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.