RESEARCH

Predicting disease associations via biological network analysis

Kai Sun[1], Joana Gonçalves[1], Chris Larminie[2] and Nataša Pržulj[1*]

[*]Correspondence:
natasha@imperial.ac.uk
[1]Department of Computing,
Imperial College London, London,
SW7 2AZ, UK
Full list of author information is
available at the end of the article

Abstract

Background: Understanding the relationship between diseases based on the underlying biological mechanisms is one of the greatest challenges in modern biology and medicine. Exploring disease-disease associations by using system-level biological data is expected to improve our current knowledge of disease relationships, which may lead to further improvements in disease diagnosis, prognosis and treatment.

Results: We took advantage of diverse biological data including disease-gene associations and a large-scale molecular network to gain novel insights into disease relationships. We analysed and compared four publicly available disease-gene association datasets, then applied three disease similarity measures, namely annotation-based measure, function-based measure and topology-based measure, to estimate the similarity scores between diseases. We systematically evaluated disease associations obtained by these measures against a statistical measure of comorbidity which was derived from a large number of medical patient records. Our results show that the correlation between our similarity measures and comorbidity scores is substantially higher than expected at random, confirming that our similarity measures are able to recover comorbidity associations. We also demonstrated that our predicted disease associations correlated with disease associations generated from genome-wide association studies significantly higher than expected at random. Furthermore, we evaluated our predicted disease associations via mining the literature on PubMed, and presented case studies to demonstrate how these novel disease associations can be used to enhance our current knowledge of disease relationships.

Conclusions: We present three similarity measures for predicting disease associations. The strong correlation between our predictions and known disease associations demonstrates the ability of our measures to provide novel insights into disease relationships.

Keywords: disease classification; network analysis; graph theory; topology; protein-protein interaction

Background

Correct diagnosis is critical for effective treatment and prevention of disease. As a result, disease classification has become a key cornerstone of modern medicine. Disease may be classified by any one of a number of criteria: topographic, anatomic, pathological, physiological, etiological, juristic, epidemiological or statistical approaches. However, without considering the molecular mechanisms driving diseases, such knowledge is limited and can even be misleading. For example, a common phenotype can be caused by different underlying mechanisms, such as breast cancer,

which can be divided into several subgroups that are characterized by distinct patterns of pathway activation [1]. However, a common mechanism may lead to different phenotypes. For example, a mutation at the β-globin locus may lead to sickle-cell anemia with different phenotypes such as bony infarcts, acute chest syndrome and stroke [2].

During the past decade, a wealth of biological data has been generated from various large-scale genomic studies, prompting the scientific community to gain deeper insight into disease relationships based on their underlying biological mechanisms. Various types of biological data have been used to infer associations between diseases. One of the most commonly used biological data is disease-gene association. In a broad definition, a disease-gene association is a connection reported in the literature, which can be a genetic association (i.e., mutations in that gene may lead to that disease), or a connection inferred from other aspects. Disease-gene associations can be obtained from large-scale knowledge-bases such as the Online Mendelian Inheritance in Man (OMIM) [3]. Early studies used text mining to infer similarities between phenotypes contained in OMIM, and found those similarities were positively correlated with a number of measures of gene functions [4] and could be used to predict disease-causing genes [5]. Also by using OMIM, Goh *et al.* [6] constructed the human diseasome by connecting diseases that share a disease-causing gene. Other types of biological data such as biological pathways [7], gene expression data [8, 9], biomedical ontologies [10, 11], and genome-wide association study (GWAS) data [12, 13, 14], have also been used to improve the current understanding of disease relationships from different aspects. Recently, networks have been used to model large-scale biological data, and network topology is beginning to provide insights into diseases and their associations [6, 15, 16, 17]. By considering the interconnectivity of biomolecules in the cell, the topology of biological networks is expected to have various biological and clinical applications [18, 19].

Despite these advances, early studies have several limitations when inferring disease associations from biological data. First, some studies only considered several specific diseases, rather than giving a global comparison among all diseases (e.g., [12, 13, 14, 9]). This is the case for GWAS-based studies, since a small number of GWAS studies have been completed to date in a relatively small proportion of the total disease population. Furthermore, most studies solely used OMIM as the source of disease-gene association data. OMIM is a catalogue of mendelian disorders and as a result, most diseases are annotated with few genes in OMIM [20]. Limitations of using OMIM have also been discussed previously [21, 22]. Finally, most computationally predicted disease associations were not systematically evaluated due to the difficulty in identifying a suitable benchmark of known disease associations. In particular, most studies were only able to validate part of their results by comparing them with phenotypic similarities (e.g., [12]) or mining the literature manually (e.g., [13]). A comparison of previous studies can be found in Table 1.

In our study, we used diverse biological data from a number of repositories to gain novel insights into the relationship of over 500 known human diseases by considering their underlying biological mechanisms. We used disease-gene associations obtained from four different sources to avoid the bias introduced by a single dataset. Moreover, we took advantage of the topology of a large-scale molecular network to

examine its use for inferring disease associations. We applied three different disease similarity measures, namely annotation-based measure, function-based measure and topology-based measure, to estimate similarity scores between diseases. The disease associations obtained by the three measures were systematically evaluated against the standard disease classification system, namely the International Classification of Diseases[1], 9^{th} revision (ICD-9), and a statistical measure of comorbidity derived from a large number of medical patient records. In addition, we evaluated our predicted disease associations by using disease associations generated from GWAS studies, which represent one of the most robust routes for identifying causal relationships between genes and diseases. To our knowledge, this is the first time comorbidity and GWAS data have been used to evaluate computationally predicted disease associations.

In the rest of this paper, we will start with a description of the biological data we analysed, followed by details of our methodology of measuring disease associations. Then we will show and discuss the evaluation of disease associations predicted by our similarity measures against known disease associations derived from ICD-9, comorbidity data and GWAS data. Finally, we will present case studies to demonstrate the ability of our similarity measures to predict novel disease associations.

Methods

Biological data

Three types of biological data were used in this study: protein-protein interactions (PPIs), Gene Ontology (GO) annotations and disease-gene associations.

PPI network.

We modelled PPI data as a network. A *network* or *graph* $G(V, E)$ consists of two types of elements, a set V of nodes and a set $E \subseteq V \times V$ of edges connecting them. A PPI network models the physical interaction among proteins in the cell, in which a node represents a protein, and an undirected edge exists between a pair of nodes if their corresponding proteins can physically bind to each other. Currently available PPIs are mostly yielded from various high throughput proteomics experiments, such as yeast two-hybrid screening (e.g., [23]) and affinity capture mass spectrometry (e.g., [24]). We constructed a human PPI network using data obtained from BioGRID [25] version 3.1.93 (released in October 2012). All self-loops, duplicate interactions were removed since we considered only simple, undirected graphs. We also removed the cross-species interactions (i.e., interactions between human proteins and proteins of other species) because we focused on the physical interactions between human proteins in our study. The PPI network we constructed contained 11,375 nodes and 66,317 edges, while its largest connected component contained 11,261 nodes and 66,253 edges. Note that the second largest connected component only contained 5 nodes and 5 edges. There were also 7 isolated triangles and 43 isolated edges in the PPI network. The presentence of these small components may be due to the incompleteness of the PPI data. In addition, the topology of these small components is not as informative as that of the largest connected component. For these reasons, we only used the largest connected component of the PPI network in our analysis.

