Chapter 5

Gene Module Extraction and Analysis-An Application to Breast Cancer

5.1 Introduction

Genes are at the core of the living body. Their interactions determine the various genetic and phenotypic changes in the body. The role of a gene is to produce certain proteins by means of expressing themselves in the form of mRNA. At this stage, some measuring techniques are used to record the expression level of genes giving rise to the gene expression data. These data can be analyzed to determine the role of genes at various molecular levels. A human cell comprises of nearly 30,000 genes. However, the expression or repression of genes at different cells delineate their role in the living body [1]. A gene coexpresses itself along with some other genes at a time to achieve certain functions. Analyzing such coexpressed genes, technically known as modules, will help in revealing the inherent properties of genes both from genetic and phenotypic perspectives. Identification of modules is generally done using clustering, which is an unsupervised approach. In order to obtain group of functionally related genes, the first requisite is to obtain a gene-gene network from the recorded coexpression values. Coexpression network construction is done using statistical techniques such as Pearson correlation or Spearman correlation measure, which quantifies the correlation between genes. An adaptive threshold is used to determine if correlation between two genes can be represented in the form

[1] https://en.wikibooks.org/wiki/Human_Physiology/Genetics_and _inheritance, accessed-31.8.2017

of an edge or not in the coexpression network. Once the coexpression network is obtained, extraction of modules is carried out based on some subjective criterion.

There are cases when a gene undergoes changes at its genetic or functional level, leading to some form of distortion which is shown by certain characteristic changes during their expression. This imbalance in expression level of genes give rise to certain types of phenotypic changes, which can be defined as some kind of disease or disorders. For many years, researchers have been trying to understand the causation of diseases and ways to cure or prevent them. However, this task is far from incomplete. This may be due to lack of knowledge about the properties of genes, their affinity with each other or may be due to certain external factors. The severity of diseases is increasing day by day and so are the causes. Earlier, it was thought that certain disorders were caused due to some abnormality in one of the genes in the body. But this notion was ruled out in [124] work. They believed that phenotypic changes cannot be due to impairment in one of the genes [48]. Mutation in one gene spreads to other genes as well, thereby disrupting their normal functioning and ultimately leading to some disease. Such disorders are referred to as non mendelian disorders, these abnormalities show heredity recurrence and have defined genetic patterns, but they cannot be easily delineated.

Diseases can be grouped into two categories- hereditary, which means a person develops higher probability of getting a disease if he inherits certain dysfunctional genes from his parents or genetic, which means that certain disturbances in the genetic organization of genes may lead to the occurrence of the disease. Cancer, the most deadly and widespread disease is mainly genetic in nature with some exceptions such as retinoblastoma (a tumor in eye developed during childhood). Cancer does not occur due to a single mutation in some particular gene, rather it is a disease which affects a large community of cells in the body. A group of damaged cells which grows and divides uncontrollably are referred to as cancerous cells. Figure 5.1 shows the uncontrolled nature of growth of cells leading to a cancerous situation.

These cancerous cells may be benign or malignant in nature. Benign ones do not disrupt the normal functioning of the body. On the other hand malignant tumors are dangerous and needs to be treated properly. There are situations where the outgrowth of cells in one part gets spread to other parts as well thereby disrupting the normal functioning of the newly affected part too. In such a case, the disease is said to metastasize. For example, breast cancer and brain cancer are different in

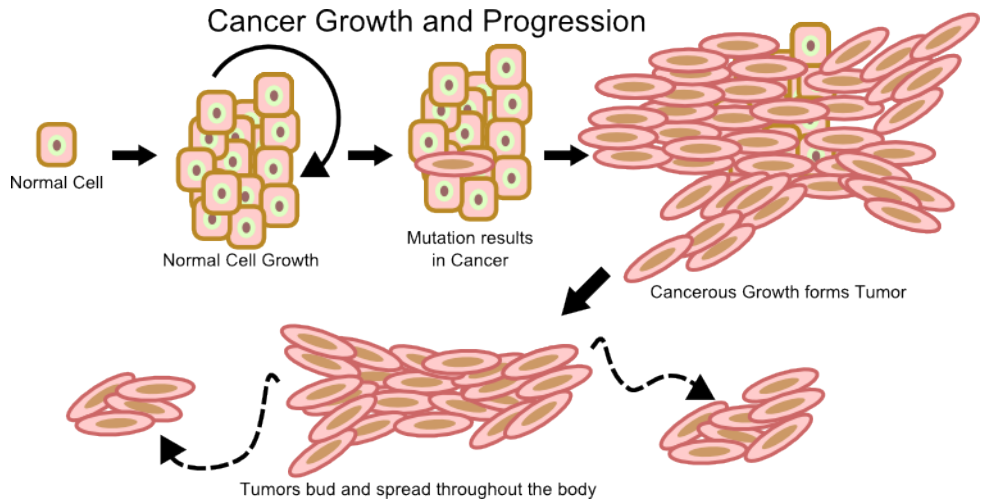


Figure 5.1. Cancer cell growth leading to tumor and its spreading (source-http://www. unc.edu/depts/our/hhmi/hhmi-ft_learning_modules/cancermodule /pages/cancer.html, accessed on-20.09.2017)

nature [2]. If a woman is detected with breast cancer at an early stage, the disease is likely to get cured and the lady can live a normal lifestyle. However, if the disease has spread to brain, the chances of suvival is almost nil even if detected at preliminary stage. Breast cancer is one among the deadliest disease affecting a large number of population. The rate of occurrence of the disease varies as per age and ethnicity of population. In United States, the death rates caused due to breast cancer is highest as compared to other cancers. By the end of 2017, the number of metastasized breast cancer cases has reached 252710 in the US [3]. Although treatment of breast cancers are available, the likelihood of a woman surviving metastasized situation is less than 16%. This is because the disease starts spreading by affecting the lymph nodes in the axilla area [4]. The lymph nodes are mainly associated with the defence mechanism of the body, which produces and carries lymphocytes produced by the bone marrow across the body. If certain disregulation occurs in this mechanism, the body looses its capacity to defend itself from any kind of attacks and therefore the situation leads to fatal consequences. Hence, a thorough study on the progression of the disease from non metastasis to metastasis stage to analyze the different trends in the properties and affinities of genes can be a useful support for the biologists to explore various ways to lower down the rate of its progression.

