Based on Bipartite Graph Label Gene Extraction Algorithm of Network Structure

Jiming Li
Business school, University of Shanghai for Science and Technical
516 Jungong Road, Shanghai, China
Email: usstsnow@126.com

Ning Zhang
Business school, University of Shanghai for Science and Technical
516 Jungong Road, Shanghai, China
Email: zhangning@usst.edu.cn

Zhaoxing Liu
Business school, University of Shanghai for Science and Technical
516 Jungong Road, Shanghai, China
Email: liouzhaoxing@126.com

Gengsheng Zhao
Business school, University of Shanghai for Science and Technical
516 Jungong Road, Shanghai, China
Email: usstemlab@usst.edu.cn

Received: April 11, 2011 Accepted: May 10, 2011 doi:10.5539/ijb.v3n4p64

The research is financed by Shanghai Leading Academic Discipline Project (S30501), National Natural Science Foundation of China (70971089), Selection and training fund of outstanding young teachers in colleges and universities of Shanghai

Abstract

Firstly we make a pretreatment to the original gene data, and analyze the information of sample gene graph by two steps. First step is removing the unrelated genes; Second step is use an extraction algorithm of label gene based on bipartite network structure to handle the candidate gene, and get gene interactive relationship network. Finally extracts some label gene form the gene interactive relationship network.

Keywords: Gene selection, SAM algorithm, Bipartite network structure, Gene interactive relationship network

1. Introduction

Occurrence and development of cancer is a complex process of regulation of gene expression, and with the development of large-scale gene expression profiling technology. Thousands of gene expression levels can be obtained simultaneously in the experiment from tissue samples by using DNA Chips, but the way to find a set of genes which decide the sample gene's character from thousands of genes which measured by DNA chip, that is, " Gene information " or " Label gene "(informative genes) , which is the key factor for establish an effective classification model, properly identify tumor type, give to reliable diagnostic and simplify experimental analysis, and has important significance in the study of cancer pathogenesis, diagnosis and treatment.

Given the importance of tumor classification information selection, now a substantial research literature for this problem has already appeared. In 1999 Golub (T. R. Golub, *et al.* (1999) used "signal to noise ratio"as an indicator feature to extraction, and studied the classification of the two subtypes of leukemia with voting method. In 2000 Guyon(2000) put forward a new features extraction method—use support vector machine as classification tools which based on Golub's study, only with eight genes as classification feature had achieved 100% classification effect. In 2002, Sigh(2002) and others use the same feature extraction indicators as Golub, using k-nearest neighbor method as classification of prostate cancer gene research. Since then, Ruan Xiaogang and others are using pattern recognition methods, calculate technologies, floating order search algorithm and support vector machines for more precise identification of the 5 label genes to identify subtypes of acute leukemia. For many types of tumor subtypes, Khan (2001) using neural network method on the small round blue cell tumor (SRBCT)'s four subtypes for the diagnosis, extracted 96 characteristic genes, and got a good result. Tibshirani (2002) extracted 43 characteristic genes with the recent contractions centroid algorithm, be able to identify 20 subtypes of blind samples. Yeo and Poggio (2001) decomposition the four categories of questions into multiple two-class problem with k-nearest neighbor algorithm (KNN) and weighted voting method and support vector machine. In addition, other researchers also put their respective statistical calculation method (Fu K, Iqbal J, Chan W C. 2005. Van't Veer L J, Dai H, van de Vijver M J, *et al.* 2002. Zhang H, Yu C Y, Singer B. 2003).

Genetic correlation measure is used for assess the relevance among genes, Measure quality in some extent determine the success or failure of genes select. Gene correlation measure and machine learning and data mining in statistics area are common measure methods, various statistics parameter estimated and non-parameter estimated of measure, as t-test (Baldi P, Long AD. (2001) parameter estimated, TNOM (Ben. DorA, Bruhn L, Friedman N, *et al.* (2000), and MDMR 9 Park P J , Pagano M.(2001), and WEPO (Chuang H Y, Tsai H K, Tsai Y F , *et al.* (2003), and SAM (Tusher VG, Tibshirani R , Chu G.(2001) etc. non-parameter estimated is widely used in gene micro-array column expression data. Jeffery (2006) using10 different statistics measure methods on 9 micro-array column data set, calculation the front 50, front 100, and front 200 genes of intersection respectively, eventually results display10 different measure method by produced of gene arranged just has 8%-21% same genes.

