Computers in Biology and Medicine 72 (2016) 22–29

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Identification of mutated driver pathways in cancer using a multi-objective optimization model

Chun-Hou Zheng [a], Wu Yang [a], Yan-Wen Chong [b], Jun-Feng Xia [c,*]

[a] College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230601, China
[b] State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[c] Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China

ARTICLE INFO

Article history:
Received 19 October 2015
Received in revised form
4 March 2016
Accepted 4 March 2016

Keywords:
Driver pathways
Multi-objective optimization model
Genetic algorithm
Integrative model
Driver mutation

ABSTRACT

New-generation high-throughput technologies, including next-generation sequencing technology, have been extensively applied to solve biological problems. As a result, large cancer genomics projects such as the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium are producing large amount of rich and diverse data in multiple cancer types. The identification of mutated driver genes and driver pathways from these data is a significant challenge. Genome aberrations in cancer cells can be divided into two types: random 'passenger mutation' and functional 'driver mutation'. In this paper, we introduced a Multi-objective Optimization model based on a Genetic Algorithm (MOGA) to solve the maximum weight submatrix problem, which can be employed to identify driver genes and driver pathways promoting cancer proliferation. The maximum weight submatrix problem defined to find mutated driver pathways is based on two specific properties, i.e., high coverage and high exclusivity. The multi-objective optimization model can adjust the trade-off between high coverage and high exclusivity. We proposed an integrative model by combining gene expression data and mutation data to improve the performance of the MOGA algorithm in a biological context.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Cancer is a complex disease that is highly driven by somatic mutations, such as single-nucleotide mutations, larger copy-number aberrations, and structural aberrations [1]. The hallmarks of cancer, including a limitless replicative potential and self-sufficiency in the growth signals of cancer, hinder its effective treatment. The researchers also found these cancer cells can invade other tissues and organs via the lymphatic system or blood circulation [2]. Thus, cancer is considered to be one of the most serious diseases to inflict human beings.

Researchers should understand the molecular mechanisms of cancer prior to exploring the effective treatment of cancer. Therefore, one of the remaining challenges is to identify functional mutations, which are referred to as "driver mutations", that confer a selective growth advantage during the development of cancer, and filter random "passenger mutations", which are neutral to the proliferation of cancer cells [3]. With the development of sequencing technologies, especially next-generation DNA sequencing technologies, large-scale genomic projects of cancer disease, such as The Cancer Genome Atlas (TCGA), have provided a vast number of profiles of samples for different types of cancers [4–6]. Recent studies revealed that the theories of carcinogenesis are primarily attributed to the disruption of cellular signaling and regulatory pathways. Thus, the design of an effective method for identifying the driver pathways in carcinogenesis process, which is helpful to personalized medicine programs, is important [7,8].

On the gene level, several studies have been devoted to identify genes with significantly higher mutation rates across samples than the background mutation rates in a vast number of samples. Some important gene mutations in cancer progression have been reported in these studies. However, minimal overlap among genome aberrations is observed even if cancer genomes originate from the same cancer [9,10]. The majority of these methods are unable to identify the heterogeneity of gene mutations.

Researchers have also realized that one driver pathway may be caused by different mutations [11,12]. Pathways rather than individual genes are considered to govern cancer progression. Therefore, it is necessary to shift the point of view from the gene level to the pathway level and it is critical to capturing the heterogeneous phenomenon in cancer [13,14]. The majority of studies have focused on analyzing known information about driver pathways and the detection of the significant pathways. The enrichment of somatic mutations is a routine method in most of studies [10,15]. The disadvantage of this method is that the background knowledge of

* Corresponding author.

http://dx.doi.org/10.1016/j.compbiomed.2016.03.002
0010-4825/© 2016 Elsevier Ltd. All rights reserved.

pathways remains incomplete, and existing pathway databases have distinct limitations because they frequently contain overlap and unavailable data [16–18]. Considering these limitations, it is indispensable to develop new methods to identify driver pathways without relying on prior knowledge.

A vast number of gene sets exist when exhaustively testing in the entire genome. For instance, more than 10^{26} sets of seven human genes exist [19]. Therefore, the detection of mutated driver pathways seems implausible due to the vast number of gene sets to enumerate. In recent years, some methods have been developed to solve this problem [20]. Researchers have discovered two constraints on the combination patterns of the mutations of cancer. First, they determined that a driver mutation is rare, that is, a single mutation in one group can sufficiently perturb one pathway. Driver mutations are mutually exclusive. This property is referred to as high-exclusivity. Second, a pathway that is important to cancer frequently applies to the majority of patients. Thus, mutations in a driver pathway should be contained by most patients, which is referred to as high-coverage. Recently, some new gene sets have been identified in several studies using these rules [21–23]. A novel scoring function that is based on these two properties, which is employed to identify the mutated driver pathway, has been defined by Vandin et al. [19]. It is a new and effective method that applies the somatic mutation data detected by next-generation DNA sequencing technologies. They defined the maximum weight submatrix problem as a problem in which the scoring function is maximized and solved this problem using a stochastic search algorithm. However, the solution of this problem is difficult due to the complex computations that are required. No effective measure is available to adjust the trade-off between coverage and exclusivity in this study.

