Computational Methods for Identifying Similar Diseases

Liang Cheng, Hengqiang Zhao, Pingping Wang, Wenyang Zhou, Meng Luo, Tianxin Li, Junwei Han, Shulin Liu, and Qinghua Jiang

Although our knowledge of human diseases has increased dramatically, the molecular basis, phenotypic traits, and therapeutic targets of most diseases still remain unclear. An increasing number of studies have observed that similar diseases often are caused by similar molecules, can be diagnosed by similar markers or phenotypes, or can be cured by similar drugs. Thus, the identification of diseases similar to known ones has attracted considerable attention worldwide. To this end, the associations between diseases at the molecular, phenotypic, and taxonomic levels were used to measure the pairwise similarity in diseases. The corresponding performance assessment strategies for these methods involving the terms "category-based," "simulated-patient-based," and "benchmark-data-based" were thus further emphasized. Then, frequently used methods were evaluated using a benchmark-data-based strategy. To facilitate the assessment of disease similarity scores, researchers have designed dozens of tools that implement these methods for calculating disease similarity. Currently, disease similarity has been advantageous in predicting noncoding RNA (ncRNA) function and therapeutic drugs for diseases. In this article, we review disease similarity methods, evaluation strategies, tools, and their applications in the biomedical community. We further evaluate the performance of these methods and discuss the current limitations and future trends for calculating disease similarity.

INTRODUCTION

Human disease is one of the permanent aspects of the human condition, similar to birth, aging, and death, from a philosophical point of view. The search for novel understanding of disease never stops. Although, currently, there has been great success with the development of biotechnology, the molecular basis of and therapeutic agents for most diseases remain unclear. Current studies have observed that similar diseases are often caused by similar molecules, can be diagnosed by similar markers or phenotypes, and are also cured by similar drugs. Based on this, novel functional molecules for a disease could, in theory, be revealed using prior knowledge of similar diseases. Thus, research on identifying the similarity between diseases has attracted increasing attention.

A pair of diseases with a high similarity score can be defined as being similar diseases. To measure disease similarity, prior knowledge of diseases plays a crucial role. The symptoms and signs accompanying diseases, also called phenotypes, are the intuitive characteristics of a disease. As early as 2004, Frieden and Propping used phenotypes sourced from the Online Mendelian Inheritance in Man (OMIM) website to calculate the similarity of OMIM diseases. With an ever-increasing number of phenotypes being observed by the biomedical community, abundant algorithms have been developed for measuring disease similarity at a phenotypic level.

Many studies have shown that the alterations of molecules can lead to the occurrence of diseases. Thus, the exploration of a common molecular basis is another way to measure disease similarity. With the development of next-generation sequencing technologies, a vast number of protein-coding genes (PCGs) and noncoding RNA (ncRNA) genes associated with diseases have been identified. For example, hemophilia A is an X-linked recessive bleeding disorder caused by a deficiency in the activity of coagulation factor VIII (F8), which can be affected by variations in the F8 genes. MicroRNA (miRNA)-155 is an endogenous ncRNA that regulates several mRNAs to cause B cell lymphomas. Based on the molecular basis of diseases, a large number of methods have been designed for calculating disease similarity, using this as a metric.

Recently, disease taxonomy has begun to play an important role in measuring disease similarity. One of the typical taxonomic classifiers for diseases is Disease Ontology (DO). In this, each disease term represents a disease with different names, and two terms can be linked on the basis of a set of inclusive relationships. For example,
Alzheimer’s disease can be linked to tauopathy. All of the disease terms and the set of inclusion relationships forms the disease hierarchy and directed acyclic graph (DAG) of DO (Figure 1), where a node represents a disease term, and an edge is a set of inclusive relationships between the two terms.35

Currently, dozens of methods have been designed for calculating disease similarity based on prior disease knowledge at the phenotypic, molecular, and hierarchical levels. In this article, we review the main topics of investigation in disease similarity, including the proper selection of proper data, the design and implementation of methods, the evaluation of a method’s performance, and even the application of existing methods for predicting molecular factors of diseases.

DATA SOURCES
Three types of data sources, including disease vocabularies, disease annotations, and gene functional annotations, are widely utilized for calculating disease similarity (Table 1). Here, we list and introduce these main data sources.

Disease Vocabularies
Disease vocabularies document disease terms for distinguishing between different diseases. Each disease term in a vocabulary contains a unique identifier, preferred disease name, synonyms, abbreviations, and the definition of a disease. Parts of these vocabularies even provide a hierarchy of disease terms based on a set of inclusive relationships.

OMIM
The OMIM22,36 is a comprehensive, authoritative compendium of genetic diseases, which is freely available and updated daily. It was initiated in the early 1960s by Dr. Victor A. McKusick and has been developed for online usage by the NCBI since 1985.

MeSH
The Medical Subject Headings (MeSH)37,38 provides hierarchically organized terminology for indexing and cataloging biomedical information for PubMed. MeSH divides all biomedical terms into 16 categories, in which C and F03 contain disease names, containing more than 4,600 disease terms. In addition to the terms in these categories, MeSH also contains supplementary term records, which document thousands of disease terms.

MEDIC
The “merged disease vocabulary” (MEDIC)39 was established by the Comparative Toxicogenomics Database (CTD)40 biocurators and is composed of more than 10,000 unique diseases. To take advantage of the familiarity and immediate genetic data offered by OMIM terms, as well as the navigation utility and PubMed indexing feature of MeSH terms, MEDIC integrates OMIM terms with MeSH terms and hierarchical relationships.

UMLS
The Unified Medical Language System (UMLS)41 is a repository of biomedical vocabularies developed by the U.S. National Library of Medicine (NLM). The UMLS integrates over 2 million names for some 900,000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations between these concepts. Vocabularies integrated in the UMLS Metathesaurus include MeSH, OMIM, Gene Ontology (GO),42 and so forth.

DO
The Disease Ontology (DO) database34 was developed to create a single structure for the classification of diseases that unifies the representation of disease between varied vocabularies into a relational ontology. DO terms can be linked in a hierarchy by a type of semantic association called an “IS_A” relationship43 (Figure 1). The initial builds of DO in 2003 and 2004 used the International Classification of Diseases (ICD-9)44 as the foundational vocabulary. Recent revisions have improved this with the reorganization of DO based on UMLS disease terms in conjunction with term mappings to Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT)45,46 and ICD-9. The current version of DO is organized into eight main classes to represent cellular proliferation, mental health, anatomical entity, infectious, and agent, etc.

Disease Annotations
The molecular basis and phenotypic characterization of a disease are two main aspects of prior knowledge often used for measuring disease

Figure 1. Sub-graph of the DO Hierarchy for Alzheimer’s Disease
Arrows represent an “IS_A” relationship for DO. For example, “Alzheimer’s disease” is linked to “Dementia” by an “IS_A” relationship. All of the terms that can be linked by “IS_A” relationships in the graph from “Alzheimer’s disease” are the ancestors of “Alzheimer’s disease.” All of the terms that can link to “Disease” by “IS_A” relationships are the descendants of “Disease.”
miRNAs are a class of endogenous single-stranded small ncRNAs that play a crucial role in various human diseases by negatively regulating the expression of PCGs. They have the ability to affect the expression of PCGs through competitively binding with miRNAs. Thus, it becomes important to understand the role of miRNAs in diseases. According to the theory of competing endogenous RNA (ceRNA), they can affect the expression of PCGs through competitively binding with miRNAs. Thus, it becomes important to understand the role of miRNAs in diseases.

### Disease Annotations of lncRNAs

Long ncRNAs (lncRNAs) are mRNA-like transcripts that are longer than 200 nt and have little or no protein-coding capacity. According to the theory of competing endogenous RNA (ceRNA), they can affect the expression of PCGs through competitively binding with miRNAs. Thus, it becomes important to understand the role of lncRNAs in diseases. The LncRNADisease database has a manually accumulated set of relationships between lncRNAs and diseases.

