Available online at www.sciencedirect.com
ScienceDirect
Physica A 370 (2006) 663–671
PHYSICA A
www.elsevier.com/locate/physa

Statistical analysis of gene regulatory networks reconstructed from gene expression data of lung cancer

Lanfang Sun[a,*], Lu Jiang[a], Menghui Li[a], Dacheng He[b]

[a]*Department of System Science, School of Management, Beijing Normal University, Beijing 100875, PR China*
[b]*College of Life Sciences, Beijing Normal University, Beijing 100875, PR China*

Received 27 July 2005; received in revised form 17 January 2006
Available online 22 March 2006

Abstract

Recently, inferring gene regulatory network from large-scale gene expression data has been considered as an important effort to understand the life system in whole. In this paper, for the purpose of getting further information about lung cancer, a gene regulatory network of lung cancer is reconstructed from gene expression data. In this network, vertices represent genes and edges between any two vertices represent their co-regulatory relationships. It is found that this network has some characteristics which are shared by most cellular networks of health lives, such as power-law, small-world behaviors. On the other hand, it also presents some features which are obviously different from other networks, such as assortative mixing. In the last section of this paper, the significance of these findings in the context of biological processes of lung cancer is discussed.

Keywords: Gene expression data; Gene regulatory network; Complex network; Lung cancer

1. Introduction

In these years, lung cancer has become the leading cause of cancer death worldwide [1]. Although biologists have found some sensitive biomarkers for lung cancer by conventional biological methods [2], the goal of early detection and diagnosis is still difficult to achieve, let alone cure it. This is largely due to the unclear mechanisms that underlie lung carcinogenesis.

At molecular level, the genesis of lung cancer is determined by the status of genes in vivo. These genes do not work independently, instead they interact with one another in the form of a complex network, which is called gene regulatory network, and function coordinately as an organic whole [3,4]. Hence, the status of a gene is determined by other genes that have interactions with it. So realizing the characteristics of the gene regulatory network of lung cancer is essential to understand the genesis of lung cancer. Apparently, conventional biology which deals with single molecule is not competent for unraveling such a complicate gene

*Corresponding author. Tel.: +86 10 58807876.
E-mail address: bnuslf@sohu.com (L. Sun).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.02.034

regulatory network through innumerable experiments. Therefore, development of new tools and methods is necessary to solve this problem.

Recently, novel gene chip technology [5] has provided us an effective and high-throughput tool to measure gene expression level on a large scale, while complex network [6,7] theory has provided us a new method to study a complex system at the whole level. If they are combined, the problem mentioned above can be settled in a way. So far, there have been many notable works in this direction [8–12].

Actually, information mined from the gene expression data by this kind of treatment is more profound than by conventional cluster analysis [13], which prevailed in the past few years. For example, it can tell us the possible regulatory interaction and mechanism, genes and relationships that are relatively more important among all genes and interactions, the feature of the linkage style of the network, definite clusters, etc.

In this paper, we use gene expression data obtained from normal and cancerous lung cells to reconstruct the gene regulatory network of lung cancer, and the reconstruction algorithm is based on Refs. [11,12]. The main improvement is that our algorithm can put co-express and counter-express gene pairs into the network simultaneously.

The finally obtained network of lung cancer displays scale-free, small-world behaviors which are similar to other empirically studied cellular networks of health lives, and assortative mixing degree correlation which is opposite to those normal cellular networks. What is more, large clusters separated from the network all have definite biological functionalities. In addition, the relationship between gene's ability to be candidates of biomarkers and its importance in the network are studied, and the results illustrates that there are no obvious relationships between them.

2. Materials and methods

The gene expression data used here is an expression profile of 12,600 genes for 203 samples [14], among which 17 are normal lung specimens and the other 186 are lung tumors. As we are aiming to unravel the gene regulatory network of lung cancer, we need to preprocess the original data for the purpose of selecting out the most informative genes whose expression levels are sensitive to the variation of clinical attributes of lung cell. Hence, we set up the same standard as that was mentioned in Ref. [14], i.e., a standard deviation threshold of 50 expression units, and filter out the 3312 most variable genes. Thereby, the final data used for the network construction is a 3312×203 matrix S, and its element S_{ij} denotes the expression level of gene i in sample j.

