Chapter 3

Biclustering-based biomarker identification

3.1 Introduction

Esophageal cancer (EC) occurs at the food pipe called esophagus. Symptoms often include difficulty in swallowing, enlarged glands around the collarbone and weight loss, a dry cough and possibly coughing up or vomiting blood. Esophageal cancers are extremely aggressive, ranking the eighth most common malignancy and the sixth most frequent cause of cancer death worldwide [164]. Esophageal cancer can be of various types based on the type of cells that are involved such as - (i) Esophageal Squamous Cell Carcinoma (ESCC), (ii) Esophageal Adeno Carcinoma (EAC) and (iii) other rare types e.g. Small cell carcinoma, Sarcoma, Lymphoma, Melanoma and Choriocarcinoma. ESCC is the most widely occurred deadly disease among all other Esophageal cancer types all over the world. The human tissue consists of four types of tissues such as epithelial, connective, muscular and nervous. Epithelial tissue can further be classified based on the shape of the cells, whether squamous, cuboidal, columnar or transitional. Esophageal Carcinoma is developed in the squamous cells which form the surface of the esophagus. These squamous cells are flat and thin epithelial tissue. ESCC occurs most often in the upper and middle portion of the esophagus. Since ESCC is a fatal cancer disease, it is highly essential to identify the genes which cause this cancer. Biomarkers can indicate some biological states or conditions. It might be a single or group of genes which causes cancer. Identification of cancer biomarker is a challenging task. In this work, potential biomarkers are identified for ESCC disease using appropriate biclustering approach followed by network topology analysis, pathway analysis,

and regulatory network analysis. The interestingness of the identified biomarkers has been established through a rigorous process.

3.2 Related Works

From the related survey of methods detecting potential biomarkers for ESCC datasets GSE20347 and GSE23400, it has been observed that no one has carried out an extensive experiment and analysis on ESCC datasets for potential biomarker identification using biclustering approach. Patowary et al. [165] identified ESCC biomarkers using differential and co-expression analysis; no prior knowledge has been integrated. Tung et al.[166] identified biomarkers for ESCC using feature selection and decision tree methods. Combination of differential analysis and graph clustering methods are used to identify prognostic markers of ESCC [167]. BicBioEC [168] identified ESCC biomarkers by using a parallel biclustering method, whereas our method uses a sequential biclustering approach. BicBioEC divides the genes based on their expression values into three categories: upward, downward, and mixed trend while extracting the initial set of biclusters. However, our approach is dependent on multiple existing biclustering algorithms. Unlike BicBioEC, our method uses regulatory network analysis to identify the potential biomarkers. We are motivated to work on this live problem because (i) ESCC is one of the most commonly occurred cancer types in the North East of India, (ii) almost 480,000 new patients are identified every year in India, (iii) early and accurate identification of the potential biomarkers for ESCC could help to decrease the mortality, and (iv) appropriate analysis of gene expression data could help uncover interesting biomarkers for ESCC.

The three major contributions of this work are (a) Highly correlated and enriched bicluster extraction from ESCC microarray data (GSE20347 and GSE23400) using appropriate biclustering techniques, (b) Biological networks (co-expression) construction to enable topological analysis, and (c) Identification of four interesting biomarkers such as IFNGR1, CLIC1, CDK4, and COPS5 which have been established to have high relevance to ESCC.

3.3 Proposed Method

3.3.1 Microarray Data Collection

Gene expression profiles from the GSE20347 (34 samples) and GSE23400 (106 samples) datasets between ESCC samples and matched normal controls are collected from the GEO (http://www.ncbi.nlm.nih.gov/geo/) database. The specification of the dataset is available at Table 3.1.

Table 3.1: A brief description of datasets used for evaluation of proposed framework for ESCC disease

GEO ID	Organism	States	Size	Sample	Summary
GSE20347	Homo sapiens	Normal, Tumor	22278, 34	17 (M), 17 (NM)	Gene expression was examined in tumor and matched normal adjacent tissues from 17 ESCC patients from a high-risk region of China. Affymetrix HG-U133A 2.0 gene expression arrays were performed. Experiment type: Expression profiling by the array.
GSE23400	Homo sapiens	Normal, Tumor	22349, 106	53 (M), 53 (NM)	Gene expression was examined in 53 ESCC samples and 53 matched normal samples. Affymetrix U133A/B chip were performed. Experiment type: Expression profiling by array.

3.3.2 PD_BiBIM framework

The following definitions are useful in describing our method.

Definition 1: Primary gene: For a given disease, a gene is referred to as a primary or elite gene if it is significant w.r.t. the existing benchmark resources.

Some of the prominent genes for ESCC are shown in Table 3.2 and descriptions are given in Table 3.3.

Definition 2: Secondary gene: A gene is referred to as a secondary gene for a given disease w.r.t. a given set of primary genes, if and only if it shows sufficient

evidence to be considered as a causal gene based on a) topological analysis, b) co-expression analysis, c) pathway analysis, and d) established wet-lab results.

Definition 3: Gene Bicluster: A gene bicluster is a subgroup of genes that exhibits similar expression patterns across a subset of experimental conditions or samples.

Definition 4: Enriched Bicluster: An enriched gene bicluster refers to a subgroup of genes that show a statistically significant enrichment with similar expression patterns within a given biological context or experimental condition.

Definition 5: Gene Co-expression Network: A gene co-expression network is a graphical representation of the relationships and connections between genes based on their expression patterns across different samples or experimental conditions. In a gene co-expression network, nodes represent individual genes, and edges represent the strength or degree of correlation between pairs of genes.

Definition 6: Hub Gene: In a GCN, a gene or a set of genes with maximum degree is known as a hub gene(s).

Table 3.2: Primary Genes associated with ESCC

Official Name	**Dataset Gene Id**	**Gene name**
RUNX3	204197_s_at	Runt related transcription factor 3; HGNC:10473
CDH1	201131_s_at	Cadherin 1; HGNC:1748
VIM	201426_s_at	Vimentin; HGNC:12692
WWOX	215526_at	WW Domain Containing Oxidoreductase; HGNC:12799
CTTN	214073_at	Cortactin; HGNC:3338
CCND1	208711_s_at	Cyclin D1; HGNC:1582

This piece of work provides a multi-objective and comprehensive analysis of microarray gene expression data to identify biomarkers for ESCC. To extract the genes associated with the development of ESCC, a topological based analysis has been carried out. The conceptual framework of the method PD_BiBIM has been shown in Figure 3-1. This framework consists of five modules i.e. M1, M2, M3, M4, and M5.

In M1, the preprocessed datasets were used for extraction of different biclusters using different biclustering methods that correspond to tumor and normal samples of ESCC. These biclusters are validated using the p-value. In this method, I have considered six referred genes, considered here as the 'primary genes' related

Table 3.3: Descriptions of Primary Genes associated with ESCC based on existing literature

Gene name	Descriptions
RUNX3	Overexpression of the RUNX3 gene has been observed in esophageal squamous cell carcinoma (ESCC) cells, indicating its significant role in the progression of tumorigenesis in ESCC. [169].
CDH1	The methylation status of the CDH1 gene has been found to be a valuable marker for predicting invasion, metastasis, and prognosis in esophageal squamous cell carcinoma (ESCC). This is attributed to its association with tumor development and the migration of tumor cells. Restoring CDH1 expression could potentially offer a novel approach in the development of cancer therapeutics. [170].
VIM	Patients with esophageal squamous cell carcinoma (ESCC) often exhibit abnormal and inverse changes in the expression levels of the VIM gene. Additionally, the enrichment of VIM in cancer cells has been linked to the invasion and metastasis of ESCC, ultimately affecting the prognosis of ESCC patients [171].
WWOX	The expression of the WWOX gene is significantly reduced in esophageal squamous cell carcinoma (ESCC) tumor tissues. This reduction has been associated with loss of heterozygosity (LOH) and hypermethylation of the gene. These findings indicate that inactivation of the WWOX gene may play a crucial role in the development of ESCC. [172]).
CTTN	The possibility of CTTN as a valuable marker of ESCC. It is found over-expressed and has a significant association with poor prognosis in patients of ESCC [173]).
CCND1	The experimental study evidences the important role of CCDN1 in ESCC [173]). After evaluating CCND1 expression in the tumor tissues from ESCC patients, it is observed that the expression of CCND1 is significantly up-regulated in ESCC tissues compared to the adjacent non-tumorous sample [174].

to Esophagus Squamous Cell Carcinoma as evidenced from multiple benchmark tools, databases, and literary sources and selected a few biclusters based on the presence of primary genes [175] [176] [177] [178] [179] [174] [169].

