MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-667

Identifying Biomarkers of Hepatocellular Carcinoma Based on Gene Co-Expression Network from High-Throughput Data

Ying Zhang[a,b], Zhiping Liu[b], Jing-song Li[a]

[a] *Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, Zhejiang Province, China*
[b] *Biomedical Engineering, College of Control Science and Engineering, Shandong University, Jinan, Shandong Province, China*

Abstract

In this paper, we proposed an approach systematically based on the use of gene co-expression network analyses to identify potential biomarkers for Hepatocellular Carcinoma (HCC). With the analysis of differential gene expression, we first selected candidate genes closely related to HCC from the whole genome on a large scale. By identifying the relationships between each two genes, we built up the gene co-expression network using Cytoscape software. Then the global network was clustered into several sub-modules by Markov Cluster Algorithm (MCL). And, GO-Analysis was carried out for these identified gene modules to further explore the genes obviously associated with the dysfunctions of HCC, and in result we find Hexokinase 2 (HK2) and Krüppel -like Factor 4 (KLF4) as potential candidate biomarkers to provide insights into the mechanism of the development of HCC. Finally, we evaluated the disease classification results via an SVM-based machine learning method to verify the accuracy of the classification

Keywords:
Hepatocellular Carcinoma; Cluster Analysis; Machine Learning

Introduction

HCC is an aggressive disease with a high morbidity rate and mortality rate, which is a serious threat to people's lives and health. At present, the treatment of HCC is still limited. More than half of the patients will relapse in the case of surgical resection and will suffer from more and more serious postoperative complications [1]. We can see the overall therapeutic effect of HCC is not optimistic. In addition, two thirds of patients with HCC are in the advanced stage when they are treated, then surgical resection is not the best choice. Therefore, it is particularly important to strengthen the research on the occurrence and development mechanism of HCC to effectively identify the biomarkers for clinical detection of liver cancer, so that we will achieve early detection and treatment of HCC and prolong survival time of patients.

The information of HCC genes is abundant in the gene expression profile. Over the past two decades, more and more high-throughput techniques have been developed to do the genome-wide analysis of gene expression and their interactions [2]. The gene co-expression network of HCC provides a tool for studying the gene regulation based on gene expression data, which provides favorable conditions for finding more stable biomarkers and gene targets.

Many investigators have made preliminary progress in the search for biomarkers of HCC. Through the study of HCC samples, the researchers have found a number of biomarkers which have significant changes in the expression, such as alpha-fetoprotein (AFP). Tada et al found that the percentage of AFP-13 is strongly related to the staging and size of HCC [3]. These tasks have much significance to clearing pathogenesis of HCC. At present, analysis of gene co-expression network has been widely used in complex disease research. Gene network is a complex dynamic system. In order to achieve genetic relationship mining, we need to select the community structure from the complex network, which is called gene co-expression module. The genes within these modules share similarities in physiology and gene function [4]. Therefore, analyzing the co- expression network of HCC and enriching the gene modules with similar function by clustering can help us to deeply understand the gene interaction in the process of HCC, so as to get the disease-related gene modules and comprehend its pathological mechanism systematically.

Methods

Process of Research

As shown in Figure 1 we first obtained gene expression data from the public database of HCC, which contains 20,673 genes in the sample data. After the data were normalized, 328 genes closely related to HCC were selected on the basis of Student's Test and correlation analysis. Then, the co-expression network of different genes was constructed according to the relativity of these genes, and the network was clustered to divide the network into several functional gene modules. Based on the systematical study of these gene modules, the functional modules of the selected genes were analyzed for enrichment and further exploration of genes related to HCC. Next, two candidate biomarkers were identified. Finally, the support vector machines (SVM) machine learning method was used to evaluate the disease classification of the selected gene modules and candidate biomarkers.

