4.29E-15	5.20E-12
Apoptosis	14	1.81E-14	2.17E-11
Proteoglycans in cancer	20	3.53E-14	4.24E-11
ErbB signaling pathway	15	8.03E-14	9.63E-11
PI3K-Akt signaling pathway	24	9.04E-14	1.08E-10
HIF-1 signaling pathway	15	3.32E-13	3.98E-10

TERM	NUMBER OF GENE	PVALUE	FDR
TNF signaling pathway	15	1.55E-12	1.86E-09
Melanoma	13	3.08E-12	3.70E-09
Bladder cancer	11	4.60E-12	5.52E-09
Pancreatic cancer	12	2.64E-11	3.17E-08
Glioma	12	2.64E-11	3.17E-08
Thyroid hormone signaling pathway	14	7.06E-11	8.47E-08
Chronic myeloid leukemia	12	8.45E-11	1.01E-07
Choline metabolism in cancer	13	2.28E-10	2.73E-07
Thyroid cancer	9	2.45E-10	2.94E-07
FoxO signaling pathway	14	4.94E-10	5.93E-07
Prolactin signaling pathway	11	1.50E-09	1.80E-06
Non-small cell lung cancer	10	3.13E-09	3.76E-06
Insulin signaling pathway	13	8.72E-09	1.05E-05
Toxoplasmosis	12	8.85E-09	1.06E-05
Central carbon metabolism in cancer	10	1.07E-08	1.28E-05
MicroRNAs in cancer	17	1.37E-08	1.64E-05
Tuberculosis	14	1.54E-08	1.84E-05
Neurotrophin signaling pathway	12	2.23E-08	2.67E-05
AMPK signaling pathway	12	2.89E-08	3.47E-05
Estrogen signaling pathway	11	4.05E-08	4.86E-05
mTOR signaling pathway	9	8.71E-08	1.04E-04
Focal adhesion	14	9.41E-08	1.13E-04
Signalling mechanisms controlling stem cells' pluripotency	12	1.11E-07	1.34E-04
HTLV-I infection	15	1.54E-07	1.85E-04
Sphingolipid signaling pathway	11	2.56E-07	3.07E-04
Signalling from B cell receptors	9	3.48E-07	4.17E-04
Adherens junction	9	4.36E-07	5.23E-04
Pathway for T cell receptor signalling	10	5.55E-07	6.66E-04
Viral carcinogenesis	13	7.04E-07	8.45E-04
Pathway for Toll-like receptor signalling	10	9.14E-07	0.001096202
Influenza A	12	1.01E-06	0.001210839
Small cell lung cancer	9	1.76E-06	0.002116165
Chemokine signaling pathway	12	1.95E-06	0.002343662
Ras signaling pathway	13	1.99E-06	0.002382254
Progesterone-mediated oocyte maturation	9	2.11E-06	0.002529006
VEGF signaling pathway	8	2.13E-06	0.002552589
Epstein-Barr virus infection	10	2.99E-06	0.003586265
Fc epsilon RI signaling pathway	8	4.47E-06	0.005366021
Melanogenesis	9	6.06E-06	0.007263961
Hepatitis C	10	6.11E-06	0.007332616
Rap1 signaling pathway	12	6.36E-06	0.007634182
Chagas disease (American trypanosomiasis)	9	8.12E-06	0.009740059
Pertussis	8	8.66E-06	0.010393349
Amyotrophic lateral sclerosis (ALS)	7	9.17E-06	0.011001012
Insulin resistance	9	1.08E-05	0.012895281
Non-alcoholic fatty liver disease (NAFLD)	10	1.72E-05	0.020602242
NOD-like receptor signaling pathway	7	1.79E-05	0.021439755
Prion diseases	6	1.95E-05	0.023354667
MAPK signaling pathway	12	3.68E-05	0.044169268
Transcriptional misregulation in cancer	10	3.83E-05	0.045986006

Table 2a: Displays the functional annotations overlap in the Gene Ontology (GO) biology process (BP) for genes in module 9. The P-value, also known as the EASE score, indicates the significance of gene-term enrichment. Smaller P-values indicate higher significance. The False Discovery Rate (FDR) controls the expected proportion of false discoveries and is set at < 0.05.

