CSIAM Trans. Life Sci.
doi: 10.4208/csiam-ls.SO-2024-0002

Vol. **1**, No. 1, pp. 22-44
March 2025

Identifying Rarely Mutated Cancer Genes by Heterogeneous Network Embedding

Yurun Lu[1,2], Songmao Zhang[1] and Yong Wang[1,2,3,*]

[1] *CEMS, NCMIS, HCMS, MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.*
[2] *School of Mathematics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100049, China.*
[3] *Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, 310024, China.*

Received 11 December 2023; Accepted 9 May 2024

Abstract. Cancer is a multifaceted disease caused by dynamic interaction between genetic mutations and environmental factors. Understanding the genetic mutations underlying the development and progression of cancer is the stepstone for developing effective treatments and therapies. However, these mutations occurred in only a small fraction of cancer patients and it is extremely difficult to associate with cancer. Here, we propose MutNet, a heterogeneous network embedding method which integrate biomolecular network with cancer genomics data. Using pan cancer genomic data from The Cancer Genome Atlas program and public protein-protein interaction and pathway data, MutNet identifies rarely mutated cancer genes often overlooked by conventional genetic studies. In addition, the unified vector representation of biological entities allows us to reveal the tumor type specific cancer genes, cancer gene modules, and potential relationships among different tumor types. Our heterogeneous network embedding method holds the promise for the underlying mechanisms of cancer and potential therapeutic targets.

AMS subject classifications: 92B20, 92-08, 68T07

Key words: Cancer genomics, cancer gene, network embedding.

1 Introduction

Cancer is a complex disease caused by a combination of genetic and environmental factors. The genetic alterations (mutation, amplification, deletion, etc.) can change gene's

*Corresponding author. *Email address:* ywang@amss.ac.cn (Y. Wang)

normal function, which in turn lead to the uncontrolled growth and division of cancer cells [3]. Identification of these cancer genes is a key goal of cancer genomic analysis and stepstone in the development of precision oncology and cancer therapeutics [15,33,37,47]. Some cancer genes are frequently altered across many different types of cancer and can be easily identified, such as TP53, KRAS, and BRAF [47]. They are well studied in cancer development and progression by disrupting cell cycle regulation, DNA repair, or signaling pathways [37] in biomolecular network. However, some mutated cancer genes are altered in only a small fraction of cancer patients or a particular cancer type or subtype and are often overlooked in genetic studies [15,47]. These rarely mutated genes may play a significant role in the development and progression of cancer through propagating information via biomolecular network. Therefore, it is in pressing need to develop novel methods to identify these rarely mutated genes.

Whole exome sequencing (WES) allows researchers to sequence all of the protein-coding regions of the genome, known as the exome, to identify genetic mutations associated with cancer [33]. With the rapid accumulation of whole exome sequencing data, several algorithms have been developed to detect genes that are significantly mutated in cancer, such as ActiveDriver [33], TUSON [9], MuSiC [10] and MutSigCV [25]. They used statistical framework to identify significantly mutated genes based on the frequency and distribution of somatic mutations across tumor samples. OncodriveFM [16], OncodriveFML [29], OncodriveCLUST [42], and OncodriveCLUSTL [1] calculate the likelihood that a given gene is under positive selection in tumor samples based on the frequency and distribution of somatic mutations as well as gene-gene interaction information. 20/20+ [45] uses a machine learning algorithm to analyze WES data and predict driver genes that are likely to contribute to cancer development.

Despite the success of the above computational methods, there is still room to borrow information from gene interactions to boost rarely mutated cancer gene discovery. It's well known that genes play crucial roles in various biological processes by acting in concert with each other within signaling and regulatory pathways, as well as in protein complexes [22]. The coordinated action of multiple genes allows cells to carry out complex functions such as growth, differentiation, and response to environmental stimuli. Therefore, network-based methods, such as HotNet2 [26] and OMEN [46] have been developed to detect cancer genes based on the interaction network and mutation patterns observed in WES data. Recently, EMOGI [36] integrates genomic data with other multi-omics data with graph convolutional networks to identify cancer genes. P-NET exploits hierarchical structure of biological pathways with convolution neural network to reveal molecularly altered candidates [13].

However, the above network-based methods are limited to their capacity of single type of molecules and associations, such as protein-protein interactions (PPI) and gene co-expression relationships, and cannot capture the full complexity of biological systems operating at many different levels [4, 24, 52]. We note that gene functional databased including pathways [48], gene ontology [44] provide rich information from different aspects. One way to capture all the information is constructing a heterogeneous network to