[2]http://www.medschool.lsuhsc.edu/genetics_center/louisiana/article_cancer.htm
[3]http://www.breastcancer.org/symptoms/understand_bc/statistics
[4]http://breast-cancer.ca/prog-untreated/

5.2 Related Works

Interactions between biological molecules is manifested in the form of modules, whether it be in gene-gene network, protein-protein network, metabolic network or any other combination of biological network. The basic network consists of the gene-gene network, which holds the responsibility for exhibiting different types of phenotypic features during the life cycle of a living being. Gene module analysis can reveal many interesting facts that a single gene cannot describe. Researchers have been using module level analysis since ages in systems biology and in genome-level analysis [127, 158]. A module may be obtained using various motives such as the module should have higher number of interconnections among themselves, should correspond to group of elements having similar functions. In earlier times, researchers hold the idea that a module should be well delineated from any other module. However, with the passage of time, it was discovered that a single biological entity can participate in multiple functionalities thereby resulting in overlapping modules, hence recent works have started analyzing modules from this perspective. Analyzing gene expression data from module analysis perspective serves as a bridge between gene and phenotypic characteristics. A person suffering from some disease exhibits different physical composition than a normal one. A bioinformaticists can investigate the expression trend observed in a healthy patient w.r.t. a diseased patient and can generalize their outcomes.

The class of cancers has been widely explored and many drugs have been designed to make it a curable disease today. Despite all these efforts, one cannot control the movement of cancerous cells to other parts of the body. This situation, known as metastasis pose life threats on the patient. Usually, metastasized cancer cells develops immunity against the normal treatment and hence keeps on spreading uncontrollably. Many researchers have emphasised the role of genes, their expression values and pathways to study the association of diseases and their progression pattern. For example, gene expression data has been successfully analyzed to differentiate the heterogenity in Diffuse large B-cell lymphoma (DLBCL) [2]. The heterogenity is mainly due to host response, the tumor's propagation rate and the tumor's differentiation style. Understanding the molecular diversity of such tumors help in better prediction of survival rates of patients. Another work [12] predicts the survival rate of patients suffering from initial stage of lung adenocarcinomas This work ranks the genes based on a risk factor depending on the survival trend. The demarcation line between the low risk and high risk genes involved in stage I of lung adenocarcinoma can be cleverly used to aid the therapy process. Few researchers

have proposed a class predictor based on gene expression profiling to differentiate between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) [39]. This work has paved way for effective and target specific treatments for the two classes of cancer and can be succesfully used to differentiate between similar looking cancer cases. Some amount of work has also been done to understand and realize the metastasis and non metastasis stages of cancer, particularly breast cancer. It was found that different patients respond differently towards treatment administered to them even if they are in the common disease stage. This difference is due to various reactions which occurs between drug molecules and other non mutated genes. In [147], researchers have used supervised learning to identify a signature gene expression to mark the patients having distant metastasis. Usually, breast cancer starts spreading through the lymph node, but there are exceptions to this. The findings of this work can be used to find such exceptional cases of patients. These patients have higher recovery rate due to adjuvant treatment as compared to other patients, where cancer has spread to the lymph nodes. Another work [155] uses training and testing concept on 286 lymph-node negative patients. This work suggests 76 gene signatures which could predict occurrence of metastasis within five years of the disease. Prediction of patients likely to develop metastasis at a later age is not easy. Moreover, cure at metastatic stage of cancer is still a dream, but researchers are trying their best to design drugs which could be effective in killing this metatstasized cancer cells.

5.3 Proposed Method

Literature has been loaded with evidences indicating that genes with similar functions can be studied from coexpression analysis point of view. The essence of such a study is the use of clustering to identify gene coexpression modules. The method proposed here is based on a seed expansion strategy. During the module growing process, it uses the strength of relation between two genes in terms of their functional similarity to extract functionally enriched modules. The symbols used for my method is givenat the beginning of the thesis and details of the method is discussed next.

5.3.1 Preprocessing

The dataset used for my work is GSE 20304, whose details have been discussed in Chapter 2. A log2 transformation is used to normalize the expression value of

the 22,273 genes. A parameter tuning of the variance threshold is carried out to identify subset of genes actively involved during the disease progression stage. The variance parameter was set from 0.9-1.5 and the number of genes were recorded in Table 5.1.

Table 5.1 Number of gene samples drawn at varying variance threshold, V_{th}.

	V_{th} = 0.9	V_{th} = 1.0	V_{th} = 1.1	V_{th} = 1.2	V_{th} = 1.3	V_{th} = 1.4	V_{th} = 1.5
Number of genes	7900	7000	6129	5292	4529	3903	3356

The appropriate range of V_{th} was found to lie between 1.1-1.3. I used the genes obtained in this range to extract gene modules. However, for my extensive analysis purpose, I used the results of modules obtained at $V_{th} = 1.2$ as it is the average of the range obtained and is also supported by literature sources [154]. Another reason for choosing this threshold was the occurrence of higher number of common elements in the modules obtained at both the stages.

5.3.2 Gene Co-expression network construction and module extraction

The coexpression network construction process is carried out for the subset of genes at the specified threshold. I have used 5292 genes for the network construction. In this phase, two networks are constructed based on the patients' characteristics, whether they belong to the metastatic or non-metastatic group. The correlation between two genes is based on the Pearson measure (PCC), which determines how the behaviour of one gene changes w.r.t. other gene simultaneously. It takes a value between -1 to +1. A positive value of PCC indicates that the expression level of one gene increases with the expression value of its co-expressed gene whereas a negative value indicates that expression of one gene is suppressing the expression level of the other coexpressed gene. Thus, it can be used to determine the regulation pattern of genes, which can be used during network construction. However, it is highly parametric dependent and follows the hard thresholding approach, i.e, a sample with PCC value greater than PCC_{th} will be considered to be connected in the network whereas one below PCC_{th} will be depicted by the absence of edge between the sample elements. Using this measure, two networks- one for the non-metastatic

stage, (Adj_{nm}) and the other for the metastatic stage (Adj_m) is constructed.

$$Adj_{(g_i,g_j)} = \begin{cases} 1 \text{ if } PCC(g_i, g_j) \geq PCC_{th} \\ 0 \text{ otherwise} \end{cases}$$

The value for PCC_{th} is chosen to be 0.5 in order to give equal weightage to both gene expression values and semantic similarity during module extraction. Each of the matrix then undergoes the module extraction process. To understand the module extraction process, some definitions needed are discussed next.