In the last ten years, complex network as an emerging cross-discipline, with strong correlations to physical, biological, and social problems have aroused more and more people's attention. Understanding the topology structure of complex network, dynamics and function is the current main objective of the study of complex networks. Complex networks can be divided into social networks (such as a film actor collaboration network), the information network (such as the World Wide Web), technology networks (such as power networks), and biological networks (such as the protein network) (Li Ze, Bao Lei, Huang Y.W.(2002) etc. according to research area.

In complex networks, based on the nature of different networks can be divided into single mode networks, dual mode network and multi-mode networks. In single mode networks, only a kind of nodes, and in the dual-mode network, contains two types of distinct nodes, at the same time, similar node does not exist edge between nodes, edges can only exist between different kinds of nodes, this network also known as the bipartite network. Multi-mode includes more different kinds of node set. For bipartite network widely exist in real life, such as (film actor network), recently got a lot of attention, such as Zhou Tao (2007)consider the bipartite network topology, design a personalized recommendation algorithm based on network structure. Is aware of the importance in bipartite graph structure in the network, in this article also attempts to using bipartite network search label genes.

2. Classification independent gene filtering

In gene expression data, differences in gene expression levels of data showing a different. The distribution of some gene of expression level in ALL and AML categories, its value also variance are no obvious differences, these gene will increased the calculation complexity of search information genes, and does not provides useful information on sample type distinguish, to these gene and sample category have nothing relation, so is necessary remove these independent genes.

T-test as a representative of parametric statistics measure, if assumes that information gene expression data exists two types of samples; the t value $t(g_i)$ of information gene g_i can be given by the following formula:

$$t(g_i) = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 - \sigma_2^2/n_2}} \qquad (1)$$

In the formula: μ_r And σ_r are respectively the mean and standard deviations of the gene g_i to class *r*. n_r is the number of samples Class r, $r = 1, 2$ .The higher value of g_i , indicated that gene information g_i has stronger similarity. t-test can only be used for two types of situations samples. There are some similar measures like t-test such as: Fisher, F-test, S2N (signal-to-noist ratio) and other similar measures.

Microarray gene expression data contain large volumes of data, usually has tens of thousands of genes and of very limited sample size, such as the t-test methods are too simple and parameter estimation of measurement is instability in selection process, so the mean and standard deviation statistics is likely to cause loss of statistical information, which is not conducive to label gene extraction. Considering the problems of t -test, Tusher(2001) put forward SAM algorithm to defined statistics:

$$d = \left| \frac{\mu_1 - \mu_2}{\sigma + c} \right| \tag{2}$$

In the denominator adding a small constant c to avoid the FC small problem, because the statistics *d* is not subject to t Distribution, Tusher using the displacement test to directly control FDR. Efron discussed the selection of constant c, and confirmed in most of the test, select all the standard gene expression value standard deviation of the 90% quantile can achieve good results. In our tests, for simplicity, set c=1.

Calculate the statistics for each gene d, and gives the results of the histogram. We can find from figure1, the weights of most of these gene property classification less than 0.4. In article we choose 127 genes which weights greater than 0.4, other genes were deemed useless genes and removed from the original sample.

3. Tag extraction and result analysis

Based on different nature of nodes, network can be divided into two different sets of nodes - gene sets and sample set, respectively, $G = \{g_1, g_2, \ldots, g_n\}$ And $S = \{s_1, s_2, \ldots, s_m\}$, Where the *n* represents the number of genes, *m* representative number of samples. $E_{m\times n}$ is $m \times n$ adjacency matrix, e_{ij} representative the expression level of gene *i* in sample *j*. Figure2 is a sample bipartite network.

In the first step, disease-related genes have been filtered, assume that all the genes are associated with disease, and interactions of these genes will only strengthen without offset. Disease always the result of the interaction of multiple genes, in order to better confirm the label gene, you first need to determine the strength of the interaction between genes, search for the existence of a particular gene has to do with many genes, in that gene interaction network, is there some core node. This core node is essential for robustness of gene interaction network (Xiaofeng Liu, Chen guohua. (2007).

Through the bipartite network projection you can get a gene interaction networks. In this article, we only need to study interactions between genes, the projection only require projection from a sample set to gene set. If two genes have common neighboring nodes, then these two genes be joined, otherwise the two genes are not joined. See Figure3.