TERM	NUMBER OF GENE	PVALUE	FDR
Absence of ligand inhibits the extrinsic apoptotic signalling pathway	8	7.13E-11	1.17E-07
apoptotic process	16	7.40E-09	1.21E-05
regulation of apoptotic process	11	1.55E-08	2.53E-05
cellular response to mechanical stimulus	8	1.56E-08	2.56E-05
DNA-templated positive transcriptional regulation	15	1.80E-08	2.94E-05
Activation of the apoptotic process's cysteine-type endopeptidase activity by cytochrome c	5	3.24E-08	5.31E-05
negative regulation of apoptotic process	14	3.47E-08	5.69E-05
reaction of the intrinsic apoptotic signalling system to DNA damage	7	3.63E-08	5.94E-05
Positive control of transcription by the RNA pol II promoter	19	5.44E-08	8.91E-05
response to estradiol	8	8.92E-08	1.46E-04
cellular response to DNA damage stimulus	10	1.72E-07	2.82E-04
Endoplasmic reticulum stress triggers the intrinsic apoptotic signalling pathway.	6	2.19E-07	3.60E-04
epithelial to mesenchymal transition	6	2.56E-07	4.20E-04
positive regulation of cell proliferation	13	3.80E-07	6.23E-04
apoptotic signaling pathway	7	4.48E-07	7.33E-04
execution phase of apoptosis	5	4.58E-07	7.50E-04
canonical Wnt signaling pathway	7	1.14E-06	0.001859839
liver regeneration	5	5.74E-06	0.00940152
transcription is negatively regulated by the RNA polymerase II promoter	14	6.24E-06	0.010217122
response to gamma radiation	5	7.55E-06	0.012377954
positive regulation of epithelial to mesenchymal transition	5	9.76E-06	0.01599782
positive regulation of cell migration	8	1.03E-05	0.016864229
peptidyl-serine phosphorylation	7	1.23E-05	0.020190248
phosphatidylinositol 3-kinase signalling regulation	6	1.69E-05	0.02772181
DNA damage response and signal transduction are negatively regulated by p53 class mediators.	4	1.84E-05	0.030210334
Cysteine-type endopeptidase activity implicated in the apoptotic process is activated.	6	2.29E-05	0.037524605
response to toxic substance	6	2.57E-05	0.042124358
positive regulation of neuron apoptotic process	5	2.86E-05	0.046771016

Table 2b: Presents the statistical representation of KEGG pathways for genes in module 9. It showcases the functional enrichment of unique KEGG pathways that are significant in the core gene module of glioblastoma. The FDR value was set at < 0.05 to determine significance.

TERM	NUMBER OF GENE	PVALUE	FDR
Colorectal cancer	14	1.37E-16	1.33E-13
Prostate cancer	15	4.94E-16	5.22E-13
Endometrial cancer	13	6.60E-16	7.77E-13
Pathways in cancer	23	3.03E-15	3.52E-12
Apoptosis	13	6.66E-15	7.80E-12
Hepatitis B	16	2.96E-14	3.47E-11
Melanoma	11	4.26E-11	4.99E-08
Proteoglycans in cancer	14	7.52E-10	8.82E-07
Glioma	9	1.36E-08	1.59E-05
Pancreatic cancer	9	1.36E-08	1.59E-05
Adherens junction	9	2.77E-08	3.25E-05
Chronic myeloid leukemia	9	3.10E-08	3.63E-05
Thyroid cancer	7	4.00E-08	4.69E-05

TERM	NUMBER OF GENE	PVALUE	FDR
TNF signaling pathway	10	4.65E-08	5.45E-05
PI3K-Akt signaling pathway	15	6.67E-08	7.83E-05
Thyroid hormone signaling pathway	10	8.73E-08	1.02E-04
Non-small cell lung cancer	8	1.06E-07	1.25E-04
Acute myeloid leukemia	8	1.06E-07	1.25E-04
ErbB signaling pathway	9	1.40E-07	1.64E-04
HIF-1 signaling pathway	9	3.02E-07	3.54E-04
FoxO signaling pathway	10	3.28E-07	3.84E-04
Bladder cancer	7	3.55E-07	4.16E-04
Signalling mechanisms controlling stem cells' pluripotency	10	4.76E-07	5.59E-04
Toxoplasmosis	9	8.66E-07	0.001014969
HTLV-I infection	12	1.17E-06	0.001377372
Amyotrophic lateral sclerosis (ALS)	7	1.19E-06	0.001400172
Viral carcinogenesis	11	1.32E-06	0.001549174
Focal adhesion	11	1.38E-06	0.001619667
Small cell lung cancer	8	1.93E-06	0.002267602
Tuberculosis	10	3.42E-06	0.004004002
MicroRNAs in cancer	12	3.76E-06	0.004408373
Central carbon metabolism in cancer	7	5.23E-06	0.006126757
Prolactin signaling pathway	7	9.61E-06	0.011270722
Rap1 signaling pathway	10	1.38E-05	0.016168085
Neurotrophin signaling pathway	8	1.93E-05	0.022635223
Epstein-Barr virus infection	8	2.15E-05	0.025214174
Progesterone-mediated oocyte maturation	7	3.11E-05	0.036457628
Hepatitis C	8	3.76E-05	0.04412865

Table 3a: Shows the functional annotations overlap in the Gene Ontology biology process (BP) for genes in module 22. The P-value, also known as the EASE score, indicates the significance of gene-term enrichment. Smaller P-values indicate higher significance. The False Discovery Rate (FDR) controls the expected proportion of false discoveries and is set at < 0.05.

TERM	NUMBER OF GENE	PVALUE	FDR
TOR signaling	3	3.87E-06	0.004661964

Table 3b presents the statistical representation of KEGG pathways for genes in module 22. It showcases the functional enrichment of unique KEGG pathways that are significant in the core gene module of glioblastoma. The FDR value was set at < 0.05 to determine significance.