The network construction algorithm is as follows: each vertex represents a gene. For any two given genes i and j, we can calculate their Pearson correlation coefficient

$$r_{ij} = \frac{\sum_{k=1}^{203}(S_{ik} - \overline{S_i})(S_{jk} - \overline{S_j})}{\sqrt{\sum_{k=1}^{203}(S_{ik} - \overline{S_i})^2 \sum_{k=1}^{203}(S_{jk} - \overline{S_j})^2}}, \quad i,j = 1,2,\ldots,3312,$$

where $\overline{S_i}$ is the mean value of S_{ik} taken over all $k = 1,2,\ldots,203$. If $|r_{ij}|$ is larger than the given threshold W_0, then connect these two vertices by an edge. Hence, the topology of the network depends strongly on the parameter W_0.

In order to select a reasonable threshold W_0, we systematically investigated the formation of the largest cluster of the network by increasing W_0. The size of the largest cluster $N_{\max}$ is plotted against the threshold W_0 in Fig. 1a. It illustrates that the network structure will not vary dramatically when W_0 takes a value higher than 0.72. Combined with consideration of a proper size of the network, we selected $W_0 = 0.75$ as a reasonable and typical threshold. The corresponding network has 3312 vertices and 6724 edges, among which 2050 vertices are isolated. Consequently, we select the 1262 non-isolated vertices and their 6724 linkages as the components of the final network, and its largest cluster contains 412 vertices.

We can see that this kind of algorithm for network construction is better than algorithm mentioned in Refs. [11,12] in the facet that it can take co-express and counter-express gene pairs into consideration simultaneously by using $|r_{ij}|$ as the connecting criterion.

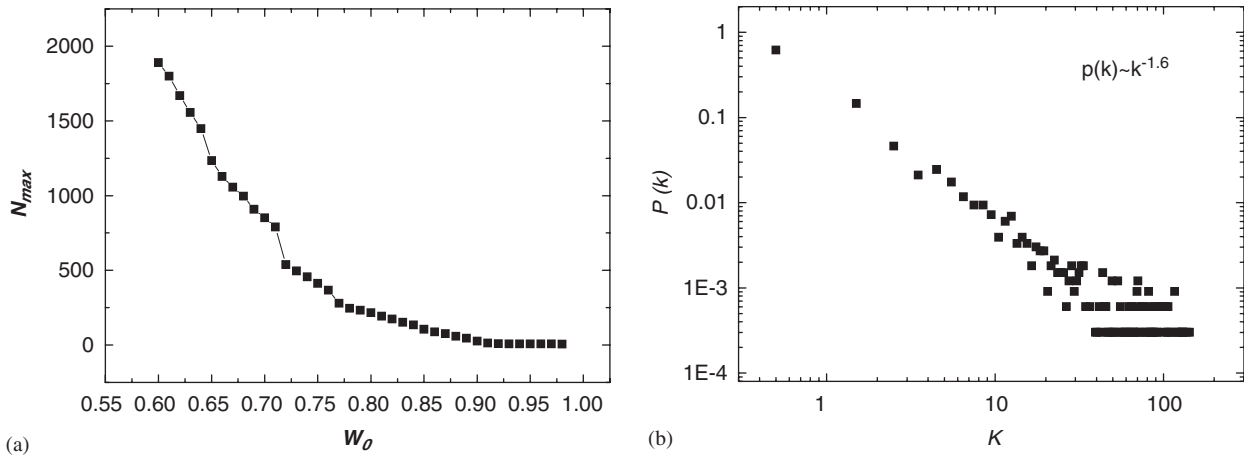

Fig. 1. (a) Size of the largest cluster $N_{\max}$ as a function of the threshold W_0. (b) Degree distribution of the network under $W_0 = 0.75$.

