Mathematical Biosciences 276 (2016) 1–7

Contents lists available at ScienceDirect

Mathematical Biosciences

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

Finding gastric cancer related genes and clinical biomarkers for detection based on gene–gene interaction network

Xuesong Wu[a], Haoran Tang[a], Aoran Guan[b], Feng Sun[a,*], Hui Wang[c], Jie Shu[a]

[a] Department of Gastrointestinal Surgery, The Second Affiliated Hospital of Kunming Medical University, Kunming 650101, Yunnan, China
[b] Department of General Surgery, Yan'an Hospital of Kunming Medical University, Kunming 650051, Yunnan, China
[c] Department of Gastroenterology, Yan'an Hospital of Kunming Medical University, Kunming 650051, Yunnan, China

ARTICLE INFO

Article history:
Received 2 September 2015
Revised 27 October 2015
Accepted 4 December 2015
Available online 14 December 2015

Keywords:
Gastric cancer
Prognostic gene for clinical detection
Gene ontology

ABSTRACT

Background/objective: Gastric cancer (GC) is the second leading cause of death resulted from cancer globally. The most common cause of GC is the infection of Helicobacter pylori, approximately 11% of cases are caused by genetic factors. The objective of this study was to develop an effective computational method to meaningfully interpret these GC-related genes and to predict potential prognostic genes for clinical detection.

Methods: We employed the shortest path algorithm and permutation test to probe the genes that have relationship with known GC genes in gene–gene interaction network. We calculated the enrichment scores of gene ontology and pathways of gastric cancer related genes to characterize these genes in terms of molecular features. The optimal features that primly representing the gastric cancer related genes were selected using Random Forest classification and incremental feature selection. Random Forest classification was also used for the prediction of the novel gastric cancer related genes based on the selected features and the identification of novel prognostic genes based on the expression of genes.

Results: Based on the shortest path analysis of 36 known GC genes, 39 genes occurring in shortest path were identified as GC-related genes. In subsequent classification, 4153 gene ontology terms and 157 pathway terms were identified as the optimal features to depict these gastric cancer related genes. Based on them, a total of 886 genes were predicted as related genes. These 886 genes could serve as expression biomarkers for clinical detection and they achieved a 100% accuracy for distinguishing gastric cancer from a case-control dataset, better than any of 886 random selected genes did.

Conclusion: By analyzing the features of known GC-related genes, we employed a systematic method to predict gastric cancer related genes and novel prognostic genes for accurate clinical detection.

1. Introduction

With the increasing improvement of medical conditions and healthier methods to preserve and process food, a dramatic decline of the incidence of gastric cancer has been achieved in the past few decades [1]. In 2003, about 10% of patients with invasive cancers were gastric cancers all around the world, of which 700, 349 people were killed. Gastric cancer has become one of the leading causes of death second to lung cancer and also one of the main enemies to human health since more and more patients died of gastric cancer than the other kinds of cancers [2]. However, an epidemiological survey concerning cancers showed that the death toll of gastric cancer has dropped to the third place in male patients with cancers and the fifth place in female patients with cancers in the past ten years. Moreover, the declining number of deaths was accompanied by the increasing number of 5-year life expectancy of patients with gastric cancer. Of the patients with gastric cancer in America, about 29% of them might be expected to maintain 5-year life expectancy after diagnosis while this percentage was only 22% six years ago [3].The fear is that patients with gastric cancer still account for a large proportion of patients with cancers in developing countries, especially China, although the threat of gastric cancer is decreasing significantly in developed countries. A statistic report published in Chinese Journal of Cancer Research indicated that 48.41 patients out of 10,000 died of gastric cancer, which is no better than the lung cancer death rate of 48.58/105 in China's remote rural areas. What is worse, only 20% of these patients with gastric cancer are likely to enjoy the 5-year life expectancy after diagnosis [4]. Thus, we can see that gastric cancer is still a considerable threat to the lives and health of people of our country.

Considering the serious situation of gastric cancer incidence in China, methods like early diagnosis and targeted therapy are necessary to be adopted for the purpose that the patients with gastric cancer may get timely and effective treatments [5]. However, these

[*] Corresponding author. Tel.: +86 87165351281.
E-mail address: waiycmarker@aliyun.com, fengsun76@yahoo.com (F. Sun).

http://dx.doi.org/10.1016/j.mbs.2015.12.001
0025-5564/© 2016 Published by Elsevier Inc.

methods are not mature enough to face all varieties of challenges. The first one is the problem of individual specificity [6]. Since each patient has his/her own special genetic background and eating habit that might affect the development and progression of gastric cancer in their body modes, thus how to adjust the general mode of treatment to adapt to each specific biological characteristic is the key to bring our treatments into the fullest play. In view of this, quite a lot of hospitals and laboratories are doing relevant researches, unfortunately, their research progress is very slow due to the long experiment period and high cost [7, 8]. Even if considerable manpower and material resources have already been devoted, the gastric cancer related genes discovered are few and far between. What is more, most of them are just doing the repetition of one or two studies of genes and many useful resources are exceedingly wasted [9–12]. Gastric cancer is a kind of relatively complex disease. The causes of gastric cancer include the direct pathogenic genes and abnormal expression of regulatory proteins. In addition, the coverage of gastric cancer related genes is more general while its metabolic mechanism is more complex because of the interference of various subtypes and external factors. Therefore, there might exist considerable gastric cancer related genes that have not been found yet. In order to expand our understanding of gastric cancer related genes, a large number of genome-wide association studies were published in all kinds of magazines with each of which listing from dozens to hundreds of so-called gastric cancer related genes since 2007, and the only parameter to prove the correlation is the significance degree of statistical test [13, 14]. However, in 2010, McGlellan and other scientists wrote in their article, published in the journal Cell, that the genes found by such simple and crude way of analysis were difficult to identify their true biological relationship with associated diseases, so they meant nothing to actual clinical diagnosis and treatments [15]. In consequence, we urgently need a swift, economic and effective approach to help us find the right gastric cancer related genes for the purpose of future drug target screening, determination of individualized treatment plan and timely diagnosis and prognosis for patients with early-stage cancers.