Definition 27 (Seed node)**.** *A node $g_v \in V$ is considered to be a seed node iff $CCf(g_v) > CCf(g_w)$ $\forall g_w \in \{V - g_v\}$. Clustering coefficient, CCf of a node is calculated as per Definition 3 of Chapter 3.*

Definition 28 (Semantically connected)**.** *In a network, $V = \{g_1, g_2, ...\}$, if two genes g_v and g_w satisfy (i) $PCC(g_v, g_w) \geq CCfT$ and (ii) $SemSim(g_v, g_w) \geq SemSim_{th}$, then the two genes g_v and g_w are said to be semantically connected.*

Definition 29 (Gene module)**.** *A group of nodes $\{g_1, g_2, ...g_v\} \in M_i$, where M_i is any i^{th} module iff all members of M_i are semantically connected among each other.*

In order to choose the seed node, I have taken up the clustering coefficient measure. Among all the nodes in the network, the one having the highest clustering coefficient value is chosen as the seed node. The use of clustering coefficient during seed selection is purely based on its topological significance, as it determines the essentiality of nodes in the network based on their common neighbors. A node with higher clustering coefficient implies that it has more number of interconnections among its neighbors, i.e., in a biological network it suggests that the node along with its neighbors are more or less regulated in a similar fashion. In order to grow the seed, nodes sharing biological similarity with the seed nodes are considererd. The biological similarity between nodes is measured in terms of semantic similarity. In this case, I have used the Wang's semantic similarity. The node having maximum semantic similarity with the seed node is considered to be the first member for module expansion. The same process is repeated to get the other module members too for both the networks. The pesudocode for the method is given in Algorithm 4.

A module having higher biological relevance is considered to be more closely related to the disease provided that it contains one or more causal genes among its members. This statement can be supported by Proposition 6.

Input : $Adj_m = A = \{G, E\}$ or $Adj_{nm} = A = \{G', E'\}$ (Adjacency matrix representation of gene gene network); $CCfT$ (Clustering coefficient threshold); $SemSim_{th}$ (Semantic similarity threshold); $SemSimM$ (Matrix containing semantic similarity values fo all pairs)

Output: $Modules = \{M_1, M_2, \cdots, M_N\}$, (a set of N modules)

Initialize RemList $= V$, Modules $= NULL, mcount = 1$;
while |RemList| > 4 **do**
  choose $g_a \in RemList$ such that $\forall g_b \in RemList$, $CCf(g_a) \geq CCf(g_B)$ and $CCf(g_a) \geq CCfT$; $pC = pC \bigcup g_a$;
  while g_a *exists* **do**
    choose another g_i from $N_{s(g_a)}$ if and only if $\exists g_x \in pC$ such that $SemSimM(g_i, g_x) \geq SemSim_{th}$
    $pC = pC \bigcup g_i$;
    $RemList = RemList - g_i$;
    $N_{s(g_a)} = (N_{s(g_a)} \bigcup N_{s(g_i)})$ choose next g_i;
  end
  Mark pC as M_{mcount} only when $|pC| \geq 3$;
  $Modules = Modules \bigcup M_{mcount}$;
  $mcount++$;
end
Return $Modules$;

Algorithm 4: Network module extraction algorithm

Proposition 6. *For a module M_i with very high biological significance (low p-value), if any gene $g_a \in M_i$ is established to be a causal gene for some disease, say D from the disease repository, then $\forall g_b \in M_i$ such that $g_b = \{M_i - g_a\}$ will also have high correspondence to the disease, D.*

Explanation: Any gene g_a and g_b will be members of the same module M_i iff they satisfy $PCC(g_a, g_b) \geq PCC_{th}$ and $SemSim(g_a, g_b) \geq SemSim_{th}$. So if $g_a \in M_i$ is known to be associated with disease D, then definitely any other member, $g_b \in M_i$ also has to show similar type of nature in order to belong to the same module. Thus $\forall g_b \in M_i$ such that $g_b = \{M_i - g_a\}$ can also be said to be closely related to the disease D [59, 67, 87], hence the proof.

An example to illustrate this fact can be taken from the module whose members are $M_{m_1} = \{CCL5, CCND2, WARS, SRGN, TRAC\}$. Out of the five genes, CCND2 is known to be associated with the disease as given in GeneCard. The pathways associated with these genes are reported in Table 5.5. The disease gene, **CCND2** is associated with *p53 signaling pathway, Wnt signalling pathway, focal adhesion, Jak-STAT signaling pathway.* **CCL5**, which is a non causal gene is involved in *cytokine-cytokine receptor interaction, chemokine signalling pathway.* An overview of the chemokine signalling pathway (Figure 5.2) given in the KEGG database shows the association of these two pathways with that of the *Jak-STAT signalling pathway*, which is already established to be associated with the disease. Another non-causal gene, **WARS** is involved in the *trytophan metabolism path-*

way. This pathway is known to coordinate the working of glycolysis (Figure 5.3), which is indirectly regulated by the *Wnt signalling pathway* [103]. This implies that **WARS** gene plays an indirect role in monitoring the *Wnt signalling pathway*, which is also established to be closely associated with the disease. Therefore, it can be emphasized that members of a module other than causal gene can also have an impact on the severity of the disease.