TERM	NUMBER OF GENE	PVALUE	FDR
Choline metabolism in cancer	4	1.22E-05	0.0096387
AMPK signaling pathway	4	2.20E-05	0.017459937
Insulin signaling pathway	4	3.11E-05	0.024682925

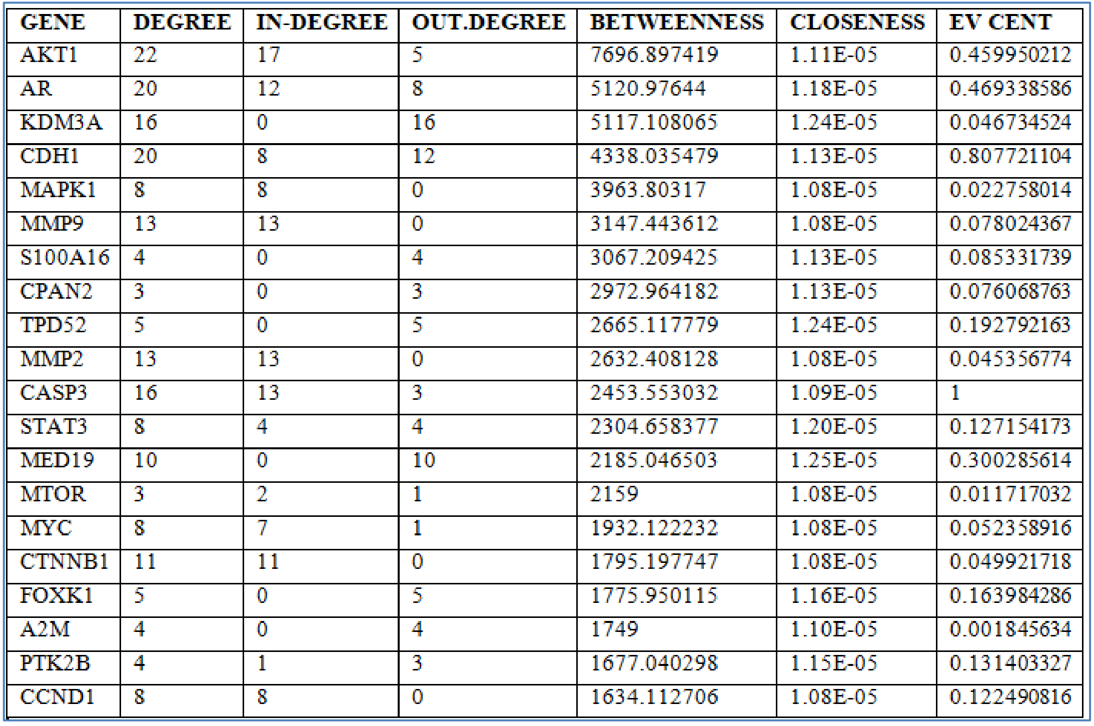
Table 4a: Displays the functional annotations overlap in the Gene Ontology (GO) biology process (BP) for genes in module 66. It includes the P-values, also known as EASE scores, which indicate the significance of gene-term enrichment. Smaller P-values indicate higher significance. The False Discovery Rate (FDR) is set at < 0.05 to control the expected proportion of false discoveries.

TERM	NUMBER OF GENE	PVALUE	FDR
inverse control of gene expression	5	7.15E-06	0.009923283
the positive control of the ERK1 and ERK2 cascade	5	1.88E-05	0.026098184

Table 4b: Shows the statistical representation of the module 66 genes in the KEGG pathways. The functional dominance of various KEGG pathways that are significant in the glioblastoma essential gene module is brought to light. The FDR value was set at 0.05 to determine significance.

TERM	NUMBER OF GENE	PVALUE	FDR
Hepatitis B	7	2.07E-07	2.38E-04
Estrogen signaling pathway	5	3.62E-05	0.041704753

