RESEARCH ARTICLE

HPOSim: An R Package for Phenotypic Similarity Measure and Enrichment Analysis Based on the Human Phenotype Ontology

Yue Deng[1,2], Lin Gao[1]*, Bingbo Wang[1], Xingli Guo[1]

1 School of Computer Science and Technology, Xidian University, Xi'an, People's Republic of China
2 Institute of Software Engineering, Xidian University, Xi'an, People's Republic of China

* lgao@mail.xidian.edu.cn

OPEN ACCESS

Citation: Deng Y, Gao L, Wang B, Guo X (2015) HPOSim: An R Package for Phenotypic Similarity Measure and Enrichment Analysis Based on the Human Phenotype Ontology. PLoS ONE 10(2): e0115692. doi:10.1371/journal.pone.0115692

Academic Editor: Yu Xue, Huazhong University of Science and Technology, CHINA

Received: September 20, 2014

Accepted: November 25, 2014

Published: February 9, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by grants from National Natural Science Foundation of China (Grant Numbers: 61402349 to YD, 91130006 to LG, 61303122 to BW and 61303118 to XG), and the website is http://www.nsfc.gov.cn/publish/portal1/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Background

Phenotypic features associated with genes and diseases play an important role in disease-related studies and most of the available methods focus solely on the Online Mendelian Inheritance in Man (OMIM) database without considering the controlled vocabulary. The Human Phenotype Ontology (HPO) provides a standardized and controlled vocabulary covering phenotypic abnormalities in human diseases, and becomes a comprehensive resource for computational analysis of human disease phenotypes. Most of the existing HPO-based software tools cannot be used offline and provide only few similarity measures. Therefore, there is a critical need for developing a comprehensive and offline software for phenotypic features similarity based on HPO.

Results

HPOSim is an R package for analyzing phenotypic similarity for genes and diseases based on HPO data. Seven commonly used semantic similarity measures are implemented in HPOSim. Enrichment analysis of gene sets and disease sets are also implemented, including hypergeometric enrichment analysis and network ontology analysis (NOA).

Conclusions

HPOSim can be used to predict disease genes and explore disease-related function of gene modules. HPOSim is open source and freely available at SourceForge (https://sourceforge.net/p/hposim/).

Introduction

Phenotypic similarity plays an important role in different biological and biomedical applications. Previous studies prove that genes with similar phenotypes yields biological modules in terms of diseases, thus it can be used in predicting disease-causing genes [1][2]. Furthermore, it is crucial for understanding the relationships between different diseases [3].

Most current methods for measuring phenotypic similarity [4][5] are based on the Online Mendelian Inheritance in Man (OMIM) database [6] that contains textual records representing genetic disorders. However, the absence of a controlled vocabulary makes it difficult to analyze the OMIM data using a computational approach [7]. The Human phenotype ontology (HPO) [8] provides a controlled and standardized vocabulary of phenotypic abnormalities annotating all clinical entries in OMIM, which sheds light on the large-scale computational analysis of the human phenome, i.e., DECIPHER [9], ECARUCA [10] and Bridge [11].

Several tools using HPO-based semantic similarity are currently available. Phenomizer [12] is the first tool for semantic similarity search over HPO, in which users input the phenotypic abnormalities of a patient as HPO IDs, and obtain a list of diagnoses in OMIM IDs. Other tools include OwlSim [13], PhenoDigm [14], PhenomeNET/PhenomeBrowser [15] and OntoSIML [16]. The detailed comparison of HPOSim and other HPO-based tools is shown in Table 1. It can be seen from the table that most of the existing tools share one drawback: the calculations of phenotypic similarity for HPO terms, genes and diseases are not well supported. Although OntoSIML and OwlSim provide functions for calculating semantic similarity, users are required to manually input the mapping from entities (gene or disease) to HPO terms, which entails additional preprocessing effort.

In addition, there exist several tools for HPO-based enrichment analysis. OntoFUNC [17] performs functional enrichment analysis over ontologies in OWL format. It is based on FUNC [18] and users need to manually input the mapping data, which is the same as OntoSIML. STOP [19] is an online tool and can be used as a Cytoscape plug-in. It can be used in the enrichment analysis of gene sets, but does not support the analysis of disease set.

Several R packages for semantic similarity and enrichment analysis are available, including GOSim [20], GOSemSim [21], DOSim [22], DOSE [23] and topGO [24]. However, these packages mainly use gene ontology (GO) [25] and disease ontology (DO) [26]. To the best of our knowledge, there is no R package that focuses on HPO-based semantic similarity and enrichment analysis.

Table 1. Comparison of HPOSim and other HPO-based tools.

Name	Release Type	Open Source	Term-Term Similarity	Gene-Gene Similarity	Disease-Disease Similarity	Gene-Disease Similarity	Similarity Measures	Combine Methods
HPOSim	Stand Alone (R)	√	√	√	√	×	Resnik, Lin, Jiang-Conrath, relevance, information coefficient, graph IC, Wang	Max, Mean, funSimMax, funSimAvg, BMA
Phenomizer [12]	Web	×	×	×	×	√	Resnik	symmetric, unsymmetric
OWLSim [13]#	Stand Alone (Java)	√	√	√	√	√	Jaccard, Resnik, overlap/ normalized overlap, GIC	Max, Mean, BMA
PhenoDigm [14]	Web	×	×	×	×	√	Mean of Jaccard and Resnik	Max, Mean
PhenomeNET [15]	Web	×	×	√*	√	√*	simGIC	Unknown
OntoSIML [16]#	Web	×	√	√	√	√	Jaccard, simGIC, Resnik	Unknown

* PhenomeNET only supports human genes included in OMIM.

