

UFO: a tool for unifying biomedical ontology-based semantic similarity calculation, enrichment analysis and visualization

Duc-Hau Le^{1,2,*}

¹Department of Computational Biomedicine, Vingroup Big Data Institute, Hanoi, Vietnam.

²School of Computer Science and Engineering, Thuyloi University, Hanoi, Vietnam.

* To whom correspondence should be addressed. Tel: +84 912 324564; Email: hauldhut@gmail.com

Supplementary File 2

User Manual & Case Studies

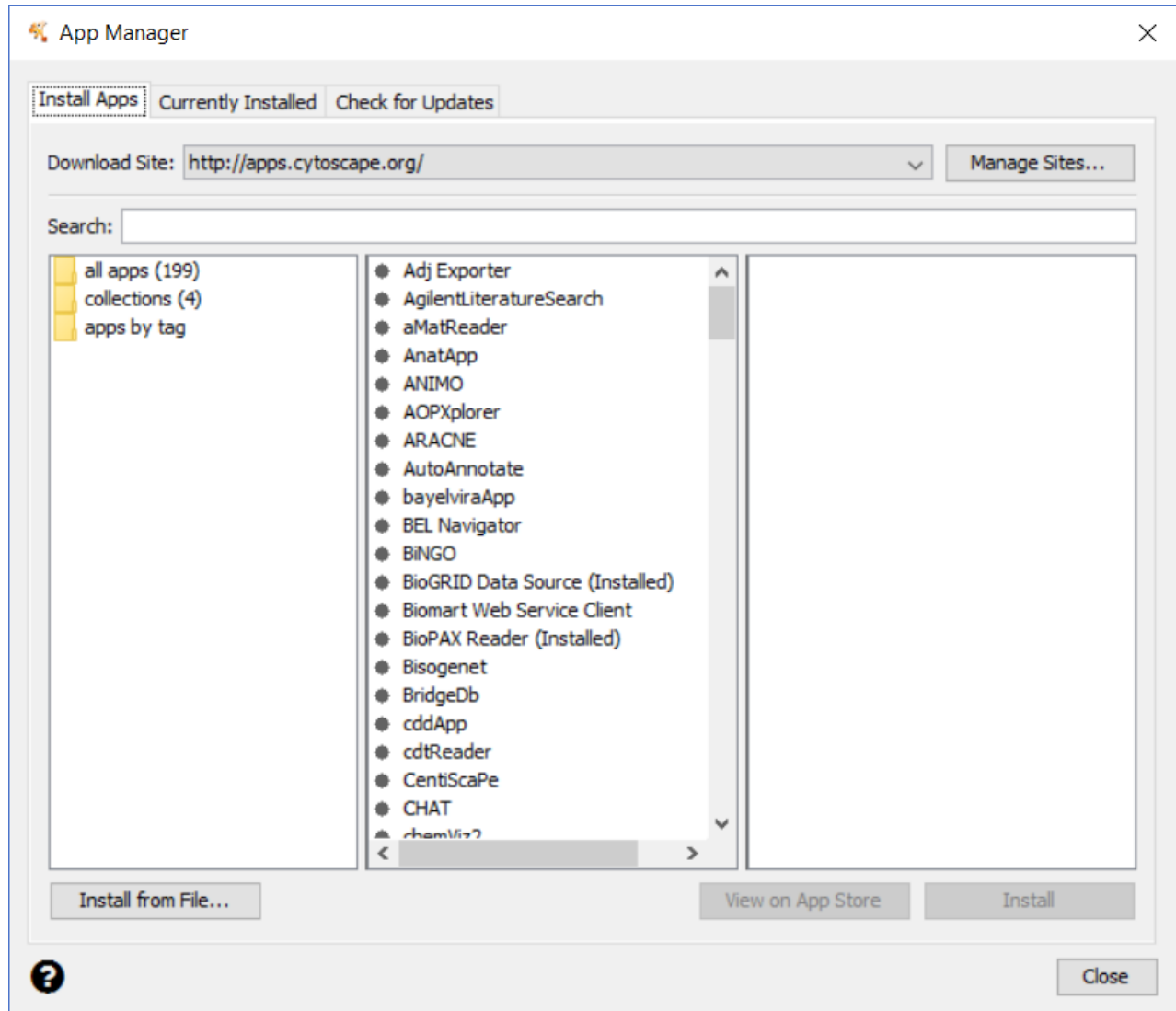
(UFO was tested to be worked well on both Windows, Mac and Ubuntu operating systems)

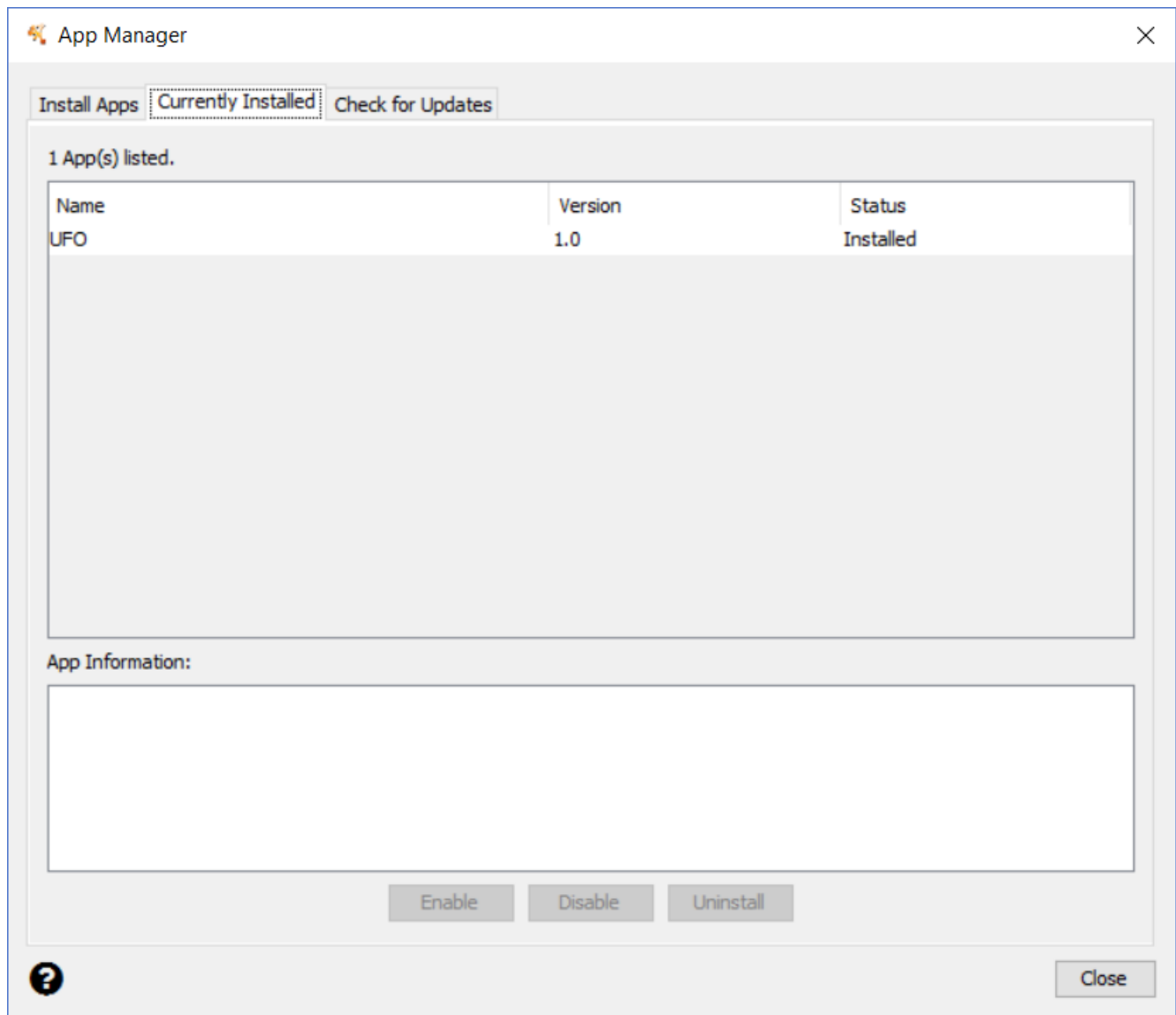
The following demonstration was performed on Windows 10

1	Setting up	2
2	How to use.....	4
2.1	Prepare data.....	4
2.2	Menu Items	4
2.3	Load and Prepare data.....	6
2.4	Main functions	10
2.4.1	Term-related Functions	10
a.	Calculate Semantic Similarity Matrix.....	10
b.	Visualize terms and annotated entities.....	12
2.4.2	Entity-related Functions	15
a.	Calculate Functional Similarity Matrix.....	16
c.	Analyze term enrichment.....	18
d.	Visualize terms and annotated entities.....	20
e.	Assess functional similarity between entity sets.....	24
2.5	Utilities.....	25
2.5.1	Extract annotation data	25
a.	Gene Ontology (GO) To Gene.....	25
b.	Human Phenotype Ontology (HPO) To Phenotype	26
c.	Disease Ontology (DO) To Gene	26
2.5.2	Weigh entity network.....	27
3	CASE STUDY	31
3.1	Assessing human disease phenotype similarity based on ontology	31
3.2	Construct gene and protein complex similarity network using GO for predicting disease-associated genes and protein complexes.....	32
3.3	Construct disease similarity network using HPO and DO for predicting disease-associated genes and lncRNAs	33
4	Reference.....	36

1 SETTING UP

- Install Cytoscape version 3.x (3.6.0 or later).
- Open **App Manager...** in **Apps** menu
- Click on button **Install from File...**
- Choose file **UFO.jar** to install app in Cytoscape
 - **UFO.jar** can be downloaded at <https://sourceforge.net/projects/ufo-cytoscape/files/>
- The app will be displayed in **Currently Installed** tab





2 HOW TO USE

2.1 Prepare data

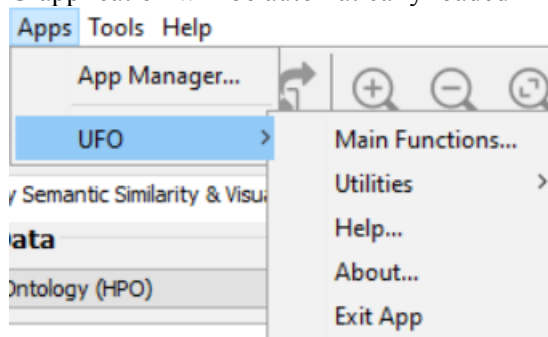
- Create a **Data** folder in Cytoscape folder (e.g., “C:\Program Files\Cytoscape_v3.5.1” in Windows, “Applications\Cytoscape_v3.5.1” in Mac)
- Download ontology and annotation data from public resources to **Data** folder
 - All the data can be downloaded at <https://github.com/hauldhut/ufo>
 - GO (go.obo), HPO (hp.obo) and DO (HumanDO.obo) in **Ontology** folder
 - Original annotation data in **Annotation** folder
 - Preprocessed annotation data in **Annotation-Preprocessed** folder

→ *Copy all ontology and preprocessed annotation data to Data folder*

- Otherwise, user can download the all data from original resources
 - For Ontology data.
 - Download Gene Ontology (go.obo) at <http://www.geneontology.org/>
 - Human Phenotype Ontology (hp.obo), Human Disease Ontology (doid.obo) and other ontologies at The Open Biological and Biomedical Ontologies (<http://www.obofoundry.org/>)
 - For Annotation data
 - User can download Gene Ontology to Gene (gene2go) at NCBI FTP site (<ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz>), Disease Ontology to Gene (ID-Mappings.rdf) at <http://dga.nubic.northwestern.edu/> and Human Phenotype Ontology to Phenotype (phenotype_annotation_hpoteam.tab) at <http://www.human-phenotype-ontology.org/>.
 - Construct other annotation data upon on your needs as following format
 - EntityID<tab>OntologyTermID<tab>EvidenceCode (optional)
 - Where Entity can be Gene, Protein, Phenotype, etc...
 - Ontology can be GO, DO, HPO, etc...
 - EvidenceCode is evidence code of that annotation. Leave blank when no evidence can be specified

2.2 Menu Items

- Run Cytoscape, and the **UFO** application will be automatically loaded in the **Apps** menu



- Menu Items
 - Main Functions...
 - Main functions of this app include calculation of between-term, between-entity similarity matrix; visualization of ontology and annotation data; analysis of term enrichment of a set of entities (e.g., genes); comparison of two entity sets; construction of functional similarity entity network.
 - Utilities

- Contains a set of utilities which facilitates users to pre-process some data
- Help...
 - Link to user manual
- About...
 - Brief information of this app
- Exit Application
 - To exit this app from Cytoscape panel

