Chapter 10
Pathway Analysis and Its Applications

Ravi Mathur
North Carolina State University, USA

Alison Motsinger-Reif
North Carolina State University, USA

ABSTRACT

As the scale of genetic, genomic, metabolomics, and proteomic data increases with advancing technology, new approaches leveraging domain expert knowledge, and other sources of functional annotation have been developed to aid in the analysis and interpretation of such data. Pathway and network analysis approaches have become popular in association analysis – connecting genetic markers or measures of gene product with phenotypes or diseases of interest. These approaches aim to leverage big data to better understand the complex etiologies of these traits. Findings from such analyses can help reveal interesting biological traits and/or help identify potential biomarkers of disease. In the current chapter, the authors review broad categories of pathway analyses and review advantages and disadvantages of each. They discuss both the analytical methods to detect phenotype-associated pathways and review the key resources in the field of human genetics that are available to investigators wanting to perform such analyses.

INTRODUCTION

Recent technological developments in high-throughput genetic, genomic, and metabolomics profiling techniques has greatly expanded the potential for more systems level analysis. As such data becomes more readily available experimentally, and the scale of the data increases, this creates exciting new challenges in analyzing this "big data." Handling and summarizing such high dimensional data in an efficient and interpretable way is crucial to making good use of this data. While there are a number of different strategies for handling big "-omics" data, pathway and network analysis approaches are becoming standard approaches for discovering and summarizing the underlying relationships in the data. The pathway and network approaches rely on either external knowledge bases or strong correlation structure in the data to collapse the data from thousands or millions of variables, to hundreds or thousands of pathways/networks for analysis. The results

DOI: 10.4018/978-1-4666-6611-5.ch010

of such analyses are valuable to the process of discovering the underlying mechanism of disease or a phenotype of interest, including the events leading up to initiation, progression, and treatment of a disease.

As stated by the National Human Genome Research Institute (NHGRI), "A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell" (www.genome.org). Such a pathway can describe the function of molecules in a cell (regulatory pathway) or the change in chemical elements throughout the cell (metabolomics pathway) or a description of the initiation of a disease (disease pathway). It is now understood that biological pathways in the cell interact with one another to carry on the actions of the cell. Therefore, a group of interacting pathways comprise a biological network. There is a wealth of knowledge on such interactions that have been curated in pathway/knowledge bases that can be leveraged in statistical analysis of specific datasets. Many pathway based analysis approaches have been developed to use these databases to aid in gene function prediction, discover new associations with the trait of interest, and even to better classify patients or sample. Other approaches, more typically referred to as network approaches, focus on quantifying the connections between the gene, proteins, or metabolites to better understand the connections between the molecules that result in the phenotype/disease. A biological pathway and network can be displayed and analyzed in a graph form with vertices and edges. In such a form a vertex represents each element contained in the pathway or network and an edge represents an interaction (activation, repression, methylation, series of chemical reactions, etc.) between those elements. Figure 1 shows a graph representation of the Glycolysis Pathway, which displays many of the properties (scale-free degree distribution, high clustering coefficient, and characteristic path length) displayed in a variety of biological pathways. In the current chapter, we review many of the major categories of pathway and network analysis tools.

Most of the pathway analysis tools developed rely on a knowledge base that defines gene sets or pathways that are statistically tested to be influential in the dataset as opposed to random chance. While the fundamental statistical approaches of the methods are largely independent of the database, in practicality many of the knowledge based and analysis approaches have been developed in tandem. Therefore the software implementations that are popularly used rely heavily on specific databases. The analysis of pathway starts by the data collection, thus we begin the chapter with a short discussion of the data collection technologies that are frequently used in human genetics. Many of these resources are also available for model organisms, but we focus on tools with direct relevance for research directly studying human health. While many of the approaches were originally developed to gene expression (particularly microarray) data, extensions of the approaches are now frequently used with deoxyribose nucleic acid (DNA) level variation, and with biochemical/metabolomics data. Due to the reliance on knowledge bases, we continue with a discussion of the most commonly used resources in human genetics. Then we will discuss the major classes/categories of gene set and pathway analysis tools, and discuss details of some of the most commonly used methods. Many of the implementation challenges will be discussed, with an emphasis on analyzing big data where often the number of samples is much lower than the number of variables or features. The chapter will conclude by discussing future research avenues and overall lessons learned from these methods so far.

BACKGROUND

Pathway analysis consists of the statistical analysis of a series of steps, all of which influence each other. The basic workflow of any pathway/network