# Although OntoSIML and OwlSim provide functions for calculating semantic similarity, users are required to manually input the mapping from entities (gene or disease) to HPO terms, which entails additional preprocessing effort.“√” means the tool provides the function and “×” means the tool does not.

doi:10.1371/journal.pone.0115692.t001

Thus, we developed an R package HPOSim with an immediate purpose to capturing phenotypic similarities between genes and diseases. The framework of HPOSim is shown in Fig. 1. HPOSim analyzes semantic similarity for HPO terms, genes and diseases. Functional enrichment analysis of gene set and disease set are also provided, including the classic hypergeometric enrichment analysis and the novel network ontology analysis (NOA) [27].

Implementation

Data

HPO contains over 10000 terms (10686 terms in the HPO build #1042 released in September 2014) in three sub-ontologies, which are phenotypic abnormality (PA), onset and clinical course (OC) and mode of inheritance (MI). Approximately 99% of the HPO terms are in the PA sub-ontology. In each sub-ontology, terms are arranged in a directed acyclic graph (DAG) and are related to their parent terms by “is a” relationships. The structure of the HPO allows a term to have multiple parent terms, which enables different aspects of phenotypic abnormalities to be explored. Diseases and genes are annotated to the most specific terms possible, which means that if a disease or a gene is annotated to a term then all of the ancestors of this term also apply (see Fig. 2 for an example).

The official ontology file provided by the HPO Consortium is in obo format, which is plain text-based. Thus, like other widely used R package for biomedical ontologies, e.g. GO.db, we constructed an R package termed HPO.db. HPO.db provided programmatic interfaces to the hierarchical structure of HPO terms. HPOSim uses HPO.db to obtain information about terms and relationships between terms. HPO.db can be used by other R packages that use HPO data.

HPOSim provides two kinds of pre-calculated data within the package: the association between HPO terms, as well as association between genes and diseases (gene-to-phenotype, phenotype-to-gene, disease-to-phenotype and phenotype-to-disease). The associations between HPO terms are obtained from the original ontology and annotation data provided by the HPO Consortium, and the information content (IC) of the HPO terms is pre-calculated based on

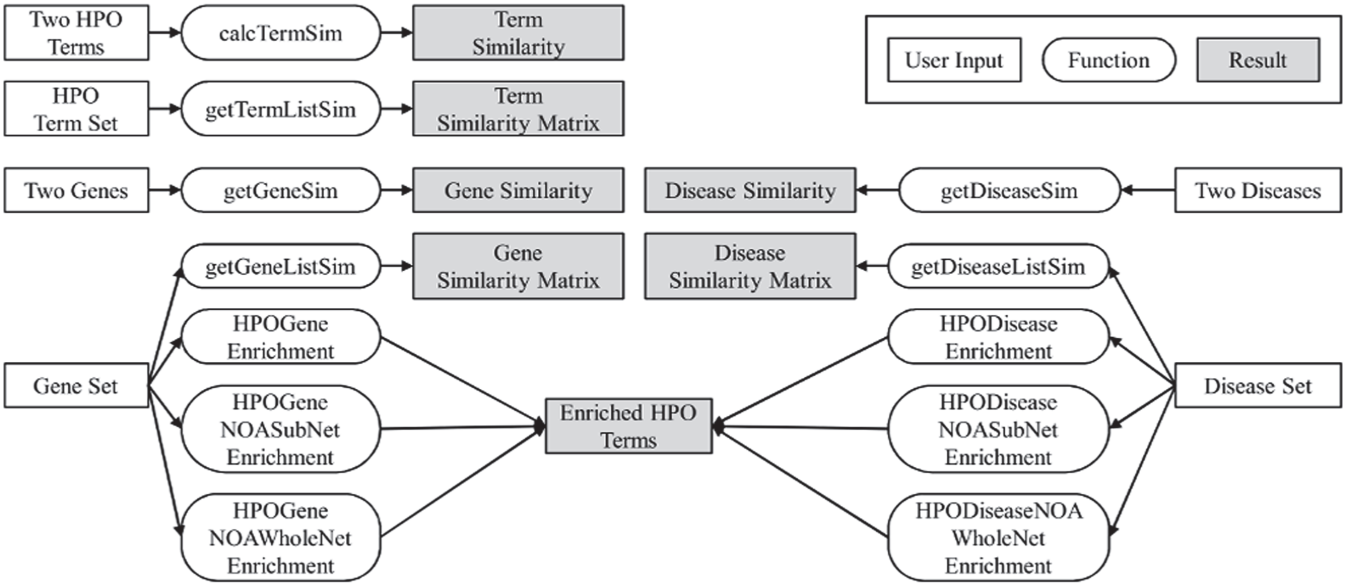

Figure 1. Framework of HPOSim. Users can use HPOSim to calculate semantic similarity for HPO terms, genes and diseases. HPOSim can also be used to identify enriched HPO terms for gene set and disease set.

