AN R-PACKAGE FOR SNP SIMILARITY CALCULATION
Chapter 1: An introduction- what is an ontology?

Chapter 2: Ontologies in biology?

Chapter 3: Research on biomedical ontologies

Chapter 4: The SNP ontology- a new adventure!

Chapter 5: Conclusion

Chapter 6: Thank you!
ontology (noun): the branch of metaphysics dealing with the nature of being.
The philosophical study of the nature of being, becoming, existence, or reality, as well as the basic categories of being and their relations.
In other words, ontologies . . .

- classify/categorize things
- allow us to compare within groups
- allow us to determine level of relatedness between groups
- ontologies alone do not imply annotation!
Biological science: the problem

- inherent complexity in ideas
- difficult to break down mathematically
- classification done in natural language
- even within ‘objective’ fields like genetics:
  - genotype = align sequences
  - phenotype = ???
CHAPTER 2: ONTOLOGIES IN BIOLOGY?

Why do we need them?

- genomics = large amount of data
- need to objectify annotations
- smoothened knowledge sharing
- better computer reasoning
What’s been done?

- biomedical ontologies do exist
- well known ones include Gene Ontology, Sequence Ontology, etc.
  - GO is a great example because it annotates within ontology
- will summarize relevant Pesquita et al., 2009 review sections
CHAPTER 3: RESEARCH ON BIOMEDICAL ONTOLOGIES

Gene ontology example

Figure 1. Section of the GO graph showing the three aspects (molecular function, biological process, and cellular component) and some of their descendant terms. The fact that GO is a DAG rather than a tree is illustrated by the term "transcription factor activity" which has two parents. An example of a part of relationship is also shown between the terms cell part and cell.

doi:10.1371/journal.pcbi.1000443.g001
### Annotating objects in an ontology

#### Found entities

<table>
<thead>
<tr>
<th>Genes/Products</th>
<th>Gene/Product Name</th>
<th>Qualifier</th>
<th>Direct Annotation</th>
<th>Annotation Extension</th>
<th>Assigned by</th>
<th>Taxon</th>
<th>Evidence</th>
<th>Evidence with</th>
<th>PANTHER Family</th>
<th>Isoform</th>
<th>Reference</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>HERC2P3</td>
<td>Putative HERC2-like protein 3</td>
<td></td>
<td>ubiquitin-protein transferase activity</td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR010606</td>
<td>GO_REF:0000002</td>
<td>20151001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HERC2P3</td>
<td>Putative HERC2-like protein 3</td>
<td></td>
<td>protein ubiquitination</td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR010606</td>
<td>GO_REF:0000002</td>
<td>20151001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HERC2P3</td>
<td>Putative HERC2-like protein 3</td>
<td></td>
<td>metal ion binding</td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR010606</td>
<td>GO_REF:0000002</td>
<td>20151001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>ubiquitin-protein transferase activity</td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR000569</td>
<td>GO_REF:0000002</td>
<td>20151001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>protein ubiquitination</td>
<td>GOCC</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR000569</td>
<td>GO_REF:0000002</td>
<td>20151001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>ligase activity</td>
<td>UniProt</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>UniProt:K3-KW/KW-0436</td>
<td>GO_REF:00000038</td>
<td>20151010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>mitochondrial inner membrane</td>
<td>Ensembl</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>Ensembl:ENSMUSP0000075579 hect domain ubiquitin-protein ligase pthv11254</td>
<td>GO_REF:00000019</td>
<td>20151010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>centriole</td>
<td>UniProt</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>UniProt:K3-SubCell:SL-0046 hect domain ubiquitin-protein ligase pthv11254</td>
<td>GO_REF:00000039</td>
<td>20151010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>spermatogenesis</td>
<td>Ensembl</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>Ensembl:ENSMUSP0000075579 hect domain ubiquitin-protein ligase pthv11254</td>
<td>GO_REF:00000019</td>
<td>20151010</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Annotating objects in an ontology

AmiGO is temporarily experiencing an issue with the load and is falling back to this auxiliary site.