Data Source

HCC gene expression profile data were obtained from Gene Expression Omnibus (GEO) database of the National Center of Biotechnology Information (NCBI). Its number is GSE20948, containing 20,673 genes in 14 experimental samples and 14 control samples. The experimental samples and the control samples correspond to each other, named from GSM523800 to GSM523827. Fourteen experimental samples among them are

infected cells. They were infected with the genotype 2a HCV clone, JFH-1 at a multiplicity of infection (MOI) of 3. At 6, 12

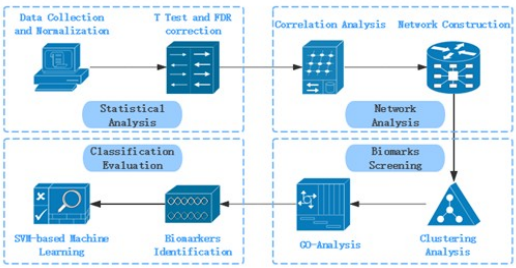


Figure 1- Map of research process

18, 24 and 48 hours' post-infection, cellular RNA was extracted and the gene expression value was measured from the gene expression profile. Another fourteen control samples were normal cells without HCV infection under the same conditions as the experimental group.

Overview of Research Methods

Statistical Analysis

The expression data are preprocessed for normalization with the gene expression console software. We apply Student's Test to achieve the initial screening of differentially expressed genes on a large scale. Using the R language as the data processing tool, we calculate the P value of two vectors of the experimental group and control group. The smaller the P value, the greater difference between the two vectors, and the greater difference in gene expression between the infected samples (disease state) and the non-infected samples (healthy state). Then False Discovery Rate (FDR) correction is performed by using the p.adjust function to obtain the FDR corresponding to each gene data. The significance of FDR is the proportion of false conclusions that are claimed to be significant. The goal is to control the false discovery rate below a certain value.

Gene Co-expression Network Analysis and Clustering

We use Pearson Correlation to calculate the Pearson Correlation Coefficient (PCC) of the expression vector between different genes to identify the first 1% of the genes with the greatest difference in disease and health status [5]. We then use Cytoscape to map all co-expressed gene pairs into a gene co-expression network (Figure 1).

In the co-expression network, the nodes represent the genes, and the edges represent the co-expression interactionscc [6]. The information can be transferred smoothly on these edges. A collection of nodes means that a set of co-expressed genes collectively carries out certain functions. In this paper, Markov clustering algorithm (MCL) is used to cluster [7]. In order to get the appropriate number of community sets, we choose the Granularity Parameter of 1.8 and the other parameters of the default parameters. After the clustering algorithm is implemented, the clustering results are represented by a number of modules, each of which contains a number of genes.

Identifying Candidate Biomarkers by GO-Analysis

For the four gene modules with the highest correlation, the functional analysis of the gene network is carried out through Network Ontology Analysis (NOA) method and Gene Card website [8]. The NOA online website provides functional gene analysis for free which gives static and dynamic analysis for gene networks. Through the functional analysis of genes, we identify two important genes, which are defined as candidate biomarkers.

Determine the Accuracy of Classification by SVM

Our candidate biomarkers are determined by our functional enrichment analysis of the genes in the gene modules. In order to further confirm the relevance of HK2, KLF4 and HCC and the accuracy of the classification, we use SVM to evaluate the effect of the classification of disease for the four most relevant gene modules. We choose the method of Leave-one-out to classify the existing data and make the ROC curve to get the accuracy index of classification [9].

Results

Screening of Differentially Expressed Gene

The t.test () function is used to test the data and get the P value of each gene data. According to the size of the P value in ascending order, we pick out the smallest P value of the first 500 genes as candidate genes. All the 500 genes selected with the smallest P value meet the required FDR range (FDR <5%). We list the first ten genes and their P value and FDR in Table 1.

Table 1- P value and FDR of the selected genes

Gene Name	*p-value*	FDR
PDGFRA	6.28E-16	1.30E-11
PRR3	5.65E-13	5.84E-09
NFKBIZ	1.02E-12	7.00E-09
SYNGR2	4.19E-12	2.16E-08
RCAN3	5.84E-12	2.41E-08
TUBB6	7.99E-12	2.43E-08
SLC25A33	8.24E-12	2.43E-08
KLF2	1.43E-11	3.68E-08
CRIP3	1.76E-11	4.03E-08
PHYHIPL	2.20E-11	4.47E-08

Correlation Analysis

In the R software, we use the cor () function to obtain the PCC matrix of two pairs of genes to take out the correlation coefficient values of disease and health state and subtract the difference to get the absolute value of the relative PPC, which ranges between 0 to 2. There are 124750 values and the size of the difference value means the strength of the correlation between the two genes. Additionally, we use hist() function to make the frequency distribution and use the lines() function to make the fitting curve (Figure 2). The horizontal axis is the absolute value of the relative PPC between 0 and 2, and the vertical axis is the number of occurrences of this value.

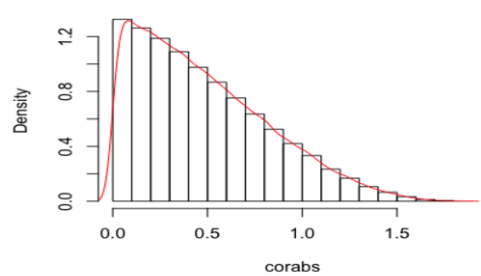


Figure 2- Frequency distribution of the correlation

Figure 2 shows that the greater the absolute value of the relative PPC, the smaller the data, that is, the weaker the correlation between the two genes. Therefore, with the frequency distribution map, the absolute value of the relative PPC is arranged in descending order and the maximum 1% of the data is retained. The first 1200 pair pairs are screened and their absolute values of the relative PPC are obtained. The first twelve gene pairs are choosen to be listed in Table 2.

Table 2- Absolute values of the relative PPC of gene pairs

Gene1 Name	Gene2 Name	Cor（abs）
DTX3L	PLEKHA2	1.841587988
SLC7A2	SMAD6	1.828728355
FGG	IRF9	1.826512894
RPF2	SMAD6	1.814375402
IRF9	RCAN2	1.81315706
RAB27B	MAFF	1.806385076
FGG	SMAD6	1.805861711
IRF9	PLEKHA2	1.791235586
TMEM99	PLEKHA2	1.790806898
RCAN2	DTX3L	1.790245603
VCAN	PSAT1	1.785278759
BICD2	RASSF5	1.783599316

Construction of Gene Coexpression Network and Clustering Analysis

To explore the co-expression changes induced by the highly relevant gene pairs, the 1200 pairs of genes selected above are mapped into the gene co-expression network with the help of the Cytoscape software (Figure 3). Each vertex in the graph represents a gene. When the co-expression of the two genes (ie, the absolute value of the relative PPC) is greater than the selected threshold, an edge is connected to indicate the significant correlation between them.

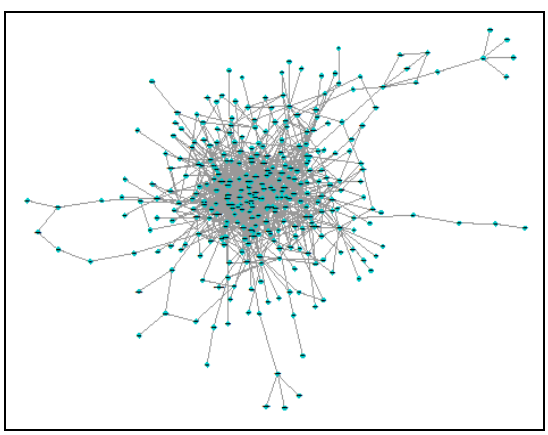


Figure 3 - Gene co-expression network

GO-Analysis

We select the four gene modules with the strongest correlations and analyze them with the NOA online website to get the functional analysis of the genes within the modules (Table 3). In combination with the analysis of Gene Card website, we identify HK2 and KLF4 as potential candidate molecular biomarkers.