doi:10.1371/journal.pone.0115692.g001

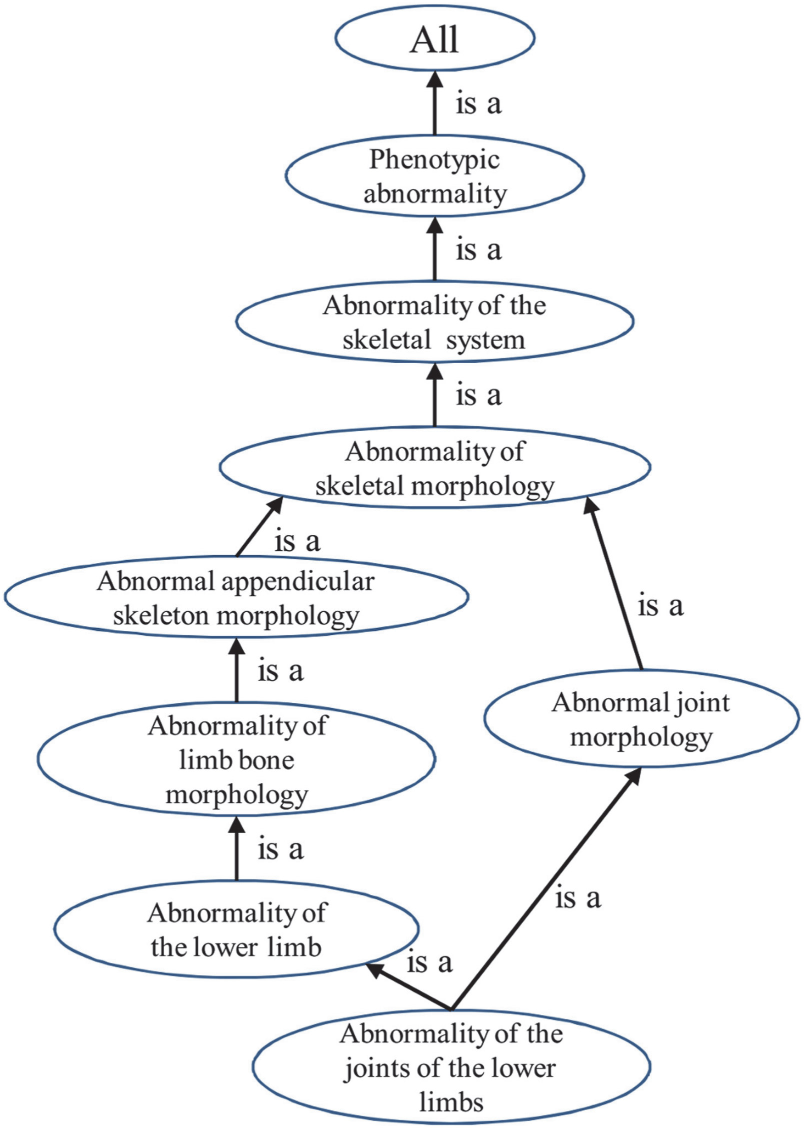
All
is a
Phenotypic abnormality
is a
Abnormality of the skeletal system
is a
Abnormality of skeletal morphology
is a
Abnormal appendicular skeleton morphology
is a
Abnormality of limb bone morphology
is a
Abnormality of the lower limb
is a
Abnormal joint morphology
is a
is a
Abnormality of the joints of the lower limbs

Figure 2. Example of the structure of HPO. HPO term *Abnormality of the joints of the lower limbs* (HP:0100491) and all its ancestor terms are shown. Each term in the HPO describes a phenotypic abnormality. Terms are related to parent terms by "is a" relationships in the form of a directed acyclic graph. If a disease or a gene is annotated to a term, it will also be annotated to all of its ancestors.

doi:10.1371/journal.pone.0115692.g002

both genes and diseases annotated to a certain term, while semantic similarity between genes and diseases are based on the IC of HPO terms.

The IC of a term t in HPO can be defined as follows:

$$IC(t) = -\log(p(t)) \tag{1}$$

where $p(t)$ is the probability of observing t and its descendants in all genes/diseases annotated to a certain sub-ontology of HPO.

Measuring the similarity between HPO terms

Recently, several metrics that measure the semantic similarity between ontology annotations have been proposed [28]. In HPOSim, we implement seven commonly used semantic similarity measures to measure the similarity between HPO terms: the Resnik measure [29], Lin measure [30], Jiang–Conrath measure [31], relevance measure [32], information coefficient measure [33], graph IC measure [34] and Wang measure [35]. The first six measures are based on IC, while the Wang measure uses both IC and graph structure.

The Resnik measure defines the similarity between terms as the IC of their most informative common ancestor (MICA):

$$sim_{Resnik}(t_1, t_2) = IC(t_{MICA}) \tag{2}$$

where t_{MICA} is the MICA of term t_1 and t_2.

The Lin and Jiang–Conrath measures consider the IC of the two terms besides the IC of their MICA:

$$sim_{Lin}(t_1, t_2) = \frac{2 \times IC(t_{MICA})}{IC(t_1)IC(t_2)} \tag{3}$$

$$sim_{JC}(t_1, t_2) = \ 1 - (IC(t_1) + IC(t_2) - 2 \times IC(t_{MICA})) \tag{4}$$

The relevance measure and the information coefficient measure are based on Lin's measure:

$$sim_{Rel}(t_1, t_2) = sim_{Lin}(t_1, t_2) \times (1 - p(t_{MICA})) \tag{5}$$

$$sim_{IC}(t_1, t_2) = sim_{Lin}(t_1, t_2) \times \left(1 - \frac{1}{1 + IC(t_{MICA})}\right) \tag{6}$$

The graph IC measure takes all the common ancestors of the two terms into account:

$$sim_{GraphIC}(t_1, t_2) = \frac{\sum\limits_{t \in (A(t_1) \cap A(t_2))} IC(t)}{\sum\limits_{t \in (A(t_1) \cup A(t_2))} IC(t)} \tag{7}$$

where $A(t)$ is the ancestors of term t in HPO.

