Journal of Biomedical Informatics 45 (2012) 363–371

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Finding disease similarity based on implicit semantic similarity

Sachin Mathur, Deendayal Dinakarpandian *

School of Computing & Engineering, University of Missouri – Kansas City, 5100 Rockhill Rd., Kansas City, MO 64110, United States

ARTICLE INFO

Article history:
Received 27 May 2011
Accepted 28 November 2011
Available online 7 December 2011

Keywords:
Ontology terms
Similarity measure
Disease similarity
Semantic similarity
Gene Ontology
Ontology perturbation
Ontology based disease similarity

ABSTRACT

Genomics has contributed to a growing collection of gene–function and gene–disease annotations that can be exploited by informatics to study similarity between diseases. This can yield insight into disease etiology, reveal common pathophysiology and/or suggest treatment that can be appropriated from one disease to another. Estimating disease similarity solely on the basis of shared genes can be misleading as variable combinations of genes may be associated with similar diseases, especially for complex diseases. This deficiency can be potentially overcome by looking for common biological processes rather than only explicit gene matches between diseases. The use of semantic similarity between biological processes to estimate disease similarity could enhance the identification and characterization of disease similarity. We present functions to measure similarity between terms in an ontology, and between entities annotated with terms drawn from the ontology, based on both co-occurrence and information content. The similarity measure is shown to outperform other measures used to detect similarity. A manually curated dataset with known disease similarities was used as a benchmark to compare the estimation of disease similarity based on gene-based and Gene Ontology (GO) process-based comparisons. The detection of disease similarity based on semantic similarity between GO Processes (Recall = 55%, Precision = 60%) performed better than using exact matches between GO Processes (Recall = 29%, Precision = 58%) or gene overlap (Recall = 88% and Precision = 16%). The GO-Process based disease similarity scores on an external test set show statistically significant Pearson correlation (0.73) with numeric scores provided by medical residents. GO-Processes associated with similar diseases were found to be significantly regulated in gene expression microarray datasets of related diseases.

1. Introduction

A disease is usually caused by congenital or acquired mutations, or by the action of external agents that disrupt gene regulation [1]. This disrupts biological processes in which the affected genes participate. Disruption of the biological processes results in phenotypes that characterize each disease, with the phenotype depending on the influence of the affected biological processes on the larger biological network. Single gene diseases such as sickle cell anemia are easier to decode than polygenic diseases that involve multiple variable genes [2]. The latter scenario is analogous to the '*k*-out-of-*n*' model in engineering that is used to build fault-tolerant subsystems. Systems with '*n*' partially redundant components fail only if at least '*k*' components are defective, $k \leqslant n$ [3]. Multifactorial disease can be conceptualized in similar fashion, where a module fails only when '*k*' out of '*n*' genes in the module are mutated/differentially regulated. Since '*k*' can be variable, it is often hard to find a reproducible set of genes across multiple microarray analyses [4] of a disease state. This suggests that representations that summarize the contributions of groups of genes rather than match genes explicitly could be useful in understanding disease. Examples of such representations are biological processes and network modules.

The use of representations (e.g., GO-Processes) other than sets of genes has helped determine signatures in meta-analysis of studies on breast cancer [5]. It can further our understanding of diseases by offering alternative ways to study the similarity between them. Studying disease similarities can yield insight into etiology, reveal common pathophysiology and/or suggest treatment that can be appropriated from one disease to another [6]. A broad analysis of the "diseasome" showed that different diseases in the same disorder class exhibited concordance in protein networks and biological processes [2]. With the high cost of drug development and lower approval rates [7], drug repositioning opportunities can be effectively explored by studying disease similarities. Several diseases previously thought to be distinct have been found to share biological processes either in their etiology or in manifestation of symptoms [8]. Genetic, symptom and phenotype information along with penetrance models has been used to find comorbidity between diseases [9]. Medicare data has been used to study co-occurrence of diseases [10,11]. Microarray gene expression data has been

* Corresponding author. Fax: +1 816 2355159.
E-mail address: dinakard@umkc.edu (D. Dinakarpandian).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.11.017

used to find modules affected by diseases and to find similarity between diseases by measuring the correlation between affected modules [12]. Shared pathways [13] and the gene–phenotype network [14] have also been used to compute similarity between diseases.

A concomitant trend in research has been the increasing use of data annotated with biomedical ontologies. In particular, such datasets can be exploited to reveal relationships between biological entities (Gene Ontology) [15] and disease pathology (UMLS) [16]. Data can be combined across ontologies and similarity between terms quantified by using computational methods. Methods to estimate similarity between ontology terms are based mainly on the co-occurrence of terms in annotation, information content or the ontology graph structure [17]. In turn, semantic similarity between diseases can be assessed by computing the similarity between the sets of associated ontological terms.

Studies on disease similarity often lack systematic assessment and evaluation of the result, in part due to the difficulty in identifying a validation set. In this paper, we present a systematic approach to evaluate the benefit of using an ontological approach for inter-disease similarity. We first present a measure to represent the similarity between terms in an ontology that combines information from both co-occurrence in annotations and ontological hierarchy [18]. Its performance is compared to well-known similarity measures using a subset of KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [19] pathways as the benchmark. We exploit UMLS and open knowledge-sources to augment the existing Disease Ontology (DO) [20] with missing terms and synonyms. We then contrast the performance of a gene-based approach to disease similarity with a biological process-based (GO biological processes) one by using a predefined set of diseases as the benchmark. We further demonstrate the benefits of considering the semantic similarity between GO-Processes rather than only exact process matches. The measure is also used on an external data set previously rated by human experts to verify the accuracy of the predictions. We verify GO-Processes predicted to be involved in each disease using gene expression microarray datasets.

2. Method

The similarity between two diseases is computed as a function of the genes or, alternately, the biological processes associated with them. To be able to do this, existing gene–disease associations were expanded by augmenting the Disease Ontology as described in Section 2.1. The metric used to compute the similarity between two genes or two processes is developed in Section 2.2 and validated in Sections 2.3 and 2.4. The approaches taken to compare diseases, based on the equations in Section 2.2, are described in Section 2.5. The benchmark sets used to evaluate the prediction of disease similarity are described in Section 2.6, and the use of microarray data to evaluate underlying GO processes in Section 2.7.

2.1. Sources of annotated data

Gene–GO Process annotations were downloaded in March 2010. Gene–disease associations were pooled from multiple sources since there is variable coverage of disease terms in different ontologies. For example, though MeSH has broad coverage on a variety of subjects, it has several missing terms and lacks detail in the disease section. SNOMED-CT has a large collection of terms but the focus is on medical billing. ICD has a broad classification of diseases and lacks detail. To overcome this, the Disease Ontology (DO ver. 3) vocabulary was augmented using UMLS (MeSH, SNOMED-CT, ICD9) as described in [21]. The DO consisting of 12,082 terms was augmented with synonyms to a total of 33,085 terms. To increase the amount of annotated data available, disease–gene associations from OMIM, Swissprot and GeneRIF were pooled together. Swissprot records were matched against UMLS using MetaMap. OMIM records were mapped against UMLS AUI records from which Concept Identifiers (CUIs) were extracted. Protein identifiers from Swissprot, gene identifiers from OMIM and GeneRIF identifiers were matched with corresponding NCBI Entrez identifiers. In the final annotation, each DO identifier was annotated with NCBI Entrez identifiers.