Information about Annotations search

---

**Found entities**

Total: 61; showing 1-10

Results count: 10

**HERC2**

<table>
<thead>
<tr>
<th>Gene/product name</th>
<th>Qualifier</th>
<th>Direct annotation</th>
<th>Annotation extension</th>
<th>Assigned by</th>
<th>Taxon</th>
<th>Evidence</th>
<th>Evidence with</th>
<th>PANTHER family</th>
<th>Isoform</th>
<th>Reference</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>HERC2</td>
<td>Putative HERC2-like protein 3</td>
<td>ubquitin-protein transferase activity</td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR010606</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>Putative HERC2-like protein 3</td>
<td>protein ubiquitination</td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR010606</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>Putative HERC2-like protein 3</td>
<td>metal ion binding</td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR010606</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td>ubquitin-protein transferase activity</td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR00569</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td>protein ubiquitination</td>
<td>GOC</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR00569</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td>ligase activity</td>
<td>UniProt</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>UniProt:K3-KW:KW-0436</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td>mitochondrial inner membrane</td>
<td>Ensembl</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>Ensembl:ENSMUSP00000075579</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td>centriole</td>
<td>UniProt</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>UniProt:K3-SubCell:SL-0046</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td>spermatogenesis</td>
<td>Ensembl</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>Ensembl:ENSMUSP00000075579</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td>centriole</td>
<td>UniProt</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>UniProt:K3-SubCell:SL-0046</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td>spermatogenesis</td>
<td>Ensembl</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>Ensembl:ENSMUSP00000075579</td>
<td>PANTHER family</td>
<td>Isoform</td>
<td>Reference</td>
<td>Date</td>
<td></td>
</tr>
</tbody>
</table>
Annotating objects in an ontology

<table>
<thead>
<tr>
<th>Gene/product name</th>
<th>Qualifier</th>
<th>Direct annotation</th>
<th>Annotation assigned by</th>
<th>Taxon</th>
<th>Evidence</th>
<th>Evidence with</th>
<th>PANTHER family</th>
<th>Isoform</th>
<th>Reference</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>HERC2P3</td>
<td>Putative HERC2-like protein 3</td>
<td></td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR010606</td>
<td></td>
<td></td>
<td>GO_REF:0000002</td>
<td>20151010</td>
</tr>
<tr>
<td>HERC2P3</td>
<td>Putative HERC2-like protein 5</td>
<td></td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR010606</td>
<td></td>
<td></td>
<td>GO_REF:0000002</td>
<td>20151010</td>
</tr>
<tr>
<td>HERC2P3</td>
<td>Putative HERC2-like protein 3</td>
<td></td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR010606</td>
<td></td>
<td></td>
<td>GO_REF:0000002</td>
<td>20151010</td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>InterPro</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR000569</td>
<td></td>
<td></td>
<td>GO_REF:0000002</td>
<td>20151010</td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>GOCC</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>InterPro:IPR000569</td>
<td></td>
<td></td>
<td>GO_REF:0000002</td>
<td>20151010</td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>UniProt</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>UniProt:K9-KW-KW-0436</td>
<td></td>
<td></td>
<td>GO_REF:0000038</td>
<td>20151010</td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>Ensembl</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>Ensembl:ENSMUSP00000075579</td>
<td></td>
<td></td>
<td>GO_REF:0000019</td>
<td>20151010</td>
</tr>
<tr>
<td>HERC2</td>
<td>E3 ubiquitin-protein ligase HERC2</td>
<td></td>
<td>Ensembl</td>
<td>Homo sapiens</td>
<td>IEA</td>
<td>Ensembl:ENSMUSP00000075579</td>
<td></td>
<td></td>
<td>GO_REF:0000019</td>
<td>20151010</td>
</tr>
</tbody>
</table>
Annotating objects in an ontology

ubiquitin-protein transferase activity

Accession GO:0004842
Name ubiquitin-protein transferase activity
Ontology molecular_function
Synonyms alt. id: GO:0004840, GO:0004841
E2, E3, anaphase-promoting complex activity, ubiquitin conjugating enzyme activity, ubiquitin lig...
Definition Catalysis of the transfer of ubiquitin from one protein to another via the reaction X-Ub + Y → Y-Ub + X, where both X-Ub and Y-Ub are covalent linkages. Source: PMID:9635407, GOC:jht2, GOC:BioGRID
Comment None
History See term history for GO:0004842 at QuickGO
Subset None
Community ON Add usage comments for this term on the GONUTS wiki.
Related to all genes and gene products annotated to ubiquitin-protein transferase activity.
Link to all direct and indirect annotations to ubiquitin-protein transferase activity.
Link to all direct and indirect annotations download (limited to first 10,000) for ubiquitin-protein transferase activity.
Feedback Contact the GO Helpdesk if you find mistakes or have concerns about the data you find here.