[1] http://www.who.int/classifications/icd/en/

GO annotations.

Genes are annotated with GO terms to represent their biological properties [26]. All GO terms are organised in three domains: cellular component, molecular function and biological process. We downloaded the ontology file and annotations of *Homo sapiens* from the Gene Ontology database[2] in November 2012. We removed annotations with evidence code 'Inferred from Electronic Annotation' (IEAs), since IEAs are computationally inferred annotations which have not been reviewed by curators. In total, we collected 171,888 annotations between 13,166 genes and 10,787 GO terms.

Disease-gene associations.

Disease-gene associations can be modelled as a graph containing both known human diseases and disease-related genes in the human genome. The degree of a disease is the number of genes associated with that disease, while the degree of a gene is the number of diseases annotated with that gene. We used four disease-gene association datasets obtained from different sources: OMIM, Comparative Toxicogenomics Database (CTD) [27], Functional Disease Ontology annotations (FunDO) [28] and Human Genome Epidemiology Network (HuGENet) [29]. Among these datasets, OMIM, CTD, and FunDO contain curated associations, while HuGENet contains computationally inferred associations. Details of these disease-gene association datasets are described below.

- OMIM is considered to be the best-curated resource of known phenotype-genotype relationships, and it has been used in various disease-related studies (discussed in the Background section). We downloaded the OMIM database in November 2012. In total, it contains 3,537 diseases (annotated by OMIM IDs), 2,862 genes and 4,337 disease-gene associations.
- CTD provides scientific data describing relationships between chemicals, genes, and human diseases, with the goal of improving the understanding of environmental chemicals' effects on human health. It contains both curated and inferred disease-gene associations, but we only used curated associations as they have higher confidence than inferred associations. Disease-gene associations directly derived from OMIM were excluded to reduce the dependency between datasets. We downloaded the data from CTD in November 2012 and obtained 17,754 associations between 2,761 diseases (annotated by Medical Subject Heading (MeSH) terms[3]) and 5,828 genes.
- FunDO contains disease-gene associations extracted from the NCBI Gene Reference Into Function (GeneRIF) database. A GeneRIF is a brief statement about the function of a gene, along with information of its association with diseases. We downloaded the latest stable version of FunDO (released in October 2008) and obtained 1,854 diseases (annotated by Disease Ontology (DO) terms), 4,781 genes and 28,442 disease-gene associations.
- HuGENet is known as an integrated knowledge-base on human genome epidemiology. The Phenopedia collection [29] of HuGENet contains disease-gene associations obtained by text-mining of abstracts on PubMed using machine

[2] http://www.geneontology.org
[3] http://www.nlm.nih.gov/mesh/

learning techniques. Disease-gene association data were downloaded via HuGE Navigator in September 2012. We obtained 353,883 associations between 2,387 diseases (annotated by Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs)[4]) and 11,915 genes.

Since disease names or IDs used in these datasets are based on different labelling schemes, we mapped all disease names or IDs to ICD-9 codes, for the purpose of comparing these datasets and further evaluation (also see the Results and Discussion section for details). We used the mapping manually constructed by [6] and [30] to convert OMIM IDs to ICD-9 codes, and used the corresponding mapping provided in Disease Ontology version 3 (the latest stable version of DO, released in May 2007) to map DO IDs, MeSH terms and UMLS CUIs to ICD-9 codes. In total, 1,467 OMIM IDs in OMIM, 423 MeSH terms in CTD, 806 DO IDs in FunDO and 693 UMLS CUIs in HuGENet were mapped to ICD-9 codes.

Disease similarity measures

We applied three similarity measures to estimate similarity scores between diseases. These measures include standard methods (i.e., Jaccard index) and novel measures proposed in this study (i.e., graphlet-based measure). Considering the information used in calculation, the similarity score of a pair of diseases was measured in three different ways: annotation-based, function-based and topology-based.

Annotation-based measure.

The annotation-based measure solely used the information obtained from disease-gene association data. We applied the Jaccard index, which is known as a standard method for comparing the similarity between two sets, to estimate the similarity score between diseases as follows. Let G_{D_i} be the set of genes associated with a disease D_i. We computed the annotation-based similarity score of two diseases D_i and D_j as the Jaccard index (or Jaccard similarity coefficient) of G_{D_i} and G_{D_j}:

$$Sim_{annotation}(D_i, D_j) = \frac{|G_{D_i} \cap G_{D_j}|}{|G_{D_i} \cup G_{D_j}|}. \quad (1)$$

Function-based measure.

The function-based similarity measure used both GO term annotations and disease-gene associations to estimate the similarity score between a pair of diseases. We first propagated the GO annotations upwards through the GO hierarchy, i.e., when a gene was annotated with a GO term, we assumed associations between the gene and the term's parents. For each disease D_i annotated in a specific disease-gene association dataset, we then identified the set of GO terms that were overrepresented within G_{D_i}, denoted by GO_{D_i}. The statistical significance (p-value) of the enrichment of a GO term was computed according to the hypergeometric distribution for sampling without replacement, and was corrected for multiple testing using the Benjamini-Hochberg test. Only overrepresented GO terms from the 'biological process' domain of GO and having a p-value less than 0.05 were considered to be in

[4] http://www.nlm.nih.gov/research/umls/

GO_{D_i}. For a pair of diseases D_i and D_j, we computed the Jaccard index of GO_{D_i} and GO_{D_j} as their function-based similarity score, defined as:

$$Sim_{function}(D_i, D_j) = \frac{|GO_{D_i} \cap GO_{D_j}|}{|GO_{D_i} \cup GO_{D_j}|}. \quad (2)$$

Topology-based measure.
Many studies have shown the relationship between topological properties of proteins in the PPI network and the involvement of proteins in diseases [6, 31, 32]. Topological similarities of proteins in a PPI network are considered as a complementary information to sequence similarities [33]. Thus in this study, we took advantage of the topology of the human PPI network along with disease-gene association data to examine the use of network topology for uncovering novel disease associations. In particular, we proposed a measure to estimate the similarity score between a pair of diseases based on the topological similarity of their annotated genes.