### Disease Annotations of Phenotypes

Phenotypes are documented in the Clinical Synopsis section of the textual descriptions of each OMIM disease. Robinson et al. extracted all of the phenotypes from this text and constructed a human phenotype ontology (HPO) to annotate human diseases.

### Integrated Resources of Disease Annotations

In previous efforts, we developed two integrated resources for disease annotations. integrated resource for annotating human genes with multi-level ontologies (OAHG) focused on the disease annotations of PCGs, miRNAs, and lncRNAs; and a semantically integrated database towards a global view of human disease (SIDD) documented disease-related molecular, phenotypic, and environmental features. The data sources integrated by OAHG involved OMIM, GAD, CTD, LncRNADisease, and HPO. SIDD integrated up to 18 different data sources, including OMIM, GAD, CTD, LncRNADisease, and HPO.

### Gene Functional Annotations

Similar molecular foundations of diseases may be influenced not only by common genes but also by different genes with common functions. Recently, associations between genes from gene functional annotation resources have been introduced for calculating disease similarity. Here, we list resources for the identification of gene functional annotations.

### GOA

Disease-related PCGs can possess similar molecular functions (MFs), and may be involved in similar biological processes (BPs). This type of functional association of genes often exposes the similarity of different diseases. The GO annotation (GOA) of PCGs provides assignments of MF and BP terms of GO to gene products, in a project run by the European Bioinformatics Institute (EBI).

### HumanNet

In addition to the GOA of PCGs, functional relationships between disease-related genes can also be reflected by protein-protein interactions, mRNA co-expression, and so forth. By integrating all of this data, HumanNet provides a more comprehensive relative score of pairwise PCG relationship.

### Disease Similarity Measures

The similarity between diseases can be reflected by their common phenotypic characteristic, molecular basis, and hierarchy structures. Therefore, we have classified the disease similarity methods into phenotype-based, molecule-based, hierarchy-based, and hybrid methods (Table 2).

---

### Table 1. Summary of Data Sources

| Category and Name | Creation Date | Initiator | PMID         |
|------------------|---------------|-----------|--------------|
| OMIM             | 1960s         | McKusick  | 17357067     |
| MeSH             | 1960s         | Winifred Sewell | 14119288   |
| UMLS             | 1980s         | Olivier Bodenreider | 14681409  |
| SNOMED-CT        | 2001          | Wang et al. | 11825284    |
| DO               | 2003          | Schriml et al. | 22080554   |
| MEDIC            | 2012          | Davis et al. | 22434833   |

**Disease Annotations**

- GeneRIF: 2007, 17990498
- CTD: 2003, 27651457
- GAD: 2004, 15118671
- miR2Disease: 2008, 18927107
- HPO: 2008, 18950739
- SpliceDisease: 2011, 22139928
- LncRNADisease: 2012, 23758614
- HMDD v2.0: 2013, 24194601
- SIDD: 2013, 24146757
- OAHG: 2016, 27703231

**Gene Functional Annotations**

- GOA: 2003, 12654719
- HumanNet: 2011, 21536720
Phenotype-Based Methods

Figure 2 shows the schematic process of phenotype-based methods. First, qualitative associations between phenotypes and diseases are extracted from phenotype data sources. Then, each pair of qualitative associations is quantified as a disease-phenotype score or phenotype-phenotype score. Finally, these scores are utilized for calculating disease similarity.

Freudenberg’s Method

OMIM diseases were originally attributed manually by Freudenberg and Propping according to their phenotypic appearance, using the indices “periodicity,” “etiology,” “tissue,” “age of onset,” and “mode of inheritance.” The index “periodicity” is a Boolean variable, indicating an episodic occurrence of a disease in contrast to a linear progression. The index “etiology” is based on clinical signs and laboratory or pathological findings related to a disease. The index “tissue” is compiled as the anatomic location of phenotype. The index “inheritance” indicates whether a disease is inherited in an autosomal-dominant, autosomal-recessive, X chromosome, mitochondrial, or complex manner. The index “age of onset” refers to the age of a patient when symptoms are generally first noticed. Then, the similarity of diseases $d_1$ and $d_2$ is defined as the following:

$$\text{sim}(d_1, d_2) = \sum_{i=1}^{5} w_i \cdot \text{sim}(d_1.\text{index}_i, d_2.\text{index}_i),$$  \hspace{1cm} (Equation 1)

where $w_i$ represents the contribution of a single index to the total similarity score, and $\text{sim}(d_1.\text{index}_i, d_2.\text{index}_i)$ indicates the similarity between the $i$th indexes of $d_1$ and $d_2$.

van Driel’s Method

van Driel et al.\textsuperscript{67} calculated the similarity between over 5,000 diseases based on phenotypic features of OMIM records. For each OMIM disease, its phenotypic descriptions were extracted from “TX” and “CS” fields. Then, the OMIM diseases and phenotypic descriptions were mapped to the anatomy (category A) and the disease (category C) sections of MeSH to establish disease-term associations. Each disease-term association was then defined as a vector with three features as follows:

$$f_1(t_1, d_1) = \frac{\text{counted}(t_1, d_1) + \text{descendant}(t_1)}{\text{descendant}(t_1, d_1)},$$  \hspace{1cm} (Equation 2)

$$f_2(t_1, d_1) = \log_2 \frac{N}{n_{t1}},$$  \hspace{1cm} (Equation 3)

and

$$f_3(t_1, d_1) = 0.5 + \frac{\text{counted}(t_1, d_1)}{\max_{s=1}^{5} \text{counted}(t_s, d_1)},$$  \hspace{1cm} (Equation 4)

where $t_1$ and $d_1$ represent a phenotype term and a disease, respectively. In Equations 2 and 4, $\text{counted}(t_1, d_1)$ means the occurrence number of $t_1$ in the OMIM records of $d_1$. In Equation 3, $N$ is the total number of records analyzed, and $n_{t1}$ is the number of records that contain the term $t_1$. In Equation 4, $\text{descendant}(t_1)$ is the number of descendant terms in the hierarchy of MeSH, and $\text{descendant}(t_1, d_1)$ is the number of descendant terms in the OMIM records of $d_1$. The
similarity between diseases $d_1$ and $d_2$ is then defined as Equation 5 below:

$$sim(d_1, d_2) = \frac{\sum_{i=1}^{m} t_{i1} \cdot t_{i2}}{\sqrt{\sum_{i=1}^{m} t_{i1}^2} \cdot \sqrt{\sum_{i=1}^{m} t_{i2}^2}},$$

(Equation 5)

where $t_{i1}$ and $t_{i2}$ mean the $i$th term vector of $d_1$ and $d_2$, respectively; and $m$ is the total number of phenotypic terms.

**Freudenberg’s Method**

Phenotypic terms of the “CS” field of OMIM records were also manually extracted to construct an HPO by Freudenberg. Then, the similarity of pairwise phenotypic terms was calculated based on Resnik’s method as follows:

$$sim(p_1, p_2) = \max_{a \text{ ancestor}(p_1, p_2)} \log \frac{N}{n(a)},$$

(Equation 6)

where $a$ is the ancestor of phenotypes $p_1$ and $p_2$, $N$ is the total number of genes associated with the phenotypes, and $n(a)$ is the number of genes associated with $a$. Then, the similarity of pairwise diseases $d_1$ and $d_2$ is defined as follows:

$$sim(d_1 > d_2) = \frac{\sum_{i=1}^{n} \max_{1 \leq j \leq m} sim(p_i, p_j)}{n},$$

(Equation 7)

and

$$sim(d_1, d_2) = \frac{sim(d_1 > d_2) + sim(d_2 > d_1)}{2},$$

(Equation 8)

where $n$ and $m$ represent the number of phenotypes associated with $d_1$ and $d_2$, respectively.