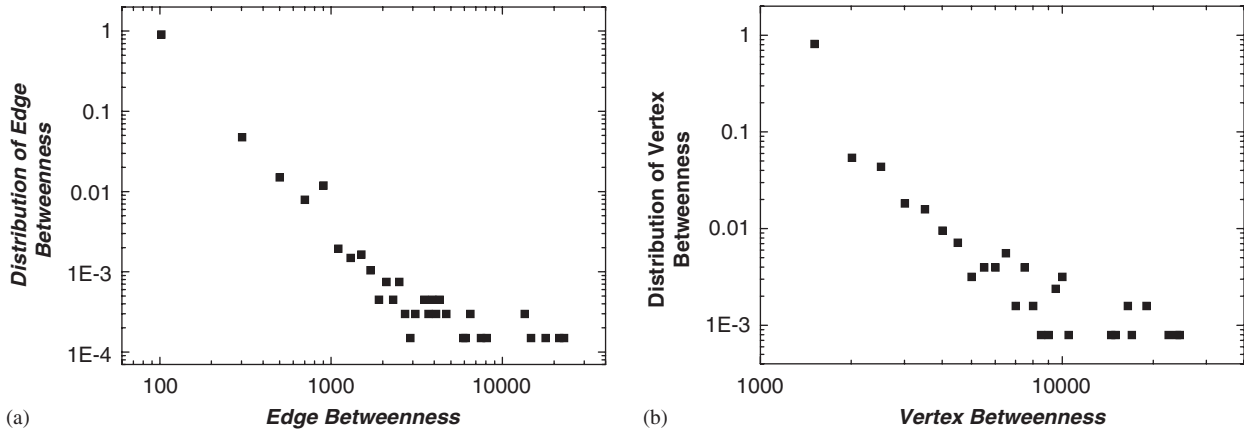

Fig. 2. The distribution of edge (a) and vertex (b) betweenness.

3. Statistical results of the network

3.1. Basic statistical features of the network

3.1.1. Degree distribution

Fig. 1b shows the degree distribution of the network. It is obvious that it follows power-law with an exponent of 1.6. Therefore, the network is *scale-free*, and this means that there exist a few genes with significantly high degrees play important roles in the network.

3.1.2. The distribution of vertex and edge betweenness

We analyze another statistical quantity of the network–betweenness [15]. It describes the importance of the vertex or edge in connecting different groups of the whole network. Surprisingly, it is found that the distribution of both vertex and edge betweenness also follows power law as illustrated in Fig. 2. That is to say, there exist a few genes and co-regulatory relationships playing crucial roles in the communication between different groups in the network. Therefore, deletion of these genes or relationships may not destroy the function of each group, but will surely ruin the whole network's functionality.

3.1.3. Average shortest path and average cluster coefficient

The average shortest distance $\langle l \rangle$ of the largest cluster is 3.89, and the corresponding average cluster coefficient $\langle C \rangle$ is 0.61. As distance is relative to the number of edges between two vertices, this low $\langle l \rangle$ tells us that there are only 4 steps in average from one vertex to another. The large $\langle C \rangle$ suggests that the network is quite compact. Here, low $\langle l \rangle$ and large $\langle C \rangle$ together indicate the small-world behavior of the network. This implies that once a gene is interrupted, the disturbance will quickly spread through the whole network.

3.1.4. Scaling of the cluster coefficient $C(k)$

Fig. 3a displays the average value of the clustering coefficient C (the average is taken over vertices with degree k) as a function of k, and we can see that it distributes approximately linearly rather than $C(k) \sim k^{-1}$ in a hierarchical network. That is to say, the network may not have the hierarchical structure.

3.1.5. Degree correlations

We study the correlations between connectivity of interacting vertices [16] by investigating the relationship between the average degrees k_i^{nn} of vertices in the neighborhood of a vertex i and the degree k_i, and find that they are positively correlative as illustrated in Fig. 3b. Such assortative mixing implies that high-degree vertices are mostly connected to other high-degree vertices and low-degree vertices are connected to other low-degree vertices. This is similar to the phenomena found in social networks such as scientific collaboration network [17,18], but is opposite to the most cellular networks such as the protein interaction network [19]. We guess that this kind of discordance may be one of the important origins of cancer. We can imagine that, in normal state, active genes (namely high degree genes) tend to interact with inactive ones, so the whole in vivo system can maintain in a well-balanced state; whereas in abnormal state, active genes tend to congregate so that the balance of the system will easily be broken up and the system will be led into a disordered status.

3.1.6. Predominant linkage style

We also analyze the feature of the linkage between vertices, and find that there are a significant number of edges connect vertices that were all with degree 2 as displayed in Table 1. That is to say, vertices with 2 edges prefer to link with each other, and this linkage style is predominant in the whole network. We suppose that this kind of structural pattern may have its superiority in fast signal transduction.

