Computers in Biology and Medicine 76 (2016) 173–177

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Cancerome: A hidden informative subnetwork of the diseasome

Mahdi Jalili [a], Ali Salehzadeh-Yazdi [a,b], Marjan Yaghmaie [a], Ardeshir Ghavamzadeh [a], Kamran Alimoghaddam [a,*]

[a] *Hematology, Oncology and SCT Research Center, Tehran University of Medical Sciences, Tehran, Iran*
[b] *Department of Systems Biology and Bioinformatics, University of Rostock, 18051 Rostock, Germany*

ARTICLE INFO

Article history:
Received 8 May 2016
Received in revised form
30 June 2016
Accepted 18 July 2016

Keywords:
Cancerome
Diseasome
Integrated human disease network
Disease–disease network
Disease similarity network

ABSTRACT

Neoplastic disorders are a leading cause of mortality and morbidity worldwide. Studying the relationships between different cancers using high throughput-generated data may elucidate undisclosed aspects of cancer etiology, diagnosis, and treatment. Several studies have described relationships between different diseases based on genes, proteins, pathways, gene ontology, comorbidity, symptoms, and other features. In this study, we first constructed an integrated human disease network based on nine different biological aspects, including molecular, functional, and clinical features. Next, we extracted the cancerome as a cancer-related subnetwork. Further investigation of cancerome could reveal hidden mechanisms of cancer and could be useful in developing new diagnostic tests and effective new drugs.

1. Introduction

Recently advances in our knowledge of the genetic and molecular features of human diseases have improved our understanding of the basis of these diseases and have enabled us to categorize diseases based on underlying molecular and genetic mechanisms and uncover relationships among them. We are now in the post-genomic era and a large amount of genome-wide association, transcriptomic, proteomic, and metabolomic data has been assembled. Now it is time to put together the puzzle pieces using a new approach, systems biology, to investigate human biology and diseases as a whole [1,2]. A holistic analysis of relationships among human diseases has provided novel insights into disease etiology and begun a new era in drug development.

Previous studies have investigated the relationships among diseases using several different approaches. One of the most commonly used approaches is linking diseases through gene–disease relationships [3,4]. Goh et al. [3] first introduced the term “diseasome” and constructed a human disease network by connecting two diseases if they shared at least one causative gene in the Online Mendelian Inheritance in Man (OMIM) database [5]. They found that disease similarities reflect functional modules among human disease genetics and correlate positively with several gene functions. Genome-wide association studies (GWAS) are the most reliable methods for uncovering causal relations between genes and diseases. Some studies have used GWAS results to construct disease–disease network [6,7], which have revealed hidden relationships between dissimilar diseases. Another approach is analyzing semantic similarities extracted from disease ontology (DO) and gene ontology (GO) trees [8,9]. In protein–protein interaction (PPI) networks if two diseases are similar, then the disease genes associated with them may be close to each other. Using this logic, Suratanee et al. [10] analyzed a PPI network and discovered relationships among different diseases. It is well-known that several cellular functions are carried out through the interaction of a collection of different genes, which are known as modules, complexes, or (most commonly) pathways. Some studies have used disease involved pathways to construct human disease–disease networks [11–13] and uncover novel disease relationships, which could lead to new treatment options. The human phenome, which is the collection of phenotypes that are expressed by human genes [14], could be used to predict candidate disease-associated genes and functional relationships between genes and proteins and has been used to construct a disease network [15]. Other approach, which is based on disease signs and symptoms, led to the development of a Human Symptoms Disease Network (HSDN) [16] and a human disease network [17]. These networks assesse interactions of clinical features and the underlying molecular mechanisms of diseases, which could help to identify the etiology of diseases, as well as new drugs. Finally, some studies have integrated different concepts and/or databases to create a more complete disease similarity network [18–21]. These integrative approaches could lead to more effective identification of disease relationships.

* Correspondence to: Shariati Hospital, Karegar Ave. P.C. 14114, Tehran, Iran.
E-mail address: alimgh@tums.ac.ir (K. Alimoghaddam).

http://dx.doi.org/10.1016/j.compbiomed.2016.07.010
0010-4825/© 2016 Published by Elsevier Ltd.

Cancer is a major public health problem and a leading cause of death worldwide. Despite decreases in the incidence and mortality rates of some cancers, death rates for cancers such as liver, pancreas, and uterine corpus cancer are increasing [22]. In contrast to monogenic diseases with obvious etiology and phenotypic features, cancer is a complex, polygenic disorder that arises from mutations in multiple genes. In fact, cancer is a result of systematic interactions among several biological processes, rather than a single alteration [23,24]. With the rapid growth of high throughput techniques to generate biological data, we can now investigate cancer with a systemic vision using a holistic approach.

In this study, we first constructed an integrative human disease–disease network (IHDN) using nine different data types, including molecular, functional, and clinical disease features. In the following, we constructed a cancer related disease network as cancerome via extraction of a subset of data from the IHDN that included cancers and its first neighbors.