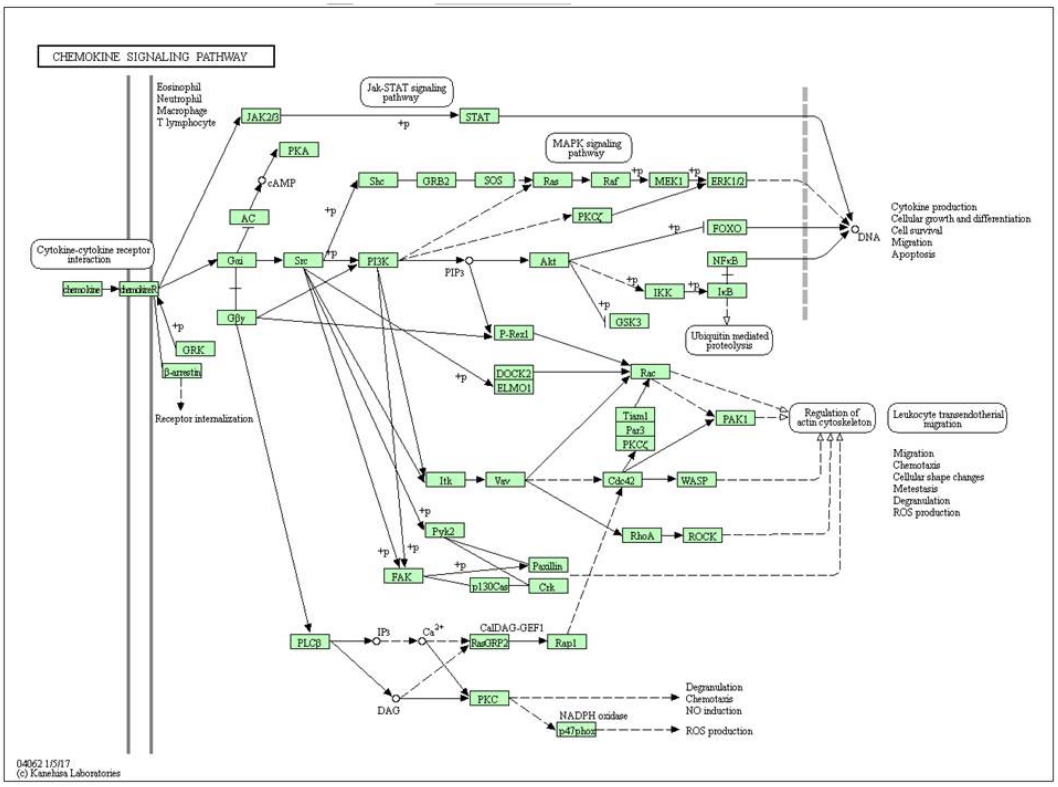


Figure 5.2. Chemokine signalling pathway (source-http://www.kegg.jp/kegg /kegg1.html).

5.4 Experimental Results

I implemented my gene network construction and module extraction on MATLAB running on HP Z 800 workstation with 12 GB RAM. The method has been validated on GSE 20304 dataset. In order to get the semantic similarity between gene pairs, we use the GOSemSim package, whose details have been discussed in Chapter 2. The estimation of semantic similarity between gene pairs is an essential step during my module extraction process.

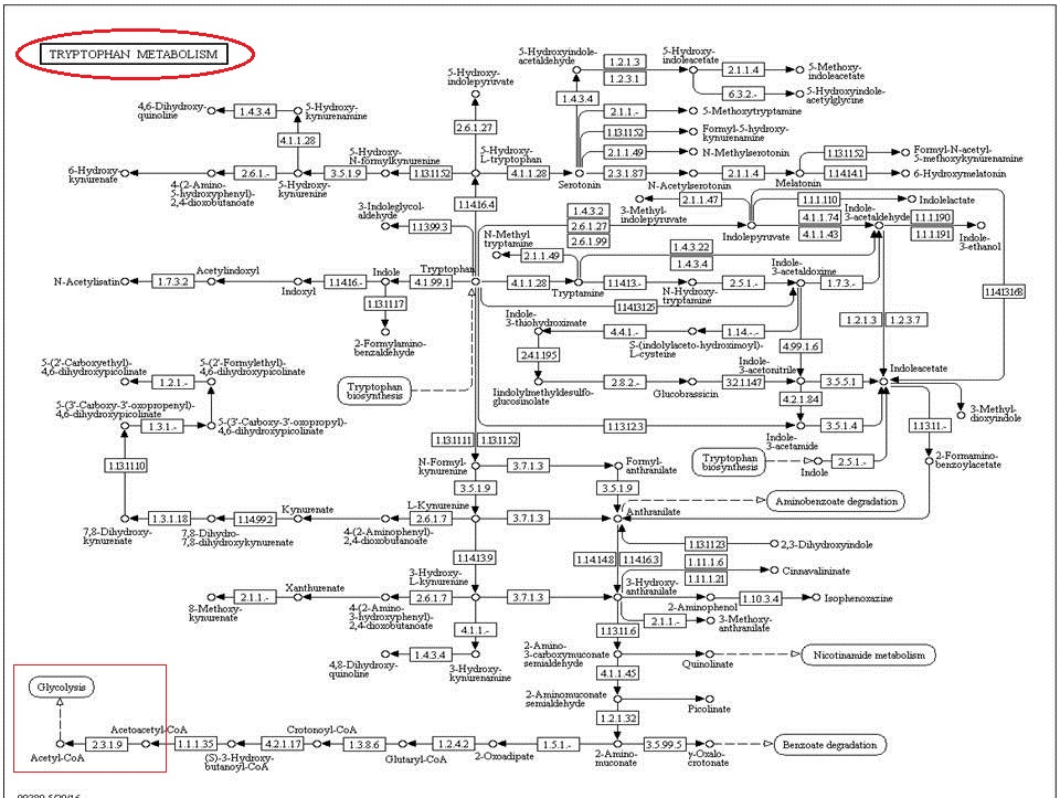


Figure 5.3. Tryptophan metabolism (source-http://www.kegg.jp/kegg /kegg1.html)

5.4.1 Parameter tuning

The proposed method relies on the appropriate use of two parameters viz., $CCfT$ and $SemSim_{th}$, where $CCfT$ takes care of the topological interactome whereas $SemSim_{th}$ quantifies the functional similarity between genes in the interactome. In order to get an optimal range of the two parameters, I have taken the help of p-value, which gives the functional similarity of a group of genes in any random environment. More details of this concept is given in Chapter 2. Each of the parameters are tuned in the range of 0.3-0.7. Table 5.2 reports the p-value of top three modules at each run of the experiment with varying thresholds.