**Zhang’s Method**

Zhang et al. extracted phenotypic terms from the “TX” and “CS” fields of OMIM’s disease records using a MetaMap transfer tool. As a result, each disease could be represented as a set of phenotypes. Then the weights of phenotypic terms for diseases were calculated based on a term frequency-inverse document frequency (TF-IDF) weighting scheme. Subsequently, each disease was represented as a weighted vector of these phenotypic terms. Finally, the similarity of pairwise diseases was defined as the cosine of their corresponding phenotypic vectors.

**Zhou’s Method**

Zhou et al. define a disease as a set of symptoms, which were extracted from PubMed. Each disease was described as a weighted vector of phenotypic terms. Here the weight was calculated by a TF-IDF weighting scheme. The similarity of a pairwise disease was then defined as the cosine of their vectors.

**Chen’s Method**

Chen et al. extracted the disease-phenotype relationships from the UMLS file MRREL.RRF where disease-phenotype relationships were documented based on OMIM, Ultrasound Structured Attribute Reporting, and Minimal Standard Digestive Endoscopy Terminology. This group then used the information content (IC) to weight each phenotype concept as follows:

$$w_i = \log \frac{N}{n_i},$$

(Equation 9)
where $N$ is the total number of diseases, and $n_i$ is the number of diseases associated with a phenotype $p_i$. Then they modeled the phenotype similarity of pairwise diseases by the cosine of their feature vectors.

**Molecule-Based Methods**

The schematic process of molecule-based methods is analogous to that of the previously stated phenotype-based methods. Here, genes are the mainly disease-related molecules. Phenotypic-based methods always utilized the semantics associations between phenotypes. In comparison, genes can be associated in more ways, such as in terms of protein-protein interactions (PPIs), co-expression, and so forth.

**Mathur’s Method**

SwissProt contains proteins that have been manually annotated with diseases, which were mapped to DO terms using MetaMap by Mathur and Dinakarpandian. Then, the similarity of diseases $d_1$ and $d_2$ was calculated based on their corresponding genes as follows:

$$sim(d_1, d_2) = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|},$$

(Equation 10)

where $G_1$ and $G_2$ are gene sets of diseases $d_1$ and $d_2$, respectively, $|$ is the number of terms in the specified set, and $N$ is the total number of genes.

**Suthram’s Method**

Suthram et al. compared diseases using an integrated analysis of disease-related mRNA expression data and the human protein interaction network. First, they identified conserved functional modules of genes using PathBLAST based on PPI data from the Human Protein Reference Database (HPRD). Next, they normalized the gene expression data in each microarray sample using a Z-score transformation and computed the activity level of each gene in a disease. Then, the module response score for each module in a disease was assigned to be the mean of the gene activity score of its component genes. Finally, they calculated the partial correlation coefficient between diseases based on the corresponding module response score and defined it as the disease similarity.

**Gottlieb’s Method**

Gottlieb et al. presented four algorithms for calculating disease similarity using the genetic signatures of diseases from gene expression experiments, which involved signature-based, signature sequence-based, signature PPI-based, and signature GO-based methods. The signature-based method utilized a Jaccard index between every pair of disease signatures to calculate disease similarity as follows:

$$sim_{pss}(d_1, d_2) = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|},$$

(Equation 11)

where $G_1$ and $G_2$ are the signatures of diseases $d_1$ and $d_2$, respectively, and $|$ is the number of terms in the specified set.

The signature PPI-based method calculated the distances between each pair of disease signatures based on their corresponding proteins using an all-pairs shortest paths algorithm on the human PPI network. Distances were transformed into similarity values using the following formula:

$$sim_{ppe}(d_1, d_2) = Ae^{-D(P_1, P_2)},$$

(Equation 12)

where $P_1$ and $P_2$ are the corresponding proteins of diseases $d_1$ and $d_2$, respectively, and $D(P_1, P_2)$ is the shortest path between these proteins in the PPI network. $A$ is a parameter chosen to be $0.9 \times e$ by Perlman et al.

The signature sequence-based method calculated the Smith-Waterman sequence alignment score between disease signatures and then divided the score by the geometric mean of the scores from aligning each sequence against itself. In addition, the signature GO-based method calculated the similarity between each pair of disease signatures based on their corresponding GO terms.

**Hamaneh’s Method**

Hamaneh and Yu devised a network-based measure to calculate disease similarity. First, they assigned weights to all proteins by using information flow from a disease to the human PPI network and back. As a result, each disease was represented as a weighted vector whose dimension is the number of proteins in the network. Then, the similarity of two diseases was defined as the cosine of the angle between their corresponding vectors.

**Kim’s Method**

Kim et al. extracted disease-gene pairs and disease-drug pairs from the literature and used the frequencies of co-occurrence relationships as features to calculate disease similarity. In this work, disease names, gene symbols, and drug names were from the Pharmacogenomics Knowledgebase (PharmGKB). This assumes that $G_1$ and $G_2$ are genes that occurred in the same sentence as diseases $d_1$ and $d_2$, respectively. $D_1$ and $D_2$ are drugs that occurred in the same sentence as diseases $d_1$ and $d_2$, respectively. The similarity of $d_1$ and $d_2$, therefore, can be defined as the following:

$$sim(d_1, d_2) = \frac{MI_c(d_1, d_2) + MI_d(d_1, d_2)}{2},$$

(Equation 13)

and

$$MI_c(d_1, d_2) = \frac{|G_1 \cap G_2|}{|N|} \times \log\frac{|G_1 \cap G_2|}{N},$$

(Equation 14)

and

$$MI_d(d_1, d_2) = \frac{|D_1 \cap D_2|}{|M|} \times \log\frac{|D_1 \cap D_2|}{M},$$

(Equation 15)

where $N$ and $M$ are the total number of genes and drugs, respectively.

**Hierarchy-Based Methods**

Hierarchy-based approaches are based only on the hierarchical structure of disease-related ontologies. In the previously mentioned
studies, multiple methods have been presented for calculating the similarity of ontology terms using shared path and distance based on hierarchical structures. However, currently only Wang’s method is widely utilized for calculating disease similarity.

**Wang’s Method**

Assuming that \( D_i \) is the set including \( d_i \) and all of its ancestor terms in an ontology-based “IS_A” relationship, the hierarchical contribution of the terms \( d \) to \( d_i \) is represented as follows:

\[
S_{d_i}(t) = \begin{cases} 
1 & d = d_i \\
\max\{w \cdot S_{d_i}(d') | d' \in d_i \} & d \neq d_i 
\end{cases}
\]  

(Equation 16)

where \( w \) is a hierarchical contribution factor for hierarchical association. According to Wang et al. and Cheng et al., \( w \) is defined as 0.5 for an “IS_A” relationship of DO. Then, the value of the summation of all of the hierarchical contributions of \( D_i \) to \( d_i \) is \( SV(d_i) \), which is defined as follows:

\[
SV(d_i) = \sum_{d \in D_i} S_{d_i}(d). 
\]  

(Equation 17)

Assuming that \( D_j \) is the set including \( d_j \) and all of its ancestor terms, the similarity between \( d_i \) and \( d_j \) is defined by Wang’s method as follows:

\[
\text{Sim}_{Wang}(d_i, d_j) = \frac{\sum_{d \in D_i \cap D_j} (S_{d_i}(d) + S_{d_j}(d))}{SV(d_i) + SV(d_j)} 
\]  

(Equation 18)

**Mabotuwana et al.’s Method**

Mabotuwana et al. defined similarity of pairwise terms as inversely proportional to the distance between terms, as follows:

\[
\text{Sim}(d_i, d_j) = \frac{1}{d} 
\]  

(Equation 19)

where \( d \) is the number of nodes in the shortest path between two diseases based on the DAG of ontology.

