A Novel Analytical Method for Structural Characteristics of Gene Networks and Its Application

Shudong Wang, Yuanyuan Zhang, Kaikai Li, Dazhi Meng
College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, China
Email: wangshd2008@yahoo.com.cn

Received July 29, 2012; revised August 30, 2012; accepted September 10, 2012

ABSTRACT

Analyzing gene network structure is an important way to discover and understand some unknown relevant functions and regulatory mechanisms of organism at the molecular level. In this work, mutual information networks and Boolean logic networks are constructed using the methods of reverse modeling based on gene expression profiles in lung tissues with and without cancer. The comparison of these network structures shows that average degree, the proportion of non-isolated nodes, average betweenness and average coreness can distinguish the networks corresponding to the lung tissues with and without cancer. According to the difference of degree, betweenness and coreness of each gene in these networks, nine structural key genes are obtained. Seven of them which are related to lung cancer are supported by literatures. The remaining two genes AKT1 and RBL may have important roles in the formation, development and metastasis of lung cancer. Furthermore, the contrast of these logic networks suggests that the distributions of logic types are obviously different. The structural differences can help us to understand the mechanism of formation and development of lung cancer.

Keywords: Systems Biology; Gene Network; Logic Network; Structural Parameter; Lung Cancer

1. Introduction

Lung cancer is one of the most common and lethal malignancy carcinomas, and its formation, development and metastasis is an extremely complex polygene regulatory process [1]. Studies have shown that formation of lung cancer involves activation of oncogenes, inactivation of cancer suppressor genes and mutation of apoptosis-related genes resulting in cell proliferation and abnormal death. Therefore, finding the genes critical to the formation and development of a disease from potentially disease-related genes is of significance to the diagnosis and cure of the disease and drug design. This is an important component in the research of bioinformatics [2]. Now, many researchers recognized genes closely related to lung cancer classification, early diagnosis, prognosis and treatment using microarray technology. Battacharjee et al. [3] analyzed gene expression of tissues with and without cancer using Affymetrix U195A oligonucleotide microarray, and divided pulmonary adenocarcinoma into four subtypes. Hayes et al. [4] determined three pulmonary adenocarcinoma subtypes with clinical significance using DNA microarray. Borczuk et al. [5] built class prediction models for lung cancer histology and for cancer outcome and found that gene expression profiles of biopsy specimens of lung cancers can identify unique tumoral signatures that provide information about tissue morphology and prognosis. In order to identify a gene signature predictive of recurrence in primary lung adenocarcinoma, Larsen et al. [6] analyzed gene expression profiles in a training set of 48 node-negative tumors (stage I-II) and compared tumors from cases who remained disease-free for a minimum of 36 months with those from cases whose disease recurred within 18 months of complete resection, and obtained a 54-gene signature that predicts the risk of recurrent disease independently of tumor stage. Also, Chen et al. [7] identified sixteen genes correlated with survival among patients with non-small-cell lung cancer (NSCLC) by analyzing microarray data and risk scores and found five-gene (DUSP6, MMD, STAT1, ERBB3, and LCK) signature is closely associated with relapse-free and overall survival among patients with NSCLC. Tsai et al. [8] studied overexpression or knockdown of gene HLJ1 in human lung adenocarcinoma CL1-5 cells and discovered that HLJ1 expression inhibits lung cancer cell proliferation, anchorage-independent growth, tumorigenesis, cell motility, and invasion, and slowed cell cycle progression.

In post-genomic era, biologists paid more attention to functions and dynamic laws of whole genome, further understood the essence of life from an integrative and systemic perspective [9]. In recent years, scientists have developed many models of gene regulatory network. For
instance, Boolean network, Bayesian network, linear model and differential equation model [10-13]. Jang et al. [14] applied decision tree method to construct gene network for disease tissues. They focused mainly on the relationships between coactions of genes and pathogenesis of a complex disease and explored a new way to study genetic complexity of disease from the view of relationship and pathway. Bowers et al. [15] proposed a computational approach-logic analysis of phylogenetic profiles (LAPP) to identify detailed triplet relationships among proteins on the basis of genomic data. It can help biomedical researchers to understand potential logical relationships and discover novel biological mechanism. Now, this method has been studied in both theories and applications [16-21].