To adjust the trade-off between higher coverage and higher overlap in the identification of gene sets for solving the maximum weight submatrix problem, we proposed a novel multi-objective optimization model that is based on a genetic algorithm (MOGA) for driver pathway detection. We integrated differential expression data to improve the performance of our MOGA in a biological context. The results show that the proposed MOGA method is effective and the integrated model can detect biologically significant gene sets.

The remainder of our paper is divided into three sections. In Section 2, the methods that was utilized in this study are briefly introduced. In Section 3, to assess the performance of our MOGA method, we apply MOGA to two biological data sets to compare our results with the Markov Chain Monte Carlo (MCMC) method. Section 4 briefly outlines our MOGA method in future studies.

2. Materials and methods

2.1. Glioblastoma and lung datasets

For the glioblastoma and lung datasets, all genes with non-synonymous single nucleotide mutations or small indels in at least two patients were used. For glioblastoma dataset, copy number variants (CNVs) were included. If a gene had a CNV of the same type (amplification or deletion) in at least 90% of the patients with a CNV, we add a CNV type for the gene. More details about the pipeline for building mutation matrices from somatic mutation data are prepared was discussed in Ref. [24].

2.2. Maximum weight submatrix problem

In recent years, researchers have employed two important characteristics to detect mutated driver pathways. They transformed this problem into the maximum weight submatrix problem by considering two constraints, i.e., 'high coverage', which indicates that the majority of patients have a minimum of one mutation in a driver pathway, and 'high exclusivity', which indicates that the mutations in one pathway should be heterogeneous [19]. The identification of driver pathways is extremely difficult because the maximum weight submatrix problem is NP-hard. Considering this point, they proposed the Markov Chain Monte Carlo (MCMC) method, which is based on these two properties, to detect effective gene sets. A binary mutation matrix A with m rows (samples) and n columns (genes), which is based on the somatic mutation data, is constructed, where $A_{ij}=1$ means gene j is mutated in patient i and $A_{ij}=0$ indicates the absence of a mutation. The fitness function is defined as

$$W(M)=|\Gamma(M)|-\omega(M)=2|\Gamma(M)|-\sum_{g\in M}|\Gamma(g)| \quad (1)$$

where M *is a* submatrix of mutation matrix A with m rows and k columns, $\omega(M)=\sum_{g\in M}|\Gamma(g)|-|\Gamma(M)|$ measures the coverage overlap in a set M of genes. $\Gamma(M)=\cup_{g\in M}\Gamma(g)$ is defined to measure the coverage of submatrix M. $\Gamma(g)=\{i:Aig=1\}$ indicates that the gene g is mutated in the set of samples. In other words, given a mutation matrix A and an integer $k>0$, the the maximum weight submatrix problem is defined as finding an $m\times k$ submatrix of A

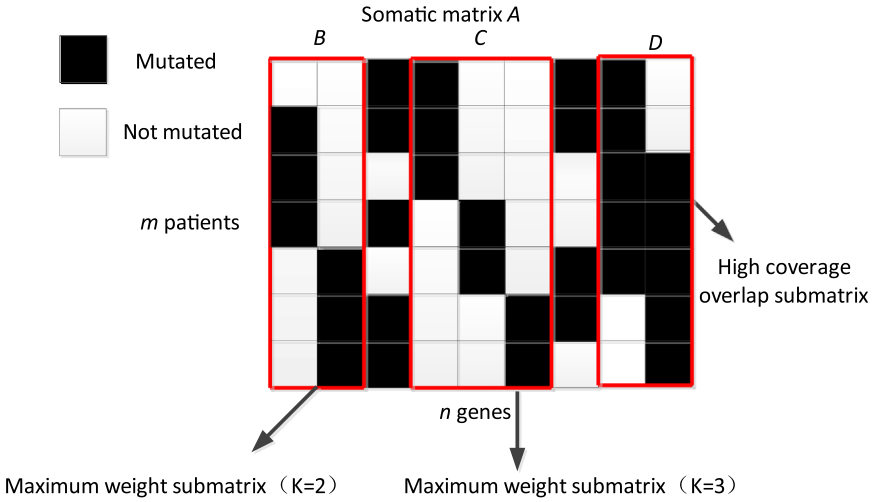

Fig. 1. Illustration of the binary somatic mutation matrix A. The mutated driver pathway (gene sets) are expected to be identified as submatrices meet two criteria.

Download English Version:

https://daneshyari.com/en/article/6920793

Download Persian Version:

https://daneshyari.com/article/6920793

Daneshyari.com