2.3 Load and Prepare data

- Click on **Main Functions...** menu item

Control Panel ▼ □ ✕

Network Style **Select** Ontology Semantic Similarity & Visualization

Ontology & Annotation Data

Ontology Human Phenotype Ontology (HPO) ▼

File (*.obo) Data\hp.obo

Annotation Human Phenotype Ontology (HPO) To Phenotype ▼

File (tab) Data\Annotation_OMIM2HPO_OMIM.txt

Evidence

<input checked="" type="checkbox"/>	All	<input checked="" type="checkbox"/>	User-d...	<input checked="" type="checkbox"/>	ICE
<input checked="" type="checkbox"/>	IEA	<input checked="" type="checkbox"/>	ISS	<input checked="" type="checkbox"/>	TAS
<input checked="" type="checkbox"/>	ND	<input checked="" type="checkbox"/>	IDA	<input checked="" type="checkbox"/>	IMP
<input checked="" type="checkbox"/>	IPI	<input checked="" type="checkbox"/>	NAS	<input checked="" type="checkbox"/>	IEP
<input checked="" type="checkbox"/>	IGI	<input checked="" type="checkbox"/>	IC	<input checked="" type="checkbox"/>	RCA

Load & Prepare Data

Similarity Calculation, Visualization & Enrichment Analysis

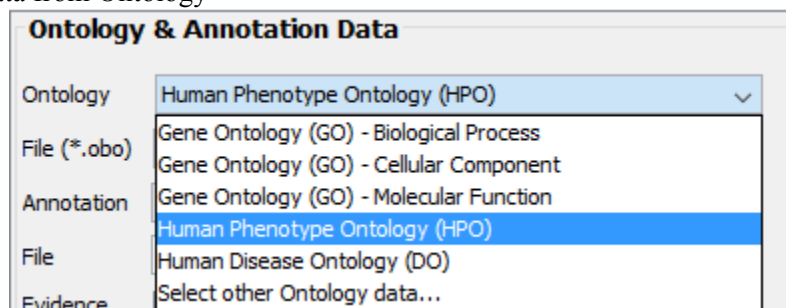
Term (Set) Entity (Set) Between Entity Sets

Input Semantic Similarity Visualization

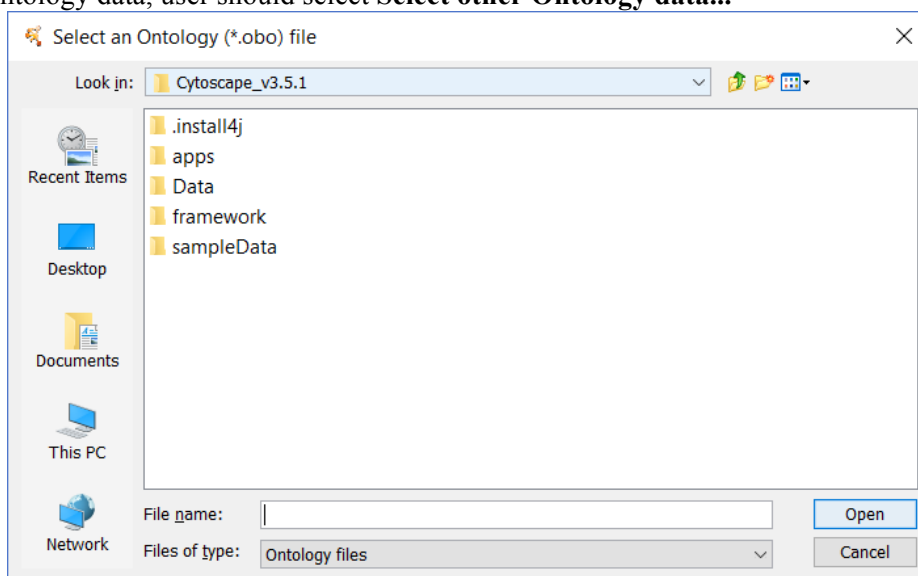
Select a Term/Term Set ☐ All ...

Sel	ID	Name	IC	Annotate...

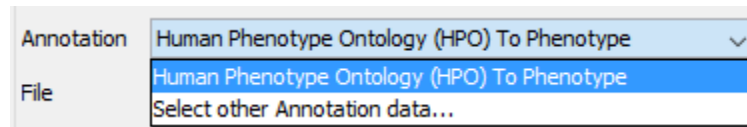
- Select Ontology data from Ontology



- We listed some popular Biomedical Ontologies such as Gene Ontology, Human Disease Ontology and Human Phenotype Ontology. Users should download them from public resources as shown in **2.1 Prepare data** section.
- For other ontology data, user should select **Select other Ontology data...**

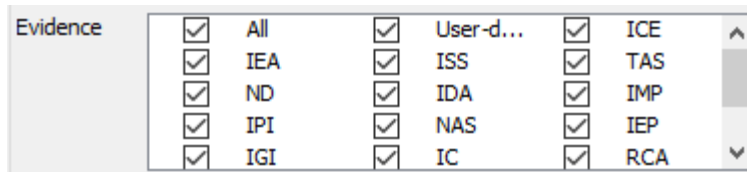


- Select relevant Annotation data from



- We also listed some Annotation data such as Gene Ontology to Genes, Disease Ontology to Genes and Human Phenotype Ontology to Phenotypes for ontologies such as Gene Ontology, Human Disease Ontology and Human Phenotype Ontology, respectively.
- For other ontology data, user should select **Select other Annotation data...**

- Select Evidence codes



- Select **All** for all annotations from selected Annotation data, otherwise select individual Evidence code. In addition, user can define their own Evidence codes accordingly by modifying **User-defined** cell or any other cells if they are not presented in the list.

- Load and Prepare Data for Analyses by clicking on

Load & Prepare Data

- This will load Ontology and Annotation data
- Calculate Information Content (IC) for each Ontology term
- Term ID, Term Name, IC and annotated Entities will be displayed in

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) Entity (Set) Between Entity Sets

Input Semantic Similarity Visualization

Select a Term/Term Set ☐ All **Total of 10934**

Select	ID	Name	Informati...	Annotat...
<input type="checkbox"/>	HP:0000001	All	-0	
<input type="checkbox"/>	HP:0000002	Abnormali...	0.893	
<input type="checkbox"/>	HP:0000003	Multicystic...	2.735 107480, 1...	
<input type="checkbox"/>	HP:0000005	Mode of in...	0.045	
<input type="checkbox"/>	HP:0000006	Autosomal...	0.387 100050, 1...	
<input type="checkbox"/>	HP:0000007	Autosomal...	0.377 100100, 1...	
<input type="checkbox"/>	HP:0000008	Abnormali...	1.513 101200, 1...	
<input type="checkbox"/>	HP:0000009	Functional...	1.694 105210, 1...	
<input type="checkbox"/>	HP:0000010	Recurrent...	2.492 109820, 1...	
<input type="checkbox"/>	HP:0000011	Neurogeni...	3.036 164200, 1...	
<input type="checkbox"/>	HP:0000012	Urinary ur...	2.352 146500, 1...	
<input type="checkbox"/>	HP:0000013	Hypoplasia...	2.638 119500, 1...	
<input type="checkbox"/>	HP:0000014	Abnormali...	1.632	
<input type="checkbox"/>	HP:0000015	Bladder di...	2.814 109820, 1...	

- Entity ID, Entity Name and annotating Terms will be displayed in

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) Entity (Set) Between Entity Sets

Input Functional Similarity Enrichment Visualization

Select an Entity/Entity Set ☐ All **Total of 6521**

Select	ID	Name	Annotating ...
<input type="checkbox"/>	100050	AARSKOG SY...	HP:0000049, ...
<input type="checkbox"/>	100070	AORTIC ANE...	HP:0004953
<input type="checkbox"/>	100100	ABDOMINAL ...	HP:0000007, ...
<input type="checkbox"/>	100200	ABDUCENS P...	HP:0000006, ...
<input type="checkbox"/>	100300	ADAMS-OLIV...	HP:0000006, ...
<input type="checkbox"/>	100600	ACANTHOSIS...	HP:0000006, ...
<input type="checkbox"/>	100640	ALDEHYDE DE...	HP:0000006, ...
<input type="checkbox"/>	100650	ALDEHYDE DE...	HP:0001033, ...
<input type="checkbox"/>	100675	ACETAMINOP...	HP:0000006, ...
<input type="checkbox"/>	100700	ACHARD SYN...	HP:0000006, ...
<input type="checkbox"/>	100800	ACHONDROP...	HP:0000006, ...
<input type="checkbox"/>	100820	ACHOO SYND...	HP:0000006, ...
<input type="checkbox"/>	101000	NEUROFIBRO...	HP:0000006, ...
<input type="checkbox"/>	101120	ACROCEPHAL	HP:0000006

These two lists of Terms and Entities will be used to select terms and entities of interest for further analyses

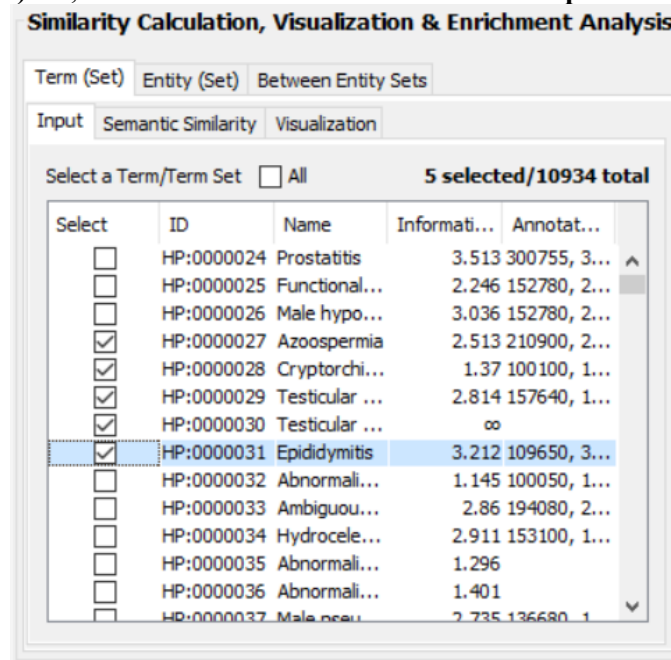
2.4 Main functions

2.4.1 Term-related Functions

This includes *Semantic Similarity Calculation* and *Visualization*

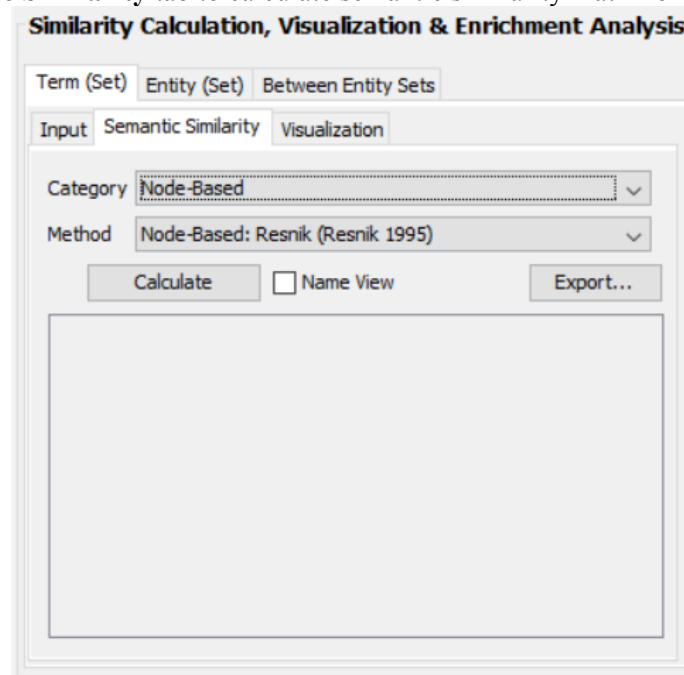
a. *Calculate Semantic Similarity Matrix*

Step 1: Select **Term (Set)** tab, then select a set of terms of interest in **Input** tab

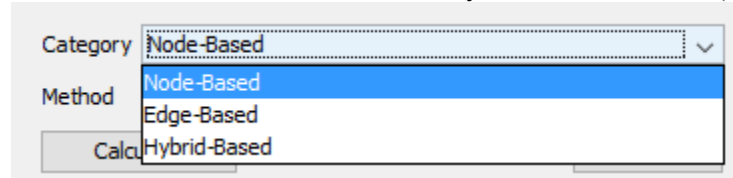


- Note that
 - For *Semantic Similarity Calculation*: Select at least two terms.