2. Materials and methods

2.1. Data sources

Two main categories of comprehensive data were collected: first, data that described molecular-based relations with diseases, such as genes involved in a disease and second, functional and clinical data related to diseases, such as pathways, comorbidities, signs and symptoms. The source data are described in details belows.

For disease–gene, disease–pathway, and disease–drug associations, the Comparative Toxicogenomics Database (CTD) human dataset was used [25]. The CTD contains both manually curated information and inferred relationships with a focus on understanding the effects of environmental chemicals on human health. Only curated data from the CTD resource were used. Curated disease–gene associations with direct evidence for marker, mechanism, or therapeutic effects were extracted manually by CTD curators from the published literature or from the OMIM database [5]. The disease–pathway associations obtained from the CTD dataset were based on the genes that occurred both in CTD-curated disease–gene associations and in pathway–gene associations established by KEGG [26] and REACTOME [27]. Disease–drug associations were extracted from the CTD disease–chemicals resource. Only curated associations with therapeutic evidence that had been extracted from the published literature were selected.

Disease–protein associations were obtained from the UniProt/SwissProt database [28]. The UniProt database contains curated information about protein sequences, structure, and function. Human disease-associated proteins were extracted from related files.

Disease–genomic variant associations were obtained from ClinVar [29]. ClinVar is a database of relationships among medically relevant human genomic variants and phenotypes. The relationships in ClinVar are supported by the existence of a variant–phenotype relationship based on family studies, population analyses, or functional assays. ClinVar is hosted by the National Center for Biotechnology Information and funded by the intramural National Institutes of Health, and is freely accessible.

Disease–SNP associations data were downloaded from the Disease-Connect database [20]. These data comprise manually curated causative SNPs from published GWAS, as cataloged by the National Human Genome Research Institute [30]. Only genes with significant p-values ($<1\mathrm{E}-6$) were selected.

Disease–miRNA association data were obtained from the miRTarBase database [31]. The miRTarBase database is the largest and most updated collection of miRNA–target interactions (MTIs), which are curated manually and derived from functional miRNA studies.

Only MTIs that have been validated experimentally by reporter assay and/or Western blot (supported by strong experimental evidence) were selected. The miRNA gene targets were mapped to diseases using the CTD disease–gene database.

Comorbidity was defined as the coexistence or presence of different diseases in relation to a disease in a patient, and a comorbidity relationship between two diseases exists when there is a statistically significant difference between appearing simultaneously or by chance [32]. These disease comorbidity data could reveal important relationships between diseases, but there are several major limitations to their use. For instance, the data were collected from medical records of elderly hospitalized patients; thus, they do not contain information on out-patients, or patients with infectious diseases or pregnancy-related conditions. In addition, because the patients were elderly, most of the diseases were age-related, such as heart diseases and cancers, which are highly prevalent in this age group. Disease–comorbidity associations data from a study by Hidalgo et al., downloaded from the HuDiNe database, was used [32]. Only disease–disease associations with a t-value greater than 1.96 (p-value <0.05) were selected, and assigned lowest weight to comorbidity sharing relationships.

For disease–symptom associations, data were obtained from the HSDN [16], which was derived from a large number of medical bibliographic records. The HSDN was constructed using 322 symptoms and 4219 diseases with 147,978 disease–symptom links and 7,488,851 disease–disease similarity links. Only significant links with similarity scores greater than 0.1 (1,121,899 links) were selected.

2.2. Diseases identifier mapping

Each data source such as MeSH, OMIM, and ICD9CM uses own its vocabulary. The unified medical language system (UMLS) [33] version 2015AB concepts was used to construct disease–disease relationships and to convert other vocabulary concepts to Concept Unique ID (CUI) using the UMLS Metathesaurus. To construct the disease–disease network, two disease concepts were considered to be linked if they were associated with the same feature in each of used data.

2.3. Disease relationship assessment

The Jaccard similarity coefficient [34,35] was used as an association index (AI) to calculate the proportion of overlap between two sets of features, such as gene and pathway, between two diseases. The Jaccard index is the proportion of shared elements between set A and set B relative to the total number of the union of the elements. The AI is defined as below:

$$\mathit{Association\ Index\ (AI)}_{i,j} = \frac{\left|N(D_i)\ \cap N(D_j)\right|}{\left|N(D_i)\ \cup N(D_j)\right|}$$

where $N(D_i)$ were genes (or other features) of disease i and $N(D_j)$ were genes (or other features) of disease j. The range was $0 \le \mathrm{AI} \le 1$.

A disease–disease association score (AS) was developed for integration and ranking disease–disease links. The AS integrates the available AI of each association based on the weight of each AI. The AI weights were assigned to each feature according to their importance, which was determined by experts. As mentioned above, GWAS have the most impact and therefore the highest weight, while comorbidity has the lowest weight due to the “noisy” data. The AS ranges from 0 to 1, and is computed as below: