Similarity Calculation, Enrichment Analysis, and Ontology Visualization of Biomedical Ontologies using UFO

Quang-Huy Nguyen[1] and Duc-Hau Le[1,2,3]

[1]Department of Computational Biomedicine, Vingroup Big Data Institute, Hanoi, Vietnam
[2]School of Computer Science and Engineering, Thuyloi University, Hanoi, Vietnam
[3]Corresponding author: *hauldhut@gmail.com*

The rapid growth of biomedical ontologies observed in recent years has been reported to be useful in various applications. In this article, we propose two main-function protocols—term-related and entity-related—with the three most common ontology analyses, including similarity calculation, enrichment analysis, and ontology visualization, which can be done by separate methods. Many previously developed tools implementing those methods run on different platforms and implement a limited number of the methods for similarity calculation and enrichment analysis tools for a specific type of biomedical ontology, although any type can be acceptable. Moreover, depending on each application, methods have distinct advantages; thus, the greater the number of methods a tool has, the better decisions that users make. The protocol here implements all the analyses above using an advanced popular tool called UFO. UFO is a Cytoscape app that unifies most of the semantic similarity measures for between-term and between-entity similarity calculation for biomedical ontologies in OBO format, which can calculate the similarity between two sets of entities and weigh imported entity networks, as well as generate functional similarity networks. The complete protocol can be performed in 30 min and is designed for use by biologists with no prior bioinformatics training. © 2021 Wiley Periodicals LLC.

Basic Protocol: Running UFO using a list of input Gene Ontology, Disease Ontology, or Human Phenotype Ontology data

Keywords: Cytoscape App • enrichment analysis • ontology • ontology visualization • similarity calculation

How to cite this article:
Nguyen, Q., & Le, D. (2021). Similarity calculation, enrichment analysis, and ontology visualization of biomedical ontologies using UFO. *Current Protocols*, *1,* e115. doi: 10.1002/cpz1.115

INTRODUCTION

The establishment of numerous biomedical ontologies (Smith et al., 2007)—e.g., Gene Ontology (GO; see Gene Ontology, 2004), Human Phenotype Ontology (HPO; see Köhler et al., 2014), and Disease Ontology (DO; see Kibbe et al., 2015)—has been confirmed to be useful in many biomedical applications, such as annotating biomedical entities (Hoehndorf, Schofield, & Gkoutos, 2015): e.g., genes (Barrell et al., 2009; Peng et al., 2013) or phenotypes (Köhler et al., 2014). To extract information on these data, the three most widely utilized analyses employed are: similarity calculation between ontology terms and between the annotated biomedical entities, enrichment analysis on a set of biomedical entities, and ontology visualization. Several computational methods have been developed to implement these analyses (Bindea et al., 2009; Deng, Gao, Wang, &

CURRENT PROTOCOLS
A Wiley Brand

Current Protocols e115, Volume 1
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpz1.115
© 2021 Wiley Periodicals LLC

Table 1 Definitions

Ontology	A framework in which to represent our knowledge in a specific scientific domain that is convenient for sharing our information using familiar language. In case of this protocol, we see it in a context of biomedicine, and we call it a biomedical ontology. There are many types of biomedical ontology data, and users can explore them at *http://www.obofoundry.org/*.
Open Biomedical Ontology (OBO) foundry (Smith et al., 2007)	This is a format type for biology and biomedical ontology data that is widely observed. It presents the data in graph-based knowledge representation language.
Directed Acyclic Graphs (DAG)	This is a common presentation of ontology data in which each node of the graph represents each term, and a node connects with others by edges. Relationships of each node may involve several parents and several children. Accordingly, there are two indicators for this relationship appearing in the graph—"is-a" and "part-of"—which show, respectively, that the child is a subclass or a component of the parent.

Guo, 2015; Du et al., 2009; Fröhlich, Speer, Poustka, & Beißbarth, 2007; Köhler et al., 2009; Li et al., 2011; Martin et al., 2004; Yu, Wang, Yan, & He, 2014) based on various platforms like Cytoscape (Shannon et al., 2003), R, Python, or the web, and using a limited set of measures for a specific type of biomedical ontology (e.g., GO, HPO, DO). However, there are some options available that seek to combine the power of these separate platforms into one, so that multiple analyses can be applied to a single dataset, such as UFO (Le, 2020) and SML-toolkit (Harispe, Ranwez, Janaqi, & Montmain, 2014). Also, a small number of tools are available with a visualization implementation, such as UFO (Le, 2020), Gorilla (Eden, Navon, Steinfeld, Lipson, & Yakhini, 2009), Golorize (Garcia et al., 2006), and g:Profiler (Reimand, Kull, Peterson, Hansen, & Vilo, 2007). It is of note that, because all biomedical ontologies exist is in the same structure, i.e., directed acyclic graph, any semantic similarity measure can be applied to them. Furthermore, depending on each application, different methods can have distinct advantages (Mazandu & Mulder, 2014; Pesquita, Faria, Falcão, Lord, & Couto, 2009); thus, the more methods a tool uses, the better decisions users can make.

The following Basic Protocol provides a step-by-step guide on how to use UFO to compute the similarity, analyze the enrichment, visualize the ontology, and weigh an entity network. To compute the similarity, our tool uses the existing similarity measures that are defined for ontology terms and annotated entities (see the review in Pesquita et al., 2009). More specifically, UFO possesses 11 between-term similarity measures—there are four pairwise between-entity similarity measures requiring the user to choose one of the between-term similarity measures before using them—and seven groupwise between-entity similarity measures. All of the between-term and between-entity measures, as well as technical and mathematical details, are extensively discussed in our original paper (Le, 2020).

Overview of the procedure

This section summarizes the two major stages of UFO—definition of ontology data of interest as well as its corresponding annotation data—and offers step-by-step implementation details on the two main functions with their accompanying analysis types [see Fig. 1, Table 1, and our original paper (Le, 2020) for basic definitions]. In the first stage, all necessary ontology data (e.g., GO, DO, HPO, …) and its corresponding annotation data are downloaded and input into UFO in order to serve for downstream analyses (step

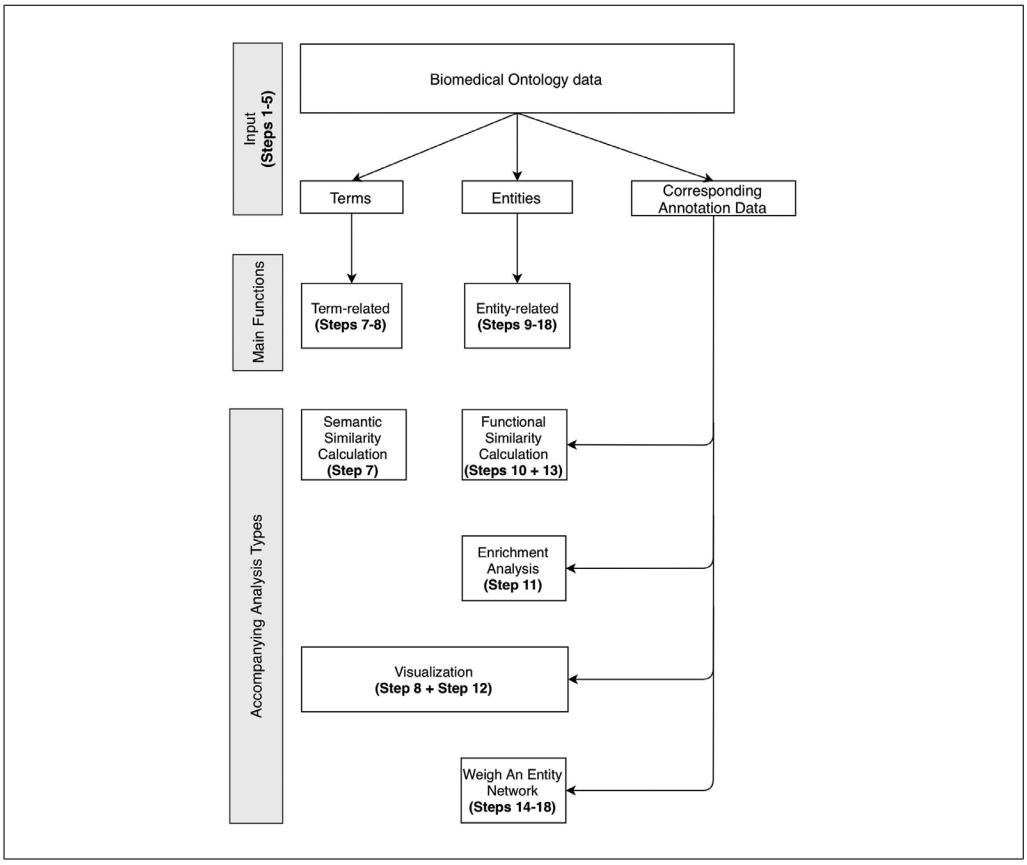


Figure 1 Protocol overview. Biomedical ontology data (e.g., GO, HPO, DO) and its corresponding annotation data are input into UFO, which runs entirely on the Cytoscape platform, to calculate the similarity, analyze the enrichment, visualize the relationships, and weigh the entity network. A complete description is given in the Basic Protocol.

1-5). Here, UFO has two main functions: term-related and entity-related. For the term-related function, ontology terms undergo between-term similarity calculation (step 7), visualization between relationships between terms, and visualization between terms and their annotated entities (step 8). For entity-related function, there are four accompanying analysis types for biomedical entities: between-entity similarity calculation (steps 10 and 13), entity set enrichment analysis (step 11), visualization of functional relations between entities and the relationship between entities and their annotating terms (step 12), and weighted entity network construction (steps 14-18). In addition, when it comes to the between-entity similarity calculation, there are two sub-types: calculation of functional similarity among entities (step 10) and between two entity sets (step 13).