2.2. Estimation of semantic similarity

Measures that quantify semantic similarity can be broadly classified as node-based, edge-based and hybrid that combine node and edge-based measures. Node-based measures use the properties of terms, their ancestors and descendents. The most commonly used approach is Information Content, which is defined as the negative logarithm of probability of occurrence of a term in a given corpus. Similarity between a pair of terms is often measured by the information content in the most specific common ancestor. This approach has limited ability to distinguish between similarity among descendants of a pair of terms and similarity between the terms themselves. Further, information content relies on the annotation density of terms, which is biased towards well-studied nodes of the graph. Another node-based approach uses the degree of co-occurrence of attributes of the terms [22]. A vector-based method by Patwardhan and Pedersen [23] is a purely corpus-based method to measure semantic relatedness using second-order context vectors.

Edge-based measures [24] use the structure of the graph (path length) to measure similarity between terms in an ontology. The main assumption is that an edge represents the same semantic distance anywhere in the graph, which is not true as sections of the graph may be finely classified and others only coarsely defined. Hybrid methods such as Wang et al. [25] use a combination of node-based and edge-based measures by computing the contribution of all ancestors. Similarity between a pair of terms is found by weighing the type of edges that connect the terms (IS-A has a greater weight than PART-OF) along with the information content of the nodes.

Existing approaches estimate similarity between a pair of terms by relying either on the structure of the graph or co-annotation, but not both, to measure similarity. Biomedical ontologies like the Gene Ontology (GO), although widely used among the community, are a work in progress. Its structure and use in annotation are frequently updated. Hence, measures relying solely on graph structure or annotation will not be able to capture similarity when either of them is inaccurate. To mitigate this, we propose a similarity measure that takes the graph structure as well as co-occurrence in annotation to quantify similarity between terms.

Given two terms x and y in an ontology (e.g., GO Biological Process), $n(x)$ = number of genes annotated with x, $n(x \cap y)$ = number of genes annotated with both x and y, the extent of co-occurrence in annotation between x and y is captured by the following equation:

$$sc(x,y) = \frac{n(x \cap y)}{n(x \cup y)} \tag{1}$$

This is the essentially the well-known Jaccard Index. It measures the strength of co-annotation between two terms based on their joint use in annotation. Note that this captures evidence from annotation in addition to relatedness that is a consequence of the ontological structure (if a gene is annotated with term x, it is assumed that it is also annotated with all the ontological ancestors of x). For example, the fact that the gene *MMP9* is annotated with the

biological process "GO:0006508: proteolysis" implies that is also annotated with all its ancestral terms, i.e., GO:0008152: metabolic process, GO:0044238: primary metabolic process, and GO:0019538: protein metabolic process. The values of the raw score can range from zero (when x and y are never used together) to 1 in case of identical annotation. Eq. (1) does not distinguish between co-annotation by a pair of related abstract terms and co-annotation by a pair of related specific terms. Ideally, co-annotation by highly specific terms should count towards a higher degree of similarity than co-annotation by abstract concepts. To overcome this limitation, the score is subsequently multiplied by the average information content of the terms. The resultant semantic similarity between x and y is given by the following equation:

$$sim(x,y) = sc(x,y) * Avg(IC(x), IC(y)) \tag{2}$$

where $IC(x)$ is the information content of x estimated as $(-log_2\, p)$, and 'p' is the ratio of the number of genes annotated with term x to the total number of genes annotated with any term. The average information content of the terms is based on the frequency of explicit use in annotation, as well as the implicit frequency of use inferred from the ontological hierarchy. This is combined with the extent of co-occurrence from Eq. (1) to measure similarity between a given pair of terms.

The similarity Mb between two entities (e.g., genes or diseases) A and B that are each annotated with a set of descriptors (e.g., processes or genes) is computed using the following equation:

$$Mb(A,B) = \frac{1}{2}\left[\frac{\sum_{1\leqslant i\leqslant m} msim(T_{Ai}, T_B)}{m} + \frac{\sum_{1\leqslant j\leqslant n} msim(T_{Bj}, T_A)}{n}\right] \tag{3}$$

where A is annotated with m terms, B with n terms, T_{Ai} are terms annotating entity A and $msim(T_{Ai}, T_B)$ is the maximum semantic similarity between the ith term of A and all terms in B (Eq. (2) is used to compute the semantic similarity between each of the m terms describing A and each of the n terms describing B). The similarity score thus obtained is a measure of involvement in similar biological processes. The maximum similarity of the pair of terms is chosen so that the best alignment between the entities (vectors) is obtained or the best-case similarity between entities is measured without loss of information. There are other ways to measure similarity between entities, for example using average or minimum among all pairs of terms, but a recent study has shown that using the maximum similarity is a better approach in a biological context [17].

2.3. Comparison with other similarity measures

We compared the proposed measure Mb with similarity measures representing different approaches: co-occurrence of annotation (PMI), node-based measures like Resnik [26], and Lin [27]; edge-based measure Leacock and Chodrow [24] (lc) and a hybrid measure Wang et al. [25]. Resnik et al. similarity measures were used from GOSemSim [28] implemented as part of the Bioconductor package, while PMI and LC were implemented using R.

Some of the ways to evaluate similarity measures are based on gene co-expression, membership in regulatory pathways, membership in protein families or sequence similarity. We have chosen to base our comparisons on the pathway membership of annotated genes. Genes involved in a common pathway are expected to yield a higher similarity score than the average similarity score within a set of randomly selected genes. A set of pathways from KEGG [19], version December 2010, was used to compare existing similarity measures with the proposed measure. Only genes having GO-Process annotation were included in the analysis. Further, associations inferred from electronic annotation (IEA evidence code) were ignored in order to minimize the presence of false positives. Corresponding p-values were computed with respect to a null model analogous to a previously published study [29]. In a given pathway (n genes), pairwise similarities of $n*(n-1)/2$ combinations between genes were computed using Eq. (3) with GO-Process as the ontology. These pairwise scores were averaged and assigned to each pathway to obtain a pathway score. The membership of KEGG pathways ranges from 3 to over 1000 genes. Pathways whose membership ranged from 5 to 50 were extracted, totaling 71 pathways. A minimum membership of 5 was chosen to ensure meaningful comparison between pathways, and a maximum of 50 was chosen to limit the computation time – computing the null model in particular requires substantial computational resources because of the quadratic complexity (doubling the number of genes takes four times as much to compute). To create a null model for each pathway, 5000 average similarity scores of k randomly selected genes were computed, where k corresponds to the number of genes in the corresponding pathway. The p-value for each pathway was computed as the proportion of null-scores greater than the pathway score.

