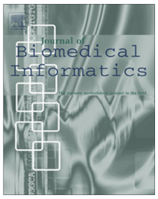
Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



Comparative analysis of a novel disease phenotype network based on clinical manifestations

Yang Chen [a,b], Xiang Zhang [a], Guo-qiang Zhang [a,b], Rong Xu [b,*]

[a] *Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH 44106, United States*
[b] *Division of Medical Informatics, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, United States*

ARTICLE INFO

Article history:
Received 16 June 2014
Accepted 21 September 2014
Available online xxxx

Keywords:
Ontology
Disease phenotype network
Network analysis

ABSTRACT

Systems approaches to analyzing disease phenotype networks in combination with protein functional interaction networks have great potential in illuminating disease pathophysiological mechanisms. While many genetic networks are readily available, disease phenotype networks remain largely incomplete. In this study, we built a large-scale Disease Manifestation Network (DMN) from 50,543 highly accurate disease-manifestation semantic relationships in the United Medical Language System (UMLS). Our new phenotype network contains 2305 nodes and 373,527 weighted edges to represent the disease phenotypic similarities. We first compared DMN with the networks representing genetic relationships among diseases, and demonstrated that the phenotype clustering in DMN reflects common disease genetics. Then we compared DMN with a widely-used disease phenotype network in previous gene discovery studies, called mimMiner, which was extracted from the textual descriptions in Online Mendelian Inheritance in Man (OMIM). We demonstrated that DMN contains different knowledge from the existing phenotype data source. Finally, a case study on Marfan Syndrome further proved that DMN contains useful information and can provide leads to discover unknown disease causes. Integrating DMN in systems approaches with mimMiner and other data offers the opportunities to predict novel disease genetics. We made DMN publicly available at nlp/case.edu/public/data/DMN.

1. Introduction

Linking complex human diseases to their genetic basis remains a challenging task. For computational strategies to discover candidate disease genes, incorporating new data may lead to new discoveries. Traditional methods prioritized genes for a disease if the genes have similar functions with the known disease genes [2,38,44,39,32,17,48]. Recent studies incorporate disease phenotype similarities in addition to the genomic data to increase the ability of identifying new disease genes [19,23,43,46,47,16,35,37], assuming that similar phenotypes and overlapping genetic causes are correlated [5,29,15,2,9,10].

However, the disease phenotype networks used in current gene prediction approaches remain largely incomplete. Most phenotype databases were constructed through mining textual phenotype descriptions [18,6]. For example, van Driel and the colleagues extracted disease-phenotype associations from OMIM through text mining, calculated the pairwise disease similarities, and stored them in the database called mimMiner [42], which is one of the most widely-used phenotype networks in recent disease gene discovery methods [23,43,33,36,16]. Combining different phenotype data has the potential to reduce the bias in each data source and improve the network-based prediction models [26,30]. Therefore, we explored new accurate and publicly accessible disease phenotype data in addition to the existing phenotype networks.

In this study, we created Disease Manifestation Network (DMN), using the highly accurate and structured clinical manifestation data from Unified Medical Language System (UMLS) [24,4,25]. Clinical manifestation captures a major aspect of disease phenotype and can predict disease causes [5]. For example, the Stickler syndrome, Marshall syndrome and Otospondylomegaepiphyseal dysplasia (OSMED) have highly similar manifestations and also involve mutations in interacting collagen genes COL2A1, COL11A2, and COL11A1, respectively [1]. The UMLS semantic network currently uses 50,543 disease-manifestation semantic relationships to explicitly link 2,305 diseases to their clinical manifestations. In this knowledge base, all disease and manifestation terms are formally represented by unified concepts and the semantic relationships between concepts were collected from multiple different ontologies.

[*] Corresponding author.
E-mail address: rxx@case.edu (R. Xu).

http://dx.doi.org/10.1016/j.jbi.2014.09.007
1532-0464/© 2014 Published by Elsevier Inc.

We hypothesized that DMN not only reflects known disease-gene relationships, but also contains different phenotypic knowledge compared with mimMiner. We tested the hypothesis through network comparative analysis between DMN, mimMiner [42], and the two variants of human disease network (HDN) [12], which connects diseases if they share genes. The correlation between DMN and HDNs indicated that DMN reflects existing knowledge on genetic relationships among diseases. The comparison between DMN and mimMiner demonstrated that the two phenotype networks are largely complementary in nodes, edges and community structures. The overall analysis suggests that combining DMN with previous phenotype data sources, such as mimMiner, may potentially improve the data-driven methods for biomedical applications, such as disease gene discovery and drug repositioning.

2. Data and methods

Our study consists of the following steps (Fig. 1): (1) Constructed DMN using the disease-manifestation associations from UMLS; (2) compare phenotypic relationships in DMN and genetic relationships among diseases; (3) compared DMN with mimMiner [42]; and (4) conducted a case study on the phenotypic relationships of Marfan Syndrome in DMN.

2.1. Construct DMN using disease-manifestation associations in UMLS

We first extracted disease-manifestation relationships from the UMLS file MRREL.RRF (2013 version). The file contains 647 different kinds of semantic relationships between biomedical concepts. We collected the concepts pairs linked by the “has manifestation” relationship, and obtained 50,543 disease-manifestation pairs. The disease-manifestation relationships come from OMIM [14], Ultrasound Structured Attribute Reporting [3], and Minimal Standard Digestive Endoscopy Terminology [40]. OMIM is the major contributor among these data sources.