Strategic Planning

Application to diverse ontology data

This protocol uses three biomedical ontologies, GO, HPO, and DO, and their corresponding annotation data as examples, since they have been employed popularly. Nevertheless, many other ontologies are available and can be downloaded at *http://www.obofoundry.org*, and additional annotation data may be applicable to the current protocol, such as Drug Ontology (Sharp, 2017) or Cell Ontology (Diehl et al., 2016; Osumi-Sutherland, 2017).

Experimental design

The prerequisite for ontology data is in a standard format, OBO. In contrast, the annotation data is in a format similar to the following (for each line): `EntityID<tab>`

`OntologyTermID<tab>EvidenceCode` (optional), where `EntityID` is the identifier of biomedical entities; `OntologyTermID` is the identifier of the annotating ontology term; and `EvidenceCode` is evidence code of that annotation. When evidence cannot be determined, please leave blank.

Advantages and limitations of UFO

UFO is one of the rare Cytoscape apps that has the ability to perform all of the three most common ontology analysis types with a large number of similarity calculation and enrichment analysis methods, for all types of biomedical ontologies. Benefitting from that, users can seek the absolutely most optimal similarity calculation measures for their biological application, as demonstrated in Le (2020) and Le, Pham, & Dao (2016). Furthermore, the Cytoscape platform is a powerful tool for integration and visualization, and many other apps have been designed to run on this platform to solve real-world biomedical problems. The development of UFO as an app on this platform takes advantage of these strengths.

The following limitations should be considered before applying this protocol to one's research. Firstly, both the ontologies themselves and the level of annotation of biomedical entities are constantly evolving; thus, future investigations should take the amount of annotation and the ontology's latest release date (Kulmanov & Hoehndorf, 2017) into consideration when evaluating the performance of the similarity measures. Secondly, UFO can only work with biological ontologies in OBO format; however, other ontology formats [e.g., Resource Description Framework (RDF), Web Ontology Language (OWL),…] also exist . However, this limitation can be solved by format-conversion tools such as ROBOT (Jackson et al., 2019). Lastly, the computation time when using measures of high computational complexity: e.g., the hybrid semantic similarity calculation method (Wang, Du, Payattakool, Yu, & Chen, 2007) when performing batch jobs (e.g., functional similarity calculation for thousands to millions of entity pairs) is a potential restriction.

BASIC PROTOCOL

RUNNING UFO USING A LIST OF INPUT GENE ONTOLOGY, DISEASE ONTOLOGY, OR HUMAN PHENOTYPE ONTOLOGY DATA

This step-by-step protocol describes how to compute the similarity between terms, between entities, and between two entity sets; analyze the enrichment for a set of phenotypes belonging to a phenotypic series; visualize semantic relationships between ontology terms, annotating relationships between entities and terms, and functional similarity between entities; and weigh a physical protein interaction network,

Necessary Resources

Hardware

A personal computer with internet access and ≥8 GB of RAM (required to run Cytoscape)

Software

Java Standard Edition (*http://www.oracle.com/technetwork/java/javase/downloads/index.html*), version 8 or higher, is needed to run Cytoscape
The Cytoscape desktop application (*http://www.cytoscape.org/download.php*), version 3.6.0 or higher, is the platform on which to run UFO
UFO app is available at Cytoscape App Store (*http://apps.cytoscape.org/apps/ufo*)

Input Data

Users can find downloadable example files that are mentioned throughout the protocol on our Github repository (*https://github.com/hauldhut/ufo*). We

recommend creating a `Data` folder in the Cytoscape folder (e.g., `C:\Program Files\Cytoscape_v3.7.2` in Windows, `Applications\Cytoscape_v3.7.2` in Mac) to store all these files. We also recommend creating an additional `Results` folder in the same working directory above to preserve the analysis results generated while performing the protocol.

For Ontology data: Users have two ways to download:

- Download example Ontology files at the Github repository (*https://github.com/hauldhut/ufo*). In the `Data/Ontology` folder, the `go.obo`, `hp.obo`, and `HumanDO.obo` files correspond to the GO, HPO, and DO data in OBO Foundry.
- Download example files from original resources: Gene Ontology (`go.obo`) from *http://www.geneontology.org/*, and Human Phenotype Ontology (`hp.obo`) and Human Disease Ontology (`doid.obo`) at The Open Biological and Biomedical Ontologies (*http://www.obofoundry.org/*)

For Annotation data: Users also have two ways to download:

- Download corresponding Annotation data from the `Data/Annotation` and the `Data/Annotation-Preprocessed` folders at the Github repository, in which the former includes original annotation data, whereas the latter includes preprocessed data
- Download example files from original resources: Gene Ontology to Gene (`gene2go`) at the NCBI FTP site (*ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz*), Disease Ontology to Gene (`IDMappings.rdf`) at *https://academic.oup.com/nar/article/41/D1/D553/1071527*, and Human Phenotype Ontology to Phenotype (`phenotype_annotation_hpoteam.tab`) at *http://www.human-phenotype-ontology.org/*

For a protein interaction network: Users also have two ways to download:

- Download example file (`interactions_extracted.txt`), which we pre-processed by only preserving interactions for human (i.e., `#tax_id = 9606`) and removing those where one of the interactants is missing at the Github repository. Proteins are identified by Entrez Gene ID.
- Download example file (`interactions.gz`) from original resource at the NCBI FTP site (*ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interactions.gz*)

NOTE: It is fine for the user to use their own interaction network. In this protocol, we use a gene rif file as an example to visualize how to use UFO. The user can use any other sources, the only condition being that proteins in the PPI network must be identified by Entrez Gene ID.

Install software

1. Install Java following the link provided above in Necessary Resources, "Software".
2. Download the recommended version of Cytoscape from *http://www.cytoscape.org/download.php*, then put it into the default folder.
3. Create the two folders (i.e., `Data` and `Results` folders), download required input files, and place them into the `Data` folder (see Necessary Resources, "Input Data").
4. Install the UFO app in Cytoscape.
   i. Launch Cytoscape.
   ii. Open the *Apps* menu and select the *App Manager.*
   iii. Click on *Install from File…* and find the `UFO.jar` file to install the app in Cytoscape.
   iv. Confirm that UFO is installed in Cytoscape by viewing the *Currently Installed* tab.

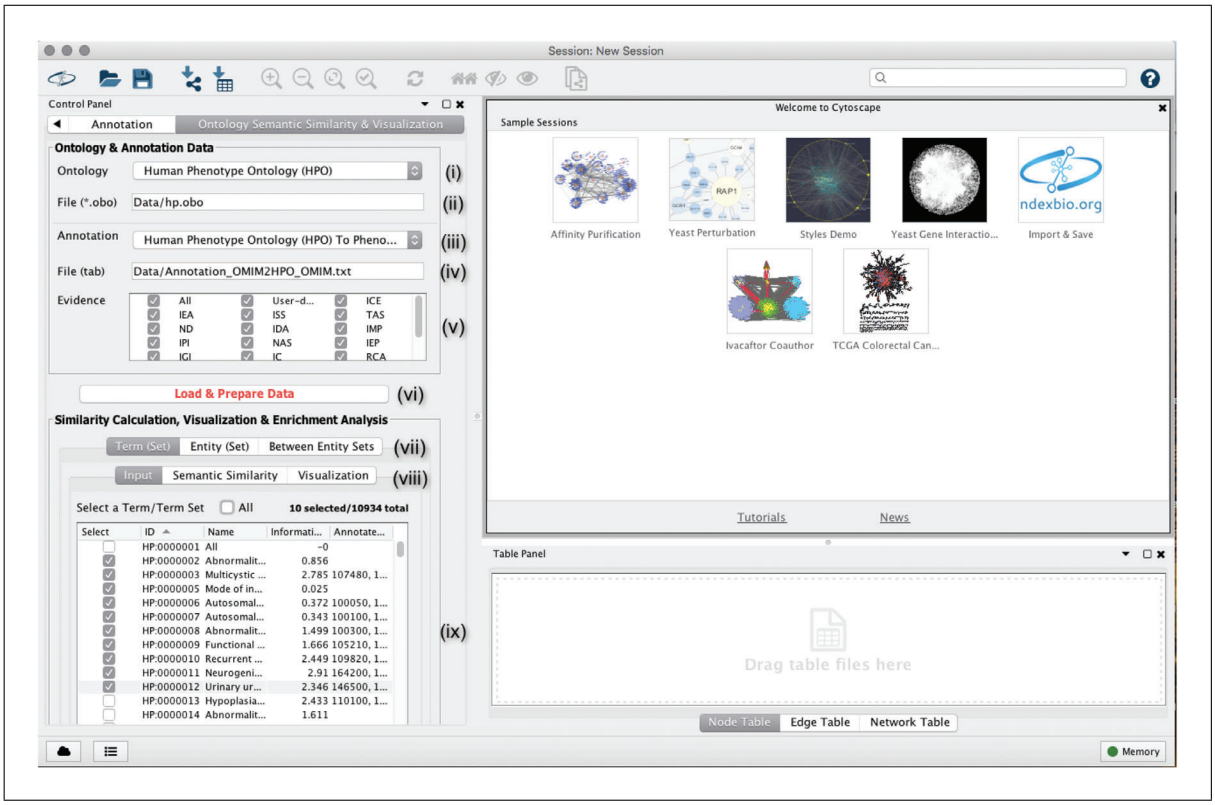


Figure 2 Screenshot of UFO in Cytoscape. Steps are marked by roman numerals (step 5, **i-ix**). Boxes (**ii**), (**iii**), and (**iv**) should be filled out automatically when the user selects ontology data of interest.

Load and prepare data

We assume that all essential files have been downloaded, including the three ontology data files from GO, DO, and HPO and corresponding annotation data, and placed in the `Data` folder as suggested in the steps above.