In this work, according to the viewpoint that the structures of matters can determine their functions, we can catch disease-risk genes through analyzing the differential structure of gene network in lung tissues with and without cancer. Different from other methods [10-13], we can find the structural difference of gene networks between normal and lung cancer easily. First, mutual information networks of carcinoma-related genes are constructed using their expression profiles under the contexts of normal and two types of lung cancer (adenocarcinoma and small cell lung cancer, abbreviated as AC and SCLC respectively). From the analysis of several structural parameters for these networks, it shows that average degree, the proportion of non-isolated nodes, average betweenness and average coreness can distinguish significantly the networks corresponding to normal and diseased tissues. Computing the degree, betweenness and coreness of each gene in the three gene networks, we find nine structural key genes. Seven of them which are related to lung cancer are supported by literatures. The validity of this approach is about 78%, while the experimental methods only can determine one or two disease-risk genes. Furthermore, we predict the remaining two genes AKT1 and REL may have important roles in the formation, development and metastasis of lung cancer. Second, gene logic networks of normal and two types of lung cancer are constructed. Comparing the distributions of 2-order logic types in these networks, we find that logic types in gene logic network of normal tissues include logic types 1, 3, 5, 8, while logic types 5, 8 and logic type 8 do not appear in that of AC and SCLC respectively. Also, logic types 2, 6 and 2, 4, 6 emerge in that of AC and SCLC respectively. The difference of the distribution of 2-order logic types in the above gene logic networks of normal and two types of lung cancer may be a reason or result of the formation and development of lung cancer. It provides some enlightenment roles to study the intrinsic mechanism of lung cancer.

2. Reverse Network Modeling Structural Parameters of Network

Now, there are two ways to build a network model for a biological system, forward and reverse network modelings. The former is based on studies of individual components of the system and, hence, is appropriate to a system for which there exist lots of experiment data in literature. However, reverse network modeling is based on the analysis of high-throughput data (from DNA chips and some other recently developed techniques) to mine regulatory mechanisms among the components of the system. Recently, reverse network modeling has been broadly applied to the studies of various biological systems [10-14, 22-24]. In 2006, Werhli et al. [22] compared the accuracy of reconstructing gene regulatory networks with three different modeling and inference paradigms Relevance networks, graphical Gaussian models and Bayesian networks through constructing Raf signal transduction network, and found that Bayesian networks and graphical Gaussian models outperform Relevance networks. Using the method of reverse network modeling, Perkins et al. [23] investigated gene regulatory network of drosophila and explained the activation of genes. Wang et al. [24] constructed gene logical network for Arabidopsis subject to external stimuli and further analyzed dynamical behaviors of the logical networks.

2.1. Mutual Information

A gene expression profile is a vector whose components are its expressions in many different experiments. For convenience, we denote the gene expression profiles by their corresponding genes. We describe the relationships between genes with mutual information. For example, the mutual information of genes $A$ and $B$ means that of their expression profiles and is quantified as follows:

$$I(A;B) = H(A) + H(B) - H(A,B)$$

where

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

is the Shannon entropy of vector $X$, $H(A,B)$ is the joint entropy of genes $A$ and $B$. Larger values of $I(A;B)$ imply closer interrelation between genes’ expression profiles. In case of $I(A;B) = 0$, genes’ expression profiles are independent.

2.2. LAPP Method

Bowers et al. [15-17] utilized uncertainty coefficient (abbreviated as $U$) to measure the logical relationships and identify logic types of 1-, 2-order among genes $A$, $B$, $C$. The relationships between/among two and three genes are called 1-, 2-order one, respectively. In this work, we
identify logical relationships among genes by LAPP method based on 0 - 1 expression profile data of carcinoma-related genes.

