Computers and Electrical Engineering 67 (2018) 267–277

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Integration of heterogeneous 'omics' data using semi-supervised network labelling to identify essential genes in colorectal cancer☆

David Chisanga[a], Shivakumar Keerthikumar[b], Suresh Mathivanan[b], Naveen Chilamkurti[a,*]

[a] Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria 3086, Australia
[b] Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria 3086, Australia

ARTICLE INFO

Keywords:
Proteomics
Protein-protein interactions
Big data
Genomics
Networks
Colorectal cancer
Biomarkers
Machine learning

ABSTRACT

Colorectal cancer (CRC) is the third most common form of cancer and has the fourth highest mortality rate in the world. To understand the origin and progression of this disease, biomedical researchers undertake global analyses of omics data of CRC patient samples and representative cell lines. However, due to the heterogeneity and high dimensionality nature of `omics' data, traditional tools for analysing this sort of data are inadequate and the heterogeneous nature of cancer makes the process of identifying essential genes very difficult. 'Omics' is a term that is used to refer to areas of study in biology that end with the ending 'omics' such as genomics, proteomics and metabolomics. This paper uses network theory-based methods to address the problem of high dimensionality in omics datasets and applies network propagation to address the problem of heterogeneity in both omics datasets and cancer in identifying the essential genes. The method successfully identifies known essential genes in CRC as well as a new set of genes that are likely to be essential in the study of CRC.

1. Introduction

Network theory, the study of how complex systems interact is widely applied in fields such as computer networks, social networks, and interactome networks in systems biology [1]. Network metrics such as node degree are often used to prioritise nodes within a network. Similarly, one of the main goals in cancer research is the identification of biomarkers or essential genes that can be used to understand the development or progression of a specific cancer type such as Colorectal cancer (CRC).

To prioritise these genes, researchers often study the complex interactions between the numerous molecules within cells such as proteins, deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and other small molecules. The molecules are obtained from the global profiling of patient samples as well as representative cell lines at multiple layers, these layers constitute what is today referred to as 'omics' data. 'Omics' is an informal term that is used to refer to areas of study in biology that end with the term 'omics' such as genomics, proteomics and metabolomics [2]. The interactions, on the other hand, are collectively known as interactome networks and provide a global picture of how molecular interactions influence cellular behaviour, an example being protein-protein interactions (PPI) [3].

Omics data is highly dimensional in nature, coupled with this, is the heterogeneity of cancer whereby two individuals with the same type of cancer may have a different set of biomarkers. This makes identifying and prioritising cancer-related genes a challenging

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. A. Sangaiah.

* Corresponding author at: Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria 3086, Australia.
E-mail address: N.Chilamkurti@latrobe.edu.au (N. Chilamkurti).

https://doi.org/10.1016/j.compeleceng.2018.03.039
Received 4 July 2017; Received in revised form 19 March 2018; Accepted 19 March 2018

and daunting task that cannot be achieved using traditional statistical methods. As such, network theory provides a means by which complexity in such instances can be used to model the cellular system behaviour. Barabási, et al. [4] provides a summation of the application of network-based metrics in associating omics-related molecules to disease. Other works in [5–7] applied network-based methods in areas such as identifying and associating genes to disease as well as identifying drug targets in various cancer types. In [8,9], integrated network-based methods with machine learning techniques are applied in reducing the dimensionality of omics data and building models to predict genes associated with the disease as well as classify multiple cancer types. While the integration of omics data with networks has been gaining momentum over the years, a typical recurring theme in most of the research has been the use of a single type of omics data as opposed to integrating the various types of omics data which are heterogeneous in nature.

In this paper, we used an integrated approach to identify essential genes in colorectal cancer, a type of cancer that originates in the bowel, is the third most common form of cancer and has the fourth highest cancer mortality rate in the world [10]. The integrated approach employed a semi-supervised learning algorithm to propagate heterogeneous omics data into a protein-protein interaction network, which was followed by a downstream enrichment analysis to validate and understand the role of the predicted potential essential genes in CRC.

The rest of the paper is organised as follows: Section 2 provides a description of the materials and methods used as well as an overview of related works, Section 3 provides a discussion of the experimental results and the implications of the findings. The paper concludes with a summary of the findings and the future directions of the research.

2. Materials and methods

2.1. Proteomics data

We used proteomics and genomics data as the input to our method. Proteomics data consisted of protein-protein interactions. Weighted protein-protein interactions were downloaded from HIPPIE Version 2.0 [11], an online web-based database resource for weighted protein-protein interactions. The weights in the interactions show the confidence in the interaction between two proteins and are calculated by the authors based on the amount and reliability of evidence supporting an interaction. The protein-protein interaction dataset was then filtered to leave out interactions with a confidence score of 0 after which 16,728 number of unique proteins and 276, 183 number of interactions remain. These were then assembled into a network using NetworkX, a Python package for network manipulation and analysis.

2.2. Genomics data

Genomics data comprised gene somatic mutations and gene differential expression status for CRC patients and representative cell lines. Previously, we collated genomics data related to CRC into a web-based resource called the Colorectal Cancer Atlas [12]. It is this data together with The Cancer Genome Atlas (TCGA) patient data obtained from COSMIC [13] that we used as the genomics input data to our method. Using the corresponding genes for the proteins identified above, we obtained gene mutation details of 564 CRC patients from TCGA.

From the mutation dataset, we then filtered out all silent mutations and for each gene with a mutation in each sample, we represented its status using a binary number (1 if a mutation was present and 0 if not present) regardless of the number of mutations for a gene in each sample. The mutation data were then represented as a matrix, $\mathbf{M}$ (16,728 × 564) with rows representing genes and columns representing a gene's mutation station status in each sample. The same was repeated for gene differential expression status in TCGA patient data. This was then represented as a matrix, $\mathbf{D}$ (16,728 × 564) with rows representing genes and columns representing the differential expression status of genes in each sample. The gene differential expression status was denoted 1 for under-regulated or up-regulated genes and 0 for genes not differentially expressed.

2.3. Theory/calculation

To identify essential genes, we use a method that integrates the different datasets discussed in the materials and methods section. Fig. 1 provides a summary of the approach taken in this paper.

2.4. Disease gene prioritisation using network theory methods

A network or a graph is defined as a set of objects (nodes) linked together by lines (edges) [1]. A network is, therefore, represented as an ordered pair G = (V, E) where V is the set of nodes and E is the set of edges. By grouping a collection of objects as a set of nodes and using edges to represent relationships between these objects, researchers have used networks to reduce the complexity of large systems. Molecular networks in biology provide a global representation of the complex interactions between various molecules within a cell such as DNA, RNA and other small molecules.

When it comes to disease-gene prioritisation, many researchers use networks to associate genes with diseases. A naïve approach that is usually taken is to predict those genes that have neighbours associated with a disease as being more likely to be implicated in such a disease, that is using the concept of “guilty by association”. Such methods that implicate neighbours as having the likelihood of being associated with a disease include node degree as well as shortest path methods. However, these methods are prone to false positives because of the biases that exist in current molecular networks’ datasets where proteins which are well studied tend to have