5. Re-open the *Apps* menu, select *UFO*, and then select *Main Function…* . Figure 2 shows initial requirements on input selection and selection of the three main analysis types, as well as the calculation of similarity, enrichment analysis, and the ontology visualization. Choose input and analysis types based on users' research questions and ontology data.
   i. Select from the types of ontology data that are listed, i.e., Gene Ontology (GO)–Biological Process; Gene Ontology (GO)–Cellular Component; Gene Ontology (GO)–Molecular Function; Human Disease Ontology (DO); and Human Phenotype Ontology (HPO).

      Users may choose Select other Ontology data… for their own other ontology data.

   ii. The path to the selected ontology data is shown, e.g., `Data/go.obo`, `Data/doid.obo`, and `Data/hpo.obo`.
   iii. Select the corresponding annotation data, i.e., Gene Ontology (GO) To Gene; Disease Ontology (DO) To Gene; and Human Phenotype Ontology (HPO) To Phenotype.

      Users may choose Select other Annotation data… for their own other annotation data.

   iv. The path to the selected annotation data is shown, e.g., `Data/Annotation_Gene2GO_9606_BP.txt`, `Data/Annotation_Gene2GO_9606_CC.txt`, `Data/Annotation_Gene2GO_9606_MF.`

`txt`; `Data/Annotation_OMIM2HPO_OMIM.txt`; and `Data/Annotation_Gene2DO.txt`.

v. Select the evidence code for that annotation. Check the first box, All, to have all annotations from selected Annotation data chosen; conversely, users can specifically choose the evidence code of their interest. In addition, users can define their own Evidence codes accordingly by modifying the User-defined box

vi. Click on Load & Prepare Data to move selected ontology data and corresponding annotation data to UFO, and subsequently compute Information Content (IC) for each Ontology term.

vii. There are two main functions: term-related includes Term (Set), or entity-related includes Entity (Set) and Between Entity sets.

viii. Show Input (i.e., terms or entities) and accompanying analysis types for that input. Specifically, for the Term (Set), UFO provides two analysis types: Semantic Similarity and Visualization (steps 7 and 8, respectively). For Entity (Set), UFO provides three analysis types: Semantic Similarity, Enrichment, and Visualization. Lastly, for Between Entity Sets, the only analysis type that UFO provides is the calculation of functional similarity between two entity sets.

ix. Show the information of terms and entities (e.g., ID, Name, Information Content, …) and the results of accompanying analyses above.

Implementation of each main function in UFO

UFO provides two main functions (i.e., term-related and entity-related) with accompanying analysis types. Depending on the experimental questions interrogated and the ontology data chosen, users can select one of these. Users do not need all of the main functions and analysis types provided in UFO.

A. Term-related functions (sematic similarity calculation and visualization)

6. After clicking on *Load & Prepare Data* (step 5, vi), the information of a term/terms is shown in Figure 2 (**ix**). The HPO data and corresponding annotation data, and the first ten terms are selected as example selections, also shown in Figure 2.

   Semantic similarity calculation: TIMING 3s–18s/10 terms/each measure.

   Computation time depends greatly on the number of terms the user selects. The more terms selected, the more computation time increases. In this protocol, we only choose the first ten terms to perform between-term similarity calculation (step 6). Also, we only perform semantic similarity calculation using HPO data; the user may use alternative ontology data (e.g., GO, DO,…) when performing the semantic similarity calculation.

7. Move to tab *Semantic Similarity* and start to calculate the between-term similarity (Fig. 3A).

   i. UFO provides users with three approaches: *Node-based*, *Edge-based*, and *Hybrid-based* (details can be seen in our original paper: Le, 2020).

   ii. UFO also provides corresponding measures of similarity calculation in each of the approaches (details can be seen in our original paper: Le, 2020).

   iii. Click on *Calculate* to calculate semantic similarity matrix.

   *As we mentioned earlier in "**Advantages and limitations of UFO**," there can be a great difference in computation time between similarity measures, with the hybrid semantic similarity approach taking the longest. Users should consider this before choosing which measure to use in their research. In this case, we choose the Resnik method (Resnik, 1995) assigned to the Node-based category as an example.*

   iv. Check the *Name View* box to turn Term ID (column *ID*; Fig. 2, **ix**) to Term Name (column *Name*; Fig. 2, **ix**), placed at row and column names of the similarity matrix. The opposite will occur if the box is left blank.

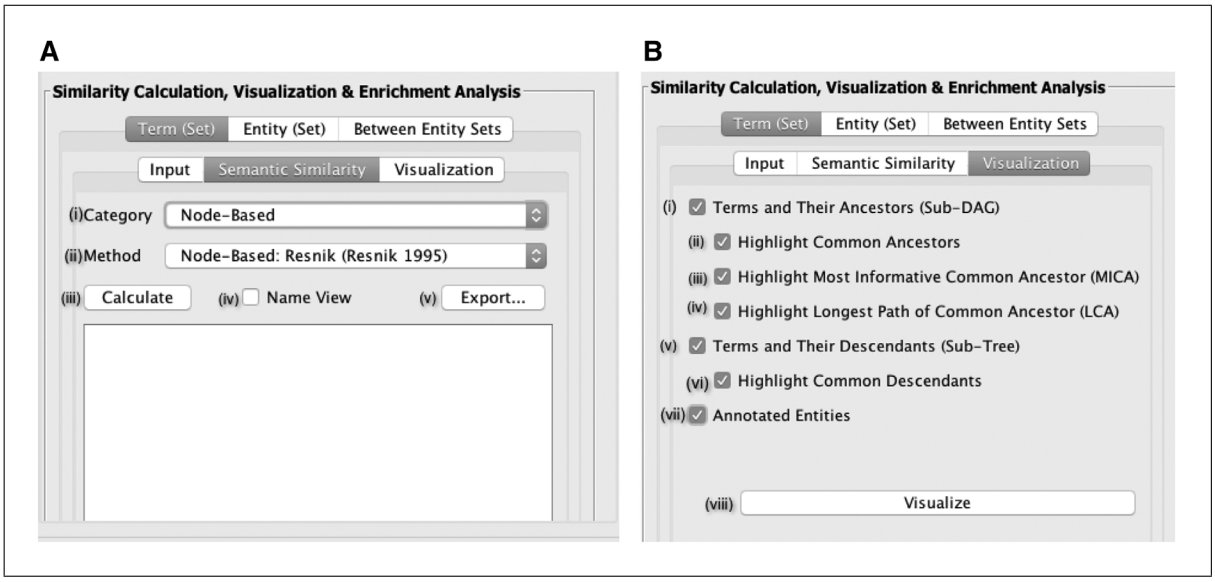


Figure 3 Two accompanying analysis types provided in term-related function. (**A**) Steps to perform between-term similarity calculation are marked by roman numerals (step 7, **i-v**). The number of measures of similarity varies (step 7, **ii**), depending on which category is selected by the user (step 7, **i**). (**B**) Steps to visualize terms of interest and annotated entities are marked by roman numerals (step 8, **i-viii**).

A

ID	HP:0000002	HP:0000003	HP:0000005	HP:0000006	HP:0000007	HP:0000008	HP:0000009	HP:0000010	HP:0000011	HP:0000012
HP:0000002	0.61771575	0.00652007	0	0	0	0.00652007	0.00652007	0.00652007	0.00652007	0.00652007
HP:0000003	0.00652007	1.74694303	0	0	0	0.63735676	0.808836	0.808836	0.808836	0.808836
HP:0000005	0	0	0	0	0	0	0	0	0	0
HP:0000006	0	0	0	0.02498776	0.02498776	0	0	0	0	0
HP:0000007	0	0	0	0.02498776	0.02498776	0	0	0	0	0
HP:0000008	0.00652007	0.63735676	0	0	0	1.44757382	0.63735676	0.63735676	0.63735676	0.63735676
HP:0000009	0.00652007	0.808836	0	0	0	0.63735676	1.6113613	0.808836	1.6113613	1.6113613
HP:0000010	0.00652007	0.808836	0	0	0	0.63735676	0.808836	1.21487947	0.808836	0.808836
HP:0000011	0.00652007	0.808836	0	0	0	0.63735676	1.6113613	0.808836	1.66555724	1.66555724
HP:0000012	0.00652007	0.808836	0	0	0	0.63735676	1.6113613	0.808836	1.66555724	1.66555724

B

ID	Abnormality	Multicystic k	Mode of inhe	Autosomal d	Autosomal re	Abnormality	Functional al	Recurrent ur	Neurogenic b	Urinary urgency
Abnormality of body height	0.61771575	0.00652007	0	0	0	0.00652007	0.00652007	0.00652007	0.00652007	0.006520073
Multicystic kidney dysplasia	0.00652007	1.74694304	0	0	0	0.63735676	0.808836	0.808836	0.808836	0.808835999
Mode of inheritance	0	0	0	0	0	0	0	0	0	0
Autosomal dominant inheritance	0	0	0	0.02498776	0.02498776	0	0	0	0	0
Autosomal recessive inheritance	0	0	0	0.02498776	0.02498776	0	0	0	0	0
Abnormality of female internal genitalia	0.00652007	0.63735676	0	0	0	1.44757382	0.63735676	0.63735676	0.63735676	0.63735676
Functional abnormality of the bladder	0.00652007	0.808836	0	0	0	0.63735676	1.6113613	0.808836	1.6113613	1.611361299
Recurrent urinary tract infections	0.00652007	0.808836	0	0	0	0.63735676	0.808836	1.21487947	0.808836	0.808835999
Neurogenic bladder	0.00652007	0.808836	0	0	0	0.63735676	1.6113613	0.808836	1.66555724	1.665557244
Urinary urgency	0.00652007	0.808836	0	0	0	0.63735676	1.6113613	0.808836	1.66555724	1.665557244

Figure 4 Resulting similarity matrix. (**A**) Leave the *Name View* box blank in step 7, **iv**. (**B**) Check the *Name View* box in step 7, **iv**. Row and column names are terms selected in step 6. Each cell is a similarity value.

v. Click on *Export…* to save your analysis results. Please name it (e.g., `SS-HPOterm`) and place it in the `Results` folder. Figure 4 issues an illustration of the given similarity matrix.