To check if Mb can discriminate between pathways, the inter-pathway similarity score was obtained by randomly choosing a pair of pathways and calculating the average similarity between the genes of one pathway and the genes of the other. The inter-pathway scores for 71 KEGG pathways were compared with their intra-pathway scores using the Wilcoxon rank-sum test.

2.4. Ontology perturbation

Ontologies are often incomplete and subject to continuous revision, e.g., GO is updated almost daily. Examining the historical log of changes shows, for example, deletion of an existing IS-A relation between the terms GO:0016485 (protein processing) and GO:0043687 (post-translational protein modification), with a new IS-A relation added with GO:0051604 (protein maturation). Subsequently, a new HAS-PART relation has been added with GO:0080120 (CAAX-box protein maturation). GO:0016485 has many descendants. Therefore, removing and adding links can change the location of the entire sub-graph (http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0051605#term=history). Ideally, in addition to being accurate, a similarity measure should be robust to minor perturbations in the ontology.

Since genes involved in the same pathway can be assumed to have collaborative functional relationships, the average gene-similarity scores within pathways can be used to test the robustness of similarity measures. To simulate this, partially altered versions of the GO-Process ontology were created. In the procedure, a pair of nodes was chosen and the nodes along with their descendents were swapped. To avoid drastic transformation of the ontology, nodes that were at a distance of at least 6 from the root node of the GO-Process graph were chosen (so that relatively small subsections of the ontology were swapped). This was repeated a number of times resulting in a perturbed ontology. The amount of perturbation was measured by comparing the pairwise graph matrix (direct or descendent association in "graph_path" GO table) with the "graph_path" table of the perturbed ontology. The pathway similarity scores were measured for each perturbation and compared to that of the unperturbed GO pathway scores by measuring the deviation of the recomputed pathway score from the score based on the unperturbed ontology, normalized by the latter.

$$Deviation = |\, pert - un\_pert \,| \,/ un\_pert$$

where $pert$ = average score of pathways from the perturbed ontology and un_pert = average score of pathways based on the original.

Perturbation results in a change of annotation, with new genes associated to ancestors of swapped GO-terms. The change is proportional to the amount of perturbation of the ontology. For example, if gene G is annotated with terms x and y that have a common

ancestor z, the terms might or might not have a different common ancestor in the perturbed ontology, depending on how abstract (position in the hierarchy) the term z is. Ideally, the similarity scores of terms should not change much (deviation ~ 0) for small perturbations and should progressively deviate with an increase in perturbation.

2.5. Gene-based and process-based disease similarity

The similarity for a given pair of diseases was calculated in four different ways: Gene-Identity based (GIB), Process-Identity based (PIB) and Process-Similarity based (PSB) (Fig. 1). As the name suggests, GIB computes inter-disease similarity based only on the set of genes known to be implicated in both diseases. PIB computes disease similarity based on shared processes, while PSB also takes the similarity between related processes into consideration. The three approaches are illustrated using a pair of diseases, Alzheimer's (240 genes) and Schizophrenia (200 genes) with 39 genes in common.

2.5.1. Gene-based disease similarity

In the GIB approach, the similarity between a pair of diseases was calculated as the hypergeometric probability of shared gene associations. For example, the p-value for Alzheimer's and Schizophrenia, based on the 39 shared genes, is 5.61×10^{-14}.

2.5.2. Process-based disease similarity

GO-Processes associated with each disease were identified by measuring the over-representation of GO-Processes in the corresponding gene set by using the hypergeometric test, and corrected for multiple testing using the Benjamini–Hochberg test. To minimize false-positives, a minimum membership of at least three genes was required for a GO-Process to be considered significant as it was observed that GO-Processes with one or two genes were highly likely to have artificially low p-values. To find a suitable p-value cut-off for associating a GO-Process with a disease, GO-Processes were first extracted at multiple p-value cut-offs of 0.05, 0.01, 0.005, 0.001 and 0.0001. This resulted in a disease being annotated with sets of biological processes at different stringencies.

A common denominator of various diseases is the set of genes that participate in the immune response such as B/T-cell proliferation, chemotaxis and regulation of isotype switching. These processes, though not very specific to disease etiology, can give artificially high scores if raw frequencies are used. To minimize the obfuscating effect of such commonly used terms on the similarity score between diseases, each GO-Process used in the calculation of the similarity score was normalized by its information content in the GO-Process graph and its information content in disease space.

$$NF = \frac{IC_{GO}(P)}{MaxIC_{GO}} * \frac{IC_{DIS}(P)}{MaxIC_{DIS}}$$

where NF is the Normalizing Factor for the GO-Process, IC_{GO} is the information content of the GO-Process P in the entire GO-Graph, and IC_{DIS} is its information content in disease space. $MaxIC_{GO}$ and $MaxIC_{DIS}$ are maximum information contents in GO-Graph and Disease Space respectively. The *msim* values from Eq. (3) were multiplied by corresponding NF values.

In the PIB approach, the similarity between a pair of diseases was calculated by considering only the common processes. A self-similarity score was computed for each process using Eqs. (1) and (2), and then summed to yield a disease similarity score. An example of a GO-process that would be included in PIB estimation of the similarity between Alzheimer's and Schizophrenia is 'neurotransmitter biosynthetic process' (GO:0042136) as both the corresponding gene sets are enriched in this. These diseases have a total of nine identical processes.

In the PSB approach, in addition to the common GO-processes, the similarity between non-identical processes between two diseases was also used to calculate similarity. The similarity between each pair of GO-Processes was computed using Eqs. (1) and (2). The similarity between a pair of diseases (entities) was computed by integrating the similarity between corresponding GO-Processes (attributes) using Eq. (3). For example, 'synaptic transmission, dopaminergic' (GO:0001963) in Alzheimer's is estimated by PSB to be similar to 'dopamine receptor signaling pathway' (GO:0007212) in Schizophrenia, and would therefore be included, in addition to common processes, in calculating the similarity between the two diseases. It is important to note that, even though the similarity between these terms appears to be trivially intuitive, these terms are not obviously related in the GO hierarchy; their nearest common ancestor is the root term 'biological process.' These diseases have 20 pairs of similar processes, in addition to the nine common processes.

PSB and PIB scores were converted to p-values by comparing them with a corresponding null model – 20,000 random disease pairs in DO-Lite. Cross-validation on a benchmark set of diseases was used to determine the optimal combination of disease similarity p-value cut-off and hypergeometric cutoff for process enrichment (Table 2 in Supplementary material).