The Wang measure is based on the graph structure of HPO DAG. In Wang's measure, a weight is given to each edge according to its type. $DAG_t = (t,T_t,E_t)$ represents the subgraph made up of term t and its ancestors, where T_t is the set of the ancestor terms of t and E_t is the set of edges in DAG_t.

In DAG_t, $S_t(n)$ measures the semantic contribution of term n to term t, which is defined as:

$$\begin{cases} S_t(t) = 1 \\ S_t(n) = \max\{w_e * S_t(n') | n' \in childrenof(n)\} \text{if} \, t \neq n \end{cases} \tag{8}$$

The similarity between HPO term t_1 and termt_2 is defined as:

$$sim_{Wang}(t_1,t_2) = \frac{\sum\limits_{t \in T_{t_1} \cap T_{t_2}} S_{t_1}(t) + S_{t_2}(t)}{SV(t_1) + SV(t_2)} \tag{9}$$

where $SV(m)$ is the sum of the semantic contributions of all the terms in DAG_m.

Combining term-term similarity into gene-gene and disease-disease similarity

In HPOSim, the similarity between two genes is calculated based on the pairwise similarity of the two HPO term sets annotating these two genes. HPOSim provides five methods to combine multiple term-term similarities into one gene-gene similarity, which are "Max" [36], "Mean" [36], "funSimMax" [32], "funSimAvg" [32], and "BMA" [35].

Given gene g_1 annotated by HPO term set $HPO_1 = \{t_{11}, t_{12}, \ldots, t_{1m}\}$ and g_2 annotated by $HPO_2 = \{t_{21}, t_{22}, \ldots, t_{2n}\}$. The similarity matrix $S=[s_{i\ j}]_{m\times n}$ contains all pairwise similarity scores of terms in HPO_1 and HPO_2.

The "Max" method calculates the maximum semantic similarity score over all pairs of HPO terms in the two term sets, and is defined as follows.

$$Sim_{Max}(g_1, g_2) = \max_{1\le i\le m, 1\le j\le n} s_{ij} \tag{10}$$

The "Mean" method calculates the average semantic similarity score over all pairs of HPO terms in the two term sets, and is defined as follows.

$$Sim_{Mean}(g_1, g_2) = \frac{1}{m\times n}\sum_{i=1}^{m}\sum_{j=1}^{n} s_{ij} \tag{11}$$

The "funSimMax", "funSimAvg" and "BMA" methods are based on the maximum value in each row and column of similarity matrix S.

The "funSimMax" and "funSimAvg" methods [32] use the arithmetic maxima and average between similarities for two directional comparisons of the similarity matrix S.

$$Sim_{funSimMax}(g_1, g_2) = \max\left\{\frac{1}{m}\sum_{i=1}^{m}\max_{1\le j\le n} s_{ij}, \frac{1}{n}\sum_{j=1}^{n}\max_{1\le i\le m} s_{ij}\right\} \tag{12}$$

$$Sim_{funSimAvg}(g_1, g_2) = \frac{1}{2}\times\left(\frac{1}{m}\sum_{i=1}^{m}\max_{1\le j\le n} s_{ij} + \frac{1}{n}\sum_{j=1}^{n}\max_{1\le i\le m} s_{ij}\right) \tag{13}$$

The "BMA" method uses the best-match average strategy, which calculates the average of all maximum similarities on each row and column of the similarity matrix S.

$$Sim_{BMA}(g_1, g_2) = \frac{\sum_{i=1}^{m}\max_{1\le j\le n} s_{ij} + \sum_{j=1}^{n}\max_{1\le i\le m} s_{ij}}{m+n} \tag{14}$$

The calculation of the similarity between diseases is the same as that between genes. The similarity between two diseases is calculated based on the pairwise similarity of the two term sets annotating these two diseases.

HPO-based Enrichment Analysis

HPOSim provides HPO-based enrichment analysis to investigate the phenotypic features of gene sets or disease sets. Two enrichment analysis methods are provided: hypergeometric test and the NOA method [27].

Given an HPO term t and a gene set with T genes, assuming that there are R genes/diseases annotated in the whole HPO in which G genes/diseases are annotated to t. In addition, there are O genes/diseases in the gene set that are annotated to t. The hypergeometric enrichment p-value for t is calculated as follows:

$$p - value = \sum_{x=O}^{\min(G,T)} \frac{\binom{G}{x}\binom{R-G}{T-x}}{\binom{R}{T}} \tag{15}$$

In NOA, users input a gene or disease network. For each edge in the network, the HPO terms annotating this edge are defined as the intersection of the two term sets annotating the two nodes of the edge. NOA uses HPO terms annotating the edges to perform the enrichment analysis. Two alternative strategies, "sub-net" and "whole-net", are applied to choose the reference set. In the "sub-net" strategy, users are required to provide the reference set. While in the "whole-net" strategy, the complete graph on the nodes of the input network is used as the reference set.

