*Research Article*

# Modules Identification in Gene Positive Networks of Hepatocellular Carcinoma Using Pearson Agglomerative Method and Pearson Cohesion Coupling Modularity

## Jinyu Hu[1] and Zhiwei Gao[2]

[1] *School of Electrical Engineering and Automation, Tianjin University, Tianjin 300072, China*
[2] *School of Computing, Engineering and Information Sciences, Northumbria University, Newcastle Upon Tyne NE1 8ST, UK*

Correspondence should be addressed to Zhiwei Gao, zhiwei.gao@northumbria.ac.uk

In this study, a gene positive network is proposed based on a weighted undirected graph, where the weight represents the positive correlation of the genes. A Pearson agglomerative clustering algorithm is employed to build a clustering tree, where dotted lines cut the tree from bottom to top leading to a number of subsets of the modules. In order to achieve better module partitions, the Pearson correlation coefficient modularity is addressed to seek optimal module decomposition by selecting an optimal threshold value. For the liver cancer gene network under study, we obtain a strong threshold value at 0.67302, and a very strong correlation threshold at 0.80086. On the basis of these threshold values, fourteen strong modules and thirteen very strong modules are obtained respectively. A certain degree of correspondence between the two types of modules is addressed as well. Finally, the biological significance of the two types of modules is analyzed and explained, which shows that these modules are closely related to the proliferation and metastasis of liver cancer. This discovery of the new modules may provide new clues and ideas for liver cancer treatment.

## 1. Introduction

Hepatocellular carcinoma (HCC) is one of the most common malignant tumors in the world. Most of the liver cancer patients are in advanced states when they are firstly clinically diagnosed, which leads to poor treatment and high mortality. It is known that the nature of liver cancer is abnormal expression of genes caused by a variety of reasons. There are a lot of modules in a cell, and these modules work together to implement a function of the cell. The functional modules are composed of genes which are similar to each other

in physiological or functional aspects. When the gene functional modules receive impacts, they may lead to disease [1]. Microarray data are the results of many gene expressions, which consist in the information of the gene function modules [2]. There is a very important biological significance to identify gene functional modules in terms of a large-scale gene expression profiling. Cancer gene therapy has become a new treatment method following surgical resection, radiotherapy and chemotherapy, and interventional therapy. For instance, the recently discovered adeno-associated virus AAV3 may be useful for attacking human liver cancers.

The gene network is a complex dynamic system. Therefore, the process of finding gene modules is actually a process of discovering community structure from a complex network. The correlation between genes may be strong or weak, leading to a variant of collections of genes. Clearly, a strong community of a collection of related genes is what we are looking for. Currently, there are a number of community discovery methods of complex networks such as GN-splitting algorithm [3] and NEWSMAN cohesion algorithm [4], both of them use graphs without weights. In order to reflect the size of the gene intensity, it motivates us to use graphs with weights. It is worthy to mentions that Pearson values can be used to measure the correlation between genes. As Pearson values may be positive or negative, the absolute of Pearson was used to express the weights in [5] so that the intensity of correlation was obtained. Unfortunately, the positive correlation (mutual promotion) and negative correlation (mutual inhibition) were not considered in the method [5]. The genes in functional modules are a collaboration which may have mutually reinforcing relationships. Therefore it motivates us to use the Pearson values greater than 0 as the weights to reflect the concerned positive correlations of genes.

In this study, we are looking for the genes with similar function; therefore we will use an undirected weighted graph to describe gene network relationship. To the best of our knowledge, the present functions of modules are mainly for gene network graphs without weights, for example, Q function [6], which are invalid for our weighed graphs. Motivated by this, we design a PCC modularity algorithm to measure the performance of the modular decomposition.

According to experiments, if the module decomposition is optimized without considering the size of the threshold, it may lead to the obtained decompositions make no practical sense. In order to overcome this drawback, we propose a modified algorithm, that is, interval PCC modularity (IPM). For instance, in order to obtain a set of very strong modules, we preset the threshold range at the interval $[0.8, 1]$, and we can find a maximum modularity in the interval.

## 2. Construction of the GPN Network

### 2.1. Definition of Network and Storage

A connection matrix $C$ is used to store gene community networks (GCN), whose element is $c_{ij}$, defined by

$$c_{ij} = \begin{cases} p_{ij} & i > j \\ 0 & i \leq j, \end{cases} \tag{2.1}$$

where $p_{ij}$ is the Pearson correlation coefficient of the nodes $i$ and $j$, and $p_{ij} \in [-1, 1]$.
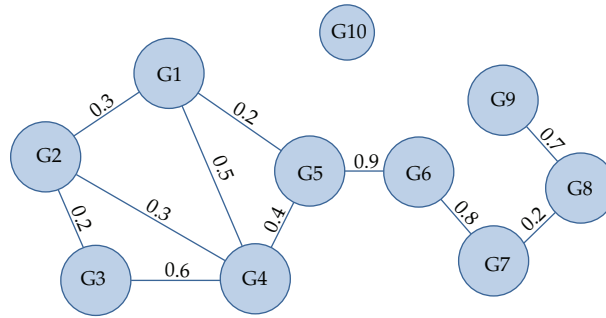
**Figure 1:** A GPN network: GPN10.

*Remark 2.1.* (1) $p_{ij} = 0$ indicates there is not linear relationship between two genes.

(2) $p_{ij} > 0$ implies a positive correlation between genes. Particularly when $p_{ij} = 1$, it indicates a completely positive correlation between genes.

(3) $p_{ij} < 0$ means a negative correlation between genes, and $p_{ij} = -1$ represents a completely negative correlation between genes.

Therefore from (2.1), one has $c_{ij} \in [-1, 1]$. When $c_{ij} = 0$ there is no edge between nodes $i$ and $j$.

The connection matrix is a lower triangular matrix, which stores an undirected weighted graph. It is noticed that no closed loops and no two-directional sides exist in the graph. This matrix is named as Pearson connection matrix (PCM).

### 2.2. The Type of the Network

The GCN networks can be divided into three kinds.

(1) Gene positive network (GPN): extract a network from the GCN where the values of all the edges in the network are greater than 0.

(2) Gene negative network (GNN): remove the edges with the weights greater than 0 to form a network.

(3) Gene absolute network (GAN): the weights of the edges are taken as the absolute values of the GCN network.

Genes in a module should reinforce mutually, which means the Pearson value of genes should be greater than 0. As a result, the GPN network will be used in this study, which is defined as follows:

$$c_{ij} = \begin{cases} p_{ij} & i > j, \ p_{ij} \in (0,1] \\ 0 & i \leq j. \end{cases} \tag{2.2}$$

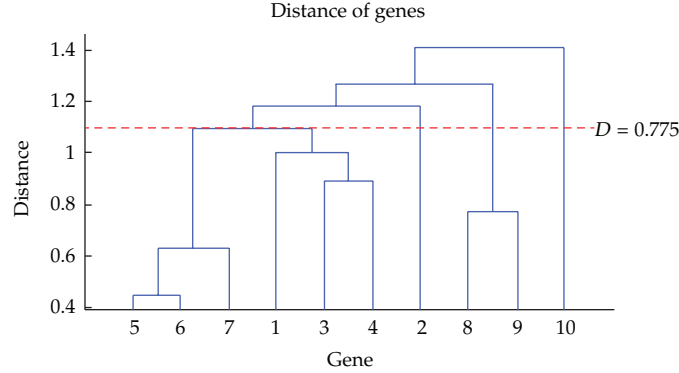A GPN network of 10 genes is depicted by **Figure 1**.

