Available online at www.sciencedirect.com

ScienceDirect

Journal of Genetics and Genomics 43 (2016) 349−367

JGG

REVIEW

Integrated Genomic and Network-Based Analyses of Complex Diseases and Human Disease Network

Olfat Al-Harazi [a], Sadiq Al Insaif [a], Monirah A. Al-Ajlan [a,c], Namik Kaya [b], Nduna Dzimiri [b], Dilek Colak [a,*]

[a] *Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Saudi Arabia*
[b] *Department of Genetics, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Saudi Arabia*
[c] *College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia*

Received 11 June 2015; revised 22 October 2015; accepted 20 November 2015
Available online 15 December 2015

ABSTRACT

A disease phenotype generally reflects various pathobiological processes that interact in a complex network. The highly interconnected nature of the human protein interaction network (interactome) indicates that, at the molecular level, it is difficult to consider diseases as being independent of one another. Recently, genome-wide molecular measurements, data mining and bioinformatics approaches have provided the means to explore human diseases from a molecular basis. The exploration of diseases and a system of disease relationships based on the integration of genome-wide molecular data with the human interactome could offer a powerful perspective for understanding the molecular architecture of diseases. Recently, subnetwork markers have proven to be more robust and reliable than individual biomarker genes selected based on gene expression profiles alone, and achieve higher accuracy in disease classification. We have applied one of these methodologies to idiopathic dilated cardiomyopathy (IDCM) data that we have generated using a microarray and identified significant subnetworks associated with the disease. In this paper, we review the recent endeavours in this direction, and summarize the existing methodologies and computational tools for network-based analysis of complex diseases and molecular relationships among apparently different disorders and human disease network. We also discuss the future research trends and topics of this promising field.

KEYWORDS: Protein−protein interaction; Subnetwork marker; Diseasome; Network medicine; Systems biology; Genomics

INTRODUCTION

"*It is the last lesson of modern science that the highest simplicity of structure is produced, not by few elements, but by the highest complexity*", Ralph Waldo Emerson, 1850. Traditionally, human disease classifications have been based on crude observational correlations between pathological analysis and clinical syndromes (Loscalzo et al., 2007). However, this classification suffers from lack of sensitivity and specificity in detecting diseases before known symptoms and ambiguity in clinical diagnosis become evident (Loscalzo et al., 2007; Butte, 2008). Thus, for example, while a disease can be caused by mutations in various genes, the clinical, anatomical and functional pathophenotypes could be essentially indistinguishable from one another (Ho and Seidman, 2006; Colak et al., 2011). With recent advances in high-throughput molecular assay technologies, a growing body of transcriptomic, proteomic, and metabolomic datasets has been generated. This has enabled the defining of human diseases with relatively higher sensitivity and specificity than the traditional approaches. Indeed, the genomics, transcriptional profiling, or proteomic datasets have led to uncovering of biomarkers for subcategorizing histologically similar diseases, and provided better information about diagnosis, prognosis and response to

* Corresponding author. Tel: +966 1 464 7272x39211, fax: +966 1 442 4542.
E-mail address: dcolakkaya@kfshrc.edu.sa (D. Colak).

http://dx.doi.org/10.1016/j.jgg.2015.11.002

therapy for several human malignancies (Hedenfalk et al., 2001; van't Veer et al., 2002; Dave, 2006; Dave et al., 2006; Finak et al., 2008; Colak et al., 2009, 2010, 2011).