We applied a graphlet-based method to assess the topological similarity of genes in the human PPI network. A *graphlet* is defined as a small, connected and induced subgraph of a larger network [34]. Within each graphlet, some nodes are topologically identical to each other, and such identical nodes are said to belong to the same *automorphism orbit* [35]. The *graphlet signature* of a node u is a 73-dimensional vector, whose i^{th} element u_i counts the number of times the node u is touched by the particular automorphism orbit i [36]. According to [36], the signature similarity of a pair of nodes u and v is defined as:

$$SigSim(u, v) = 1 - \frac{1}{\sum_{i=0}^{72} w_i}(\sum_{i=0}^{72}(w_i \times \frac{|\log(u_i + 1) - \log(v_i + 1)|}{\log(\max\{u_i, v_i\} + 2)})) \quad (3)$$

where w_i is a weight assigned to orbit i defined as $1 - log(o_i)/log(73)$ (o_i is the dependency count of orbit i, see [36] for details). $SigSim(u, v)$ ranges between 0 and 1, where the value of 1 means that the two nodes, u and v, are considered to be topologically identical. This measure is a highly constraining measure of local topological similarity between two nodes in a network as it compares the nodes based on local structures of their neighbourhoods, which describe their interconnectivities out to a distance of four [36]. Signature similarities have been applied to measure the topological similarities between proteins in a PPI network [36, 37, 31, 33, 38, 39, 40]. It has been shown that topologically similar proteins are likely to belong to the same protein complexes, perform the same biological functions, be localised in the same subcellular compartments and have the same tissue expressions [36]. Signature similarities have also been used to relate the network structure around a protein in a PPI network to homology [33] and its involvement in diseases [31]. For these reasons, we hypothesize that the topology around disease genes in the PPI network can reflect the underlying biological mechanisms of diseases.

We calculated the signature similarity of each pair of genes in the human PPI network. Note that the network has an edge density (the proportion of the number of edges to the maximum possible number of edges) of 0.001, which for its size (11,261 nodes and 66,253 edges) is dense enough to avoid low edge density regions

in which the topology of networks is unstable (see [41] for details). Here we extended the use of graphlet-based method to measure disease similarities. We introduced two terms to quantify the topology-based similarity score between diseases D_i and D_j. The first term, denoted by $AllSig$, is the maximum of the signature similarity between a gene in G_{D_i} and a gene in G_{D_j}:

$$AllSig(D_i, D_j) = \max_{\substack{g_m \in G_{D_i} \\ g_n \in G_{D_j}}} SigSim(g_m, g_n). \tag{4}$$

The second term, denoted by $ShareSig$, focuses on the topological similarity between genes shared with both diseases:

$$ShareSig(D_i, D_j) = \max_{\substack{g_m \neq g_n \\ g_m \in G_{D_i} \cap G_{D_j} \\ g_n \in G_{D_i} \cap G_{D_j}}} SigSim(g_m, g_n). \tag{5}$$

Finally we defined the topology-based similarity score between D_i and D_j as the average of these two terms:

$$Sim_{topology}(D_i, D_j) = \frac{1}{2} \times (ShareSig(D_i, D_j) + AllSig(D_i, D_j)). \tag{6}$$

Evaluation

Comorbidity associations of diseases.

The availability of electronic patient records facilitates studies into disease comorbidity, which indicates the potential for co-occurrence of two given diseases in the same individual. Comorbidity can be considered as a type of disease association derived from electronic medical record, but the underlying driver for comorbidity may be very different from one another. Comorbidity and its correlation with other types of disease associations such as genetic associations [42] and evolutionary associations [43] have previously been studied. Unlike these studies, we used comorbidity data to evaluate disease associations predicted by our similarity measures. Comorbidity associations were downloaded from the Human Disease Network (HuDiNe, [44]), which were obtained from the disease history of 32 million American patients. Diseases were annotated using ICD-9 codes in HuDiNe, and as many diseases in patient records were not specific enough to map to 4-digit or 5-digit codes, we used the comorbidity data annotated using 3-digit level ICD-9 codes for our analysis. The strength of comorbidity association between a pair of diseases can be measured by the Relative Risk and ϕ-correlation [44]. Because comorbidity associations quantified by ϕ-correlation were reported to contain more connections across different ICD-9 categories [44], we chose ϕ-correlation as the measure of comoridity. The ϕ-correlation score between D_i and D_j was defined as the Pearson's correlation for binary variables, given by:

$$\phi(D_i, D_j) = \frac{C_{ij}N - P_iP_j}{\sqrt{P_iP_j(N - P_i)(N - P_j)}} \tag{7}$$

where C_{ij} is the number of individuals affected by both D_i and D_j, N is the total number of individuals in the population, P_i and P_j are the prevalences of D_i and

D_j respectively. A ϕ-correlation higher than 0 indicates the co-occurrence of D_i and D_j is more frequently than expected by random. The statistical significance of ϕ-correlation was determined by using a t-test,

$$t = \frac{\phi\sqrt{n-2}}{\sqrt{1-\phi^2}} \tag{8}$$

where $n = max(P_i, P_j)$ is the number of observations used to calculate ϕ. We used significant associations at 5% level ($t \geq 1.96$) for our analyses.

GWAS data.
GWAS is a powerful method to identify genetic variations associated with diseases and is one of the most robust routes for identifying causal relationships between genes and diseases [45, 46]. GWAS studies examine the genome for single-nucleotide polymorphisms (SNPs) that occur more frequently in people with a particular disease than in people without it. GWAS studies have enabled exploration of gene association in complex diseases in a systematic way on a genome scale. Whilst individual studies are extremely powerful, only a small number of diseases have been studied thus far using GWAS. Hence the GWAS database as a whole is only able to contribute a relatively small component to the overall knowledge base of general disease-gene associations. For this reason, we did not use GWAS data as a source of disease-gene association to measure disease similarity scores, but used them to evaluate our predicted disease associations. We downloaded GWAS data from the National Human Genome Research Institute (NHGRI) GWAS catalog [47] in May 2013. This resource collects significant associations between traits (or diseases) and SNPs from the literature. Similar to [48], we only considered highly confident associations with p-value lower than 10^{-7}. We also eliminated not replicated associations to minimise false-positives. For all disease-SNP associations in our analysis, we used the corresponding disease-gene associations reported by the authors in the original publications as recorded in the GWAS Catalog. After mapping diseases to ICD-9 codes, we obtained 1,756 genetic associations (from 478 publications) between 126 diseases and 1,298 genes.

Results and discussion

Comparison of disease-gene association datasets

We analysed four different disease-gene association datasets: three curated datasets, namely OMIM, CTD and FunDO, and one computationally predicted dataset, HuGENet (details of these datasets can be found in the Methods section). Although these datasets focus on different aspects of the connections between diseases and genes, they are not fully independent since information contained in these datasets is extracted from the literature. For example, disease-gene associations contained in CTD and FunDO were extracted from 9,269 and 48,436 publications respectively, and they have 799 publications in common. We mapped all disease names or IDs annotated in these datasets to ICD-9 codes for a correct comparison (see the Methods section for more details). If several diseases were mapped to a common ICD-9 code, we assigned the union of genes associated with those diseases to that ICD-9

code. In order to evaluate our measures using comorbidity data, we further limited the ICD-9 codes to 3-digit level. We are aware that noise may be introduced when merging diseases into 3-digit level. Generally speaking, a 3-digit level ICD-9 code is always associated with more than one disease, thus the average degree of diseases increased after mapping. Note that it is possible that two diseases may share clinical traits but have different 3-digit level ICD-9 codes, e.g., acute bronchitis (ICD-9: 466) and chronic bronchitis (ICD-9: 491). However, in most cases if two diseases have different ICD-9 codes at 3-digit level, they always have different clinical phenotypes and they are unlikely to share similarity traits.

