A Network Partition Algorithm for Mining Gene Functional Modules of Colon Cancer from DNA Microarray Data

Xiao-Gang Ruan, Jin-Lian Wang*, and Jian-Geng Li

Institute of Artificial Intelligence and Robotics, School of Electronic Information & Control Engineering, Beijing University of Technology, Beijing 100022, China.

Computational analysis is essential for transforming the masses of microarray data into a mechanistic understanding of cancer. Here we present a method for finding gene functional modules of cancer from microarray data and have applied it to colon cancer. First, a colon cancer gene network and a normal colon tissue gene network were constructed using correlations between the genes. Then the modules that tended to have a homogeneous functional composition were identified by splitting up the network. Analysis of both networks revealed that they are scale-free. Comparison of the gene functional modules for colon cancer and normal tissues showed that the modules' functions changed with their structures.

Key words: DNA microarray data, colon cancer, gene functional module, GN algorithm

Introduction

Cancer is a systemic disease originating in different tissues and requiring "cooperation" by other non-cancer tissues (*1*). It has been suggested that cancer can only be fully understood at the system level (*2*). The complexity of cancer necessitates the use of a system-wide approach to data analysis, which can be defined as the integration of genomics, proteinomics, and metabonomics data using computational methods (*3*). According to the previous study (*4*), integrative approaches can simplify complex cancer signatures into coordinately regulated modules. This process transforms one-dimensional cancer signatures into multi-dimensional interaction networks and extracts the regulatory mechanisms encoded in cancer gene expression (*4*). Therefore, identifying the modular organization of a metabolic network by network decomposition methods can help us better understand the organizational principles of complex biological systems.

In the past years, a number of network decomposition methods have been developed and applied to identify modules in various biological systems, including the Monte Carlo optimization method for finding tightly connected clusters of nodes (*5*), k-means and hierarchical clustering graph theory (*6*), and complex networks (*7*, *8*). The complex networks have been used to describe systems in many fields, such as the Internet, the World Wide Web, and social and biological interaction networks (*9*). Recently Girvan and Newman (*10*) have proposed a computer algorithm (GN algorithm) based on the iterative removal of edges with high betweenness scores. The GN algorithm has been exploited in the social and ecological sciences to study communities as well as in the study of biochemical pathways (*11*). However, since the GN algorithm has drawbacks in module detection, it has been improved and extended in other studies and then applied to biological networks (*12*, *13*).

Compared with other conventional methods, the GN algorithm has a demonstrated capability for accurately identifying networks of communities with known community structures (*14*). In this study, the community consists of strongly interrelated genes, and such a community corresponds to a module in most references (*12*, *13*). Since the GN algorithm is a graph-theoretical instrument relying on network properties, it does not pay any attention to the roles that the nodes play or the annotations of the nodes; however, these factors are very important from a biological view. Nodes with different roles are affected by different evolutionary constraints and pressures, thus the roles and annotations of the nodes within a module as well as the similarity between modules must be considered. Since it is a common situation that the structure of modules in a large network is unknown in advance, in order to make the GN algorithm applicable to the gene networks with unknown community

***Corresponding author.**
E-mail: wjinlian1999@emails.bjut.edu.cn

structures where both the nodes' roles and annotations must be considered, here we propose a method based on the GN algorithm for detecting gene functional modules in cancer networks.

Algorithm

Method overview

First, the expected gene sets are obtained from DNA microarray data. The Pearson's correlation coefficient (PCC) (*15*) is used to calculate each pair of genes, and the value lies between -1 and 1. The closer the value is approaching to 1 or -1, the stronger the linear relationship between the two variables. Only those genes with the absolute PCC value higher than a given threshold are selected as members of the gene sets.

Undirected weighted networks are then created from these gene sets. In the network, a vertex set $V = \{g_1, g_2, \ldots, g_i\}$ represents the genes. $E = \{\{g_i, g_j\} \mid g_i, g_j \in V \wedge |P_{ij}| \geq T\}$ is the edge set that represents the relationship between the genes, where T is the given threshold of PCC, and P_{ij} is the PCC value of g_i and g_j. A weight is assigned to each edge in the network, which is the reciprocal of its absolute PCC value, namely $w_{ij} = 1/|P_{ij}|$. Therefore, the constructed gene network is an undirected weighted network. We then use this network as the subject of our research.