**Figure 2:** The PAM clustering tree of GNP10.

## 3. Pearson Agglomerative Method (PAM)

### 3.1. The Basic Idea of the PAM Algorithm

Hierarchical clustering is a conventional method to find a social network community structure, which can be classified into two types [7]: agglomerative method and divisive method. The main principle of the PAM proposed here is to first calculate the functional similarity between nodes using the Pearson method. Then to add new edges to a raw network composed of $n$ nodes and 0 edge starting from the node pairs with the most similarity. This process is repeated and may end in any node. This procedure from an empty graph to form a resulting graph can be described by Figure 2. In this figure, $x$-axis is the node and $y$-axis is the distance between nodes. We call this tree structure as PAM Clustering tree.

The distance of the genes is defined by [8]

$$d_{ij} = \sqrt{1 - c_{ij}}, \quad c_{ij} \in [0, 1]. \tag{3.1}$$

The larger the distance is, the relationship between genes become farther and vice versa. In our GPN, the weights are positives. Therefore the definition (3.1) can be used to measure disturbances of genes in this study.

In Figure 2, when the dashed line moves up from the bottom of the tree gradually, the variant nodes can be integrated into a greater community. The whole network becomes one community as the dashed line move up to the top. The tree structure corresponds to different community structure when it is cutoff from any location using dashed line.

If the red dotted line is placed at 0.775, we can obtain five modules as follows: {G5, G6, G7}, {G1, G3, G4}, {G2}, {G8, G9}, and {G10}.

### 3.2. Algorithm Implementation

In terms of the definition of the distance matrix $D = [d_{ij}]$, the PAM algorithm can be addressed in the following steps.

(1) Initialize the network with $n$ communities, where every node is a unique community.

(2) Calculate the minimum distance using Dijkstra algorithm [9], and then combine the minimum distance node pairs.

(3) Repeat step 2 until the whole network is merged to one community. The maximum execution times are $n - 1$.

We call this tree structure as "PAM clustering tree".

## 4. PCC Modularity Assessments

Similar to the GN algorithm [3] and other decomposition methods, the PAM algorithm cannot determine what kind of decomposition is optimal. It is of significance to value the performance of the decompositions. In [6], the modularity Q function was used to measure the quality of network partitioning. Unfortunately, for communities with big differences, the Q function is not ideal to assess the performance of module partitioning [9]. For the weighted networks, we introduce a standard function to measure the quality of network partitioning, namely, the PCC module function.

Here, the network is assumed to be divided into $N$ modules: $C_1, C_2, \ldots, C_N$.

### 4.1. Cohesion (Coh)

Cohesion (Coh) is the measure of the relevance of the internal nodes in a module. For a module with $n$ nodes, the maximum edges are $(n(n-1))/2$ and the maximum weight of the edge is 1. We use the ownership of the internal value which is divided by $(n(n-1))/2$ to represent the cohesion, described by

$$
\mathrm{Coh}(C_i) = \begin{cases} \dfrac{I(C_i)}{n_i(n_i - 1)/2} & n_i > 1 \\ 1, & n_i = 1, \end{cases} \tag{4.1}
$$

where $C_i$ is the $i$th module; $n_i$ is the number of nodes of the module; $I(C_i)$ is the sum of the ownership values in the module, expressed as $I(C_i) = \sum_{k=1}^{m_i} w_k$, where $m_i$ is the number of edges in the module and $w_k$ is the weight for the $k$th edge. The weights are not greater than 1, so $I(C_i) \leq m_i$. If the module has only one node, its cohesion is defined as 1.

### 4.2. Coupling (Cou)

The coupling (Cou) is a measure of the degree of association between modules, defined by

$$
\mathrm{Cou}(C_i) = \begin{cases} \dfrac{O(C_i)}{I(C_i) + O(C_i)}, & n_i > 1 \\ 1, & n_i = 1, \end{cases} \tag{4.2}
$$

where $O(C_i)$ is the sum of the weights of external edges connected to the module, expressed as $O(C_i) = \sum_{r=1}^{s_i} w_r$, in which $s_i$ is the number of external edges connected to the module, and $w_r$ is the weight of the $r$th edge. If the module has only one node, its coupling is defined

as 1. When $1/2 < \mathrm{Cou}(C_i) < 1$, it is equivalent to $I(C_i) < O(C_i)$, which indicates the internal strength of the module is less than the external strength, and the division of this module is generally unreasonable.

Equation (4.2) reflects the dependence of a module to the other. The lower is the coupling, the higher is the independence of the modules.

### 4.3. PCC Modularity

It is evident that the model partition needs high cohesion and low coupling. The formula $\mathrm{Coh}(C_i)[1 - \mathrm{Cou}(C_i)]$ can be used to describe the tightness of connection within the community $C_i$ and reflect the independence of the community. Thus, the new modularity is defined as

$$\mathrm{PCC}(C_i) = \mathrm{Coh}(C_i)[1 - \mathrm{Cou}(C_i)] \tag{4.3}$$

The PCC can be understood as "Pearson", "Cohesion", and "Coupling". Substitution (4.1) and (4.2) into (4.3) yields:

$$\mathrm{PCC}(C_i) = \begin{cases} \dfrac{2I(C_i)^2}{n_i(n_i - 1)[I(C_i) + O(C_i)]}, & n_i > 1 \\ 0, & n_i = 1. \end{cases} \tag{4.4}$$

If a module has only one node, it is called outlier module, whose PCC is 0.

The average of all the modules of PCC is used to measure the division of the entire network, which is defined as

$$\mathrm{PCC}(C_1, C_2, \ldots, C_N) = \frac{\sum_{i=1}^{N} \mathrm{PCC}(C_i)}{N}, \tag{4.5}$$

where $N$ denotes the number of modules. When the number of nonisolated point modules is $R$, the number of outlier modules is $N - R$. Since the PCC value of outlier module is 0, (4.5) can be rewritten as

$$\mathrm{PCC}(C_1, C_2, \ldots, C_N) = \frac{\sum_{i=1}^{R} \mathrm{PCC}(C_i)}{N}. \tag{4.6}$$

In other words, the network is divided into $N$ modules. Since the PCC value of each module is not greater than 1, one thus has $\mathrm{PCC}(C_1, C_2, \ldots, C_N) \leq R/N$.

## 5. The Relationship of Threshold and Modularity

### 5.1. One-to-One Map of Threshold to Modularity

From Figure 2, when the dotted line is set as $D = 0.775$, the network can be divided into five modules. In terms of (3.1), the threshold can be calculated as $T = \sqrt{1 - D^2} = 0.41$. Removing the edges whose PCC values are less than 0.41, Figure 1 can be transformed into Figure 3.
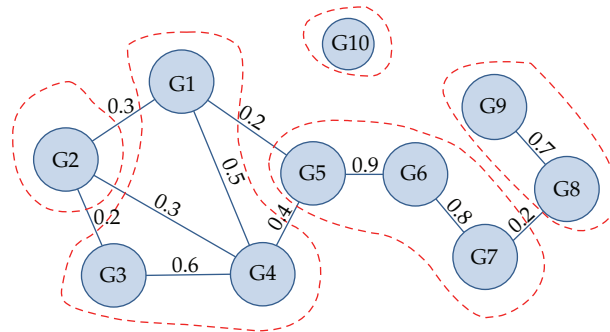
**Figure 3:** The decomposition of the network GPN10 by using $T = 0.41$.

**Table 1:** The relationships among $T$, PCC, and $N$.

| $T$ | 0 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| $N$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| PCC | 0.07 | 0.24 | 0.18 | 0.22 | 0.19 | 0.13 | 0.05 | 0.04 | 0 |

As a result, when the threshold is $T = 0.41$, the network GPN10 in Figure 1 is divided into five modules: {G5, G6, G7}, {G1, G3, G4,}, {G2}, {G8, G9}, and {G10}. This partition is the same as Figure 2.

Rather than building PAM-tree to divide the modules, we can decompose the network by using threshold value. A modular decomposition corresponds to a module function; therefore a threshold only has a corresponding module function.