Interestingly, the overlap among the four disease-gene association datasets is unexpectedly small, as shown in Figure 1. While a considerable number of diseases (120 diseases in total, that is, 50.21%, 47.43%, 26.20% and 33.33% of diseases annotated in OMIM, CTD, FunDO and HuGENet, respectively) have gene annotations in all four datasets, few disease-gene associations (159 associations in total, that is, 7.05%, 1.99%, 0.92% and 0.11% of associations in OMIM, CTD, FunDO and HuGENet, respectively) can be found in all datasets. Figure S1 further demonstrates the difference between these datasets according to the degree distribution of diseases. In general, these distributions follow power law distributions, indicating that most human diseases are associated with only a few disease genes, while a small number of diseases relate to many genes. However, this scale-free topology may also be an artifact of sampling: several diseases are better studied than others [49]. We notice that in OMIM, most diseases are associated with fewer genes compared with other datasets. The average number of genes associated with a disease in OMIM is 9.43, while in the two other curated datasets CTD and FunDO, these numbers are 31.59 and 37.80. On the other hand, on average a disease in HuGENet is annotated with more than 300 genes: HuGENet has a higher false positive rate compared to other datasets, since its associations were derived from computational predictions rather than manual curations.

The difference and inconsistency discussed above indicate that currently available disease-gene association datasets are still noisy and incomplete. The incompleteness may be due to the focus of the datasets and the nature of the curation process. For example, OMIM mainly focuses on mendelian diseases and traits. Meanwhile, many false positives may be introduced by text-mining the literature (e.g., HuGENet). However, there is no single standard and systematic method to assess the quality of these data. Therefore, to gain a more comprehensive view of human diseases and to test the robustness of our methods, we used all four disease-gene association datasets along with the intersection/union of the three curated datasets in further computation and evaluation.

Evaluation of similarity measures

Correlation with ICD-9.

The results obtained by these measures were first evaluated against the standard disease classification system ICD-9. We say that two diseases are associated according to ICD-9, if they are classified under the same ICD-9 category[5]. For example, diabetes mellitus (ICD-9 code: 250) and thyroiditis (ICD-9 code: 245) are classified

[5] http://www.icd9data.com/2013/Volume1/default.htm

under the same category 'endocrine, nutritional and metabolic diseases, and immunity disorders'. To investigate the correlation between our similarity measures and the ICD-9 classification, we tested whether a pair of diseases from the same ICD-9 category tends to have a higher similarity score than diseases from different ICD-9 categories (Table 2). Since similarity scores obtained by our measures are not normally distributed, we used a non-parametric test, namely the Mann-Whitney U test, to assess the statistical significance (p-value). Our results show that for all three similarity measures and all four disease-gene association datasets, similarity scores of diseases from the same ICD-9 category are significantly higher than those from different ICD-9 categories.

Correlation with comorbidity.
As the goal of our study is to uncover novel disease associations that may reflect common underlying mechanisms, we are more interested in the associations between diseases that belong to different ICD-9 categories. For this reason, we systematically evaluated our similarity measures against a statistical measure of comorbidity. We say two diseases are associated according to comorbidity if they are reported to have a significant co-occurrence in the same individual. In particular, their ϕ-correlation score should be higher than a chosen threshold and statistically significant at 5% level. Figure S2 shows the distribution of ϕ-correlation scores for all pairs of diseases we analysed. Note that even though the comorbidity associations we used for evaluation contained disease associations across different ICD-9 categories, there was overlap between associations derived from ICD-9 and comorbidity associations. For example, the association between diabetes mellitus and obesity was supported by both ICD-9 classification and comorbidity data. Since ICD-9 and comorbidity describe the relationship between diseases from different aspects, we believe the evaluations against ICD-9 classification and comorbidity do not contradict each other, but are complementary to each other.

To assess the ability of our measures to uncover highly confident comorbidity associations, we used Receiver Operating Characteristic (ROC) curves, in which we plotted the *True Positive Rate* (TPR, also known as *sensitivity*) versus the *False Positive Rate* (FPR, also known as $1-$*specificity*) for different thresholds of similarity score. TPR is defined as the fraction of true positives (that is, all pairs of diseases having a similarity score higher than a chosen threshold and having comorbidity association) out of the positives (all pairs of diseases having comorbidity association), while FPR is defined as the fraction of false positives (all pairs of diseases having a similarity score higher than a chosen threshold but having no comorbidity association) out of the negatives (all pairs of diseases excluding those having comorbidity association). Figure 2, Figure S5 and Table 3 show the ROC curves and Area Under Curve (AUC) values obtained by the three disease similarity measures. To illustrate that our results cannot be obtained by chance, we assigned a randomised score which was drawn from the same distribution of the similarity scores to each pair of diseases, and evaluated associations derived from these randomised scores against comorbidity. We show that the correlation between our similarity measures and comorbidity scores is substantially higher than expected at random for all disease-gene association datasets we analysed. In particular, diseases

yielding a high similarity score are very likely to have comorbidity associations, thus confirming that our measures are able to uncover known comorbidity relationships.

While varying the ϕ-correlation threshold, we obtained higher AUC values for higher thresholds (the ROC curves are not shown in the paper due to space limitations). For example, when the ϕ-correlation threshold was set to 0.06 (49 comorbidity pairs), the AUC value was 0.7580 ± 0.0024 (using the topology-based measure and FunDO as the source of disease-gene associations). When the ϕ-correlation threshold was set to 0.08 (33 comorbidity pairs) and 0.10 (25 comorbidity pairs), the AUC value increased to 0.7669 ± 0.0027 and 0.7996 ± 0.0060, respectively. This indicates our similarity measures tend to detect strong comorbidity associations with high ϕ-correlation. Meanwhile, when we decreased the number of false negatives in the comorbidity data by lowering the ϕ-correlation threshold from 0.06 to 0.02, the AUC values we obtained were still higher than expected at random. For example, when the ϕ-correlation threshold was set to 0.04 (93 comorbidity pairs) and 0.02 (300 comorbidity pairs), the AUC values we obtained were 0.7064 ± 0.0019 and 0.6017 ± 0.0015, respectively. These results suggest our similarity measures are robust to high false negatives in the comorbidity data. Better ROC curves can also be obtained by evaluating diseases annotated with higher numbers of genes (Figure S5). From Table 3, we observed that best performances of our similarity measures are achieved by using highly confident curated disease-gene associations (i. e. the intersection set of OMIM, CTD and FunDO), with AUC values higher than 0.98.

Note that our approach is robust to the incompleteness presented in disease-gene association datasets and PPI networks. We downloaded the disease-gene association data from OMIM and the PPI data from BioGRID (version 3.2.112) in June 2014 to re-examine whether we obtained the same results when we used the latest biological data. In total, the OMIM data contained 4,002 diseases (annotated by OMIM IDs), 3,218 genes and 4,816 disease-gene associations. The PPI network we constructed contained 14,089 nodes and 126,891 edges. By re-computing the similarity scores and evaluating the results against comorbidity on these latest biological data, we showed that we were able to obtain results (shown in Figure S4) that agree with the ones reported in Table 3 and Figure S5. These results further validated the robustness of our approach.