Next, a partitioning process is applied to the network. Our algorithm based on the GN algorithm is used to divide the network into modules. The quality of the division is evaluated using the Q function developed by Girvan and Newman (*10*), where the division corresponds to the different values of Q, and the best division is considered to be the maximal value of Q. However, the modules detected from the networks are not assumed to be biologically functional modules. They are simply interesting artifacts within the network that provide a powerful method for organizing and presenting information from the genomic data. To assess whether the gene modules have predictive and functional relevance, we examine the gene modules' functional compositions by mapping them using the Gene Ontology (GO) database (*16*). A p-value calculated using a hypergeometric distribution is used to measure the significance of the functional modules.

Network partition algorithm

The general idea of our process is based on the GN algorithm, but the details are different as described below.

Firstly, we define A_{ij} as the element of the adjacent matrix of a network, thus

$$A_{ij} = \begin{cases} w_{ij} & \text{for } \{g_i, g_j\} \in E \text{ and } g_i \neq g_j \\ \infty & \text{for } \{g_i, g_j\} \notin E \end{cases}$$

where E is the set of edges and w_{ij} is the weight of the edge $\{g_i, g_j\}$ as described above. The Floyd's algorithm (*17*) is used to compute the length of the shortest path between every pair of vertices in the network.

Secondly, the edge betweenness of all edges are computed using the breadth-first search. The betweenness of an edge is the number of all the shortest paths running through it. We find the edge $e = \{g_1, g_2\}$ with the highest betweenness and remove it from the network, then compute the shortest paths from all other nodes to the two vertices g_1 and g_2. The nodes that are nearer to node g_1 (or g_2) than to node g_2 (or g_1) are classified into the module M_1 (or M_2). If the distance of a node g to node g_1 is the same as that to node g_2, then g is randomly classified to M_1 or M_2. Note that g_1 and g_2 are called the center of M_1 and M_2, respectively.

Thirdly, we recalculate the betweenness for all the remaining edges. If the two nodes of an edge with the highest betweenness belong to the module M_1 or M_2, we remove the edge and split M_1 or M_2 into two subnetworks, with the center on the two end nodes of the removed edge, respectively. If not, we only remove the edge. By continuing the removal of the edge with the highest betweenness from the network, the network is divided into small subnetworks and isolated nodes. The algorithm stops when all the edges are removed from the network. The algorithm of determining module structure is as follows.

Step 1: For a gene network $G(V, E)$, compute the shortest path between every two vertices using the Floyd's algorithm, and find out the edge $e = \{g_1, g_2\}$ with maximal betweenness using the breadth-first search algorithm. Delete this edge e and let $E = E - \{e\}$.

Step 2: For the current gene network $G(V, E)$, partition the vertex set V into two modules M_1 and M_2 with center g_1 and g_2, respectively. For vertex g, if the length of the shortest path between g and g_1 is less than that between g and g_2, then put g into M_1; otherwise put g into M_2. Let $S_0 = M_1$ and $S_1 = M_2$.

Step 3: Let $i = 0$, $N_1 = M_1$, and $N_2 = M_2$.

Step 4: If $|E| = 0$ then stop; otherwise go to Step 5.

Step 5: For the current gene network $G(V, E)$, compute the shortest path between every two vertices using the Floyd's algorithm, and find out the edge $e = \{g_1, g_2\}$ with maximal betweenness using the breadth-first search algorithm. Delete this edge e and let $E = E - \{e\}$.

Step 6: For the current gene network $G(V, E)$, if g_1 and g_2 belong to the same module N_j, then partition the vertex set N_j into two modules M_1 and M_2 with center g_1 and g_2, respectively. For vertex $g \in N_j$, if the length of the shortest path between g and g_1 is less than that between g and g_2, put g into M_1; otherwise put g into M_2. Let $i = i + 1$, $S_{2i} = M_1$, $S_{2i+1} = M_2$, $N_j = M_1$, and $N_{i+2} = M_2$.

Step 7: Go to Step 4.

Finally, the resulting output is the gene modules N_j $(j = 1, 2, 3, \ldots)$.