In M2, a gene co-expression network has been constructed to support topological and other gene-gene associative analysis for each module where such primary genes are present.

Based on the results obtained, in M3, genes with the highest connectivity with primary genes are identified as the 'secondary genes'. Such genes have also been found to follow common genetic pathways with the primary genes in the context of ESCC.

And in M4, gene regulatory network analysis, gene enrichment analysis, pathway, and literature evidence are carried out rigorously on the identified sec-

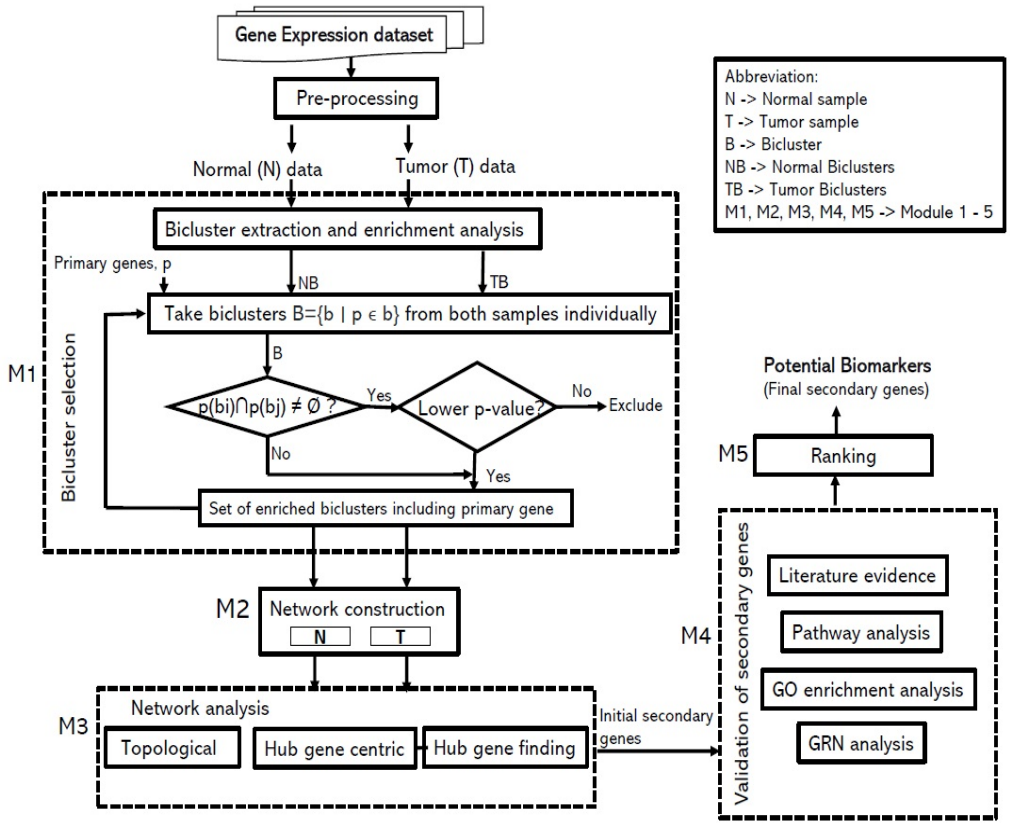


Figure 3-1: PD_BiBIM framework: This framework consists of five modules - M1, M2, M3, M4, and M5.

ondary genes for their establishment as potential biomarkers. Gene Regulatory Network has been constructed for further analysis to investigate the behaviour of those secondary genes (transcription factors) in the normal and tumor samples. Gene ontology (GO) enrichment analysis is performed to investigate the functions of secondary genes and literature mining reveals the supportive evidence for ESCC biomarkers.

Finally, in M5, genes with higher rank are considered as the potential biomarkers based on the highly enriched biclustering results.

Proposition 1: A gene identified using PD_BiBIM framework as potential biomarker is topologically significant.

Proposition 2: A gene identified using PD_BiBIM framework as potential biomarker is biologically significant.

The main step in the PD_BiBIM method is bicluster extraction. The complexity of this step depends on the number of genes and samples in the dataset, as well as the chosen biclustering algorithm. Overall, the computational com-

plexity of the method is likely to be dominated by the most extensive biclustering method used. Finding biclusters can be computationally expensive for large datasets. However, there are a number of ways to reduce the complexity of this step, such as using a more efficient biclustering algorithm or pre-filtering the data.

3.4 Experimental Results

The experiments are carried out on a machine with 32 GB main memory, Intel(R) Xeon(R) CPU E5-1650 v3 processor and 64-bit Windows 10 operating system. Several tools are used such as Matlab15 for mapping gene id, BicAT-plus[180] for finding biclusters, David 6.7/6.8 [181] for finding p-values and gene id conversion, Expander [182] for extracting biclusters from method Samba, GeneAnalytics (S Ben-Ari Fuchs et al. 2016) for finding pathways and related disease name, Cytoscape GeneMANIA [183] for building co-expression network and for topological analysis, R studio [184] for extracting biclusters and Microsoft Excel tools for generating the graph.

Microarray gene expression dataset is represented in terms of a matrix where rows represent genes and columns represent samples or conditions. To identify the correlation patterns of the gene expression matrix in an unsupervised framework with high accuracy, a good number of clustering techniques have been introduced. To eliminate biases of individual algorithm, several biclustering algorithms are applied in two separate ESCC datasets, as shown in Table 3.1. Biclusters or groups of genes are extracted with seven biclustering techniques such as - OPSM [185] xMotif [186], Qubic [187], Bimax [188], IBBIG [189], SAMBA [190], and Plaid [191].

Several human disease databases and tools are used such as Gene MalaCards [178], DISEASES [179]) databases, and literary sources to identify a set of reference genes, closely associated with Esophageal Squamous Cell Carcinoma. These genes are referred to here as 'primary genes' and they are - RUNX3, CDH1, VIM, WWOX, CTTN, and CCND1. Before consideration of these genes, for the subsequent study, an initial experimental study is carried out to observe the variations in expression levels of these primary genes for normal and tumor samples. It is shown in Figure 3-2a– 3-2f . Brief descriptions of these primary genes and their importance in the context of ESCC have been reported in Table 3.3.

