Genomics xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Identifying colon cancer risk modules with better classification performance based on human signaling network

Xiaoli Qu [a,1], Ruiqiang Xie [a,1], Lina Chen [a,*,1], Chenchen Feng [a], Yanyan Zhou [a], Wan Li [a], Hao Huang [a], Xu Jia [a], Junjie Lv [a], Yuehan He [a], Youwen Du [a], Weiguo Li [a], Yuchen Shi [a], Weiming He [b,*]

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang Province Postal Code: 150081, China
[b] Institute of Opto-electronics, Harbin Institute of Technology, Harbin, Heilongjiang Province Postal Code: 150080, China

ARTICLE INFO

Article history:
Received 31 January 2013
Accepted 1 November 2013
Available online xxxx

Keywords:
Colon cancer
Module

ABSTRACT

Identifying differences between normal and tumor samples from a modular perspective may help to improve our understanding of the mechanisms responsible for colon cancer. Many cancer studies have shown that signaling transduction and biological pathways are disturbed in disease states, and expression profiles can distinguish variations in diseases. In this study, we integrated a weighted human signaling network and gene expression profiles to select risk modules associated with tumor conditions. Risk modules as classification features by our method had a better classification performance than other methods, and one risk module for colon cancer had a good classification performance for distinguishing between normal/tumor samples and between tumor stages. All genes in the module were annotated to the biological process of positive regulation of cell proliferation, and were highly associated with colon cancer. These results suggested that these genes might be the potential risk genes for colon cancer.

1. Introduction

Colon cancer is a complicated disease, the mechanisms of which remain largely unclear [1]. Efforts have been made in genome-wide analysis of gene expression profiles to identify novel cancer-related genes and to improve our understanding of the relevant molecular processes. For example, Smith et al. predicted recurrence and death in patients with colon cancer based on the metastasis-associated gene expression profile [2].

Although gene expression profiles can explore the pathogenesis of tumors at the microcosmic level, gene expression observations alone are generally insufficient to indentify causative or responsive roles of genes in complicated diseases [3]. It is well accepted that genes and proteins within a cell do not function alone, but interact with each other to form networks to carry out biological functions [4]. These networks help us to understand how complex molecular processes are activated in the cell, and reveal how cells respond to various conditions and environments [5]. Many methods have recently been developed to identify biomarkers based on gene expression datasets [6,7]. Gene expression data, combined with information on the interaction networks in which genes participate, may provide insights into the dynamic molecular mechanisms of cancers.

A study shows that the disturbances of signal transduction in cancer state are closely related to cell differentiation, proliferation and infection [8]. Biological signal transduction networks play a key role in modulating cell functions in response to extracellular and intracellular stimuli [9]. In signal transduction processes, a stimulus could be transformed into a cellular response through network modules that ultimately alter the function and behavior of the cell [10]. A previous study showed that consideration of signaling network modules can shed significant light on the mechanisms responsible for disease development [11].

In this study, we developed an expression-correlation method by integrating human signaling network and expression data for colon cancer, to identify risk modules and evaluate the classification performance in colon cancer.

This integrated analysis could provide new insights into complex diseases at the system level through the identification of the signal network modules.

2. Materials and methods

In this article, we developed an expression-correlation method to identify risk modules for the classification of colon cancer by integrating the signaling network and gene expression profiles. We compared our expression-correlation method with average expression-value and unweighted methods and evaluated their classifying performances (Fig. 1).

* Corresponding authors.
E-mail addresses: quxiaoli1030@gmail.com (X. Qu), xrq1989@126.com (R. Xie), chenlina@ems.hrbmu.edu.cn (L. Chen), fengchenchen.finally@gmail.com (C. Feng), yanyanzhou2011@163.com (Y. Zhou), wendyliwan@gmail.com (W. Li), huanghaobio@gmail.com (H. Huang), jiaxu.happy@gmail.com (X. Jia), lvjunjie525@126.com (J. Lv), heyuehan56@126.com (Y. He), duyouwen215@163.com (Y. Du), lwgbioinfor@gmail.com (W. Li), shiyuchen1988@hotmail.com (Y. Shi), hewm@hit.edu.cn (W. He).
[1] These authors contributed equally to the work.