Module quality

In order to objectively identify the number of modules inherent in the network, Girvan and Newman (*10*) defined a measure Q, namely "modularity", which is used to evaluate the quality of the divisions generated by network partitioning:

$$Q = Tre - \left\|e^2\right\| \tag{1}$$

where e is a $g \times g$ matrix whose component e_{ij} is the fraction of edges in the original network connecting vertices in module i to those in module j, Tre is the trace of matrix e, and $\|x\|$ indicates the sum of the elements of matrix x. $Tre = \sum_i e_{ii}$, which gives the fraction of edges in the network connecting vertices in the same cluster. A good partition of a network must comprise the most within-module links and the least between-module links, and the objective of a module algorithm is to find the partition with the largest modularity.

Here we construct an undirected weighted network with the quality function a little different from Equation 1. We modify the "modularity" to include the weight:

$$\begin{aligned} Q &= \frac{1}{2m}\sum_{ij}(w_{ij} - \frac{w_i w_j}{2m})\ \delta(c_i, c_j) \\ &= \frac{1}{2m}\sum_{i}(w_{ij} - \frac{w_i^2}{2m}) \end{aligned} \tag{2}$$

where w_{ij} is the weight of the edge-linked nodes i and j; w_i is the weight of node i, $w_i = \sum_j w_{ij}$; the function $\delta(c_i, c_j) = 1$ if $i = j$ and otherwise $\delta(c_i, c_j) = 0$; and m is the number of edges in the network.

The functional modules are extracted from these subnetworks by mapping to the GO database. A gene is assigned with the ontology term for the process it belongs to, as well as all the parental process terms in the GO directed acyclic graph. To find out the probability of given genes occurring in certain category by chance, we use a hypergeometric distribution with 1,000 permutations to get p-values. The GO terms that are most specific for the analyzed genes will have the lowest p-value. The hypergeometric distribution is given by:

$$P(X = k \mid G, c, n) = \frac{\begin{pmatrix} c \\ k \end{pmatrix}\begin{pmatrix} G - c \\ n - k \end{pmatrix}}{\begin{pmatrix} G \\ n \end{pmatrix}} \tag{3}$$

where G is the total number of annotated genes in all the modules, n is the number of genes in a module, c is the number of selected GO classes, and k is the number of genes annotated within a module. The p-value of k genes occurring in a certain category can be calculated by summing the probabilities of a random set of n genes having $1, 2, \ldots, k$ genes of a category:

$$p = \sum_{i=0}^{k} \frac{\begin{pmatrix} c \\ i \end{pmatrix}\begin{pmatrix} G - c \\ n - i \end{pmatrix}}{\begin{pmatrix} G \\ n \end{pmatrix}} \tag{4}$$

Application

Data filtering

We then applied this method to a colon cancer microarray dataset that is publicly available at http://microarray.princeton.edu/oncology (*18*). It consists of gene expression profiles in 40 colon cancer and 22 normal colon tissue samples, with an Affymetrix Hu6800 oligonucleotide array set complementary to more than 65,000 probes. This dataset had been preprocessed before we obtained it and the number of differential genes analyzed in our study was 2,000. We deleted 12 probes used for quality control and filtered the probes for which there was at least a single measure greater than 100. Then the probes were mapped to Locus ID using NetAffx (http://www.affymetrix.com/analysis), and those probes without an assigned Locus ID were discarded. As a result, 1,707 genes met the filtering requirement.

Network construction and statistics

We first selected the expected gene set for constructing the network. We computed the PCC value of each pair of genes for this dataset (Figure 1), and set the threshold to 0.95 for colon cancer and normal tissue samples so that the selected genes are strongly correlated. The genes with the absolute PCC value larger than the threshold were selected as members of the expected gene set.

Then we constructed the network as stated before. Each gene of the dataset was represented by a vertex in the network. If the PCC value of each pair of genes satisfied the threshold, then the two genes were connected. As a result, we constructed a colon cancer gene network containing 284 vertices and 408 edges, and a normal colon tissue gene network consisted of 277 vertices and 386 edges.

We next analyzed the network features from statistical and topological perspectives. The degree distribution is shown in Figure 2, and it can be seen that both networks reveal scale-free characteristics. The degree distribution of both networks satisfies the power law. The probability of finding a node with k connections is represented by $P(k) \propto k^{-\gamma}$, where the scaling exponent γ correlates with the degree distribution of the network. On a log-log scale, the distribution is approximately a straight line, the slope of which is γ (Figure 2).