Correlation with GWAS data.

We further examined the correlation between our predicted disease associations and currently available highly confident GWAS data (see the Methods section for details) to see whether our findings are supported by GWAS studies. A gene is said to be associated with a disease according to GWAS, if the occurrence of genetic variants (SNPs) within that gene is significantly higher in people with that disease than in people without it. We say that two diseases are associated according to GWAS if they share at least one gene in GWAS data. Since disease-gene associations collected in the four datasets we analysed were extracted from the literature, genetic associations reported in GWAS studies may also be collected in these datasets. To avoid bias in evaluation, we chose FunDO as the source of disease-genes associations, as it has few overlap with GWAS data. In particular, since most GWAS data were published after FunDO's last stable release (October 2008), only 42 out of 48,436

publications in FunDO were also found in GWAS data. We removed disease-gene associations collected from the common 42 publications before computing similarity scores between diseases using FunDO. Similar to our evaluation against comorbidity, we used ROC curve analysis to assess the ability of our similarity measures to recover disease associations derived from GWAS (Table 4). For each of the three measures, we found that the correlation between our similarity measures and GWAS data is substantially higher than expected at random. This result further confirms the validity of our methods.

Comparison of similarity measures

The three similarity measures, namely annotation-based measure, function-based measure, and topology-based measure, use different biological information to predict disease associations. For a pair of diseases, the annotation-based measure estimates their similarity score based on the overlap of their annotated genes, while the function-based measure estimates their similarity score based on the overlap of their associated biological functions derived from GO annotations. The topology-based measure makes use of the topology information derived from the underlying PPI network, and estimates disease similarity scores based on the topological similarity of their annotated genes. Based on our evaluation, the three similarity measures perform well in recovering known disease associations. Note that since all three measures compare diseases based on information derived from their associated genes, the three measures are not independent from each other. Diseases that have many shared genes are likely to have common biological processes and have high topological similarities. In addition, a part of the GO annotations is inferred from PPIs (i.e., annotations with evidence code 'inferred from physical interactions'). However, even though dependency between the three measures exists, the three measures uncover different aspects of disease-disease associations. In fact, the predictions derived from them can differ from each other, demonstrating that the three measures give different insights despite being dependent. Figure S3 shows the overlap of disease associations predicted by the three measures. When considering the top 5% of the most associated disease pairs as our predicted disease associations, 14% $\sim$ 38% of the predictions are supported by all three similarity measures.

In the topology-based measure, we used two terms, namely $AllSig(D_i, D_j)$ and $ShareSig(D_i, D_j)$, to measure the topological simialrity of disease genes. Since the term $AllSig(D_i, D_j)$ is defined as the maximum of the signature similarity between a gene associated with disease D_i and a gene associated with disease D_j, we have $AllSig(D_i, D_j) = 1$ if the two diseases D_i and D_j have at least one common genes. The term $ShareSig(D_i, D_j)$ is defined as the maximum of the signature similarity between genes that are shared between diseases D_i and D_j, thus we have $ShareSig(D_i, D_j) = 0$ if the two diseases share no genes. Therefore, the topology-based similarity score for a pair of diseases that share genes is always higher than a pair of diseases that do not share genes. To assess the contribution of the two terms, $AllSig$ and $ShareSig$, in predicting disease associations, we evaluated the performance of the topology-based similarity measure for predicting comorbidity associations by solely using $AllSig$ and $ShareSig$ as the disease similarity score. The good performance of the topology-based similarity measure is mainly attributed to

the term $AllSig$ when using OMIM or CTD as the disease-gene association dataset (Table S3). Since in these two datasets, only 2.69% (OMIM) and 16.62% (CTD) disease pairs have common genes, we have $Sim_{topology} = AllSig$ for most disease pairs. On the other hand, the good performance of the topology-based measure is mainly caused by $ShareSig$ when using FunDO or HuGENet, as in these two datasets 31.41% (FunDO) and 80.57% (HuGENet) of disease pairs have common genes.

Our similarity measures are sensitive to the noise in disease-gene association data. We notice that prediction performances of our similarity measures generally decrease with the increase of noise level, thus using the intersection of curated disease-gene association datasets results in the best performance when predicting comorbidity associations (Table 3 and Figure S5). Both the annotation-based measure and the topology-based measure have better performances by using curated disease-gene associations (i.e., OMIM, CTD and FunDO) than computationally predicted associations (i.e., HuGENet). However, the function-based measure obtains lower AUC values for curated datasets CTD and FunDO than the two other similarity measures, but higher AUC values for HuGENet. In this regard, the function-based measure may be more appropriate for analysing predicted datasets, while the annotation-based measure and topology-based measure may be more appropriate for analysing curated datasets.

The annotation-based measure is straightforward, but has relatively good performance according to our evaluation. However, as it only uses disease-gene associations to estimate similarity scores, for a pair of diseases sharing few genes, their annotation-based similarity score may be low, even if their annotated genes are closely related. In particular, the annotation-based measure is highly affected by the occurrence of pleiotropic genes (genes that cause multiple phenotypes) in the dataset. We obtained the list of 802 pleiotropic genes from the OMIM Morbidmap by identifying genes that associated with more than one disease (similar approach was used in [50]). To examine the influence of pleiotropic genes on our measures, we excluded these genes from OMIM and evaluated the performances of our similarity measures against comorbidity. Note that when pleiotropic genes were excluded from OMIM, there were no disease pairs that had any common genes. Therefore, the annotation-based similarity score for a pair of diseases became 0 in this case and no predictions could be derived from the annotation-based measure. On the other hand, since both the function-based measure and the topology-based measure use additional data sources (GO annotations or network topology) to estimate similarity scores, they are less affected by pleiotropic genes. AUC values obtained by the function-based measure and the topology-based measure dropped to 0.7816 and 0.7199 respectively, after removing pleitropic genes from OMIM. These results show the contribution of similarities between specific genes (genes associated with only one disease) to the prediction performances of our similarity measures.

Since disease-gene association datasets were obtained by different research groups and approaches, good performances for all datasets confirm the robustness of our similarity measures in predicting disease associations. In addition, the topology-based measure is also robust to the noise and incompleteness presented in PPI networks. We evaluated this by using PPI data obtained from different releases of

BioGRID database (see Table S1 for details). Generally speaking, the performance of the topology-based measure slightly decreases when using early PPI networks (Table S2). However, AUC values obtained by using these early PPI networks are still substantially higher than expected at random. These results suggest that the ability of the topology-based measure to predict disease-disease associations may increase with more accurate and complete PPI data.

Case studies

To demonstrate how our similarity measures can be used for uncovering novel disease associations, we present a case study for diabetes mellitus (DM, ICD-9 code: 250). DM is a metabolic disease that affects the body's ability to produce or use insulin, a hormone for regulating carbohydrates. It causes hyperglycemia and may lead to severe consequences such as brain damage, amputations and heart disease [51]. Table 5 lists the top 10 diseases associated with DM using the topology-based measure and FunDO as the source of disease-gene associations. Results obtained by other measures and data are not shown here due to space limitations.