The manifestation terms vary greatly in abundance. For example, common manifestations such as “seizures” are associated with many diseases, while rare manifestations such as “Amegakaryocytic thrombocytopenia” are only associated with one disease. We used the information content (1) into weight each manifestation concept.

$$w_c = -log(n_c/N) \tag{1}$$

Variable w_c is the weight of the manifestation concept c, n_c is the number of diseases associated with manifestation c, and N is the total number of diseases. Then we modeled the manifestation similarity between disease x and y by the cosine of their feature vectors in (2), in which the feature vectors consist of manifestations x_i and y_i for disease x and y. The cosine similarity was used before [19,42] to quantify phenotype overlaps.

$$s(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} \tag{2}$$

We constructed DMN as a weighted network with the manifestation similarities. The edges weights are in the range (0, 1].

2.2. Compare phenotypic relationships in DMN with genetic disease associations

We conducted two experiments to evaluate whether the phenotypic relationships in DMN reflect genetic associations among diseases. The first experiment is to calculate the correlation between the disease similarities in DMN and two quantified measures of genetic associations. We first ranked the edges (disease pairs) in DMN by their weights (disease similarities) from large to small. For top N disease pairs, we counted the percentage of disease pairs that share associated genes in OMIM and the average number of genes shared by the N disease pairs. Then we calculated the Pearson’s correlations between N and the genetic measures.

In the second experiment, we compared the network topologies between DMN and two genetic disease networks. A well-studied genetic disease network is HDN, in which diseases were connected if they share associated genes in OMIM and edges were weighted

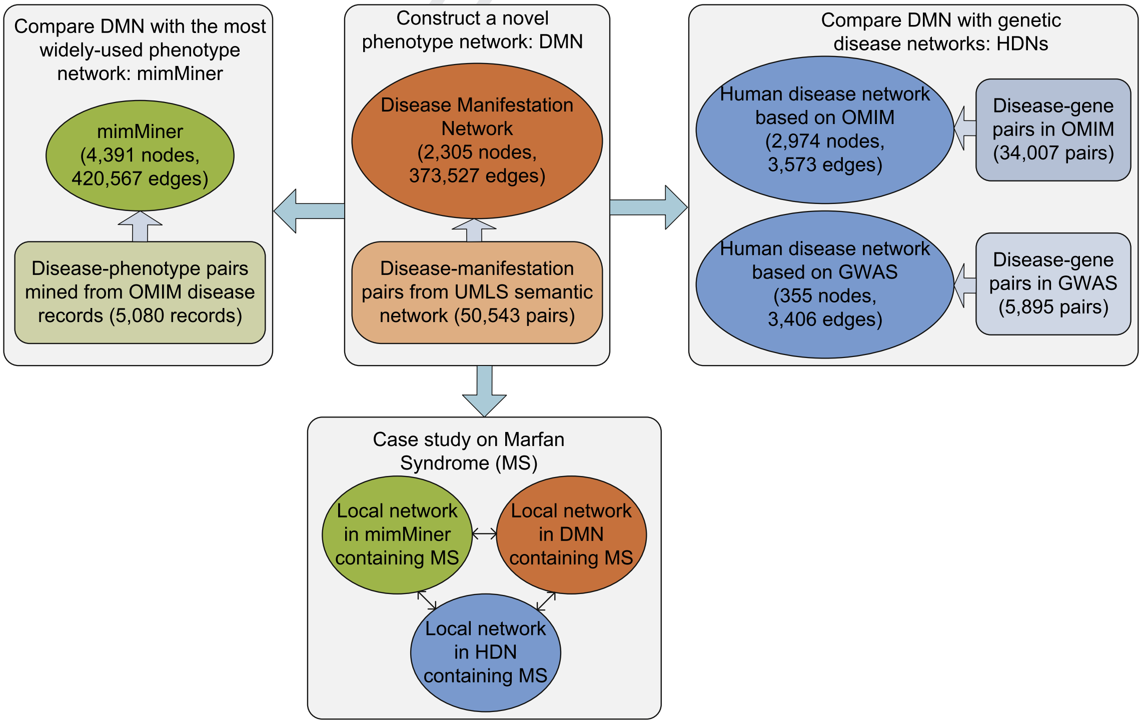

Compare DMN with the most widely-used phenotype network: mimMiner
mimMiner (4,391 nodes, 420,567 edges)
Disease-phenotype pairs mined from OMIM disease records (5,080 records)
Construct a novel phenotype network: DMN
Disease Manifestation Network (2,305 nodes, 373,527 edges)
Disease-manifestation pairs from UMLS semantic network (50,543 pairs)
Compare DMN with genetic disease networks: HDNs
Human disease network based on OMIM (2,974 nodes, 3,573 edges)
Disease-gene pairs in OMIM (34,007 pairs)
Human disease network based on GWAS (355 nodes, 3,406 edges)
Disease-gene pairs in GWAS (5,895 pairs)
Case study on Marfan Syndrome (MS)
Local network in mimMiner containing MS
Local network in DMN containing MS
Local network in HDN containing MS

Fig. 1. The four steps of network analysis for DMN.