Extracting Biological Significant Subnetworks from Protein-Protein Interactions Induced by Differentially Expressed Genes of HIV-1 Vpr Variants

Bandana Barman, Department of Electronics and Communication Engineering, Kalyani Government Engineering College, Kalyani, India
Anirban Mukhopadhyay, Department of Computer Science and Engineering, University of Kalyani, Kalyani, India

ABSTRACT

Identification of protein interaction network is very important to find the cell signaling pathway for a particular disease. The authors have found the differentially expressed genes between two sample groups of HIV-1. Samples are wild type HIV-1 Vpr and HIV-1 mutant Vpr. They did statistical t-test and found false discovery rate (FDR) to identify the genes increased in expression (up-regulated) or decreased in expression (down-regulated). In the test, the authors have computed q-values of test to identify minimum FDR which occurs. As a result they found 172 differentially expressed genes between their sample wild type HIV-1 Vpr and HIV-1 mutant Vpr, R80A. They found 68 up-regulated genes and 104 down-regulated genes. From the 172 differentially expressed genes the authors found protein-protein interaction network with string-db and then clustered (subnetworks) the PPI networks with cytoscape3.0. Lastly, the authors studied significance of subnetworks with performing gene ontology and also studied the KEGG pathway of those subnetworks.

Keywords: Differential Gene Expression, False Discovery Rate, Gene Ontology, KEGG Pathway, Microarray Data, Normalization, P-Value, Permutation Test, PPI Networks, Q-Value

INTRODUCTION

Protein protein interaction (PPIs) is the intentional physical contacts established between two or more proteins. It is a result of biochemical events and/or electrostatic forces. Proteins are macromolecules in both cellular and systemic levels but they cannot act alone. These interactions studied from different perspectives: biochemistry, quantum chemistry, molecular dynamics, signal transduction with others. The molecular processes within a cell are carried out by molecular
machines and that is built from a large number of protein components organized by their PPIs. These interactions are at the core of entire interactomics system of any living cell. So, aberrant PPIs are cause of multiple diseases, like Creutzfeld-Jacob, Alzheimer’s disease, and cancer.

The prediction of protein interaction is a field of combined bioinformatics and structural biology. It is important to understand the protein protein interactions to investigate the intracellular signaling pathways, to model protein complex structures and for gaining insights into various biochemical processes. The physical interactions between pairs of proteins can be found from a variety of experimental techniques. The experiments include yeast two-hybrid systems, protein fragment complementation assays (PCA), affinity purification/mass spectrometry, protein microarrays, fluorescence resonance energy transfer (FRET), and Microscale Thermophoresis (MST).

Clustering of PPI network is useful to isolate groups of interacting proteins which participate in same biological processes or perform specific biological functions together. To identify the functionally related proteins in protein protein interaction network is very important in the proteomic community. Clustering of PPI network becomes the focus of research as network structure reveals crucial information for nodes i.e. connectivity and information sharing. Clustering is done with concept of graph representations for protein interaction network. Let a network N, is there with V number of nodes or proteins and A number of arcs. So, the network N = (V, A). The nodes’ degree is number of arcs connected to that node. After clustering of network, the subnetwork is N’ = (V’, A’), where V’ ⊆ V and A’ ⊂ A. Network order is total number of nodes present in the network and network size is total number of edges that network contains. The network path is a sequence of arcs. Clustering coefficient is defined as the density of network and it is a dependent of nodes’ degree. Network clustering coefficient will be a/V C². Where, C, is the network cluster, a is number of arcs, V C² is the maximum number of arcs. Network cycle means closed path in which starting and ending nodes are same.

With finding ratios of expression levels of different samples differential expression is assessed. The variability of ratios is not a constant parameter. If a gene shows both statistical and biological significance that gene is considered to be differentially expressed between two sample groups. So many methods have been developed to study the microarray sample data and to detect differentially expressed genes (Dudoit et al., 2002, Buechler, 2007). Now a days, a various microarray experiments have been used to analyze multivariate microarray data (Bean, 2012). The quality of microarray datasets are too much essential to analyze and it is not so easy task. Before or after data processing, the quality verification or assessment may be done. The quality control improves detection of differentially expressed genes (Kaufmann, 2010). Research works are done for DEG combinations searching, for detection of DEG from GeneChip arrays (Hein, 2006), and also in networking of DEGs in human cancer (Selga, 2009). Statistical models to assess differential expression in Microarray Experiments (Smyth, 2004) is developed and using public databases identification of DEGs are improved (Kim, 2004). The statistical methods with comparing two Poisson means (rates) to detect differentially expressed genes for RNA-seq data (Chen, 2011) is introduced.

Differentially expressed genes are found with Non-parametric methods such as nonparametric t-test, Wilcoxon ranksum test and heuristic idealized discriminator method (Troyanskaya, 2002). Statistical methods to detect differentially expressed genes in replicated cDNA microarray experiments (Dudoit et al, 2000) are developed. First microarray data are normalized to scale the data, so that the mean and variance of data becomes 0 and 1 respectively. For multiple testing, correction is made by adjusted pvalue in univariate testing problem in identified differentially expressed genes. To find differentially expressed genes, the principal component analysis (PCA) in space and classification of genes between two conditions are done (Ospina, 2013). Characterization in cancer pathways (Li et al, 2015) and also meta-analysis of DEGs (Yang. et al, 2014)
The Residence Time of the Water in Lake MAGGIORE. Through an Eulerian-Lagrangian Approach
www.igi-global.com/chapter/residence-time-water-lake-maggiore/69464?camid=4v1a