Annotations
Graph Views
Inferred Tree View
Ancestors and Children
Mappings

Found entities
Total: 19355; showing 1-10
Gene/product name | Qualifier | Direct annotation | Annotation extension | Assigned by | Taxon | Evidence | Evidence with | PANTHER family | Isoform | Reference
---|---|---|---|---|---|---|---|---|---|---
TRIM24 | Uncharacterized protein | ubiquitin protein ligase activity | | Ensembl | Gallus gallus | Ensembl:ENSMUSP00000031859 | family not named pthr24103 | GO_REF:0000019
UBR4A | Ubiquitin | negative | | Ensembl | Gallus gallus | Ensembl:ENSMUSP00000038480 | family not named | GO_REF:0000019
Annotating objects in an ontology

ubiquitin-protein transferase activity

Definition: Catalysis of the transfer of ubiquitin from one protein to another via the reaction X-Ub + Y \rightarrow Y-Ub + X, where both X-Ub and Y-Ub are covalent linkages. Source: PMID:29836447, GO:in2, GO-BioGRID.

History: See term history for GO:0004842 at QuickGO.

Related to:
- All genes and gene products annotated to ubiquitin-protein transferase activity.
- All direct and indirect annotations to ubiquitin-protein transferase activity.

Feedback: Contact the GO Helpdesk if you find mistakes or have concerns about the data you find here.
Semantic similarity

- biomedical ontology annotations use semantic similarity measures
- objectifies measurement of similarities between terms
- based on ontology usage (will explain later)
Important concepts in open biomedical ontologies

- **true path rule**: annotation of a term applies to all ancestors (direct vs indirect annotation)

- **edge-based approach**: number of edges between two terms *(we will ignore)*

- **node-based approach**: comparison of properties between terms

- **information content (IC)**: how informative a term is

- **MICA**: most informative common ancestor
Node-based approach

- uses IC in the calculations

- $IC = -\log p(c)$, where $p(c)$ is the probability that a term occurs

- can use IC of the MICA to determine semantic similarity
  - high IC of a MICA between terms = closely related
Semantic similarity calculation example
Semantic similarity calculation example

- let’s say ‘root’ encompasses 1000 gene database
Semantic similarity calculation example

- let’s say ‘root’ encompasses 1000 gene database
- ‘root’ annotates all 1000 genes; therefore $p(root) = 1000/1000 = 1$
- therefore $IC(root) = -\log(1) = 0$, root is not informative at all!
Semantic similarity calculation example
CHAPTER 3: RESEARCH ON BIOMEDICAL ONTOLOGIES

Semantic similarity calculation example

- let’s look at similarity between C1 and C2
Semantic similarity calculation example

- let’s look at similarity between C1 and C2
- MICA is the most informative link between C1 and C2
- let’s say MICA is the ‘apoptosis’ category, annotates 100 genes
Semantic similarity calculation example

- let’s look at similarity between C1 and C2
- MICA is the most informative link between C1 and C2
- let’s say MICA is the ‘apoptosis’ category, annotates 100 genes
let’s look at similarity between C1 and C2

MICA is the most informative link between C1 and C2

let’s say MICA is the ‘apoptosis’ category, annotates 100 genes

\( p(\text{apop}) \) is \( \frac{100}{1000} = 0.1 \)

therefore IC = \(-\log(0.1) = 1\)
Semantic similarity calculation example

Graph:
- 1000 gene db
- apoptosis
- C1
- C2
Semantic similarity calculation example

now let’s say C2 is apoptotic bodies, MICA of C3 and C4
CHAPTER 3: RESEARCH ON BIOMEDICAL ONTOLOGIES

Semantic similarity calculation example

- now let’s say C2 is apoptotic bodies, MICA of C3 and C4
now let’s say C2 is apoptotic bodies, MICA of C3 and C4

let’s say 22 genes are annotated by ‘apoptotic bodies’

\[ p(\text{apopbod}) = \frac{22}{1000} = 0.022 \]

therefore IC = \(-\log(0.1) = 1.67\)

C3/C4 (IC=1.67) are more closely related than C1/C2 (IC=1)
CHAPTER 3: RESEARCH ON BIOMEDICAL ONTOLOGIES