For instance, by setting different thresholds, we can obtain the resulting decompositions (see Table 1), and each module corresponds to the modularity of a PCC. Under the same modularity premise, in order to ensure internal correlation of each module stronger, we choose a larger threshold.

Based on the decomposition of the network GPN10, Table 1 reflects the relationships among "$T$", "PCC", and "$N$", where "$T$" is threshold, "PCC" is PCC modularity, and "$N$" is the number of modules.

As each threshold corresponds to one decomposition, each threshold corresponds to one modularity as well. From Figure 4, when the threshold $T = 0.2$, the decomposition of the network GPN10 is optimal. In this case, the network GPN10 is broken down into three modules: {G1, G2, G3, G4, G5, G6, G7}, {G8, G9}, and {G10}.

### 5.2. The Definition of Interval PCC Modularity (IPM)

The absolute value of correlation coefficient is greater and the correlation is stronger. The correlation coefficient is close to 1 or −1, the correlation is very strong. The correlation coefficient is close to 0, the correlation is weak.

Generally, we judge the intensity of two variables by the range of correlation coefficients (see Table 2).

According to Table 2, we define five different ranges of modularity.

(1) Modularity of very weak correlation: $PCC(T)$, $T \in [0.0, 0.2)$.

(2) Modularity of weak correlation: $PCC(T)$, $T \in [0.2, 0.4)$.
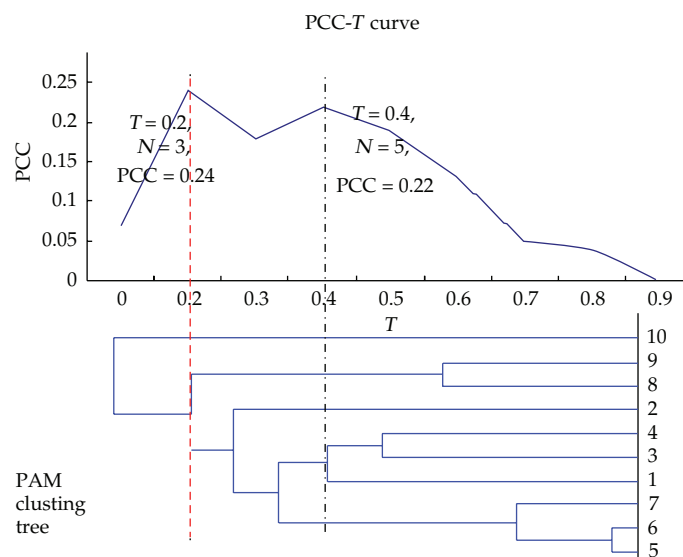
**Figure 4:** The relationship of thresholds and modularity under the different decomposition of GPN10.

**Table 2:** The intensity of Pearson correlation coefficient.

| | | |
|---|---|---|
| | 0.8–1.0 | Very strong correlation |
| | 0.6–0.8 | Strong correlation |
| Correlation coefficient | 0.4–0.6 | Moderate correlation |
| | 0.2–0.4 | Weak correlation |
| | 0.0–0.2 | Very weak or no correlation |

    (3) Modularity of moderate correlation: $PCC(T)$, $T \in [0.4, 0.6)$.

    (4) Modularity of strong correlation: $PCC(T)$, $T \in [0.6, 0.8)$.

    (5) Modularity of very strong correlation: $PCC(T)$, $T \in [0.8, 1.0]$.

    Generally, we find a strong correlation or strong related modules by using (4.1) and (4.2).

## 6. Results

### 6.1. Obtain the HCC Gene Modules

The liver cancer microarray data is taken from Chen et al. [10], which is available at http://genome-www.stanford.edu/hcc/supplement.shtml. The 1648 genes are differentially expression in HCC and nontumor liver in 156 liver tissues (74 nontumor liver and 82 HCC). We only study the gene expression of HCC. The Missing values are replaced by the average of the gene expression data under corresponding data column or sequence.

    We build the GPN network of 1648 HCC genes. Next we, respectively, test the PCC value in the threshold interval [0.8, 1] and [0.6, 0.8). According to the maximum of PCC value, GPN network is, respectively, divided into the HCC very strong correlation modules and the HCC strong correlation modules.
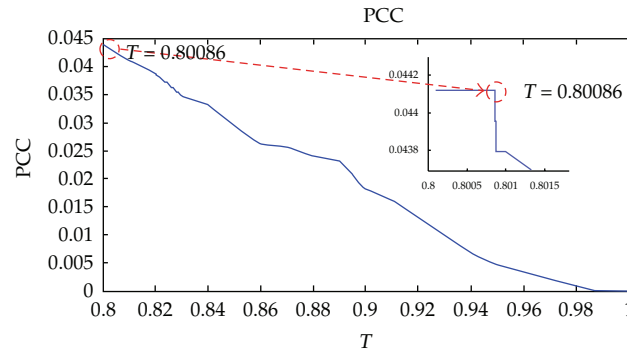
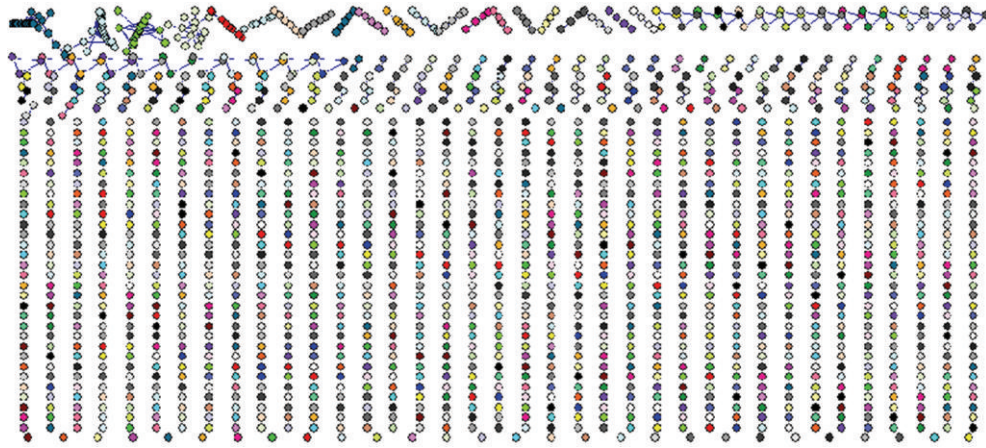**Figure 5:** The movements of PCC modularity at the threshold interval [0.8, 1].



**Figure 6:** The decomposition diagram of HCC GPN network when the threshold $T = 0.80086$.

### 6.1.1. Very Strong Correlation Modules (VSCM) of HCC

In the threshold interval [0.8, 1], the PCC curve is given by Figure 5, which shows that the PCC is downtrend. Within the threshold interval [0.8, 1], as the threshold is greater, the modular decomposition is getting worse. When the threshold is between 0.8 and 0.80086, the modularity PCC values are equal. In order to make the module correlation coefficient greater, we choose the threshold $T = 0.80086$. In this case, the modularity PCC = 0.0441.

In Figure 6, when the threshold $T = 0.80086$, the network is broken down into 1360 modules, including 150 nonisolated point module. According to formula (4.6), the PCC < 150/1360 = 0.1103.

In Figure 7, there are 13 modules and 121 genes in total, where each module is not less than five nodes. The modules numbered and arranged according to gene-related strengths from strong to weak. In order to distinguish the Very Strong Correlation Modules (VSCM) from the strong correlation modules (SCM), we mark VSCM and SCM with "S" and "W" respectively, which means "strong" and "weak".