Disease Occurrence or not are closely related with the level of gene expression. For genes *i* and *j* in the sample α on the expression levels respectively $e_{i\alpha}$ and $e_{j\alpha}$, Then get interaction strength $\min(e_{i\alpha}, e_{j\alpha})$ between gene *i* and gene *j* from sample α . Therefore, in gene interaction networks, gene interaction strength between:

$$I_{ij} = \frac{1}{m}\sum_{\alpha=1}^{m} \min(e_{i\alpha}, e_{j\alpha}) \qquad i \neq j \tag{3}$$

$$I_{ij} = 0 \qquad i = j \tag{4}$$

$\{I_{ij}\}$ Constitute an $n \times n$ interaction strength matrix I , and the matrix is a symmetric matrix. From upper triangular of gene interaction matrix, there are some genes have very large strength, which means very high level of gene expression, these genes is exactly what we need to search.

Using SAM algorithms remove unrelated genes, get 127 relate genes. Get Gene interaction strength matrix I_{sam} from (3) and (4), and then extract the upper triangular matrices, in row order transform into a one-dimensional gene sequences as shown below:

Can be seen some of the gene pairs have very strong interaction strength from Figure4, and then the probability of occurrence of diseases is very high. We are given a threshold to determine these notable gene pairs. When we take threshold value for 0.6, get gene pairs on sequence: (2 3), (2 4), (2 5), (2 11), (2 14), (2 20), (2 21), (2 23), (2 43), (2 70), (2 76), (4 11), (5 11), (5 20), (11 20), (11 21), (11 23), (11 70), (11 76), (14 20), (16 76), (20 21), (20 23), (20 70), (20 76), (21 76), (23 76), (43 70), (43 76), (62 70), (62 76), (70 76), (76 94), (76 105), total of 27 pairs (the order in which the label is not in the source data, only serial number). According to the relationship between 27 gene pairs, by the strength of the interaction between them to build the network as in Figure5 shows:

From Figure5 we can clearly find that the network constructed by filtered genes exist some very large degree nodes, these nodes of a greater degree may have maximum influence in induce disease based on the assumption, according to the topology of the network, by the naked eye can be find label gene. In Figure5, H05899 and L11706 are the label genes what we need. So after removing these nodes in the network, network robustness obvious variation, sees Figure6, appeared many isolated nodes in the network, there is reason to believe that gene H05899 interactions with other genes induced the disease.

Did bipartite network projection with the filtered 127 Candidate genes, received 16 characteristic genes (table 1).

Existing in the research literature on colon cancer (Gennadi V. Glinsky, Yelena A. Ivanova, Anna B. Glinskii. (2003). Li Jiangeng, Gao Zhikun, Ruan Xiaogang, Yan Chi. (2009). Ian W Taylor, Rune Linding, *et al*. (2009). Han-Yu Chuang, Eunjung Lee, *et al*. (2007), many scholars have found some labels gene. By comparing with literature, based on the network structure labels gene extraction method of bipartite network (SIGABN) resulting in colon label gene, many of them are coinciding with the tabs for the other documents found in the genes, this shows that the method is effective, as for the other genes, pending further validation of medical workers.

4. Analysis and discussion

In this article classification information gene selection is divided into two steps, independent gene filtering and removal of redundant genes (label gene extraction). Redundancy removal will not increase characteristics of classification information contained. Therefore, in conducting the independent gene filtering should conduct a comprehensive analysis of gene samples contained classified information, so as not to filter out genes that contain important information. Through bipartite network projection of the filtered candidate label genes, get a related network. Form gene related network, impact of gene on the network can recognize the importance of genes and to identify a label gene. Although the algorithm in the article is simple, test results in better, whether other using pending further validation of medical workers.

Reference

Baldi P, Long AD. (2001). A Bayesian framework for the analysis of microarray expression data. Regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, 17(16) : 509-519. http://dx.doi.org/10.1093/bioinformatics/17.6.509

Ben. DorA, Bruhn L, Friedman N, *et al*. (2000). Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7 (324) : 559-583. http://dx.doi.org/10.1089/106652700750050943

Chuang H Y, Tsai H K, Tsai Y F , *et al*. (2003). Ranking genes for discriminability on microarray data. *Journal of Information Science and Engineering*. 19 (6)953-966.