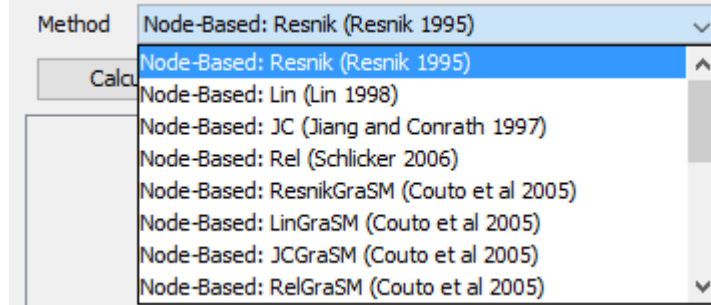
Step 2: Select **Semantic Similarity** tab to calculate semantic similarity matrix of the selected terms



- Select **Category** of methods to calculate semantic similarity between two terms (Pesquita, et al., 2009)



- Select **Method** in the selected **Category**



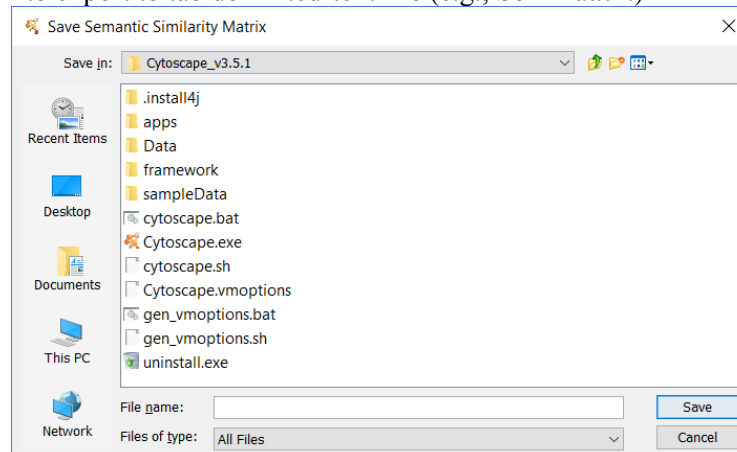
- Click **Calculate** to calculate semantic similarity matrix
➔ Result (e.g., for *Node-based: Lin (Lin 1998)* method)

	HP:0000027	HP:0000028	HP:0000029	HP:0000031
HP:0000027	2.3993408...	0.9289529...	0.9289529...	0.9289529...
HP:0000028	0.9289529...	1.2958002...	1.2958002...	1.1339786...
HP:0000029	0.9289529...	1.2958002...	1.2958002...	1.1339786...
HP:0000031	0.9289529...	1.1339786...	1.1339786...	2.9692161...

- Check ☐ **Name View** to turn to Term Name view of the semantic similarity matrix

	Azoosper...	Cryptorc...	Testicular...	Epididymitis
Azoospermia	2.3993408...	0.9289529...	0.9289529...	0.9289529...
Cryptorchid...	0.9289529...	1.2958002...	1.2958002...	1.1339786...
Testicular ...	0.9289529...	1.2958002...	1.2958002...	1.1339786...
Epididymitis	0.9289529...	1.1339786...	1.1339786...	2.9692161...

- Click **Export...** to export to tab delimited-text file (e.g., SemMat.txt)



Detail of SemMat.txt file

ID → HP:0000027 → HP:0000028 → HP:0000029 → HP:0000031 **CRLE**
 HP:0000027 → 2.3993408522366417 → 0.9289529801759476 → 0.9289529801759476 → 0.9289529801759476 **CRLE**
 HP:0000028 → 0.9289529801759476 → 1.2958002603295722 → 1.2958002603295722 → 1.1339786867928963 **CRLE**
 HP:0000029 → 0.9289529801759476 → 1.2958002603295722 → 1.2958002603295722 → 1.1339786867928963 **CRLE**
 HP:0000031 → 0.9289529801759476 → 1.1339786867928963 → 1.1339786867928963 → 2.9692161601932026 **CRLE**

b. Visualize terms and annotated entities

Step 1: Select **Term (Set)** tab, then select a set of Terms of interest in **Input** tab

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) Entity (Set) Between Entity Sets

Input Semantic Similarity Visualization

Select a Term/Term Set ☐ All 5 selected/10934 total

Select	ID	Name	Inform...	Ann...
<input type="checkbox"/>	HP:0000024	Prostatitis	3.513 3007...	
<input type="checkbox"/>	HP:0000025	Functional a...	2.246 1527...	
<input type="checkbox"/>	HP:0000026	Male hypog...	3.036 1527...	
<input checked="" type="checkbox"/>	HP:0000027	Azoospermia	2.513 2109...	
<input checked="" type="checkbox"/>	HP:0000028	Cryptorchidi...	1.37 1001...	
<input checked="" type="checkbox"/>	HP:0000029	Testicular a...	2.814 1576...	
<input checked="" type="checkbox"/>	HP:0000030	Testicular g...	∞	
<input checked="" type="checkbox"/>	HP:0000031	Epididymitis	3.212 1096...	
<input type="checkbox"/>	HP:0000032	Abnormality...	1.145 1000...	
<input type="checkbox"/>	HP:0000033	Ambiguous ...	2.86 1940...	
<input type="checkbox"/>	HP:0000034	Hydrocele t...	2.911 1531...	
<input type="checkbox"/>	HP:0000035	Abnormality...	1.296	
<input type="checkbox"/>	HP:0000036	Abnormality...	1.401	
<input type="checkbox"/>	HP:0000037	Male pseud...	2.735 1366...	
<input type="checkbox"/>	HP:0000039	Epispadias	3.337 2255...	
<input type="checkbox"/>	HP:0000040	Delayed as...	2.011 1255...	

Step 2: Select **Visualization** tab to visualize terms and annotated entities

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) Entity (Set) Between Entity Sets

Input Semantic Similarity Visualization

☒ Terms and Their Ancestors (Sub-DAG)

☐ Highlight Common Ancestors

☐ Highlight Most Informative Common Ancestor (MICA)

☐ Highlight Common Disjunctive Ancestor (DCA)

☐ Highlight Longest Path of Common Ancestor (LCA)

☐ Terms and Their Descendants (Sub-Tree)

☐ Highlight Common Descendants

☐ Annotated Entities

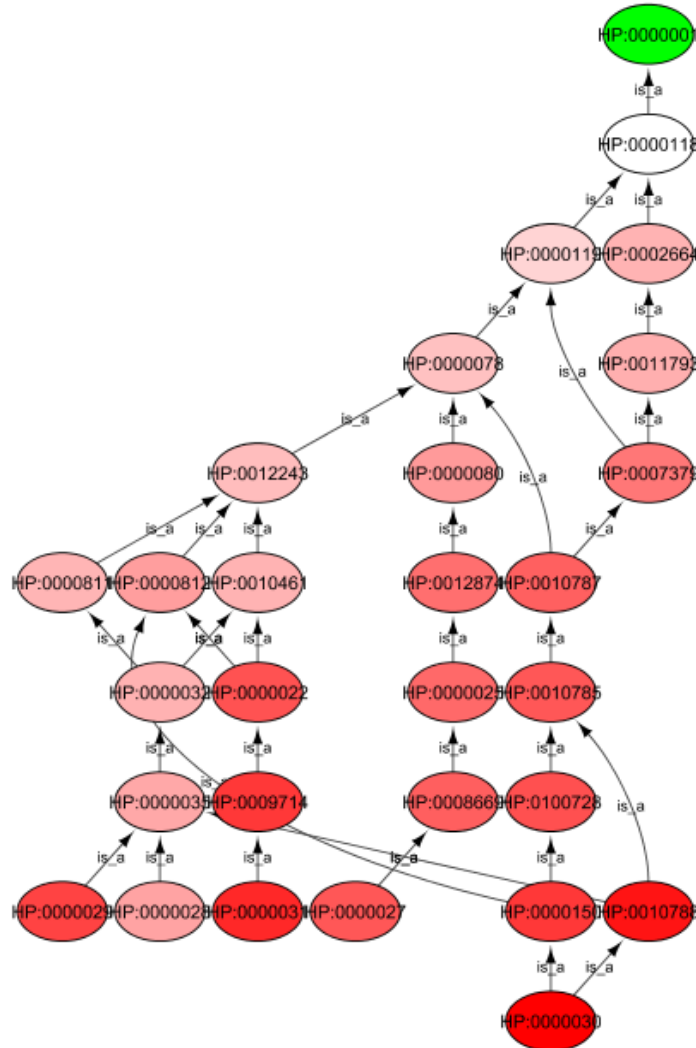
Visualize

- Select options then click

Visualize

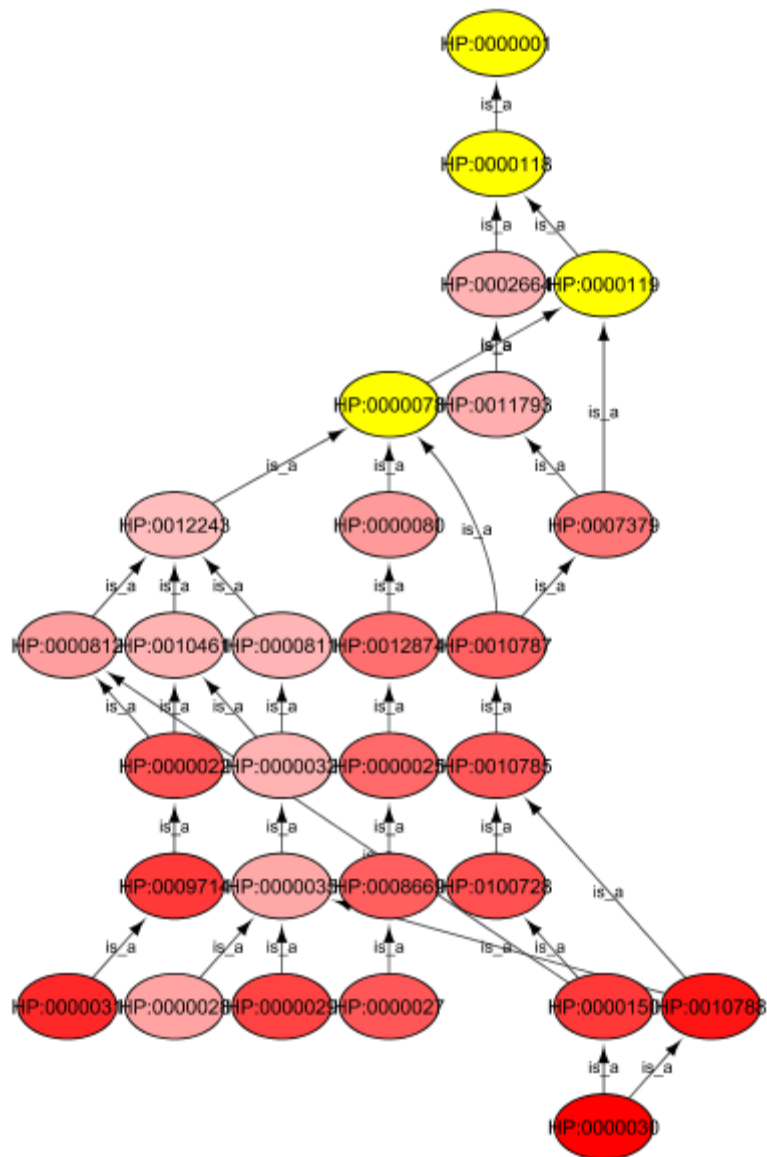
➔ Result

- With **Terms and Their Ancestors**



Five terms HP:0000027, HP:0000028, HP:0000029, HP:0000030 and HP:0000031 and their ancestors

- Check on ☐ Highlight Common Ancestors to highlight Common Ancestor terms of the selected terms.



Common ancestors of five terms HP:0000027, HP:0000028, HP:0000029, HP:0000030 and HP:0000031

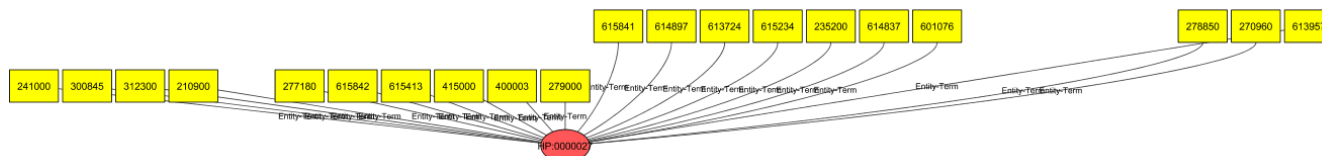
- With **Terms and Their Descendants**



Five terms HP:0000027, HP:0000028, HP:0000029, HP:0000030 and HP:0000031 and their descendants

- Check on ☐ Highlight Common Descendants to highlight Common Descendant terms of the selected terms.

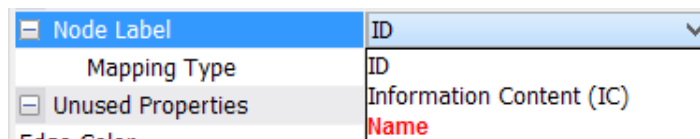
- With **Annotated Entities** (e.g., with Term HP: 0000027). User can view all annotated entities of the selected terms from Annotation data.



