Gene 590 (2016) 68–78

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Network analysis of genes and their association with diseases

Panagiota I. Kontou [a,1], Athanasia Pavlopoulou [a,1], Niki L. Dimou [a], Georgios A. Pavlopoulos [b], Pantelis G. Bagos [a,*]

[a] Department of Computer Science and Biomedical Informatics, University of Thessaly, Greece
[b] Lawrence Berkeley Lab, Joint Genome Institute, United States Department of Energy, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form 20 May 2016
Accepted 30 May 2016
Available online 2 June 2016

Keywords:
Human disease network
Human disease-related genes network
Gene-disease associations
Drug design

ABSTRACT

A plethora of network-based approaches within the Systems Biology universe have been applied, to date, to investigate the underlying molecular mechanisms of various human diseases. In the present study, we perform a bipartite, topological and clustering graph analysis in order to gain a better understanding of the relationships between human genetic diseases and the relationships between the genes that are implicated in them. For this purpose, disease-disease and gene-gene networks were constructed from combined gene-disease association networks. The latter, were created by collecting and integrating data from three diverse resources, each one with different content covering from rare monogenic disorders to common complex diseases. This data pluralism enabled us to uncover important associations between diseases with unrelated phenotypic manifestations but with common genetic origin. For our analysis, the topological attributes and the functional implications of the individual networks were taken into account and are shortly discussed. We believe that some observations of this study could advance our understanding regarding the etiology of a disease with distinct pathological manifestations, and simultaneously provide the springboard for the development of preventive and therapeutic strategies and its underlying genetic mechanisms.

1. Introduction

The majority of human genetic disorders are rarely attributed to the activity of a single gene, but rather to the combinatorial activity of more than one gene (Cordell and Clayton, 2005; Goldstein, 2009). Disease-associated genes have been traditionally studied using family-based linkage methods (Oti et al., 2006). However, during the last decade, the deep-sequencing technologies allowed for the sequencing of the entire human genome and the subsequent identification of millions of single nucleotide polymorphisms (SNPs), which can be spotted in a single scan in a timely affordable manner within the genome-wide association studies (GWAS) framework (Ajay et al., 2011; Meyerson et al., 2010). This way, a large number of genes and genetic variations that contribute to disorders (Hirschhorn, 2009; Manolio, 2010), such as various cancers (Ioannidis et al., 2010) or common diseases (Wellcome Trust Case Control Consortium, 2007, 2010), have been detected.

Network-based approaches for the discovery of gene-disease associations have enabled biomedical researchers to not only investigate the genetic complexity of a particular disease, but also the relatedness among apparently discrete disease phenotypes (Barabasi et al., 2011; Pawson and Linding, 2008). Disease networks can provide the foundation for predicting causative genes, unravelling a disease's molecular mechanisms and designing new therapeutic strategies (Barabasi et al., 2011; Pawson and Linding, 2008). Genes associated with similar disease phenotypes have a higher propensity to physically interact with each other, forming distinct disease-specific functional modules (Hartwell et al., 1999; Oti and Brunner, 2007). Conversely, compared to diseases with different phenotypes, diseases with similar phenotypes have an increased tendency to share genes (Goh et al., 2007).

In the present study, we facilitate network-based approaches in order to explore the associations between the human genetic diseases and the relations between their effector genes. In a previous considerable work, Goh et al. (2007), created such a network based on disease-gene associations obtained from the OMIM database (Amberger et al., 2015). However, although today OMIM is one of the major repositories holding genetic association data for Mendelian diseases, it mainly archives rare disorders of high penetrance (Amberger et al., 2015). This parameter is of importance, since multigenic diseases of low penetrance may have different properties which have to be taken into account. Other studies, such as the ones conducted by Barrenas et al. (2009) and Liu et al. (2014), partially overcome this issue by integrating disease-gene association data from multiple resources. Of importance, in the study conducted by Goh et al., neither the disease concepts not the gene terms were standardized. However, in the studies conducted by Barrenas et al. (2009) and Liu et al. (2014), an effort was made to homogenize the disease concepts but not the gene terms.

[*] Corresponding author at: Department of Computer Science and Biomedical Informatics, University of Thessaly, Papasiopoulou 2-4, Lamia 35100, Greece.
E-mail address: pbagos@compgen.org (P.G. Bagos).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.gene.2016.05.044
0378-1119/© 2016 Elsevier B.V. All rights reserved.

The present study was based on raw data from three primary resources containing information for gene-disease associations. These are: *i*) The Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2015). It is maintained by NCBI and is a curated knowledgebase of human genes and genetic phenotypes. OMIM's content is extracted from the published biomedical literature. *ii*) NIH's Genetic Association Database (GAD) (Becker et al., 2004). It comprises complex genetically associated non-Mendelian disorders and polymorphisms that affect gene functions which are implicated in these disorders. The information in GAD originates from published papers of genetic association studies which mostly target multigenic diseases of low penetrance. *iii*) The National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies (GWAS) (Welter et al., 2014). It includes a manually curated collection of published GWAS, with more than 100,000 assayed SNPs and SNP-disease associations. GAD and GWAS are, therefore, complementary to OMIM since they contain up-to-date information for multigenic diseases of low penetrance, and to some extent information for low frequency variants (Becker et al., 2004; Welter et al., 2014).