Among these 10 diseases, both ovarian dysfunction (ICD-9 code: 256) and obesity (ICD-9 code: 278) are classified under the same ICD-9 catalogue 'Endocrine, nutritional and metabolic diseases, and immunity disorders' with DM. In addition, both obesity and essential hypertension (ICD-9 code: 401) have highly confident comorbidity associations with DM. Note that among all disease pairs that we analysed, only 0.74% of them have a ϕ-correlation score higher than 0.06. Therefore, the ϕ-correlation scores reported in the case study (see Table S4 and Table S5 for details) are relatively high compared with the ϕ-correlation scores of all disease pairs. Moreover, 6 out of 10 associations are supported by the GWAS data, e.g., rheumatoid arthritis shares 8 genes with DM according to GWAS data. Apart from the above, associations between DM and the remaining diseases listed in the table are considered as novel associations predicted by the topology-based measure. We evaluated the top 14 novel associations via mining the literature on PubMed[6] (see Table S4 for details). We are able to confirm all of these associations, including surprising associations such as DM and 'other cerebral degenerations' (ICD-9 code: 331). This result highlights the power of our approaches to identify novel associations between diseases. Further exploration of potential underlying mechanisms shared by these diseases may lead to improvement in disease diagnosis, prognosis and treatment.

Another case study (Parkinson's disease, ICD-9 code: 332) can be found in the supplemental material.

Conclusions

In this study, we gained novel insights into the relationship between human diseases by considering their molecular causes and underlying physical interactions. We used information derived from latest biological data, including disease-gene associations, gene functions and the topology of the human PPI network in our analysis. We applied three different measures to estimate the similarity score of diseases, and these measures were systematically evaluated against ICD-9 classification system,

[6] http://www.ncbi.nlm.nih.gov/pubmed

a statistical measure of comorbidity and GWAS data. Our results showed the correlation between associations predicted by our measures and known disease associations, and we also demonstrated the use of our measures in discovering novel disease associations and validated it via literature curation.

Novel disease associations uncovered in this study can be further used to improve our understanding of disease classification. For example, a human disease network that models the relationship of diseases can be constructed based on these similarity measures, and computational approaches, such as clustering, can be applied to detect communities in the disease network. This may provide the opportunity to redefine the current disease classification and further lead to improvements in disease diagnosis, prognosis and treatment.

List of Abbreviations

GWAS: Genome-Wide Association Studies; OMIM: Online Mendelian Inheritance in Man; ICD: International Classification of Diseases; PPI: Protein-Protein Interaction; GO: Gene Ontology; IEA: Inferred from Electronic Annotation; CTD: Comparative Toxicogenomics Database; FunDO: Functional Disease Ontology; HuGENet: Human Genome Epidemiology Network; MeSH: Medical Subject Heading; GeneRIF: Gene Reference Into Function; DO: Disease Ontology; SNP: Single-Nucleotide Polymorphism; ROC: Receiver Operating Characteristic; TPR: True Positive Rate; FPR: False Positive Rate; AUC: Area Under Curve; DM: Diabetes Mellitus.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Designed the experiments: KS, JG, CL, NP. Collected the data: KS, JG. Performed the experiments: KS. Interpreted the results: KS, JG, CL, NP. Wrote the manuscript: KS, CL, NP. Conceived and directed the research: NP. All authors read and approved the final manuscript.

Acknowledgements

We thank members of GlaxoSmithKline (GSK) Computational Biology group, specifically Dr. Hannah Tipney and Dr. Peter Woollard for their helpful comments. This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212, the National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) OIA-1028394, ARRS project J1-5454, GSK Research & Development Ltd., and the Serbian Ministry of Education and Science Project III44006.

Author details

[1]Department of Computing, Imperial College London, London, SW7 2AZ, UK. [2]Computational Biology, GlaxoSmithKline, Stevenage, Hertfordshire, SG1 2NY, UK.

References

1. Gatza, M.L., Lucas, J.E., Barry, W.T., Kim, J.W., Wang, Q., Crawford, M.D., Datto, M.B., Kelley, M., Mathey-Prevot, B., Potti, A., *et al.*: A pathway-based classification of human breast cancer. Proceedings of the National Academy of Sciences **107**(15), 6994–6999 (2010)
2. Loscalzo, J., Kohane, I., Barabasi, A.-L.: Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Molecular systems biology **3**(1) (2007)
3. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A.: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research **33**(suppl 1), 514–517 (2005)
4. van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G., Leunissen, J.A.: A text-mining analysis of the human phenome. European journal of human genetics **14**(5), 535–542 (2006)
5. Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., *et al.*: A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature biotechnology **25**(3), 309–316 (2007)
6. Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabasi, A.-L.: The human disease network. Proceedings of the National Academy of Sciences **104**(21), 8685–8690 (2007)
7. Li, Y., Agarwal, P.: A pathway-based view of human diseases and disease relationships. PloS one **4**(2), 4346 (2009)

8. Hu, G., Agarwal, P.: Human disease-drug network based on genomic expression profiles. PLoS One **4**(8), 6536 (2009)
9. Suthram, S., Dudley, J.T., Chiang, A.P., Chen, R., Hastie, T.J., Butte, A.J.: Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS computational biology **6**(2), 1000662 (2010)
10. Mathur, S., Dinakarpandian, D.: Finding disease similarity based on implicit semantic similarity. Journal of biomedical informatics **45**(2), 363–371 (2012)
11. Žitnik, M., Janjić, V., Larminie, C., Zupan, B., Pržulj, N.: Discovering disease-disease associations by fusing systems-level molecular data. Scientific Reports **3** (2013)
12. Huang, W., Wang, P., Liu, Z., Zhang, L.: Identifying disease associations via genome-wide association studies. BMC bioinformatics **10**(Suppl 1), 68 (2009)
13. Kim, S., Sohn, K.-A., Xing, E.P.: A multivariate regression approach to association analysis of a quantitative trait network. Bioinformatics **25**(12), 204–212 (2009)
14. Lewis, S.N., Nsoesie, E., Weeks, C., Qiao, D., Zhang, L.: Prediction of disease and phenotype associations from genome-wide association studies. PloS one **6**(11), 27175 (2011)
15. Lee, D.-S., Park, J., Kay, K., Christakis, N., Oltvai, Z., Barabási, A.-L.: The implications of human metabolic network topology for disease comorbidity. Proceedings of the National Academy of Sciences **105**(29), 9880–9885 (2008)
16. Milenković, T., Memišević, V., Bonato, A., Pržulj, N.: Dominating biological networks. PloS one **6**(8), 23016 (2011)
17. Janjić, V., Pržulj, N.: The core diseasome. Molecular BioSystems **8**(10), 2614–2625 (2012)
18. Ideker, T., Sharan, R.: Protein networks in disease. Genome research **18**(4), 644–652 (2008)
19. Barabási, A.-L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature Reviews Genetics **12**(1), 56–68 (2011)
20. Bauer-Mehren, A., Bundschus, M., Rautschka, M., Mayer, M.A., Sanz, F., Furlong, L.I.: Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. PloS one **6**(6), 20284 (2011)
21. Becker, K.G., Barnes, K.C., Bright, T.J., Wang, S.A.: The genetic association database. Nature genetics **36**(5), 431–432 (2004)
22. Oti, M., Huynen, M.A., Brunner, H.G.: Phenome connections. Trends in Genetics **24**(3), 103–106 (2008)
23. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., *et al.*: A human protein-protein interaction network: a resource for annotating the proteome. Cell **122**(6), 957–968 (2005)
24. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., *et al.*: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature **440**(7084), 637–643 (2006)
25. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic acids research **34**(suppl 1), 535–539 (2006)
26. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene Ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)
27. Davis, A.P., King, B.L., Mockus, S., Murphy, C.G., Saraceni-Richards, C., Rosenstein, M., Wiegers, T., Mattingly, C.J.: The comparative toxicogenomics database: update 2011. Nucleic acids research **39**(suppl 1), 1067–1072 (2011)
28. Osborne, J., Flatow, J., Holko, M., Lin, S., Kibbe, W., Zhu, L., Danila, M., Feng, G., Chisholm, R.: Annotating the human genome with Disease Ontology. BMC genomics **10**(Suppl 1), 6 (2009)
29. Yu, W., Clyne, M., Khoury, M.J., Gwinn, M.: Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. Bioinformatics **26**(1), 145–146 (2010)
30. Park, J., Lee, D.-S., Christakis, N.A., Barabási, A.-L.: The impact of cellular networks on disease comorbidity. Molecular systems biology **5**(1) (2009)
31. Milenković, T., Memišević, V., Ganesan, A.K., Pržulj, N.: Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. Journal of The Royal Society Interface **7**(44), 423–437 (2010)
32. Sarajlić, A., Janjić, V., Stojković, N., Radak, D., Pržulj, N.: Network topology reveals key cardiovascular disease genes. PloS One **8**(8), 71537 (2013)
33. Memišević, V., Milenković, T., Pržulj, N.: Complementarity of network and sequence structure in homologous proteins. Journal of Integrative Bioinformatics **7**(3), 135 (2010)
34. Pržulj, N., Corneil, D.G., Jurisica, I.: Modeling interactome: scale-free or geometric? Bioinformatics **20**(18), 3508–3515 (2004)
35. Pržulj, N.: Biological network comparison using graphlet degree distribution. Bioinformatics **23**(2), 177–183 (2007)
36. Milenkoviæ, T., Pržulj, N.: Uncovering biological network function via graphlet degree signatures. Cancer informatics **6**, 257 (2008)
37. Milenković, T., Ng, W.L., Hayes, W., Pržulj, N.: Optimal network alignment with graphlet degree vectors. Cancer Informatics **9**, 121 (2010)
38. Ho, H., Milenković, T., Memišević, V., Aruri, J., Pržulj, N., Ganesan, A.K.: Protein interaction network topology uncovers melanogenesis regulatory network components within functional genomics datasets. BMC Systems Biology **4**(1), 84 (2010)
39. Kuchaiev, O., Milenković, T., Memišević, V., Hayes, W., Pržulj, N.: Topological network alignment uncovers biological function and phylogeny. Journal of the Royal Society Interface **7**(50), 1341–1354 (2010)
40. Kuchaiev, O., Pržulj, N.: Integrative network alignment reveals large regions of global network similarity in yeast and human. Bioinformatics **27**(10), 1390–1396 (2011)

41. Hayes, W., Sun, K., Pržulj, N.: Graphlet-based measures are suitable for biological network comparison. Bioinformatics **29**(4), 483–491 (2013)
42. Davis, D.A., Chawla, N.V.: Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. PloS one **6**(7), 22670 (2011)
43. Park, S., Yang, J.-S., Kim, J., Shin, Y.-E., Hwang, J., Park, J., Jang, S.K., Kim, S.: Evolutionary history of human disease genes reveals phenotypic connections and comorbidity among genetic diseases. Scientific reports **2** (2012)
44. Hidalgo, C.A., Blumm, N., Barabási, A.-L., Christakis, N.A.: A dynamic network approach for the study of human phenotypes. PLoS computational biology **5**(4), 1000353 (2009)
45. Hirschhorn, J.N., Daly, M.J.: Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics **6**(2), 95–108 (2005)
46. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., Hirschhorn, J.N.: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics **9**(5), 356–369 (2008)
47. Hindorff, L.A., Junkins, H.A., Mehta, J., Manolio, T., et al.: A catalog of published genome-wide association studies (2011)
48. Sanseau, P., Agarwal, P., Barnes, M.R., Pastinen, T., Richards, J.B., Cardon, L.R., Mooser, V.: Use of genome-wide association studies for drug repositioning. Nature biotechnology **30**(4), 317–320 (2012)
49. Han, J.-D.J., Dupuy, D., Bertin, N., Cusick, M.E., Vidal, M.: Effect of sampling on topology predictions of protein-protein interaction networks. Nature biotechnology **23**(7), 839–844 (2005)
50. Chavali, S., Barrenas, F., Kanduri, K., Benson, M.: Network properties of human disease genes with pleiotropic effects. BMC systems biology **4**(1), 78 (2010)
51. Samreen, R.: Diabetes mellitus: A review. Sci. Res. Essay **4**(5), 367–373 (2009)

Figures

Figure 1 The overlap of datasets. The overlap of diseases (denoted by 'D'), genes (denoted by 'G') and their associations (denoted by 'A') between the four disease-gene association datasets we analysed. Boxes on the left list the sizes of the datasets. The size of the intersection of the datasets is marked in bold.

Figure 2 Evaluation against comorbidity. ROC curves obtained by evaluating the three disease similarity measures against comorbidity. Due to space limitations, only ROC curves of FunDO are shown here (see Figure S5 for ROC curves of other datasets). The ϕ-correlation threshold was set to 0.06 (the same threshold was used in [44]). We evaluated diseases annotated with at least 1, 3, 5, 7, 10, 15 genes, shown by curves with different colours in each plot.

Tables

Additional Files

Additional file 1 — Supplementary Information

The Supplementary Information file contains all additional figures and tables mentioned in the manuscript.

Table 1 Comparison of studies on inferring disease association. The comparison is based on the data used to derive associations (denoted by 'Data'), number of diseases evaluated (denoted by 'Size') and benchmarks used for evaluation (denoted by 'Evaluation'). The number of diseases evaluated in our study is computed as the union of diseases annotated in the four disease-gene association datasets we analysed, given in Figure 1.