Proteins do not function independently in the cell. Thus, a cellular phenotype is the result of proteins interacting with each other and with other molecules both inside and on the cell surface. Comprehensive protein-interaction mapping projects are currently underway for many model species and humans. During the last decade, significant efforts have been invested into annotating and cataloguing individual interactions between proteins and protein−protein interactions (PPI) into databases for generating large networks of graphical interactions (Bader, 2003; Ideker and Sharan, 2008). One such example is the human protein interaction network (interactome) which has marked an important milestone for biomarker discovery in cancer (Chuang et al., 2007; Taylor et al., 2009). Such integration of all cellular information into a functional model should enhance the robustness and reliability of disease markers. Besides, the availability of high-throughput datasets, such as mRNA expression profiles, proteomics data and protein interaction networks will further advance our current understanding of diseases. Indeed, several reports have shown the effectiveness of such network-based approaches to identifying disease markers in various disorders (Chuang et al., 2007; Ergun et al., 2007; Taylor et al., 2009). For example, Chuang et al. (2007) integrated gene expression profiles with interaction networks to predict breast cancer metastasis. The authors discovered that subnetwork markers are more reproducible than individual marker genes selected without network information, and that they achieve higher accuracy in the classification of metastatic *versus* non-metastatic breast tumours. The network of diseases, or diseasome, is a disease−disease interaction network, which has emerged to investigate the disease relationships. Thereby, the nodes are diseases/disorders and links represent various molecular relationships between diseases, such as shared familiar disease genes from the Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2015) or metabolites (Goh et al., 2007; Lee et al., 2008), or based on the directly observed comorbidity between them (Rzhetsky et al., 2007; Hidalgo et al., 2009).

However, most of the existing approaches are limited to employing single molecular data, such as mRNA expression, which captures only a restricted measure of cellular activity. Integrating gene expression with other types of biomolecular datasets, such as the copy number variation (CNV) data, would reveal the chromosomal regions with concordantly altered genomic and transcriptional status in disease (Pollack et al., 2002; Garraway et al., 2005; Patil et al., 2005). These integrated approaches have been demonstrated to be useful in identifying the likely drivers of cancer, for example (Akavia et al., 2010; Colak et al., 2010). Hence, inclusion of multiple high-throughput molecular data (proteomics, CNV, and others) in the search for subnetwork markers has the potential to find better and more robust disease markers (Joyce and Palsson, 2006; Colak et al., 2010, 2011; Nibbe et al., 2010).

In this review, we provide a comprehensive overview of the recent literature on network-based studies of complex human diseases and networks (diseasome), as well as existing methodologies and computational tools developed thus far for identifying genes and subnetwork markers for various disorders. We first define the interactome, and provide the related databases in the subsequent section. This is followed by the review of the network-based methods developed for prioritizing disease genes as well as data integration, and a discussion on the subnetwork or gene module identification approaches and the developed tools. The application of network-based approaches to identify disease modules/subnetworks is then reviewed. To this extent, we also present an application of one of these subnetwork methodologies to the microarray gene expression data that we have generated at our institution for identifying significant subnetworks associated with idiopathic dilated cardiomyopathy (IDCM). Finally, we discuss the human disease network "diseasome" and the related methodologies for creating the diseasome, followed by some concluding remarks.

THE HUMAN PROTEIN−PROTEIN INTERACTION NETWORK: "INTERACTOME"

To date, comprehensive protein-interaction mapping efforts have yielded experimentally validated or expert-annotated interactions in many model species and humans. High-throughput methods have accelerated the generation of PPI data on a large scale. It is estimated that more than 100,000 protein interactions exist in the human body (Bonetta, 2010). Significant efforts have been invested into annotating and cataloguing the PPIs through database analyses that have been utilized to construct large network graphs of interactions (Bader, 2003; Ideker and Sharan, 2008). In the PPI network maps, nodes represent proteins and edges represent interactions between the proteins. Recently, the human protein interaction network (interactome) has been increasingly utilized to unravel the molecular basis of diseases (Chuang et al., 2007; Taylor et al., 2009). Together with other studies, these endeavours demonstrate that in disease, co-expressed genes tend to cluster non-randomly in certain 'hotspots', and when mapped to the interactome, they reveal well-connected sets of proteins (modules), thereby enhancing the exploitation of both the functional and topological modularity of the network. A number of databases assemble various types of information about protein−protein interactions (Table 1). These include the Biological General Repository for Interaction Datasets (BioGRID) (Chatr-Aryamontri et al., 2015), Human Protein Reference Database (HPRD) (Prasad et al., 2009), Molecular INTeraction database (MINT) (Licata et al., 2012), Online Predicted Human Interaction Database (OPHID) (Brown and Jurisica, 2005), Molecular Interaction Database (IntAct) (Orchard et al., 2014), Database of Interacting Proteins (DIP) (Salwinski et al., 2004), Biomolecular Interaction Network Database (BIND) (Isserlin et al., 2011), Search Tool for Retrieval of Interacting Genes/Proteins (STRING) (Franceschini et al., 2013), Mammalian Protein−Protein