Visualization of terms of interest and annotated entities

Again, we just choose the first ten terms and HPO data to perform the visualization of selected terms (step 6).

8. Move to the *Visualization* tab and start to visualize terms of interest and annotated entities (Fig. 3B).

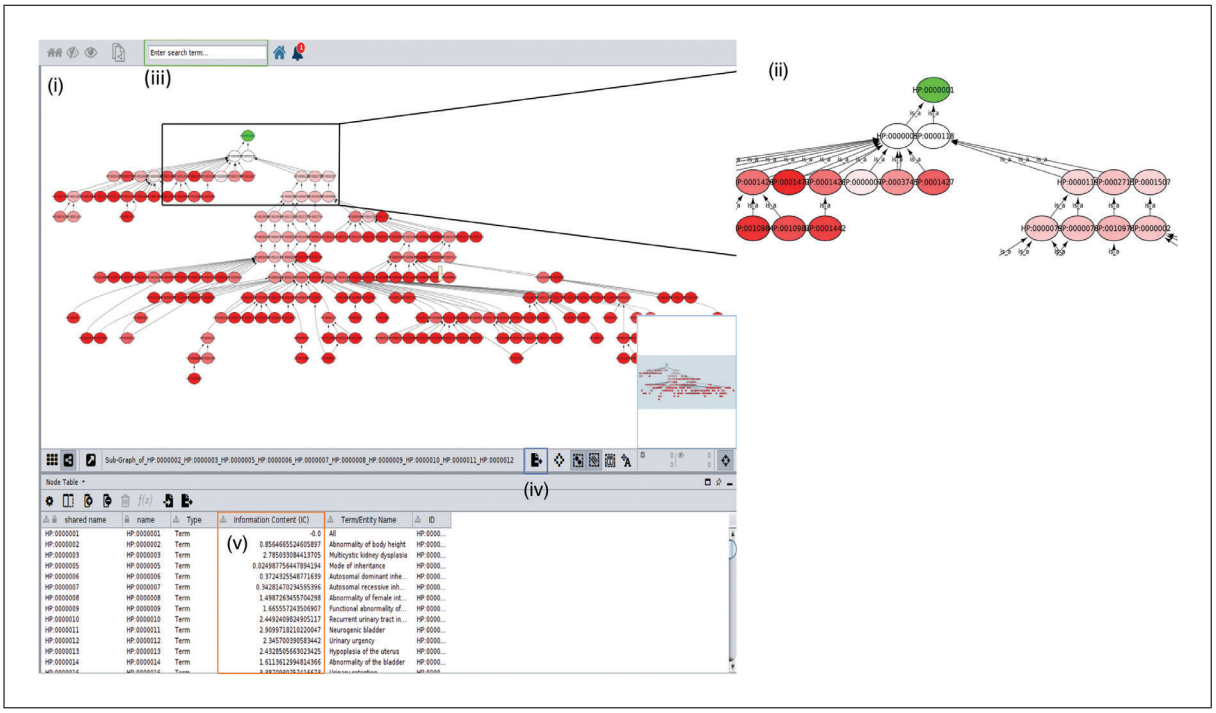


Figure 5 Resulting publication-ready network in a directed acyclic graph. (**i**) Overall thumbnail view of publication-ready network created with all of the checked boxes. (**ii**) Zoomed-in section of publication-ready network, in which the green node is the root, whereas color levels (white to bold red) indicate a corresponding increase in the IC of the terms. (**iii**) Cytoscape search bar, which can be used to search for terms of interest in the network. (**iv**) Save the resulting relationship with two options: *Export as Network…* or *Export as Image…* . (**v**) ICs of HPO terms.

9. Depending on what you want to visualize, check the appropriate box provided in UFO to show Common Ancestors (**ii**), MICA (**iii**), LCA (**iv**), Common Descendants (**vi**), and Annotated Entities (**vii**). Each of these are thoroughly discussed in our original paper (Le, 2020).

   *In this protocol, we choose all existing boxes, although this will make our resulting network extremely complicated. Alternatively, users may check individual boxes [Fig. 3B, (**i**), (**v**), or (**vii**)] or a combination of two out of three boxes to successfully grasp the information on a relationship among the terms of interest and annotated entities.*

10. Click on *Visualize* to show your analysis results. Figure 5 illustrates the relationship/network in a directed acyclic graph. To save, select *Export as Image…*. Please name it (e.g., `Viz-HPOterm`) and place it in the `Results` folder. Note that users also can seek terms of interest in the network using the search bar (Fig. 5, **iii**).

B. Entity-related functions

The HPO data, corresponding annotation data, and the first ten entities are selected as example selections, shown in Figure 6A.

Functional similarity calculation

11. Move to the *Functional Similarity* tab and start to calculate the between-entity similarity (Fig. 6B).

    i. Select one of the two approaches: Pairwise and Groupwise (descriptions of these two approaches can be seen in our original paper; Le, 2020).

    *For the Pairwise approach, since the semantic similarity calculation for a pair of terms annotating to each entity must be performed first, users need to predetermine a certain semantic similarity measure of interest in the Semantic Similarity sub-tab of the Term (Set) tab, as seen in step 7, a-b). Afterwards, step 9 (**i-iii**) is performed. In contrast, the Groupwise category does not require any precalculation.*

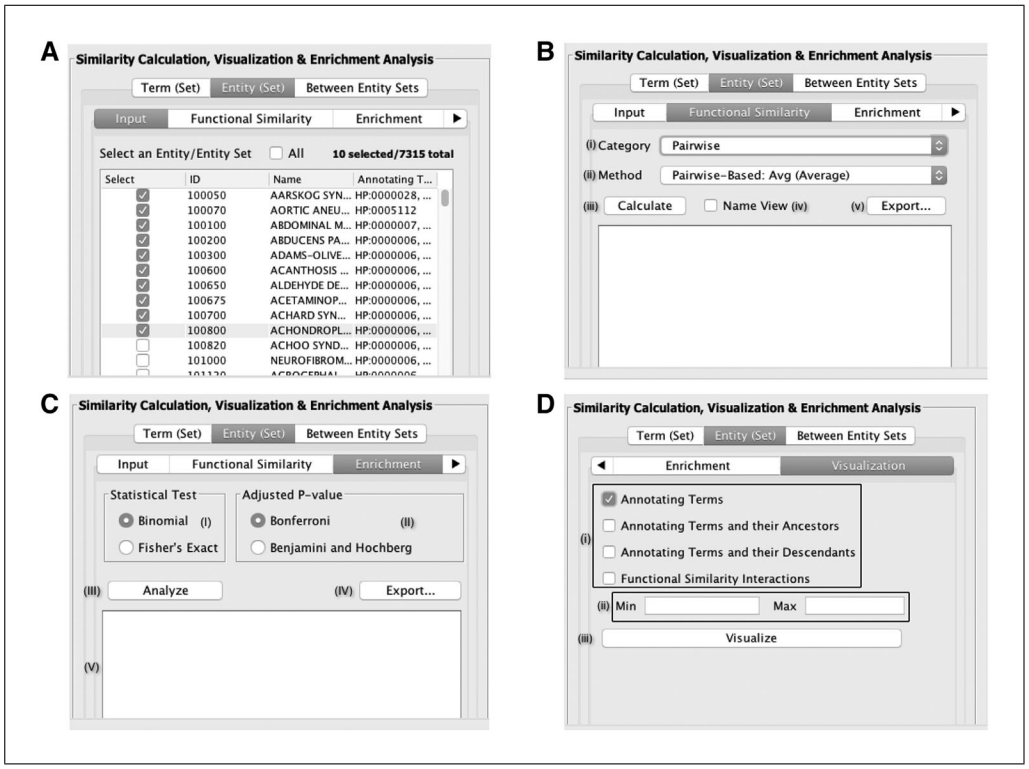


Figure 6 Input entities and three accompanying analysis types provided in Term (Set) function. (**A**) The first ten entities of interest are selected as examples. (**B**) *Functional Similarity* sub-tab to perform functional similarity calculation (step 10, **i-v**). (**C**) *Enrichment* sub-tab to perform term enrichment analysis (step 11, **i-iv**). (**D**) *Visualization* sub-tab to perform visualization of terms and annotated entities (step 12, **i-iii**). Analysis steps are marked by roman numerals and/or rectangles.

UFO also provides corresponding measures of similarity calculation in each the approaches (details can be seen in our original paper; Le, 2020).

ii. Click on *Calculate* to compute the functional similarity of entities of interest.

In this case, a predetermined semantic similarity method is the Node-based JC (Jiang & Conrath, 1997) to generate a semantic similarity matrix. Then, we choose the Average method belonging to the functional similarity method to compute the between-entity similarity.

iii. Check the *Name View* box to turn Term ID (column *ID*; Fig. 2, **ix**) to Term Name (column *Name*; Fig. 2, **ix**), placed at row and column names of the similarity matrix,. The opposite will occur if the box is left blank.

iv. Click on *Export…* to save your analysis results. Please name it (e.g., `FS-HPOentity`) and place it in the `Results` folder.

Entity set enrichment analysis

12. Move to the *Functional Similarity* tab and start to perform the enrichment for the set of selected entities (Fig. 6C). For detailed explanation of enrichment, please see our original paper (Le, 2020).

    i. Choose the appropriate *Statistical Test* (*Binomial* or *Fisher's Exact*) and *Adjusted p-value* (*Bonferroni* or *Benjamini-Hochberg*) for your needs (details can be found in our original paper; Le, 2020).