- Note: We provided both ID and Name for each entity/term as their attributes in **Node Table** tab of Cytoscape

ID /	Name
HP:0000027	Azoospermia
615842	SPERMATOGENIC FAILURE 14; SPGF14
615841	SPERMATOGENIC FAILURE 13; SPGF13
615413	SPERMATOGENIC FAILURE 12; SPGF12
615234	ANEMIA, HYPOCHROMIC MICROCYTIC, WITH IRON OVERLOAD 2; AHMIO2
614897	HYPOGONADOTROPIC HYPOGONADISM 16 WITH OR WITHOUT ANOSMIA; HH16
614837	HYPOGONADOTROPIC HYPOGONADISM 8 WITH OR WITHOUT ANOSMIA; HH8
613957	SPERMATOGENIC FAILURE 8; SPGF8
613724	LEUKOENCEPHALOPATHY WITH DYSTONIA AND MOTOR NEUROPATHY
601076	MULLERIAN DUCT APLASIA, UNILATERAL RENAL AGENESIS, AND CERVICOTHORACIC SOMITE ANOMALIES; MURCS
415000	SPERMATOGENIC FAILURE, Y-LINKED, 2; SPGFY2
400003	DELETED IN AZOOSPERMIA; DAZ
312300	ANDROGEN INSENSITIVITY, PARTIAL; PAIS
300845	MOYAMOYA DISEASE 4 WITH SHORT STATURE, HYPERGONADOTROPIC HYPOGONADISM, AND FACIAL DYSMORPHISM; MYMY4
279000	YOUNG SYNDROME
278850	46,XX SEX REVERSAL 2; SRXX2
277180	VAS DEFERENS, CONGENITAL BILATERAL APLASIA OF; CBAVD
270960	SPERMATOGENIC FAILURE 4; SPGF4
241000	HYPOGONADISM WITH LOW-GRADE MENTAL DEFICIENCY AND MICROCEPHALY
235200	HEMOCHROMATOSIS; HFE
210900	BLOOM SYNDROME; BLM

- Therefore, user can view entity/term by their Name by **VizMapper** function of Cytoscape



2.4.2 Entity-related Functions

For convenience, Entity-oriented Functions are organized on two tabs:

- **Entity (Set)**
 - o This tab includes Functional Similarity Matrix Calculation among selected entities, Enrichment Analysis of selected entities and Visualization.
- **Between Entity Sets**
 - o This tab is for calculating functional similarity between entity sets

a. *Calculate Functional Similarity Matrix*

Step 1: Select **Entity (Set)** tab, then select a set of entities of interest in **Input** tab.

Select	ID	Name	Annotating ...
<input type="checkbox"/>	100050	AARSKOG SY...	HP:0000049, ...
<input type="checkbox"/>	100070	AORTIC ANE...	HP:0004953
<input checked="" type="checkbox"/>	100100	ABDOMINAL ...	HP:0000007, ...
<input checked="" type="checkbox"/>	100200	ABDUCENS P...	HP:0000006, ...
<input checked="" type="checkbox"/>	100300	ADAMS-OLIV...	HP:0000006, ...
<input checked="" type="checkbox"/>	100600	ACANTHOSIS...	HP:0000006, ...
<input type="checkbox"/>	100640	ALDEHYDE DE...	HP:0000006, ...
<input type="checkbox"/>	100650	ALDEHYDE DE...	HP:0001033, ...
<input type="checkbox"/>	100675	ACETAMINOP...	HP:0000006, ...
<input type="checkbox"/>	100700	ACHARD SYN...	HP:0000006, ...
<input type="checkbox"/>	100800	ACHONDROP...	HP:0000006, ...
<input type="checkbox"/>	100820	ACHOO SYND...	HP:0000006, ...
<input type="checkbox"/>	101000	NEUROFIBRO...	HP:0000006, ...
<input type="checkbox"/>	101120	ACROCEPHAL	HP:0000006

- Note that
 - o For *Functional Similarity Calculation* and *Enrichment*: Select at least two entities.

Step 2: Calculate Functional Similarity Matrix

Category: Pairwise

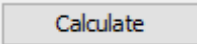
Method: Pairwise-Based: Avg (Average)

Calculate ☐ Name View Export...

- Select **Category** of methods to calculate functional similarity between two terms (Pesquita, et al., 2009)

- Select **Method** in the selected **Category**

- Note that
 - For **Pair-wise** methods, a pair of terms annotating to each entity must be calculated their semantic similarity based on the selected **Method** which is selected in **Semantic Similarity** tab of **Term (Set)** tab. For example, the selected semantic similarity method is *Node-based: Lin (Lin 1998)*
 - For **Group-wise** methods, users do not need to specify the semantic similarity method for term.

- Click  to calculate functional similarity matrix

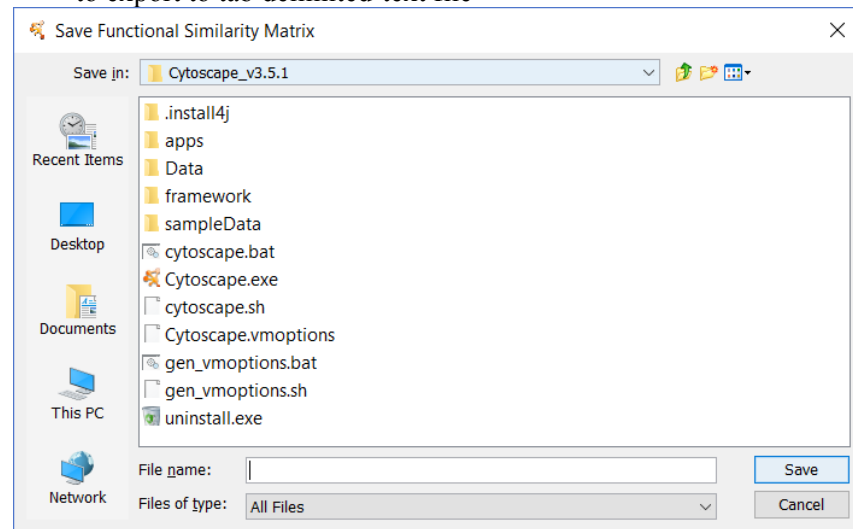
➔ Result

	100100	100200	100300	100600
100100	1.0	0.0044147...	0.0470141...	0.0211084...
100200	0.0044147...	1.0	0.0652407...	0.0093561...
100300	0.0470141...	0.0652407...	1.0	0.0474116...
100600	0.0211084...	0.0093561...	0.0474116...	1.0

Check  to turn to Entity Name view of the functional similarity matrix

	ABDOMIN...	ABDUCE...	ADAMS-O...	ACANTH...
ABDOMIN...	1.0	0.0044147...	0.0470141...	0.0211084...
ABDUCENS...	0.0044147...	1.0	0.0652407...	0.0093561...
ADAMS-OL...	0.0470141...	0.0652407...	1.0	0.0474116...
ACANTHO...	0.0211084...	0.0093561...	0.0474116...	1.0

- Click **Export...** to export to tab delimited-text file

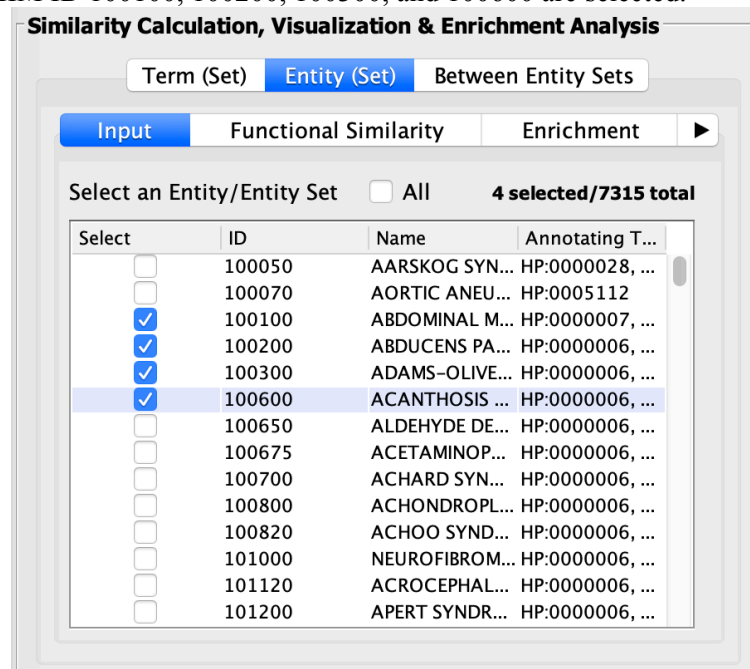


Detail of FunSimMat.txt file

```
ID→ABDOMINAL MUSCLES, ABSENCE OF, WITH URINARY TRACT ABNORMALITY AND CRYPTORCHIDISM→ABDUCENS PALSY→ADAMS-OLIVER SYNDROME 1; AOS1→ACANTHOSIS NIGRICANS
ABDOMINAL MUSCLES, ABSENCE OF, WITH URINARY TRACT ABNORMALITY AND CRYPTORCHIDISM→1.0→0.004414741161548392→1.0→0.0652407231735359→0.009356199289017913
ABDUCENS PALSY→0.004414741161548392→1.0→0.0652407231735359→0.009356199289017913
ADAMS-OLIVER SYNDROME 1; AOS1→0.04701416855323756→0.0652407231735359→1.0→0.047411637177936644
ACANTHOSIS NIGRICANS→0.021108486965106463→0.009356199289017913→0.047411637177936644→1.0
```

c. Analyze term enrichment

Step 1: Select **Entity (Set)** tab, then select a set of entities of interest in **Input** tab. For example, four phenotypes with OMIM ID 100100, 100200, 100300, and 100600 are selected.



- Note that
 - For *Functional Similarity Calculation* and *Term Enrichment*: Select at least two entities.

Step 2: Select **Enrichment** tab, then choose methods of Statistical Test and Adjusted P-value

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) **Entity (Set)** Between Entity Sets

Input Functional Similarity **Enrichment** ▶

Statistical Test

☒ Binomial

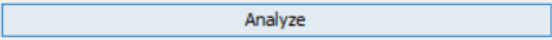
☐ Fisher's Exact

Adjusted P-value

☒ Bonferroni

☐ Benjamini and Hochberg

Analyze

- Note that
 - Enrichment Analysis is performed based on semantic similarity and functional similarity measures, which are selected in **Semantic Similarity** tab in **Term (Set)** tab, and **Functional Similarity** tab in **Entity (Set)** tab.
- After that, click  to do Enrichment Analysis. For example, with semantic similarity *Node-based: Resnik (Resnik 1995)* and functional similarity *Pairwise-Based: Avg (Average)*

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) **Entity (Set)** Between Entity Sets

Input Functional Similarity **Enrichment** ▶

Statistical Test

☒ Binomial

☐ Fisher's Exact

Adjusted P-value

☒ Bonferroni

☐ Benjamini and Hochberg

Analyze

Term ID	Name	P-value	Adjusted ...
HP:0005199	Aplasia of the ...	5.4659736971...	0.0311560500...
HP:0007589	Aplasia cutis c...	5.4659736971...	0.0311560500...
HP:0007590	Aplasia cutis c...	5.4659736971...	0.0311560500...
HP:0006970	Periventricular...	0.0016384472...	0.0933914950...
HP:0010957	Congenital po...	0.0016384472...	0.0933914950...
HP:0030011	Imperforate hy...	0.0016384472...	0.0933914950...
HP:0004392	Prune belly	0.0021837001...	0.1244709110...
HP:0001362	Skull defect	0.0022728628...	0.1865531820...

- Terms having Adjusted P-value ≤ 0.05 (HP: 0005199, HP: 0007589 and HP:0007590) could be of interest and represent functions of the selected set (four phenotypes with OMIM ID 100100, 100200, 100300, and 100600)

d. Visualize terms and annotated entities

Step 1: Select **Entity (Set)** tab, then select a set of entities of interest in **Input** tab.