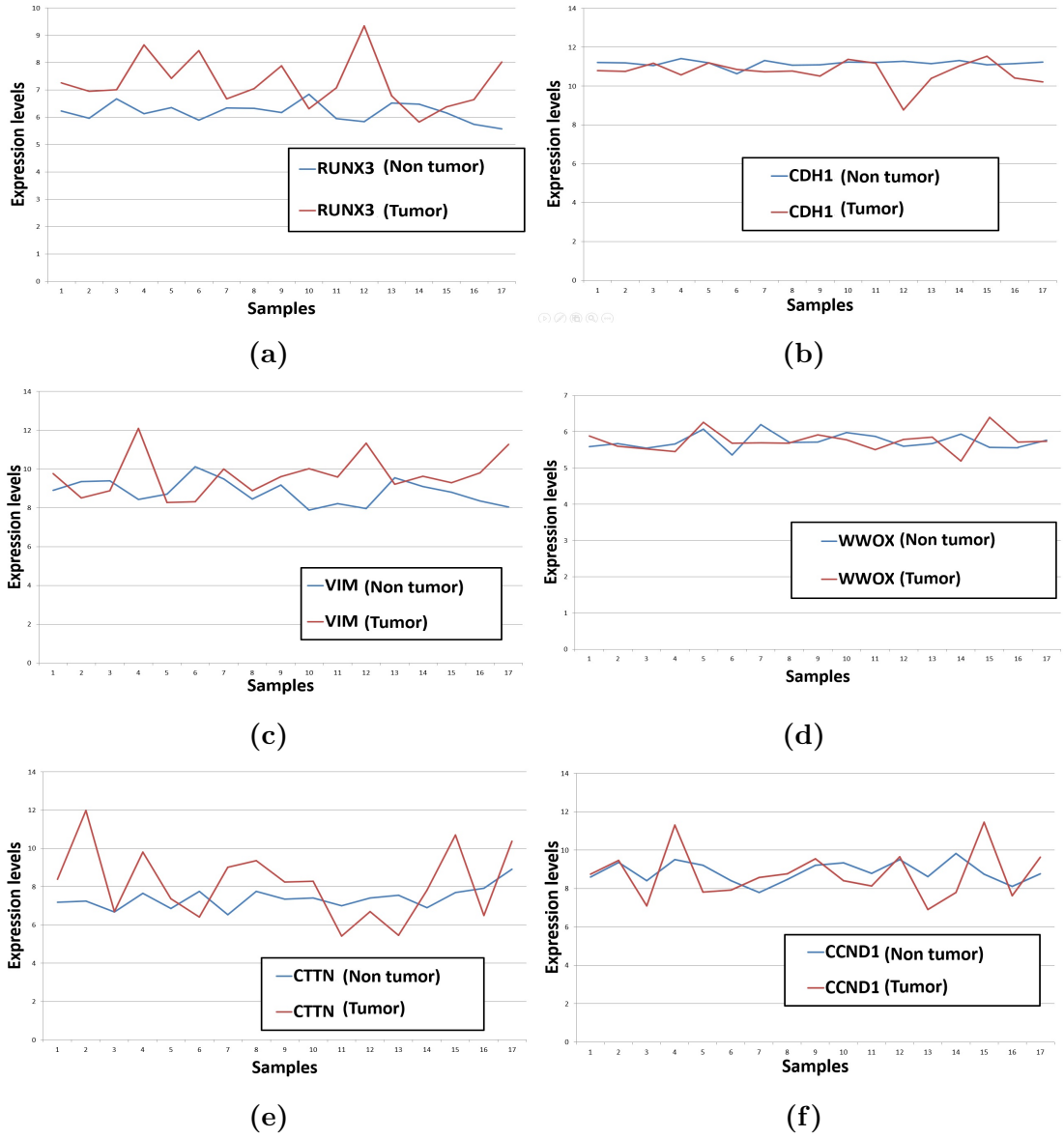


Figure 3-2: The different patterns of gene expression profiles of primary genes across all the normal samples vs tumor samples for dataset GSE20347 (a) RUNX3, (b) CDH1, (c) VIM, (d) WWOX, (e) CTTN, and (f) CCND1.

3.4.1 Pre-processing of dataset

The gene expression dataset - GSE20347 and GSE23400 are pre-processed using the following steps: normalization, non-annoted probe removal, and averaging duplicated gene. GSE20347 and GSE23400 datasets are normalized across all samples by the Robust Multiarray Average (RMA) algorithm. In this work, the samples of GSE20347 and GSE23400 have been subdivided into two separate datasets respectively, one for normal condition and the other one for tumor condition.

3.4.2 Performance evaluation of PD_BiBIM using GSE20347 and GSE23400 datasets

The results obtained by our method for both datasets are reported and analyzed in the subsequent subsections.

3.4.2.1 Bicluster Extraction and their Enrichment Analysis

The pre-processed datasets are used as input for further analysis. We consider ten popular biclustering techniques viz. Bimax [188], IBBIG [189], Qubic [187], Plaid [191], xMotif [186], OPSM [185], SAMBA [190]), CC [192], Spectral [193], and ISA [194] for extraction of a set of unbiased biclusters of high biological significance. It is observed that Bimax, IBBIG, Qubic, Plaid, xMotif, OPSM, and SAMBA are more effective than others, however, the number of biclusters extracted by these algorithms for ESCC dataset is different for each condition. Here, for different datasets GSE20347 or GSE23400, biclustering results are varying in terms of the number of the biclusters and the size of biclusters. Although we experimented with more than seven biclustering techniques, those techniques have been eliminated for subsequent processing which did not generate an adequate number and useful biclusters. Cheng and Church algorithm (CC) finds a single bicluster for GSE20347 and GSE23400 datasets with all genes and conditions rendering the bicluster useless since all genes are present. Similarly, spectral finds one bicluster for both datasets. For enrichment analysis of the bicluster results, we use the p-value of each bicluster. In statistics, p-value [195] is a well-established validation technique and lower p-value for a cluster signifies higher coherency. It means that more genes present in the bicluster with lower p-value are related to the GO term. The p-values for each bicluster given by each algorithm for both the conditions are calculated using DAVID [181] and the lowest p-values are considered to see the performance of different methods. BIMAX can generate biclusters with the lowest p-value among all other methods in this analysis. Though Plaid and Qubic are not performing well for our datasets comparatively with other methods, the p-values of each bicluster with primary genes are investigated for these two methods and reported in Table 3.4 and 3.5 . For normal sample of GSE20347, BIMAX extracts 100 biclusters whereas IBBIG gives 10 biclusters. Similarly, Qubic, Plaid and xMotif extract 100, 1 and 100 biclusters respectively. Similarly, OPSM is also able to extract 100 biclusters, whereas SAMBA generates 13 biclusters. For tumor sample of GSE20347, BIMAX gives 40 biclusters, whereas IBBIG, Qubic, Plaid and xMotif generate 10, 100, 4 and 100 biclusters, respectively. OPSM extracts

Table 3.4: P-values of identified biclusters with primary genes in normal and tumor samples for dataset GSE20347. Note: SN: Serial Number, G: Gene, S: Sample

Biclusters for normal sample of GSE20347						
SN	**Method**	**P-value**	**GO Term**	**Identified Primary Gene**	**Size of BC**	
					G	**S**
1	BIMAX	7.90E-250	Acetylation	CDH1	2841	2
2	BIMAX	2.50E-243	Acetylation	CDH1, CCND1	2844	2
3	BIMAX	2.00E-220	Acetylation	VIM	2514	4
4	IBBIG	1.30E-263	Acetylation	CDH1, VIM	2973	17
5	rQubic	2.00E-08	Alternative splicing	CTTN	343	1
6	rQubic	1.30E-06	Phosphoprotein	CTTN	492	1
7	xMotif	7.50E-227	Phosphoprotein	WWOX	13541	6
8	xMotif	3.10E-18	Phosphoprotein	CDH1	494	6
9	OPSM	1.80E-33	Glycoprotein	VIM	2362	4
10	OPSM	1.90E-63	Phosphoprotein	WWOX	6269	3
11	SAMBA	8.50E-133	Extracellular exosome	CDH1	1134	16
12	SAMBA	1.80E-134	Extracellular exosome	CDH1	1151	16
13	SAMBA	1.10E-109	Extracellular exosome	CDH1	714	5
14	SAMBA	1.40E-133	Extracellular exosome	CDH1	1144	16
15	SAMBA	3.20E-138	Extracellular exosome	CDH1	1162	16
16	SAMBA	3.20E-138	Extracellular exosome	CDH1	1155	16
Biclusters for tumor sample of GSE20347						
1	Bimax	2.20E-274	Acetylation	CTTN	2452	2
2	Bimax	6.20E-264	Acetylation	CDH1, VIM	2472	2
3	IBBIG	5.80E-277	Acetylation	CDH1	2846	17
4	IBBIG	9.20E-43	Acetylation	CCND1	744	3
5	IBBIG	2.30E-26	Phosphoprotein	CTTN	550	3
6	Qubic	1.10E-09	Isopeptide bond	VIM	495	1
7	Plaid	2.30E-110	Phosphoprotein	RUNX3, VIM	3383	2
8	xMotif	2.00E-90	Phosphoprotein	WWOX	2349	6
9	xMotif	1.60E-04	Phosphoprotein	CDH1	59	6
10	OPSM	4.70E-115	Phosphoprotein	RUNX3, WWOX, CCND1, CTTN	13438	2
11	OPSM	1.90E-79	Glycoprotein	RUNX3, WWOX	7446	3
12	OPSM	2.90E-71	Glycoprotein	RUNX3	3500	4
13	SAMBA	1.10E-135	Acetylation	CDH1	1075	12
14	SAMBA	1.30E-128	Acetylation	CDH1	1087	14
15	SAMBA	7.90E-145	Acetylation	CDH1	973	15
16	SAMBA	4.90E-131	Acetylation	CDH1	978	10
17	SAMBA	3.30E-150	Acetylation	CDH1	1009	15
18	SAMBA	4.10E-134	Acetylation	CDH1	1009	10

12 biclusters and SAMBA generates 13 biclusters. Total number of biclusters extracted by each algorithm for GSE20347 and GSE23400 are presented in Table 3.6. The performance evaluation for both the datasets is carried out separately.