3.2. Modules of the network

Apart from those statistical features mentioned above, we also investigated modules in the network by conducting Newman's edge betweenness cluster algorithm [20], and finally gained 12 clusters. After searching

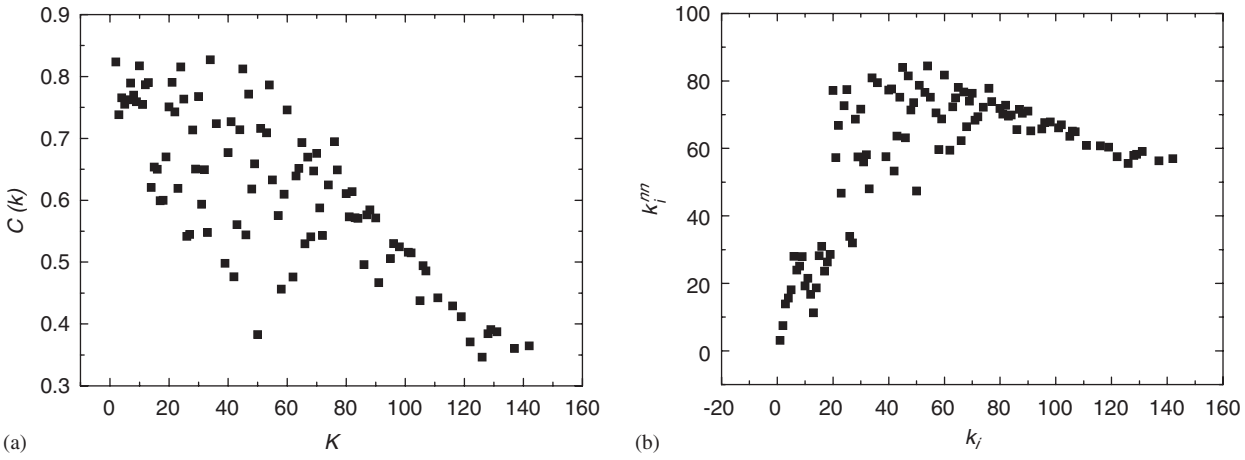

Fig. 3. (a) Scaling of cluster coefficient $C(k)$. (b) Average degrees k_i^{nn} of vertices in the neighborhood of a vertex i as a function of degree k_i.

Table 1
Top 6 linkage styles in the network

Linkage style	Recurrent times
(2,2)	82
(12,13)	44
(4,4)	38
(5,5)	34
(5,6)	32
(12,12)	32

Elements in the linkage style column (i, j) mean the mode that vertices with degree i are connected with vertices with degree j in a network. Recurrent times refer to times that the corresponding linkage style appears in the network.

Table 2
Result of the cluster analysis

Community no.	Biological function
1	Relative to the drug resistance of the cancer cells
2	Relative to inflammation reaction
3	Relative to cell apotosis and invade
4	Relative to cell proliferation, differentiation, and individual development
5	Relative to the proliferation and metastasis of cancer cells
6	Relative to cell proliferation
7	Relative to the squama of epithelia
8	Relative to the nerve system and metabolism
9	Relative to the metastasis of tumor
10	Relative to immune process
11	Relative to the interaction between cells
12	Relative to the cell cycle

for each gene's biological function in the OMIM dataset (see http://www.ncbi.nlm.nih.gov), it is found that these clusters all have definite biological functions as listed in Table 2. Fig. 4 illustrates the detail of two of the clusters, and we can easily find that they are relative to the immune reaction and tumor metastasis, respectively (see Supplementary Material for complete data sets).

3.3. Biomarker genes and their importance in the network

In Sections 3.1 and 3.2, we have investigated the network at macro and medium level, so we will turn to the micro level research in this section. Since we have known that there are some functionally important genes either with high degrees or betweenness in the network from Section 3.1, we will naturally wonder if such important genes can be candidates for biomarkers [21] (A kind of molecules including genes, proteins, metabolites, whose intensities change sensitively in response of the variation of clinical attributes, and can be used to distinguish between patients and healthy persons) of lung cancer. Here, we use Ref. [22] as a reference, and take Shannon entropy [23] as a measurement of the ability of genes to be biomarkers. For instance, to get the Shannon entropy of gene i, we first transform the nominal variable which describes the specimen's normal and abnormal state into a boolean variable which takes value of 0 and 1, respectively, and then sort specimens by increasing gene i's expression values. After such operation, we can get several "families" which contain nothing but normal specimens or nothing but abnormal specimens, and simultaneously we can obtain a series of expression values $\{S_{ij}^*\}$ which are boundaries of these families. For each S_{ij}^*, we can divide all specimens into two parts, one of which (denoted as specimen set 1) is composed of those whose expression level of gene i is higher than S_{ij}^* and the other (denoted as specimen set 2) is lower thanS_{ij}^*. Then we can give a quantity taking