Another two parameters that describe the network's topology are the cluster coefficient C and the average length L. The cluster coefficient of the undirected network is defined as:

$$C = \frac{1}{n}\sum_{i=1}^{n}\frac{C_i}{N_i(N_i-1)/2}$$

where n is the number of nodes in the network, C_i is the number of connections between neighbors of node i, and N_i is the number of neighbors of node i. The cluster coefficient is 0.41 and 0.5 for the colon cancer gene network and the normal colon tissue gene network, respectively. The average length L is the average shortest path length between each pair of nodes, with the value of 5.75 and 5.15 for the colon cancer gene network and the normal colon tissue gene network, respectively. Therefore, both networks indicate small-world network characteristics as revealed by combination of C and L.

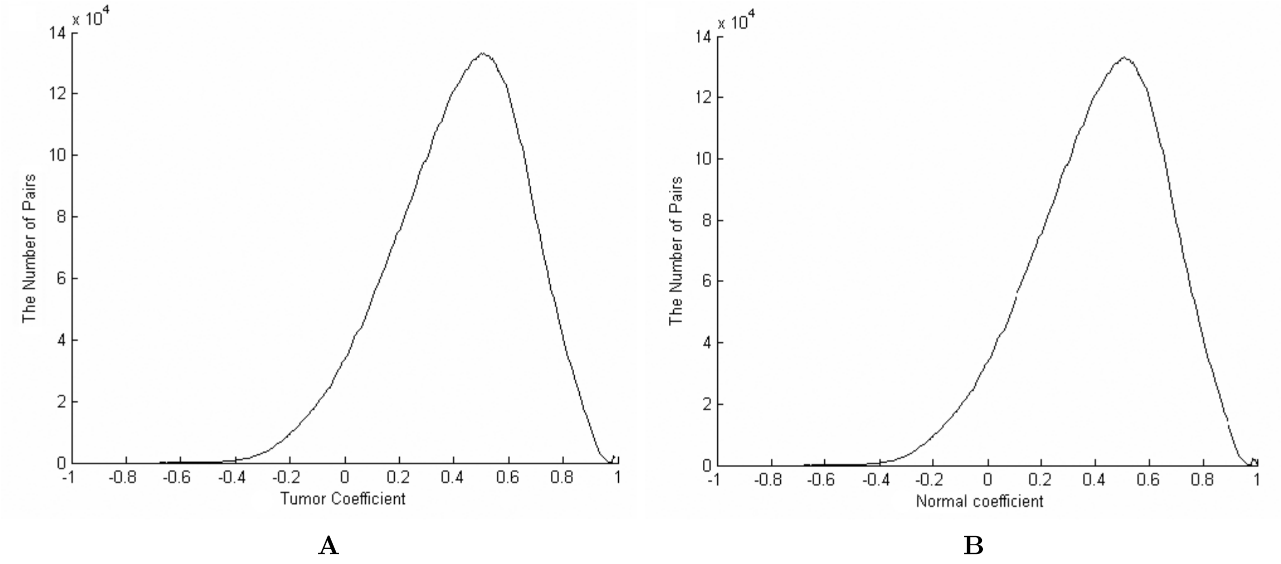

Fig. 1 Distribution of PCC value vs. the number of pairs of genes for colon cancer (**A**) and normal (**B**) tissue samples. The threshold is set to 0.95.