In Table 3, "NO" is the abbreviation for "No Gene information" and "Trans" means "Transcribed locus". The number of genes is given in the bracket, for example, SERPINA5(5)
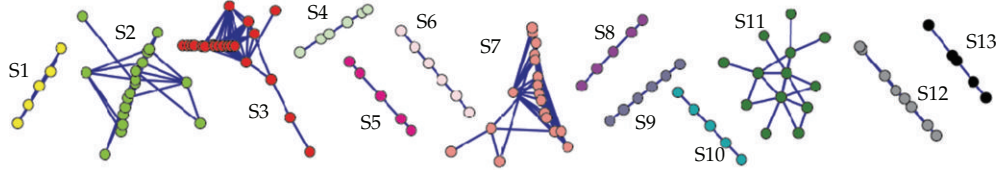
**Figure 7:** Thirteen gene modules of HCC very strong correlation modules with no less than five nodes.

**Table 3:** Thirteen gene modules of HCC very strong correlation modules (VSCM).

| Module | Number | The detailed genes of each module |
|---|---|---|
| S1 | 5 | SERPINA5(5), LOC100507281 |
| S2 | 18 | WNT4, SLU7, CSF2RA, IGKC(3), NCF1, EPB72, IGHG3, Trans, IGL(5), TNFSF10, NAPA, CSF2 |
| S3 | 21 | NO(11), Trans(7), C2orf55, SLC35E1, TRIOBP |
| S4 | 6 | Trans, ZFP92(2), TAGLN(2), AEBP1 |
| S5 | 5 | C1R(2), C1S(2), FGA |
| S6 | 6 | CKAP2, AQP4, HAMP, Transcribed locus(3) |
| S7 | 18 | LRRC8C, EDNRA, BIRC5, MT1B(2), Trans, AGXT, MT1H, MT1G, MT1F(2), MT1E, MT1L, NO, LARP4(3), DLG4 |
| S8 | 5 | NO, Trans, RS10(2), CDNA |
| S9 | 6 | GRN, C19orf6, RAD23A, ZNF451, RER1, ABCF1 |
| S10 | 5 | TUBA2, TUBA1(2), TUBA3, Trans |
| S11 | 13 | PLK, TROAP, Trans, CENPM, MYBL2, PTTG1, NUSAP1, CDC20, FOXM1, UBE2C, CDC2, KIAA0101, IFIT1 |
| S12 | 8 | RPS20(2), EIF3S6(2), NO(2), RPL30, Trans |
| S13 | 5 | SPARC(2), THY1(2), COL4A2 |

means that there are five SERPINA5 genes. Green denotes the genes in the module with low expression and red indicates the genes with high expression.

### 6.1.2. Strong Correlation Modules (SCM) of HCC

In the threshold interval (0.6, 0.8], we obtain strong modularity PCC curve as found in Figure 8.

When $T = 0.67302$, the optimal PCC is 0.0687. In this case, the HCC GPN network is divided into $N = 955$ modules (see Figure 9) and the number of the nonisolated modules is $R = 164$. By the formula (4.6), we can get that the PCC $< 164/955 = 0.172$.

We selected the modules with more than five nodes and arranged according to the order of strength from strong to weak order as follows (see Figure 10).

In Figure 10, there are 14 modules and 505 nodes in total. The modules with fewer nodes are W1, W6, W8, W12, W13, and W14. The modules containing a large number of nodes are W11, W4, W3, W2, W5, and W7.

In Table 4, there are 14 strong modules in total, involving 504 genes. The number of gene duplication is marked in brackets. Red means the gene is highly expressed in HCC and green indicates genes in low expression. The genes in W3 have both high expression and low expression, therefore it is not colored.
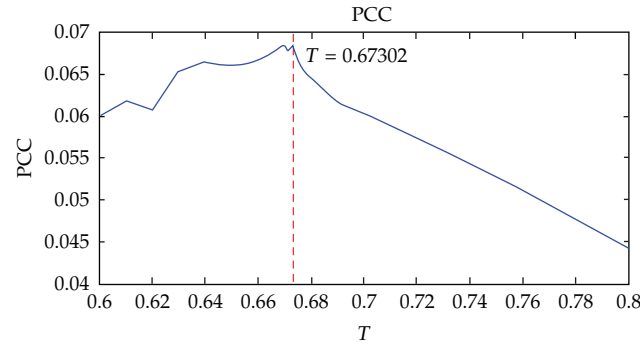
**Figure 8:** The PCC modularity trend on the threshold interval [0.6, 0.8].
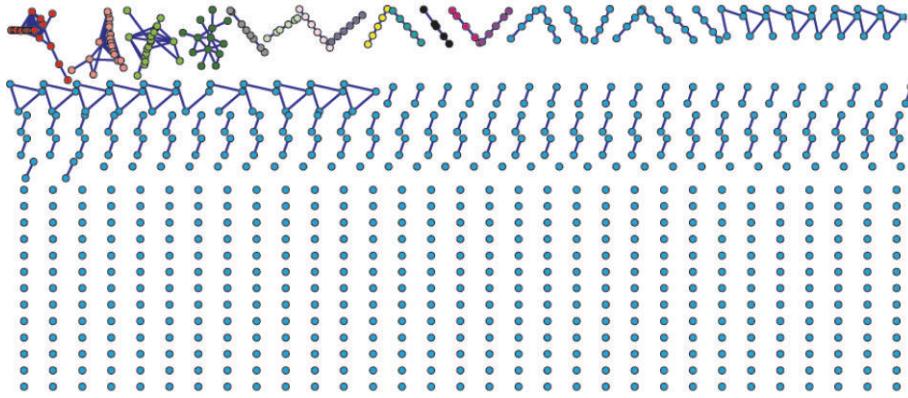


**Figure 9:** When the threshold $T = 0.67302$, the network is broken down into 955 modules in which there are 164 nonisolated point modules.

Generally, HCC strong correlation modules (SCMs) may include the genes of HCC with very strong correlation modules (VSCMs). In Table 4, blue bolds mark genes which appeared in the VSCMs. This kind of inclusion relations are shown in Table 5.

Table 5 shows the inclusion relationship of the VSCM and the SCMs. The table can also be described as W5 ≥ S3; W7 ≥ S7; W2 ≥ B2; W11 ≥ S11 + S12 + S10 + S8; W10 ≥ S6; W13 ≥ S9; W1 = S1; W3 ≥ S4 + S13; W4 ≥ S5.

### 6.2. Biological Explanation of the HCC Gene Modules

The biological explanation of the HCC gene modules, if not otherwise specified, all refers to the Stanford gene database: http://smd.stanford.edu/cgi-bin/source/sourceSearch. As there are too many genes, we only provide the biological explanation for the important genes in the each module. According to the major functions of the genes in every group, we name each module.