**Hybrid Methods**

Molecular and hierarchical associations between diseases have been combined as hybrid methods for calculating disease similarity. These methods often utilize disease-related genes to define the IC of diseases as follows:

\[
\text{IC}(d) = \log_{2} \frac{n_d}{N} 
\]  

(Equation 20)

where \( N \) denotes the total number of genes, and \( n_d \) represents the number of genes of \( d \). Here, disease-related genes are often based on OMIM, CTD, SIDD, OAHG, and so on.

**Resnik’s Method**

Early in 1995, Resnik presented a method for calculating the similarity between ontology terms. In 2002, this method was introduced for calculating the similarity between GO terms. In 2011, Li et al. utilized this method for calculating the similarity between DO terms. According to Resnik’s method, the similarity of pairwise diseases \( d_i \) and \( d_j \) equals the IC of the most informative common ancestor (MICA) of these two diseases as follows:

\[
\text{sim}_{\text{Resnik}}(d_i, d_j) = \text{IC}(\text{MICA}). 
\]  

(Equation 21)

**Lin’s Method**

Concerned that the similarity between ontology terms should also be decided by the IC of the two terms, Lin improved Resnik’s method in 1998. According to Lin’s method, the similarity of pairwise diseases \( d_i \) and \( d_j \) can be reflected by both the MICA of the disease pair and the IC of each disease as follows:

\[
\text{sim}(d_i, d_j) = \frac{2 \cdot \text{IC}(\text{d}_{\text{MICA}})}{\text{IC}(d_i) + \text{IC}(d_j)}. 
\]  

(Equation 22)

**Schlicker’s Method**

Schlicker et al. improved Resnik’s method from the same perspective as Lin, and they defined disease similarity as follows:

\[
\text{sim}(d_i, d_j) = \max_{d_{\text{ancestors}}(d_i, d_j)} \left( \frac{2 \cdot \text{IC}(d)}{\text{IC}(d_i) + \text{IC}(d_j)} \cdot \left(1 - \frac{n_d}{N}\right) \right).
\]  

(Equation 23)

In this equation, \( d_{\text{ancestors}}(d_i, d_j) \) represents the common ancestor of diseases \( d_i \) and \( d_j \).

**Mathur’s Method**

In 2012, Mathur et al. designed a new method named PSB for calculating the similarity between DO terms. According to this method, the significance of related BPs terms from GO should be computed for each disease using a hypergeometric test. Assuming that \( d_i \) and \( d_j \) can be associated with \( m \) and \( n \) BP terms, respectively, the similarity of \( d_i \) and \( d_j \) is defined as follows:

\[
\text{sim}(d_i, d_j) = \frac{1}{2} \left( \sum_{i=1}^{m} \max_{j \leq m} \left( \frac{\text{Sim}(p_{i1}, p_{j1})}{m} \right) + \sum_{i=1}^{n} \max_{j \leq n} \left( \frac{\text{Sim}(p_{i2}, p_{j2})}{n} \right) \right).
\]  

(Equation 24)

where \( \text{Sim}(p_i, p_j) \) represents the similarity between two BPs \( p_{i1} \) and \( p_{j2} \) as follows:

\[
\text{Sim}(p_i, p_j) = \frac{1}{2} \cdot \left( \text{IC}_{\text{GO}}(p_i) + \text{IC}_{\text{GO}}(p_j) \cdot \frac{n(p_i \cap p_j)}{n(p_i \cup p_j)} \cdot \frac{\text{IC}_{\text{GO}}(p_i)}{\text{Max}(\text{IC}_{\text{GO}})} \cdot \frac{\text{IC}_{\text{GO}}(p_j)}{\text{Max}(\text{IC}_{\text{GO}})} \right).
\]  

(Equation 25)
Here, $IC_{GO}$ and $IC_{DO}$ represent the IC based on GO and DO, respectively. $n(p_1 \cap p_2)$ and $n(p_1 \cup p_2)$ denote the number of common genes of $p_1$ and $p_2$, and the number of total genes of $p_1$ and $p_2$, respectively.

**Cheng’s Method**

In addition to related BP, genes can be associated by PPI, co-expression, and so forth. Therefore, Cheng et al.\(^91\) presented the SemFunSim method to improve Mathur’s method by incorporating the gene functional network from HumanNet,\(^66\) which reflects the comprehensive gene associations from PPI, co-expression, BP, and so on. This assumes that $G_1$ and $G_2$ represent related gene sets of $d_1$ and $d_2$, respectively. Then, the similarity between $t_2$ and $t_2$ by Cheng et al.’s\(^91\) method is described by the following:

$$Sim_{SemFunSim}(t_1, t_2) = \frac{\sum_{i=1}^{m} \max_{1 \leq j \leq n} \left( Sim \left( g_{1i}, g_{2j} \right) \right) + \sum_{i=1}^{n} \max_{1 \leq j \leq m} \left( Sim \left( g_{2i}, g_{1j} \right) \right)}{m + n} \left| G_{MICA} \right| \left| G_{MICA} \right|$$

(Equation 26)

where $|G_{MICA}|$ represents the number of genes of MICA for $t_1$ and $t_2$ and $m$ and $n$ denote the number of genes in $G_1$ and $G_2$, respectively. $Sim(g_{1i}, g_{2j})$ is the functional similarity score between genes $g_{1i}$ and $g_{2j}$ from HumanNet.\(^66\)

**PERFORMANCE EVALUATION**

The performance of a disease similarity method can be affected by the quality of the prior knowledge it is based on. Most of the methods that utilize a manually curated dataset is high reliability. Some of the methods mentioned here use data from the literature extracted using text-mining tools. Data obtained in an unsupervised way should always be evaluated. In Mathur’s method,\(^77\) disease-related genes were mined from literature using MetaMap.\(^70\) The recall and precision were calculated based on a benchmark dataset from Monttaz et al.,\(^100\) which contained 200 records that were manually annotated by experts. The identified similarity pairs of diseases should always be then evaluated to measure the performance of the method used. Three types of classical evaluation strategies are introduced here (Figure 3).

**Simulated-Patient-Based Strategy**

In consideration of the difficulty in obtaining phenotypic information about a large number of patients, Sebastian et al.\(^48\) presented a simulated-patient-based method to evaluate their phenotype-based disease similarity method. We used 44 complex dysmorphology syndromes for which adequate frequency phenotypes were available, and then 100 virtual patients for each disease were generated on the basis of the frequency of phenotypes among persons diagnosed with a certain disease. For example, to generate patients with phenotypes A and B, in which A occurs in 40% and B occurs in 60% of patients, a random number generator was utilized to generate two random numbers uniformly distributed between 0 and 100. Subsequently, the similarity of the simulated patient to each of the OMIM diseases was calculated and then ranked. The average rank of all of the patients was returned to assess the performance of the original method.

**Term-Category-Based Strategy**

Sun et al.\(^101\) utilized information on disease-related molecules to design a disease similarity measurement method. Their results were evaluated using the disease classification terminologies found in the ICD-9. Their assumption was that two similar diseases should be subjected to the same categories in the ICD-9. Therefore, the correlation between the similarity of diseases and their classifications can reflect the performance of this method. Since similarity scores are not normally distrib-
SemFunSim were 0.6209, 0.6351, 0.6849, 0.8843, and 0.9849, respectively.

