Gene Reports 12 (2018) 276–288

Contents lists available at ScienceDirect

Gene Reports

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

Biological networks integration based on dense module identification for gene prioritization from microarray data

S. Mahapatra*, B. Mandal, T. Swarnkar

Department of Computer Application, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India

ARTICLE INFO

Keywords:
Gene expression
PPI networks
Dense modules
Gene correlation network
Gene co-expression network

ABSTRACT

Background: Selection of genes associated with disease plays an important role in understanding the disease pathogenesis and discovering the therapeutic targets. Network based approaches have been widely used for gene selection or functional module detection. Although microarray technology promotes the study of global changes in the gene expression, it is still not ready for clinical research because of lack of meticulous standards for data collection, analysis and validation. Integrating multiple omics data type can compensate for missing and unreliable information of any single data type and thus increases the reliability of newly discovered knowledge.

Objective: The work aims to design a novel dense subgraph based framework to select informative genes from the given gene expression data. The goal here is achieved by integrating the multiple types of biological networks and extracting densely connected components. The work focuses on identifying statistically significant and biologically enriched genes given a gene expression data.

Methods and result: The proposed framework identifies the modules of co-expressed genes and further integrates it with protein-protein interaction (PPI) network to identify modules of strongly interacting genes. The dense gene modules (DGM) embedded in the integrated co-expressed gene network are extracted representing tightly co-expressed gene modules. The proposed approach has been compared with the existing clustering based integration of expression data with PPI network. Besides achieving high prediction accuracy, our proposed approach underlines the robustness and consistency of identifying DGM and selects putative and functionally coherent genes.

Discussion: The co-expression based integration infers interesting biological information associated with the disease and the hub genes that are involved in many biological processes. Such a study reveals that integration of different biological networks for gene selection may provide greater insight in disease related biological process.

1. Introduction

In humans, despite the rapid increase in the discovery of disease associated genes, the molecular basis of many diseases is still unknown (Yang et al., 2014). Hence, selection of significant and differentially expressed genes associated with complex diseases becomes important in order to understand the underlying biological process and disease progression (Xu and Zhang, 2005; Ghosh et al., 2014; Swarnkar and Mitra, 2012). Microarray data is limited with noise present in data and small sample size, which gives an incomplete understanding and analysis of the disease (Trajkovski et al., 2008; He et al., 2014). The analysis of microarray data based on the existing knowledge of genetic networks provides in depth insight into molecular mechanism underlying the disease phenotypes (Swarnkar et al., 2015).

Gene co-expression network (GCN) is the network of genes showing similar expression pattern across samples. Modules in GCN corresponds to group of genes that have similar function and are involved in common biological processes that may cause many interactions among them (Weirauch, 2011). In comparison with other biological networks, GCN has nearly complete coverage of human genes and reduced bias due to knowledge obtained from published literature (Yang et al., 2014; Langfelder et al., 2008).

Restrictive single omics data type analysis leads to incomplete exploration of genetic aetiology of many complex traits and biological networks. Integrative analysis of multiple data types at different levels of genetic, genomic and proteomic regulation can compensate for missing or unreliable information in any single data type. Jin Li et al. constructed eQTL based gene-gene co-regulation network and further

Abbreviations: GCN, Gene Co-expression network; PPI, Protein-protein interaction; DGM, Dense Gene Module

* Corresponding author.
E-mail address: saswatimohapatra@soa.ac.in (S. Mahapatra).

https://doi.org/10.1016/j.genrep.2018.07.008
Received 16 February 2018; Received in revised form 25 June 2018; Accepted 13 July 2018
Available online 18 July 2018

integrated with PPI network and have used random walk with restart method for mining candidate genes for Alzheimer's disease (Li et al., 2015).

Diseases caused by functionally related genes are found to be phenotypically similar. Genes which are functionally related are co-expressed and also found to be closely connected in a known biological network (Barabási et al., 2011). PPI network represents physical binding events between protein pairs and can be used to uncover the biological mechanisms behind the genetic interactions. Dense gene modules are expected to be functionally associated and largely similar across independent data sets of same disease. A set of tightly connected modules in PPI network correspond to dense clusters with strong co-expression which are also involved in common biological process (Ma and Gao, 2012; Mitra et al., 2013). Hence, densely connected genes/proteins in PPI network corresponds to genes with strong co-expression and involved in a common biological process. Study of these densely connected components gives greater insight into molecular basis of the disease (Swarnkar et al., 2015).

In this study we have proposed a dense sub graph based approach to identify the differentially expressed genes by integrating different biological networks. The gene correlation information extracted from GCN and protein interaction network available in known public databases are integrated to identify densely connected subnetworks containing functionally enriched genes. The study compares the significant dense gene modules extracted from proposed approach with the prominent modules of the existing clustering based approach of gene selection from literature. The results are compared on the bases of i) Graph density, ii) Persistency of genes in the modules, iii) Classifier performance iv) Biological significant analysis.