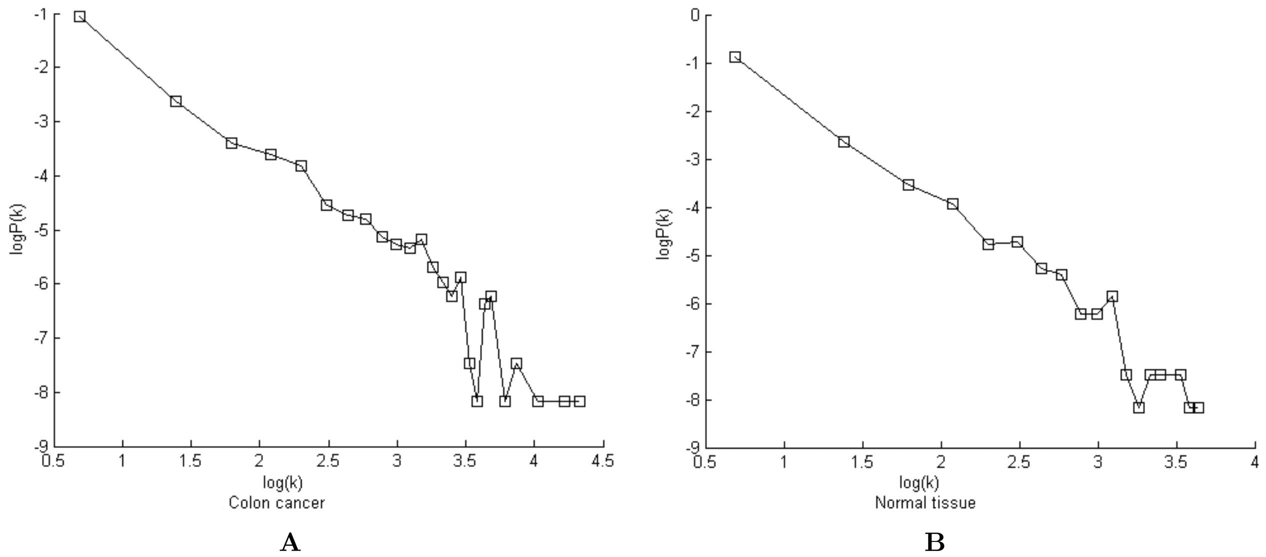

Fig. 2 The log-log scale plot of the probability $p(k)$ of finding a node with k connections for the colon cancer gene network (**A**) and the normal colon tissue gene network (**B**).

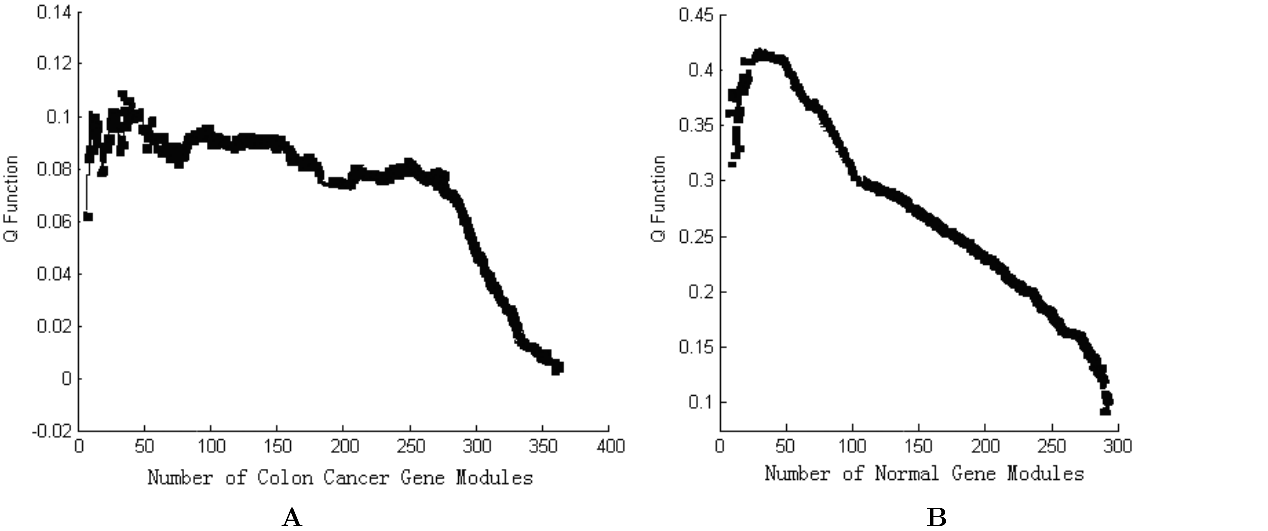

Fig. 3 Modularity distribution and the number of modules detected in the colon cancer gene network (**A**) and the normal colon tissue gene network (**B**).

