Chapter 4
Large-Scale Regulatory Network Analysis from Microarray Data: Application to Seed Biology

Anamika Basu
Gurudas College, India

Anasua Sarkar
SMIEEE Government College of Engineering and Leather Technology, India

ABSTRACT

The inference of gene networks from gene expression data is known as "reverse engineering." Elucidating genetic networks from high-throughput microarray data in seed maturation and embryo formation in plants is crucial for storage and production of cereals for human beings. Delayed seed maturation and abnormal embryo formation during storage of cereal crops degrade the quality and quantity of food grains. In this chapter, the authors perform comparative gene analysis of results of different microarray experiments in different stages of embryogenesis in Arabidopsis thaliana, and to reconstruct Gene Networks (GNs) related to various stages of plant seed maturation using reverse engineering technique. They also biologically validate the results for developing embryogenesis network on Arabidopsis thaliana with GO and pathway enrichment analysis. The biological analysis shows that different genes are over-expressed during embryogenesis related with several KEGG metabolic pathways. The large-scale microarray datasets of Arabidopsis thaliana for these genes involved in embryogenesis have been analysed in seed biology. The chapter also reveals new insight into the gene functional modules obtained from the Arabidopsis gene correlation networks in this dataset.

INTRODUCTION

Recent advances in microarray technologies have made it possible to routinely measure the expression levels of tens or even hundreds of thousands of genes simultaneously. Such high-throughput experimental data have initiated much recent research on large-scale gene expression data analysis. Various data mining techniques (e.g., clustering and classification) have been employed to uncover the biological functions of genes from microarray data. Recently, these techniques have included a

DOI: 10.4018/978-1-4666-6611-5.ch004

reverse engineering approach to extracting gene regulatory networks from microarray data in order to reveal the structure of the transcriptional gene regulation processes.

The general purpose of gene regulatory network analysis is to extract pronounced gene regulatory features (e.g., activation and inhibition) by examining gene expression patterns. Changes of expression levels of genes across different samples provide information that allows reverse engineering techniques to construct the network of regulatory relations among those genes. Many studies have shown that these learned networks have the potential to help researchers propose and evaluate new hypotheses in basic research of genetic regulatory process. From various issues like noise, incompleteness, multiple cluster belongingness etc., several different trends evolve in research fields over networks. To explore different approaches with or without using prior biological knowledge, we have compared several existing approaches to find out the probable transcriptional gene network for seed maturation in *Arabidopsis thaliana.*

For multicellular organisms for systemic and better understanding of their gene regulation involved in a specific developmental stage of life cycle, Gene Network [GN] is the best way. In plant after fertilization under suitable environmental condition embryogenesis occurs. Not only for new plant generation but also for the food habit of world population, construction of a GN for seed maturation and embryogenesis of our food crops e.g. rice, maize is necessary. Here we use model plant *Arabidopsis thaliana* to identify the gene regulators for different stages of embryogenesis. For construction of GRN for different species with huge microarray datasets for seed development and maturation we use that reverse engineering as a powerful method.

BACKGROUND

Network Analysis on Microarray Data

The network and co-expression analysis of microarray data finds out probable transcriptional gene interactions and co-regulation. Network analyses typically work with the gene–gene co-expression matrix, based on the correlation between each pair of genes in the dataset. Hypothetically the amount of co-expression between two genes is associated with an increasing probability that these two genes interact. Thus, the co-expression matrix infers the networks of interactions (Pavlopoulos, 2008).

A network is the same as a mathematical concept of a graph, denoted as a pair G= {V, E}, where V is a set of vertices (nodes) and E is a set of links (edges) that connect pairs of nodes. When constructing a network, one needs to know how biological elements are related among themselves to represent nodes and edges. For example, an edge in a gene interaction network can be placed between two genes if they are functionally associated. This results in a gene interaction network. In a metabolic network, nodes correspond to metabolites and enzymes and directed edges are metabolic reactions to combine all metabolic pathways possible within a cell. Gene regulatory interactions builds a transcriptional regulatory network. Quality of data is the most limiting factor for the network analyses, because of noisy and incomplete biological data covering a significant part in public databases.

There are a number of tools available for visualizing and analysing the biological network from a microarray data set (Pavlopoulos, 2008),(Thomas, S., 2010). Table 1 shows a comparison of available tools online for analysing biological networks.

Several different methods have been used to construct networks from microarray data. Pearson