    *Implementation of the enrichment analysis is on the basis of selection of both semantic similarity and functional similarity measures, which are predetermined. In this protocol, we choose the Jiang-Conrath method assigned to the Node-based category (step 7, a-b) and the Average method assigned to the Pairwise category (step 9, **i-ii**) as an example. Also, we choose the statistical test of Fisher's Exact and the adjusted p-value of Benjamini-Hochberg.*

A

Term	Enriched Phenotype	P-value	Q-value
HP:0002023	Anal atresia	0.00419764	0.02938347
HP:0000394	Lop ear	0.00544805	0.02984585
HP:0002677	Small foramen magnum	0.00544805	0.02984585
HP:0002682	Broad skull	0.00544805	0.02984585
HP:0003093	Limited hip extension	0.00544805	0.02984585
HP:0004392	Prune belly	0.00544805	0.02984585
HP:0002055	Curved linear dimple below the lower lip	0.00273074	0.03127943
HP:0002512	Brain stem compression	0.00273074	0.03127943
HP:0003318	Cervical spine hypermobility	0.00273074	0.03127943
HP:0003533	Delayed oxidation of acetaldehyde	0.00273074	0.03127943
HP:0004792	Rectoperineal fistula	0.00273074	0.03127943
HP:0005733	Spinal stenosis with reduced interpedicular distance	0.00273074	0.03127943
HP:0001033	Facial flushing after alcohol intake	0.00409107	0.03436503
HP:0006970	Periventricular leukomalacia	0.00409107	0.03436503
HP:0010957	Congenital posterior urethral valve	0.00409107	0.03436503
HP:0030011	Imperforate hymen	0.00409107	0.03436503
HP:0010886	Osteochondrosis dissecans	0.00680168	0.03570883
HP:0000484	Hyperopic astigmatism	0.00815197	0.03804252
HP:0001362	Skull defect	0.00815197	0.03804252
HP:0002781	Upper airway obstruction	0.00815197	0.03804252
HP:0005112	Dilatation of the abdominal aorta	0.00815197	0.03804252
HP:0011349	Abducens palsy	0.00815197	0.03804252
HP:0000951	Abnormality of the skin	0.00762991	0.03845475
HP:0001388	Joint laxity	0.00989431	0.04021558
HP:0002761	Generalized joint laxity	0.01084254	0.04269249
HP:0001156	Brachydactyly syndrome	0.00247806	0.04460509
HP:0001470	Sex-limited autosomal dominant	0.01218283	0.04651627

B

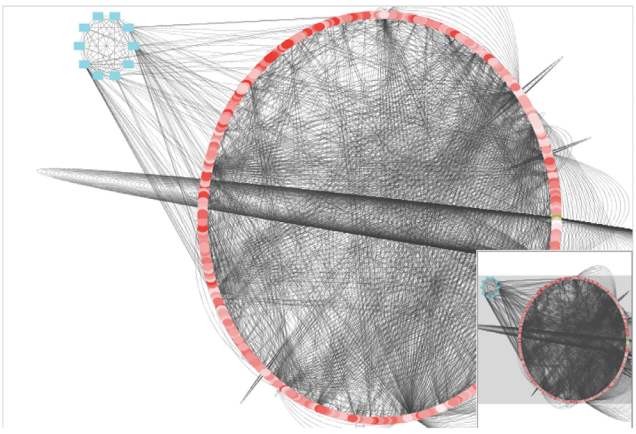


Figure 7 Analysis results from entity-related function. (**A**) Phenotypic series reported only to be enriched significantly for the term set if Q-value $\leq$0.05. p-values and Q-values are computed by Fisher's exact test and the Benjamini-Hochberg procedure, respectively (**B**) Publication-ready functional similarity network between ten entities of interest (turquoise rectangles creating a left top circle) annotating relationships between entities and terms, and a directed acyclic graph of the annotating ontology terms (red spheres creating a right bottom circle).

ii. Click on *Analyze* to perform the enrichment analysis.
iii. To create a table comprising analysis results in terms of significantly enriched terms (Q-value $\leq$0.05), the user clicks on the 'Export' button. At last, please name it (e.g., `EA-HPOterm`) and place it in the `Results` folder. Figure 7A shows an illustration of the results.

Visualization of terms and annotated entities

13. Move to the *Functional Similarity* tab and start to visualize terms of interest and annotated entities (Fig. 6D).
    i. Check boxes of interest. In this protocol, we choose all of them.
    ii. If the user checks the *Functional Similarity Interactions* box, then set thresholds (Min, Max) to visualize functional similarities among entities of interest in the form of a network.
    iii. Click on *Visualize* to see the result (Fig. 7B). To save, select *Export as Image*.... Please name it (e.g., `Viz-HPOterm`) and place it in the `Results` folder.

Similarity calculation between entity sets

14. Select entities of your interest in two sets and click on *Calculate Entity Set Functional Similarity* to calculate the functional similarity between two entity sets (Fig. 8).

Weigh entity network

15. Input the protein interaction network file (`Data/interactions_extracted.txt`) into UFO with the four steps as followings:
    i. Open the *File* menu, select the item *Import*, and then select the item *Network from File*... .
    ii. Find and choose the `interactions_extracted.txt` file and click on the *Open* button.
    iii. Define source nodes and target nodes (Fig. 9A). Specifically, click on the *Column 1* (Fig. 9A, blue rectangle) and set as *Source Node*; do the same with the *Column 2* (Fig. 9A, orange rectangle) and set as *Target Node*.
    iv. Click on the *OK* button. The network is shown in Figure 9B.

16. Choose a suitable ontology and annotation data.

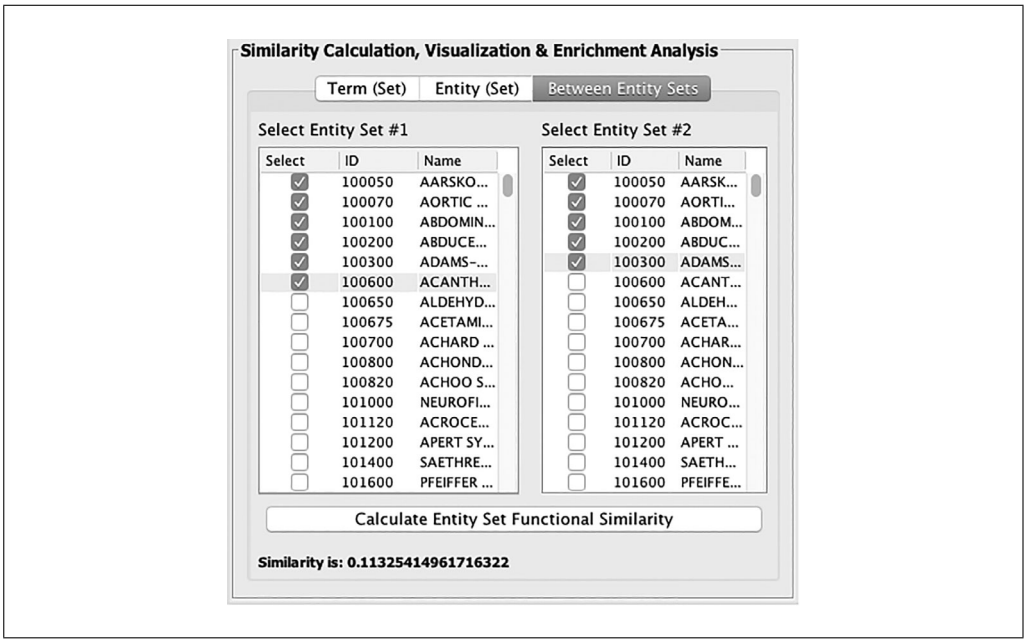


Figure 8 Implementation of functional similarity calculation between two entity sets. Six entities in the first set and five entities in the second set are randomly selected. Estimated functional similarity between two entity sets is approximately 0.11.

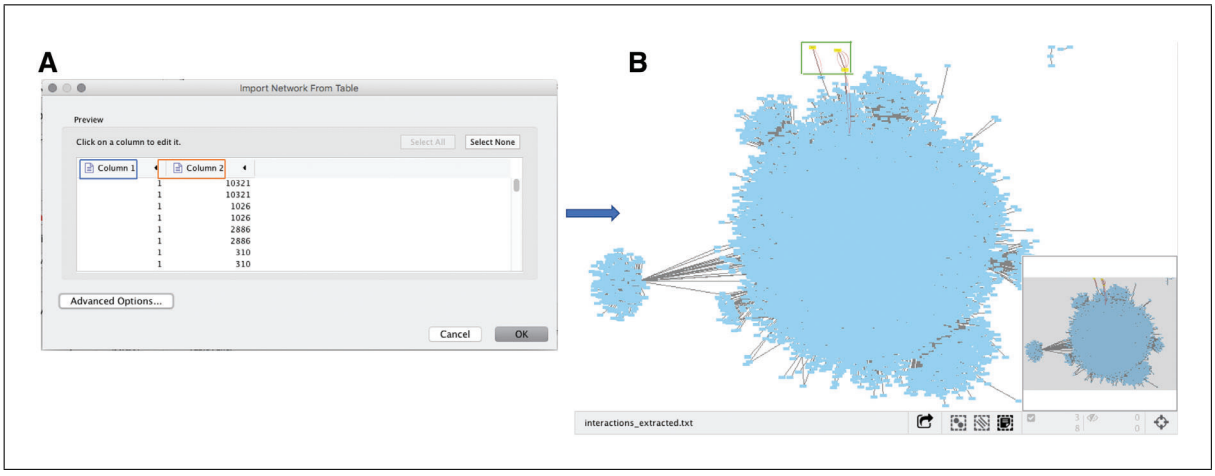


Figure 9 Creation of a publication-ready entity network. (**A**) Define source nodes and target nodes. Click on *Column 1* (blue rectangle) and set as *Source Node*; do the same with *Column 2* (orange rectangle) and set as *Target Node.* (**B**) Publication-ready network. Each entity is one turquoise node. Yellow nodes placed in a green rectangle are selected interactions.

*The example file is a gene ontology-protein network having Entrez Gene ID; therefore, the user must choose Gene Ontology Annotation data using these identifiers. We select ontology Gene Ontology (GO)–Biological Process and annotation data Gene Ontology (GO) To Gene, as in step 5, **i-v**).*

17. Specify a certain semantic similarity measure between terms for GO and a functional similarity measure between entities for genes.