Module detection

We applied the modified GN algorithm to split the network into modules as mentioned before. Each module consisted of highly correlated genes. The number of modules ranged from one to the number of all genes in the network. The modularity function was computed using Equation 2, and the distribution of modularity is shown in Figure 3. For the colon cancer gene network (Figure 3A), the highest modularity is $Q = 0.11$, which corresponds to 18 modules. For the normal colon tissue gene network (Figure 3B), the highest modularity is $Q = 0.427$, corresponding to 22 modules.

Correlation with biological processes

To assess whether the obtained gene modules have functional relevance, we mapped the genes in each module to GO functional categories. The GO hierarchy and functional information was acquired from the GO database (February 2006, http://www.geneontology.org). Other annotation information including Locus ID, gene symbol, gene title, and OMIM was acquired from NCBI (February 2006, http://www.ncbi.nlm.nib.gov/LocusLink). The p-value was computed using Equations 3 and 4. As a result, we obtained the functional categories significantly ($p <$ 0.05) correlated with the gene modules (Figure 4).

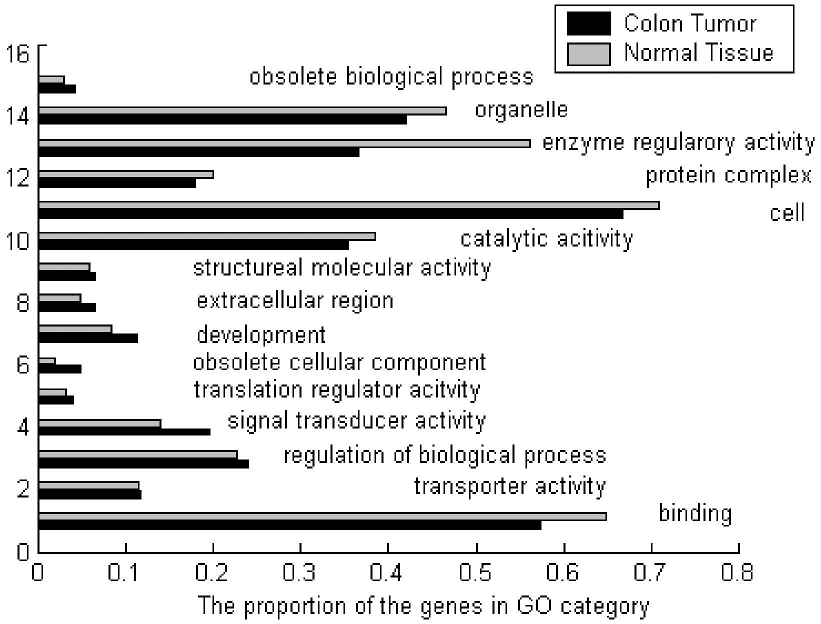

Fig. 4 The GO categories that are significantly (p <0.05) correlated with the extracted gene modules in colon cancer and normal tissues. The length of the bars represents the proportion of the annotated genes with the biological process, molecular function, and cellular component.

We then compared the differences between colon cancer and normal gene modules with respect to biological processes (Figure 5). To further quantify the annotated genes of these modules, we also analyzed the functional categories of 144 genes shared by both colon cancer and normal tissues (Figure 6). From Figures 5 and 6, it can be seen that all the functional categories that correlate with the colon cancer gene modules are well-known to be associated with cancer, such as signal transduction, regulation of cell cycle, and cell proliferation. We thus deduced that the reason for the occurrence of colon cancer is related to the dysfunction of normal colon gene modules. Most of the genes in the colon cancer gene network are involved in the functional categories correlated with colon cancer and are also highly expressed. Therefore, we conclude that normal colon cells tend to grow cancerous if the genes correlated with these functional categories are highly expressed.

Conclusion

We have presented a method for finding gene functional modules in a complex cancer gene network. Statistic analysis of the gene network suggests that it possesses small-world and scale-free characteristics. This method can be used to divide the gene network into modules containing highly correlated genes. The gene functional modules produced by this method contain groups of genes that are known to cooperate to perform common functions. The smaller the modules generated by this method, the higher the average number of significant annotations.

We have applied this method to the colon cancer microarray dataset. By comparing the colon cancer gene modules with the normal colon gene modules, we find that their differences are significantly affected by several processes known to be associated with cancer. The results demonstrate that this method is feasible for finding gene functional modules in biological networks.