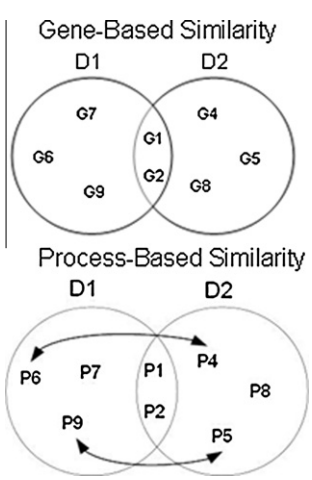

Fig. 1. Contrasting approaches used to compute similarity between diseases. D1 and D2 are diseases, Gi are associated genes, Pi are enriched GO-Processes. G1 and G2 are genes common to both diseases. P1and P2 are biological processes common to both diseases; P4 is similar to P6.

2.6. Validation and test sets for disease similarity

2.6.1. Validation set

A set of 36 diseases with 68 known disease similarities (see Supplementary material) was used as the benchmark. This was based on the diseases analyzed in the study by Suthram et al. [12] with the addition of common diseases like asthma, hypertension and lipid disorders. Cancers were omitted as a large number of secondary biological processes are affected. The guiding principle for classifying a pair of diseases as being related was that the knowledge of one could potentially help in the management or treatment of the other. A disease pair was marked as similar if it was reported in a textbook, review paper in a peer-reviewed mainstream journal, or multiple independent journal articles and met at least one of the following conditions: (a) Have common underlying pathophysiology that causes or increases the risk of both diseases. (b) Both diseases could result in a common metabolic signature or biochemical phenotype. (c) One of the diseases could increase the risk of the second one, e.g., diabetes mellitus increasing the risk of nephropathy or HIV infection increasing the risk of tuberculosis.

Isolated reports of weak correlations based on statistical studies of clinical records were not taken as evidence of disease similarity.

2.6.2. Test set

A published dataset (available at http://rxinformatics.umn.edu/SemanticRelatednessResources.html) based on the judgment of medical residents [30] on the semantic relatedness of medical terms was used as an external test set. Out of the 587 pairs of UMLS concepts, the majority involve symptoms and drugs, for example, "Lipitor-Zocor" or "Heartburn-Protonix." Only 76 were found to be disease pairs, where disease is defined as a term that maps to the Disease Ontology [20]. Of these, 27 pairs of diseases remained after filtering for GO Process enrichment (Processes having at least 3 associated genes AND p-value $\leqslant 0.005$). UMLS-Similarity tools [31] were used to estimate semantic similarity and semantic relatedness [32] for these disease pairs and compared with the process-based measured proposed here.

2.7. Fold change calculation

Microarray datasets mentioned in Suthram et al. [12] that had both disease and control (normal) data were extracted from NCBI-GEO [33]. In addition, a dataset for Asthma was also extracted (GSE470). Expression values were calculated using RMA [34] if CEL files were available; otherwise the existing values were transformed to *log* 2. In case of multiple probesets for a gene, the probeset with the highest inter-quartile range was selected. To find if a GO-process is significantly regulated in a disease, the fold change (average expression value in diseased samples/average expression value in control samples) for a GO-process was compared with average fold change of random sampling of genes from the experiment. Fold change provides a rough estimate of the change in gene expression in diseased samples compared to control samples. Genes with high fold change in positive or negative directions have a high likelihood of being involved in the disease. The fold change of a GO-Process was calculated by averaging the fold changes (fc) of the individual genes. As a null model, fold changes for 1000 randomly selected gene-sets of variable size were estimated for each disease. A fold change p-value for a GO-Process was calculated as the percentage of random gene-set fold changes greater/lower than that of the fold change of the GO-Process, depending on whether the value was positive/negative. To further contrast the GO-Process p-values of similar diseases detected by PSB and PIB approaches, a null model was constructed using 1000 random GO-Processes. The p-value for each random GO-Process was determined using the fold change method described earlier for each disease. The median p-value among diseases was associated with the random GO-Process. Finally, the set of p-values obtained for random GO Processes were compared to the p-values obtained from PSB and PIB approaches.

3. Results

3.1. Performance of similarity metric

Pathways from KEGG were used as a benchmark to evaluate the accuracy of the proposed measure (*Mb*). The intra-pathway scores for *Mb* (mean = 1.9, SD = 0.68) were significantly higher than the inter-pathway scores (mean = 0.08, SD = 0.21), p-value = 1.88e−17, demonstrating that *Mb* distinguishes between unrelated and related genes. The performance of the proposed similarity measure (*Mb*) for genes is contrasted with five other measures in Table 1. A representative sample of the KEGG pathways used for evaluation is shown. The first two columns show the pathway ID and corresponding number of genes. As detailed in Methods, the p-values reflect the probability that the group of genes in each pathway could be assembled by chance; low p-values imply that the genes within a pathway are related.

The proposed metric's p-value (*Mb* in column 1) is consistently lower than that of the other five similarity measures. The average p-values of 71 pathways for the six similarity measures in the order listed in Table 1 are 0.02, 0.31, 0.12, 0.26, 0.17, 0.19 respectively. The number of pathways that were significantly distinguished by the similarity measures ($p \leqslant 0.05$) was 68, 27, 51, 35, 22 and 55 respectively. The next best performer after *Mb* is "Res". One explanation for the lower p-values for *Mb* compared to the other metrics is that, in many instances, GO-Processes with similar function have abstract terms as common ancestors. In other words, the similarity between processes is not obvious from the ontological hierarchy as the terms appear to be far apart. So measures that use the information content in common ancestors or the graph structure could fail to capture similarity in such instances.

Since *Resnik* performed the best among existing metrics, we further compared it to '*Mb*' as follows. The GO-Process ontology was perturbed up to 2%, 5%, 10%, 15% and 20% (Fig 2) and average deviation of pathway scores assessed. Ideally, the resulting deviation should be proportional to the amount of error introduced into the ontology.

The Resnik pathway scores seem unaffected by the degree of perturbation, implying that the information content in the common ancestors does not undergo sizable changes. This is expected as the common ancestors of many pairs of terms are abstract terms. In contrast, the deviation for '*Mb*' is proportional to the extent of perturbation. It shows a small deviation at low perturbations (2% and 5%) but rises on large perturbations (10–20%). This suggests that '*Mb*' is robust to small errors in an ontology and degrades gracefully with increasing noise.

To check for the bias introduced by variable cardinality of annotation as reported by [35], i.e., any effect of the number of GO Processes or genes associated with diseases on the disease similarity scores, the Pearson correlation coefficient was calculated between the following: the number of genes associated with a disease, number of associated biological processes and the average disease similarity scores. As expected, higher the number of genes annotating a disease, higher the number of enriched GO Processes (correlation = 0.93). The accompanying increase in score is smaller as the correlation of the number of processes to similarity scores is 0.56. This indicates that the increase in similarity score is only weakly proportional to gene/bioprocess cardinality.