**Table 4:** Fourteen gene modules of HCC strong correlation modules (SCM).

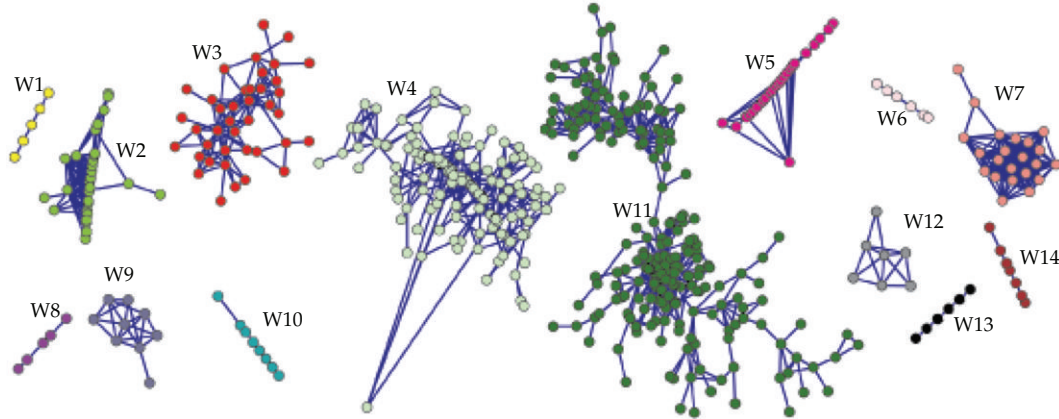| Module | Number | The detailed genes of each module |
|---|---|---|
| W1 | 5 | SERPINA5(5), LOC100507281 |
| W2 | 29 | WNT4, SLU7, CSF2RA, IGKC(3), NCF1, EPB72, IGHG3, Trans, IGL(5), TNFSF10, NAPA, CSF2, CD69, TF, NO, PSMF1, EDR2, HNT, KLRK1, SYT6, ID4, HCLS1, CD53 |
| W3 | 40 | Trans, ZFP92(2), TAGLN(2), AEBP1, SPARC(2), THY1(2), COL4A2, ID3, COL6A1, FGF12B, TMEM204, MYO10, CSNK2B, PDGFRA, SVEP1, SVEP1, SRPX, CRISPLD2, RBMS3, PYGM, MFAP4, COL6A2, PODN, LAMA2, NGFR, NRG2, CYR61, SLC15A2, SCYA2, TSPYL1, ID4, CRHBP, THY1, NOTCH3, COL15A1, LOXL2 |
| W4 | 123 | CPS1, ZNF248, HRSP12, PCBD1, ALDOB(2), ENC1, APOC3(2), PAH, CD302, POR, Trans(7), SERPING1(2), IVD, APOH, SCYA14, PBP, SORD, EVX1, UGP2(2), C21ORF4, GALE, HSD17B6, CYP2A7, MST1(2), APOA1, C1R(2), C1S(2), FGA, C1RL, PROML1, LRRN3, LANCL1, ACOX1, CYP2C, BDH1, PIPOX, MPDZ, HSD11B1, RGN, PCK1, CHD9, ACAA2(2), FACL2, PON3(2), D4S234E, AZGP, RNAC(2), ADH6(2), ADH4(2), ADH2, APOC4, SLC27A5, MMSDH(3), PCK2, CPB2(2), CPN2, DEPDC7, CYP4V2, LY9, GRHPR, AMDHD1(2), ACADSB, ST3GAL6, SPRYD4, CYB5, ADI1, NO(4), QDPR, PLG, CYP27A1, GYS2, CTH, SHMT, ARHB(2), OGDHL, ACY1, APCS, PXMP2, EDNRB, C14orf45(2), SCP2(2), DHTKD1, KNG, ALAS1, MARC2, SULT2A1, CYP2J2, CTSO, SOD1, MYO1B(2), SYBU, PVRL3, PDK, KIAA0317 |
| W5 | 25 | NO(13), Trans(9), C2orf55, SLC35E1, TRIOBP, L3MBTL4 |
| W6 | 6 | PTMS, SDHAP1, RBP5(2), IKBKAP, HAAO |
| W7 | 22 | LRRC8C, EDNRA, BIRC5, MT1B(2), Trans(2), AGXT, MT1H, MT1G(2), MT1F(2), MT1E, MT1L, NO, LARP4(3), DLG4, CDK5, TFG |
| W8 | 5 | AFF4, NUFIP2, LEAP2(3) |
| W9 | 9 | SERPINA3, FGB(4), Trans, FGA, FGG, CFI |
| W10 | 8 | CKAP2, AQP4, HAMP, Trans(4), LOC257396 |
| W11 | 212 | EIF4B, RB1CC1(3), MTF2, MAL2, MAL2, CDNA(2), HMG17, CSNK2B, CUTA, ASAP1(2), RCC2, LMNB2, MAPK13, HJURP, SMC4, CMTM1, NO(6), SEMG2, 14ORF4, RPS10(3), Trans(6), RPS16, KLK3, HBG1, RPLP0, RPS5, RPS19, CPNE1, ETV1, TUBB(2), WNK1, RTN3, C1orf43, PAX8, FAM83H, TUBG2, TUBG1, TSEN54, UBE2M, TRIM28, SNRPB, HGS, STARD3, GPS1, CLPTM1, ARF3, ASNA1, TAF2E, USP5, SHC1, VARS2, ASF1B, PKMYT1, SERPINB3, E2F1, NLRP2, H2AFX, MLF1IP, ILF3(2), C1orf9, NAP1L1, SCNM1, LAPTM4B(2), TOP2A, HN1, TUBA2, TUBA1(2), TUBA3, BUB1, HSU, CKS1, CBX1, SLC1A4, KPNA2, EIF4A2, TMEM106C, EHMT2, SF3B4, SCAMP3, FLAD1, TCFL1, UBAP2L, PRCC, UBE2Q1, HTCD37, SNX27(3), PYGO2, FAM189B, NCSTN, RPRD2, USP21, MCM4, SNHG10, GMNN(2), PLK, TROAP, CENPM, MYBL2, CENPW, TPX2, ZNF261, ZWINT, LAP18, PTTG1, NUSAP1, CDC20, FOXM1, UBE2C, CDC2, MAD2L1, KIAA0101, CDKN3(2), RRM2(2), IFIT1(2), FAM72B(2), CEP55(3), KNSL5(2), NUDT1(2), TRIP13, MCM5(2), NRM, CDK4, KIAA1522, RDBP, PEA15, NPM1, UBR5, MRPL42(2), XPOT(2), MZT1, ACLY, PAPPA, ILF2(2), TCEB1, ASPH, ATP6V1H, YWHAZ(3), ZNF706, RPS20(2), EIF3S6(2), RPL30, RAD21, BIG1, MTDH, POLR2K, ARMC1(3), COPS5, CANX, KIAA0196, PTK2, TCEA1, NSMCE2, ZHX1, UQCRB, NBS1, FAM49B, DEK, UBA2, TIMP1, PSPH, LAMB1, SRXN1, PIR, TACC3, MCM3, DR1, CDC7, MCM6, RASSF3, POLA, YKT6 |
| W12 | 7 | TIGD5, MAF1, PUF60, CYC1, SHARPIN, GPAA1, Trans |
| W13 | 6 | GRN, C19orf6, RAD23A, ZNF451, RER1, ABCF1 |
| W14 | 7 | HIST1H2AC, HIST1H2BK, HIST1H2BC, HIST2H2BE(2), CPS1, Trans |

**Figure 10:** Fourteen gene modules of HCC strong correlation modules with no less than five nodes.

**Table 5:** The inclusion relationship of the VSCM and the SCMs.

| SCM | W5 | W7 | W2 | | W11 | | | W10 | W13 | W1 | | W3 | W4 |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| VSCM | S3 | S7 | S2 | S11 | S12 | S10 | S8 | S6 | S9 | S1 | S4 | S13 | S5 |

### 6.2.1. Comparison with Other Results

In this section, we will show the comparison of our results with the experimental results of Chen et al. [10] and Yan et al. [1].

The module S2 (or W2) agrees with the module D in Chen et al. [10] and the module C in Yan et al. [1], which is in relation to B lymphocytes. The disorder of B-cell immune function has a lot to do with liver cancer.

The module S4 agrees with the module E in Chen et al. [10], which is stroma cell module. The function of the module S13 relates to the endothelial cell, and the module G in Chen et al. [10] has the similar capability.

The module W3 contains the genes both from the module S4 and S13. Since S4 genes means stroma cells, while the S13 genes are all located in the stroma, the module W4 is a generalized stroma cell module. The W4 functions as the module D given by Yan et al. [1]. In the HCC, the genes of S4 are in lower level expression, while the expression level of S13 is higher. The module W3 not only contains the highly expressed genes, but also includes the gene with low expression. From this viewpoint, the model W3 is different from any other modules.