Results and Discussion

HPOSim consists of two parts: (i) the similarity measures between phenotypes (HPO terms), between human genes (Entrez IDs) and between diseases (OMIM IDs), and (ii) HPO-based enrichment analysis (NOA and the hypergeometric method) for gene set and disease set.

Application on gene similarity and gene set enrichment analysis

We used the aging network [37] to demonstrate the application of gene semantic similarity provided by HPOSim. The aging network was constructed by identifying genes related to aging and adding edges between interacting gene pairs. After removing the genes that are not annotated in the PA sub-ontology of HPO, 102 genes and 293 interactions were remained (see S1 Dataset for detail).

First, the semantic similarity matrix of the 102 genes was constructed using the Resnik measure and "funSimMax" combining method (see S2 Dataset for detail). A hierarchical clustering was then performed using the R package stats, and six modules were detected using the R package dynamicTreeCut. HPO enrichment analysis (hypergeometric test) was then performed using HPOSim. GO enrichment analysis and pathway enrichment analysis based on KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database [38] were performed using DAVID [39]. The results are shown in Table 2.

It can be seen that the enriched GO and HPO annotations are largely different among these modules. For example, the enriched GO annotations of module M2 implied that aging is associated with radiation including ultraviolet (UV), which has been verified by previous study in skin aging [40]. While the enriched GO annotations of module M3 implied that aging is associated with hormone stimulus, and literature mining showed that older women require a greater parathyroid hormone stimulus than younger women [41]. The enriched HPO annotations of the module M3 implied that aging are associated with abnormality of the pituitary, which has been verified by Sano *et al.* [42]. Disease enrichment analysis based on OMIM was then performed on genes in M3 using DAVID [39] and showed that term "Pituitary hormone deficiency, combined" was representative (p-value = 8.2E-3).

The enriched pathways of different modules are closely related to cancer, however various among different modules. Jak-STAT signaling pathway was found to be representative in modules M3 and M4. In a previous study by Fulop *et al.* [43], it was found that the signalling of

Table 2. Gene modules of the aging network.

Module	Size	Genes (Entrez ID)	TOP 5 Enriched GO BP Terms	TOP 5 Enriched HPO Terms	TOP 5 Enriched KEGG Pathways
M1	36	25, 207, 472, 581, 596, 641, 672, 675, 701, 1029, 1050, 1499, 1956, 2064, 2308, 3265, 4193, 4292, 4609, 5159, 5422, 5728, 5781, 5925, 6794, 7015, 7157, 7486, 9184, 1385, 7153, 627, 1649	regulation of apoptosis, cell cycle process, regulation of programmed cell death, regulation of cell death, regulation of cell cycle	Neoplasm, Neoplasm by anatomical site, Neoplasm by histology, Sarcoma, Hematological neoplasm	Pathways in cancer, Prostate cancer, Endometrial cancer, Glioma, Bladder cancer
M2	26	545, 1387, 2010, 2033, 2068, 2073, 2074, 2260, 3479, 3480, 4000, 4036, 4792, 4803, 5979, 7020, 7314, 7341, 7415, 7507, 5830, 1950, 1161, 847, 1490, 2067	DNA metabolic process, response to UV, response to radiation, DNA repair, nucleotide-excision repair	Intrauterine growth retardation, Aplasia/Hypoplasia of the mandible, Micrognathia, Defective DNA repair after ultraviolet radiation damage, Abnormality of the mandible	Nucleotide excision repair, Prostate cancer, Pathways in cancer, Melanoma, Adherens junction
M3	17	367, 2099, 2353, 2690, 2908, 3630, 3643, 3952, 3953, 5449, 5578, 6777, 7040, 8626, 8820, 2688, 5626	response to hormone stimulus, response to endogenous stimulus, response to organic substance, positive regulation of macromolecule metabolic process, response to estrogen stimulus	Abnormality of the anterior pituitary, Abnormality of the pituitary gland, Abnormality of the endocrine system, Abnormality of the hypothalamus-pituitary axis, Anterior hypopituitarism	Jak-STAT signaling pathway, Neuroactive ligand-receptor interaction, Cytokine-cytokine receptor interaction, Aldosterone-regulated sodium reabsorption, Pathways in cancer
M4	11	355, 2071, 3561, 3575, 4683, 4791, 5295, 5580, 6774, 6929, 5336	cell activation, B cell activation, lymphocyte activation, leukocyte activation, immune system development	Abnormality of lymphocytes, Abnormal immunoglobulin level, Abnormality of B cell physiology, Abnormality of B cells, Abnormality of humoral immunity	Pathways in cancer, Jak-STAT signaling pathway, Fc epsilon RI signaling pathway, Fc gamma R-mediated phagocytosis, Neurotrophin signaling pathway
M5	9	3064, 4001, 4137, 5155, 6872, 6908, 5663, 6647, 1938	negative regulation of neuron apoptosis, regulation of neuron apoptosis, positive regulation of MAP kinase activity, behavior, regulation of membrane potential	Abnormality of extrapyramidal motor function, Personality changes, Adult onset, Dysarthria, Parkinsonism	Huntington's disease, Basal transcription factors
M6	5	348, 351, 3717, 2876, 5328	regulation of response to external stimulus, induction of apoptosis, induction of programmed cell death, positive regulation of apoptosis, positive regulation of programmed cell death	Long-tract signs, Abnormal bleeding, Abnormalities of the peripheral arteries, Arterial stenosis, Cerebral inclusion bodies	N/A*

* N/A indicates that there are no enriched KEGG pathway (p-value<0.05) for this module.Module M5 only have two enriched KEGG pathway (p-value<0.05).