The performance of these methods is subject to the prior knowledge they used. Wang’s method only used the entire structure of the ontology; therefore, its performance is limited by the comprehensive structure of the ontology. Although Resnik’s and Lin’s methods incorporated the structure of ontology and ontology annotation, they do not utilize all the “IS_A” relationships of ontology. Thus, the performance of these three methods is not very good. In comparison with Resnik’s and Lin’s methods, PSB introduced GOA for associating disease-related genes. Thus, its performance improved a lot. Since disease-related genes could be associated in terms of PPIs, co-expression, and so on, the performance of PSB is improved much more by the SemFunSim method.

APPLICATIONS

Disease similarity can be determined at the molecular, phenotypic, and hierarchical levels. Conversely, similar diseases reflect the correlations of their inducing molecules, phenotypes, and classifications. Therefore, disease similarity has been widely applied in the functional prediction of molecules, clinical diagnosis, and the establishment of disease associations.

The Functional Prediction of Molecules

This is based on the observation that genes causing similar diseases tend to lie close to one another in a network of PPI.90-105 Vanunu et al.106 constructed a comprehensive network using gene-disease association, disease similarity, and PPI data to predict disease-related PCGs using a random walk method.106

In comparison with PCGs, it is not easy to determine the function of ncRNAs due to limited knowledge with regard to their impact on proteins from wet lab experiments with these ncRNAs. Fortunately, disease similarity has been useful for this in previous investigations.90,107-110 Based on prior knowledge of the associations between ncRNAs and diseases, functional similarity of ncRNAs can be calculated based on the similarities of their related diseases to construct a network in which an ncRNA is represented as a node and the similarity of pairwise ncRNAs is represented as edges.89 Just such a network was then utilized for predicting novel ncRNA-disease associations by the random walk with restart (RWR) method.106,108,109

Recently, disease similarity has been utilized for mining potential therapeutic drugs for diseases. Based on the observation that similar diseases can often be treated with similar drugs, Cheng et al.91,111 prioritized potential drugs for a disease based on their results with similar diseases. Gottlieb et al.8 combined disease similarity and drug similarity to predict novel drug indications.

Clinical Diagnosis

The diagnosis process can be a challenging undertaking, given the large number of hereditary disorders and the range of partially overlapping clinical features associated with them. To resolve this problem, Robinson et al.5 established an HPO to calculate the disease similarity and diagnose diseases according to clinical phenotype. According to Equations 6, 7, and 8, disease similarity can be
calculated based on their phenotype sets. For an individual patient, the similarity between OMIM diseases and clinical features could also be calculated based on this method. The similarity score in this case then reflects the probability of a potential disease in the patient.

Construction of Qualitative Associations of Diseases
In 2006, Goh et al.112 utilized the common genetic origin of diseases to construct a human disease network (HDN) from the molecular level based on OMIM. This was an early study that established a qualitative association between diseases from a quantitative perspective. A portion of each disease stems not as the consequence of the single genetic defects but, rather, the breakdown in molecular interaction networks. Thus, their associations cannot be reflected by this network. Therefore, the network was extended based on PPIs, metabolic networks, and different pathways.113–115

Recently, Zhou et al.72 established an HDN at the phenotypic level, where the link weight between two diseases quantified the disease similarity. Here, the symptoms of diseases were extracted from literature in PubMed. Each disease was described as a vector of phenotypes. Then, the similarity between diseases was defined as the cosine similarity of their vectors.

TOOLS FOR CALCULATING DISEASE SIMILARITY
Inspired by the wide recent application of machine learning methods in bioinformatics,116–118 various algorithms have been implemented for calculating disease similarity using R and web-based programs67,80,97,111,119–124 (Table 3). These tools play important roles in disease diagnosis, the prediction of drugs, and so forth. Here, we introduce four frequently used tools in detail.

MimMiner
van Driel et al.67 designed a phenotype-based method and implemented it as a tool—namely, MimMiner—for calculating the similarity of OMIM diseases. This tool provides interfaces to query the similar diseases related to an input diseases and is widely used in bioinformatics community. It should be noted that this tool needs to be updated due to the rapid increase in the size of the OMIM disease database.

Phenomizer
Phenomizer is an online tool that can be helpful in the diagnosis processes and is based on disease similarity.68 Currently, thousands of genetic disorders characterized by specific combinations of phenotypic features are documented in OMIM. The diagnosis process based on phenotypes is difficult without computer-based tools. Phenomizer allows an automatic correlation between phenotypic abnormalities and hereditary disorders found in OMIM. The p values are generated to evaluate the statistical significance of those correlation scores given by Phenomizer. This tool is also useful for suggesting additional possible phenotypic alterations for further evaluation in a patient of interest.

DOSim
DOSim is an R package used for computing the similarity between DO terms97 based on Wang’s method35 and nine hybrid methods involving Resnik’s method, Lin’s method, and so forth.93–95,98,125–127. This tool also implements utilities to calculate the similarity of genes based on their inducing diseases and conduct DO enrichment analysis.

DisSim
DisSim111 is an online system for exploring similar diseases in DO. It provides both the similarity of pairwise diseases and the significance of their similarity score. In addition, the system integrates therapeutic drugs for known diseases to predict potential drugs for other human diseases based on the observation that similar diseases can be treated with similar drugs.78

DISCUSSION
Most disease similarity methods depend on disease vocabularies and their annotations. Phenotype-based methods extract disease annotations of phenotypes from PubMed and OMIM. Disease names from these data sources are from MeSH and OMIM. Hierarchy-based
methods utilize the structure of ontology from MeSH and DO. Current molecule-based methods mainly used the DO annotations of genes. In summary, DO, MeSH, and OMIM contain the most frequently used vocabularies for calculating disease similarity. However, not all disease terms are contained in any one of these vocabularies. For comparison, OMIM documents more specific disease terms, such as TYPE III SYNDACTYL Y (OMIM: 186100). MeSH and DO involve classification of diseases, such as cancer (DOID: 162). Figure 5 shows the number of disease terms distributed across the different vocabularies. In total, 958 common disease terms are documented in DO, MeSH, and OMIM, which covers 8.8%, 8.5%, and 11.4% of DO, MeSH, and OMIM terms, respectively. Although OMIM and MeSH terms have been integrated into MEDIC, MEDIC lacks many DO terms and disease classifications. Therefore, combining all of the disease terms of DO, MeSH, and OMIM is critical for calculating disease similarity using the same vocabulary. In addition, a unified disease annotation database based on this integrated vocabulary is indispensable for improving the universality of similarity determining algorithms. In our previous studies, we provided a global view of human diseases by annotating disease-related molecule and phenotype features with DO. However, the absence of disease terms in DO limits its applications.

Disease-related ontologies only contain “IS_A” relationships, which limits the performance of hierarchy-based methods. For example, Wang’s method could be applied to multiple term associations of ontology, such as “IS_A,” “PART_OF,” “LOCATE_IN,” and so on. The performance evaluation results in Figure 4 shows that Wang et al.’s method could be improved, which may be achieved with the occurrence of more types of disease associations than the “IS_A” relationship.

Data quality and the quantity of disease annotations of phenotypes and molecules are crucial for the performance of molecule-based, phenotype-based, and hybrid-based methods. OMIM documents close but few disease-gene associations. Contrary to this, GeneRIF and SIDD retain loose but abundant associations. All of these datasets were combined together without distinction for calculating disease similarity in most cases. These methods could be improved by ranking all of the associations. For example, we can improve the disease annotations by adding the evidence for each disease-gene association such as that found in the GOA database.

In general, newer methods should consider more types of prior knowledge, leading to better performance. Wang’s method, which is a hierarchy-based method, was presented in 2007. The SemFunSim method was presented in 2014, and it incorporates the hierarchical structure of DO, disease annotations of genes, and gene associations. The evaluation results in Figure 4 show that SemFunSim achieves a higher AUC than Wang’s method. Although hybrid methods integrate more types of prior knowledge of diseases, molecular and phenotypic associations of diseases were ignored. Therefore, it is possible that the performance of disease similarity methods could be further improved by fusing more disease knowledge types.