    *In this protocol, we choose the Jiang-Conrath method assigned to the Node-based category (step 7, (a-b)) and the Average method assigned to the Pairwise category (step 10, (**i-ii**)) as an example.*

18. Select some interactions in the loaded network (Fig. 9B) by simultaneously pressing the left button on your mouse and clicking CTRL in Windows (or clicking

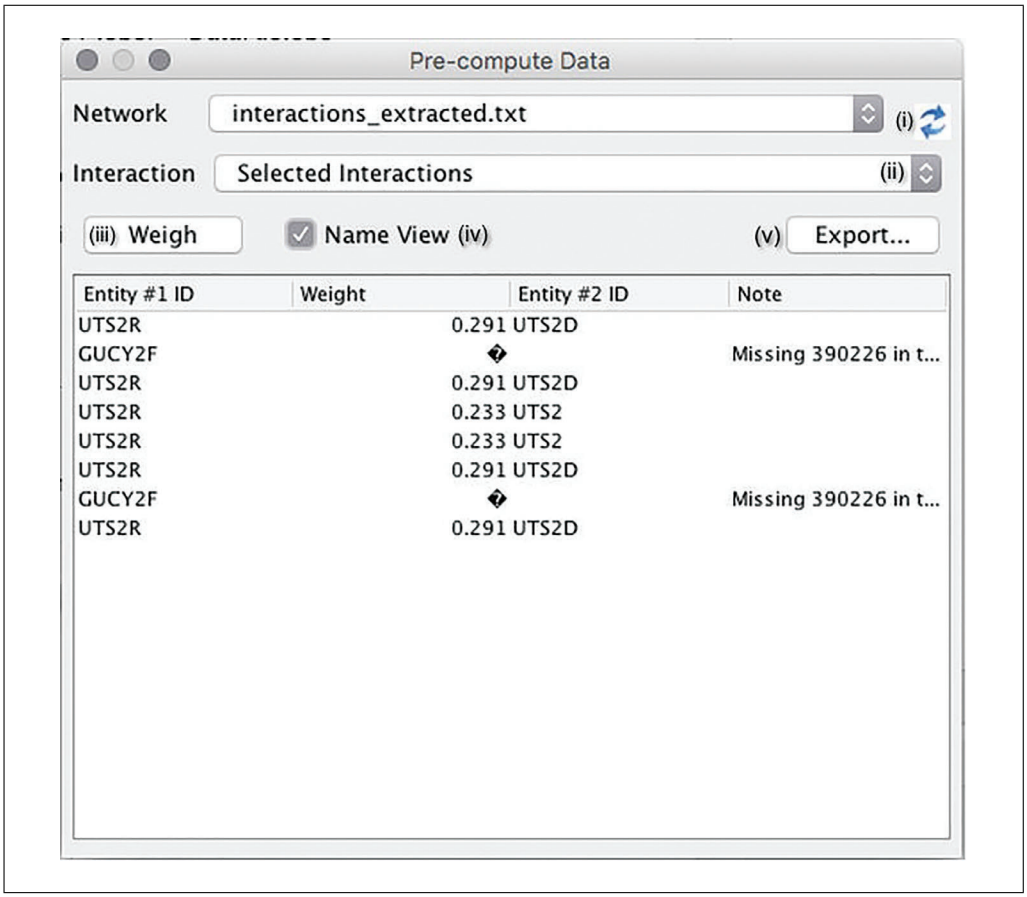


Figure 10 Weigh selected interactions of interest. Obtained weight of an interaction is the similarity between two interacting entities.

COMMAND in Mac), and then move the mouse pointer over several interactions of interest.

19. Weigh the entity network. Open the *Apps* menu, select the item *UFO*, then select the item *Utilities*, and select the item *Weigh Entity Network…* . Figure 10 shows the place we weigh the entity network.
    i. Upload the loaded network.
    ii. Choose *Selected interactions* or *All network interactions.* We will choose *Selected interactions* to weigh selected interactions (step 17 and Fig. 9B).
    iii. Click on *Weigh* to weigh selected interactions.
    iv. Check *Name View* box to turn Entity ID to Term Name, placed under the two columns *Entity #1 ID* and *Entity #2 ID* (Fig. 10). The opposite will occur if the box is left blank.
    v. Click on *Export…* to save your analysis results. Please name it (e.g., `Sim_Network`) and place it in the `Results` folder.

COMMENTARY

Background Information

Throughout the early implementations of UFO, we improved our tool to predict disease-associated genes (Le & Dang, 2016; Le & Kwon, 2013), protein complexes (Le, 2015), and lncRNAs (Le & Dao, 2018) by constructing gene similarity networks, protein complex similarity networks, and disease similarity networks using GO, HPO, and DO, respectively. Also, in another study (Le et al., 2016), an HPO-based human disease phenotype similarity assessment was performed using a wide variety of similarity measures. The results were then considered guidelines for researchers interested in this field to select the most appropriate semantic similarity method to assess

the phenotypic similarity between diseases. To this end, the tool has been additionally proven able to analyze the enrichment and visualize the ontology of HPO terms for a list of similar phenotypes (Le, 2020).

As we mentioned earlier, many tools have been introduced (Bindea et al., 2009; Deng et al., 2015; Du et al., 2009; Fröhlich et al., 2007; Köhler et al., 2009; Li et al., 2011; Martin et al., 2004; Yu et al., 2014) to calculate the similarity, analyze the enrichment, and visualize the ontology. Major differences between those tools include the platform (e.g., Cytoscape, R, Python, or the web) that they were developed for, the number of available analysis types (i.e., similarity calculation, ontology visualization, enrichment analysis), and the number of measures (see details in our original paper; Le, 2020) included in each of the analysis types (Table 2). Table 2 shows a comprehensive comparison of UFO and 33 alternative ontology tools based on the criteria above. In general, UFO runs on the Cytoscape platform widely used to build apps for various biomedical problems (Lotia, Montojo, Dong, Bader, & Pico, 2013), implements three of the most common ontology analysis types, and far exceeds the similarity measure numbers included in the other state-of-the-art tools. In addition, all the similarity measures for the calculation of between-term (semantic similarity) and between-entity (functional similarity) measures in UFO can be used for any type of biomedical ontology in the OBO Foundry (Smith et al., 2007).

Troubleshooting

Table 3 presents problems that may arise when running UFO, along with their possible causes and solutions.

Understanding Results

(A) Term-related functions

Two main example files are produced in the analysis:

`SS-HPOterm` *(given in step 7, v; also see* Fig. 4*):* The result represents a similarity matrix of terms whose rows and columns are terms; each cell is a calculated similarity value between a pair of terms. The bigger the similarity value, the stronger the similarity level between a pair of terms.

`Viz-HPOterm` *(given in step 8, viii; also see* Fig. 5*):* A publication-ready network, in which a green node is the root while color levels (white to bold red) indicate a corresponding increase in the IC of the terms. Users can learn more about the Information Content (IC) in our original paper (Le, 2020). For the relationship "is-a", please see the **Box 1**.

(B) Entity-related functions

`FS-HPOentity` *(given in step 10, v:* The result represents a similarity matrix of entities with an interpretation is the same as SS-HPOterm above.

`EA-HPOterm` *(given in step 11, iii; also see* Fig. 7A*):* The result file includes four columns: Term, Enriched Phenotype, *p*-value, and *Q*-value. The Term column lists all of the terms, and the Enriched Phenotype column lists the corresponding phenotypic series. The two *p*-value and *Q*-value columns indicate corresponding *p*-values and *Q*-values at each row to reject the null hypothesis and control the false discovery rate (see details in our original paper; Le, 2020).

`Viz-HPOentity` *(given in step 12, iii; also see* Fig. 7B*):* The result is a publication-ready functional similarity network between ten entities of interest (turquoise rectangles creating a left top circle), annotating relationships between entities and terms, with a directed acyclic graph of the annotating ontology terms (red spheres creating a right bottom circle).

Figure 8 *(given in step 14):* This is the result of functional similarity calculation between two entity sets. The range of a similarity value is between 0 and 1. The bigger the similarity value (closer to 1), the stronger the similarity level between two sets.

`Sim_Network` *(given in step 18, v; also see* Fig. 10*):* We weigh selected interactions in an entity network by the similarity between interacting entities. The range of a similarity value is between 0 and 1. The bigger the similarity value (closer to 1), the higher the functional similarity level between interacting entities.

Time Considerations

The time required for this protocol depends mainly on your hardware, the number of inputs selected (i.e., number of entities, number of terms) for the analysis process (steps 6 and 9), and the type of similarity calculation measure (steps 7 and 10). The above example analysis was performed on a Mac OS X machine with 8 GB of RAM and Java 8.