Table 3- Result of GO-Analysis

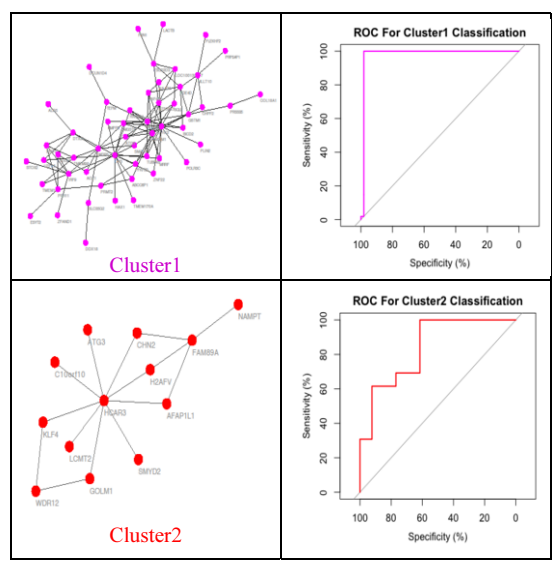
GO: term	p-value	Corrected p-value	Term name
GO:0000460	9.4E-4	0.1551	Maturation of 5.8S rRNA
GO:0000466	9.4E-4	0.1551	Maturation of 5.8S rRNA from tricistronic rRNA transcript
GO:0000676	9.4E-4	0.1551	Nicotinamide metabolic process
GO:0010999	5.4E-4	0.1022	Regulation of eIF2 alpha phosphorylation by heme
GO:0045993	5.4E-4	0.1022	Negative regulation of translational initiation by iron
GO:0046984	5.4E-4	0.1022	Regulation of hemoglobin biosynthetic process
GO:0046986	5.4E-4	0.1022	Negative regulation of hemoglobin biosynthetic process

SVM-based machine learning method

Through the systematic study of these functional modules and candidate biomarkers, we use the method of Leave-one-out to do the machine learning of the existing data in the R software and get the corresponding value of accuracy (AUC) (Table 4) and ROC curves (Figure 4).

Table 4- Number and AUC of selected gene modules

Module	Gene Number	Auc
Cluster1	54	0.98
Cluster2	13	0.84
Cluster3	12	0.97
Cluster4	11	0.98



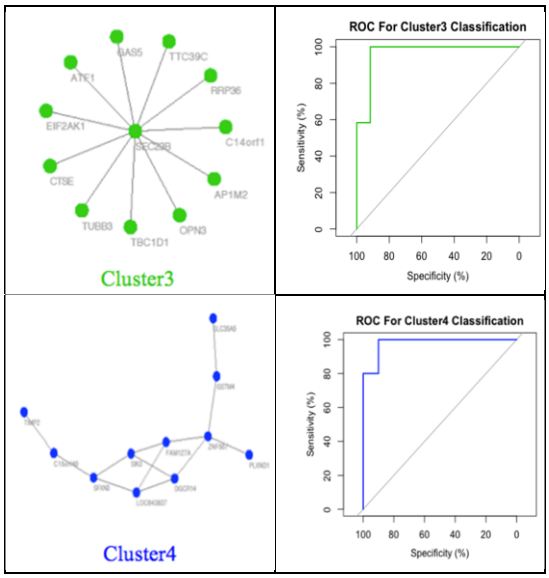


Figure 4- Gene modules and corresponding ROC curves

From Table 4 and Figure 4, we can see that the SVM classifier has a good AUC of 0.98, 0.84, 0.97 and 0.98, which are very close to 1 after training and learning for the four gene modules. And the corresponding ROC curves are very close to the upper left corner of the highest threshold of the specificity and sensitivity. It can be seen that the SVM classifier shows good performance in the classification of the four gene modules, indicating that the four gene modules can distinguish disease state and health state. And the gene HK2 and gene KLF4 contained show significant differences in the two states and most likely can be identified as biomarkers of HCC.

Discussion

Looking for biomarkers of HCC and exploring the molecular mechanism of HCC are of great significance for early prediction and early treatment of HCC. HK2 is a Protein Coding gene which is associated with the occurrence of some tumors. Among its related pathways are Regulation of Glucokinase by Glucokinase Regulatory Protein and Translation Insulin regulation of translation. KLF4 is a tumor suppressor gene which participates in cell differentiation, cell proliferation, necrosis and angiogenesis. The encoded protein is thought to control the G1-to-S transition of the cell cycle following DNA damage by mediating the tumor suppressor gene p53. These two genes are to some degree, related with the occurrence or metastasis and invasion of the tumor and cell differentiation, angiogenesis and so on. This also confirms the reliability and accuracy of our approach from another side.