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) Entity (Set) Between Entity Sets

Input Functional Similarity Enrichment Visualization

Select an Entity/Entity Set ☐ All **4 selected/6521 total**

Select	ID	Name	Annotating ...
<input type="checkbox"/>	100050	AARSKOG SY...	HP:0000049, ...
<input type="checkbox"/>	100070	AORTIC ANE...	HP:0004953
<input checked="" type="checkbox"/>	100100	ABDOMINAL ...	HP:0000007, ...
<input checked="" type="checkbox"/>	100200	ABDUCENS P...	HP:0000006, ...
<input checked="" type="checkbox"/>	100300	ADAMS-OLIV...	HP:0000006, ...
<input checked="" type="checkbox"/>	100600	ACANTHOSIS...	HP:0000006, ...
<input type="checkbox"/>	100640	ALDEHYDE DE...	HP:0000006, ...
<input type="checkbox"/>	100650	ALDEHYDE DE...	HP:0001033, ...
<input type="checkbox"/>	100675	ACETAMINOP...	HP:0000006, ...
<input type="checkbox"/>	100700	ACHARD SYN...	HP:0000006, ...
<input type="checkbox"/>	100800	ACHONDROP...	HP:0000006, ...
<input type="checkbox"/>	100820	ACHOO SYND...	HP:0000006, ...
<input type="checkbox"/>	101000	NEUROFIBRO...	HP:0000006, ...
<input type="checkbox"/>	101120	ACROCEPHAL	HP:0000006

Step 2: Visualize terms and annotated entities

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) Entity (Set) Between Entity Sets

Input Functional Similarity Enrichment Visualization

☒ Annotating Terms

☐ Annotating Terms and their Ancestors

☐ Annotating Terms and their Descendants

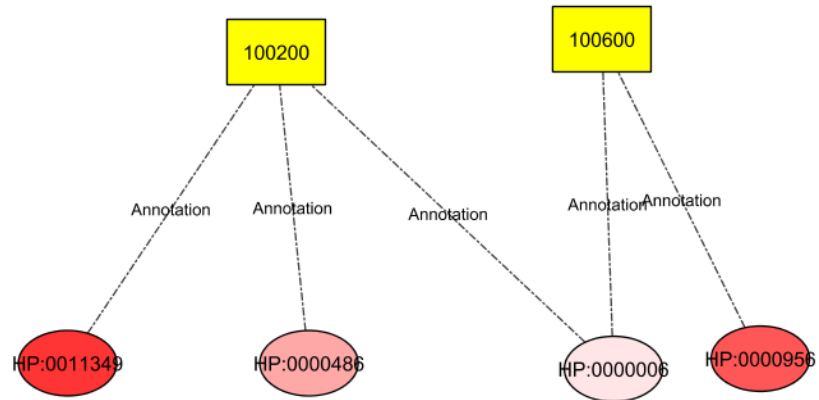
☐ Functional Similarity Interactions

Min Max

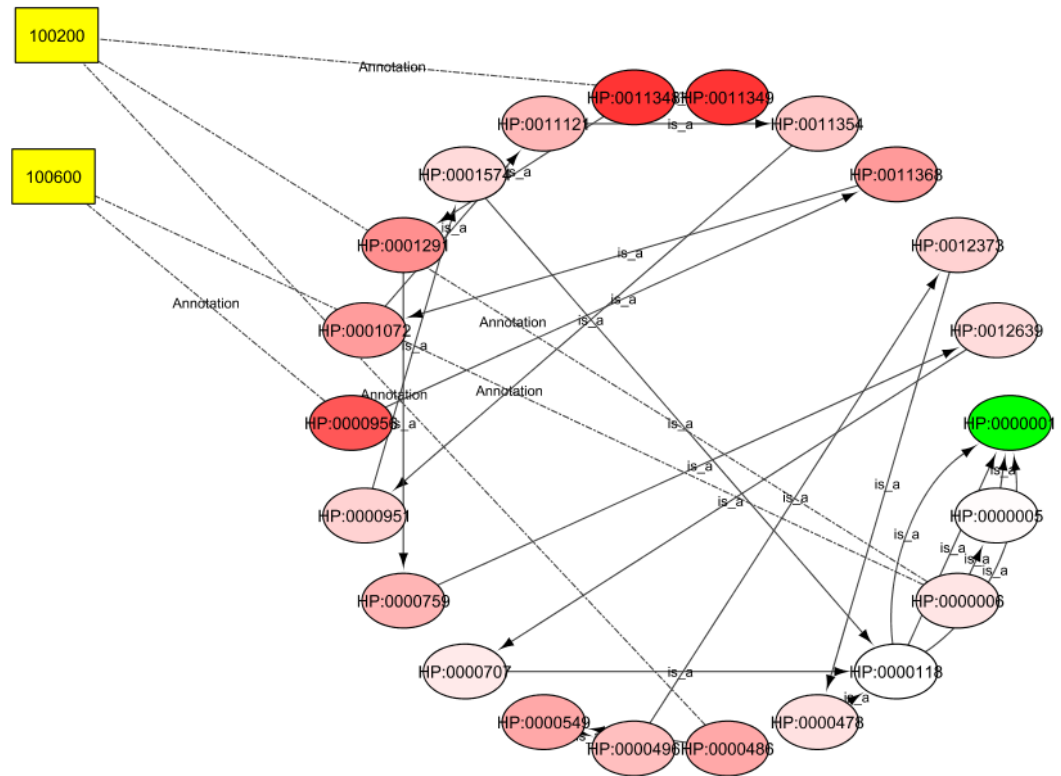
- Select options then click

➔ Result

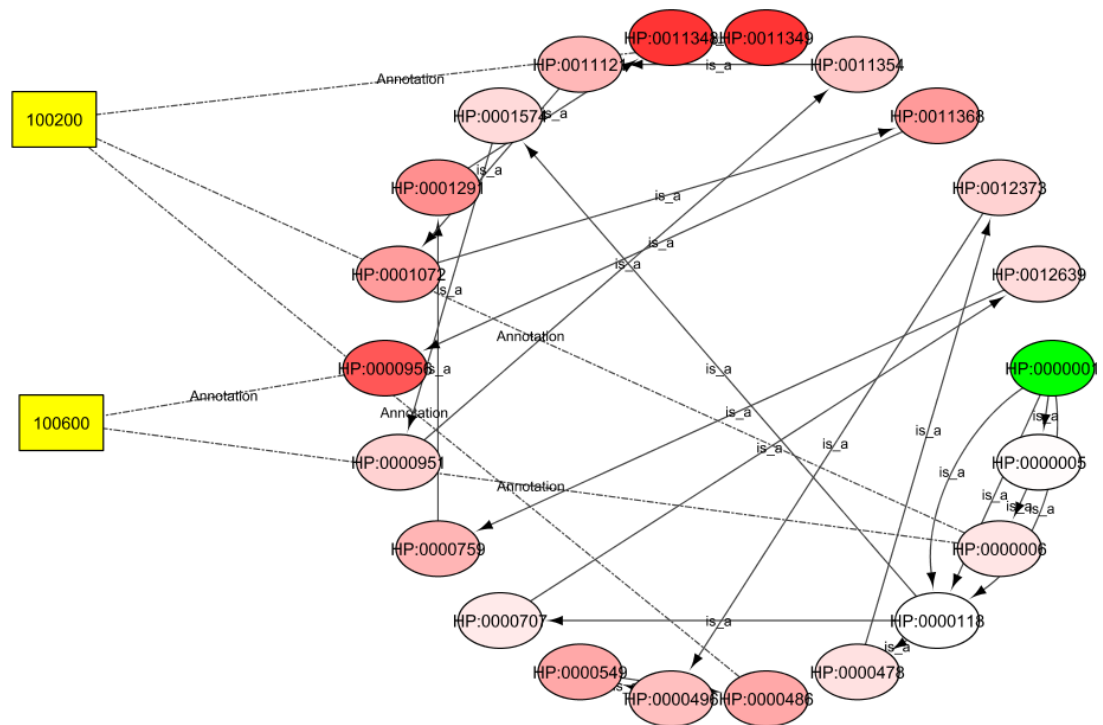
- With **Annotating Terms** (e.g., with 2 entities 100200 and 100600)



- With **Annotating Terms and Their Ancestors**



- With **Annotating Terms and Their Descendants**



- With **Functional Similarity Interactions**
 - Set thresholds (Min, Max) to visualize functional similarities among selected entities in a form of network

☒ Functional Similarity Interactions

Min Max



- Note that:
 - To show more detail information of nodes and links in each visualization, users should select **Node Table** and **Edge Table** tabs of Cytoscape.

For example: Detail information of Term and Entity

ID ▾	Information Content (IC)	Name
100200		ABDUCENS PALSY
100600		ACANTHOSIS NIGRICANS
HP:0000001	-0.0	All
HP:0000005	0.045010740018377814	Mode of inheritance
HP:0000006	0.3869904138502123	Autosomal dominant inheritance
HP:0000118	0.005563227857864829	Phenotypic abnormality
HP:0000478	0.4552379741481966	Abnormality of the eye
HP:0000486	1.2997664475471735	Strabismus
HP:0000496	0.9551759029129288	Abnormality of eye movement
HP:0000549	1.2958002603295722	Abnormal conjugate eye movement
HP:0000707	0.3149020745351841	Abnormality of the nervous system
HP:0000759	1.1136104830624403	Abnormal peripheral nervous system morphology
HP:0000951	0.6666378759663608	Abnormality of the skin
HP:0000956	2.5132842045434782	Acanthosis nigricans
HP:0001072	1.4839004268582687	Thickened skin
HP:0001291	1.680775291837242	Abnormality of the cranial nerves
HP:0001574	0.5525756528549217	Abnormality of the integument
HP:0011121	1.0630350962241173	Abnormality of skin morphology
HP:0011348	3.036162949823816	Abnormality of the sixth cranial nerve
HP:0011349	3.036162949823816	Abducens palsy

Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser

For example: Detail information of Term-Term and Entity-Term relations

ID	interaction /
HP:0011349 (Object-Term) 1002...	Object-Term
HP:0000006 (Object-Term) 1006...	Object-Term
HP:0000486 (Object-Term) 1002...	Object-Term
HP:0000006 (Object-Term) 1002...	Object-Term
HP:0000956 (Object-Term) 1006...	Object-Term
HP:0001291 (is_a) HP:0000759	is_a
HP:0011368 (is_a) HP:0001072	is_a
HP:0000006 (is_a) HP:0000005	is_a
HP:0000118 (is_a) HP:0000001	is_a
HP:0011121 (is_a) HP:0011354	is_a
HP:0012639 (is_a) HP:0000707	is_a
HP:0000707 (is_a) HP:0000118	is_a
HP:0000956 (is_a) HP:0011368	is_a
HP:0000549 (is_a) HP:0000496	is_a
HP:0011349 (is_a) HP:0011348	is_a
HP:0000486 (is_a) HP:0000549	is_a
HP:0001574 (is_a) HP:0000118	is_a
HP:0000478 (is_a) HP:0000118	is_a
HP:0011354 (is_a) HP:0000951	is_a
HP:0011348 (is_a) HP:0001291	is_a

Node Attribute Browser
Edge Attribute Browser
Network A

- e. *Assess functional similarity between entity sets*
- Define elements of two sets, then click to calculate the functional similarity between two selected sets. For example, here is the functional similarity between two gene sets as following:

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) Entity (Set) **Between Entity Sets**

Select Entity Set #1

Select	ID	Name
<input type="checkbox"/>	10	NAT2
<input type="checkbox"/>	100	ADA
<input type="checkbox"/>	1000	CDH2
<input type="checkbox"/>	10000	AKT3
<input type="checkbox"/>	10001	MED6
<input checked="" type="checkbox"/>	10002	NR2E3
<input checked="" type="checkbox"/>	10003	NAAL...
<input checked="" type="checkbox"/>	10004	NAAL...
<input checked="" type="checkbox"/>	1000...	SIGL...
<input type="checkbox"/>	10005	ACOT8

Select Entity Set #2

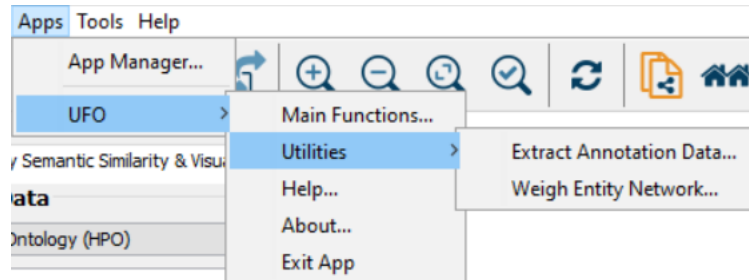
Select	ID	Name
<input checked="" type="checkbox"/>	10	NAT2
<input checked="" type="checkbox"/>	100	ADA
<input checked="" type="checkbox"/>	1000	CDH2
<input checked="" type="checkbox"/>	10000	AKT3
<input checked="" type="checkbox"/>	10001	MED6
<input type="checkbox"/>	10002	NR2E3
<input type="checkbox"/>	10003	NAAL...
<input type="checkbox"/>	10004	NAAL...
<input type="checkbox"/>	1000...	SIGL...
<input type="checkbox"/>	10005	ACOT8

Calculate Entity Set Functional Similarity

Similarity is: 0.3193651136887307

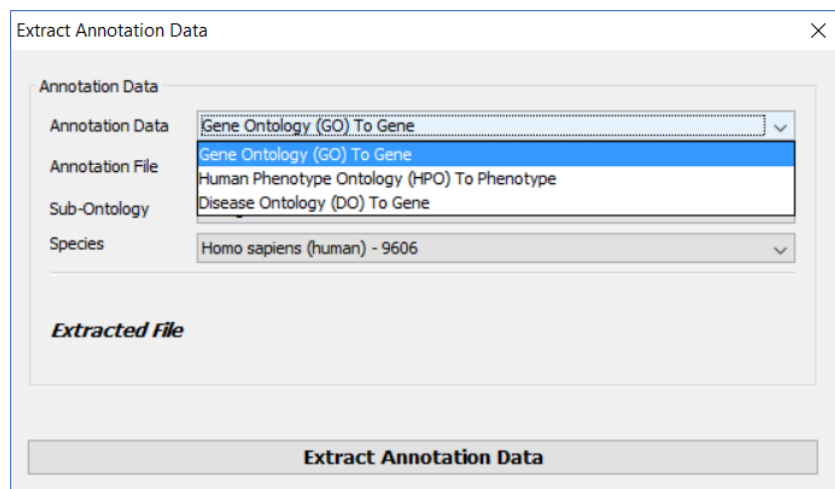
2.5 Utilities

Select Apps → UFO → Utilities...