From Table 5.2, it can be seen that among the top three modules at each run, only two module is having p-value of *5.10E-7* and *7.14E-7*, which is the two best values obtained during the experimentation. These values were obtained for parameter pair ($CCfT = 0.5, SemSim_{th} = 0.7$). Therefore, the optimal range of parameter can be fixed at this value. This is in line with the assumption that a lower p-value implies the similar functionality of the group of genes involved and hence supports the ultimate target of module extraction.

Table 5.2 p–value of top 3 modules obtained at different threshold value of $CCft$ and $SemSim_{th}$ in metastasis stage.

ModuleType	**p-value**		
	$CCfT = 0.3$		
	$SemSim_{th}$ = 0.3	$SemSim_{th}$ = 0.5	$SemSim_{th}$ = 0.7
M1	5.41E-6	2.25E-5	5.41E-6
M2	3.41E-5	3.41E-5	4.14E-6
M3	6.25E-5	1.68E-4	3.41E-5
ModuleType	**p-value**		
	$CCfT = 0.5$		
	$SemSim_{th}$ = 0.3	$SemSim_{th}$ = 0.5	$SemSim_{th}$ = 0.7
M1	4.14E-6	5.41E-6	5.10E-7
M2	5.41E-6	3.41E-5	7.14E-7
M3	3.41E-5	2.25E-5	8.41E-6
ModuleType	**p-value**		
	$CCfT = 0.7$		
	$SemSim_{th}$ = 0.3	$SemSim_{th}$ = 0.5	$SemSim_{th}$ = 0.7
M1	5.41E-6	2.25E-5	5.41E-6
M2	3.41E-5	1.56E-5	3.41E-5
M3	1.68E-4	1.68E-4	2.25E-5

5.4.2 Comparison with existing work

In order to validate our module extraction technqiue, we need to compare it with some other existing technique. This comparison is based on the p-value which denotes the enrichment level of a group of genes. We have used Module Miner [90] technique which uses NMRS as the similarity measure and is based on a spanning tree concept to identify modules. Table 5.3 reports the p-value of top three modules obtained using Module Miner.

Table 5.3 Comparison based on p–value of top 3 modules obtained in metastasis stage.

Method	**M1**	**M2**	**M3**
Proposed method	5.10E-7	7.14E-7	8.41E-6
Module Miner	3.76E-4	1.534E-2	1.69E-2

From Table 5.3, it can be seen that our proposed method gives modules with better p-value as compared to Module Miner. This indicates the superiority of our method over the existing method.

5.4.3 Pathway identification from module members

A pathway in a biological network reveals the sequence of interactions occurring within the molecules in cells. These series of interactions lead to different changes in the cell, for example, the expression or suppression of certain genes in the cell is controlled by these interactions or formation of certain products is regulated by them [5]. Identification of pathways associated with different members in a module may help unravel the mysteries associated with the progression of the disease from non metastatic to the metastatic stage. To reduce the search space for module members, I have used only the top five modules in terms of p-value in both the stages. In order to identify the pathway associated to each gene in a module, I have used the DAVID tool, whose details have been discussed in Chapter 2. Table 5.4 and 5.5 lists down the associated pathway information for each module along with their p-value in non-metastasis stage and Table 5.6 reports the same information for the metastasis stage respectively. A careful analysis of the two tables unfolds three new pathways namely, *Glycerophospholipid metablism, h-Efp pathway* and *CARM1 and Regulation of Estrogen Receptor* in the metastasis stage. These three new pathways can be studied individually to trace down the progression of the disease from nonmetastatic to metastatic stage.

5.4.4 Role of common genes during disease progression

In order to substantiate the role of common disease genes during the spread of this type of cancer, I have used a Venn diagram to illutsrate the common genes found among the stages. Figure 5.4 and 5.5 shows the common genes among the top five modules in both the stages.

From Figure 5.4, it is clear that three genes, viz., CCL5, CCND2 and WARS are found among all the top five modules in the nonmetastasis stage, however, when Figure 5.5 is considered, these three genes are found only among two modules out of the top five modules in the metastatic stage. This may be caused due to their low semantic similarity value with the seed node during module extraction process. Among the three common genes, CCND2 is already established to be aassociated with the disease as given in the database repository GeneCard [120]. This gene has been shown to have higher meddling capacity as compared with other genes and therfore leads to the spread of the disease to other organs as well. Studies have found that over expression of this gene causes an aggressive growth of cells in

[5]https://en.wikipedia.org/wiki/Biological_pathway, accessed on-08.09.2017

Table 5.4 Non-Metastasis modules

Module No.	Members	Pathways	Pathway associated gene names	p-value
47	SRGN, MX1, GBP1, PLEK, PDE4B, SLA, IL32, CXCL9, RUNX3, IFI44L, CD52, LRMP, TRBV19, CTSS, PDE4B, CCL5, CXCL10, CCND2, CYP1B1, WARS, MMP9, PFKP, TAP1, ARHGAP4, SLC2A3	Antigen processing and presentation	CTSS	5.38E-6
		Purine metabolism	PDE4B	
		cytokine-cytkine receptor interaction, chemokine signalling pathway, NOD-like receptor signaling pathway, Cytosolic DNA-sensing pathway, Toll-like receptor signaling pathway.	CCL5	
		cytokine-cytkine receptor interaction, chemokine signalling pathway, NOD-like receptor signaling pathway, Cytosolic DNA-sensing pathway, Toll-like receptor signaling pathway,RIG-I-like receptor signaling pathway	CXCL10	
		p53 signaling pathway, Wnt signaling pathway, Focal adhesion, Jak-STAT signaling pathway, Cyclins and cell cycle regulation	CCND2	
		Steroid hormone biosynthesis	CYP1B1	
		Tryptophan metabolism, Aminoacyl-tRNA biosynthesis	WARS	
		Leukocyte transendothelial migration, Pathways in cancer	MMP9	
		Glycolysis/Gluconeogenesis Pentose Phosphate pathway, Fructose and mannose metabolism, Galactose metabolism	PFKP	
		Antigen processing and presentation	TAP1	
		Rho cell motility signaling pathway	ARHGAP4	
		Facilitated glucose transporter	SLC2A3	
14	CCL5, CCND2, WARS, LAG3	cytokine-cytokine receptor interaction, chemokine signalling pathway, NOD-like receptor signaling pathway, cytosolic DNA-sensing pathway, Toll-like receptor signaling pathway.	CCL5	3.08E-5
		p53 signaling pathway, Wnt signaling pathway, Focal adhesion, Jak-STAT signaling pathway, Cyclins and cell cycle regulation	CCND2	
		Tryptophan metabolism, Aminoacyl-tRNA biosynthesis	WARS	