E.J.Newman. (2001). *Phys. Rev. E*64 016131,016132.

Fu K, Iqbal J, Chan W C. (2005). Recent advances in the molecular diagnosis of diffuse large B - cell lymphoma, *Expert Rev Mol Diagn*, 5(3): 397-408. http://dx.doi.org/10.1586/14737159.5.3.397

Gennadi V. Glinsky, Yelena A. Ivanova, Anna B. Glinskii. (2003). Common malignancy-associated regions of transcriptional activation (MARTA) in human prostate, breast, ovarian, and colon cancers are targets for DNA amplification. *Cancer Letters,* Volume 201, Issue 1: 67-77. http://dx.doi.org/10.1016/S0304-3835(03)00419-1

Guyon I, Weston J, Barnhill S, *et al*. (2000). Gene selection for cancer classification using support vector machines, *Machine Learning* , 46(13):389-422.

Han-Yu Chuang, Eunjung Lee, *et al*. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, Volume 3. http://dx.doi.org/10.1038/msb4100180

Ian W Taylor, Rune Linding, *et al*. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology,* Volume 27, No 2: 199-204. http://dx.doi.org/10.1038/nbt.1522

Jeffery Ian B, Desmond G Higgins, Aedin C Culhane. (2006). Comparison and evaluation of methods for generating differentially expressed gene list s from microarray data. *BMC Bioinformatics*, 7:359. http://dx.doi.org/10.1186/1471-2105-7-359

Khan J, Wei J S, Ringner M, *et al.* (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks , *Nature Medicine*, 7(6):673-679. http://dx.doi.org/10.1038/89044

Li Jiangeng, Gao Zhikun, Ruan Xiaogang, Yan Chi. (2009). Colon cancer marker gene selection based on double-gene analysis. *Chinese Journal of biomedical engineering*, Vol.28, No.5.

Li Ze, Bao Lei, Huang Y.W. (2002). Based on the characteristics of tumor classification based on gene expression and gene selection. *Interface science.* 18(4):413-417.

Park P J , Pagano M. (2001) , Bonetti M. A nonparametric scoring algorithm for identifying informative genes from microarray data[C] *PPProc of Pacific Symp on Biocomputing*. Singapore: World Scientific Pablishing Company, 52 – 63.

Singh D, Febbo P G, Ross K, *et al.* (2002). Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, 1: 203-209.

T. R. Golub, *et al.* (1999). Monitoring and Class Prediction by Gene Expression, *Science,* Vol.286, pp.531-537. http://dx.doi.org/10.1126/science.286.5439.531

Tibshirani R, Hastie T, Narasimhan B, *et al.* (2002). Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression, *PNAS,* 99:6567-6572. http://dx.doi.org/10.1073/pnas.082099299

Tusher VG, Tibshirani R , Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response [C] *PPProc of the National Academy of Sciences*. Washington :National Academy of Science , 5116 – 5121.

Van't Veer L J, Dai H, van de Vijver M J, *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, 415 (6871): 484-485.

Xiaofeng Liu , Chen guohua. (2007). Based on supply chain robustness analysis of complex networks. *Journal of Southeast University (Natural science Edition),* (S2).

Yeo G, Poggio T. (2001). Multiclass Classification of SRBCT, *CBCL Meno,* 18:83-84.

Zhang H, Yu C Y, Singer B. (2003). Cell and tumor classification using gene expression data: construction of forests, *Proc Natl Acad Sci USA*, 100 (7): 4168-4172.

Zhou T, Ren J, Medo M and Zhang Y C. (2007). Bipartite network projection and personal recommendation. *Physical Review E,* Vol. 76 (4): 46115. http://dx.doi.org/10.1103/PhysRevE.76.046115

Table 1. Select a tag in the article gene and literature which can verify

	Label gene
Filters out gene	H05899 H23544 X12466 H40095 R84411 J05032 U14631 D59253 U09587 H55916 R33367 T83368 T51571 L41559 L11706 T57468
Document verifiable gene	H05899 H23544 X12466 H40095 R84411 J05032 U14631