Gene FOXO4 (Entrez ID: 4303) could not be grouped into a certain module.

doi:10.1371/journal.pone.0115692.t002

IL-2 receptors is altered in T cells and macrophages with aging, mainly in relation to the Jak-STAT pathway.

These results above indicate that HPO-based semantic similarity can provide a different aspect in disease-related studies other than GO.

NOA and hypergeometric gene set enrichment analysis were then performed on the aging network. The "whole-net" strategy [27] was used to choose the reference set in NOA. The top 10 enriched HPO terms in the two enrichment methods are shown in Table 3. It can be seen that both enrichment methods identify neoplasm-related HPO terms as the top hits. However, these two methods give different terms and different ranks of terms. When dealing with gene/disease sets from biological networks, users are suggested to use the NOA method. If the gene sets are not from network data, users can use either hypergeometric or NOA enrichment method.

Application on disease similarity and disease set enrichment analysis

HPOSim can also be used to investigate the phenotypic relationships between diseases. First, 115 cancer related entries were obtained by searching the OMIM database [6] using "cancer"

Table 3. Top 10 enriched HPO terms by the NOA method and hypergeometric enrichment.

Rank	NOA(whole-net)			Hypergeometric Enrichment		
	HPO ID	Description	q-value	HPO ID	Description	q-value
1	HP:0011793	Neoplasm by anatomical site	<1E-14	HP:0002664	Neoplasm	<1E-14
2	HP:0002664	Neoplasm	4.8E-14	HP:0011792	Neoplasm by histology	1.2E-13
3	HP:0007379	Neoplasm of the genitourinary tract	1.6E-5	HP:0011793	Neoplasm by anatomical site	1.1E-12
4	HP:0001156	Brachydactyly syndrome	4E-5	HP:0100242	Sarcoma	3.1E-10
5	HP:0010787	Genital neoplasm	5.1E-5	HP:0004377	Hematological neoplasm	6.9E-8
6	HP:0008069	Neoplasm of the skin	5.7E-4	HP:0000008	Abnormality of female internal genitalia	7.7E-7
7	HP:0001909	Leukemia	3.6E-3	HP:0004375	Neoplasm of the nervous system	7.7E-7
8	HP:0000006	Autosomal dominant inheritance	4.2E-3	HP:0002665	Lymphoma	7.7E-7
9	HP:0000008	Abnormality of female internal genitalia	4.2E-3	HP:0000812	Abnormal internal genitalia	8.2E-7
10	HP:0000812	Abnormal internal genitalia	4.4E-3	HP:0010460	Abnormality of the female genitalia	8.6E-7

Both enrichment methods identify HPO terms related to neoplasm as the top hits. However, these two methods give different enriched terms and different ranks of terms.

doi:10.1371/journal.pone.0115692.t003

or "carcinoma" as the key word. After removing the diseases that are not annotated in the PA sub-ontology of HPO and all the genes, 55 disease entries were remained (see S3 Dataset for detail).

The semantic similarity matrix of the 55 disease entries was constructed using the Resnik measure and "funSimMax" combining method (see S4 Dataset for detail). A hierarchical clustering was then performed and four modules were detected using the same routine as used in the previous case study. HPO enrichment analysis (hypergeometric test) was also performed using HPOSim. The results are shown in Table 4.

The results showed that these four disease modules had different phenotypic features. For example, module M3 included several types of women-only cancer, including breast cancer (OMIM:114480), breast-ovarian cancer (OMIM:604370, OMIM:612555), ovarian cancer (OMIM:167000) and cervical cancer(OMIM:603956). And lung cancer (OMIM:211980) in M3 was the second most commonly diagnosed types of cancer among women in 2013[44].

The result above indicated that HPO-based semantic similarity had potential ability to play an important role in disease classification and other disease-related studies.

Table 4. Disease modules of the cancer entries in OMIM.

Module	Size	Diseases (OMIM ID)	TOP 5 Enriched HPO Terms
M1	22	OMIM:246470, OMIM:114550, OMIM:120435, OMIM:133239, OMIM:137215, OMIM:148500, OMIM:260350, OMIM:276300, OMIM:601228, OMIM:606719, OMIM:608615, OMIM:609310, OMIM:612229, OMIM:612591, OMIM:613244, OMIM:613347, OMIM:613659, OMIM:614331, OMIM:614337, OMIM:614350, OMIM:614385, OMIM:615083	Neoplasm by anatomical site, Neoplasm, Abnormality of the large intestine, Neoplasm of the large intestine, Neoplasm of the gastrointestinal tract
M2	13	OMIM:109400, OMIM:109800, OMIM:114500, OMIM:144700, OMIM:150800, OMIM:176807, OMIM:273300, OMIM:300854, OMIM:312300, OMIM:601518, OMIM:603688, OMIM:605074, OMIM:608089	Neoplasm of the genitourinary tract, Neoplasm, Neoplasm by anatomical site, Genital neoplasm, Urinary tract neoplasm
M3	12	OMIM:603641, OMIM:114480, OMIM:158320, OMIM:167000, OMIM:211980, OMIM:260500, OMIM:275355, OMIM:603956, OMIM:604370, OMIM:612555, OMIM:614456, OMIM:614564	Breast carcinoma, Neoplasm, Neoplasm of the breast, Neoplasm by anatomical site, Abnormality of the breast
M4	6	OMIM:155240, OMIM:171400, OMIM:188470, OMIM:188550, OMIM:202300, OMIM:608266	Neoplasm of the endocrine system, Thyroid carcinoma, Neoplasm of the thyroid gland, Abnormality of thyroid morphology, Neoplasm by anatomical site