Table 5.5 Non-Metastasis modules

Module No.	**Members**	**Pathways**	**Pathway associated gene names**	**p-value**
6	CCL5, CCND2, WARS, IGK@, IGLV5-45	cytokine-cytokine receptor interaction, chemokine signalling pathway, NOD-like receptor signaling pathway, cytosolic DNA-sensing pathway, Toll-like receptor signaling pathway.	CCL5	1.68E-4
		p53 signaling pathway, Wnt signaling pathway, Focal adhesion, Jak-STAT signaling pathway, Cyclins and cell cycle regulation	CCND2	
		Tryptophan metabolism, Aminoacyl-tRNA biosynthesis	WARS	
15	CCL5, CCND2, WARS, IGLV3-19	cytokine-cytokine receptor interaction, chemokine signalling pathway, NOD-like receptor signaling pathway, cytosolic DNA-sensing pathway, Toll-like receptor signaling pathway.	CCL5	1.68E-4
		p53 signaling pathway, Wnt signaling pathway, Focal adhesion, Jak-STAT signaling pathway, Cyclins and cell cycle regulation	CCND2	
		Tryptophan metabolism, Aminoacyl-tRNA biosynthesis	WARS	
21	CCL5, CCND2, WARS, NKG7	cytokine-cytokine receptor interaction, chemokine signalling pathway, NOD-like receptor signaling pathway, cytosolic DNA-sensing pathway, Toll-like receptor signaling pathway.	CCL5	1.68E-4
		p53 signaling pathway, Wnt signaling pathway, Focal adhesion, Jak-STAT signaling pathway, Cyclins and cell cycle regulation	CCND2	
		Tryptophan metabolism, Aminoacyl-tRNA biosynthesis	WARS	

Table 5.6 Metastasis modules

Module No.	**Members**	**Pathway**	**Gene names**	**p-value**
12	CCL5,CCND2,WARS, SRGN, TRAC,	cytokine-cytkine receptor interaction, chemokine signalling pathway, NOD-like receptor signaling pathway, cytosolic DNA-sensing pathway, Toll-like receptor signaling pathway. p53 signaling pathway, Wnt signaling pathway, Focal adhesion, Jak-STAT signaling pathway, Cyclins and cell cycle regulation Tryptophan metabolism, Aminoacyl-tRNA biosynthesis	CCL5 CCND2 WARS	5.10E-7
13	WARS, PDYN, NUCB1, GUSBP3	Tryptophan metabolism, Aminoacyl-tRNA biosynthesis Opioid prodynorphin pathway, Signaling by GPCR	WARS PDYN	7.14E-7
8	WARS, ESR1,NUCB1, ASCL1	Tryptophan metabolism, Aminoacyl-tRNA biosynthesis CARM1 and Regulation of the Estrogen Receptor, h-Efp Pathway	WARS ESR1	8.41E-6
15	CCL5, CCND2, WARS, SRGN, TRBV19	cytokine-cytkine receptor interaction, chemokine signalling pathway, NOD-like receptor signaling pathway, cytosolic DNA-sensing pathway, Toll-like receptor signaling pathway. p53 signaling pathway, Wnt signaling pathway, Focal adhesion, Jak-STAT signaling pathway, Cyclins and cell cycle regulation Tryptophan metabolism, Aminoacyl-tRNA biosynthesis	CCL5 CCND2 WARS	9.27E-6
10	CCL5, WARS, LCAT, MFGE8	cytokine-cytkine receptor interaction, chemokine signalling pathway, NOD-like receptor signaling pathway, cytosolic DNA-sensing pathway, Toll-like receptor signaling pathway. Tryptophan metabolism, Aminoacyl-tRNA biosynthesis Glycerophospholipid metabolism	CCL5 WARS LCAT	1.52E-5

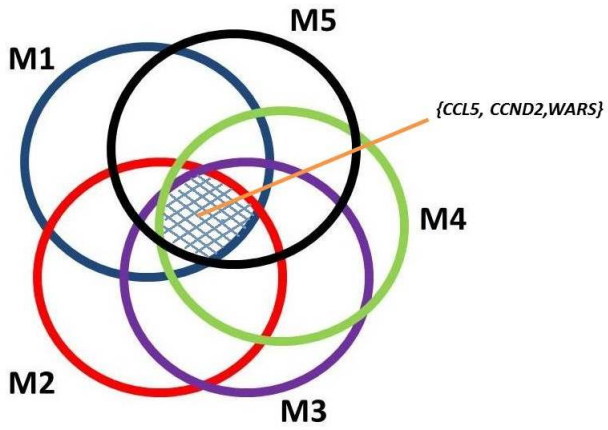


Figure 5.4. Common genes found among the five modules in nonmetastasis stage.

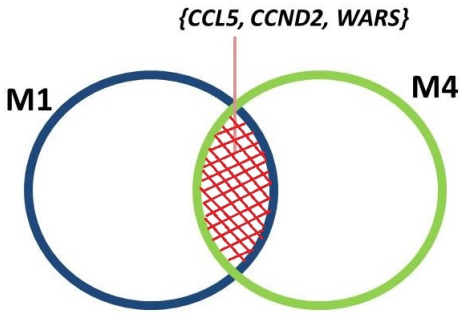


Figure 5.5. Common genes found among the two modules in metastasis stage.

an in vivo environment [84]. Apart from this causal gene, two other genes - CCL5 and WARS were found to be strongly correlated with the disease. Available sources reveals that polarization of CD+T cells caused by the active nature of CCL5/CCR3 gene in luminal breast cancer promotes the spread of the disease to other parts [170]. Another gene called WARS, which is the only common gene among all top five modules in both the stages is commonly known as Tryptophanyl-tRNA synthetase, corresponding to the aminoacyl-tRNA synthetase family. The role of this gene is seen during RNA transcription, angiogenic signalling pathways and also during the synthesis of many proteins [37]. Manifestation of tRNA synthetase favours movement of carcinogenic cells [78] thereby leading to the spread of the disease. Its presence in both the stages can be well defined due to its nature for promoting the movement of cancer causing cells over the body.