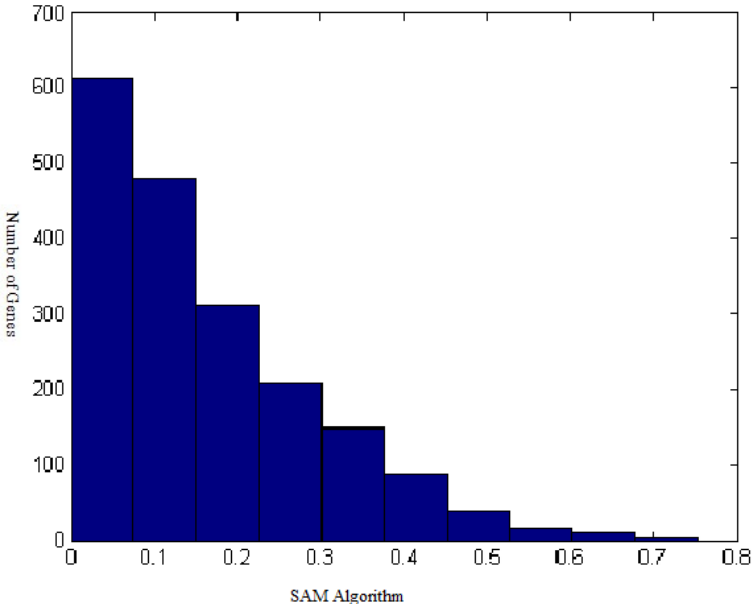

Figure 1. Gene SAM Algorithm of histogram

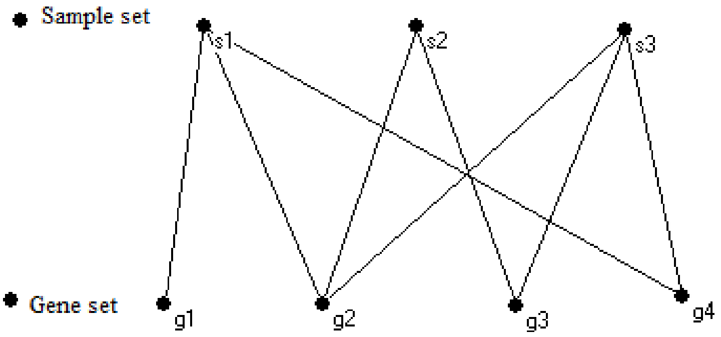

Figure 2. sample bipartite network of sample set–gene set

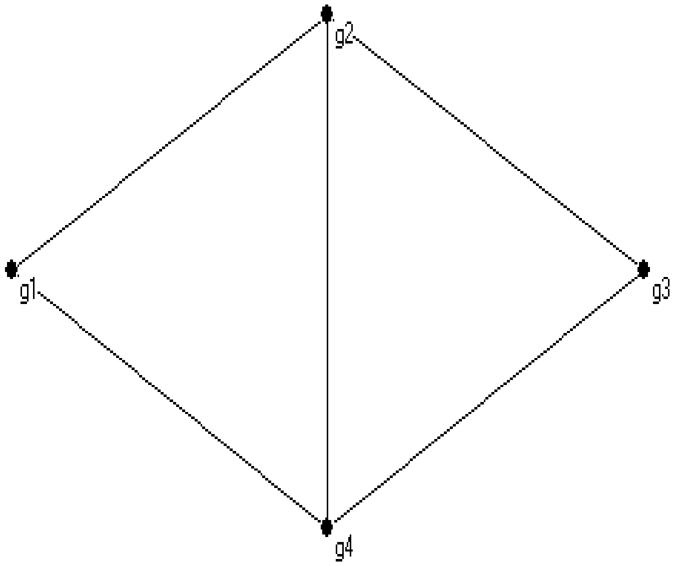

Figure 3. gene interaction network