To detect the cancer biomarkers, first, those biclusters are considered for an

Table 3.5: P-values of identified biclusters with primary genes for normal and tumor samples for dataset GSE23400. Note: SN: Serial Number, G: Gene, S: Sample

Biclusters for normal sample of GSE23400						
SN	**Method**	**P-value**	**GO Term**	**Identified Primary Gene**	**Size of BC**	
					G	**S**
1	BIMAX	2.43E-137	Acetylation	CDH1, VIM, CTTN	1558	8
2	BIMAX	4.15E-151	Extracellular	CDH1, VIM, CTTN, WWOX	1558	8
3	BIMAX	4.97E-141	Acetylation	CDH1, VIM, CTTN	1623	8
4	Plaid	7.80E-36	Extracellular exosome	VIM	1290	6
5	IBBI G	2.40E-188	Acetylation	CCND1	2112	53
6	rQubic	6.30E-21	Extracellular matrix	VIM	163	10
7	rQubic	3.60E-19	Extracellular matrix	VIM	478	3
8	xMotif	5.60E-32	Phosphoprotein	CTTN	2486	6
9	xMotif	5.90E-07	Phosphoprotein	CCND1	826	6
10	OPSM	1.00E-50	integral component of plasma membrane	RUNX3	4894	4
11	OPSM	4.20E-37	Glycoprotein	WWOX	2609	5
12	SAMBA	3.77E-106	extracellular exosome	CDH1	1255	40
Biclusters for tumor sample of GSE23400						
1	BIMAX	7.39E-155	Acetylation	VIM, CTTN, CDH1, WWOX	1636	8
2	BIMAX	1.81E-156	Acetylation	WWOX, CDH1, VIM	1629	8
3	IBBIG	4.60E-18	Phosphoprotein	CCND1	399	7
4	IBBIG	2.45E-188	Acetylation	VIM, CDH1, WWOX, CTTN	2624	53
5	Plaid	1.50E-03	Cell adhesion	VIM	34	11
6	xMotif	3.50E-10	Phosphoprotein	CTTN	589	6
7	xMotif	1.00E-05	DNA helicase activity	VIM	373	6
8	xMotif	2.50E-05	Extracellular exosome	CCND1	476	6
9	xMotif	1.90E-04	Membrane	RUNX3	295	6
10	xMotif	2.00E-03	Endoplasmic reticulum	CDH1	116	6
11	OPSM	6.20E-29	Glycoprotein	RUNX3	1827	5
12	OPSM	8.60E-53	Glycoprotein	WWOX	4224	4
13	SAMBA	2.40E-121	Acetylation	CDH1	1058	17

algorithm which includes at least one primary gene. Such biclusters are identified and calculated p-values for them in DAVID. If a common primary gene, (say, P1) is present in the different biclusters of same biclustering method, say b1 and b2 then that bicluster is selected based on the lower p-value. P-values of all biclusters with the primary gene for normal and tumor samples for both datasets are listed in Table 3.4- 3.5, respectively. In Table 3.4, it is observed that BC serial no. 1, 2, 4 for normal samples and BC serial no. 1, 2, 6, 7, 10 for tumor sample have

Table 3.6: Number of Biclusters extracted by different reported biclustering methods for GSE20347 and GSE23400 datasets. Note: T: Tumor sample, N: Adjacent normal sample.

Biclustering method	**Total number of Biclusters**			
	GSE20347		**GSE23400**	
	N	**T**	**N**	**T**
BIMAX	100	40	100	100
IBBIG	10	10	10	10
OPSM	100	12	13	13
Plaid	1	4	3	3
Qubic	100	100	10	10
xMotif	100	100	91	100
SAMBA	13	13	15	15

sample size less than 3. In this case, the biclusters with sample size at least 3 and comparatively lower p-value are considered. Therefore, these biclusters are not included for further analysis of GSE20347 dataset. From N sample, BC serial number 3, 4, 7, 8, 9, 10, and 15 are considered and from T samples BC number 3, 4, 5, 8, 11, 12, 17 are considered for network construction (Table 3.4). From Table 3.5, BC no. 2, 4, 5, 6, 8, 10, 11, 12 for N sample and BC no. 1, 2, 4, 12, 13 are considered for further investigation to identify potential biomarkers for GSE23400.

3.4.2.2 Co-expression Network Construction and Topological analysis

By considering the selected biclusters as modules, different co-expression networks are constructed using Cytoscape plugin GeneMANIA. Here, genes are considered as nodes and each edge represents the co-expression between the genes. From the co-expression network, the degree and weight of each node are calculated and compared. Highly connected nodes (hubs) in biological networks are topologically significant to the structure of the network [196]. Highly connected nodes are statistically important and functionality more relevant than other nodes in a network [197]. This analysis aims to identify topological insights of the biclusters obtained from the normal and tumor samples of GSE20347 and GSE23400 datasets. The statistics based on normal/tumor sample topological analysis is discussed below and reported in Table 3.7-3.8.

(A) Topological analysis on GSE20347: For co-expression network analysis, selected biclusters of considered biclustering algorithms are given as input to GeneMANIA to build the gene co-expression networks. The gene with the highest connectivity in a selected bicluster and is directly or indirectly connected to at

Table 3.7: Degree analysis of primary genes and their respective hub genes (secondary genes) for the dataset GSE20347 along with association type (Direct (D) or Indirect (I)) between the primary gene and hub genes are also mentioned. Note: P: Primary Gene, N: Normal sample, T: Tumor sample, NF: Not found, AT(H,T): Association Type between H and T

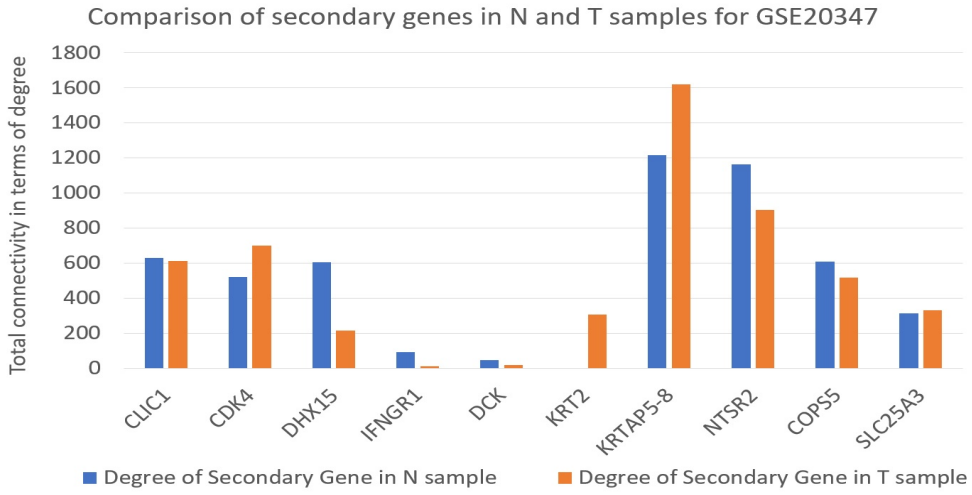
Method	P	Degree(P)		Hub gene (H)		Degree(H)		AT(H,T)	
		N	T	N	T	N	T	N	T
Bimax	VIM	315	181	COPS5	CDK4	608	558	I	I
IBBIG	CDH1	325	264	CLIC1	CDK4	629	699	I	I
IBBIG	CTTN	36	Nf	DRG1	Nf	99	-	I	-
xMotif	CDH1	50	3	IFNGR1	DCK	93	21	D	I
xMotif	WWOX	391	91	NTSR2	KRT2	1818	307	I	I
OPSM	VIM	247	Nil	KRTAP5-8	Nf	686	-	I	-
OPSM	WWOX	190	244	KRTAP5-8	KRTAP5-8	1216	1620	I	I
OPSM	RUNX3	203	Nf	NTSR2	Nf	905	-	I	-
Samba	CDH1	145	84	SLC25A3	SLC25A3	316	320	I	I