Table 2 Comparison of UFO and 33 Alternative Ontology Tools[a]

		Between-term similarity				Between-entity similarity			Visualization			
		Node-based					Groupwise		Term-term	Entity-term	Entity-entity	
Tool	Platform	MICA	CDA	Edge-based	Hybrid	Pairwise	Vector-based	Graph-based	DAG	Annotation	Functional similarity network	Enrichment
All biomedical ontologies												
UFO (Le, 2020)	Cytoscape	R,L,J,S	R,L,J,S	Yu, Wu	Wang	M,A,B,rc	C, K	T,N,U,G,L	Yes	Yes	Yes	Yes
SML-Toolkit (Harispe et al., 2014)	Java	R,L,J,S	R,L,J,S	Wu		M,A,B,rc		T,N,U,G,L				
Gene ontology												
GOSim (Fröhlich et al., 2007)	R	R,L,J,S	R,L,J,S			M,A,B,rc	C		Yes			Yes
GOSemSim (Yu et al., 2010)	R	R,L,J,S	R,L,J,S		Wang							
SemSim	R	R,L,J,S			Wang	M,A,rc						
GOvis/GOstats	R							U,L	Yes			Yes
csbl.go (Ovaska, Laakso, & Hautaniemi, 2008)	R	R,L,J,S	R,L,J,S				C, K	G				Yes

(*Continued*)

Table 2 Comparison of UFO and 33 Alternative Ontology Tools[a], *continued*

		Between-term similarity				Between-entity similarity			Visualization			
		Node-based					Groupwise		Term-term	Entity-term	Entity-entity	
Tool	Platform	MICA	CDA	Edge-based	Hybrid	Pairwise	Vector-based	Graph-based	DAG	Annotation	Functional similarity network	Enrichment
Golorize (Garcia et al., 2006)	Cytoscape									Yes		Yes
ClueGO (Bindea et al., 2009)	Cytoscape						K					Yes
BiNGO (Maere, Heymans, & Kuiper, 2005)	Cytoscape								Yes			Yes
SimTrek (Wang, Zheng, & Azuaje, 2010)	Cytoscape	R,L,J				M,A						
DaGO-Fun (Mazandu & Mulder, 2013)	Web	R,L,S	R,L,J		Wang	M,A,B		U,G				
FuSSiMeG	Web	R,L,J	L,J,S			M						Yes
GOToolBox (Martin et al., 2004)	Web								Yes			Yes

(*Continued*)

Table 2 Comparison of UFO and 33 Alternative Ontology Tools[a], *continued*

		Between-term similarity				Between-entity similarity			Visualization			
		Node-based					Groupwise		Term-term	Entity-term	Entity-entity	
Tool	Platform	MICA	CDA	Edge-based	Hybrid	Pairwise	Vector-based	Graph-based	DAG	Annotation	Functional similarity network	Enrichment
ProteInOn	Web	R,L,J	R,L,J			B		U,G				Yes
G-SESAME (Du, Li, Chen, Yu, & Wang, 2009)	Web	R,L,J			Wang				Yes			
FunSimMat (Schlicker & Albrecht, 2008)	Web	R,L,J,S				M,A,rc		T,N,U,G				
GossTo (Caniza et al., 2014)	Web	R,L,J,S	R,L,J,S			M		U,G				
KU-GOAL (Jeong & Chen, 2015)	Web	R,L,J,S						T,N				
AmiGO (Carbon et al., 2008)	Web								Yes	Yes		Yes
Gorilla (Eden et al., 2009)	Web								Yes			Yes
g:Profiler (Reimand et al., 2007)	Web									Yes		Yes

(*Continued*)

Table 2 Comparison of UFO and 33 Alternative Ontology Tools[a], *continued*

		Between-term similarity				Between-entity similarity			Visualization			
		Node-based					Groupwise		Term-term	Entity-term	Entity-entity Functional similarity network	
Tool	Platform	MICA	CDA	Edge-based	Hybrid	Pairwise	Vector-based	Graph-based	DAG	Annotation		Enrichment
DynGO (Liu, Hu, & Wu, 2005)	Standalone					A			Yes	Yes		
FastSemSim	Python	R,L,J,S			Wang	M,A,B	C	T,N,U,G				
A-DaGO-Fun (Mazandu & Mulder, 2013)	Python	R,L,J,S			Wang	M,A,B,rc	C	T,N,U,G				
Disease ontology												
DOSim (Li et al., 2011)	R	R,L,J,S	R,L,J,S		Wang	M,A,B		G	Yes			Yes
DOSE (Yu et al., 2014)	R	R,L,J,S			Wang	M						Yes
FunDO (Du et al., 2009)	Web	R,J		Wu	Wang	M						Yes

(*Continued*)

Table 2 Comparison of UFO and 33 Alternative Ontology Tools[a], *continued*

		Between-term similarity				Between-entity similarity			Visualization			
		Node-based					Groupwise		Term-term	Entity-term	Entity-entity	
Tool	Platform	MICA	CDA	Edge-based	Hybrid	Pairwise	Vector-based	Graph-based	DAG	Annotation	Functional similarity network	Enrichment
Phenotype ontology												
HPOSim (Deng et al., 2015)	R	R,L,J,S				M,A,B,rc						Yes
OWLSim (Washington et al., 2009)	Standalone	R				A,B,rc		G				
Phenomizer (Köhler et al., 2009)	Web	R										
PhenoDigm (Smedley et al., 2013)	Web	R				A,B						
PhenomeNET (Hoehndorf, Schofield, & Gkoutos, 2011)	Web							G				
OntoSIML	Web	R						G				

[a]When it comes to the measures (each cell at the third and fourth columns in Table), for Most Informative Common Ancestors (MICA), R, L, J, and S correspond to Resnik's method (Resnik, 1995), Lin's method (Lin, 1998), Jiang-Conrath's method (Jiang & Conrath, 1997), and Schlicker's method (Schlicker, Domingues, Rahnenfuhrer, & Lengauer, 2006). For Common Disjunctive Ancestors (CDA), R, L, J, and S correspond to ResnikGraSM (Couto, Silva, & Coutinho, 2005), LinGraSM (Couto et al., 2005), Jiang-ConrathGraSM (Couto et al., 2005), and RelGraSM (Couto et al., 2005). For Edge-based, Yu and Wu are Yu's method (Yu, Gao, Tu, & Guo, 2005) and Wu's method (Wu, Su, Mao, Olman, & Xu, 2005). For Hybrid, Wang is Wang's method (Wang et al., 2007). For Pairwise, M, A, B, and rc correspond to Maximum (Riensche, Baddeley, Sanfilippo, Posse, & Gopalan, 2007; Sevilla et al., 2005), Average (Lord, Stevens, Brass, & Goble, 2003), Best Match Average (Azuaje, Wang, & Bodenreider, 2005; Couto et al., 2005; del Pozo, Pazos, & Valencia, 2008), and row-column Maximum (Schlicker et al., 2006). For Vector-based, C and K are Cosine coefficient (Huang et al., 2007) and Kappa coefficient (Chabalier, Mosser, & Burgun, 2007). For Graph-based, T, N, U, G, and L correspond to Term Overlap (Lee, Hsu, Sajdak, Qin, & Pavlidis, 2004), Normalized Term Overlap (Mistry & Pavlidis, 2008), Union-Intersection (Gentleman, 2005), Graph Information Content (Pesquita, Faria, Bastos, Falcão, & Couto, 2007), and Longest Shared Path (Gentleman, 2005). Abbreviation: GraSM, Graph-based Similarity; DAG, Directed Acyclic Graph.

Table 3 Troubleshooting Guide

Basic Protocol step	Problem	Possible reason	Solution
2	User wonders if Java is installed on the computer or which version is installed.		Go to the Terminal (in Mac OS or Ubuntu) or Command line (in Windows) and type `java --version`, then press Enter
7 (**iii**), 10 (**iii**), 11(**iii**)	Error messages: "*Please select at least two terms (whose ICs are finite for Node-based methods)!*" or "*Please select at least two objects*".	User only selects one term or one entity as input data	Choose at least two terms or two entities
8 (**iii**)	User cannot find any term of interest in the network	The terms of interest used for the search are not included in the network	Check that the terms of interest appear in the network

Steps 1-4, software installation: 5 min

Step 5, load and prepare data: 5 s/Ontology data

Steps 6-7, Semantic similarity calculation: 3-18 s/10 terms/each measure

Step 8, Visualization of terms of interest and annotated entities: 10 s/10 terms

Steps 9-10, Functional similarity calculation: 3 s to 15 min/10 entities/each measure

Step 11, Entity set enrichment analysis: 3 s

Step 12, Visualization of terms and annotated entities: 10 s

Step 13, Similarity calculation between entity sets: 10 s/5 entities per set

Steps 14-18, Weigh entity network: 9 min.

Author Contributions

Q.-H.N. wrote the manuscript, which was edited by D.-H.L. Q.-H.N. ran UFO, and analyzed output data. D.-H.L. coded the UFO app. D.-H.L. supervised the work. Q.-H.N., D.-H.L. read and approved the final manuscript.

Author Contributions

Quang-Huy Nguyen: Writing-original draft, Writing-review & editing, Duc-Hau Le

Conflict of Interest

The authors have declared no conflict of interest.

Data Availability Statement

The protocol uses publicly available data that users can easily download from our Github repository (*https://github.com/hauldhut/ufo*).

Literature Cited

Azuaje, F., Wang, H., & Bodenreider, O. (2005). Ontology-driven similarity approaches to supporting gene functional assessment. Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies 2005, 9−10.

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., & Apweiler, R. (2009). The GOA database in 2009: An integrated Gene Ontology Annotation resource. *Nucleic Acids Research*, *37*(suppl 1), D396–D403. doi: 10.1093/nar/gkn803.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., … Galon, J. (2009). ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, *25*(8), 1091–1093. doi: 10.1093/bioinformatics/btp101.

Caniza, H., Romero, A. E., Heron, S., Yang, H., Devoto, A., Frasca, M., … Paccanaro, A. (2014). GOssTo: A stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics*, *30*(15), 2235–2236. doi: 10.1093/bioinformatics/btu144.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., … Web Presence Working Group. (2008). AmiGO: Online access to ontology and annotation data. *Bioinformatics*, *25*(2), 288–289. doi: 10.1093/bioinformatics/btn615.

Chabalier, J., Mosser, J., & Burgun, A. (2007). A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics*, *8*(1), 235.

Couto, F., Silva, M. r., & Coutinho, P. (2005). Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. CIKM '05: Proceedings of the 14th ACM

international conference on Information and knowledge management, Bremen, Germany.

del Pozo, A., Pazos, F., & Valencia, A. (2008). Defining functional distances over Gene Ontology. *BMC Bioinformatics*, *9*(1), 50.

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.

Diehl, A. D., Meehan, T. F., Bradford, Y. M., Brush, M. H., Dahdul, W. M., Dougall, D. S., … Mungall, C. J. (2016). The cell ontology 2016: Enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, *7*(1), 44. doi: 10.1186/s13326-016-0088-7.