LAPP method can determine logical relationships of 1-, 2-order by calculating the uncertainty coefficient of proper functions of every order among genes. The 1-order proper relationship between genes $A$ and $B$ is determined in the following formula:

$$U(B \mid f_1(A)) = \frac{H(B) + H(f_1(A)) - H(B, f_1(A))}{H(B)}$$  (2)

where $H$ refers to the entropy of the individual or joint distributions. $U(Y/X)$ denotes the uncertainty coefficient of the influence of $X$ on $Y$. The size of $U$ value denotes the statistical possibility of the uncertainty logic of 1-order of $X$ to $Y$. $f_1$ is one of 1-order proper functions of $A$ to $B$. The 1-order proper functions are divided into two types (Table 1). The first is $B = A$, namely, the presence of $A$ leads to the one of $B$ (equivalent with $\neg B = \neg A$); the second is $B = \neg A$, namely, the presence of $A$ leads to the absence of $B$ (equivalent with $\neg B = A$).

Similarly, the uncertainty coefficient of 2-order proper function is calculated in the following expression:

$$U(C \mid f_2(A,B)) = \left( H(C) + H(f_2(A,B)) - H(f_2(A,B)) \right) / H(C)$$  (3)

where $f_2$ is one of 2-order proper functions. There are ten 2-order proper functions and eight logic types (Table 1).

Let $u_i$ denote thresholds of $i$-order uncertainty coefficients ($i = 1, 2$) and $d'$ a small number and a difference threshold. When the difference of the uncertainty coefficients $U(B \mid f_1(A))$ and $U(A \mid f_1(B))$ is greater than a given value $d'$, namely: only if

$$U(B \mid f_1(A)) - U(A \mid f_1(B)) > d'$$

and

Table 1. List of 1-, 2-order proper functions and logic types. Obviously, 1-order proper functions and logic types are the same.

<table>
<thead>
<tr>
<th>order</th>
<th>Proper function and its serial number</th>
<th>Logic type and its serial number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-order</td>
<td>1. $B = A$</td>
<td>1. $B = A$</td>
</tr>
<tr>
<td></td>
<td>2. $B = \neg A$</td>
<td>2. $B = \neg A$</td>
</tr>
<tr>
<td></td>
<td>1. $C = A \land B$</td>
<td>1. $C = A \land B$</td>
</tr>
<tr>
<td></td>
<td>2. $C = \neg (A \land B)$</td>
<td>2. $C = \neg (A \land B)$</td>
</tr>
<tr>
<td></td>
<td>3. $C = A \lor B$</td>
<td>3. $C = A \lor B$</td>
</tr>
<tr>
<td></td>
<td>4. $C = \neg (A \lor B)$</td>
<td>4. $C = \neg (A \lor B)$</td>
</tr>
<tr>
<td></td>
<td>5. $C = A \land \neg B$</td>
<td>5. $C = A \land \neg B$ (5-1)</td>
</tr>
<tr>
<td></td>
<td>6. $C = \neg (A \land B)$</td>
<td>6. $C = \neg (A \land B)$ (5-2)</td>
</tr>
<tr>
<td></td>
<td>7. $C = A \lor \neg B$</td>
<td>7. $C = A \lor \neg B$ (6-1)</td>
</tr>
<tr>
<td></td>
<td>8. $C = \neg (A \lor B)$</td>
<td>8. $C = \neg (A \lor B)$ (6-2)</td>
</tr>
<tr>
<td></td>
<td>9. $C = (A \iff B)$</td>
<td>9. $C = (A \iff B)$</td>
</tr>
<tr>
<td></td>
<td>10. $C = \neg (A \iff B)$</td>
<td>10. $C = \neg (A \iff B)$</td>
</tr>
</tbody>
</table>

the relationship between genes $A$ and $B$ is $A \rightarrow B$. We set threshold $u_2$ to judge the existence of 2-order logical relationships in the following formula:

$$\begin{align*}
U(C \mid f_2(A,B)) &> u_2 \\
U(C \mid f_1(A)) &< u_1 \\
U(C \mid f_1(B)) &< u_1
\end{align*}$$  (4)

According to the above description, we can determine all the logical relationships and logic types among the investigated genes. We regard all of the logical relationships as a gene network with 1-, 2-order logics, where each node in the network represents a gene.