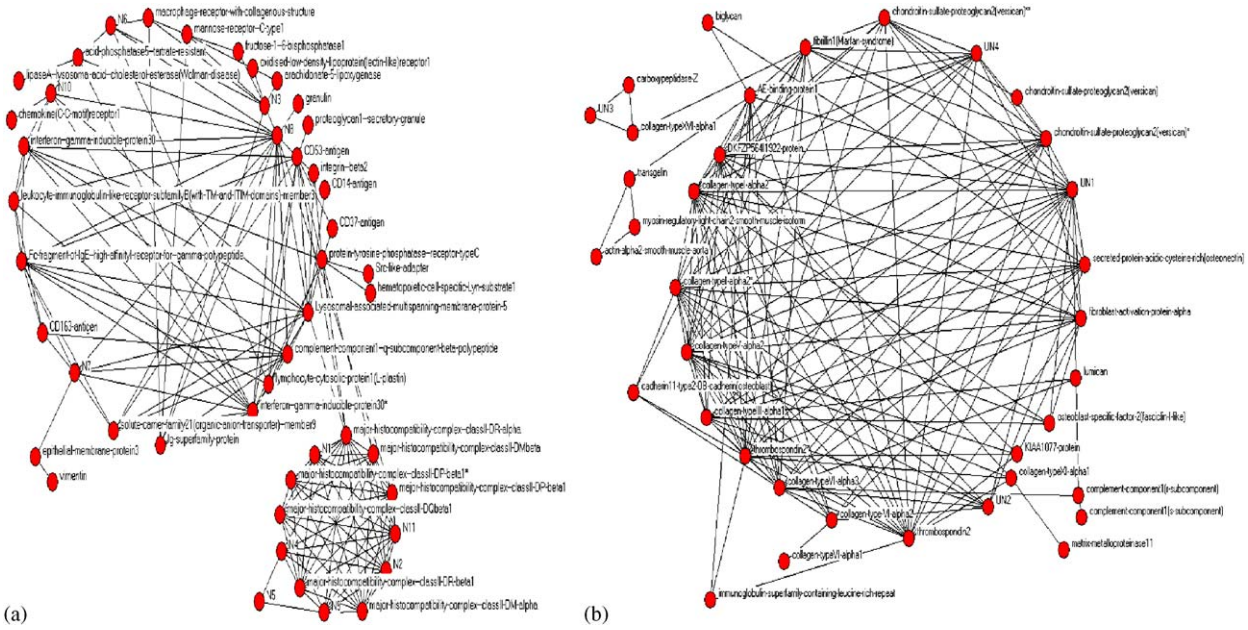

Fig. 4. (a) Detail of the cluster with 47 genes. The circle in red represents each gene, and the edge represents the link between these genes. Obviously, this large cluster can further be divided into two smaller groups. One is mainly composed of the family of major histocompatibility complex, and the other is composed of immune molecules. (b) The detail of the cluster with 36 genes. Obviously, this cluster is mainly composed of the collagen family, and is relative to tumor metastasis.

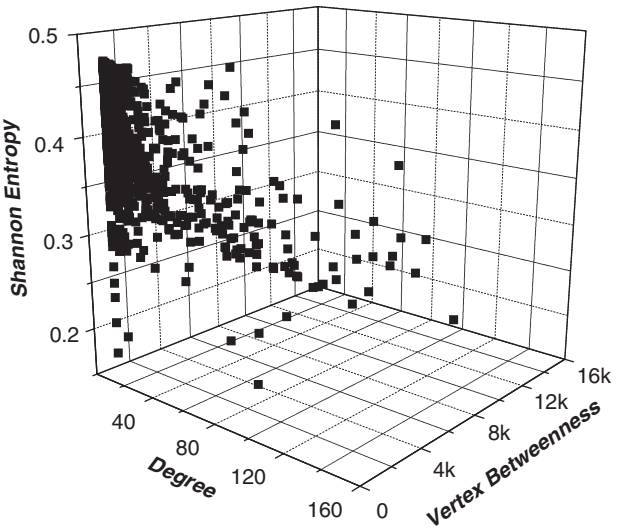

Fig. 5. Relationship between the gene's Shannon entropy, degree, and vertex betweenness.