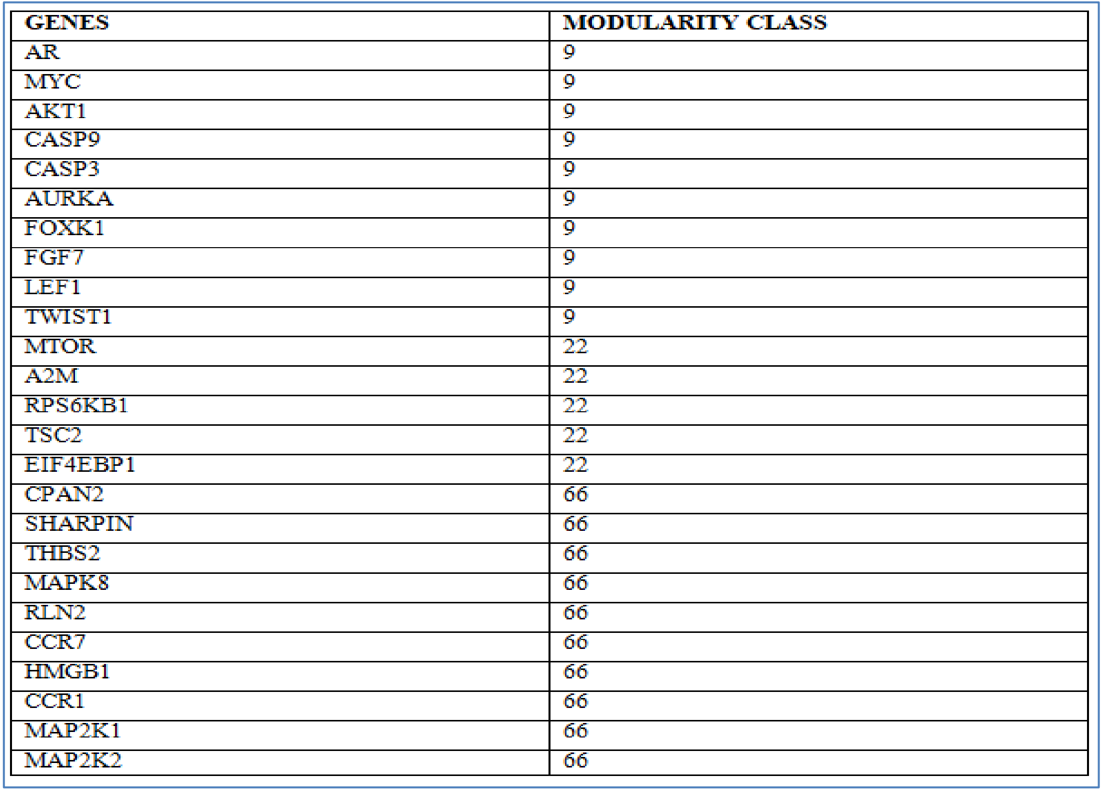
Table 5: The gene-gene network's network analysis measures



GENE	DEGREE	IN-DEGREE	OUT.DEGREE	BETWEENNESS	CLOSENESS	EV CENT
AKT1	22	17	5	7696.897419	1.11E-05	0.459950212
AR	20	12	8	5120.97644	1.18E-05	0.469338586
KDM3A	16	0	16	5117.108065	1.24E-05	0.046734524
CDH1	20	8	12	4338.035479	1.13E-05	0.807721104
MAPK1	8	8	0	3963.80317	1.08E-05	0.022758014
MMP9	13	13	0	3147.443612	1.08E-05	0.078024367
S100A16	4	0	4	3067.209425	1.13E-05	0.085331739
CPAN2	3	0	3	2972.964182	1.13E-05	0.076068763
TPD52	5	0	5	2665.117779	1.24E-05	0.192792163
MMP2	13	13	0	2632.408128	1.08E-05	0.045356774
CASP3	16	13	3	2453.553032	1.09E-05	1
STAT3	8	4	4	2304.658377	1.20E-05	0.127154173
MED19	10	0	10	2185.046503	1.25E-05	0.300285614
MTOR	3	2	1	2159	1.08E-05	0.011717032
MYC	8	7	1	1932.122232	1.08E-05	0.052358916
CTNNB1	11	11	0	1795.197747	1.08E-05	0.049921718
FOXK1	5	0	5	1775.950115	1.16E-05	0.163984286
A2M	4	0	4	1749	1.10E-05	0.001845634
PTK2B	4	1	3	1677.040298	1.15E-05	0.131403327
CCND1	8	8	0	1634.112706	1.08E-05	0.122490816

We used the network measuring techniques of proximity, betweenness, and degree to prune or analyse the gene networks. Table 5 presents the measurement's findings in descending order of closeness and betweenness values. These data allow us to identify the top 20 genes in the network, which are shown in Table 5. The betweenness values in Table 5 demonstrate which genes are critical for gene-gene interactions and can be used to identify therapeutic targets to block disease-causing pathways. AKT1, BCL2, ZEB-ASL, CCNDI, and MMP9, for instance, have the highest betweenness values and are hence essential for the gene-gene interaction.

Table 6: Modularity or Community of some Prostate-implicated genes



GENES	MODULARITY CLASS
AR	9
MYC	9
AKT1	9
CASP9	9
CASP3	9
AURKA	9
FOXK1	9
FGF7	9
LEF1	9
TWIST1	9
MTOR	22
A2M	22
RPS6KB1	22
TSC2	22
EIF4EBP1	22
CPAN2	66
SHARPIN	66
THBS2	66
MAPK8	66
RLN2	66
CCR7	66
HMGB1	66
CCR1	66
MAP2K1	66
MAP2K2	66

Similar to clustering, the modularity values assist in locating gene communities composed of grouped genes. For instance, the communities of AR, MYC, AKT1, CASP9, CASP3, and AURKA all share a modularity class of 9. These communities can be used by researchers to find genes that are indirectly connected and to use experimental data to support that connection.