0888-7543/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.ygeno.2013.11.002

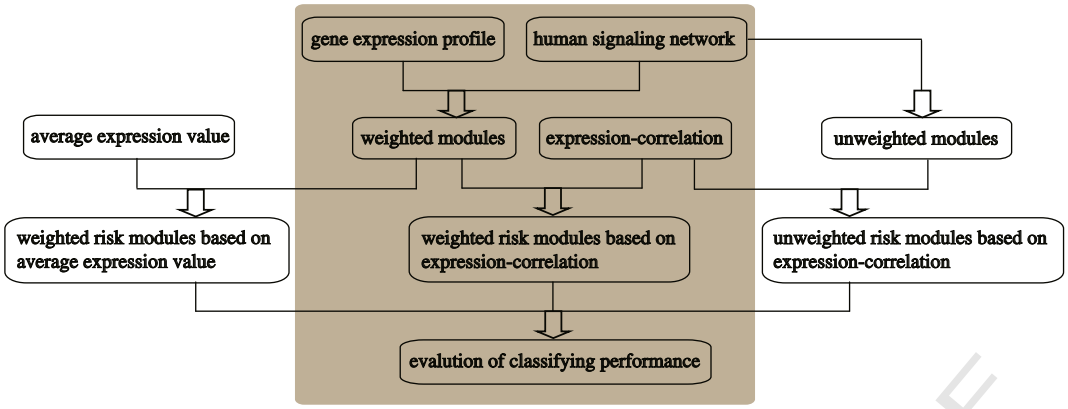

Fig. 1. The flowchart of expression-correlation method and other methods comparison. The gray background is expression-correlation method.

2.1. Data source

Human signaling network information was derived from a previous study, which contained 5089 interactions among 1634 genes [12]. Three types of interactions are recognized: including activation, inhibition, and physical interaction.

We integrated the gene expression and signaling network data by mapping the gene expression values for each gene into the network. Four colon cancer gene expression datasets (GSE10950, GSE10972, GSE24993, GSE8671) were extracted from the GEO database [13].

2.2. Network modules

The weighted signaling network was constructed by calculating Pearson correlation coefficients between genes in the human signaling network. Network modules were mined using the online tool GraphWeb [14] in the weighted signaling network. GraphWeb provides a method to identify network modules using the Markov clustering algorithm [15] (http://biit.cs.ut.ee/graphweb/). The parameter of Markov clustering parameter was therefore set to a default value 1.8. Modules containing at least four genes were selected.

2.3. Common modules

Common modules were those at the intersection of overlaps between normal and tumor modules. Common modules could reveal the difference between normal and tumor conditions. We selected the common modules using two steps. First, we determined the overlap of same condition (normal/tumor condition) modules between two expression profiles, in order to improve the reproducibility. For two modules, overlap modules were defined if the percentage of common genes was >50%. We then screened the overlap modules between normal and tumor modules (overlap genes >50%), and considered the intersections as common modules.

2.4. Cancer-associated risk modules

Cancer-associated risk modules were identified by screening for significant changes in gene expression between normal and tumor samples. This method was called the expression-correlation method. The expression-correlation differential score was used as a measure to evaluate the expression changes. Given a common module M with E1… Em representing m edges of the module M, the expression-correlation differential score S was defined by:

$$S(M) = \sum_{k=1}^{m} \left| E_k - E_k^{'} \right| \quad (1)$$

$$E_k = \text{pearson}(X, Y) = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\left(\sqrt{\sum_{i=1}^{n1} (X_i-\bar{X})^2}\right)\left(\sqrt{\sum_{i=1}^{n1} (X_i-\bar{X})^2}\right)} \quad (2)$$

$$E_k^{'} = \text{pearson}\left(X', Y'\right) = \frac{\sum \left(X'-\bar{X'}\right)\left(Y'-\bar{Y'}\right)}{\left(\sqrt{\sum_{i=1}^{n2} \left(X_i'-\bar{X'}\right)^2}\right)\left(\sqrt{\sum_{i=1}^{n2} \left(X_i'-\bar{X'}\right)^2}\right)} \quad (3)$$

where (X, Y) and (X′, Y′) are the gene expression values under normal and tumor conditions, respectively, and E_k and E_k' are the Pearson correlation coefficient of the kth edge connecting two genes under normal and tumor conditions, respectively. n1 and n2 are the number of samples for normal and tumor, respectively. For a common module, we calculated the real differential score S, and 1000 degree-conserved random modules were then constructed and the random differential scores S1…S1000 were calculated. If the real differential score was significantly greater than the random ones (permutation test, $p < 0.05$), the module was considered as a risk module for colon cancer.

2.5. Classification and evaluation of risk modules

We considered risk modules as classification features and applied the Support Vector Machine (SVM) method to classify patients with normal and tumor samples. We then applied a receiver operating characteristic (ROC) curve to estimate classification performance.

In the ROC curve, tumor samples were considered as positive and normal samples as negative. We then selected a training set for machine learning, and used a test set to evaluate the classification performance. The area under the curve (AUC) reflected the classification performance. A larger AUC represented a better classification performance.

2.6. Jonckheere–Terpstra test

Jonckheere–Terpstra test is a nonparametric test and is a test for an ordered hypothesis within an independent samples design. It is used to test whether there is a significant difference in the distribution of the

Table 1
The number of modules in normal and tumor conditions.

	GSE10950	GSE10972	Overlap
Normal	127	137	109
Tumor	133	124	112