In view of the above problem, we consider it is necessary to see the relationship between pathogenesis and susceptibility genes of gastric cancer from a different perspective of network to find more gastric cancer related genes [16, 17], starting from the development and progression of gastric cancer. Based on this idea, we developed a set of method that might be used to predict the other possible gastric cancer related genes on the basis of the existing gastric cancer related genes. First, we collected the genes whose relationship with the pathogenesis and susceptibility of gastric cancer have already been reported and verified from experiments or considerable population data analysis in recent years before the application of such method for the prediction of gastric cancer related genes. Then, we sought for some other genes closest to original gene sets in the network of gene dependency (genetic correlation) by using the shortest path algorithm. Next, we combined the two kinds of genes together with the help of a set of machine learning method based on Random Forest algorithm to find the special features that might be used to distinguish gastric cancer genes from non-gastric cancer genes to the utmost extent. The special features were divided into two categories, the GO (Gene Ontology) annotations representing gene function and the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway annotations representing metabolic pathway where the gene is located. After the proper optimal feature was selected, we continued to seek for more gastric cancer related genes in the rest of human genes. Finally, in order to verify the reliability of the genes predicted, we randomly selected some human genes with the purpose of the attempt to distinguish patients with gastric cancer from the control group. It turned out that our selected genes achieved a 100% accuracy while the randomly selected genes were powerless to this. Our predicted genes lay the foundation for conquering gastric cancer since they could not only be used as a novel prognostic feature gene, but also offer us help to screen more target drugs for gastric cancer subtypes, even individual specificity.

2. Materials and methods

2.1. Data collection

Access to relevant literatures published since 2014, we collected 36 genes related to the pathogenesis and susceptibility of gastric cancer backed by data. A study published in the International Journal of Cancer in September 2014 showed that there existed a strong statistical correlation between the SNP locating at 16 kb from the upstream of IL1A (SNP and IL1A are in the same linkage group) and the infection of diseases by making genome-wide association analysis of a total of 365 cases of gastric cancer patients from 10 European countries with 1284 cases of healthy people from the control group. Further haplotype experiments verified the significant correction between the mutation of IL1A gene region and the infection of gastric cancer [18]. Another study based on the same set of data placed close attention on miRNA and the researchers found that the SNP in the regions of miR25, miR29, miR93, miR106b, SNP had a high correlation with the incidence of gastric cancer. On the one hand, continuous reports covered the correction between these identified miRNA and gastric cancer. On the other hand, researchers found that the expression level of gastric cancer related genes made corresponding changes when down-regulating the expression of these miRNA, thereby proved that there might existed correction between miR25, mir29, miR106b, miR93 and the susceptibility of gastric cancer, and they even participated in the process of gastric carcinogenesis [19]. In addition, another research group participated in the program summarized the previous findings and made genetic typing for the 30 SNPs in distribution with CD14, TLR4, NOD2 and NFKB as the starting point. The conclusion, that ‘NOD2 and NFKB1 are in correlation with non-cardia gastric cancer while NFKB1 is in correlation with cardia gastric cancer’, was drawn from establishing a logistic regression model with multiple afterward tests correction [20]. An article published in Gastric Cancer in April 2014 also adopted such statistical method and it showed that the specific mutations of identified CYP2E1, CYP1A2 and CYP1A1 often occur in patients with gastric cancer from the 88 patients with gastric cancer and 170 healthy people. At the same time, the infection of Helicobacter pylori was also detected with the result that the mutation frequency of the three genes was higher in patients infected with Helicobacter pylori and easier to cause the susceptibility of gastric cancer to body [21]. Moreover, both NOD2 and NFKB1 were found in correction with non-cardia gastric cancer while CD14 was identified in correction with cardia gastric cancer, and the conclusion was based on a SNP test for 365 patients with gastric cancer and 1284 matched healthy people. Together with the highly credible genes related to the susceptibility or pathogenesis of gastric cancer collected from other literatures [22–28], a total of 36 genes were collected as the first gastric cancer related genes set (Supplementary Table 1).

2.2. Obtaining gene–gene interaction data from STRING

The gene–gene interaction network used in this paper could be obtained from STRING [29]. SRING (http://string-db.org/) is a kind of system used to search for the interaction between the known proteins and predicting proteins. The foregoing interaction not only includes the direct physical interaction of proteins and the indirect function relevance of proteins, but also experimental data and bioinformatics computer prediction techniques. STRING weighs these results from different methods and gives a comprehensive grade for them. There exists a correlative score which shows the possibility of interaction between two interactional genes.