Furthermore, this network partition algorithm can be used for mining functional modules of protein-protein interaction networks, as well as for mining co-regulated genes of gene regulatory networks and co-expressed genes of gene-gene networks.

Acknowledgements

We thank Drs. Rhoda Perozzi and Edmund Perozzi in Beijing University of Technology for their helpful instructions. This work was supported by the National Natural Science Foundation of China (Grant No. 60234020).

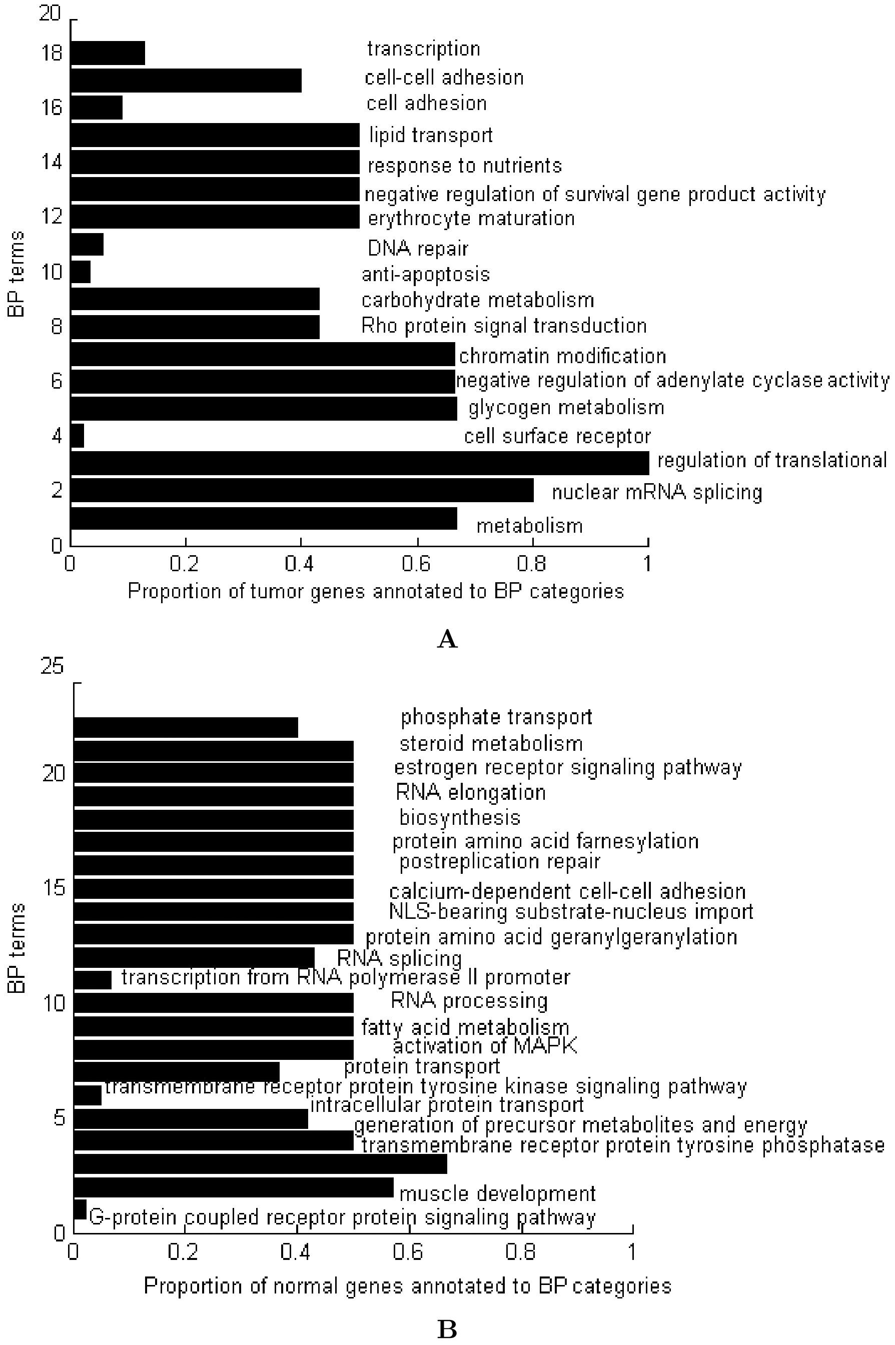

Fig. 5 Comparison of the colon cancer (**A**) and normal (**B**) gene modules correlated with significant ($p < 0.05$) biological process (BP) functional categories.