2.3. Structural Parameters of Network

Now, let us briefly introduce some structural parameters of network that are used to describe network structures in this work (see [25,26] for detailed explanations). Set $G = (V, E)$ be a complex network with node-set $V = \{1, 2, \cdots, N\}$ and edge-set $E$.

1) Average degree ($K$)

The degree of a node is the number of nodes adjacent to it. The average degree is the average value of degrees of all nodes, denoted by $K$.

2) Average path length ($L$)

The distance between nodes $i$ and $j$ in $V$ is defined as the smallest number of edges that connect nodes $i$ to $j$, denoted by $d_{ij}$. The average path length $L$ of the network is defined as

$$L = \sum_{i \neq j} d_{ij} / (0.5N(N-1))$$

3) Average clustering coefficient ($C$)

The clustering coefficient of node $i$, denoted by $C_i$, is equal to the proportion of the edges among its adjacent nodes in the possible edges. The average clustering coefficient is the average value of clustering coefficients of all the nodes.

4) Modularity ($Q$)

In a complex network, the concept of community is a good tool to describe network structures and provides better understanding of network functions. Newman [26] proposed modularity to measure the probability of a network having communities. Suppose that network $G$ contains $l$ communities $G_1, G_2, \cdots G_l$. Define symmetric matrix $M = (m_{ij})_{l \times l}$, where $h_{ij}$ represents the ratio of the number of edges between two communities $G_i, G_j$ to the total number of edges of the network. The modularity is defined as:

$$Q = \sum_i Q_i = \sum_i \left( m_{ij} - a_i^2 \right)$$
where $a_i$ is the sum of all elements in the $i$th row of $M$, representing the ratio of the number of edges adjacent to community $G_i$ to that of all edges.

5) Proportion of non-isolated nodes ($R$)

Proportion of non-isolated nodes, denoted by $R$, is the ratio of the number of the nodes with degree not 0 to that of all nodes.

6) Average betweenness ($B$)

\[ B_n = \frac{\sum_{(i,j)} d_{ij}}{d_{ij}} \]

is the betweenness of the node $n$, where $d_{ij}$ presents the number of the shortest paths between the node $i$ and $j$, and $d_{ij}$ presents that of paths which pass the node $n$. The betweenness of a node reflects its influence in the network. The average value of the betweennesses of all nodes in the network is average betweenness of the network.

7) Average coreness ($H$)

$k$-core of a graph is the subgraph after getting rid of the nodes whose degrees are smaller than or equal to $k$ repeatedly. If a node exists in a $k$-core, but is removed in $(k + 1)$-core, the coreness of the node is $k$. The average value of the corenesses of all nodes in the network is average coreness of the network.

3. Material and Numerical Experiments

3.1. Data Sources and Processing

1) Data sources

The gene expression data we work on are all from GPL96 and GPL570 in NCBI. The sample data set for normal tissue, AC and SCLC includes 73, 68 and 39 samples respectively. These data are from GSE2771, GSE12667 and GSE15240 respectively. For description convenience, we refer to the networks corresponding to these three data sets as normal network, adenocarcinoma network (abbreviated as AC network) and small cell lung cancer network (abbreviated as SCLC network). Almost all of these data sets include p-values and P-M-A (P, A and M stand for presence, absence and margin respectively) except for some samples. We process these three data sets (corresponding to normal tissue, AC and SCLC) by MAS 5.0 of Expression Console, and obtain values, p-values and P-M-A for each sample of each gene. Then we represent P with 1, A and M with 0 in these three processed data sets. Our work is based on these three databases (corresponding to normal, AC and SCLC group)\(^1\).