Table 3.8: Degree analysis of primary genes and their respective hub genes (secondary genes) for the dataset GSE23400 along with association type (Direct (D) or Indirect (I)) between the primary gene and hub genes are also mentioned. Note: P: Primary Gene, N: Normal sample, T: Tumor sample, NF: Not found, AT(H,T): Association Type between H and T

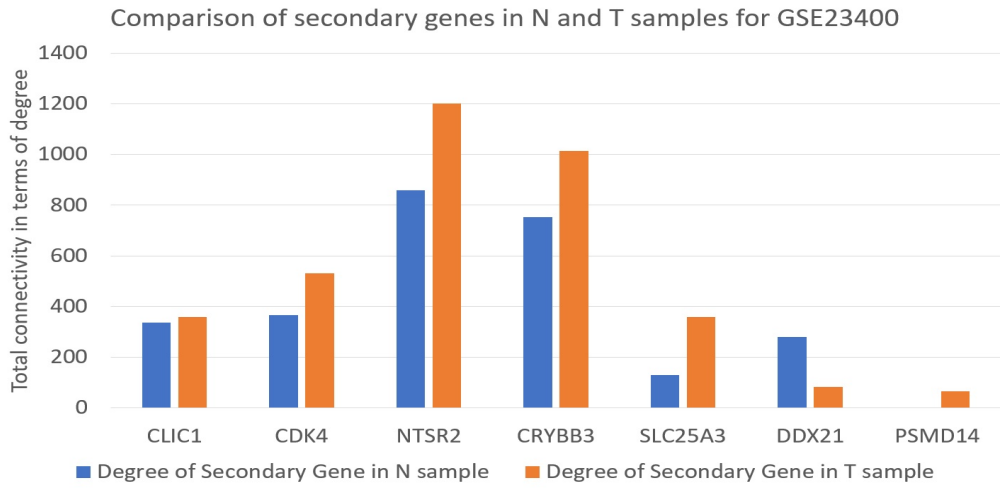
Method	P	Degree(P)		Hub Gene (H)		Degree(H)		AT(H,T)	
		N	T	N	T	N	T	N	T
Bimax	WWOX	45	40	CLIC1	SLC25A3	337	358	I	I
Bimax	VIM	203	207	CLIC1	SLC25A3	337	358	I	I
Bimax	CTTN	110	100	CLIC1	SLC25A3	337	358	I	I
IBBIG	WWOX	61	70	CLIC1	CDK4	421	531	I	I
IBBIG	VIM	252	298	CLIC1	CDK4	421	531	D	I
IBBIG	CTTN	139	158	CLIC1	CDK4	421	531	D	I
Plaid	VIM	115	Not found	RAB25	Not found	349	-	I	-
xMotif	CTTN	208	20	DDX21	PSMD14	279	66	I	I
OPSM	RUNX3	232	179	NTSR2	NTSR2	1346	1176	I	I
OPSM	WWOX	107	157	CRYBB3	NTSR2	753	1202	I	I
Samba	CDH1	134	145	CLIC1	CDK4	433	510	I	I

least one primary gene (P) will be the hub gene or secondary gene (S) for our study. It is found that WWOX is directly connected to IFNGR1 in dataset GSE20347 and in other case no genes are found to be directly connected with their respective primary genes. In Table 3.7, the present primary gene as its corresponding hub-gene (S: secondary genes) of each biclustering methods are shown for both the samples of GSE20347. In this analysis, some primary genes have been found which are not present in the particular biclusters in both the samples and therefore, its corresponding gene DRG1 from Table 3.7 is not considered for subsequent analysis. From this topological analysis, we are considering COPS5, CLIC1, CDK4, IFNGR1, NTSR2, KRTAP5-8, and SLC25A3 hub genes for further investigation.

(B) Topological analysis of GSE23400: CDK4 has been found to have a direct association with two primary genes - VIM and CTTN in the co-expression network of N sample. However, except this gene, others are indirectly connected with reported primary genes. Though Plaid was not performing well, it is considered because of one enriched bicluster with primary gene VIM in N sample, but it results in absence of the primary gene in T sample of GSE23400 dataset. Therefore, we are excluding RAB25 gene reported in Table 3.8. From this analysis, CLIC1, SLC25A3, CDK4, DDX21, PSMD14, NTSR2, and CRYBB3 hub genes are considered for subsequent study.



(a)



(b)

Figure 3-3: Variation of connectivities for each secondary gene found in matched Normal (N) vs Tumor (T) samples for the dataset (a) GSE20347, (b) GSE23400. In most cases, secondary gene's connectivities are higher in T sample than that of N sample.

After successfully finding a list of highly connected genes (nodes) for normal

and tumor samples for both the datasets, we consider them as the initial secondary genes according to their topological behaviour. From both datasets, 4 hub genes are found common viz. CDK4, CLIC1, NTSR2, and SLC25A3. The initial secondary genes selected from the topological analysis are CDK4, CLIC1, NTSR2, SLC25A3, COPS5, IFNGR1, DCK, KRT2, KRTAP5-8, DDX21, PSMD14, and CRYBB3. Thus, I have found a total of 12 initial secondary genes from dataset GSE20347 and GSE23400 considering both the samples. Further, I have searched for any secondary gene for the biclusters generated by each biclustering algorithm in an alternative tissue and if found then the degree of that secondary gene are compared for both the tissues (normal and tumor). Taking all the identified secondary genes for corresponding biclusters, a comparison in terms of degrees for both normal and tumor samples for both datasets are presented in Figure 3-3. Here, it is observed that the connectivity of the majority of secondary genes is increased in tumor tissue. Gene KRTAP5-8 shows the highest connectivity in GSE20347 dataset but in GSE23400 dataset, gene NTSR2 has the highest connectivity. From Figure 3-3, it is observed that CDK4, CLIC1, NTSR2, KRTAP5-8, and SLC25A3 are found as hub genes in both tissues of many biclusters from several biclustering methods. Hence, CDK4, CLIC1, NTSR2, KRTAP5-8, and SLC25A3 might play an important role in causing ESCC. Since, a noticeable variation of degrees are observed for remaining secondary genes presented in Figure 3-3a– 3-3b, so these genes might also play a role.