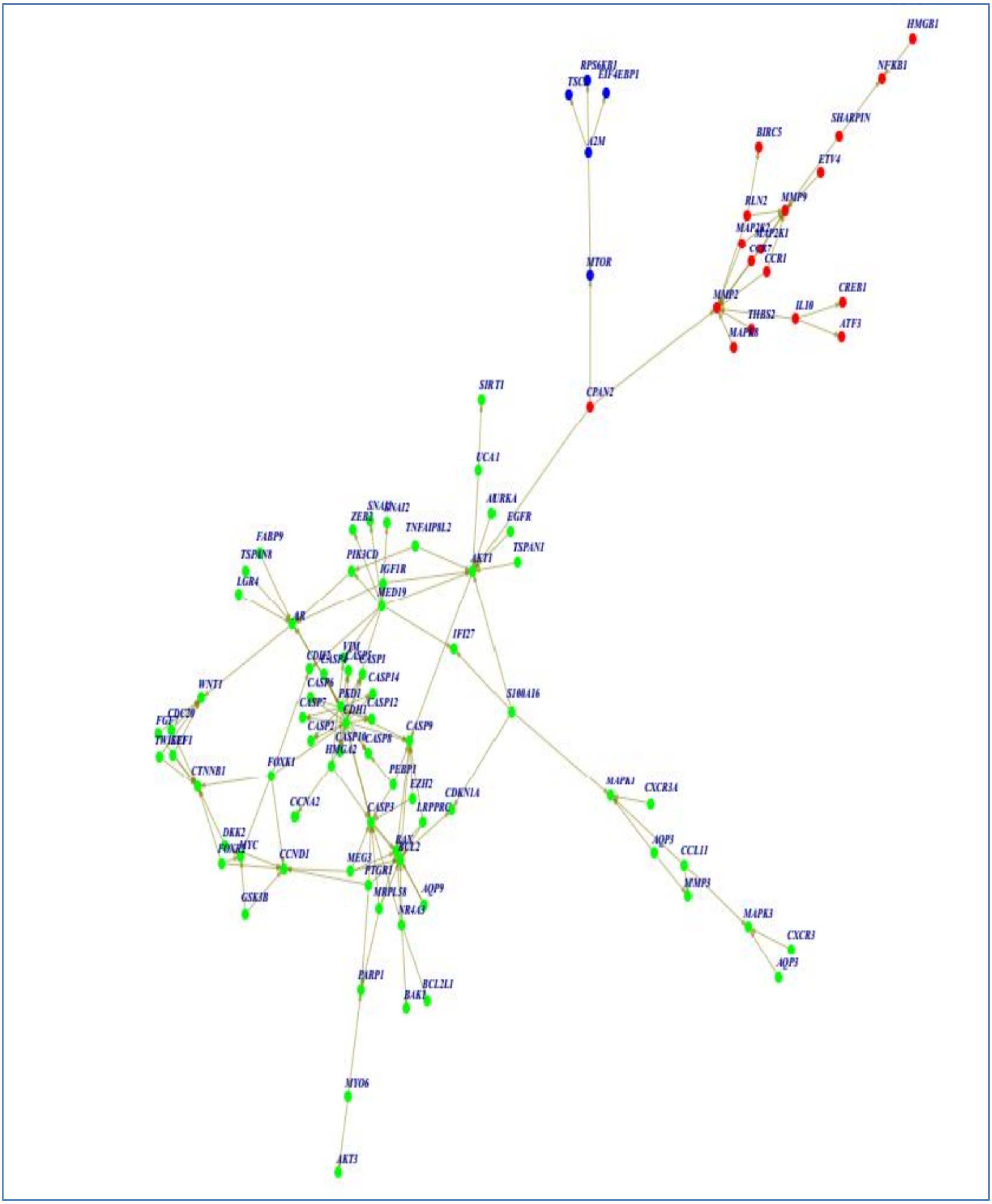


Figure 1: Prostate cancer map created utilizing information on protein-protein interactions. The links between the genes that make up each community or module from our gene network study were verified using the co-expression network. The modules are identified by a certain colour.