Methods of using IC for semantic similarity

- Method 1: $sim_{Res}(c_1, c_2) = IC(c_{MICA})$
  - but what about distance from MICA?
Factoring in distance from the MICA
CHAPTER 3: RESEARCH ON BIOMEDICAL ONTOLOGIES

Factoring in distance from the MICA

C1 and C2 will have IC of x
Factoring in distance from the MICA

- C1 and C2 will have IC of x
- C1 and C6 will also have IC of x
Factoring in distance from the MICA

- C1 and C2 will have IC of x
- C1 and C6 will also have IC of x
- C1 and C2 are (using method 1) considered equally related as C1 and C6...
**Methods of using IC for semantic similarity (improved)**

- **Method 2:**
  
  \[ sim_{Lin}(c_1,c_2) = \frac{2 \times IC(c_{MICA})}{IC(c_1) + IC(c_2)} \]

  to take into account the distance from the MICA

- **Method 3:**
  
  \[ sim_{Jc}(c_1,c_2) = 1 - (IC(c_1) + IC(c_2) - 2 \times IC(c_{MICA})) \]

- **Method 4:**
  
  \[ sim_{Rel}(c_1,c_2) = sim_{Lin}(c_1,c_2) \times (1 - p(c_A)) \]

  to weigh in absolute IC of the MICA
Deriving similarity between two genes

- Let’s say we want to compare two genes, G1 and G2.
- Each gene is annotated by several terms.
- Can use the node-based approach to compare terms.
- Based on term comparisons, can calculate gene similarity.
Important concepts in open biomedical ontologies

- pairwise comparisons: compare terms using all of their sub-terms
- group-wise comparisons: disregard sub-terms (we will ignore)
Pairwise comparisons to derive similarity between two genes

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>T2</td>
<td>T4</td>
</tr>
<tr>
<td>T3</td>
<td>T5</td>
</tr>
<tr>
<td></td>
<td>T6</td>
</tr>
</tbody>
</table>
Pairwise comparisons to derive similarity between two genes
Methods of using pairwise comparisons
CHAPTER 3: RESEARCH ON BIOMEDICAL ONTOLOGIES

Methods of using pairwise comparisons

12 comparisons in total
Methods of using pairwise comparisons

Method 1: pairwise max:

- use node-based method to calculate semantic similarity between each term
- the highest similarity value is used as overall gene similarity ‘score’
CHAPTER 3: RESEARCH ON BIOMEDICAL ONTOLOGIES

Methods of using pairwise comparisons

Gene 1

T1
T2
T3

Gene 2

T2
T4
T5
T6

12 comparisons in total
Methods of using pairwise comparisons

Method 2: pairwise average:

- use node-based method to calculate semantic similarity between each term
- find mean of similarity values
- use mean value as overall gene similarity score
Method of using pairwise comparisons (improved): best-match

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>T2</td>
<td>T4</td>
</tr>
<tr>
<td>T3</td>
<td>T5</td>
</tr>
<tr>
<td></td>
<td>T6</td>
</tr>
</tbody>
</table>
Method of using pairwise comparisons (improved): best-match
Method of using pairwise comparisons (improved): best-match
Bi-directionality of best-match pairwise comparisons

- important to calculate best match for each term, in both directions!
- e.g. G1T1 may be best matched to G2T3
- but G2T3 is best matched to G1T4 (higher IC of MICA)
CHAPTER 3: RESEARCH ON BIOMEDICAL ONTOLOGIES

Best-match (b.m.) methods for pairwise comparisons
Best-match (b.m.) methods for pairwise comparisons

- Method 1: b.m. average
- Method 2: b.m. max
- Method 3: b.m. min (complete)

7 comparisons in total
### So which method do we use?