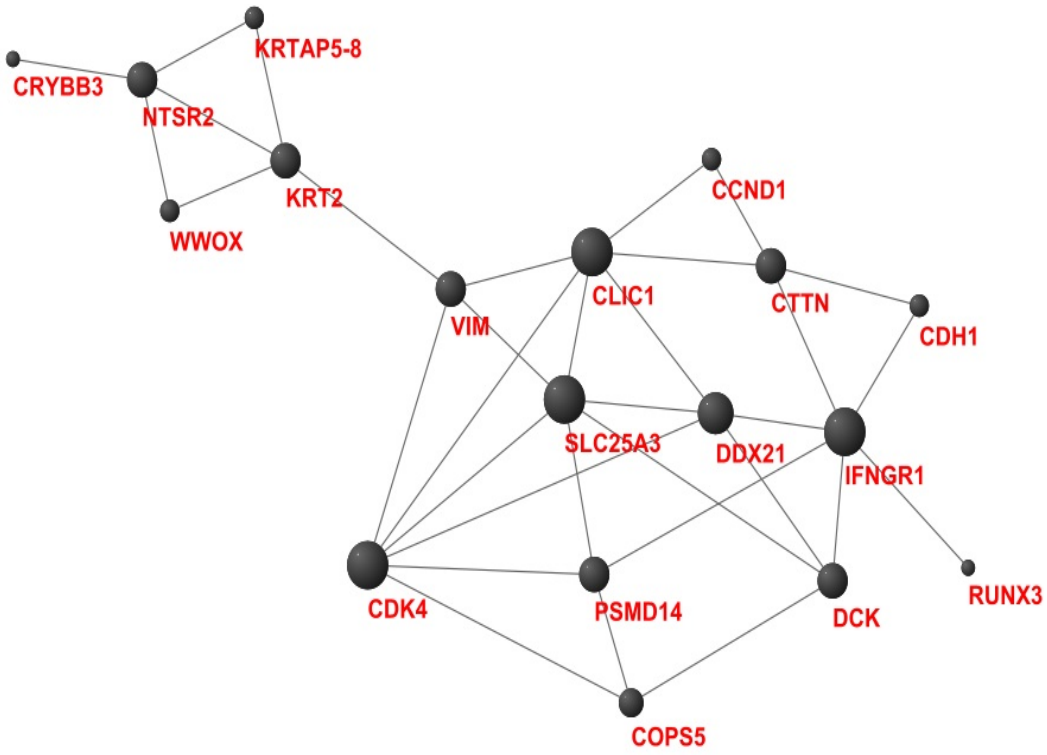


Figure 3-4: Biological interactions between primary and secondary genes

3.4.3 Biological network analysis for Primary and Secondary genes

Considering 12 secondary genes and 6 primary genes as nodes, a biological network is constructed using GeneMANIA as shown in Figure 3-4. Here, secondary genes CDK4, SLC25A3, CLIC1, IFNGR1, NTSR2, and KRT2 are directly connected with primary genes, on the other hand, DCK, COPS5, DDX21, KRTAP5-8, CRYBB3, and PSMD14 are indirectly connected to primary genes. From this scenario, we observe that CDK4, SLC25A3, CLIC1, IFNGR1, NTSR2, and KRT2 can be assumed to have might play a significant role in ESCC disease in progression.

Considering secondary and primary genes as proteins, a PPI network has been constructed in STRING [198]) and shown in Figure 3-5. The edges of this PPI indicate both functional and physical protein associations based on active interaction sources text mining, experiments, databases, coexpression, neighborhood, and Gene Fusion. The line thickness indicates the strength of data support and minimum required interaction score is 0.15. From this network, it is evidenced that CDK4, DCK, IFNGR1, DDX21, COPS5, CLIC1, CRYBB3 and PSMD14 are directly linked with primary genes - CCND1, RUNX3, and it is found that two genes CDK4 and PSMD14 are directly connected with primary genes.

3.4.4 Pathway analysis

Pathway analysis is carried out for the selected 12 secondary genes viz. CDK4, CLIC1, NTSR2, SLC25A3, COPS5, IFNGR1, DCK, KRT2, KRTAP5-8, DDX21, PSMD14, and CRYBB3 using GeneAnalytics tool as well as DAVID and for the six primary genes together. In this pathway analysis, six secondary genes out of 12 are found, which follow the same cancer pathway as the primary genes, shown in Table 3.9. They are- CDK4, DCK, COPS5, IFNGR1, SLC25A3, and KRT2. Again, using GeneAnalytics tool, it is also investigated which primary, as well as secondary genes, are associated with cancer disease. The list of cancer diseases caused by primary and secondary genes are presented in Table 3.10. It is found that the secondary gene CDK4 is directly connected to Esophageal cancer. KRT2 and DCK are related to squamous cell carcinoma and pancreatic cancer, respectively. KRTAP5-8 is responsible for renal oncocytoma. The Tissues and Cells in GeneAnalytics results gave a total of four genes which are matched to the esophagus entity type with gene matched score 0.46. These are- IFNGR1, SLC25A3, COPS5, and KRT2.

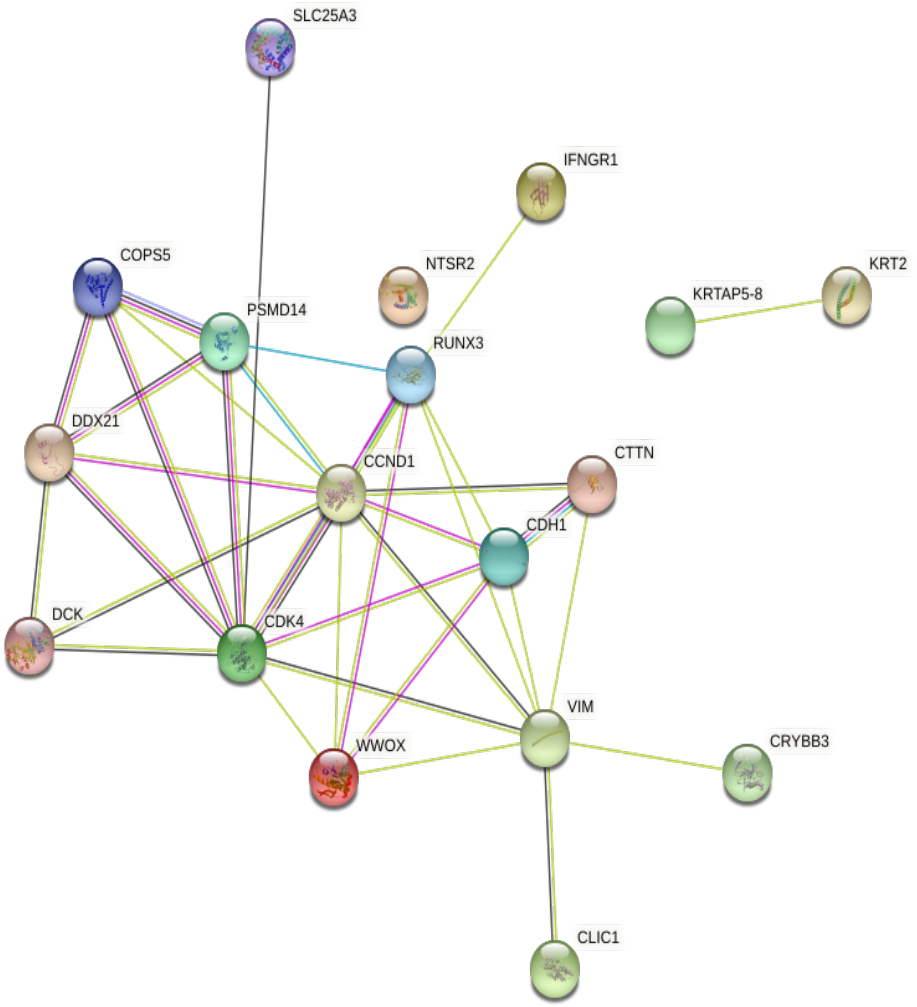


Figure 3-5: PPI network among primary and secondary genes

3.4.5 GO enrichment analysis

GO enrichment analysis is performed using DAVID on the primary and secondary genes together. Altogether 17 terms showed enriched GO term except for gene NTSR2 as reported in Table 3.11. Obtained percentage of enrichment in Biological Process (BP), Cellular Component (CP) and Molecular Function (MF) are 94.4, 94.4, and 100, respectively. As reported in Table 3.11, most genes in GO analysis are related with the biological process of regulation of protein binding, cell adhesion, positive regulation of cell cycle arrest, extracellular exosome etc.