2) Selection of carcinoma-related genes

It is out of our scope to construct and analyze the networks for all genes in the databases (each of the three databases contains expressions of more than 20,000 genes). Our main focus is on carcinoma-related genes [27]. From the databases we choose 301 genes that are known to be linked with cancers. These genes correspond to 801 probes. In the cases where several probes correspond to one gene, the highest expression value is chosen to form the gene’s expression profile. We obtain 283 genes in normal group and 286 genes in AC and SCLC group. Choosing the intersection of these genes, and using Bonferroni test to compare the difference of genes’ expression profiles between normal and lung cancer groups respectively, we choose the significant level 0.0001 and obtain 91 genes which both have significant difference between normal and two types of lung cancer. We focus on mutual information and logical networks of these genes in each database\(^2\).

3) Data processing

To calculate mutual information between arbitrary two genes, we discretize the p-values in each database as follows: a) Select the ranges [Min, Max] for p-values in the intervals [0,0.05] and (0.05,1] and divide them into 10 portions respectively such that each portion contains almost the same number of p-values. Order the portions in the number order and denote them by 1st, 2nd, ··· 20th interval, respectively; b) Replace the p-values in an interval by its labeling value. Obviously, the granularity of our discretization is finer than that of 0 - 1 discretization and hence our discretization loses less information contained in the 0 - 1 discretization. Therefore, it is reasonable to believe that the mutual information networks based on our finer discretization better reflect the nature of the gene regulatory system.

3.2. Establishment of Mutual Information Gene Network

Based on above discretized data, we can calculate mutual information values of any two genes by formula (1) and hence obtain a complete network of all genes in each database. Note that the ranges for mutual information values for our three databases are different. For the purpose of comparing the networks, we normalize the mutual information values for each database as

\[ x' = (x - \text{min})/(\text{max} - \text{min}) \]

where $x$ and $x'$ represent the original and normalized mutual information values. max and min are the maximum and minimum values of the mutual information values before normalization respectively.

In order to highlight the structural characteristics of the networks so that valuable biological conclusions can be drawn, it is necessary to choose a threshold value to carry out coarse graining on normalized mutual information. We build three gene networks (with normalized mutual
the cases of normal, AC and SCLC group. For each network, we compute seven structural parameters: average degree $K$, average path length $L$, average clustering coefficient $C$, modularity $Q$, proportion of non-isolated nodes $R$, average betweenness $B$ and average coreness $H$. These structural parameters are plotted versus threshold values $T$ in Figure 1. From Figure 1, we can see that these structural parameters $K$, $R$, $B$ and $H$ can distinguish normal network from AC and SCLC networks in a broad range of the threshold variation. It shows that the difference between normal and lung cancer which do not dependent on the choosing of threshold.

We define the degree of differentiation of each structural parameter as the proportion of the distinguishing interval of threshold in $[0, 1]$, denote the degree of differentiation that the values of structural parameters of normal tissues are greater or smaller than that of lung cancer by $G$ differentiation or $S$ differentiation respectively. The results are show in Table 2. In order to show the specific of these networks, we randomize the genes’ expression profiles keeping the equal number of p-values in $[0, 0.05]$ and $(0.05, 1]$. We obtain the changes of those structural parameters versus thresholds in randomized networks (Figure 2) and compute the degrees of differentiation of those structural parameters (Table 2). Compared with these results in real data, these structural parameters can distinguish normal and lung cancer in their structures. According to the view that system’s structure decides its function, it shows that the functions of gene networks have significant difference in normal and lung cancer.

We search genes which have largest contribution to these differences through calculating the difference of the degree, betweenness and coreness of each gene in the three gene networks. We get the top ten genes whose degrees, betweennesses or corenesses are largest (Table 3), which are called structural key genes. We get the intersection of these genes and find that there are seven genes contributing to the difference of structures of gene networks of AC and normal, also there are seven genes between SCLC and normal tissues. Genes FN1 and REL are special between AC and normal tissues, and genes TOP2A and RBL1 between SCLC and normal tissues.