Table 5.7 Expression value of genes involved in common pathways in both the stages

Pathway	Common gene(s)	Non-Metastasis		Metastasis		Change
		Module No.	Average Expression Value	Module No.	Average Expression Value	
Cytokine-cytokine pathway	CCL5	14, 6, 15, 21	9.29	12, 15, 10	8.826	Decrease
Chemokine signaling pathway	CCL5	14, 6, 15, 21	9.29	12, 15, 10	8.26	Decrease
p53signaling pathway	CCND2	47, 14, 6, 15, 21	5.961	12, 15	5.676	Decrease
Cytosolic DNA sensing pathway	CCL5	14, 6, 15, 21	9.29	12, 15, 10	8.26	Decrease
Wnt signaling pathway	CCND2	47, 14, 6, 15, 21	5.961	12, 15	5.676	Decrease
Tryptophan metabolism	WARS	14, 6, 15, 21	8.65	12, 13, 8, 15, 10	8.276	Decrease
Focal adhesion	CCND2	47, 14, 6, 15, 21	5.961	12, 15	5.676	Decrease
Toll-like receptor signaling pathway	CCL5	14, 6, 15, 21	9.29	12, 15, 10	8.26	Decrease

5.4.5 Adaption of gene expression and role of common pathways during disease progression

I also did an analysis on the pathways which were found to be common as the disease progresses from the non metastatic to the metastatic stage. Apart from this, I also tried to analyze the behavioral changes in the genes during this transformation. Table 5.7 gives a list of all the common pathways along with the expression value of the associated genes in both the stages.

From Table 5.7, it is observed that as the disease progresses, the reported common genes show a decrease in their expression value. This is in line with the established works, which suggests the decreasing trend of genes during the progression of disease. Next, I discuss the role of the obtained pathways given in Table 5.7.

1. *Cytokine-cytokine pathway:* The body's immune system releases cytokines so as to hamper the development of tumor. However, there is a deviation among the carcinogenic cells, they use cytokines in the growth and spread of disease in the host's body [27].

2. *Chemokine signaling pathway:* A disturbed chemokine signalling pathway is

attributed to the alterations in the expression value of chemokines during different malignancies. Such a dysfunctional pathway is associated with the spread of the disease [141].

3. *p53signaling pathway:* Usually, p53 loss can disturb pathways favouring metastasis. However, there is an exception to this, transcriptionally defective TP53 mutants promote the spread of carcinogenic cells [110].

4. *Wnt signaling pathway:* Studies have suggested an important role of Wnt/β-catenin signalling pathway during the developmental stages of breast cancer [58].

5. *Tryptophan metabolism:* Tryptophan degradation is catalyzed by the overexpression of a number of enzymes. The same enzymes are linked to various forms of lung cancer, breast cancer and melanoma [113].

6. *Toll-like receptor signaling pathway:* Toll ike receptors promote the over secretion of cytokines/chemokines, which are involved in the growth and movement of cancer causing cells [74].

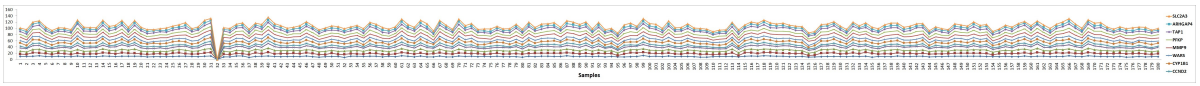
7. *Cytosolic DNA sensing pathway:* This pathway has not yet been established to be associated with the progression of the disease, but it can certainly be analyzed in context with the disease by some biologists.

5.4.6 Neutrality of few causal genes during module formation in metastasis stage

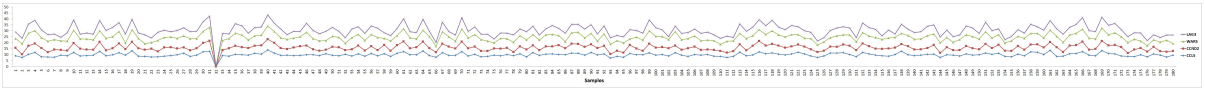
Disease genes such as CCND2, XBP1, SCGB1D2, MET, CYP1B1 and MMP9 are found to be actively involved in breast cancer. But during the transition from non metastasis to metastatic stage, only CCND2 has been found to retain its membership among the top five modules explored here in both the stages. Apart from this, the other five genes could not get a place during module formation in the metastasis stage. In order to support my finding, a proposition has been given here.

Proposition 7. *An established causal gene, $g_a \in M_{i_{nm}}$ may not show up during any module formation in metastasis stage.*

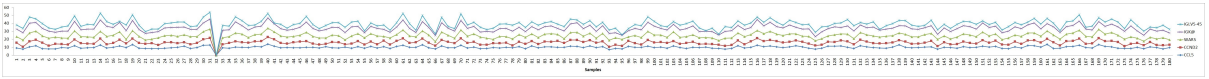
Explanation: The proposed module formation is based on two thresholds, $CCfT$ and $SemSim_{th}$. For a gene to be a member of the module, both the criteria needs to be satisfied. However, as per literature, a gene highlights significant variation



(a) Expression pattern among module members in module 1 of nonmetastasis stage.



(b) Expression pattern among module members in module 2 of nonmetastasis stage.



(c) Expression pattern among module members in module 3 of nonmetastasis stage.

Figure 5.6. Expression trend among module members in nonmetastasis stage.

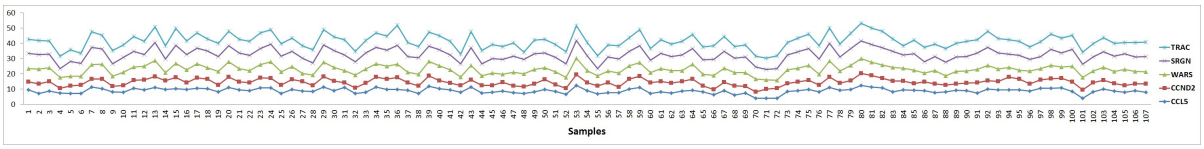
(may be fall or rise) in expression or semantic similarity value during stage transition. Suppose gene $g_a \in M_{i_{nm}}$, when g_a expresses itself in the metastatic stage, its expression value decreases. This decrease may lead to a network where g_a is completely isolated or connected with very few other nodes. In such a situation, it cannot lead to the formation of any module during this stage.