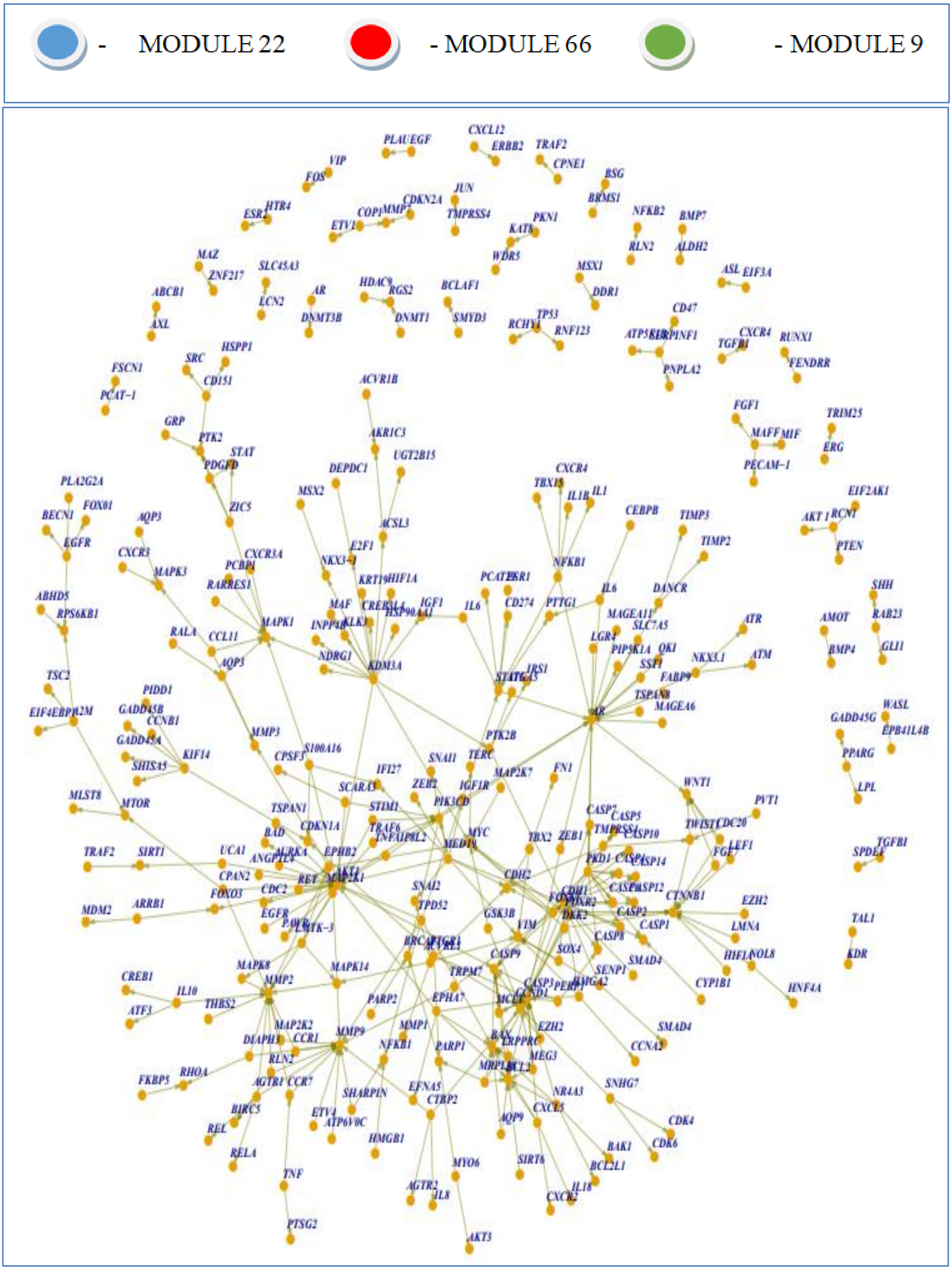


Figure 2: Gene-gene network of all the 296 identified prostate genes. The interactome is mapped out based on their different modules which connote different biological and molecular functions. There are 166 modules in the gene-gene interaction, with only 3 having significant biological annotation.