![Figure 1. Plots of seven structural parameters versus threshold values T in real networks.](image)

![Figure 2. Plots of seven structural parameters versus threshold values T in randomized networks.](image)

### Table 2. Comparison the degree of differentiation of real data with that of randomized data.

<table>
<thead>
<tr>
<th></th>
<th>$K$</th>
<th>$L$</th>
<th>$C$</th>
<th>$Q$</th>
<th>$R$</th>
<th>$B$</th>
<th>$H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real data</td>
<td>G differentiation</td>
<td>83%</td>
<td>39%</td>
<td>27%</td>
<td>30%</td>
<td>81%</td>
<td>81%</td>
</tr>
<tr>
<td></td>
<td>S differentiation</td>
<td>0</td>
<td>26%</td>
<td>5%</td>
<td>21%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Randomized data</td>
<td>12%</td>
<td>17%</td>
<td>13%</td>
<td>12%</td>
<td>0</td>
<td>14%</td>
<td>18%</td>
</tr>
</tbody>
</table>

Copyright © 2012 SciRes.
Table 3. The top ten genes of the difference of degree, betweenness and coreness between normal and lung cancer networks respectively. The negative value presents that the degree or betweenness or coreness of the gene in gene network of disease stage is smaller than that of normal tissues. The differences of degree, betweenness and coreness are abbreviated as DD, BD and CD respectively.

<table>
<thead>
<tr>
<th>Genes</th>
<th>(AC and normal) DD</th>
<th>Genes</th>
<th>(SCLC and normal) DD</th>
<th>Genes</th>
<th>(AC and normal) BD</th>
<th>Genes</th>
<th>(SCLC and normal) BD</th>
<th>Genes</th>
<th>(AC and normal) CD</th>
<th>Genes</th>
<th>(SCLC and normal) CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>FN1</td>
<td>–52</td>
<td>PIK3CA</td>
<td>–54</td>
<td>PIK3CA</td>
<td>–930</td>
<td>PIK3CA</td>
<td>–955</td>
<td>FN1</td>
<td>–41</td>
<td>PIK3CA</td>
<td>–41</td>
</tr>
<tr>
<td>MSH2</td>
<td>–50</td>
<td>MSH2</td>
<td>–50</td>
<td>FN1</td>
<td>–399</td>
<td>EMP1</td>
<td>354</td>
<td>MSH2</td>
<td>–40</td>
<td>MAPK9</td>
<td>–41</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>–49</td>
<td>NRAS</td>
<td>–49</td>
<td>MAPK9</td>
<td>–256</td>
<td>NRAS</td>
<td>–320</td>
<td>REL</td>
<td>–40</td>
<td>RBL1</td>
<td>–41</td>
</tr>
<tr>
<td>REL</td>
<td>–47</td>
<td>RBL1</td>
<td>–47</td>
<td>RBL1</td>
<td>–216</td>
<td>RBL1</td>
<td>–244</td>
<td>AKT1</td>
<td>–38</td>
<td>MSH2</td>
<td>–40</td>
</tr>
<tr>
<td>AKT1</td>
<td>–46</td>
<td>MAPK9</td>
<td>–46</td>
<td>MSH2</td>
<td>–211</td>
<td>MAPK9</td>
<td>–211</td>
<td>PIK3CA</td>
<td>–37</td>
<td>NRAS</td>
<td>–40</td>
</tr>
<tr>
<td>MAPK9</td>
<td>–45</td>
<td>AKT1</td>
<td>–46</td>
<td>AKT1</td>
<td>–180</td>
<td>AKT1</td>
<td>–185</td>
<td>PTCH1</td>
<td>–36</td>
<td>AKT1</td>
<td>–40</td>
</tr>
<tr>
<td>CXCL2</td>
<td>–43</td>
<td>CHEK1</td>
<td>–45</td>
<td>NRAS</td>
<td>–150</td>
<td>CHEK1</td>
<td>–179</td>
<td>EMP1</td>
<td>–36</td>
<td>MAFG</td>
<td>–40</td>
</tr>
<tr>
<td>ECT2</td>
<td>–42</td>
<td>SUPT4H1</td>
<td>–44</td>
<td>TOP2A</td>
<td>–145</td>
<td>SUPT4H1</td>
<td>–160</td>
<td>KDR</td>
<td>–35</td>
<td>PTCH1</td>
<td>–40</td>
</tr>
<tr>
<td>AXL</td>
<td>–42</td>
<td>PTCH1</td>
<td>–43</td>
<td>REL</td>
<td>–128</td>
<td>PTCH1</td>
<td>131</td>
<td>MAPK9</td>
<td>–35</td>
<td>REL</td>
<td>–39</td>
</tr>
</tbody>
</table>

3.3. Construction of Gene Logical Networks

Looking for the regulatory relationships among genes can help us to understand the regulatory pathways related to diseases and discover the regulatory mechanisms of organism at the molecular level. Next, we will construct and analyze gene logical networks for normal, AC and SCLC respectively.