OMIM:191600 (URETER, CANCER OF) and OMIM:610644 (PALMOPLANTAR HYPERKERATOSIS WITH SQUAMOUS CELL CARCINOMA OF SKIN AND 46,XX SEX REVERSAL) could not be grouped into a certain module.

doi:10.1371/journal.pone.0115692.t004

Conclusions

HPOSim is an open source R package that contains seven semantic similarity measures and two enrichment analysis based on HPO data. Also, it provides useful functions for disease-related research and can be integrated with other R packages. In future work, we will integrate more similarity measures and other functions, such as visualization of the HPO data.

Supporting Information

S1 Dataset. Aging network after removing the genes that are not annotated in PA sub-ontology of HPO.
(CSV)

S2 Dataset. Semantic similarity matrix of the 102 genes in the aging network.
(CSV)

S3 Dataset. Cancer entries in OMIM.
(XLSX)

S4 Dataset. Semantic similarity matrix of the 55 cancer entries.
(CSV)

Acknowledgments

We thank the Human Phenotype Ontology Consortium for the original ontology and annotation data, and also thank Dr. Xiaoke Ma for reviewing the manuscript.

Author Contributions

Conceived and designed the experiments: YD LG BW XG. Performed the experiments: YD LG BW. Analyzed the data: YD BW XG. Contributed reagents/materials/analysis tools: YD XG. Wrote the paper: YD LG BW XG.

References

1. Köhler S, Bauer S, Horn D, Robinson PN (2008) Walking the Interactome for Prioritization of Candidate Disease Genes. Am J Hum Genet 82: 949–958. doi: 10.1016/j.ajhg.2008.02.013 PMID: 18371930
2. Van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM (2006) A text-mining analysis of the human phenome. Eur J Hum Genet 14: 535–542. doi: 10.1038/sj.ejhg.5201585 PMID: 16493445
3. Oti M, Brunner H (2007) The modular nature of genetic diseases. Clinical Genetics 71: 1–11. doi: 10.1111/j.1399-0004.2006.00708.x PMID: 17204041
4. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25: 309–316. doi: 10.1038/nbt1295 PMID: 17344885
5. Freudenberg J, Propping P (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 18: S110–S115. doi: 10.1093/bioinformatics/18.suppl_2.S110 PMID: 12385992
6. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucl Acids Res 33: D514–D517. doi: 10.1093/nar/gki033 PMID: 15608251
7. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, et al. (2008) The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. The American Journal of Human Genetics 83: 610–615. doi: 10.1016/j.ajhg.2008.09.017
8. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, et al. (2013) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucl Acids Res: gkt1026.

9. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, et al. (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. The American Journal of Human Genetics 84: 524–533. doi: 10.1016/j.ajhg.2009.03.010

10. Vulto-van Silfhout AT, van Ravenswaaij CMA, Hehir-Kwa JY, Verwiel ETP, Dirks R, et al. (2013) An update on ECARUCA, the European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations. Eur J Med Genet 56: 471–474. doi: 10.1016/j.ejmg.2013.06.010 PMID: 23851227

11. The BRIDGE Project. Available: www.bridgestudy.org. Accessed 2014 December 10.

12. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, et al. (2009) Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. The American Journal of Human Genetics 85: 457–464. doi: 10.1016/j.ajhg.2009.09.003

13. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, et al. (2009) Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation. PLoS Biol 7. doi: 10.1371/journal.pbio.1000247 PMID: 19956802

14. Smedley D, Oellrich A, Kohler S, Ruef B, Sanger Mouse Genetics Project, et al. (2013) PhenoDigm: analyzing curated annotations to associate animal models with human diseases. Database 2013: bat025–bat025. doi: 10.1093/database/bat025 PMID: 23660285

15. Hoehndorf R, Schofield PN, Gkoutos GV (2011) PhenomeNET: a whole-phenome approach to disease gene discovery. Nucl Acids Res 39: e119–e119. doi: 10.1093/nar/gkr538 PMID: 21737429

16. OntoSIML. Available: http://phenomebrowser.net/ontosim/. Accessed 2014 December 10.

17. Hoehndorf R, Dumontier M, Gkoutos GV (2012) Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. Bioinformatics 28: 2169–2175. doi: 10.1093/bioinformatics/bts350 PMID: 22711793

18. Prüfer K, Muetzel B, Do H-H, Weiss G, Khaitovich P, et al. (2007) FUNC: a package for detecting significant associations between gene sets and ontological annotations. BMC Bioinformatics 8: 41. doi: 10.1186/1471-2105-8-41 PMID: 17284313

19. Wittkop T, TerAvest E, Evani US, Fleisch KM, Berman AE, et al. (2013) STOP using just GO: a multi-ontology hypothesis generation tool for high throughput experimentation. BMC Bioinformatics 14: 53. doi: 10.1186/1471-2105-14-53 PMID: 23409969