Bipartite gene-disease networks and projected monopartite gene-gene and disease-disease networks were constructed based both on data from each of these three repositories individually and data combined from all three of them. Notably, prior to our network building, we used a standardized nomenclature across the three repositories for both the disease and gene terms in order to avoid edge and node redundancies. Our analysis enabled us to identify important topological properties of the biological networks and uncover noticeable disease-disease and gene-gene associations. In particular, based on the topological analysis of networks, we provide numeric evidence on the assumption that many genes can be causative for a human disease. Moreover, we suggest that phenotypically unrelated diseases may share a common genetic background, contrary to previous reports (Bauer-Mehren et al., 2011; Goh et al., 2007), which propose that functional modules exist for diseases of similar pathophenotypes. Of particular importance, unexpected disease-disease associations were uncovered when data from the three different sources were combined further, supporting previous hypotheses (Bauer-Mehren et al., 2011) that the integration process offers a great advantage in the identification of novel associations. Finally, we present a case study where we show how a gene-gene network can be used to find associations between genes which encode proteins of similar tertiary structure.

2. Methods

2.1. Data collection

Disease-gene association data were collected from three different publicly available, comprehensive databases as shown below. These are:

i. The NCBI's OMIM (Online Mendelian Inheritance in Man) (Amberger et al., 2015) which provides information about genetically inherited diseases. The ‘genemap2.txt’ files were downloaded from ftp://ftp.ncbi.nih.gov/gene/DATA/mim2gene_medgen (August 17th, 2013) and parsed to acquire disease-gene associations of the type 3. A spreadsheet was created (Text S1) in which, the first column contains the disease names while the second column contains the disease-related gene names. In the case a gene name was not available, its corresponding genomic region was included instead.

ii. NIH's GAD (Becker et al., 2004) covers mainly multigenic diseases of low penetrance. All GAD data were downloaded from http://geneticassociationdb.nih.gov/data.zip (August 17th, 2013). The “all.txt” file was searched for disease-gene associations under the heading “Association”. Due to the vast amount of gene-disease association data available in GAD (approximately 167,132 entries) and the fact that genetic association studies are characterized by their non-replicability (Ioannidis, 2005; Ioannidis et al., 2001), only the positive associations resulted from meta-analyses were collected. In this way, we ensured that only the statistically significant and replicated associations are included in the analysis. Based on these data, a spreadsheet was created (Text S2) with three columns containing the disease name, the gene name and the genetic polymorphism which confers the particular disease. Like before, in the absence of a gene name, its chromosomal locus was included.

iii. The NHRI GWAS Catalog (Welter et al., 2014) focuses on SNPs and their association(s) to diseases. The collected dataset is similar to GAD, but the GWAS framework ensures that most of biases in genetic association studies will be negligible since GWAS operates in a rather agnostic manner and employs strict criteria for statistical significance. The catalog was downloaded from http://www.genome.gov/gwastudies (August 23rd, 2013). While only the genic SNPs (both intronic and exonic) were included, the intergenic and the possible intergenic SNPs were omitted. Subsequently, a spreadsheet (Text S3) was created including columns with the disease name, the gene name and the SNP.

The spreadsheets derived from the three individual databases (Texts S1–S3) were merged, and subsequently, a JOINT file was generated (Text S4). Duplicated entries were removed from all four files (Texts S1–S4).

2.1.1. Collection of transmembrane protein identifiers

UniProt Knowledgebase (UiprotKB) (http://www.uniprot.org/) (Poux et al., 2014) is the central repository holding curated information about a protein's name, description, its amino acid sequence, function and other annotations. A total of 5061 human transmembrane (TM) protein identifiers were extracted from UniProtKB, release 2014_05 (Text S5).

2.2. Disease and gene nomenclature

Disease name heterogeneity and ambiguity in all the three repositories would not allow for a direct data comparison across the three databases. In order to maintain a consistent nomenclature and classification for diseases in our analysis, the naming conventions described in the International Classification of Diseases (ICD) were used. ICD (http://www.who.int/classifications/icd/en/) is maintained by the World Health Organization (WHO) and is a classification system to group diseases into major categories. Prior to our network analysis, each disease term in our three datasets was searched against the ICD for synonyms. In order to include a wide range of diseases in our study, in the case where a related synonym for a specific disease was not found in ICD, the broader disease category to which the disease belonged to was instead chosen. In the worse case scenario that a disease lacked both a synonym and category match in ICD (a total of 112 diseases), it was excluded from our study. To eliminate the generic ICD concepts, we only considered ICD terms at the second depth level in the hierarchy. Finally, in order to maintain a uniform nomenclature across all datasets, all genes from our three databases along with the ones from UniProtKB were converted to the official HGNC (HUGO Gene Nomenclature Committee) (Gray et al., 2015) gene symbols.

2.3. Network analysis, clustering and visualization

Cytoscape v.3.2.1 (http://www.cytoscape.org/) (Shannon et al., 2003), an open source network visualization platform, was employed for the statistical analysis, processing, and visualization of networks. Gene-disease associations were presented as bipartite graphs, where two groups of distinct nodes (genes and diseases) are connected through edges. Notably, in the original bipartite graph, genes were connected to diseases and there were no connections between nodes of the same type (gene-gene, disease, disease). Therefore, the two projected