Du, P., Feng, G., Flatow, J., Song, J., Holko, M., Kibbe, W. A., & Lin, S. M. (2009). From disease ontology to disease-ontology lite: Statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics*, *25*(12), i63–i68. doi: 10.1093/bioinformatics/btp193.

Du, Z., Li, L., Chen, C.-F., Yu, P. S., & Wang, J. Z. (2009). G-SESAME: Web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Research*, *37*(suppl_2), W345–W349. doi: 10.1093/nar/gkp463.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*(1), 48. doi: 10.1186/1471-2105-10-48.

Fröhlich, H., Speer, N., Poustka, A., & Beißbarth, T. (2007). GOSim—an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, *8*(1), 166. doi: 10.1186/1471-2105-8-166.

Garcia, O., Saveanu, C., Cline, M., Fromont-Racine, M., Jacquier, A., Schwikowski, B., & Aittokallio, T. (2006). GOlorize: A Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics*, *23*(3), 394–396. doi: 10.1093/bioinformatics/btl605.

Gene Ontology, C. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, *32*(suppl 1), D258–D261. doi: 10.1093/nar/gkh036.

Gentleman, R. (2005). Visualizing and distances using GO. Available at http://bioconductor.org/packages/release/bioc/vignettes/GOstats/inst/doc/GOvis.pdf.

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2014). The semantic measures library and toolkit: Fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, *30*(5), 740–742. doi: 10.1093/bioinformatics/btt581.

Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2011). PhenomeNET: A whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, *39*(18), e119. doi: 10.1093/nar/gkr538.

Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: A functional perspective. *Briefings in Bioinformatics*, *16*(6), 1069–1080. doi: 10.1093/bib/bbv011.

Huang, D., Sherman, B., Tan, Q., Collins, J., Alvord, W. G., Roayaei, J., … Lempicki, R. (2007). The DAVID gene functional classification tool: A novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, *8*(9), R183.

Jackson, R. C., Balhoff, J. P., Douglass, E., Harris, N. L., Mungall, C. J., & Overton, J. A. (2019). ROBOT: A tool for automating ontology workflows. *BMC Bioinformatics*, *20*(1), 407. doi: 10.1186/s12859-019-3002-3.

Jeong, J. C., & Chen, X. (2015). A new semantic functional similarity over gene ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *12*(2), 322–334. doi: 10.1109/tcbb.2014.2343963.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the International Conference on Research in Computational Linguistics (ROCLING), Taiwan, (1997).

Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., … Schriml, L. M. (2015). Disease Ontology 2015 update: An expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research*, *43*(D1), D1071–D1078. doi: 10.1093/nar/gku1011.

Kulmanov, M., & Hoehndorf, R. (2017). Evaluating the effect of annotation size on measures of semantic similarity. *Journal of Biomedical Semantics*, *8*(1), 7. doi: 10.1186/s13326-017-0119-z.

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., … Robinson, P. N. (2014). The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, *42*(D1), D966–D974. doi: 10.1093/nar/gkt1026.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., … Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, *85*(4), 457–464. doi: 10.1016/j.ajhg.2009.09.003.

Le, D.-H. (2015). A novel method for identifying disease associated protein complexes based on functional similarity protein complex networks. *Algorithms for Molecular Biology*, *10*(1), 14.

Le, D.-H. (2020). UFO: A tool for unifying biomedical ontology-based semantic similarity calculation, enrichment analysis and visualization. *PLoS One*, *15*(7), e0235670. doi: 10.1371/journal.pone.0235670.

Le, D.-H., & Dang, V.-T. (2016). Ontology-based disease similarity network for disease gene prediction. *Vietnam Journal of Computer Science*, *3*, 197–205. doi: 10.1007/s40595-016-0063-3.

Le, D.-H., & Dao, L. T. M. (2018). Annotating diseases using human phenotype ontology improves prediction of disease-associated long non-coding RNAs. *Journal of Molecular Biology*, *430*(15), 2219–2230. doi: 10.1016/j.jmb.2018.05.006.

Le, D.-H., & Kwon, Y.-K. (2013). Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization. *Computational Biology and Chemistry*, *44*(0), 1–8. doi: 10.1016/j.compbiolchem.2013.01.001.

Le, D.-H., Pham, B.-S., & Dao, A.-M. (2016). *Assessing human disease phenotype similarity based on ontology* RIVF 2016: The 12th IEEE-RIVF International Conference on Computing and Communication Technologies, Hanoi, Vietnam.

Lee, H., Hsu, A., Sajdak, J., Qin, J., & Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Research*, *14*(6), 1085–1094.

Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., … Li, X. (2011). DOSim: An R package for similarity between diseases based on Disease Ontology. *BMC Bioinformatics*, *12*(1), 266. doi: 10.1186/1471-2105-12-266.

Lin, D. (1998). *An Information-Theoretic Definition of Similarity*. ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning, July 1998. San Francisco: Morgan Kaufman Publishers, Inc.

Liu, H., Hu, Z.-Z., & Wu, C. H. (2005). DynGO: A tool for visualizing and mining of Gene Ontology and its associations. *BMC Bioinformatics*, *6*(1), 201. doi: 10.1186/1471-2105-6-201.

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Semantic similarity measures as tools for exploring the gene ontology. *Pacific Symposium on Biocomputing*, 601–612.

Lotia, S., Montojo, J., Dong, Y., Bader, G. D., & Pico, A. R. (2013). Cytoscape app store. *Bioinformatics*, *29*(10), 1350–1351. doi: 10.1093/bioinformatics/btt138.

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, *21*(16), 3448–3449. doi: 10.1093/bioinformatics/bti551.

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., & Jacq, B. (2004). GOToolBox: Functional analysis of gene datasets based on Gene Ontology. *Genome Biology*, *5*(12), R101. doi: 10.1186/gb-2004-5-12-r101.

Mazandu, G. K., & Mulder, N. J. (2013). DaGO-Fun: Tool for Gene Ontology-based functional analysis using term information content measures. *BMC Bioinformatics*, *14*, 284. doi: 10.1186/1471-2105-14-284.

Mazandu, G. K., & Mulder, N. J. (2014). Information content-based gene ontology functional similarity measures: Which one to use for a given biological data type? *PLoS One*, *9*(12), e113859. doi: 10.1371/journal.pone.0113859.

Mistry, M., & Pavlidis, P. (2008). Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, *9*(1), 327.

Osumi-Sutherland, D. (2017). Cell ontology in an age of data-driven cell classification. *BMC Bioinformatics*, *18*(17), 558. doi: 10.1186/s12859-017-1980-6.

Ovaska, K., Laakso, M., & Hautaniemi, S. (2008). Fast gene ontology based clustering for microarray experiments. *BioData Mining*, *1*(1), 11. doi: 10.1186/1756-0381-1-11.

Peng, K., Xu, W., Zheng, J., Huang, K., Wang, H., Tong, J., … Xia, T. (2013). The disease and gene annotations (DGA): An annotation resource for human disease. *Nucleic Acids Research*, *41*(D1), D553–D560. doi: 10.1093/nar/gks1244.

Pesquita, C., Faria, D., Bastos, H., Falcão, A., & Couto, F. (2007). Evaluating go-based semantic similarity measures. Proceedings pf 10th Annual Bio-Ontologies Meeting, Vienna, Austria.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, *5*(7), e1000443.

Reimand, J., Kull, M., Peterson, H., Hansen, J., & Vilo, J. (2007). g:PRofiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, *35*(suppl_2), W193–W200. doi: 10.1093/nar/gkm226.

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, Montreal, Quebec, Canada.

Riensche, R. M., Baddeley, B. L., Sanfilippo, A. P., Posse, C., & Gopalan, B. (2007). XOA: Web-enabled cross-ontological analytics. 2007 IEEE Congress on Services, Salt Lake City, Utah.

Schlicker, A., & Albrecht, M. (2008). FunSimMat: A comprehensive functional similarity database. *Nucleic Acids Research*, *36*(Database issue), D434–439. doi: 10.1093/nar/gkm806.

Schlicker, A., Domingues, F., Rahnenfuhrer, J., & Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, *7*(1), 302.

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martinez-Cruz, L. A., … Rubio, A. (2005). Correlation between gene expression and GO semantic similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *2*(4), 330–338.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A software environment for

integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. doi: 10.1101/gr.1239303.

Sharp, M. E. (2017). Toward a comprehensive drug ontology: Extraction of drug-indication relations from diverse information sources. *Journal of Biomedical Semantics*, *8*(1), 2–2. doi: 10.1186/s13326-016-0110-0.

Smedley, D., Oellrich, A., Köhler, S., Ruef, B., Westerfield, M., Robinson, P., … Mungall, C. (2013). PhenoDigm: Analyzing curated annotations to associate animal models with human diseases. *Database (Oxford)*, *2013*, bat025. doi: 10.1093/database/bat025.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., … Lewis, S. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251–1255. doi: 10.1038/nbt1346.

Wang, H., Zheng, H., & Azuaje, F. (2010). Ontology- and graph-based similarity assessment in biological networks. *Bioinformatics*, *26*(20), 2643–2644. doi: 10.1093/bioinformatics/btq477.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, *23*(10), 1274–1281. doi: 10.1093/bioinformatics/btm087.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., & Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biology*, *7*(11), e1000247. doi: 10.1371/journal.pbio.1000247.

Wu, H., Su, Z., Mao, F., Olman, V., & Xu, Y. (2005). Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Research*, *33*(9), 2822–2837. doi: 10.1093/nar/gki573.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., & Wang, S. (2010). GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, *26*(7), 976–978. doi: 10.1093/bioinformatics/btq064.

Yu, G., Wang, L.-G., Yan, G.-R., & He, Q.-Y. (2014). DOSE: An R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, *31*(4), 608–609. doi: 10.1093/bioinformatics/btu684.

Yu, H., Gao, L., Tu, K., & Guo, Z. (2005). Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene*, *352*(0), 75–81. doi: 10.1016/j.gene.2005.03.033.