Based on three databases corresponding to normal, AC and SCLC group with 0-1 expression data in Section 3.1, we can calculate i-order uncertainty coefficients of any $I + 1$ ($i = 1, 2$) genes by formulas (2) - (4). It is considered that there is at most one 1-order logical relationship between any two genes $A$ and $B$. We set a relative difference:

$$d = \frac{U(A | f_i(B)) - U(B | f_i(A))}{\left(U(A | f_i(B)) + U(B | f_i(A))\right)/2}$$

The 1-order proper function $f_i$ between genes $A$ and $B$ is $B \rightarrow A$ (or $A \rightarrow B$) only if $d > 0.15$ and $U(A | f_i(B))$ (or $U(B | f_i(A)) > u_i$. Hence we obtain a complete logical network of all genes with 1-, 2-order logics in each database.

In order to eliminate the difference of 2-order logic types owing to choosing thresholds, we compare the distribution of 2-order logic types in the three networks at
different threshold pairs (Figure 3). It shows that the
distribution of 2-order logic types nearly doesn’t change
as the threshold pairs. In normal gene logic networks,
there are logic types 1, 3, 5, 8, while in that of AC, logic
types 5, 8 do not appear and logic types 2, 4, 6 emerge,
also in that of SCLC, logic types 8 does not appear and
logic 2, 6 emerge. The difference of 2-order logic types
between normal tissues and lung cancer may be a reason
or result of the formation of lung cancer.

4. Conclusions and Discussions

Based on the gene expression data of carcinoma-related
genomes expressed in normal tissues and diseased tissues
with AC and SCLC, we construct the corresponding mutual
information networks. From the comparison of the struc-
tural parameters of mutual information networks, we find
nine structural key genes: PIK3CA, FN1, MAPK9,
MSH2, NRAS, TOP2A, RBL1, AKT1 and REL, where
genomes FN1, REL and RBL1, TOP2A are special genes in
comparing the gene network of normal with those of AC
and SCLC respectively. In the constructed gene networks,
these genes are related to each other with higher mutual
information values in that of normal (greater than 0.7)
than in those of AC and SCLC (smaller than 0.05). We
consider that these genes may play important roles in
maintaining the normal functions of genome. Through
analyzing Gene Ontology of each genes, we find that five
genomes of them PIK3CA, MAPK9, MSH2, AKT1 and
TOP2A are involved in the regulation of cell apoptosis.
For example, gene FN1 is related to cell adhesion and
gene RBL1 is a key regulator in the regulation of cell
cycle. Therefore, we predict that the disorder of cell
apoptosis, cell adhesion and cell cycle might play impor-
tant roles in the formation of lung cancer.