The module S5 (or W4) functions as the module K by Yan et al. [1], whose function is about complement. Bacterial infection of the liver cells may be related to genetic disorders of the complement component module.

The module S10 is consistent with the module J by Yan et al. [1], in which the main genes are about the tubulins $\alpha1$, $\alpha2$, and $\alpha3$. Tubulin abnormalities have impacts on the occurrence and development of liver cancer.

The genes of module S11 all appear in the module A given by Chen et al. [10]. The function of the module is also identical to the feature of the module A from Yan et al. [1], which is related to the cell cycle and proliferation of cancer cells.
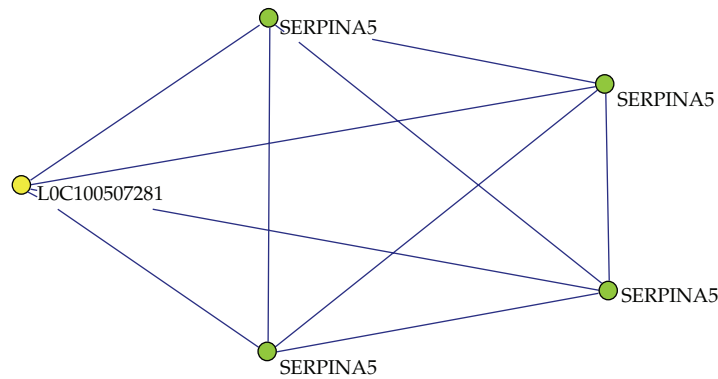
**Figure 11:** The gene network of module S1.

The genes in the module S12 are about ribosomal proteins, which functions as the same as in the module G in the reference by Yan et al. [1]. Ribosomal protein dysfunction may be related with liver cancer.

The function of module W11 is consistent with the module A given by Chen et al. [10], whose function is about cell proliferation. It is noted that the proliferation module A from Chen et al. [10] does not include S8, S10, and S12. In fact, tumor cell proliferation not only relates to cell cycle, but also relates to ribosomes, and tubulin. Because ribosomal synthesis protein and microtubule protein involves in mitosis, both of the two process are essential in cell proliferation.

The module W14 of histone is consistent with the module I from Yan et al. [1]. The occurrence of liver cancer may be associated with abnormal expression of the histone.

We have found that twelve gene modules such as S2, S4, S13, S5, S10, S11, S12, W2, W3, W4, W11, and W14 have specific functions and are largely in line with the gene clusters found by Yan et al. [1] or Chen et al. [10]. It has proven that the PAM algorithm and the PCC modularity can effectively discover gene function modules.

### 6.2.2. The Specific Modules of VSCM

In this subsection, we will focus on the modules only existing in this study. The specific modules of VSCM are S1, S3, S6, S7, S8, and S9, in which only S8, S9 are highly expressed.

### (1) Hemostasis Module S1

In Figure 11, the key gene of the module S1 is SERPINA5, which plays the hemostatic role in the blood plasma. Additionally, SERPINA5 is able to inhibit the migration of HCC cells. The low expression of SERPINA5 genes probably promotes the occurrence and development of HCC.

### (2) Transport Module S3

In Figure 12, NO means no gene information and trans represents Transcribed locus. TROPBP and SLC35E1 are the key genes in S3. TRIOB is closely related to HCC [11]. Moreover,
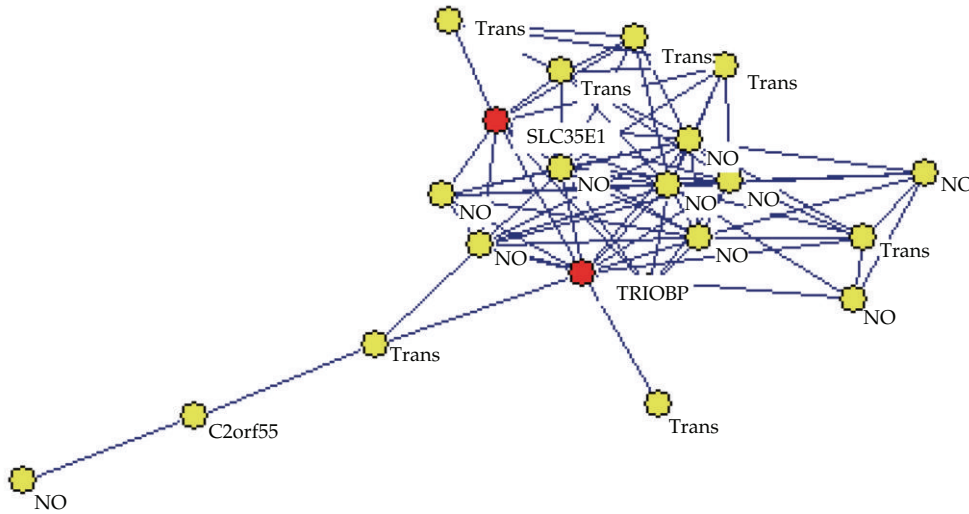
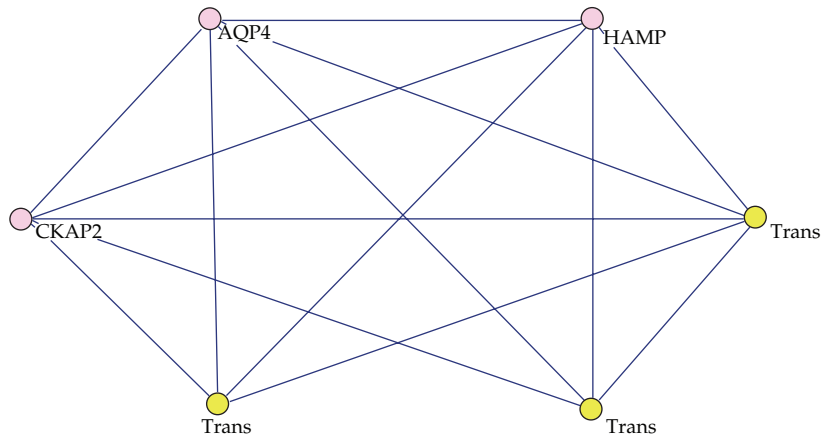**Figure 12:** The gene network of module S3.



**Figure 13:** The gene network of module S16.

the transport function of the solute carrier SLC35E1 is related to cancer [12]. Therefore the module should be responsible for transport in hepatoma cells.

*(3) Iron Regulation Module S6*

The key genes in Figure 13 are CHAP2, AQP4, and HAMP. CKAP2 is a cytoskeleton-associated protein involved in mitotic progression. AQP4 encodes a member of the aquaporin family of intrinsic membrane proteins. Decrease in aquaporin expression [13] may lead to the increase of the resistance to apoptosis in hepatocellular carcinoma. The product encoded by this gene HAMP is involved in the maintenance of iron homeostasis. This module relates to the regulation of the iron.
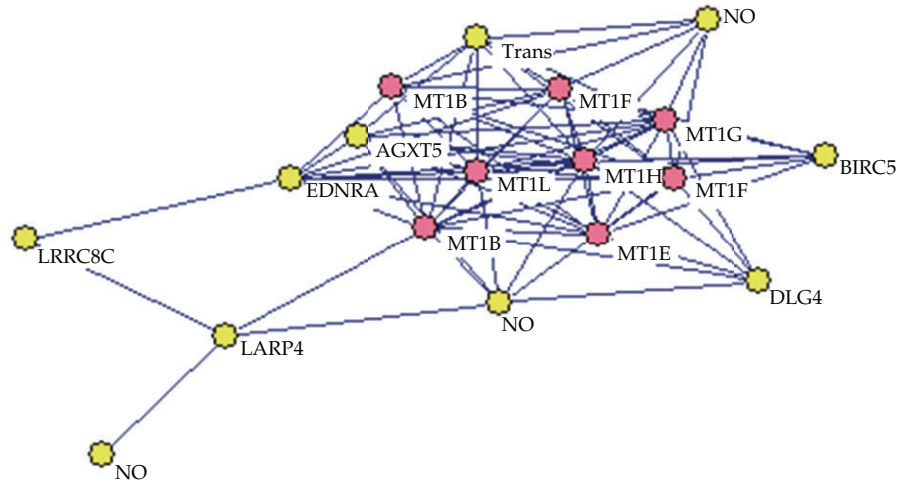
**Figure 14:** The gene network of module S7.