20. Frohlich H, Speer N, Poustka A, BeiSZbarth T (2007) GOSim—an R-package for computation of information theoretic GO similarities between terms and gene products. BMC Bioinformatics 8: 166. doi: 10.1186/1471-2105-8-166 PMID: 17519018

21. Yu G, Li F, Qin Y, Bo X, Wu Y, et al. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 26: 976–978. doi: 10.1093/bioinformatics/btq064 PMID: 20179076

22. Li J, Gong B, Chen X, Liu T, Wu C, et al. (2011) DOSim: An R package for similarity between diseases based on Disease Ontology. BMC Bioinformatics 12: 266. doi: 10.1186/1471-2105-12-266 PMID: 21714896

23. DOSE: Disease Ontology Semantic and Enrichment analysis. Available: http://www.bioconductor.org/packages/release/bioc/html/DOSE.html. Accessed 2014 December 10.

24. Alexa A, Rahnenfuhrer J (2010) topGO: topGO: Enrichment analysis for Gene Ontology. topGO: topGO: Enrichment analysis for Gene Ontology. Available: http://www.bioconductor.org/packages/release/bioc/html/topGO.html. Accessed 2014 December 10.

25. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29. doi: 10.1038/75556 PMID: 10802651

26. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, et al. (2011) Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Research 40: D940–D946. doi: 10.1093/nar/gkr972 PMID: 22080554

27. Wang J, Huang Q, Liu Z-P, Wang Y, Wu L-Y, et al. (2011) NOA: a novel Network Ontology Analysis method. Nucleic Acids Res 39: e87. doi: 10.1093/nar/gkr251 PMID: 21543451

28. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM (2009) Semantic Similarity in Biomedical Ontologies. PLoS Comput Biol 5: e1000443. doi: 10.1371/journal.pcbi.1000443 PMID: 19649320

29. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th international joint conference on Artificial intelligence—Volume 1. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. pp. 448–453. Available: http://dl.acm.org/citation.cfm?id = 1625855.1625914. Accessed 2014 December 10.

30. Lin D (1998) An Information-Theoretic Definition of Similarity. Proceedings of the Fifteenth International Conference on Machine Learning. ICML'98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. pp. 296–304. Available: http://dl.acm.org/citation.cfm?id = 645527.657297. Accessed 2014 December 10.

31. Jiang J, Conrath D (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. International Conference Research on Computational Linguistics (ROCLING X). p. 9008. Available: http://adsabs.harvard.edu/cgi-bin/nph-bib_query?bibcode = 1997cmp.lg. . ..9008J. Accessed 2014 December 10.

32. Schlicker A, Domingues F, Rahnenführer J, Lengauer T (2006) A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 7: 302. doi: 10.1186/1471-2105-7-302 PMID: 16776819

33. Li B, Wang JZ, Feltus FA, Zhou J, Luo F (2010) Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins. arXiv:10010958. Available: http://arxiv.org/abs/1001.0958. Accessed 2014 December 10.

34. Pesquita C, Faria D, Bastos H, Falcão AO, Couto FM (2007) Evaluating GO-based Semantic Similarity Measures. Proceedings of the 10th Annual Bio-Ontologies Meeting. pp. 37–40. Available: http://www.psb.ugent.be/cbd/cco/Bio-Ontologies2007.pdf. Accessed 2014 December 10.

35. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23: 1274–1281. doi: 10.1093/bioinformatics/btm087 PMID: 17344234

36. Lord PW, Stevens RD, Brass A, Goble CA (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19: 1275–1283. doi: 10.1093/bioinformatics/btg153 PMID: 12835272

37. Wang J, Zhang S, Wang Y, Chen L, Zhang X-S (2009) Disease-Aging Network Reveals Significant Roles of Aging Genes in Connecting Genetic Diseases. PLoS Comput Biol 5: e1000521. doi: 10.1371/journal.pcbi.1000521 PMID: 19779549

38. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28: 27–30. doi: 10.1093/nar/28.1.27 PMID: 10592173

39. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols 4: 44–57. doi: 10.1038/nprot.2008.211

40. Fisher GJ, Wang Z, Datta SC, Varani J, Kang S, et al. (1997) Pathophysiology of Premature Skin Aging Induced by Ultraviolet Light. New England Journal of Medicine 337: 1419–1429. doi: 10.1056/NEJM199711133372003 PMID: 9358139

41. Silverberg SJ, Shane E, de la Cruz L, Segre GV, Clemens TL, et al. (1989) Abnormalities in Parathyroid Hormone Secretion and 1,25-Dihydroxyvitamin D3 Formation in Women with Osteoporosis. New England Journal of Medicine 320: 277–281. doi: 10.1056/NEJM198902023200503 PMID: 2911322

42. Sano T, Kovacs KT, Scheithauer BW, Young WF (1993) Aging and the Human Pituitary Gland. Mayo Clinic Proceedings 68: 971–977. doi: 10.1016/S0025-6196(12)62269-1 PMID: 8412363

43. Fulop T, Larbi A, Douziech N, Levesque I, Varin A, et al. (2006) Cytokine receptor signalling and aging. Mechanisms of Ageing and Development 127: 526–537. doi: 10.1016/j.mad.2006.01.025 PMID: 16530252

44. Siegel R, Naishadham D, Jemal A (2013) Cancer statistics, 2013. CA: A Cancer Journal for Clinicians 63: 11–30.