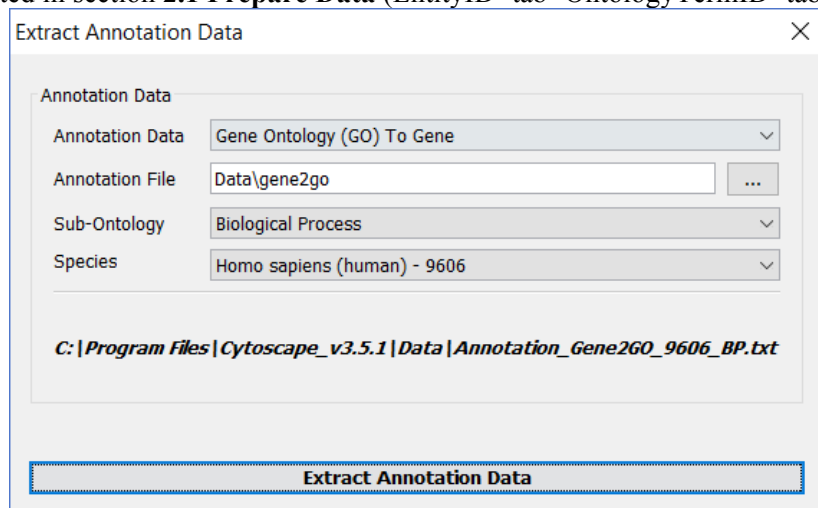
UFO have functions to extract annotation data for genes and phenotypes and to weigh entity network.

2.5.1 Extract annotation data



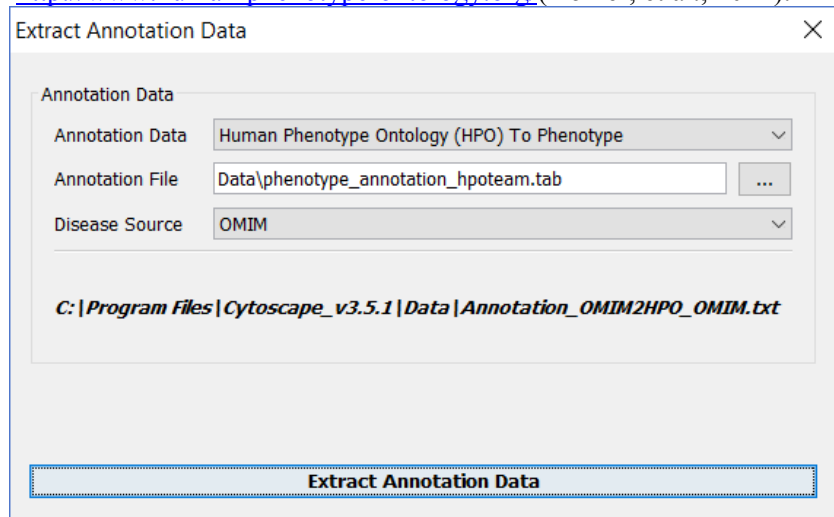
a. Gene Ontology (GO) To Gene

Annotation data of gene by Gene Ontology (gene2go) can be downloaded from NCBI FTP site (<ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz>), then user can select Sub-Ontology (Biological Process, Cellular Component, and Molecular Functions) and Species to extract annotation data to store in file with format indicated in section 2.1 Prepare Data (EntityID<tab>OntologyTermID<tab>EvidenceCode)



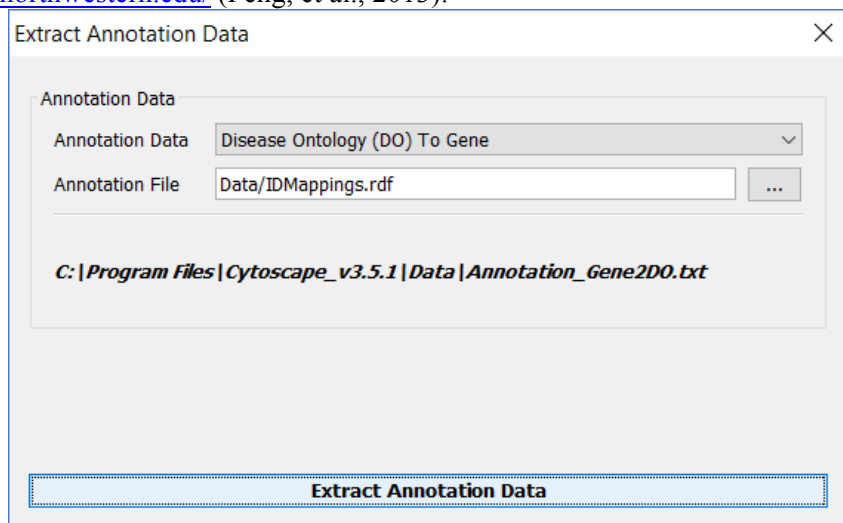
b. *Human Phenotype Ontology (HPO) To Phenotype*

Annotation data of phenotype by Human Phenotype Ontology (phenotype_annotation_hpoteam.tab) can be downloaded from <http://www.human-phenotype-ontology.org/> (Köhler, et al., 2014).



c. *Disease Ontology (DO) To Gene*

Annotation data of phenotype by Human Disease Ontology (IDMappings.rdf) can be downloaded from <http://dga.nubic.northwestern.edu/> (Peng, et al., 2013).

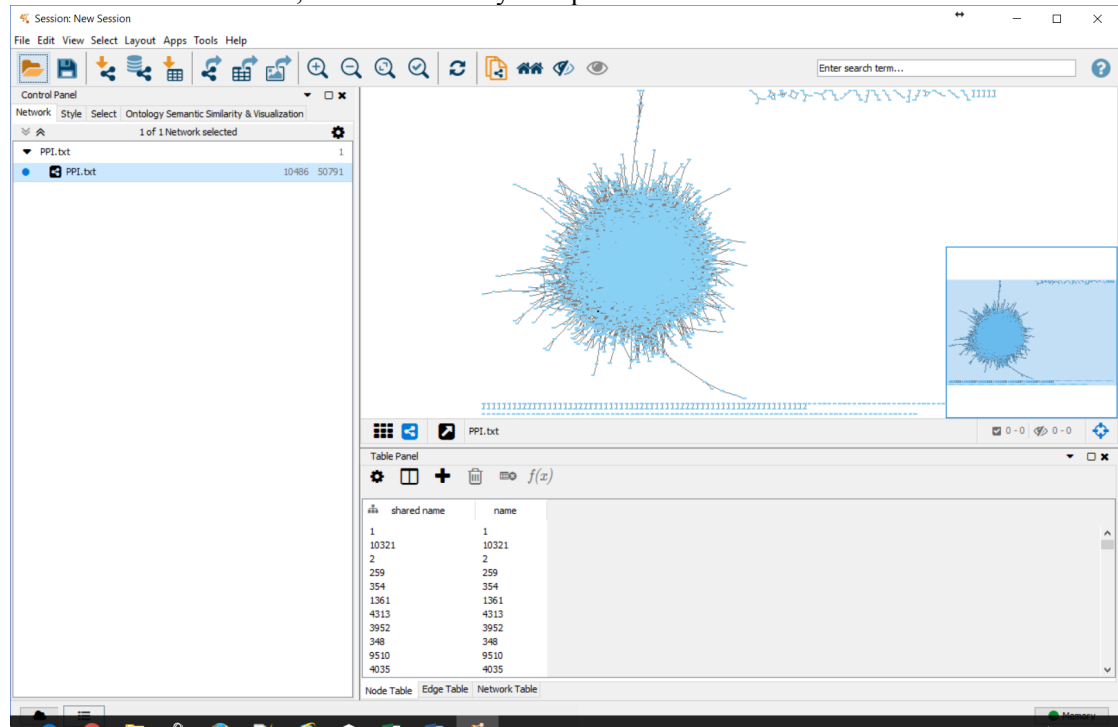


2.5.2 Weigh entity network

The entity network can be a network of genes/proteins, a network of phenotypes. In the study (Le and Kwon, 2013), we used GO-based similarity to weigh protein interaction network to create gene similarity networks using three types of gene ontology i.e., biological process, cellular component and molecular function.

To weigh an entity network

- An entity network must be loaded to Cytoscape (Note that: Network nodes must be identified by the same ID as Entity ID in the used Annotation data). For example:
 - o A physical protein interaction network is downloaded from NCBI FTP site (<ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interactions.gz>), where proteins are identified by Entrez Gene ID. Then, it is loaded to Cytoscape.



- Then, a suitable ontology and annotation data must be used. In this case, proteins in the loaded network are identified by Entrez Gene ID, therefore, Gene Ontology and Annotation must be used. For example, sub-ontology **Gene Ontology (GO) – Biological Process** and annotation **Gene Ontology (GO) To Gene** are used by clicking on **Load & Preparing Data**

Network Style Select **Ontology Semantic Similarity & Visualization**

Ontology & Annotation Data

Ontology: Gene Ontology (GO) - Biological Process

File (*.obo): Data\go.obo

Annotation: Gene Ontology (GO) To Gene

File (tab): Data\Annotation_Gene2GO_9606_BP.txt

Evidence:

<input checked="" type="checkbox"/>	All	<input checked="" type="checkbox"/>	User-d...	<input checked="" type="checkbox"/>	ICE
<input checked="" type="checkbox"/>	IEA	<input checked="" type="checkbox"/>	ISS	<input checked="" type="checkbox"/>	TAS
<input checked="" type="checkbox"/>	ND	<input checked="" type="checkbox"/>	IDA	<input checked="" type="checkbox"/>	IMP
<input checked="" type="checkbox"/>	IPI	<input checked="" type="checkbox"/>	NAS	<input checked="" type="checkbox"/>	IEP
<input checked="" type="checkbox"/>	IGI	<input checked="" type="checkbox"/>	IC	<input checked="" type="checkbox"/>	RCA

Load & Prepare Data

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) Entity (Set) Between Entity Sets

Input Semantic Similarity Visualization

Select a Term/Term Set ☐ All **Total of 26930**

Select	ID	Name	Informati...	Annotat...
<input type="checkbox"/>	GO:0000001	mitochond...		∞
<input type="checkbox"/>	GO:0000002	mitochond...	3.145 1763, 189...	
<input type="checkbox"/>	GO:0000003	reproduction	1.347 100, 1000...	
<input type="checkbox"/>	GO:0000011	vacuole in...		∞
<input type="checkbox"/>	GO:0000012	single stra...	3.3 10013331...	
<input type="checkbox"/>	GO:0000017	alpha-gluc...		∞
<input type="checkbox"/>	GO:0000018	regulation...	2.713 10111, 29...	
<input type="checkbox"/>	GO:0000019	regulation...	3.543 10111, 42...	
<input type="checkbox"/>	GO:0000022	mitotic spi...	3.844 9055, 9493	
<input type="checkbox"/>	GO:0000023	maltose m...	4.145 2548	
<input type="checkbox"/>	GO:0000024	maltose bi...		∞
<input type="checkbox"/>	GO:0000025	maltose c...		∞
<input type="checkbox"/>	GO:0000027	ribosomal l...	4.145 6152	
<input type="checkbox"/>	GO:0000028	ribosomal ...	3.668 6194, 620...	
<input type="checkbox"/>	GO:0000032	cell wall m...		∞

- After that, a semantic similarity measure between terms (GO) and a functional similarity measure between entities (Gene) must be specified. For example, semantic similarity measure *Node-based: Resnik* (Resnik 1995) and functional similarity measure *Pairwise-Based: Avg* (Average) are specified.