*(4) Metalloproteins Module S7 (see Figure 14)*

The genes in this module are most from the metalloproteins (MTs) family, such as MT1B and MT1G. MTs play a key role in the transport of essential heavy metals, detoxification of toxic metals, and protection of cells against oxidation stress. The increasing of the MTs levels in cancer cells are probably related to their increased proliferation and protection against apoptosis [14].

*(5) Antiterminator Module S8*

From Figure 15, RPS10 (ribosomal protein S10) are the most important genes in this module. RPS10 (as known as NusE) reflects another function [15] which is different from the ribosome module S12. Ribosomal S10 relates to liver cancer [16], and RPS10 play an antitermination role in the transcription process [17]. Therefore, ribosomal S10 may be the antitermination factor for liver cancer.

*(6) Immortal Module S9 (see Figure 16)*

This module is the most important module of all modules, where any two genes are strongly correlated. For each gene function, we can conclude as follows.

(1) GRN are a potent growth factor, which can promote the excessive proliferation of tumor cells.

(2) C19orf6 (also known as membralin) relates NMDAR1 receptor activity, which promotes tumor to differentiation and invasion and metastasis.

(3) RAD23A involves in negative regulation of HIV-1 replication, and VPR prevents cell division. Therefore RAD23A with high expression is to promote tumor cell division.
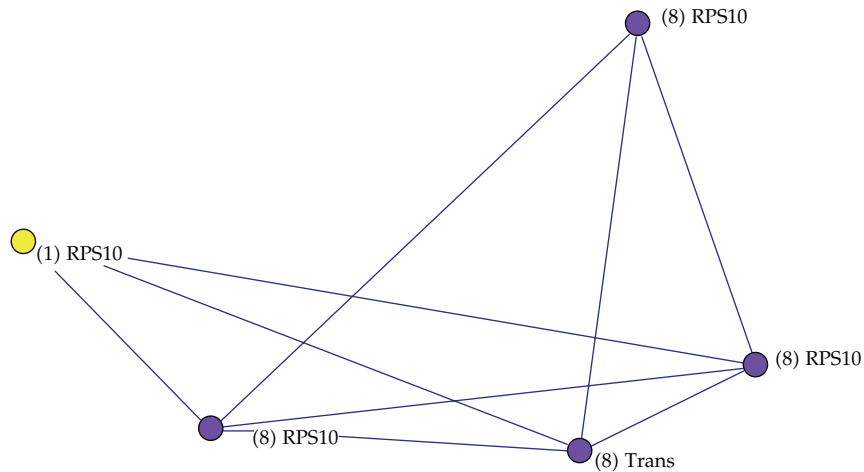
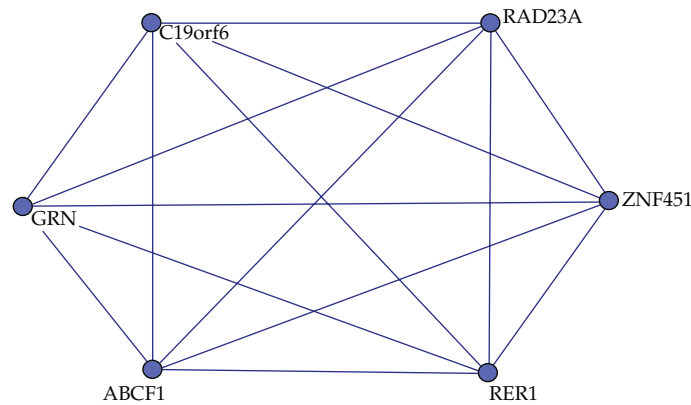**Figure 15:** The gene network of module S8.



**Figure 16:** The gene network of module S9.

(4) ZNF451 may negatively regulate the steroid hormone receptor coactivator of transcription factor (Src), where Src protein plays an important role in the proliferation of hepatoma cells during apoptosis.

(5) The protein encoded by RER1 is a multipass membrane protein, which facilitates gamma-secretase complex assembly.

(6) The protein encoded by ABCF1 is a member of the superfamily of ATP-binding cassette (ABC) transporters. This protein may be regulated by tumor necrosis factor-alpha and play a role in enhancement of protein synthesis and the inflammation process. The gene overexpression in HCC will reduce the efficiency of drug treatment.

Summing up the narrative, the module may be the secret of liver cancer cells "immortal".
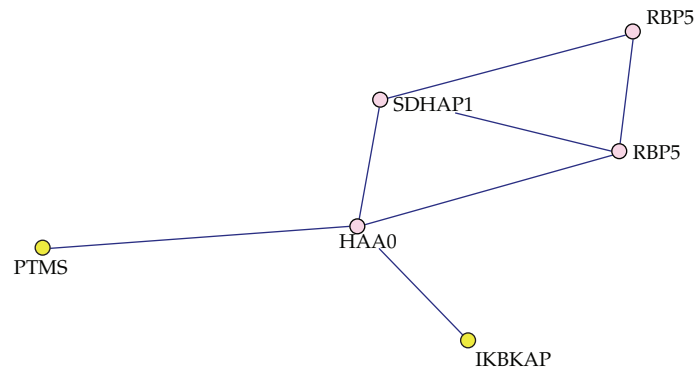
**Figure 17:** The gene network of module W6.

### 6.2.3. The Specific Modules of SCM

The specific modules of SCM are W6, W8, W9, and W12, in which only W12 is highly expressed.

### (1) Antigrowth Inhibition Module W6

From Figure 17, one can see that the important genes of this module are HAAO, SDHAP1, and RBP5. Specifically, the quinolinic acid, which is the resulting product of HAAO, inhibits the growth of hepatoma cells. SDHAP1 is a marker enzyme of mitochondrial, which provides electron to respiratory chain. Retinoic acid that produced by RBP5 after the second oxidation can inhibit the growth of hepatoma cells.

As a result, the low expression of the genes in this module is to cut off the aerobic capacity of the respiratory chain of electronic sources, making the oxidation products inhibit the growth of cancer cells to be synthesized. Therefore, this module is named as antigrowth inhibitory module.

### (2) Antimicrobial Peptides Module W8

In Figure 18, LEAP2 (liver expressed antimicrobial peptide 2) is the most important genes of the module W8, which has antibacterial activity.

### (3) Fibrinogen Module W9 (see Figure 19)

There are nine fibrinogen (FIB) genes in module W9, such as FGA, FGB, and FGG. FIB is a glycoprotein synthesized by the liver and plays an important physiological role in the coagulation process. It is worthy to point out that the FIB increases in early stage, but decreases in advanced liver cancer [18]. This module is a low expression; therefore the data should be from advanced liver cancer.