3.2. Comparison of gene-based and process-based assessments of disease similarity

A curated set of 36 diseases with 68 known pairwise similarities (see Table 1 of Supplementary material for complete list) was taken as a benchmark to compare different approaches to assess disease

Table 1
Comparison of p-values for gene similarity within KEGG Pathways from six different methods.

KEGG Pathway	No. of genes	*Mb*	Lin	Res	Wang	Lc	PMI
hsa00511	5	0.001	0.01	0.01	0.01	0.01	0.13
hsa00450	11	0.015	0.43	0.02	0.83	0.14	0.35
hsa00760	15	0.016	0.92	0.93	0.81	0.34	0.09
hsa00270	20	0.000	0.67	0.01	0.11	0.17	0.37
hsa00310	23	0.105	0.20	0.26	0.07	0.64	0.78
hsa00350	30	0.001	0.96	0.49	0.88	0.13	0.48
hsa00330	38	0.001	0.96	0.68	0.23	0.26	0.16
hsa00980	45	0.001	0.04	0.07	0.02	0.01	0.01

'*Mb*' is the proposed method.

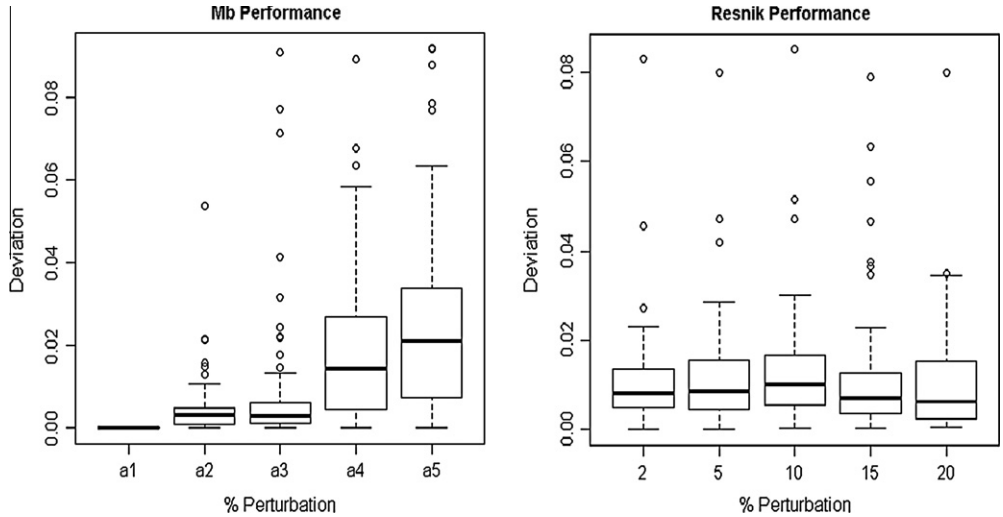

Fig. 2. Box-plots of deviation of pathway scores with increasing ontology perturbation for '*Mb*' and 'Resnik' similarity measures.

similarity. A representative sample of disease pairs with corresponding *p*-values is shown in Tables 2a and 2b.

Recall and precision were estimated based on the optimal combination of cutoffs for associating a process with a disease, and for the disease similarity score threshold (details in Table 2 of Supplementary material). In the PSB approach, hypergeometric *p*-value $\leqslant 0.005$ and score *p*-value $\leqslant 0.067$ had the best performance with *f*-score = 0.575 (Recall = 55%, Precision = 60%). In the PIB approach, hypergeometric *p*-value $\leqslant 0.001$ and score *p*-value $\leqslant 0.081$ had the best performance with *f*-score = 0.38 (Recall = 29%, Precision = 58%). Disease similarities calculated using the extent of gene overlap (GIB) (hypergeometric *p*-value < 0.05) resulted in *f*-score = 0.27 (Recall = 88%, Precision = 16%). While GIB seems to perform almost as well as PSB in detecting the presence of disease similarity, it suffers from poor precision. In other words, it suffers from a large proportion of false positives. A comprehensive summary of the performance of GIB, PIB and PSB approaches is plotted as Recall/Precision curves in Fig. 3 (Recall and Precision plot is shown rather than a receiver operating curve (ROC) as the number of True-Negatives is much higher than True-Positives). The area under the curve for the PSB approach is the highest.

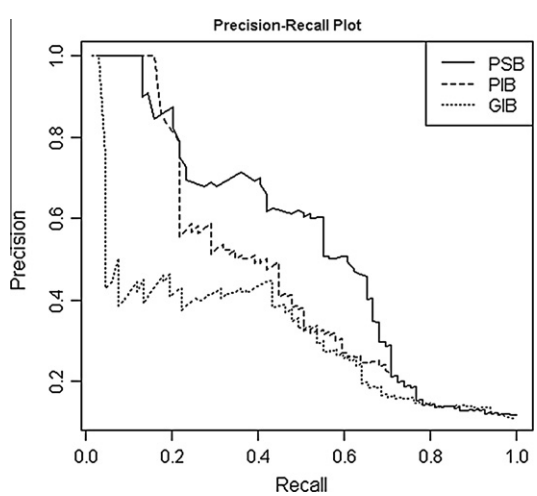

Fig. 3. Recall–precision graphs for disease similarity obtained from PSB (hypergeometric *p*-value $\leqslant 0.005$), PIB (*p*-value $\leqslant 0.001$) and GIB (*p*-value $\leqslant 0.05$) approaches.

Pubmed identifiers, with supporting evidence type in parenthesis (see Section 2.6), for the first eight disease pairs reported in Table 2a are 15,236,409 (a), 9,787,748 (c), 18,230,193 (a), 21,083,567 (a), 20,371,230 (a), 17,552,001 (a), 17,401,045 (b), and 20,157,305 (b). The last pair is trivial as diabetes is known to cause kidney failure. Table 2b shows unrelated diseases. GIB predicts all of these as having a low probability of being unrelated, i.e., makes an incorrect prediction for all of them. While PSB does better, it yields lower *p*-values for the last three entries in the table; false positives persist, though less than in a gene-based assessment.

3.3. Comparison with existing methods

Out of the 54 diseases chosen by Suthram et al. [12], 41 were found in DO. Cancers were omitted from the analysis, as were diseases with minimal annotation, resulting in 24 diseases. This

Table 2a
Sample of related disease pairs with *p*-values for gene based and process based similarity.