	Data	Size	Evaluation
van Driel *et al.* (2006) [4]	OMIM	5132 phenotypes in OMIM	Comparing results with genotypic similarities
Lage *et al.* (2007) [5]	OMIM	7000 OMIM record pairs	Evaluating results against the overlap of the OMIM record pairs
Goh *et al.* (2007) [6]	OMIM	1284 OMIM diseases	Analysing network topological properties
Huang *et al.* (2009) [12]	GWAS	7 diseases	Comparing results with phenotypic similarities
Li and Agarwal (2009) [7]	Pubmed abstracts, biological pathways	1028 diseases in MeSH	Comparing results with MeSH classification
Kim *et al.* (2009) [13]	GWAS	53 clinical traits related to severe asthma	Mining the literature manually
Hu and Agarwal (2009) [8]	Expression data	645 diseases in MeSH	Comparing results with MeSH classification
Suthram *et al.* (2010) [9]	Expression data, PPI	54 diseases	Evaluating results against genetic similarities
Lewis *et al.* (2011) [14]	GWAS	61 diseases	Comparing results with Huang *et al.* (2009) results
Mathur and Dinakarpandian *et al.* (2007) [10]	DO annotation, GO annotation	36 diseases (for evaluation)	Evaluating results using 68 curated disease associations
Our study	Disease-gene associations, GO annotation, PPI	543 ICD-9 diseases	Evaluating results against ICD-9 classification, comorbidity, and genetic similarities derived from GWAS data

Table 2 Evaluation of our measures against ICD-9 classification. Numbers in the table are similarity scores between diseases from the same ICD-9 categories, compared with those from different ICD-9 categories. P-values are calculated by using the Mann−Whitney U test.

Data	Group	Annotation-based	Function-based	Topology-based
OMIM	Same	0.0114 ± 0.0665	0.0355 ± 0.0892	0.4349 ± 0.1101
	Different	0.0010 ± 0.0139	0.0118 ± 0.0314	0.3996 ± 0.0760
	P-value	1.2785×10^{-13}	1.0423×10^{-52}	2.1257×10^{-54}
CTD	Same	0.0361 ± 0.1590	0.0728 ± 0.1754	0.4863 ± 0.1770
	Different	0.0050 ± 0.0274	0.0333 ± 0.0662	0.4408 ± 0.1368
	P-value	1.4887×10^{-23}	1.4040×10^{-9}	2.0240×10^{-25}
FunDO	Same	0.0418 ± 0.1344	0.0991 ± 0.1611	0.5560 ± 0.2214
	Different	0.0100 ± 0.0262	0.0549 ± 0.0830	0.4952 ± 0.1636
	P-value	1.7609×10^{-144}	9.6708×10^{-100}	2.7037×10^{-90}
HuGENet	Same	0.0931 ± 0.1798	0.2470 ± 0.2123	0.8031 ± 0.2248
	Different	0.0438 ± 0.0566	0.1881 ± 0.1522	0.7837 ± 0.2292
	P-value	1.4585×10^{-74}	9.9053×10^{-72}	4.5910×10^{-14}
Intersection	Same	0.0338 ± 0.1511	0.0593 ± 0.1907	0.3826 ± 0.1131
	Different	0.0024 ± 0.0329	0.0089 ± 0.0428	0.3496 ± 0.1020
	P-value	2.2667×10^{-2}	2.7448×10^{-4}	5.4716×10^{-4}
Union	Same	0.0350 ± 0.1179	0.0963 ± 0.1463	0.5680 ± 0.2226
	Different	0.0085 ± 0.0219	0.0583 ± 0.0818	0.5042 ± 0.1716
	P-value	1.3493×10^{-211}	7.1478×10^{-113}	4.1709×10^{-141}

Table 3 Evaluation of our measures against comorbidity. Numbers in the table are AUC values obtained by evaluating the three disease similarity measures against comorbidity associations. The ϕ-correlation threshold was set to 0.06 (the same threshold was used in [44]), and all diseases annotated with at least 3 genes were evaluated. Average AUC values obtained by using randomised scores are shown by numbers in brackets (standard deviations are not shown in the table due to space limitation). Each evaluation test was run 30 times to compute the statistics reported in the table.

Data	Annotation-based	Function-based	Topology-based
OMIM	0.8009 ± 0.0277 (0.5740)	0.8694 ± 0.0073 (0.5120)	0.8495 ± 0.0011 (0.5044)
CTD	0.7849 ± 0.0164 (0.5404)	0.7316 ± 0.0046 (0.5047)	0.7949 ± 0.0042 (0.5203)
FunDO	0.7426 ± 0.0088 (0.4672)	0.7142 ± 0.0017 (0.4940)	0.7497 ± 0.0016 (0.5031)
HuGENet	0.7563 ± 0.0001 (0.5084)	0.8185 ± 0.0001 (0.4987)	0.7153 ± 0.0015 (0.4922)
Intersection	0.9925 ± 0.0001 (0.6013)	0.9802 ± 0.0001 (0.5081)	0.9958 ± 0.0041 (0.4664)
Union	0.8225 ± 0.0045 (0.4704)	0.7491 ± 0.0001 (0.4999)	0.7939 ± 0.0022 (0.5008)
Average	0.8194 ± 0.0837 (0.5270)	0.8106 ± 0.0930 (0.5029)	0.8163 ± 0.0907 (0.4979)

Table 4 Evaluation of our measures against GWAS. Numbers in the table are AUC values obtained by evaluating the three disease similarity measures against disease associations derived from highly confident GWAS data. Only diseases annotated with at least 3 genes were evaluated. 'F/G' are diseases having associated genes in both FunDO and GWAS data (99 diseases in total). 'Common' are diseases having associated genes in all four disease-gene association datasets (given in Figure 1) and GWAS data (50 diseases in total). Average AUC values obtained by using randomised scores are shown by numbers in brackets (standard deviations are not shown in the table due to space limitation). Each evaluation test was run 30 times to compute the statistics reported in the table.

Data	Annotation-based	Function-based	Topology-based
F/G	0.7224 ± 0.0010 (0.4945)	0.6781 ± 0.0001 (0.4968)	0.6863 ± 0.0009 (0.5005)
Common	0.7527 ± 0.0010 (0.4926)	0.7147 ± 0.0001 (0.5005)	0.7555 ± 0.0020 (0.4951)

Table 5 List of the top 10 diseases associated with DM. The topology-based measure was used as the similarity measure, and FunDO was used as the source of disease-gene associations. Only diseases annotated in all four disease-gene association datasets are listed in the table. For a disease associated with DM according to ICD-9, comorbidity or GWAS, we added the supported evidence to the reference (the last column). The remaining disease associations were validated via mining the literature on PubMed, and for each disease only one reference (shown by PubMed ID) was listed in the table due to space limitation.

Rank	Code	Disease name	Reference
1	239	Neoplasms of unspecified nature	PMID: 23639840
2	155	Malignant neoplasm of liver and intrahepatic bile ducts	GWAS
3	710	Diffuse diseases of connective tissue	GWAS
4	714	Rheumatoid arthritis and other inflammatory polyarthropathies	GWAS
5	256	Ovarian dysfunction	ICD-9, GWAS
5	278	Overweight, obesity and other hyperalimentation	ICD-9, comorbidity, GWAS
7	401	Essential hypertension	Comorbidity
8	295	Schizophrenic disorders	PMID: 17474808
9	282	Hereditary hemolytic anemias	GWAS
10	289	Other diseases of blood and blood-forming organs	PMID: 11727971