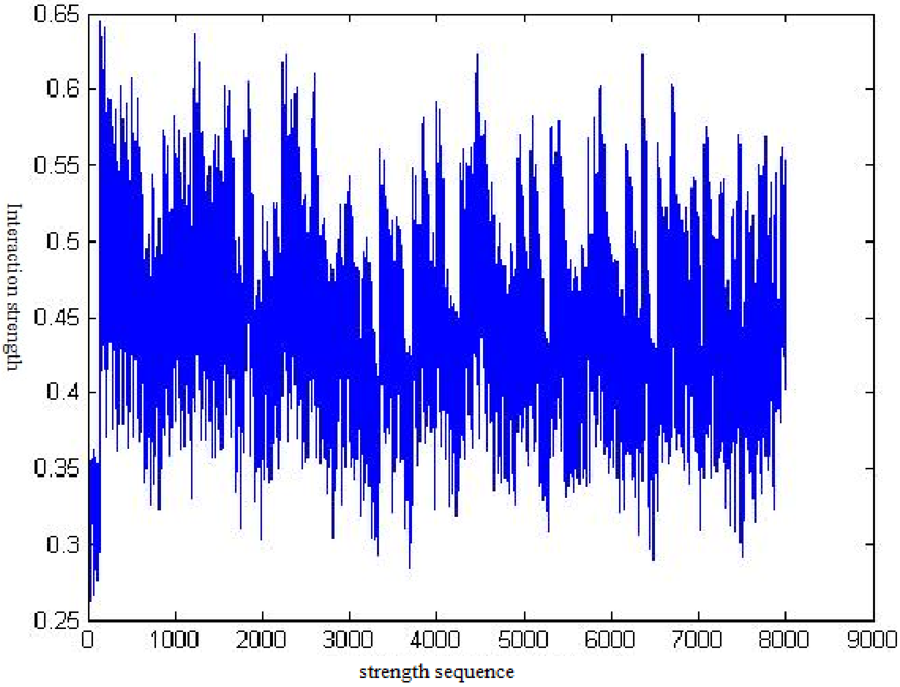

Figure 4. Based on the SAM algorithm for gene strength sequence diagram

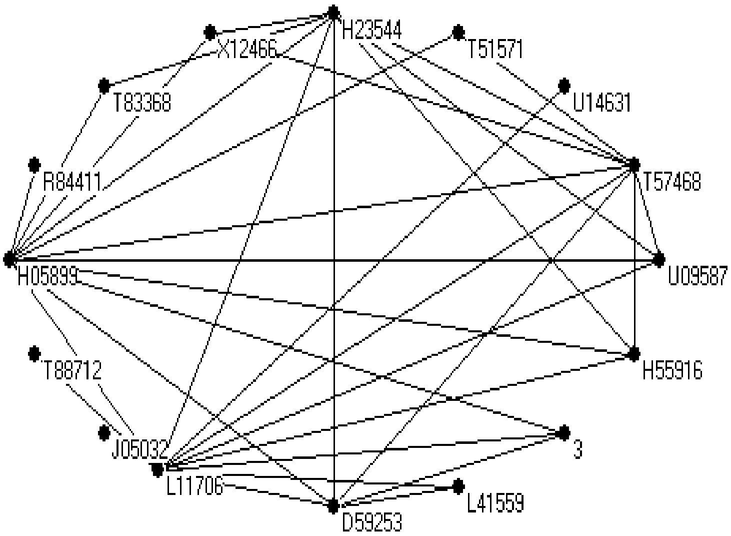

Figure 5. Gene interaction network

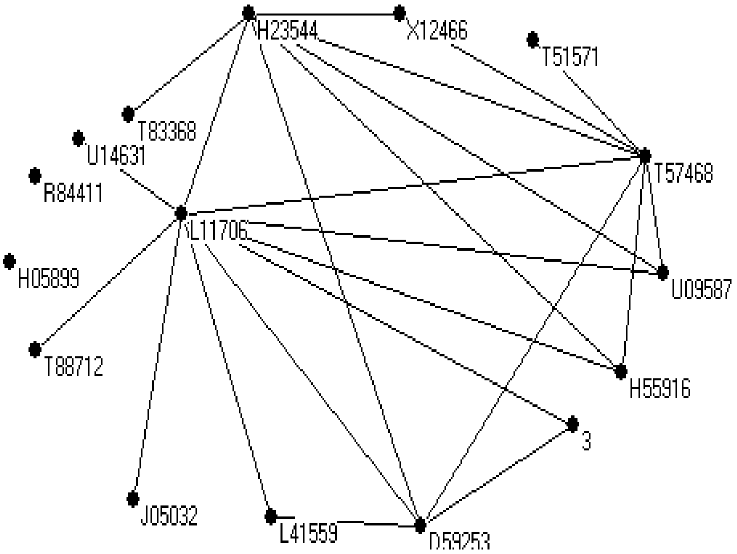

Figure 6. Gene interaction network after delete nodes H05899