Chapter 10
Knowledge-Driven, Data-Assisted Integrative Pathway Analytics

Padmalatha S. Reddy
Pfizer, USA

Stuart Murray
Agios Pharmaceuticals Inc, USA

Wei Liu
Agios Pharmaceuticals Inc, USA

ABSTRACT

Target and biomarker selection in drug discovery relies extensively on the use of various genomics platforms. These technologies generate large amounts of data that can be used to gain novel insights in biology. There is a strong need to mine these information-rich datasets in an effective and efficient manner. Pathway and network based approaches have become an increasingly important methodology to mine bioinformatics datasets derived from 'omics' technologies. These approaches also find use in exploring the unknown biology of a disease or functional process. This chapter provides an overview of pathway databases and network tools, network architecture, text mining and existing methods used in knowledge-driven data analysis. It shows examples of how these databases and tools can be used integratively to apply existing knowledge and network-based approach in data analytics.

INTRODUCTION

Target and Biomarker Selection in Drug discovery

A critical step in the drug discovery process is the effective selection of candidate molecular targets. Target identification and selection requires a thorough understanding of the cellular role of the target, the signaling and metabolic pathways it is involved in, and the network of interactions that are involved in the functional role of the target. Perturbations in one or more of these may be responsible for a disease state or an off-target effect during drug treatment. Companies must deploy effective methods to select the targets since

DOI: 10.4018/978-1-60960-491-2.ch010

the drug discovery and development process is expensive and time-consuming. Furthermore, it is essential to fully understand the target and disease pathways to minimize expensive late-stage failures and to successfully translate animal models into the clinical development of therapies. With the advent of high-throughput 'omics' technologies and the rise of informatics technologies, it has become possible to routinely and systematically explore targets and disease-related cellular pathways, as well as cross-talk between pathways and interaction networks. Thus, a rational pathway and network based approach for target and biomarker identification has begun to be adopted by pharmaceutical companies in the recent years.

Biological Networks and Their Descriptors

Cellular functions are carried out through a complex network of interactions between biomolecules (genes, transcripts, proteins, metabolites, miRNAs, etc.). The various interactions can be biochemical or physical, and the interconnected assembly of 'cellular machinery' can be effectively presented as an "interactome" or "network" that enables visualization of molecular relationships and the logic of their function. The topology and dynamics of these complex networks can be readily studied by graph theory. The terms "interactome", "graph" and "network" have been used interchangeably. However, "interactome" and "network" describe the physical or biological system, whereas "graph" denotes the mathematical object representing the topology of the system. Topological analysis of the networks provides information about the networks, and these are described by the following parameters (Zhu, Gerstein, & Snyder, 2007) (i) **Degree:** The number of edges connected to one node is defined as its degree. In directed networks these can be further subdivided into incoming degree, outgoing degree and total degree. A node with high degree is well connected (also called "hubs") and may play a role in maintaining network structure. Thus the number of interactions a node has positively correlates with its importance in the network. Hub nodes that represent essential genes/proteins are generally conserved in evolution (Barabasi & Oltvai, 2004). (ii) **Clustering coefficient:** The ratio of the actual number of links between a node's neighbors and the maximum possible number of links between them. A high clustering coefficient for a network indicates a small world network. (iii) **Shortest path:** For any pair of nodes, the minimum number of network edges that need to be traversed to travel from one node to another. (iv) **Characteristic path length:** The average length of "shortest paths" for all pairs of nodes. (v) **Diameter:** The maximum distance between any two nodes. The average shortest path length and diameter of a network measure the approximate distance between nodes in a network. A network with a small diameter is termed "small world", in which any two nodes can be connected with relatively short paths. (vi) **Betweenness:** The fraction of the shortest paths between all pairs of nodes that pass through one node or edge, and provides an estimate of the information flow through one node or edge. It is a better indicator for the essentiality of a gene than degree centrality (Han, 2008). (vii) **Eigenvector centrality:** A measure of the contribution of degree centrality by its neighbors. (viii) **Closeness centrality:** A measure of the centrality of a node based on how close it is to other nodes in a network.

It is generally hypothesized that perturbation of one or more nodes (gene or protein) in a network disrupts cellular pathways, functions and cellular processes, giving rise to various disease conditions. There are over 6000 human diseases caused by a defect in a single gene (McKusick, 2007). In these disorders, single gene defects are sufficient to perturb the network, resulting in the disruption of normal cellular, tissue and organ functions. A recent study demonstrated that disease causing alleles that result in truncated proteins lead to the removal of nodes from networks. Disease alleles