Although comprehensive knowledge benefits the calculative precision of disease similarity, these methods based on a single type of prior knowledge can also very valuable for biological applications. Diseases are often caused by the molecular mechanism and could be reflected by diverse phenotypes. Disease phenotypes can be detected from clinical diagnosis, while causal molecules are identified from wet labs. Gaps in phenotypic and molecular levels exist for understanding diseases. Here, disease similarity based on different types of knowledge could bridge the gap.

The purpose of calculating disease similarity is to identify similar diseases. However, it is not easy to determine similar diseases directly from most of the presented methods and tools. One feasible strategy...
for this purpose is provided here by DisSim,\textsuperscript{111} which provides the $p$ values for each similarity score. According to current methods, the similarity of pairwise diseases can be obtained, which are then normalized to $Z$ scores. Then, the one-side $p$ values are calculated as a significance score for each similarity score. Another way to provide $p$ values for similarity scores would be a permutation test.

Disease similarity plays important roles in mining the novel molecular features of diseases, clinical diagnosis, and so on. The exploration of the function of ncRNAs is a long-term challenge, as these RNAs do not produce proteins. Currently, disease similarity has been successful in predicting the function of ncRNAs, especially in prioritizing miRNA-disease\textsuperscript{6,128–133} and lncRNA-disease pairs.\textsuperscript{86,108} In the future, these methods can be used for comprehending the function of other types of ncRNAs, such as circular RNA (circRNAs).\textsuperscript{134} In a previous study, disease similarity was utilized for diagnosis based on phenotypes.\textsuperscript{86} This may also be helpful for molecular diagnosis. Alterations in the presence of metabolites are easily determined in the clinical, meaning metabolite-disease pairs can be prioritized based on disease similarity methods. Therefore, it is theoretically possible to predict potential diseases based on abnormalities in metabolite levels.

**AUTHOR CONTRIBUTIONS**

L.C., J.H., S.L., and Q.I. conceived and designed the experiments. L.C., H.Z., P.W., W.Z., M.L., and T.L. analyzed data. L.C. wrote the manuscript. All authors read and approved the final manuscript.

**CONFLICTS OF INTEREST**

The authors declare no competing interests.

**ACKNOWLEDGMENTS**

We thank LetPub (https://www.letpub.com) for its linguistic assistance during the preparation of the manuscript. This work was supported by the National Natural Science Foundation of China (grant nos. 61871160 and 61502125); the Heilongjiang Postdoctoral Fund (grant nos. LBH-TZ20 and LBH-Z15179); and the China Postdoctoral Science Foundation (grant nos. 2018T110315 and 2016M590291).

**REFERENCES**

1. Aerts, S., Lambrechts, D., Maity, S., Van loos, P., Coessens, B., De Smet, F., Tranchevent, L.G., De Moor, B., Marynen, P., Hassan, B., et al. (2006). Gene prioritization through genomic data fusion. Nat. Biotechnol. 24, 537–544.
2. Franke, L., van Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M., and Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am. J. Hum. Genet. 78, 1011–1025.
3. Chavali, S., Barrenas, F., Kanduri, K., and Benson, M. (2010). Network properties of human disease genes with pleiotropic effects. BMC Syst. Biol. 4, 78.
4. Robinson, P.N., and Mundlos, S. (2010). The human phenotype ontology. Clin. Genet. 77, 525–534.
5. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet. 83, 610–615.
6. Tang, W., Wan, S., Yang, Z., Tschendorff, A.E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics 34, 398–406.
7. Yu, L., Ma, X., Zhang, L., Zhang, J., and Gao, L. (2016). Prediction of new drug indications based on clinical data and network modularity. Sci. Rep. 6, 32530.
8. Gottlieb, A., Stein, G.Y., Ruppin, E., and Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. 7, 496.
9. Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F.X., and Pan, Y. (2016). Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. Bioinformatics 32, 2664–2671.
10. Yu, L., Su, R., Wang, B., Zhang, L., Zou, Y., Zhang, J., and Gao, L. (2017). Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk. IEEE/ACM Trans. Comput. Biol. Informatics 14, 966–977.
11. Yu, L., Wang, B., Ma, X., and Gao, L. (2016). The extraction of drug-disease correlations based on module distance in incomplete human interactome. BMC Syst. Biol. 10 (Suppl 4), 111.
12. Chen, X., and Huang, L. (2017). LRRSLMDA: Laplacian Regularized Sparse Subspace Learning for miRNA-Disease Association prediction. PLoS Comput. Biol. 13, e1005912.
13. Chen, W., Peng, P., Ding, H., and Lin, H. (2018). Classifying included and excluded exons in exon skipping event using histone modifications. Front. Genet. 9, 433.
14. Lai, H.Y., Feng, C.Q., Zhang, Z.Y., Tang, H., Chen, W., and Lin, H. (2018). A brief survey of machine learning application in cancerletin identification. Curr. Gene Ther. 18, 257–267.
15. Chen, X., and Yan, G.Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics 29, 2617–2624.
16. Jiang, L., Xiao, Y., Ding, Y., Tang, J., and Guo, F. (2019). Discovering cancer subtypes via an accurate fusion strategy on multiple profile data. Front. Genet. 10, 20.
17. Yu, L., Huang, J., Ma, Z., Zhang, J., Zou, Y., and Gao, L. (2015). Inferring drug-disease associations based on known protein complexes. BMC Med. Genomics 8 (Suppl 2), S2.
18. Wang, L., Ping, P.Y., Kuang, L.N., Ye, S.T., Lebal, F.M.R., and Pei, T.R. (2018). A novel approach based on bipartite network to predict human microbe-disease associations. Curr. Bioinform. 13, 141–148.
19. Albuisson, J., Isidor, B., Giraud, M., Pichon, O., Marsaud, T., David, A., Le Caignec, C., and Bezieau, S. (2011). Identification of two novel mutations in Shh long-range regulator associated with familial pre-axial polydactyly. Clin. Genet. 79, 371–377.
20. Gurnett, C.A., Bowcock, A.M., Dietz, F.R., Morcuende, J.A., Murray, J.C., and Dobbs, M.B. (2007). Two novel point mutations in the long-range SHH enhancer in three families with triphalangeal thumb and preaxial polydactyly. Am. J. Med. Genet. A. 143A, 27–32.
21. Freudenberg, J., and Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 18 (Suppl 2), S110–S115.

22. Amberger, J., Bocchini, C., and Hamosh, A. (2011). A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Hum. Mutat. 32, 564–567.

23. Mannucci, P.M., and Tuddenham, E.G. (2001). The hemophilies—from royal genes to therapy. N. Engl. J. Med. 344, 1773–1779.

24. Mazurier, C., Parquet-Gernez, A., Gaucher, C., Lavergne, J.M., and Goudemand, J. (2002). Factor VIII deficiency not induced by FVIII gene mutation in a female first cousin of two brothers with haemophilia A. Br. J. Haematol. 119, 390–392.

25. Klimmer, J., Poppema, S., de Jong, D., Blokzijl, T., Harms, G., Jacobs, S., Kroesen, B.J., van den Berg, A. (2005). BIC and miR-155 are highly expressed in Hodgkin, primary mediastinal and diffuse large B cell lymphomas. J. Pathol. 207, 243–249.

26. Eis, P.S., Tam, W., Sun, L., Chadburn, A., Li, Z., Gomez, M.F., Lund, E., and Eisen, H.N. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 18 (Suppl 10), 911.

27. Chen, X., Wang, L., Qu, J., Guan, N.N., and Li, J.Q. (2018). Predicting miRNA-disease association based on inductive matrix completion. Bioinformatics 34, 4256–4265.