Disease 1	Disease 2	GIB	PIB	PSB
Alzheimer's Disease	Schizophrenia	0	0.040	0.008
Hyperlipidemia	Diabetes Mellitus	0	0.103	0.023
Polycystic Ovary Syndrome	Diabetes Mellitus	0	0.185	0.040
Asthma	Diabetes Mellitus	0	0.010	0.041
Mental Depression	Schizophrenia	0	0.262	0.024
Hyperlipidemia	Brain Diseases, Metabolic, Inborn	0.04	0.127	0.062
Alzheimer's Disease	Bipolar Disorder	0	0.002	0.039
Infertility, Male	Obesity	0.06	0.261	0.053
Diabetic Nephropathy	Diabetes Mellitus	0	0.174	0.058

Table 2b
Sample of unrelated disease pairs with *p*-values for gene based and process based similarity.

Disease 1	Disease 2	GIB	PIB	PSB
Schizophrenia	Polycystic Kidney Diseases	0	0.050	0.133
Multiple Sclerosis	Lipid disorder	0	0.051	0.182
Endometriosis	Aortic Aneurysm	0	0.331	0.257
Multiple Sclerosis	Asthma	0	0.031	0.025
Endometriosis	Diabetes Mellitus	0.020	0.026	0.034
Sarcoidosis	Asthma	0.020	0.074	0.031

Table 3
Pearson correlation between mean scores provided by residents and semantic similarity and relatedness measures.

	PSB	Lin	Res	Lc	Jc	Lesk adapted	Vector based
Pearson correlation	0.733	0.529	0.552	0.312	0.321	0.562	0.841
p-Value	1.4e−05	0.004	0.002	0.113	0.101	0.002	3.8e−08

Jc = Jiang and Conrath [36], Lesk adapted = extended gloss overlaps [37], Vector based = semantic relatedness [38].

contains 21 true disease pairs (see section 2.6 for definition of disease similarity). The PSB method predicted 14 pairs out of which nine were true (Recall = 43% and Precision = 65%), while Suthram et al. reported 115, out which eight are true (Recall = 38% and Precision = 7%). The pairs that were correctly detected by both methods are Alzheimer's Disease & Schizophrenia, Alzheimer's & Bipolar Disorder, Schizophrenia & Bipolar Disorder and Hyperlipidemia & Diabetic Nephropathy. Examples of similarity detected by PSB but not by Suthram et al. are Obesity & Polycystic Ovary Syndrome, and Hyperlipidemia & Obesity. Out of the 20 disease pairs reported by Li et al., none were found to be significant using the PSB method. Out of 20 associations reported by Wang et al., 11 were found to be significant (six were trivially related by IS-A relations).

Table 3 lists the Pearson correlation coefficients between the mean scores of the medical resident test set and those obtained from PSB and other measures. Since the expert scores are a numerical value between 0 and 1600 rather than a Boolean judgment, correlation with the raw scores from the PSB method was computed. Table 4 shows scores and verdicts for 10 of the 27 disease pairs from the resident test set (The details of all the disease pairs and the corresponding scores are shown in Table 3 of the Supplementary file).

The vector-based relatedness measure shows the highest correlation with numerical scores provided by a group of medical residents. PSB shows the second highest correlation. As expected, the semantic similarity measures (Lin, Res, Lc and Jc) show lower or insignificant correlations when compared with semantic relatedness measures (PSB, Lesk adapted and Vector-based).

3.4. GO-Processes in microarray expression data

Among the disease pairs detected by PSB and PIB methods, 10 and 11 pairs respectively had available microarray data. *P*-values of the top 3 GO-Processes were computed using the method described in Section 2.7 and compared with random GO-Process *p*-values using the non-parametric Wilcoxon rank sum test. The GO-Process *p*-values obtained from PSB approach were found to be significantly different (*p*-value = 0.024). The distribution of *p*-values using PIB was not found to be significantly different (*p*-value = 0.178). This indicates that the GO-Processes found by the PSB approach between similar diseases are significantly regulated in gene expression microarray datasets.

Table 4
Mean scores provided by residents and the corresponding PSB scores for the five most related and five least related disease pairs in the test set.

Disease 1	Disease 2	Resident scores	PSB scores	PSB decision
Hypothyroidism	Goiter	1424	1.0494	Yes
Ischemias	Arteriosclerosis	1399.5	1.4226	Yes
Angina	Atherosclerosis	1357.75	1.4922	Yes
Pneumoniae	Influenza	1354	1.0023	Yes
Meningitis	Encephalitis	1325.75	1.0334	Yes
Ischemia	Epilepsy	477.5	0.28	No
Influenza	Atherosclerosis	416	0.1265	No
Epilepsy	Cataract	361	0.1222	No
Cataract	Pancreatitis	345.5	0.0239	No
Cardiomyopathy	Osteoporosis	326.25	0.2191	No

4. Discussion

Biomedical ontologies are growing in popularity as the usefulness of controlled vocabulary in addressing biomedical problems is being increasingly recognized. In this paper, we have demonstrated that similarity between genes, or between diseases, can be more accurately measured by combining evidence from co-occurrence, information content and the semantics embedded in the ontological hierarchies. A case is made for estimating disease similarity based not just on gene overlap, but in terms of the similarity between the underlying processes.

As noted in [22], measures like PMI that are based purely on co-occurrence in corpi are biased towards very specific terms which share only one gene. It introduces negative scores when the joint probability is lower than the product of individual probabilities, i.e., log of real numbers between 0 and 1. This can confound the results when estimating disease similarity.

The construction of ontologies in a domain is often dictated by the overarching theme of classification. There can be multiple ways to classify entities, e.g., diseases can be classified based on anatomy, symptoms or etiology. This can result in different hierarchical relationships. A similarity metric can therefore yield different estimates for the similarity of a pair of entities based on the particular ontology used. Further, ontologies are often a work in progress with terms and relationships both added and deleted over time (e.g., GO has grown in size by an order of magnitude in the decade since its inception). Thus, similarity metrics that are based on the common-ancestor of two terms in the ontology will, by definition, be unable to capture similarity between terms that are related but topologically far apart in the ontology. Such methods are also sensitive to errors in ontologies. Hybrid methods which use graph structure and node information are also inadequate as they will fail to capture similarity between terms that are far apart in the ontology.

The observed co-occurrence of ontological terms in annotated data can potentially compensate for the subjective and incomplete nature of a given ontology in estimating term-term similarity. For example, rhodopsin mediated phototransduction (GO:0009586) and rhodopsin mediated signaling pathway (GO:0016056) are closely related semantic terms. However, they share the common ancestor 'Signaling' only one level below the generic term 'Biological Process' in the GO Process ontology; their position in the ontology seems to indicate that they are not closely related. In fact, the two GO-Processes share all three genes annotated to them in the human genome. The proposed metric, based on both co-occurrence and information content of terms, gives a high score (95th percentile) while the scores for existing metrics range from 0 to 0.58 (1 being highest).