form of the Shannon entropy to measure S_{ij}^{*}'s ability to be thresholds to distinguish normal person and abnormal patients by

$$H_{ij} = q_{ij}H_{ij}^{1} + (1 - q_{ij})H_{ij}^{2},$$

$$H_{ij}^{k} = -p_{ij,k}^{0} \ln p_{ij,k}^{0} - (1 - p_{ij,k}^{0}) \ln(1 - p_{ij,k}^{0}), \quad k = 1, 2,$$

where q_{ij} means the proportion of specimens whose expression level of gene i is higher than S^*_{ij}, while $p^0_{ij,k}$ denotes the proportion of normal specimens in specimen set H_{ij}. S^*_{ij} with maximum H_{ij} can be the terminal evaluation of gene i's ability to be candidates of biomarker. Finally, as illustrated in Fig. 5, we find that there do not exist a certain kind of genes that were more likely to be candidates for biomarkers. We can see that some genes with extremely low degrees and extremely low vertex betweenness have low Shannon entropies, but there still exists the same kind of genes with extremely high Shannon entropies. We may suppose that, in the actual directed gene regulatory networks, both functionally important genes and pathologically sensitive genes can be candidates for biomarkers.

4. Hints to understand lung cancer from features of the network

Recent research has approved that most of the cellular networks, such as protein interaction network [19,24–28], gene regulatory network [11,29], metabolic network [30,31], bear some common interesting features [32] that can be summarized as follows:

1. the power-law degree distribution,
2. small average shortest path and large average cluster coefficient,
3. a certain kind of structural pattern is predominant in the network,
4. there are modules in the network,
5. vertices with high degrees tend to connect with low degree ones.

In the previous section, we have demonstrated that our network also displays features 1–4 listed above. However, as a network of cancer, which is an abnormal state of life, it also displays features obviously different from those networks in normal state, such as feature 5 in the above.

All these findings indicate that the gene regulatory network of lung cancer is very compact. In such a system, any two given genes can directly or indirectly interact with each other through several different pathways. Such redundancy of signal transduction pathways may be one of the reasons for the robustness of lung cancer, as deletion of one pathway would not interdict communications between two genes.

Several distinct gene groups with individual biological functions in the network are tightly connected by a few of important genes and links. If these genes or links are deleted, gene groups will be separated into islands, and information communication between them will be blocked. Besides, there still exist another few genes, which are called ''hubs'', playing important roles in maintaining the integrity of the whole network. Former research [33] of such scale-free network has told us that this kind of system is robust to random attacks, but intentional attacks to these ''hub'' genes will make the network break down into a number of isolated vertices quickly. These provide us a novel measure to evaluate genes' importance in the network, and suggest a new way to find drug targets.

In addition, the network also self-organizes in a unique topology, where a certain kind of structural pattern appears significantly more frequent than others for the sake of fast signal transduction. We think that these are typical features of the gene regulatory network of lung cancer. Although we do not know whether these features are of lung cancer specificity, understanding of these characteristics is undoubtedly essential in further research on lung cancer.

5. Discussion

Interactions between genes are so complicated that simple correlation coefficient is not enough to unravel the real state of their regulatory relationships perfectly, especially in that it cannot indicate the direction of the relationship. Hence, the network proposed here may not be the actual gene regulatory network of lung cancer, and edges between genes represent association rather than causation relationships. However, it does partially reflect some important characteristics of the real network, and this can be demonstrated by the result of the cluster analysis.

Furthermore, we also attempt to weight each edge of the network by $1/|r_{ij}|$, which represents the intensity of the co-regulatory relationship between two genes, so as to describe the network more detailedly. Results have

demonstrated that weight does not affect the property of the network, at least under such kind of weighting manner.

Acknowledgements

The authors want to thank A. Bhattacharjee, J. Staunton, and their partners for their experimental data. This work was supported by Beijing Science and technology committee project Y0204002040111.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at 10.1016/j.physa.2006.02.034.