2. Related work

Studies explored from literature reveals several integrative analyses of known gene networks with transcriptomics data to identify and prioritize the biomarker for drug targets, providing better biological interpretation and statistical analysis. Improved understanding of full spectrum of interaction can be gained by exploring multiple independent networks (Padi and Quackenbush, 2015). Interpolating pair wise gene association information in PPI helps to identify functionally coherent subnetworks that better generalizes the prediction of metastasis risk (Akker et al., 2011). Wie Kong et al. suggested the study of pathway analysis by integrating gene expression data and PPI network information to reconstruct inflammatory signaling pathway of known genes responsible for Alzheimer's disease (Kong et al., 2014).

A list of hypotheses on network medicine have been stated, that links topological properties of PPI network to biological functionalities (Swarnkar et al., 2015; Mitra et al., 2013; Park et al., 2014). Centrality analysis is one of the prevalent methods to identify important elements in a network, which focuses on highly interconnected nodes in the network. Hub genes in a genetic network represent a small proportion of nodes/genes showing many interactions with their neighboring genes that regulate the flow of information. The exhaustive analysis of co-expressed dense gene modules and hub genes plays a far reaching role in understanding the molecular mechanisms behind different biological functions. (Ji et al., 2014). The density or community structure representing a tight co-expression among their entities is expected to participate in common biological processes and may be further used for study of progression of a disease (Lee et al., 2010; Fortunato, 2010). Feng et al. (2011) suggested a top-down, max-flow-based approach to extract the dense gene modules by combining PPI network information with gene expression data. Study reveals, the existence of such dense gene modules containing highly interconnected nodes is based on factuality and certainly not a deviation in nature and human-planned network (Lee et al., 2010; Barabási et al., 2011). The hub molecules which form these densely connected modules exhibit diverse physiological functions and are likely to be putative mediators of disease. The existing technique (Swarnkar et al., 2015), propose a cluster based gene selection approach, integrating the gene expression data and PPI network information. Dense gene modules and dense hub modules are extracted by combining the clusters with the PPI network. Besides achieving a satisfactory predictive accuracy, the identified modules are also enriched in disease related genes which give greater biological insight into molecular mechanism underlying the complex diseases. Nevertheless, the first step of this approach relies solely on gene expression clustering analysis, without extracting gene expression network modules from that.

3. Methods and materials

3.1. Weighted gene correlation network analysis (WGCNA)

WGCNA (Padi and Quackenbush, 2015) is an established method designed for construction of gene networks which is based on marginal measure of linear associations (correlation patterns) among genes across microarray samples. Pearson correlation coefficient is used to measure the gene co-expression correlation. This method selects the threshold for constructing the network based on the scale free topology of the gene co-expression network (Langfelder and Horvath, 2008). In such a network highly interconnected genes are called hub genes, which are expected to play an important role in understanding the biological mechanism of many complex diseases. WGCNA is implemented as a package of R language.

3.2. Dense subgraph construction

The algorithm derived by Goldberg (1984) is being used to discover a dense subgraph in a given network. Density is defined as the ratio between number of edges and number of nodes in the unweighed graph. Goldberg's algorithm transforms the given graph into a new network and thus reduces the problem of finding the dense subgraph into a sequence of max flow/min cut problems. Further, fixing a threshold ‘g’ (guess), it finds the maximum density subgraph with density at least g. The algorithm has time complexity $O(|V||E|)$, where $|V|$ is the number of vertices and $|E|$ is the number of edges in the graph. So, this algorithm executes in linear time and gives the exact solution for finding the densest subgraph in given graph.

3.3. Proposed model

Our proposed approach for identification of dense gene modules goes through the following steps. Fig. 1 illustrates the steps of our proposed framework.

1) *Data Preprocessing*: the benchmark gene expression data set from NCBI is considered for the preprocessing step of our proposed model. The preprocessing of data includes cleaning of data by removal of genes with at least 30% of missing values across the samples (Swarnkar et al., 2015) and further replacing the missing values with the mean value of observed data. The genes are filtered based on their variances across diseased and normal samples producing 104 samples and 15,467 probes (Ma and Gao, 2012).
2) *Construction of GCN*: WGCNA functions are used for GCN construction and gene module formation. Hierarchical clustering is performed on normalized data set and sample outliers are removed, resulting 100 samples. GCN is constructed from resultant gene expression data. Scale free topology in the constructed network is obtained by selecting a power value of 4, which is the minimum power that resulted in a network with satisfactory scale independence according to WGCNA. The co-expressed gene modules are then formed by a robust dynamic tree cut algorithm (Langfelder et al., 2008) and adjacent modules are merged based on the parameter cut height of 0.25. By keeping minimum module size as 30,