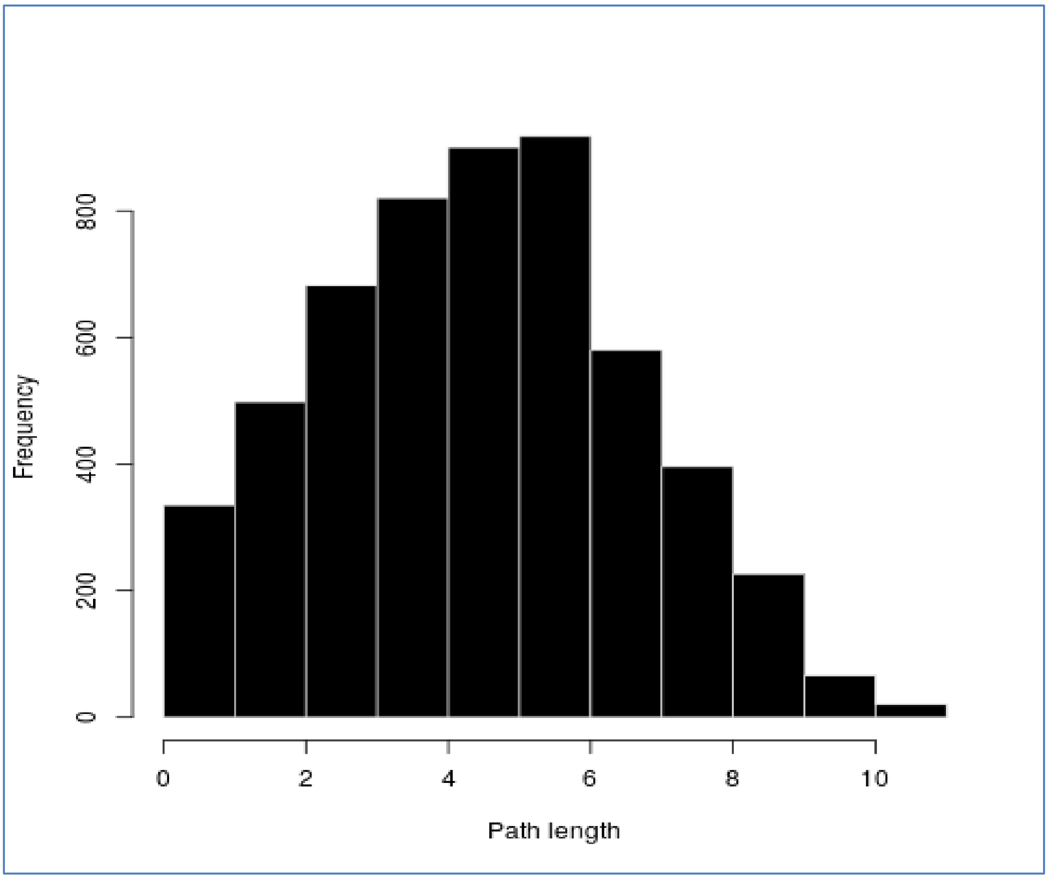


Figure 3: The short path length frequency

4. DISCUSSION

In this present study, prostate-implicated genes were identified utilizing an arrangement of bioinformatics investigation (text mining) to screen key pathways. The analyses found prostate cancer-related genes and their pathways to play important roles in cancer initiation and progression as evidenced by their processes. In known prostate cancer genes, biological processes that are enriched include those that positively regulate transcription from the RNA polymerase II promoter, positively regulate transcription from DNA templates, positively regulate cell proliferation, negatively regulate apoptosis, and respond to drugs. The unchecked activation or acceleration of RNA polymerase II's transcriptional rate can promote the growth of malignant tumours in cancer.

Additionally, the analysis of pathway enrichment in the gene modules (Table1b and 2b) showed that the genes were enriched in activities such as the NF signalling pathway, PI3K-Akt signalling pathway, thyroid hormone signalling pathway, ErbB signalling pathway, HIF-1 signalling pathway, FoxO signalling pathway, signalling pathways regulating pluripotency of stem cells, prostate cancer, mTOR signalling pathway, and AMPK signalling pathway. The molecular processes underlying the growth of malignant tumours have been linked to these pathways (Nosrati *et al.*, 2017). The network analysis shows how the genes interact to affect their biological functions. Some parameters used in describing network analysis include degree, betweenness, and closeness. The values of these parameters for the identification of the genes are shown in (Table 5). The degree can be described as the number of edges (interaction or connection between genes) connected to each gene. Betweenness is a crucial factor that counts the number of smallest routes through a gene. (Claros *et al.*, 2016). The gene with the highest betweenness is very important for effective communication between the nodes(genes), and if removed from the network, it will result in the disconnection of a lot of genes from the network. The node indicates the prostate cancer genes in the network while edges indicate the ties or connections between nodes (Claros *et al.*, 2016). The constructed genes network of all the 305 identified genes, contained a total number of 305 nodes and 373 edges.

Analysis of the signalling pathways connected to these modules revealed that the PI3k-At signalling pathway is abundant in all of them. Fundamental cellular processes like translation, transcription, cell proliferation, development, and survival are regulated by the phosphatidylinositol 3'-kinase (PI3k)-At signalling pathway, which is activated by a variety of cellular stimuli (Arcaro & Guerreiro, 2007). Class 1a and class 1b P13k isoforms are stimulated when growth factors bind to their tyrosine kinase receptor (RTK) or receptors coupled with G proteins, respectively. At the cell membrane, P13K catalyses the synthesis of PIP3 or phosphatidylinositol-3,4,5-triphosphate. In turn, PIP3 functions as a second messenger that aids in Akt activation (He *et al.*, 2021). By phosphorylation substrates involved in apoptosis, synthesis of proteins, metabolism, and cell cycle, Akt can regulate essential functions once it is active.

The histogram of the short path length frequency (fig.3) showed the statistical representation of the network cluster (Yang *et al.*, 1998). In this study, the showed path length, describes the path between each protein in a gene community relative to other communities that made up the clustering. A distance between nodes can be created by varying the width of the path that runs between them in the graph. The shortest route will be taken in many circumstances.

5. CONCLUSION

AKT1, AR, KDM3A, CDH1, MAPK1, and MMP9, which are hub genes, were identified as important genes in this study. Additionally, pathways (such as the TNF signalling pathway, PI3K-Akt signalling pathway, thyroid hormone signalling pathway, ErbB signalling pathway, HIF-1 signalling pathway, and FoxO signalling pathway) that may contribute to the development of prostate cancer were also identified. Thus, targeting these genes may have a significant effect on prostate cancer research. Furthermore, they may also be useful to build an effective computational method for the identification of novel genes related to prostate cancer.