The semantic similarity value among these non involved causal genes were observed and it could be seen that XBP1, MET and CYP1B1 could have participated during module formation in the metastatic satge provided the threshold criteria was satisfied. Their low semantic similarity value with the seed nodes as per our threshold parameter cancelled out their membership during module formation.

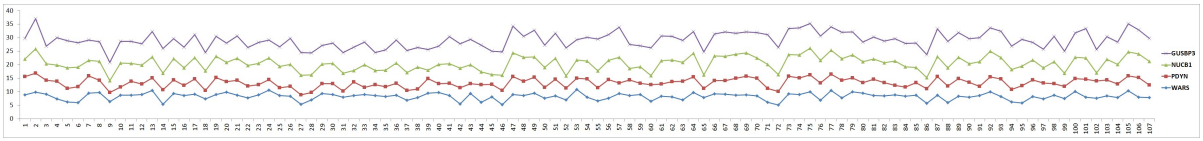
5.4.7 Expression pattern of module members

The expression trend of module members have been observed to cross check if members within a module are coherent in nature. Figures 5.6 and 5.7 shows the expression pattern of top three modules in both the stages suggesting that participants share high coherence among themselves.

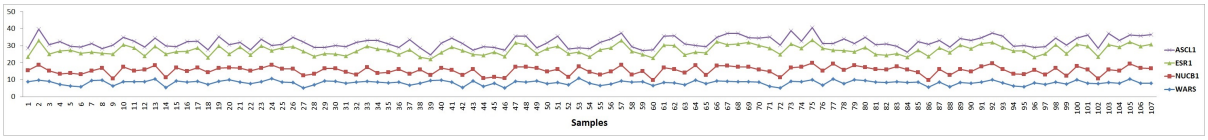
I have also analyzed the expression pattern of genes associated with the disease as listed in GeneCard. The causal genes are analyzed in terms of their average expression values across both the stages. Among the causal genes, CYP1B1 and MMP9 show a slight variation in their expression trend during the progression of disease from non metastasis to metastasis stage. Their expression value tends to increase during the disease progression, which is a deviation from the normal trend. Three other causal genes viz., XBP1, SCGB1D2 and MET are in line with



(a) Expression pattern among module members in module 1 of metastasis stage.



(b) Expression pattern among module members in module 2 of metastasis stage.



(c) Expression pattern among module members in module 3 of metastasis stage.

Figure 5.7. Expression trend among module members in metastasis stage.

the reported behavior (Figure 5.8). Another peculiar observation can be seen in XBP1's behavior across the stages. Its expression value shows very little variation over the 286 samples. The low variation in expression change can be attributed to its role in the regulation of expression level in immune system and in other cellular responses. This gene acting as a transcription factor which regulates the expression level of genes has to be involved in the same way across both the stages of disease as it is known to be associated with the immune system, which actually prompts the body to respond/fight back both in case of non metastasis and metastasis stage.

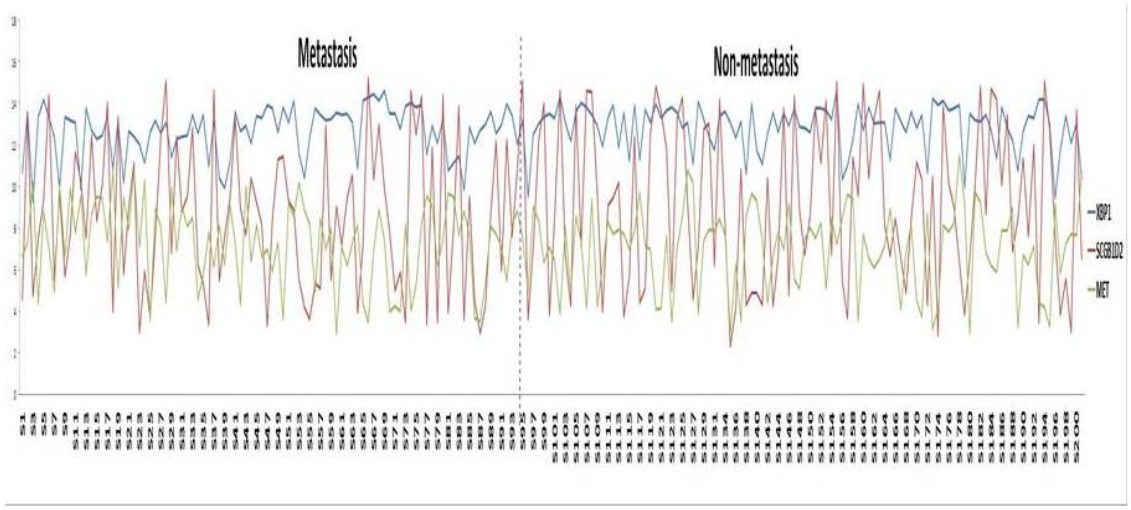


Figure 5.8. Common genes found among the two modules in metastasis stage.

5.5 Conclusion

In this work, I have proposed a module extraction technique from a gene gene network. This work has then been extended towards certain interesting biomarker discoveries during the progression of breast cancer from non metastasis to metastasis stage. The proposed module extraction technique has been validated based on the p-value concept and the newly identified suspected biomarkers have been validated based on the information available in their biological pathways along with their associations with the diseased pathways. The contribution of this chapter is given in the form a publication listed down as Publication No. 6 under the Publication section.

However, a major concern of this method lies in the use of a threshold to determine an edge between two genes based on the Pearson correlation coefficient. This filtering sometimes lead to information loss. Two genes may have strong association over a set of samples but they may be weakly related over the whole set of samples. In such a case, using a global threshold would result in the absence of an edge between these two genes which might be a significant portion during module formation. To handle this issue, the concept of association of genes over a subset of samples has been taken up in the next chapter. The next chapter discusses the association between genes in terms of a set of samples resulting in a multi-edge network. Using this network, gene modules are extracted based on the topological structure and these modules are found to be biologically significant and is at par with the modules extracted using a single edge network.