References

[1] C.A. Granville, P.A. Dennis, An overview of lung cancer genomics and proteomics, Am. J. Respir. Cell Mol. Biol. 32 (2005) 169–176.
[2] H. Uramoto, K. Sugio, T. Oyama, S. Nakata, K. Ono, T. Yoshimastu, M. Morita, K. Yasumoto, Expression of endoplasmic reticulum molecular chaperone Grp78 in human lung cancer and its clinical significance, Lung Cancer 49 (2005) 55–62.
[3] L.H. Hartwell, J.J. Hopfield, S. Leibler, A.W. Murray, From molecular to modular cell biology, Nature 402 (1999) C47–C52.
[4] Z.N. Oltvai, A.-L. Barabási, Life's complexity pyramid, Science 298 (2002) 763–764.
[5] A. Marshall, J. Hodgson, DNA chips: an array of possibilities, Nature Biotechnol. 16 (1998) 731.
[6] R. Albert, A.-L. Barabasi, Statistical mechanics of complex networks, Rev. Mod. Phys. 74 (2002) 47–97.
[7] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of networks, Adv. Phys. 51 (2002) 1079–1187.
[8] T.G. Dewey, From microarrays to networks: mining expression time series, Drug Discov. Today 7 (2002) s170–s175.
[9] S.-Y. Kim, S. Imoto, S. Miyano, Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, Biosystems 75 (2004) 57–65.
[10] N. Kasabov, Knowledge-based neural networks for gene expression data analysis, modeling and profile discovery, Drug Discov. Today: Biosilico 6 (2) (2004) 253–261.
[11] H. Agrawal, Extreme self-organization in networks constructed from gene expression data, Phys. Rev. Lett. 89 (2002) 268702.
[12] A. Lindlöf, B. Olsson, Could correlation-based methods be used to derive genetic association networks?, Inform. Sciences 146 (1–4) (2002) 103–113.
[13] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.
[14] A. Bhattacharjee, J. Staunton, et al., Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc. Natl. Acad. Sci. USA 98 (2001) 13790–13795.
[15] K.-I. Goh, E. Oh, B. Kahng, D. Kim, Betweenness centrality correlation in social networks, Phys. Rev. E 67 (2003) 017101.
[16] K. Suchecki, Scaling of distances in correlated complex networks, Physica A 351 (2005) 167–174.
[17] Y. Fan, M. Li, J. Chen, L. Gao, Z. Di, J. Wu, Network of econophysicists: a weighted network to investigate the development of econophysics, Int. J. Mod. Phys. B 18 (17–19) (2004) 2505–2512.
[18] M.E.J. Newman, J.Y. Park, Why social networks are different from other types of networks, Phys. Rev. E 68 (2003) 036122.
[19] S. Maslov, K. Sneppen, Specificity and stability in topology of protein networks, Science 296 (2002) 910–913.
[20] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (2002) 7821–7826.
[21] http://www.nature.com/nrd/journal/v3/n9/glossary/nrd1499_glossary.html
[22] W. Xizhao, H. Jiarong, On the handling of fuzziness for continuous-valued attributes in decision tree generation, Fuzzy Sets and Systems 99 (1998) 283–290.
[23] C.E. Shannon, W. Weaver, Mathematical Theory of Communication, University of Illinois Press, Champaign, IL, 1963.
[24] H. Jeong, S.P. Mason, A.-L. Barabási, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41–42.
[25] A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, Mol. Biol. Evol. 18 (2001) 1283–1292.
[26] L. Giot, et al., A protein interaction map of Drosophila melanogaster, Science 302 (2003) 1727–1736.
[27] S. Li, et al., A map of the interactome network of the metazoan, C. elegans, Science 2 (2004) (doi:10.1126/science.1091403).
[28] S.-H. Yook, Z.N. Oltvai, A.-L. Barabási, Functional and topological characterization of protein interaction networks, Proteomics 4 (4) (2004) 928–942.
[29] D.E. Featherstone, K. Broadie, Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network, Bioessays 24 (2002) 267–274.

[30] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.-L. Barabási, The large-scale organization of metabolic networks, Nature 407 (2000) 651–654.
[31] A. Wagner, D.A. Fell, The small world inside large metabolic networks, Proc. R. Soc. London B 268 (2001) 1803–1810.
[32] A.-L. Barabasi, Z.N. Oltvai, Network biology: understanding the cell's functional organization, Nat. Rev. Genet. 5 (2) (2004) 101–113.
[33] R. Albert, H. Jeong, A.-L. Barabási, Error and attack tolerance of complex networks, Nature 406 (2000) 378–382.