Compared with the previous studies, the molecular mechanism based on HCC in this paper should be a hypothesis of a network of gene dynamic regulation, and more consideration is given to the potential interconnections between genes. Through the construction of co-expression network to further explore the molecular mechanism of liver cancer, the method of mining effect should be more reasonable. Although the correlation between the two genes and the occurrence of HCC needs to be confirmed by further experiments, it is believed that the applicability of this method will be further demonstrated with the further accumulation and perfection of high-throughput experimental data.

The occurrence and development of cancer are caused by the mutual disturbance of multiple functional pathways. Therefore, during the occurrence and development of HCC, tumor-associated genes should be a dynamic process of mutual regulation of the network. Consequently, we analyze the microarray data of HCC with the combination of traditional bioinformatics analysis method and the increasingly active system biology analysis and use bioinformatics methods such as normalization, statistical analysis, clustering analysis, GO-analysis, SVM and so on. The construction of gene co-expression network is completed and the network is searched for genes related to HCC. As a result, the gene HK2 and gene KLF4 are obtained and the functional modules of these two genes are proved to possess a great ability of classification by using SVM, which will contribute to the future detection and personalized medicine of HCC. This will be of great help and significance for the development of precision medicine to improve the lives and health of patients.

Acknowledgements

This work was supported by Chinese National High-tech R&D Program (2015AA020109), National Key Scientific Instrument and Equipment Development Project (2016YFF0103200), the Fundamental Research Funds for the Central Universities of China, National Natural Science Foundation of China (61572287) and Shandong Provincial Natural Science Foundation, China (ZR2015FQ001).

References

[1] D. Joshi, C. Taylor, J. Maggs, M. Heneghan, K. Childs, I. Carey, N. Heaton, J. O'Grady, K. Agarwal, and A. Suddle, Hepatocellular carcinoma (HCC) in HIV-positive patients: a more aggressive disease course?, *HIV MEDICINE* **12** (2011), 36-36.

[2] W. Zhao, P. Langfelder, T. Fuller, J. Dong, A. Li, and S. Hovarth, Weighted Gene Coexpression Network Analysis: State of the Art, *Journal of Biopharmaceutical Statistics* **20** (2010), 281-300.

[3] T. Tada, T. Kumada, H. Toyoda, S. Kiriyama, Y. Sone,M. Tanikawa, Y. Hisanaga, S. Kitabatake, T. Kuzuya, K. Nonogaki, J. Shimizu, A. Yamaguchi, M. Isogai, Y. Kaneoka, J. Washizu, and S. Satomura, Relationship between Lens culinaris agglutinin - reactive α - fetoprotein and pathologic features of hepatocellular carcinoma, *Liver International* **25** (2005), 848-853.

[4] S. Zhang, H. Zhao, and M. Ng, Functional Module Analysis for Gene Coexpression Networks with Network Integration, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **12** (2015), 1146- 1160.

[5] M. Rahman and Q. Zhang, Comparison among pearson correlation coefficient tests, *Far East Journal of Mathematical Sciences* **99** (2016), 237-255.

[6] Z.P. Liu, Identifying network-based biomarkers of complex diseases from high-throughput data, *BIOMARKERS IN MEDICINE* **10** (2016), 633-650.

[7] S. Van Dongen and C. Abreu-Goodger, Using MCL to extract clusters from networks, in, United States, pp. 281- 295.

[8] S. Zhang, J. Cao, Y.M. Kong, and R.H. Scheuermann, GO-Bayes: Gene ontology-based overrepresentation analysis using a Bayesian approach, *Bioinformatics* **26** (2010), 905-911.

[9] W. Mao, X. Mu, Y. Zheng, and G. Yan, Leave-one-out cross-validation-based model selection for multi-input multi-output support vector machine, *Neural Computing and Applications* **24** (2014), 441-451.

Address for correspondence

Corresponding author: Jing-song Li.

Address: College of Biomedical Engineering and Instrument Science, Zhejiang University, 38 Zheda road, Hangzhou 310027, China.
Email: ljs@zju.edu.cn. Tel.: +86-571-87951564

Corresponding author: Zhiping Liu.
Address: College of Control Science and Engineering, Shandong University, 17923 Jingshi road, Jinan 250061, China.
Email: zpliu@sdu.edu.cn. Tel.: +86-137-9113-2032