Similarity between diseases has previously been computed between flat lists of shared genes [21], pathways [13] or functional modules [12]. Given the fact that many processes are related, we computed the pair-wise similarity between diseases by basing it on process similarity. In other words, instead of looking only for identical matches between biological processes, the similarity between biological processes was computed and used in assessing the similarity between a pair of diseases. The hypergeometric distribution was used to estimate overrepresentation, i.e., the

enrichment of processes in the gene set associated with a disease. To minimize spurious results, the following filter was applied. A biological process was required to have at least 3 genes to be considered for association with a disease. Further, rather than relying on the rather permissive traditional value of 0.05 for statistical significance, a curated set of 36 diseases with 68 verified pairs of similar diseases was used to find the optimal *p*-value cutoff (see Supplementary file, Table 2). The *p*-value cutoff of 0.005 was found to be the most stringent as out of 1565 diseases that were associated with at least 1 GO-Process at 0.05 significance level, 1477 diseases were found at 0.005 significance level.

To evaluate if using similarity between processes is more accurate than using identical matches to compute disease similarity, disease similarities were computed using the two approaches in parallel and compared. While process-matching performed better than gene-matching or gene-similarity, process similarity did better than all other approaches (Fig. 3 and Tables 2a and 2b). Some of the possible reasons for this are discussed below.

When a pair of diseases has few identical processes, the resulting disease similarity score may be low. For example, Diabetic nephropathy and Hyperlipidemia only share 'triacylglycerol metabolic process' (GO:0006641). A consideration of similarity, and not just identical processes reveals that 'cholesterol transport' (GO:0030301) in Diabetic Nephropathy shares similarity with 'cholesterol efflux' (GO:0033344) and 'phospholipid efflux'(-GO:0033700) in Hyperlipidemia. Another example is the association of Male Infertility with Obesity despite having no identical process matches. Process similarities between Male Infertility and Obesity involve 'triacylglycerol metabolic process' (GO:0006641), 'follicle-stimulating hormone secretion' (GO:0046884) and 'regulation of insulin secretion' (GO:0050796). Other interesting disease pairs include Rheumatoid Arthritis and Diabetes Mellitus. Patients with Rheumatoid Arthritis have increased risk to type-2 diabetes as systemic inflammation can cause insulin resistance [39]. Some of the significant process pairs between the 2 diseases are 'positive regulation of MHC class II biosynthetic process' (GO:0045348), 'negative regulation of osteoclast differentiation' (GO:0045671), 'positive regulation of B cell proliferation' (GO:0030890) and 'fibrinolysis' (GO:0042730). Another disadvantage of considering only identical processes is that it can give an artificially high score based on rare matching processes that are not relevant to disease etiology, e.g., spurious association of Huntington's disease with Hamman–Rich syndrome based on the common process 'protein oligomerization' (GO:0051259). Studying processes in common between diseases can potentially help in borrowing treatment for one disease from another. An interesting example is the drug Donepezil (DrugBank ID: DB00843), which is effective in Alzheimer's by targeting genes in the "muscarinic acetylcholine receptor signaling pathway" (GO:0007213). Process-based disease similarity shows that this is the top hit among shared processes between Alzheimer's and Schizophrenia. In fact, though its effectiveness is inconclusive (Ferreri, Agbokou et al. 2006), Donepezil has been used in several clinical trials for Schizophrenia. Thus, there are several advantages to incorporating the semantic similarity between processes in assessing similarity between diseases.

By comparing the fold changes of apparently significant GO-Processes with randomly selected processes, the set of significant findings can be minimized. GO-Processes corresponding to positively classified disease pairs were found to have significant fold changes in gene-expression microarray data.

Though some of the disease pairs reported by other studies were found to be significant by the PSB method, the overlap was poor. This can be attributed to the variety of data sources used for computing disease similarities, the nature of annotation and differences in the methods used. Wang et al. [25] used only OMIM as the source of gene annotation and looked for identical matches between pathways. Though pathways offer a rich source of annotation, much of the human genome is not represented. Comparatively, GO-Processes offer higher annotation coverage and a platform for function inference. Suthram et al. [12] used microarray gene-expression datasets for estimating disease similarity. While microarrays measure the overall response of the disease, it is susceptible to noise which is exacerbated by relatively small sample sizes for each disease. Microarrays include the measurement of many incidental features of a disease like immune response and secondary effects. This could mask the primary genes that cause the disease. The authors note that targets for potential repositioning of drugs between diseases were often immune-response related processes. Li et al. use phenotypes from OMIM to link diseases. The phenotypes can be abstract and have a strictly genetic pre-disposition. Many of the reported disease similarities share an IS-A relationship in DO. Studies on disease similarities often report their findings without using a test set to validate the accuracy of the predictive method used. A notable exception is a series of studies on semantic relatedness verified by physicians, medical residents or coders [30–32]. However, most of the terms in the benchmark used in the studies refer to non-disease terms, and the few pairs of terms that represent disease–disease pairs often represent dissimilar pairs. We were able to find 27 pairs of diseases suitable for verifying the accuracy of the PSB measure on an external test set. A significant correlation with medical resident scores validates the PSB measure. Interestingly, a vector based measure [38] that exploits second-order co-occurrence of terms in clinical notes performed even better. This suggests that combining relational clues in co-occurrence of terms with hierarchical inference from ontological structure is a reliable estimate of disease similarity. However, the vector based measure had relatively low scores for the less obvious similarities detected by the PSB measure, such as Diabetes mellitus & Polycystic Ovary Syndrome (vector score = 0.619), Asthma & Diabetes mellitus (vector score = 0.462), and Male infertility and Obesity (vector score = 0.3452). Presumably, the underlying processes need to be factored into the comparison to detect shared aspects of pathophysiology.

Understandably, it is difficult to identify true negatives as this is an active area of research with systems-level analysis continuing to reveal previously unknown relationships. However, this makes it difficult to evaluate the true effectiveness of various methods. By creating and using a benchmark set, we have systematically evaluated the performance of the proposed measure. It would be worthwhile to establish a curated data set, such as the one in this paper but larger in size, that has both true positives and negatives delineated in a causal context by collaborative efforts that could help in benchmarking different methods in the future. Although restricting the use of GO-Processes by using an arbitrary level in the GO hierarchy as cutoff has been shown to be detrimental [40,41], we observed that performance of PSB improved (*f*-score = 0.61) when GO-Processes < level 5 were discarded (data not shown). This is due to association of many abstract GO-Processes to diseases with low *p*-values.