28. Chen, X., Sun, Y.Z., Guan, N.N., Qu, J., Huang, Z.A., Zhu, Z.X., and Li, J.Q. (2019). Computational models for lncRNA function prediction and functional similarity calculation. Brief Funct. Genomics 18, 58–82.

29. Schraml, L.M., Arce, C., Nađenda, S., Chang, Y.W., Mazaitis, M., Felix, V., Feng, G., and Köhler, W.A. (2012). Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 40, D940–D946.

30. Wang, J.Z., Du, Z., Payatakeel, R., Yu, P.S., and Chen, C.F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics 23, 1274–1281.

31. McKusick, V.A. (2007). Mendelian Inheritance in Man and its online version, OMIM. Am. J. Hum. Genet. 80, 588–604.

32. Lowe, H.J., and Barnett, G.O. (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA 271, 1103–1108.

33. Sewell, W. (1964). Medical subject headings in MEDLARS. Bull. Med. Libr. Assoc. 52, 164–170.

34. Davis, A.P., Wiegars, T.C., Rosenstein, M.C., and Mattingly, C.J. (2012). MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. Database (Oxford) 2012, bar065.

35. Davis, A.P., Grondin, C.J., Johnson, R.L., Sciaky, D., King, B.L., McMorran, R., Wiegars, J., Wiegars, T.C., and Mattingly, C.J. (2017). The Comparative Toxicogenomics Database: update 2017. Nucleic Acids Res. 45 (D1), D972–D978.

36. Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, D267–D270.

37. 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. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29.

38. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kuhn, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., and Rosse, C. (2005). Relations in biomedical ontologies. Genome Biol. 6, R46.

39. Deyo, R.A., Cherkin, D.C., and Ciol, M.A. (1992). Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J. Clin. Epidemiol. 45, 613–619.

40. Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. Stud. Health Technol. Inform. 121, 279–290.

41. Wang, A.Y., Barrett, J.W., Bentley, T., Markwell, D., Price, C., Spackman, K.A., and Searns, M.Q. (2001). Mapping between SNOMED RT and Clinical Terms version 3: a key component of the SNOMED CT development process. Proc. AMIA Symp 2001, 741–745.

42. Mitchell, J.A., Aronson, A.R., Mork, J.G., Folk, L.C., Humphrey, S.M., and Ward, J.M. (2003). Gene indexing: characterization and analysis of NLM’s GeneRIFs. AMIA Annu. Symp. Proc 2003, 460–464.

43. Becker, K.G., Barnes, K.C., Bright, T.J., and Wang, S.A. (2004). The genetic association database. Nat. Genet. 36, 431–432.

44. Wang, J., Zhang, J., Zhao, W., and Cui, Q. (2012). SpliceDisease database: linking RNA splicing and disease. Nucleic Acids Res. 40, D1055–D1059.

45. Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281–297.

46. Chen, Y., Yang, X., Xu, Y., Cao, J., and Chen, L. (2017). Genomic analysis of drug resistant small cell lung cancer cell lines by combining miRNA and mRNA expression profiling. Oncol. Lett. 13, 4077–4084.

47. Chen, X., Xie, D., Zhao, Q., and You, Z.H. (2019). MicroRNAs and complex diseases: from experimental results to computational models. Brief. Bioinform. 20, 515–539.

48. Chen, X., Yin, J., Qu, J., and Huang, L. (2018). MDHG: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. PLoS Comput. Biol. 14, e1006418.

49. Jiang, Q., Wang, Y., Hao, Y., Juin, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. 37, D98–D104.

50. Li, Y., Qu, C., To, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. 42, D1070–D1074.

51. Mercer, T.R., Dingier, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. Nat. Rev. Genet. 10, 155–159.

52. Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Zhou, W., Liu, G., Jiang, H., and Jiang, Q. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of IncRNAs in human and mouse. Nucleic Acids Res. 47 (D1), D140–D144.

53. Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfo, P.P. (2011). A cmRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell J6, 353–358.

54. Voucžičević, D., Schrewe, H., and Orom, U.A. (2014). Molecular mechanisms of long ncRNAs in neurological disorders. Front. Genet. 5, 48.

55. Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., and Cui, Q. (2013). LncRNA2Disease: a database for long non-coding RNA associated diseases. Nucleic Acids Res. 41, D983–D986.

56. Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). OAHC: an integrated resource for annotating human genes with multi-level ontologies. Sci. Rep. 6, 34820.

57. Cheng, L., Wang, G., Li, J., Zhang, T., Xu, P., and Wang, Y. (2013). SIDD: a semantically integrated database towards a global view of human disease. PLoS ONE 8, e75504.

58. Camon, E., Magrane, M., Barrett, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) database: sharing knowledge in UniProt with Gene Ontology. Nucleic Acids Res. 32, D262–D266.

59. Orita, C., and Vihinen, M. (2009). Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. Nucleic Acids Res. 37, 622–628.

60. Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249–255.
66. Lee, I., Blom, U.M., Wang, P.L., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 21, 1109–1121.

67. van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G., and Leunissen, J.A. (2006). A text-mining analysis of the human phenotype. Eur. J. Hum. Genet. 14, 535–542.

68. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am. J. Hum. Genet. 85, 457–464.

69. Zhang, S., Wu, C., Li, X., Chen, X., Jiang, W., Geng, B.S., Li, J., and Yan, Y.Q. (2010). From phenotype to gene: detecting disease-specific gene functional modules via a text-based human disease phenotype network construction. FEBS Lett. 584, 3635–3643.

70. Aroonsan, A.R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc. AMIA Symp. 2001, 17–21.

71. Wilbur, W.J., and Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Comput. Biol. Med. 26, 209–222.

72. Zhou, X., Menche, J., Barabási, A.L., and Sharma, A. (2014). Human symptoms-disease network. Nat. Commun. 5, 4212.

73. Chen, Y., Zhang, X., Zhang, G.Q., and Xu, R. (2015). Comparative analysis of a novel disease phenotype network based on clinical manifestations. J. Biomed. Inform. 53, 113–120.

74. Bell, D.S., Greenes, R.A., and Doubitel, P. (1992). Form-based clinical input from a structured vocabulary: initial application in ultrasound reporting. Proc. Annu. Symp. Comput. Appl. Med. Care 1992, 789–790.

75. Tringali, M., Hole, W.T., and Srinivasan, S. (2002). Integration of a standard gastrointestinal endoscopy terminology in the UMLS Metathesaurus. Proc. AMIA Symp. 2002, 801–805.

76. UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 38, D142–D148.

77. Mathur, S., and Dinakarpandian, D. (2010). Automated ontological gene annotation for computing disease similarity. Summit Transl. Bioinform 2010, 12–16.

78. Suthram, S., Dudley, J.T., Chiang, A.P., Chen, R., Hastic, T.J., and Butte, A.J. (2010). Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput. Biol. 6, e1000662.

79. Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., et al. (2005). Human protein reference database—2005 update. Nucleic Acids Res. 37, D767–D772.

80. Permam, L., Gottlieb, A., Attias, N., Ruppin, E., and Sharan, R. (2011). Combining drug and gene similarity measures for drug-target elucidation. J. Comput. Biol. 18, 133–145.

81. Hamaneh, M.B., and Yu, Y.K. (2014). Relating diseases by integrating gene associations and information flow through protein interaction network. PLoS ONE 9, e110936.

82. Kim, H., Yoon, Y., Ahn, J., and Park, S. (2015). A literature-driven method to calculate similarities among diseases. Comput. Methods Programs Biomed. 122, 108–122.

83. Thorn, C.F., Sharma, M.R., Altman, R.B., and Klein, T.E. (2017). PharmGKB summary: pazopanib pathway, pharmacokinetics. Pharmacogenet. Genom. 27, 307–312.

84. del Pozo, A., Pazos, F., and Valencia, A. (2008). Defining functional distances over gene ontology. BMC Bioinformatics 9, 50.