Table 3.9: Common Pathway shared by Primary and Secondary genes

Category	Term	P-Value	Genes
KEGG	hsa05219:Bladder cancer	9.49E-04	CCND1, CDH1, CDK4
KEGG	hsa05218:Melanoma	0.002825	CCND1, CDH1, CDK4
KEGG	hsa05162:Measles	0.009625	CCND1, CDH1, CDK4
KEGG	hsa05216:Thyroid cancer	0.033249	CCND1, CDH1
KEGG	hsa05130:Pathogenic Escherichia coli infection	0.057823	CTTN, CDH1
KEGG	hsa05213:Endometrial cancer	0.058927	CCND1, CDH1
KEGG	hsa05223:Non-small cell lung cancer	0.063331	CCND1, CDH1
KEGG	hsa05200:Pathways in cancer	0.072512	CCND1, CDH1, CDK4
KEGG	hsa05214:Glioma	0.073175	CCND1, CDK4
KEGG	hsa05212:Pancreatic cancer	0.073175	CCND1, CDK4
KEGG	hsa04115:p53 signalling pathway	0.07535	CCND1, CDK4
KEGG	hsa05220:Chronic myeloid leukemia	0.080769	CCND1, CDK4
KEGG	hsa05100:Bacterial invasion of epithelial cells	0.087234	CTTN, CDK4
KEGG	hsa05222:Small cell lung cancer	0.094727	CCND1, CDK4
KEGG	hsa04530:Tight junction	0.096857	CTTN, CDK4
Biosystem	Colorectal cancer tumor	-	CCND1, CDH1, IFNGR1
Qiagen	Retinoblastoma(RB) in cancer	-	CDK4, DCK, CCND1
Biosystem	Allograft rejection	-	CCND1, VIM, IFNGR1
Biosystem	Cell cycle regulation	-	CCND1, COPS5, CDK4
Biosystem	C-MYB transcription factor network	-	CCND1, SLC25A3
Biosystem	Cytoskeleton remodelling neurofilaments	-	VIM, KRT2
Biosystem	Gastric cancer	-	CDH1, CCND1, IFNGR1, CDK4

3.4.6 Gene Regulatory Network (GRN) analysis

Only one secondary gene viz. COPS5 and another two primary genes viz. RUNX3 and WWOX are found as transcription factors. These three primary genes are considered as regulators and combine all primary and secondary genes as target

Table 3.10: List of Primary and Secondary genes associated with Cancer Disease

Cancer Disease Name	Matched Genes	
	Primary Gene	Secondary Gene
Esophageal cancer	WWOX,CDH1,CCND1	CDK4
Ovarian cancer	CDH1,CCND1,CTTN	CDK4
Breast cancer	CDH1,CCND1,CTTN,WWOX	CDK4
Myeloma, multiple	CCND1	CDK4
Cell Type cancer	CDH1,CCND1,WWOX	CDK4
Endometrial cancer	CDH1,CCND1	CDK4
Colorectal cancer	CCND1,CDH1, RUNX3	CDK4
Hepatocellular Carcinoma	CCND1,CTTN,CDH1	CDK4
Squamous cell carcinoma , head and neck	CCND1,CDH1,CTTN	CDK4
Cervical squamous cell carcinoma	CDH1,CCND1	KRT2
Astrocytoma	CCND1,VIM	CDK4
Mantle cell lymphoma	CCND1	CDK4
Spindle cell lipoma	VIM	CDK4
Lung cancer	CCND1,CDH1,	CDK4, DCK
Pancreatic cancer	CCND1,CDH1	CDK4, DCK
Renal Oncoytoma	CCND1	KRTAP5-8

genes. This experiment is carried out to study their associations using GENIE3 [199]. GENIE3 constructs a regulatory network by determining the tree-based ensemble methods. Various regulatory associations and their respective weights for both the samples are reported in Table 3.12- 3.13. Considering weight greater than 0.5 as a strong regulatory association between a regulator and a target gene for tumor sample. It can be observed that among the secondary genes viz. CTTN, DCK, CCND1 and RUNX3, as shown in Table 3.12, three are primary genes and DCK is a secondary gene which has been found to have a strong association with ESCC. Further, for all these four target genes, the weights are found to increase while in progression from normal tissue to tumor sample, as shown in Table 3.13.

3.4.7 Literature evidence

Here, all the secondary genes are briefly described and mentioned if their associations are found with ESCC by other authors.

CDK4: Cyclin-Dependent Kinase 4 is a Protein-Coding gene which is responsible for the phosphorylation of the retinoblastoma gene product. Diseases associated with CDK4 include Melanoma, Cutaneous Malignant 3 and Dedifferen-

Table 3.11: GO enrichment analysis of primary and secondary genes. Note: GC: GOterm Category

GC	Term	Genes
BP	GO:0045109intermediate filament organization	KRT2, VIM
BP	GO:0010971positive regulation of G2/M transition of miotic cell cycle	CCND1, CDK4
BP	GO:0071157negative regulation of cell cycle arrest	CCND1, CDK4
BP	GO:0045787positive regulation cycle	CCND1, CDK4
BP	GO:0042493response to drug signalling pathway	CCND1, CDK4, CDH1
BP	GO:0030178negative regulation	WWOX, CCND1
BP	GO:0006468protein phosphorylation	CCND1, CDK4, RUNX3
BP	GO:0009636response to toxic substance	CDH1, CDK4
BP	GO:0006366transcription from RNA polymerase II promoter	COPS5, DDX21, RUNX3
BP	GO:0000082G1/S transition	CCND1, CDK4
BP	GO:0001649osteoblast differentiation	WWOX, DDX21
BP	GO:0045471response to ethanol	CCND1, PSMD14
BP	GO:0009615response to virus	IFNGR1, DDX21
MF	GO:0016538cyclin-dependent protein kinase regulator activity	CCND1, CDK4
MF	GO:0032403protein serine/threonine	SLC25A3, CCND1, CDK4
MF	GO:0005212structural constituent of eye lens	CRYBB3, VIM
MF	GO:0098641cadherin binding involved in cell- cell adhesion	CTTN, CDH1, CLIC1
MF	GO:0005515protein binding	WWOX, PSMD14, IFNGR1, KRT2, CRYBB3, DDX21, RUNX3, COPS5, CTTN, CCND1, CDH1, CDK4, VIM, CLIC1
MF	GO:0003725double-stranded RNA binding	DDX21, VIM
MF	GO:0001948glycoprotein binding	CDH1, VIM
MF	GO:0008237metallopeptidase activity	COPS5, PSMD14
CC	GO:0005634nucleus	WWOX, SLC25A3, COPS5, CCND1, PSMD14, CDK4, KRT2, DDX21, RUNX3, DCK, CLIC1
CC	GO:0016020membrane	SLC25A3, CCND1, IFNGR1, CDH1, KRT2, DDX21, CLIC1
CC	GO:0000307cyclin-dependent protein kinase holoenzyme complex	CCND1, CDK4
CC	GO:0048471perinuclear region of cytoplasm	COPS5, CDH1, CDK4, CLIC1
CC	GO:0070062extracellular exosome	CTTN, PSMD14, CDH1, KRT2, VIM, CLIC1
CC	GO:0005913cell-cell adherens	CTTN, CDH1, CLIC1
CC	GO:0045111intermediate filament	KRT2, VIM
CC	GO:0005925focal adhesion	CTTN, CDH1, VIM
CC	GO:0005737cytoplasm	WWOX, COPS5, CTTN, CCND1, CDH1, KRT2, VIM, RUNX3, CLIC1
CC	GO:0045095keratin filament	KRT2, KRTAP5-8
CC	GO:0005923bicellular tight	CCND1, CDK4
CC	GO:0005882intermediate filament	KRT2, VIM

Table 3.12: GRN network statistics for normal sample

Regulatory Gene	Target Gene	Weight
COPS5	CTTN	0.689236165
	DCK	0.624834828
	CCND1	0.580179903
	RUNX3	0.561231

tiated Liposarcoma [200]. CDK4 is involved in cancer and HGF/MET pathways and are closely associated with a variety of tumors [201]. CDK4 is overexpressed in ESCC tissues compared with their paired adjacent non-neoplastic tissues. In this experiment, it is also found that miR-1 directly regulates CDK4. miR-1 suppresses the growth of ESCC through the downregulation of CDK4 expression [201]. CDK4 is a signaling molecule which acts as a master regulator [200]. CDK4

Table 3.13: GRN network statistics for tumor sample

Regulatory Gene	Target Gene	Weight
COPS5	RUNX3	0.471067992
	DCK	0.349884608
	CCND1	0.258790647
	CTTN	0.102428644

has a significant role in the EGFR inhibition process in esophageal squamous cell carcinoma [202].