We have shown that incorporating the similarity between biological processes yields more accurate detection of disease similarity than explicit gene or process matching. Enhancing the disease vocabulary by using a combination of UMLS (SNOMED-CT, MeSH, ICD) and DO overcomes the problem of synonyms and yields better coverage by exploiting the hierarchical structure of ontologies. However, it is important to point out the limitations of the work described here. The ontological term similarity measure tends to overestimate similarity between a pair of terms that are rare/poorly annotated as the average information content is high even though the co-occurrence score is compensated by the average of maximum scores. We have used genes from OMIM, GeneRIF and Swissprot data sets in this approach. While OMIM and Swissprot offer high quality annotation, GeneRIF covers a wide variety of

gene–disease relationships reported in the literature. Consequently, a disease is not only annotated with causal genes but also ancillary genes associated with its symptoms (e.g., immune response genes are common to a large subset of diseases, leading to false positives like the last three rows in Table 2b). As discussed above, this can confound the interpretation of observed disease similarity. The PSB approach has a bias towards well annotated diseases. To minimize false positives in using the hypergeometric distribution, diseases were required to have at least five genes [40], and GO-Processes at least 3. Hence relationships among diseases that are sparsely annotated cannot be inferred; this is unfortunate as understanding the etiology of such diseases is particularly important. At the other end, a large number of genes annotated to a disease also tend to bias the disease similarity score as many biological processes are affected.

5. Conclusion

The widespread use of biomedical ontologies demands methods that translate information between ontologies and quantify similarity between terms in an ontology. This paper presents a function to measure similarity between a pair of ontological terms, or entities annotated with them, that outperforms other well known similarity measures. Quantifying similarity between diseases has the potential to help in understanding pathophysiology and ultimately leading to clinical interventions like repositioning of drugs. By using a curated set of similar diseases to evaluate different approaches to measure similarity between diseases (gene-based and process-based), we conclude that it is important to consider the similarity between processes underlying each disease for more accurate prediction of disease similarity. Complementing inferences on diseases similarity from gene annotation with text-mining approaches that exploit phenotypic case reports [42] and clinical notes [43] can help in more comprehensive discernment of hidden similarities between apparently disparate diseases.

Acknowledgments

We would like to thank the University of Missouri Bioinformatics Consortium (UMBC) for access to the High Performance Computing Resources used for the computations in this project, Arcady Mushegian for valuable feedback on the manuscript, the lab of Ted Pedersen for access to evaluation software, and the reviewers for helpful comments and suggestions for improving evaluation of the approach.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2011.11.017.

References

[1] Chen Y et al. Variations in DNA elucidate molecular networks that cause disease. Nature 2008;452(March 27):429–35.
[2] Goh KI et al. Proc Natl Acad Sci USA 2007;104(May 22):8685–90.
[3] Derman CAL, Lieberman GJ, Ross SM. On the consecutive-*k*-out-of-*n*:*F* system. IEEE Trans Reliab 1982;31:57–63.
[4] Ein-Dor L et al. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 2005;21(January 15):171–8.
[5] Wirapati P et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res 2008;10:R65.
[6] Butte AJ, Kohane IS. Creation and implications of a phenome–genome network. Nat Biotechnol 2006;24(January):55–62.
[7] Hughes B. 2009 FDA drug approvals. Nat Rev Drug Discov 2010;9(February):89–92.
[8] Hirsch HA et al. A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases. Cancer Cell 2010;17(April 13):348–61.
[9] Rzhetsky A et al. Probing genetic overlap among complex human phenotypes. Proc Natl Acad Sci USA 2007;104(July 10):11694–9.
[10] Hidalgo CA et al. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol 2009;5(April):e1000353.
[11] Park J et al. The impact of cellular networks on disease comorbidity. Mol Syst Biol 2009;5:262.
[12] Suthram S et al. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol 2010;6(February):e1000662.
[13] Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. PLoS ONE 2009;4:e4346.
[14] Li Y, Patra JC. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. Bioinformatics 2010;26(May 1):1219–24.
[15] Ashburner M et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 2000;25(May):25–9.
[16] Humphreys BL et al. The unified medical language system: an informatics research collaboration. J Am Med Inform Assoc 1998;5(January–February):1–11.
[17] Pesquita C et al. Semantic similarity in biomedical ontologies. PLoS Comput Biol 2009;5(July):e1000443.
[18] Mathur S, Dinakarpandian D. A new metric to measure gene product similarity. In: Presented at the IEEE international conference on bioinformatics and biomedicine; 2007.
[19] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucl Acids Res 2000;28(January 1):27–30.
[20] Osborne JD et al. Annotating the human genome with disease ontology. BMC Genom 2009;10(Suppl. 1):S6.
[21] Mathur S, Dinakarpandian D. Automated ontological gene annotation for computing disease similarity. In: Presented at the AMIA summit on translational bioinformatics 2010; 2010.
[22] Church KW, Hanks P, Word Association Norms. Mutual information and lexicography. In: 27th Annual conference of the association of computational linguistics; 1989. p. 76–83.
[23] Patwardhan SBS, Pedersen T. Using measures of semantic relatedness for word sense disambiguation. In: Fourth international conference on intelligent text processing and computational linguistics, Mexico City; 2003. p. 241–57.
[24] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. In: WordNet: an electronic lexical database. Cambridge; 1998. p. 265–83.
[25] Wang JZ et al. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007;23(May 15):1274–81.
[26] Resnik P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J Artif Intell Res 1999;11:95–130.
[27] Lin D. An information-theoretic definition of similarity. In: Proceedings of the fifteenth international conference on machine learning; 1998. p. 296–304.
[28] Yu G et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 2010;26(April 1):976–8.
[29] Guo X et al. Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 2006;22(April 15):967–73.
[30] Pakhomov S, et al. Semantic similarity and relatedness between clinical terms: an experimental study. In: AMIA annu symp proc 2010; 2010. p. 572–6.
[31] McInnes BT, et al. UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity. AMIA annu symp proc, 2009; 2009. p. 431–5.
[32] Pedersen T et al. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40(June):288–99.
[33] Barrett T et al. NCBI GEO: mining millions of expression profiles – database and tools. Nucl Acids Res 2005;33(January 1):D562–6.
[34] Irizarry RA et al. Summaries of Affymetrix GeneChip probe level data. Nucl Acids Res 2003;31(February 15):e15.
[35] Wang J et al. Revealing and avoiding bias in semantic similarity scores for protein pairs. BMC Bioinform 2010;11:290.
[36] Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of 10th international conference on research in computational linguistics; 1997.
[37] Satanjeev Banerjee TP. Extended gloss overlaps as a measure of semantic relatedness. In: Eighteenth international joint conference on artificial intelligence; 2003.
[38] Patwardhan BS, Pedersen T. Using measures of semantic relatedness for word sense disambiguation. In: Fourth international conference on intelligent text processing and computational linguistics, Mexico City; 2003. p. 241–257.
[39] Doran M. Rheumatoid arthritis and diabetes mellitus: evidence for an association? J Rheumatol 2007;34(March):460–2.
[40] Myers CL et al. Finding function: evaluation methods for functional genomic data. BMC Genom 2006;7:187.
[41] Dennis Jr G et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003;4:P3.
[42] Chen L, Friedman C. Extracting phenotypic information from the literature via natural language processing. Stud Health Technol Inform 2004;107:758–62.
[43] Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010;17(September–October):524–7.