Similarity Calculation, Visualization & Enrichment Analysis

Term (Set) Entity (Set) Between Entity Sets

Input Semantic Similarity Visualization

Category Node-Based

Method Node-Based: Resnik (Resnik 1995)

Similarity Calculation, Visualization & Enrichment Analysis

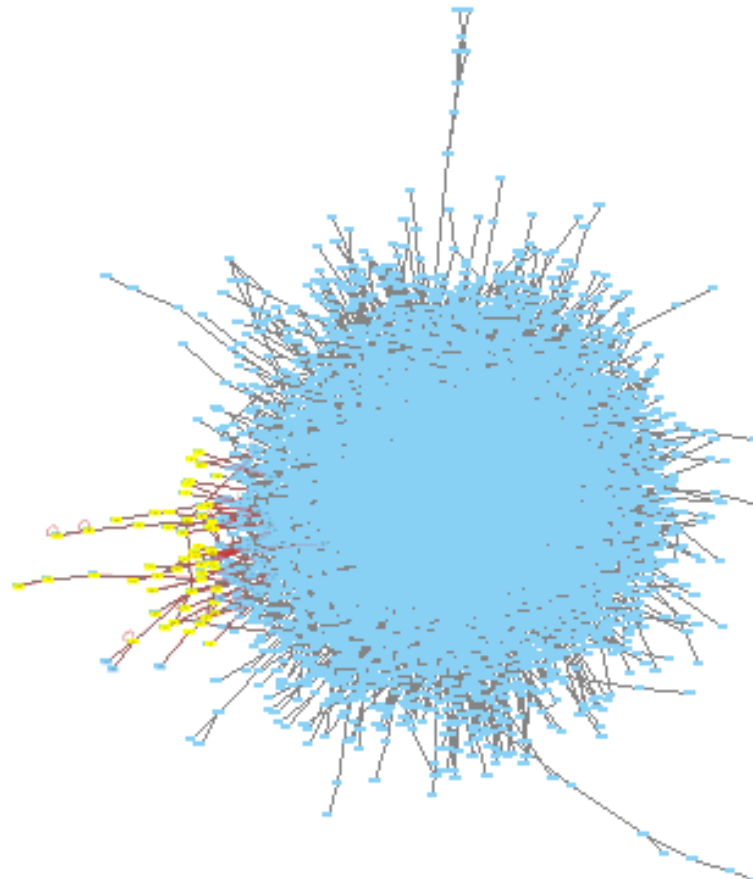
Term (Set) Entity (Set) Between Entity Sets

Input Functional Similarity Enrichment Visualization

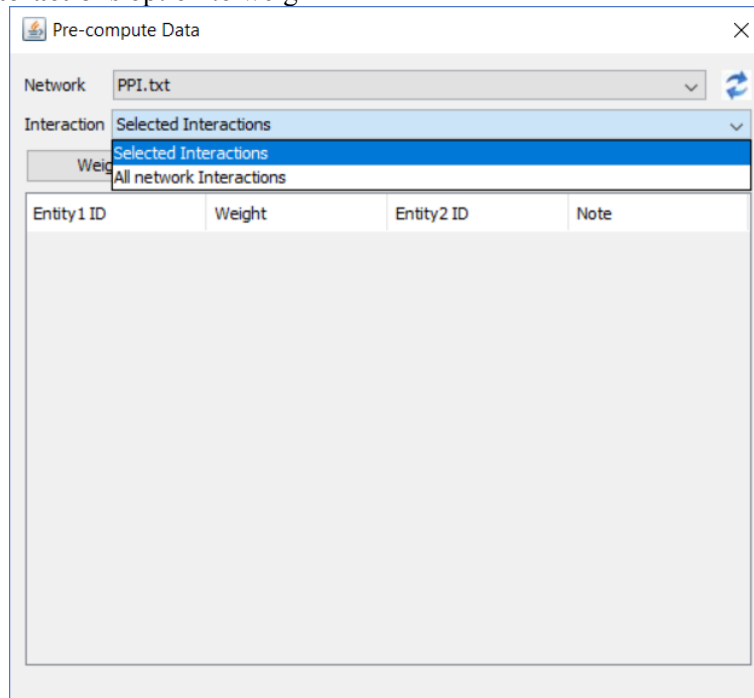
Category Pairwise

Method Pairwise-Based: Avg (Average)

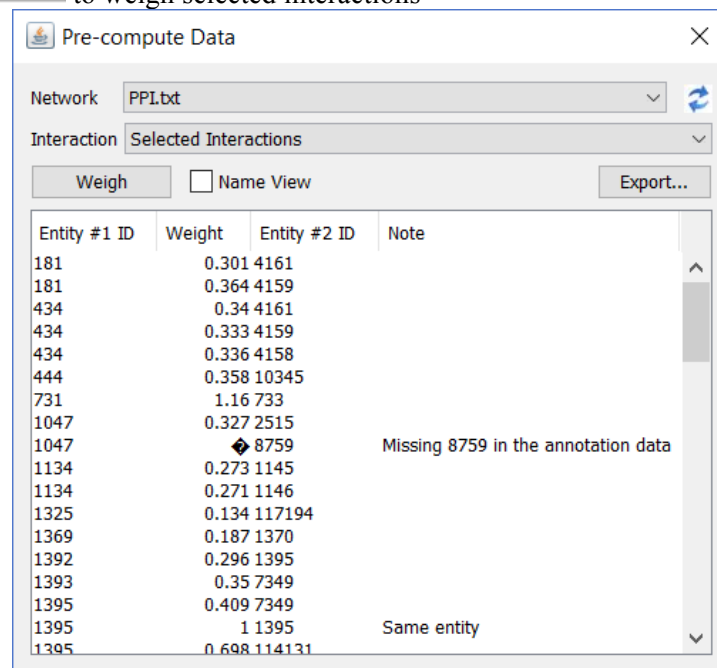
- Select some interactions in the loaded network to be weighted



- Go to Apps → UFO → Utilities... → Weigh Entity Network..., select the loaded network and **Selected Interactions** option to weigh



- Click **Weigh** to weigh selected interactions



- If option **All network interactions** is selected, then all interactions in the loaded network will be weighted.

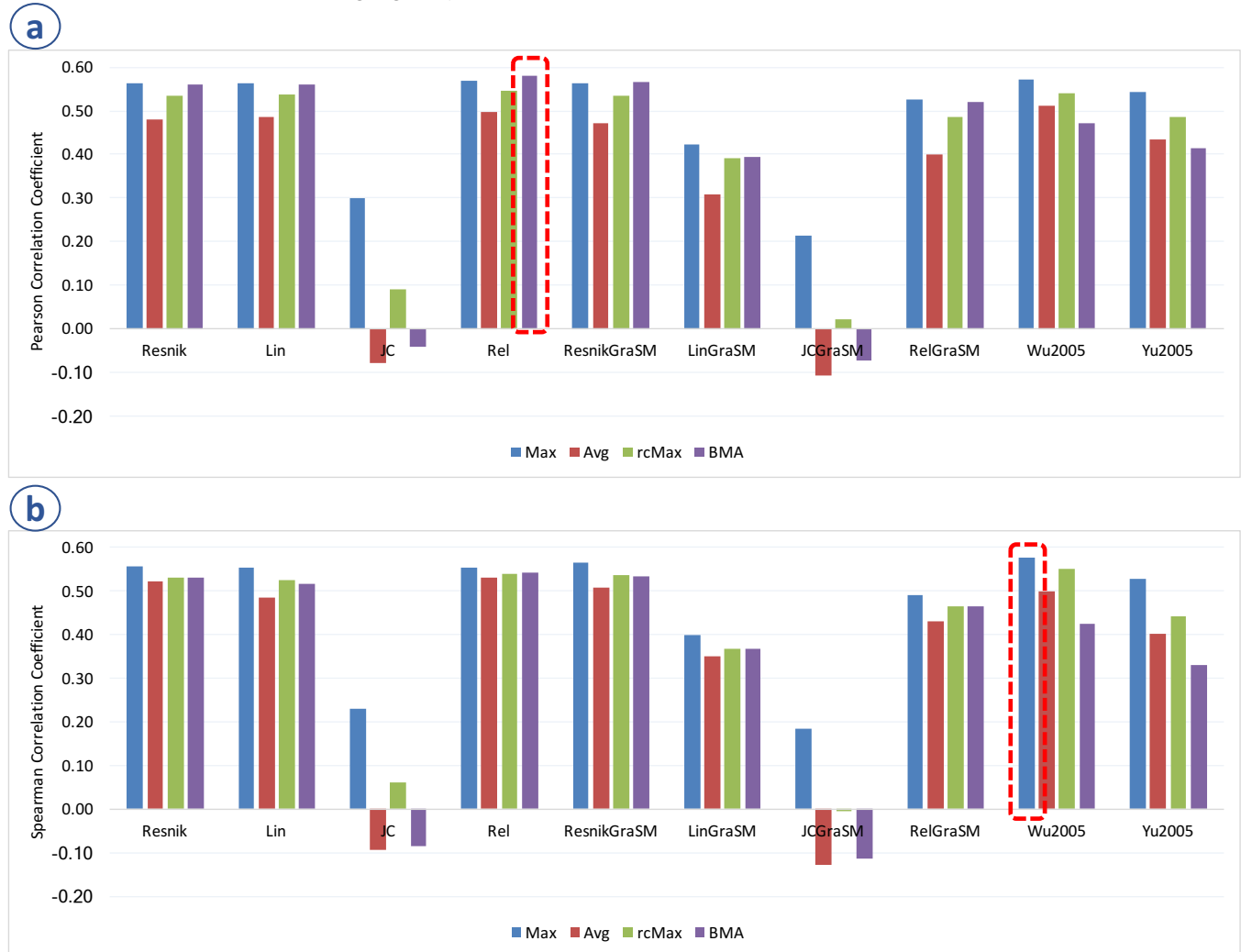
3 CASE STUDY

Comparing different similarity measures could help researchers choose the most appropriate measure for their biological application (Mazandu and Mulder, 2014). Pesquita et al., assessed similarity between genes based on gene ontology (Pesquita, et al., 2009). Since UFO provides batch calculation of similarity, thus we recently estimated the similarity of 9,221,365 pairs of 4,295 phenotypes based on HPO using 47 between-entity similarity measures (Le, et al., 2016). This provides guidelines for future studies which need to choose the most approximate semantic similarity method to assess phenotypic similarity between diseases. In addition, we have employed UFO for constructing gene and protein complex similarity network using GO for predicting disease-associated genes (Le and Kwon, 2013) and protein complexes (Le, 2015), respectively; disease similarity network using HPO for predicting disease-associated genes (Le and Dang, 2016); and disease similarity network using DO and HPO for predicting disease-associated lncRNAs (Le and Dao, 2018) (See more detail in Additional file 3).

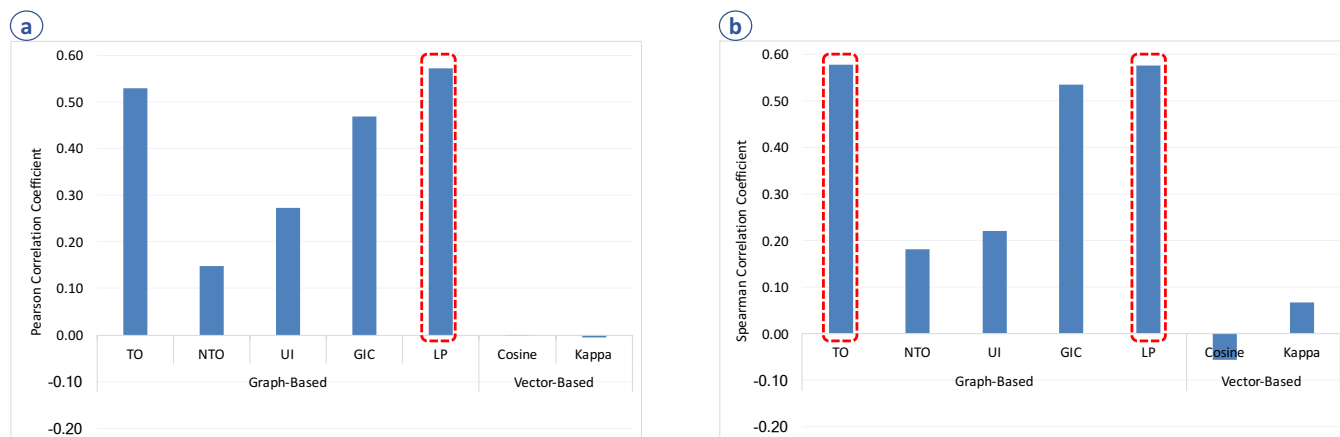
3.1 Assessing human disease phenotype similarity based on ontology

The following figures (adapted from our previous study (Le, et al., 2016)) shows comparison for pairwise between-term measures using Pearson and Spearman correlation coefficients.

For pair-wise methods, the results showed that, for both types of correlation, the largest correlation coefficient is 0.58 for BMA (Azuaje, et al., 2005; Couto, et al., 2005) and Max pairwise between-entity similarity with Rel (Schlicker, et al., 2006) and Wu2005 (Wu, et al., 2005) between-term measures, respectively (Le, et al., 2016) (see the red-dash box of the following figures).

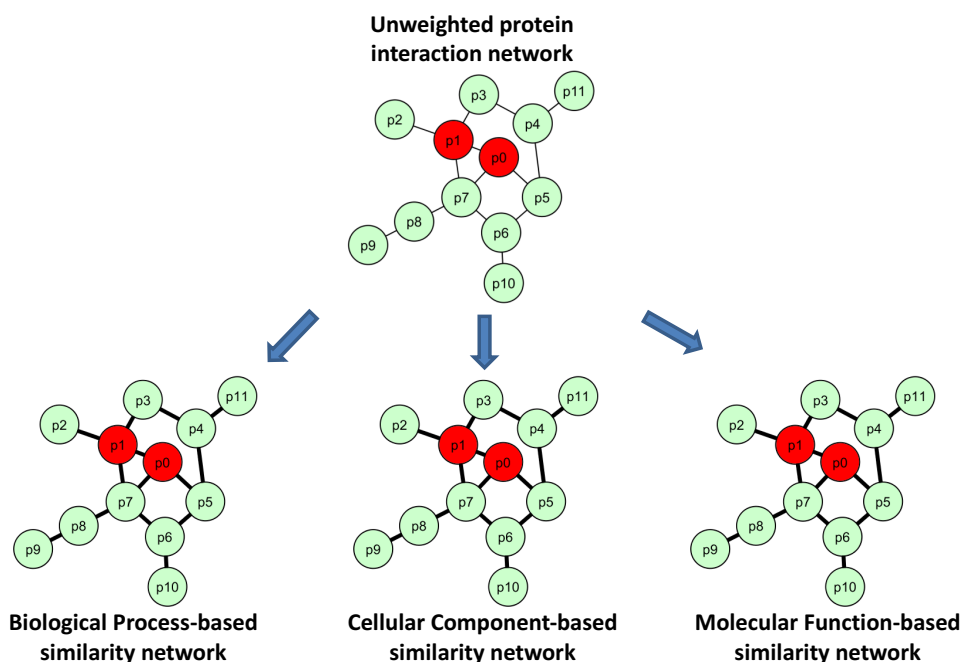


For group-wise methods, best measures were LP between-entity (Gentleman, 2005) for both correlation methods and TO between-entity method for Spearman correlation (Le, et al., 2016) (see the red-dash box of the following figures).