COPS5: COP9 Signalosome Subunit 5 is a Protein-Coding gene. Diseases associated with COPS5 include Xeroderma Pigmentosum, Complementation Group E [200]. COPS5 has the direct interaction with CDK4 [203]. COPS5 has been discovered to be a predictive biomarker for multiple types of cancers [204]. Research has found that the expression of COPS5 is notably increased in both SOC cells and tissues when compared to control tissues.[204].

KRT2: Keratin 2 is a Protein-Coding gene. Diseases associated with KRT2 include Ichthyosis Bullosa Of Siemens and Exfoliative Ichthyosis [200]. Hepatocarcinoma is the most common primary liver tumor. In accordance with mRNA level, KRT2 is found to be differentially expressed which might be a critical candidate associated with LNM of hepatocarcinoma [205]). KRT2 was found as one of the up-regulated genes among selected 41 up-regulated genes associated with the process of esophageal carcinogenesis which is experimented on mouse cell [206].

IFNGR1: IFN-g and IFNGR1 were found correlated with the progression of ESCC. The downregulation of IFNGR1 was tightly associated with clinicopathologic features of ESCCs, which suggested that the loss of IFNGR1 was involved in the development and progression of ESCCs [207].

CLIC1: Chloride Intracellular Channel 1 acts as a switch among tumor behaviours in human esophageal squamous cell carcinoma [208].

DCK: Deoxycytidine Kinase is a Protein-Coding gene required for the phosphorylation of several deoxyribonucleosides and their nucleoside analogs. Diseases associated with DCK include Periampullary Adenocarcinoma and Purine Nucleoside Phosphorylase Deficiency [200]. Prognosis of the patients with a high DCK expression suggests DCK expression is a prognostic factor of the ESCC patients [209].

SLC25A3: Diseases associated with SLC25A3 include Mitochondrial Phos-

phate Carrier Deficiency and Wheat Allergy [200].

KRTAP5-8: The related pathways of Keratin-Associated Protein 5-8 are Keratinization and Developmental Biology [200]. The increased expression of KRTAP5-8 is associated with progression of Esophageal Squamous Dysplasia [210].

NTSR2: Neurotensin Receptor 2 is a Protein-Coding gene [200]. NTSR2 is overexpressed in malignant human B lymphocytes [211].

DDX21: DDX21 interacts with the mitotic regulator PP1 and oncoprotein DEK. Additionally, upregulated DDX21 has been found to promote tumorigenesis in breast cancer by phosphory [212].

PSMD14: The epithelial-mesenchymal transition (EMT) transcription factor SNAIL is associated with distant metastasis and poor prognosis of ESCC patients. Deubiquitinating enzyme PSMD14 promotes tumor metastasis through stabilizing SNAIL in human eESCC [213].

CRYBB3: Crystall in Beta B3 is a Protein Coding gene. Cataract 22, Multiple Types and Cataract 24 are the diseases associated with this gene. Gene Ontology (GO) annotations related to this gene include structural constituent of eye [200].

3.5 Discussion

The provided experimental results offer a comprehensive analysis of gene expression data in the context of ESCC. Seven biclustering techniques are applied to identify correlation patterns in gene expression datasets. The choice of these algorithms was based on their effectiveness in extracting meaningful biclusters. It is observed that there is a variations in the number and size of biclusters across different conditions and datasets. Six genes (RUNX3, CDH1, VIM, WWOX, CTTN, and CCND1) are defined as 'primary genes' based on their association with ESCC. An initial experimental study is conducted to observe variations in the expression levels of these primary genes for normal and tumor samples.

From the topological analysis of different modules based on biclusters given by biclustering algorithms, a total of 12 genes are found as secondary genes. These 12 secondary genes are CDK4, CLIC1, NTSR2, SLC25A3, COPS5, IFNGR1, DCK, KRT2, KRTAP5-8, DDX21, PSMD14, and CRYBB3. After investigating

the network topology for both the samples (normal and tumor) w.r.t. these 12 genes, some genes are filtered out. In biological (co-expression) network analysis among the primary and secondary genes, as shown in Figure 3-4, CDK4, SLC25A3, IFNGR1, CLIC1, and NTSR2 secondary genes are directly connected to the primary genes. Considering the degrees of the nodes, those genes with higher degrees are selected. For example, IFNGR1, CLIC1, and CDK4 will get more weightage than that of remaining secondary genes due to their higher degrees. In PPI network analysis as shown in Figure 3-5, importance is given to those secondary genes which are directly interacted with any primary gene. Two secondary genes CDK4 and PSMD14 are found.

From pathway analysis, CDK4, DCK, COPS5, IFNGR1, SLC25A3, CLIC1 and KRT2 are considered as the most significant biomarkers for ESCC. In gene enrichment analysis, NTSR2 gene is not found. Gene interaction of NTSR2 with primary genes in biological network and PPI network are found satisfactory. From the literature review,it is found that some genes including CDK4, DCK, KRT2, and KRTAP5-8 are related to cancer. In GRN analysis, regulator COPS5 regulates other primary genes. The weights indicated strong regulatory associations, and an increase in weights from normal to tumor samples is observed. In Table 3.14, the average scores of secondary genes (biomarkers) are reported to rank those biomarkers. In this table, if a secondary gene is found directly associated, it is indicated as 1 else 0. Based on the evidence reported in Table 3.14, the higher-ranked biomarkers obtained (average score is greater than 0.5) from all the experiments done in this analysis are IFNGR1, CLIC1, CDK4, and COPS5.

The proposed method integrates various computational approaches, ranging from biclustering and network analysis to pathway and regulatory network exploration. The combination of these methods provides a comprehensive understanding of gene expression patterns and their regulatory mechanisms in ESCC. The identified primary and secondary genes, their associations, and their roles in pathways and networks suggest their potential significance in the context of ESCC. This information could guide further experimental validations and contribute to understanding the molecular mechanisms underlying ESCC. From this analysis, it is observed that biclustering may not be suitable for all types of dataset. It is particularly effective when there are local patterns within subsets of the data. For some datasets, traditional clustering may be more appropriate. Again, biclustering needs to optimize both row and column clusters simultaneously, leading to increased computational overhead.

Table 3.14: Biomarker ranking: Secondary genes are ranked based on their average score. Note: 1 means yes, 0 means No; S: Secondary gene, P: Primary gene.

Gene	**Direct associativity of S with P in**						**Literature evidence of associativity with ESCC**	**Average Score**	**Rank**
	Bicluster	**PPI**	**Biological interaction network**	**Pathway analysis**	**GO enrichment analysis**	**GRN**			
IFNGR1	1	1	1	1	1	0	1	0.857143	1
CLIC1	1	1	1	0	1	0	1	0.714286	2
CDK4	0	1	1	1	1	0	1	0.714286	3
COPS5	0	1	0	1	1	1	1	0.714286	4
CRYBB3	0	1	0	1	1	0	0	0.428571	5
DCK	0	1	0	0	1	0	1	0.428571	6
KRT2	0	0	1	1	1	0	0	0.428571	7
SLC25A3	0	0	1	0	1	0	0	0.285714	8
DDX21	0	1	0	0	1	0	0	0.285714	9
PSMD14	0	1	0	0	1	0	0	0.285714	10
NTSR2	0	0	1	0	0	0	0	0.142857	11
KRTAP5-8	0	0	0	0	1	0	0	0.142857	12

3.6 Conclusion

This piece of work presents an effective method to identify potential biomarker genes for a deadly disease, i.e., ESCC. The method exploits initially some biclustering techniques which have been found effective in terms of the enriched group of genes identified to support subsequent biomarker identification process. Biological networks have been constructed in the next step based on biclustering results. The topological, pathway and regulatory network analysis have been carried out finally on a filtered set of genes (based on the degree and the associativity) towards the identification of a set of biomarkers for ESCC. Out of 12 potential secondary genes, four genes viz. IFNGR1, CLIC1, CDK4, and COPS5 are found mostly associated with ESCC.

Identification of potential biomarkers from gene expression data using differential expression analysis is an another important and widely used approach. In next chapter, the dissertation proposes an ensemble based differential expression analysis method to identify potential biomakers associated with ESCC disease.