Researches show that gene FN1 encodes fibronectin
which is involved in cell adhesion and migration proc-
esses including embryogenesis, wound healing, blood
coagulation, host defense, and metastasis. Anastellin
binds fibronectin and induces fibril formation. This fi-
bronectin polymer, named superfibronectin, exhibits en-
hanced adhesive properties. Both anastellin and superfi-
bronectin inhibit tumor growth, angiogenesis and metas-
tasis. Anastellin activates p38 MAPK and inhibits lyso-
phospholipid signaling. Nashino et al. [28] found that the
disappearance of FN reflects the aggressiveness of the
tumor and presumably plays an important role in the
prognosis of AC of the lung. So gene FN1 is an AC-
suppressor gene. The protein encoded by gene RBL1 is
similar in sequence and possibly function to the product
of the retinoblastoma 1 (RB1) gene. RB1 gene product is
a tumor suppressor protein that appears to be involved in
cell cycle regulation, as it is phosphorylated in the S to M
phase transition and is dephosphorylated in G1 phase of
the cell cycle. Both the RB1 protein and the product of
this gene can form a complex with adenovirus E1A pro-
tein and SV40 large T-antigen, with the SV40 large
T-antigen binding only to the unphosphorylated form of
each protein. Due to the sequence and biochemical simi-
larities with the RB1 protein, it is thought that the protein
encoded by gene RBL1 may also be a tumor suppressor.
Baldi et al. [29] found an inverse correlation between the
histological grading of the tumors, the development of
metastasis, and the level of expression of Rb2/p130
which is structurally and functionally similar to the
product of gene RBL1. So gene RBL1 may be a tumor
suppressor gene of lung cancer. Gene TOP2A encodes a
DNA topoisomerase which functions as the target for
several anticancer agents and a variety of mutations in this gene have been associated with the development of drug resistance. Jensen et al. [30] studied the resistant to the inhibitory action of F 11782 in SCLC patients and found that TOP2A is a target of F 11782 in vivo, and that F 11782 may act as a novel topoisomerase II poison. It shows the function of gene TOP2A in SCLC. Gene PIK3CA encodes the p110alpha catalytic subunit, Samuels et al. [31] found that PIK3CA was frequently mutated in cancers of the colon, breast, brain and lung. The majority of mutations clustered near two positions within the PI3K helical or catalytic domains and at least one hotspot mutation appeared to increase kinase activity. The protein encoded by gene MAPK9 is a member of MAP kinase family. MAP kinases act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development. This gene and MAPK8 are also known as c-Jun Nterminal kinases. This kinase blocks the ubiquitination of tumor suppressor p53, and thus it increases the stability of p53 in nonstressed cells. Nitta et al. [32] investigated the role of the JNK2a (the alias of MAPK9) isoform in NSCLC formation by examining its expression in primary tumors and by modulating its expression in cultured cell lines and found Cell lines deficient in JNK2a had decreased cellular growth and anchorage-independent growth, and the tumors were four-fold smaller in mass. Gene MSH2 is a DNA mismatch repair gene, Xinarianos et al. [33] found a relationship between p53 and MSH2, there is evidence for p53 being a transcriptional activator of MSH2, and suggested a hypothesis that MSH2 acts as a DNA-damage signaler triggering p53 overexpression. Sasaki et al. [34] found NRAS mutation in 1 of 195 patients, although the rate is small, we consider that the mutation of NRAS may be a reason of the formation of lung cancer. The relationships between MSH2, NRAS and SCLC are not clear, but owing to they have important roles in structures of gene networks between normal and SCLC, we predict these two genes are related to the formation and development of SCLC. Besides, the remaining two genes AKT1 and REL are open. We predict that they may have significant roles in the formation of lung cancer based on the view that system’s structure decides its function.

Through constructing gene logic networks corresponding to normal, AC and SCLC, we find the significant difference of the distribution of 2-order logic types. Relative to normal gene logic network, logic types 5, 8 do not appear in that of AC, and logic type 8 does not appear in that of SCLC. Also, logic types 2, 6 and 2, 4, 6 emerge in that of AC and SCLC respectively. The difference of the distribution of 2-order logic types in the above gene logic networks of normal and two types of lung cancer may be a reason or result of the formation of lung cancer. It provides some enlightenment roles to study the intrinsic mechanism of lung cancer.

Finally, we have applied this approach to the gene detection of other cancers, such as breast cancer and brain tumor, and Arabidopsis thaliana response to the special external stimulus. Through analyzing the results, we find most of genes can be supported by literatures. It further approves the validity of our approach.

5. Acknowledgements

The authors would like to thank National Natural Science Foundation of China (Nos: 60874036, 61170183), Scientific Research Foundation for Excellent Middle-Aged and Youth Scientists of Shandong Province of China (BS2011SW025), SDUST Research Fund of China (2010 KYJQ104) for the support to this work.

REFERENCES


Copyright © 2012 SciRes.