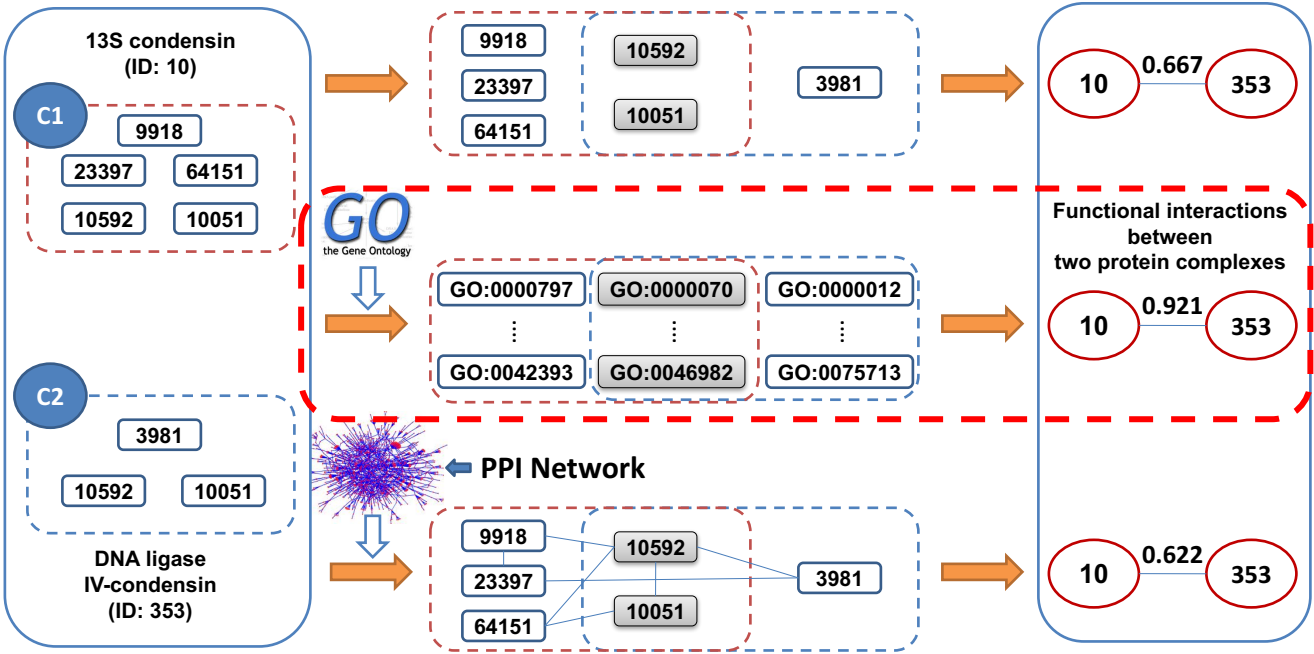


3.2 Construct gene and protein complex similarity network using GO for predicting disease-associated genes and protein complexes

The following figure illustrates the proposed method in (Le and Kwon, 2013). In the study, we weigh an unweighted protein interaction using three sub-gene ontologies (i.e., biological process, cellular component and molecular function) to create three weighted protein similarity networks as following figures:

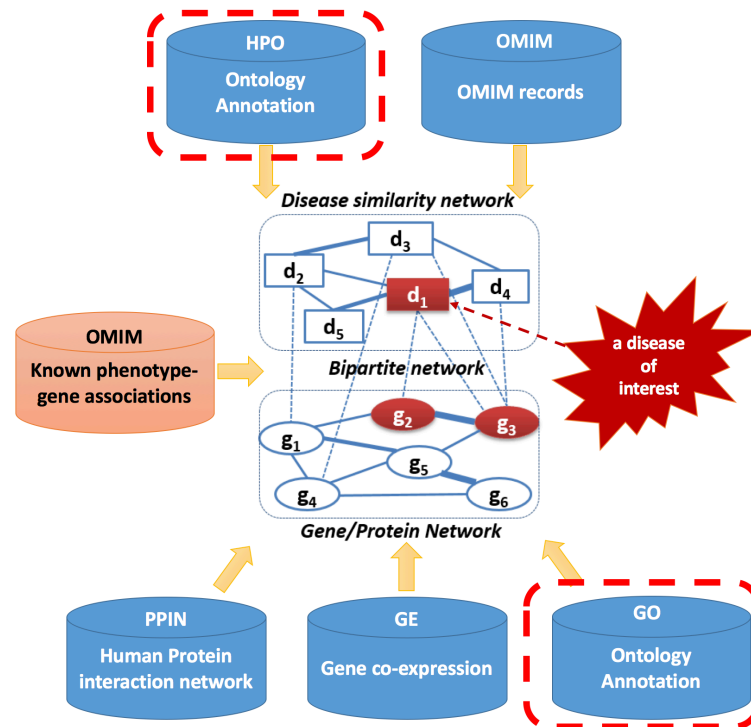


The following figure (adapted from our previous study (Le, 2015)) illustrates how a functional similarity interaction between two protein complexes was created. This interaction was created based on shared GO terms which are used to annotate protein elements in protein complexes (see the red-dash box in the following figures).

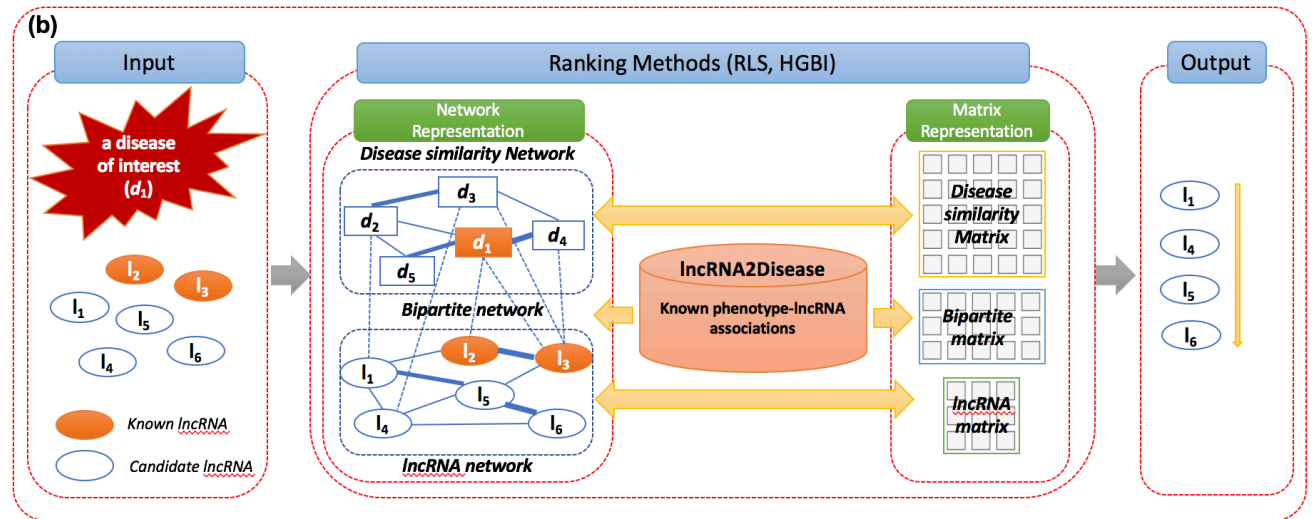
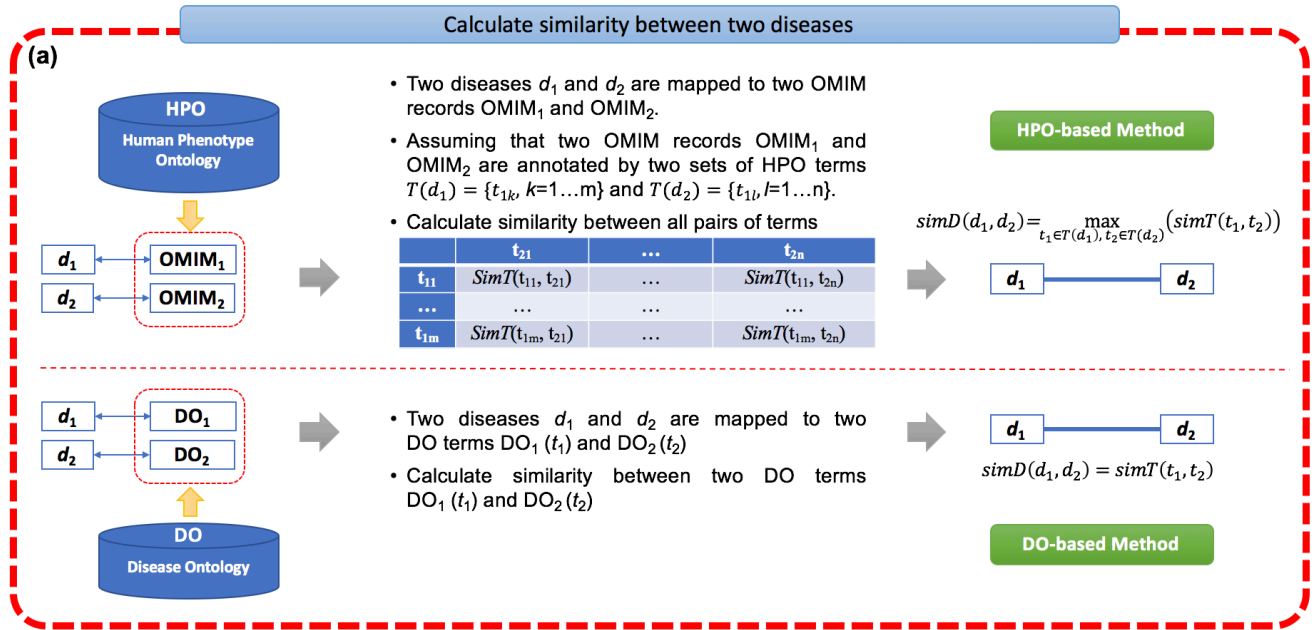


3.3 Construct disease similarity network using HPO and DO for predicting disease-associated genes and lncRNAs

The following figure (adapted from our previous study (Le and Dang, 2016)) illustrates the method proposed in the study. In the study, we calculated semantic similarity for every pair of disease phenotypes using HPO (see the red-dash box in the following figure), then a threshold was set to create a disease similarity network.



The red-dash box in the following figure (adapted from our previous study (Le and Dao, 2018)) shows how to estimate a similarity between two diseases using HPO and DO.



4 REFERENCE

- Azuaje, F., Wang, H. and Bodenreider, O. Ontology-driven similarity approaches to supporting gene functional assessment. *Proc. Of The Eighth Annual Bio-Ontologies Meeting* 2005;Michigan, 25 June.
- Couto, F., Silva, M.r. and Coutinho, P. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In, *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*. Bremen, Germany: ACM; 2005. p. 343-344.
- Gentleman, R. Visualizing and distances using GO. URL <http://www.bioconductor.org/docs/vignettes.html> 2005;38.
- Köhler, S., *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research* 2014;42(D1):D966-D974.
- Le, D.-H. A novel method for identifying disease associated protein complexes based on functional similarity protein complex networks. *Algorithms for Molecular Biology* 2015;10(1):14.
- Le, D.-H. and Dang, V.-T. Ontology-based disease similarity network for disease gene prediction. *Vietnam Journal of Computer Science* 2016;1-9.
- Le, D.-H. and Dao, L.T.M. Annotating Diseases Using Human Phenotype Ontology Improves Prediction of Disease-Associated Long Non-coding RNAs. *Journal of Molecular Biology* 2018;430(15):2219-2230.
- Le, D.-H. and Kwon, Y.-K. Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization. *Computational Biology and Chemistry* 2013;44(0):1-8.
- Le, D.-H., Pham, B.-S. and Dao, A.-M. Assessing human disease phenotype similarity based on ontology. In, *RIVF 2016*. Hanoi: IEEE; 2016. p. 211-216.
- Mazandu, G.K. and Mulder, N.J. Information Content-Based Gene Ontology Functional Similarity Measures: Which One to Use for a Given Biological Data Type? *PLOS ONE* 2014;9(12):e113859.
- Peng, K., *et al.* The disease and gene annotations (DGA): an annotation resource for human disease. *Nucleic acids research* 2013;41(D1):D553-D560.
- Pesquita, C., *et al.* Semantic Similarity in Biomedical Ontologies. *PLoS Comput Biol* 2009;5(7):e1000443.
- Schlicker, A., *et al.* A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 2006;7(1):302.
- Wu, H., *et al.* Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic acids research* 2005;33(9):2822-2837.