85. Wu, H., Su, Z., Mao, F., Olman, V., and Xu, Y. (2005). Prediction of functional modules based on comparative genome analysis and Gene Ontology application. Nucleic Acids Res. 33, 2822–2837.

86. Wu, H., Gao, L., Tu, K., and Guo, Z. (2005). Broadly predicting specific gene functions with expression similarity and taxonomy similarity. Gene 352, 75–81.

87. Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D., and Siani-Rose, M.A. (2004). A knowledge-based clustering algorithm driven by Gene Ontology. J. Biopharm. Stat. 14, 687–700.

88. Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Infering the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics 26, 1644–1650.

89. Cheng, L., Li, J., Pu, P., Peng, J., and Wang, Y. (2014). SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. PLoS ONE 9, e94915.

90. Mahotuwana, T., Lee, M.C., and Cohen-Solal, E.V. (2013). An ontology-based similarity measure for biomedical data—application to radiology reports. J. Biomed. Inform. 46, 857–868.

91. Jiang, J.J., and Comrath, D.W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv, arXiv:cmp-lg/9709098, https://arxiv.org/abs/cmp-lg/9709098.

92. Pesquisa, C., Faria, D., Bastos, H., Falco, A., and Couto, F.M. (2007). Evaluating GO-based semantic similarity measures. Ismb/ecb Sig. Meet. Program Mater. Iscb 37, 37–40.

93. Li, B., Wang, J.Z., Feltus, F.A., Zhou, J., and Luo, F. (2010). Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins. arXiv, arXiv:1001.0958, https://arxiv.org/abs/1001.0958.

94. Lord, P.W., Stevens, R.D., Brass, A., and Goble, C.A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19, 1275–1283.

95. Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., Li, C., Li, X., Rao, S., and Li, X. (2011). DOfSim: an R package for similarity between diseases based on Disease Ontology. BMC Bioinformatics 12, 266.

96. Schlicker, A., Domingues, F.S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 7, 302.

97. Mathur, S., and Dinakarpandian, D. (2012). Finding disease similarity based on implicit semantic similarity. J. Biomed. Inform. 45, 363–371.

98. Mottaz, A., Yip, Y.L., Ruch, P., and Veuthey, A.L. (2008). Mapping proteins to disease terminologies: from UniProt to MeSH. BMC Bioinformatics 9 (Suppl 5), S3.

99. Sun, K., Gonçalves, J.P., Larminie, C., and Przulj, N. (2014). Predicting disease associations via biological network analysis. BMC Bioinformatics 15, 304.

100. Nachar, N. (2008). The Mann-Whitney U: a test for assessing whether two independent samples come from the same distribution. Tutor. Quant. Methods Psychol. 4, 13–20.

101. Pakhomov, S., McNees, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G.B. (2010). Semantic similarity and relatedness between clinical terms: an experimental study. AMIA Annu. Symp. Proc 2010, 572–576.

102. Vazquez, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. 6, e1000641.

103. Ganegoda, G.U., Sheng, Y., and Wang, J. (2015). ProSim: a method for prioritizing disease genes based on protein proximity and disease similarity. BioMed Res. Int. 2015, 213750.

104. Köhler, S., Bauer, S., Horn, D., and Robinson, P.N. (2008). Walking the interactome for prioritization of candidate disease genes. Am. J. Hum. Genet. 82, 949–958.

105. Hu, Y., Zhou, M., Shi, H., Ji, H., Jiang, Q., and Cheng, L. (2016). IndelSim: a novel method for measuring disease similarity based on information flow. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine, T. Tian, Q. Jiang, Y. Liu, K. Burrage, J. Song, Y. Wang, X. Hu, S. Morishita, Q. Zhu, and G. Wang, eds. (BIBM), pp. 20–26.
108. Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., He, W., Hao, D., Liu, S., and Zhou, M. (2014). Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. Mol. Biosyst. 10, 2074–2081.

109. Chen, X., Yan, C.C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. Sci. Rep. 5, 11338.

110. Yu, L., Zhao, J., and Gao, L. (2018). Predicting potential drugs for breast cancer based on miRNA and tissue specificity. Int. J. Biol. Sci. 14, 971–982.

111. Cheng, L., Jiang, Y., Wang, Z., Shi, H., Sun, J., Yang, H., Zhang, S., Hu, Y., and Zhou, M. (2016). DisSim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs. Sci. Rep. 6, 30024.

112. Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proc. Natl. Acad. Sci. USA 104, 8685–8690.

113. Lee, D.S., Park, J., Kay, K.A., Christakis, N.A., Oltravi, Z.N., and Barabasi, A.L. (2008). The implications of human metabolic network topology for disease comorbidity. Proc. Natl. Acad. Sci. USA 105, 9880–9885.

114. Li, Y., and Agarwal, P. (2009). A pathway-based view of human diseases and disease relationships. PLoS ONE 4, e3436.

115. Zhang, X., Zhang, R., Jiang, Y., Sun, P., Tang, G., Wang, X., Lv, H., and Li, X. (2011). The expanded human disease network combining protein–protein interaction information. Eur. J. Hum. Genet. 19, 783–788.

116. Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics 33, 3518–3523.

117. Diao, F.Y., Lv, H., Wang, F., Feng, C.-Q., Ding, H., Chen, W., and Lin, H. (2018). Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics 35, 2075–2083.

118. Feng, C.Q., Zhang, Z.Y., Zhu, X.J., Lin, Y., Chen, W., Tang, H., and Lin, H. (2019). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics 35, 1469–1477.

119. Hoehndorf, R., Schofield, P.N., and Gkoutos, G.V. (2015). Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. Sci. Rep. 5, 10888.

120. Deng, Y., Guo, L., Wang, B., and Guo, X. (2015). HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. PLoS ONE 10, e0115692.

121. Yu, G., Wang, L.G., Yan, G.R., and He, Q.Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics 31, 608–609.

122. Hu, Y., Zhao, L., Liu, Z., Ju, H., Shi, H., Xu, P., Wang, Y., and Cheng, L. (2017). DisSetSim: an online system for calculating similarity between disease sets. J. Biomed. Semantics 8 (Suppl. 1), 28.

123. Hamaneh, M.B., and Yu, Y.K. (2015). DeCoaD: determining correlations among diseases using protein interaction networks. BMC Res. Notes 8, 226.

124. Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics 34, 1953–1956.

125. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 1 (Morgan Kaufmann Publishers), pp. 448–453.

126. Lin, D. (1998). An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning, Vol. 1 (Morgan Kaufmann Publishers), pp. 296–304.

127. Couto, F.M., Silva, M.J., and Coutinho, P. (2005). Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. CIKM ’05 Proceedings of the 14th ACM International Conference on Information and Knowledge Management, 343–344.

128. Li, Y., and Yu, H. (2014). A robust data-driven approach for gene ontology annotation. Database, 2014 (Oxford), p. bau113.

129. Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2016). Similarity computation strategies in the microRNA-disease network: a survey. Brief. Funct. Genomics 15, 55–64.

130. Liu, Y., Zeng, X., He, Z., and Zou, Q. (2017). Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Trans. Comput. Biol. Bioinformatics 14, 905–915.

131. Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018). EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. Cell Death Dis. 9, 3.

132. Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.H., and Liu, H. (2018). BNPMDA: Bipartite Network Projection for MiRNA-Disease Association prediction. Bioinformatics 34, 3178–3186.

133. Chen, X., Yan, C.C., Zhang, X., and You, Z.H. (2017). Long non-coding RNAs and complex diseases: from experimental results to computational models. Brief. Bioinform. 18, 558–576.

134. Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017). A comprehensive overview and evaluation of circular RNA detection tools. PLoS Comput. Biol. 13, e1005420.