Authors' contributions

XGR participated in the study and drafted the manuscript. JLW carried out the studies on colon cancer gene modules. JGL participated in the network partition algorithm. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

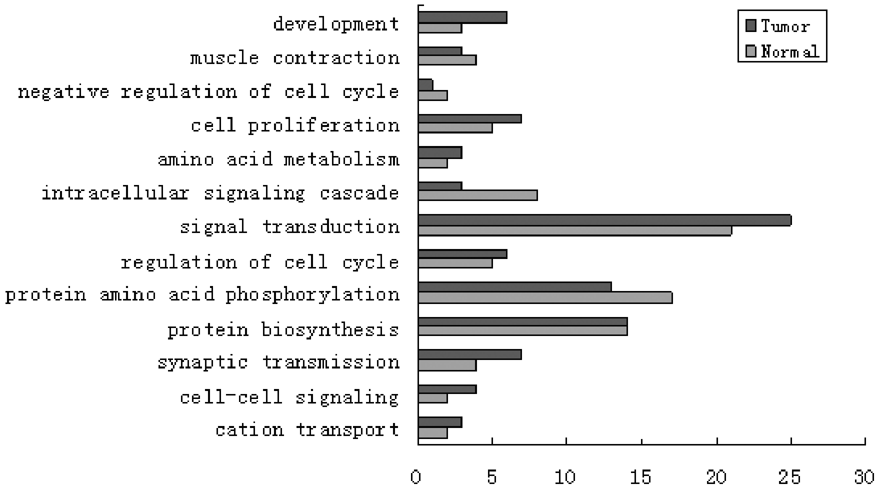

Fig. 6 Comparison of biological process (BP) functional categories of 144 genes shared by both colon cancer and normal tissue gene modules.

References

1. Fraser, A.G. and Marcotte, E.M. 2004. A probabilistic view of gene function. *Nat. Genet.* 36: 559-564.
2. Hartwell, L.H., *et al.* 1999. From molecular to modular cell biology. *Nature* 402: C47-52.
3. Nicholson, J.K., and Wilson, I.D. 2003. Understanding “global” systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* 2: 668-676.
4. Rhodes, D.R. and Chinnaiyan, A.M. 2005. Integrative analysis of the cancer transcriptome. *Nat. Genet.* 37: S31-37.
5. Spirin, V. and Mirny, L.A. 2003. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA* 100: 12123-12128.
6. Tornow, S. and Mewes, H.W. 2003. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res.* 31: 6283-6589.
7. Albert, R. and Barabasi, A.L. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phs.* 7447-7497.
8. Dorogovtsev, S.N. and Mendes, J.F.F. 2003. *Evolution of Networks: From Biological Nets to the Internet and WWW.* Oxford University Press, Oxford, UK.
9. Milo, R., *et al.* 2002. Network motifs: simple building blocks of complex networks. *Science* 298: 824-827.
10. Girvan, M. and Newman, M.E.J. 2003. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99: 7821-7826.
11. Holme, P., *et al.* 2003. Subnetwork hierarchies of biochemical pathways. *Bioinformatics* 19: 532-538.
12. Muff, S., *et al.* 2005. Local modularity measure for network clusterizations. *Phys. Rev. E* 72: 056107.
13. Guimera, R. and Nunes Amaral, L.A. 2005. Functional cartography of complex metabolic networks. *Nature* 433: 895-900.
14. Gustafsson, M., *et al.* 2006. Comparison and validation of community structures in complex networks. *Physica A* 367: 559-576.
15. Pearson, K. 1896-1897. Mathematical contributions to the theory of evolution—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc. Lond.* 60: 489-498.
16. Ashburner, M., *et al.* 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25: 25-29.
17. Floyd, R.W. 1962. Algorithm 97: shortest path. *Commun. ACM* 5: 345.
18. Alon, U., *et al.* 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96: 6745-6750.