### (4) Antiapoptotic Module W12

From Figure 20, the most important three genes in the module W12 are SHARPIN, CYC1 and PUF60. SHARPIN interferes with TNF-induced cell death [19], CYC1 access to electrons
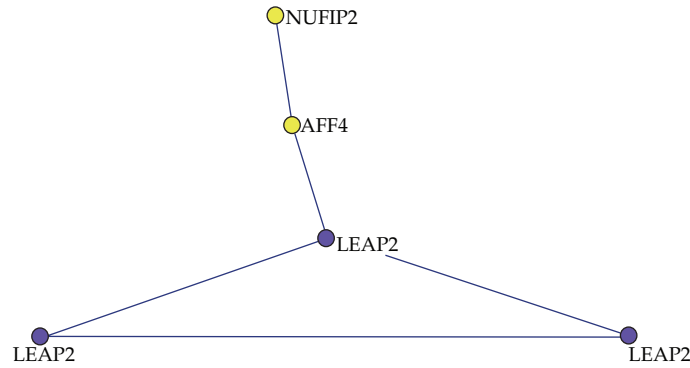
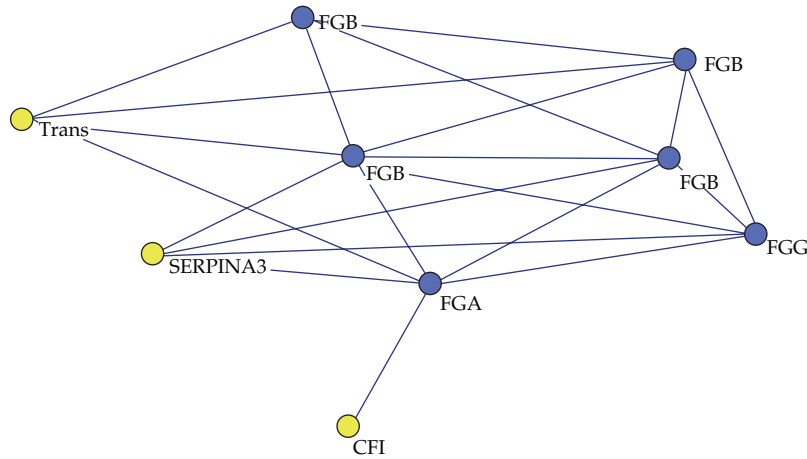**Figure 18:** The gene network of module W8.



**Figure 19:** The gene network of module W9.

for respiratory chain, and PUF60 may increase a greater degree of apoptosis resistance of cancer cells. This module provides electrons of respiratory chain to smoothly synthesis anti-apoptotic protein; therefore this module plays the role of antiapoptotic.

## 7. Conclusions

By using the Pearson agglomerative method (PAM) and Pearson correlation coefficient (PCC) modularity, we have investigated the modules decompositions and the decompositions valuations for liver cancer genes. By using the data from Chen et al. [10], and the proposed methods in this study, we have obtained 13 very strong correlation modules and 14 strong correlation modules. In addition to some common modules, we have found a number of new functional modules.

Coagulation modules are the hemostatic module S1 and fibrin module W9. It is noted that the fibrinogen will be a huge increase in the early liver cancer, but in advanced liver cancer, the fiber protein content would be down to a level slightly lower than normal. Fibrinogen can be used as one of the detection of early stage liver cancer.
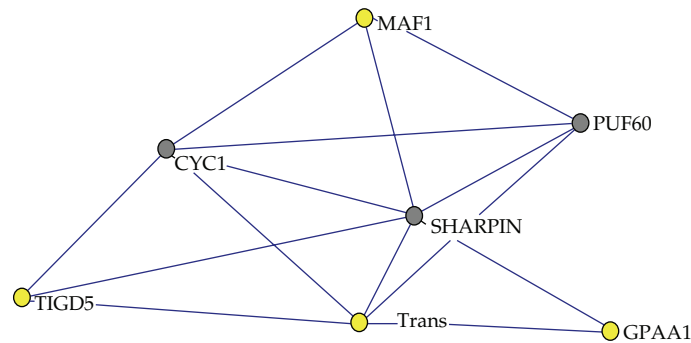
**Figure 20:** The gene network of module W12.

With respect to the reasons of the cancer cells unlimited reproduction, we have found antiterminate module S8, die module S9, antigrowth inhibition module W6, and antiapoptotic module W12. In which only W6 is in a low expression, and the others are highly expressed. One of the most modules is the immortal module S9, which may be the command center of unlimited reproduction of the entire tumor.

There are two modules associated with the metals. Iron regulation module S6 and metalloproteins module S7. These two modules are in low expression, but their functions are very different. (1) Module S6 is to increase the iron content, making more iron ions combined with more oxygen, and provide a steady stream of energy for the proliferation and metastasis of liver cancer. (2) Metalloproteins in module S7 relates to tumor differentiation, and the content of which is low in liver cancer.

## References

[1] P. Yan, R. Li. J. Chen, Y. Li et al., "Appliction of random matrix theory to microay data for discovering functional gene modules of hepatocellular carcinoma," *Acta Biophysica Sinica*, vol. 25, no. 3, pp. 192–202, 2009.

[2] C. Ling, Y. Lu, J. K. Kalsi et al., "Human hepatocyte growth factor receptor is a cellular coreceptor for adeno-associated virus serotype 3," *Human Gene Therapy*, vol. 21, no. 12, pp. 1741–1747, 2010.

[3] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[4] M. Newsman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, Article ID 066133, 2004.

[5] X. G. Ruan and J. L. Wang, "Method for finding tumor functional modules using DNA microarray data," *Journal of Beijing University of Technology*, vol. 33, no. 4, pp. 366–371, 2007.

[6] E. Newman and M. Grivan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, Article ID 026113, 2004.

[7] J. Scott, *Social Network Analysis: A Handbook*, London, UK, 2002.

[8] X. Wang, *The Application of Multivariate Analysis*, Tsinghua University Press, 3rd edition, 2009.

[9] L. Wang, G. Dai, and H. Zhao, "Research on modularity for evaluating community structure," *Computer Engineering*, vol. 36, no. 14, pp. 227–229, 2010.

[10] X. Chen, S. T. Cheung, S. So et al., "Gene expression patterns in human liver cancers," *Molecular Biology of the Cell*, vol. 13, no. 6, pp. 1929–1939, 2002.

[11] S. I. Kitajiri, T. Sakamoto, I. A. Belyantseva et al., "Actin-bundling protein TRIOBP forms resilient rootlets of hair cell stereocilia essential for hearing," *Cell*, vol. 141, no. 5, pp. 786–798, 2010.

[12] C. Rimkus, J. Friederichs, A. L. Boulesteix et al., "Microarray-based prediction of tumor response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer," *Clinical Gastroenterology and Hepatology*, vol. 6, no. 1, pp. 53–61, 2008.

[13] A. Warth, T. Muley, M. Meister et al., "Loss of aquaporin-4 expression and putative function in non-small cell lung cancer," *BMC Cancer*, vol. 11, article 161, 2011.

[14] S. Krizkova, I. Fabrik, V. Adam et al., "Metallothionein-a promising tool for cancer diagnostics," *Bratislavske Lekarske Listy*, vol. 110, pp. 429–447, 2009.

[15] R. B. Bhavsar, L. N. Makley, and P. A. Tsonis, "The other lives of ribosomal proteins," *Human Genomics*, vol. 4, no. 5, pp. 327–344, 2010.

[16] M. Shuda, N. Kondoh, K. Tanaka et al., "Enhanced expression of translation factor mRNAs in hepatocellular carcinoma," *Anticancer Research*, vol. 20, no. 4, pp. 2489–2494, 2000.

[17] X. Luo, H. H. Hsiao, M. Bubunenko et al., "Structural and functional analysis of the E.coli NusB-S10 transcription antitermination complex," *Molecular Cell*, vol. 32, no. 6, pp. 791–802, 2008.

[18] X. Guo, M. Chen, L. Ding et al., "Application of cox model in coagulation function in patients with primary liver cancer," *Hepato-Gastroenterology*, vol. 58, no. 106, pp. 326–330, 2011.

[19] B. Gerlach, S. M. Cordier, A. C. Schmukle et al., "Linear ubiquitination prevents inflammation and regulates immune signalling," *Nature*, vol. 471, no. 7340, pp. 591–596, 2011.