REFERENCES

- Arcaro, A., & Guerreiro, A. S. (2007). The phosphoinositide 3-kinase pathway in human cancer: genetic alterations and therapeutic implications. *Current genomics*, *8*(5), 271-306. https://doi.org/10.2174/138920207782446160
- Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, *5*(2), 101-113. https://doi.org/10.1038/nrg1272
- Claros, I., Cobos, R., & Collazos, C. A. (2015). An approach based on social network analysis applied to a collaborative learning experience. *IEEE Transactions on Learning Technologies*, *9*(2), 190-195. https://doi.org/10.1109/TLT.2015.2453979
- Faro, A., Giordano, D., & Spampinato, C. (2012). Combining literature text mining with microarray data: advances for system biology modeling. *Briefings in bioinformatics*, *13*(1), 61-82. https://doi.org/10.1093/bib/bbr018
- Feldman, R., & Dagan, I. (1995). *Knowledge Discovery in Textual Databases (KDT)*. www.aaai.org
- Ferreira, J. A. (2007). The Benjamini-Hochberg method in the case of discrete test statistics. *The international journal of biostatistics*, *3*(1). https://doi.org/10.2202/1557-4679.1065
- Greene, K. L., Cowan, J. E., Cooperberg, M. R., Meng, M. V., DuChane, J., Carroll, P. R., & Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE) Investigators. (2005). Who is the average patient presenting with prostate cancer? *Urology*, *66*(5), 76-82. https://doi.org/10.1016/j.urology.2005.06.082
- He, Y., Sun, M. M., Zhang, G. G., Yang, J., Chen, K. S., Xu, W. W., & Li, B. (2021). Targeting PI3K/Akt signal transduction for cancer therapy. *Signal transduction and targeted therapy*, *6*(1), 425. https://doi.org/10.1038/s41392-021-00828-5
- Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: a cancer journal for clinicians*, *61*(2), 69-90. https://doi.org/10.3322/caac.20107
- Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, *7*(2), 119-129. https://doi.org/10.1038/nrg1768
- Jurca, G., Addam, O., Aksac, A., Gao, S., Özyer, T., Demetrick, D., & Alhajj, R. (2016). Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends. *BMC research notes*, *9*, 1-35. https://doi.org/10.1186/s13104-016-2023-5
- Jurca, G., Addam, O., Aksac, A., Gao, S., Özyer, T., Demetrick, D., & Alhajj, R. (2016). Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends. *BMC research notes*, *9*, 1-35. https://doi.org/10.1186/s13104-016-2023-5
- Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, *38*(suppl_1), D355-D360. https://doi.org/10.1093/nar/gkp896
- Mishra, A., & Verma, M. (2010). Cancer biomarkers: are we ready for the prime time? *Cancers*, *2*(1), 190-208. https://doi.org/10.3390/cancers2010190
- Nosrati, N., Bakovic, M., & Paliyath, G. (2017). Molecular mechanisms and pathways as targets for cancer prevention and progression with dietary compounds. *International journal of molecular sciences*, *18*(10), 2050. https://doi.org/10.3390/ijms18102050
- Ono, T., & Kuhara, S. (2014). A novel method for gathering and prioritizing disease candidate genes based on construction of a set of disease-related MeSH® terms. *BMC bioinformatics*, *15*(1), 1-12. https://doi.org/10.1186/1471-2105-15-179
- Ploussard, G., & De La Taille, A. (2010). Urine biomarkers in prostate cancer. *Nature Reviews Urology*, *7*(2), 101-109. https://doi.org/10.1038/nrurol.2009.261

- Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., ... & Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database*, *2010*. https://doi.org/10.1093/database/baq020
- Siegel, R., Ma, J., Zou, Z., & Jemal, A. (2014). Cancer statistics, 2014. *CA: a cancer journal for clinicians*, *64*(1), 9-29. https://doi.org/10.3322/caac.21208
- Van't Veer, L. J., & Bernards, R. (2008). Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, *452*(7187), 564-570. https://doi.org/10.1038/nature06915
- Vardakas, K. Z., Tsopanakis, G., Poulopoulou, A., & Falagas, M. E. (2015). An analysis of factors contributing to PubMed's growth. *Journal of Informetrics*, *9*(3), 592-617. https://doi.org/10.1016/j.joi.2015.06.001
- Yang, Y., Pierce, T., & Carbonell, J. (1998, August). A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 28-36).
- Zhang, Y. J., Sun, Y. Z., Gao, X. H., & Qi, R. Q. (2019). Integrated bioinformatic analysis of differentially expressed genes and signaling pathways in plaque psoriasis. *Molecular medicine reports*, *20*(1), 225-235. https://doi.org/10.3892/mmr.2019.10241
- Zhao, M., Li, X., & Qu, H. (2013). EDdb: a web resource for eating disorder and its application to identify an extended adipocytokine signaling pathway related to eating disorder. *Science China Life Sciences*, *56*, 1086-1096. https://doi.org/10.1007/s11427-013-4573-2
- Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., ... & Shen, B. (2013). Biomedical text mining and its applications in cancer research. *Journal of biomedical informatics*, *46*(2), 200-211. https://doi.org/10.1016/j.jbi.2012.10.007