<table>
<thead>
<tr>
<th>Study</th>
<th>Standard</th>
<th>Best Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lord et al. [15]</td>
<td>Sequence similarity</td>
<td>Resnik (average)</td>
</tr>
<tr>
<td>Wang et al. [42]</td>
<td>Gene expression</td>
<td>None</td>
</tr>
<tr>
<td>Sevilla et al. [26]</td>
<td>Gene expression</td>
<td>Resnik (max)</td>
</tr>
<tr>
<td>Couto et al. [9]</td>
<td>Family similarity</td>
<td>Jiang and Conrath</td>
</tr>
<tr>
<td>Schlicker et al. [16]</td>
<td>Sequence and family similarity</td>
<td>Schlicker et al.</td>
</tr>
<tr>
<td>Lei et al. [29]</td>
<td>Subnuclear location</td>
<td>TO</td>
</tr>
<tr>
<td>Guo et al. [43]</td>
<td>Human regulatory pathways</td>
<td>Resnik (average)</td>
</tr>
<tr>
<td>Wang et al. [25]</td>
<td>Clustering</td>
<td>Wang et al.</td>
</tr>
<tr>
<td>Pesquita et al. [44]</td>
<td>Sequence similarity</td>
<td>simGIC</td>
</tr>
<tr>
<td>Xu et al. [45]</td>
<td>PPI/gene expression</td>
<td>Resnik(Max)</td>
</tr>
<tr>
<td>Mistry et al. [31]</td>
<td>Sequence similarity</td>
<td>TO/Resnik(Max)</td>
</tr>
</tbody>
</table>

What we can draw from these studies is that there is no clear best measure for comparing terms or gene products. Different measures have performed differently under different circumstances, and a given measure can be well suited for a specific task but perform poorly in another. For instance, simUI was found by Guo et al. to be effective for comparing terms, while Resnik was found to perform better in different scenarios.
So which method do we use?

<table>
<thead>
<tr>
<th>Study</th>
<th>Standard</th>
<th>Best Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lord et al. [15]</td>
<td>Sequence similarity</td>
<td>Resnik (average)</td>
</tr>
<tr>
<td>Wang et al. [42]</td>
<td>Gene expression</td>
<td>None</td>
</tr>
<tr>
<td>Sevillla et al. [26]</td>
<td>Gene expression</td>
<td>Resnik (max)</td>
</tr>
<tr>
<td>Couto et al. [9]</td>
<td>Family similarity</td>
<td>Jiang and Conrath</td>
</tr>
<tr>
<td>Schlicker et al. [16]</td>
<td>Sequence and family similarity</td>
<td>Schlicker et al.</td>
</tr>
<tr>
<td>Lei et al. [29]</td>
<td>Subnuclocal location</td>
<td>TO</td>
</tr>
<tr>
<td>Guo et al. [43]</td>
<td>Human regulatory pathways</td>
<td>Resnik (average)</td>
</tr>
<tr>
<td>Wang et al. [25]</td>
<td>Clustering</td>
<td>Wang et al.</td>
</tr>
<tr>
<td>Pesquita et al. [44]</td>
<td>Sequence similarity</td>
<td>simGIC</td>
</tr>
<tr>
<td>Xu et al. [45]</td>
<td>PPI/gene expression</td>
<td>Resnik(Max)</td>
</tr>
<tr>
<td>Mistry et al. [31]</td>
<td>Sequence similarity</td>
<td>TO/Resnik(Max)</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pcbi.1000443.t004

What we can draw from these studies is that there is no clear best measure for comparing terms or gene products. Different measures have performed differently under different circumstances, and a given measure can be well suited for a specific task but perform poorly in another. For instance, simUI was found by Guo et al. to be the best measure for gene products, while simGIC was preferred for checking whether two terms are related.
What do we (Hai, Katie, and I) want to do?

- put together an R package
- goal is to be able to calculate semantic similarity between SNPs- using the experimental factor ontology (EFO)
What do we (Hai, Katie, and I) want to do?

**Immune system diseases**

Experimental Factor Ontology (EFO):
controlled vocabulary (in a hierarchy) to represent **GWAS Catalog traits** (thus trait-associated genes/SNPs)
What do we (Hai, Katie, and I) want to do?

- put together an R package
- goal is to be able to calculate semantic similarity between SNPs
  - will allow user to use SNP ontology
  - will give options for any of the mentioned methods
  - will process ontologies and help analyze annotations
  - will give insight into underlying relatedness between SNPs
  - can inform of possible new research direction, validate putative SNP interaction
- will also supports enrichment analysis for SNP/genes of interest
To reiterate...

- biomedical ontologies provide objectivity and formalism
- can evaluate similarities in numerous ways
- structuring the ontology vs annotating it
- allow us to describe results that are easily shareable
- allow computers to process our results easily
CHAPTER 6: THANK YOU!

Thank you, everyone . . . literally . . .
Ignoring the edge-based approach

- assumes uniform distribution of nodes and edges
- assumes same level in graph = same semantic distance
- not ideal for biological processes
Ignoring group-wise methods

- cause technical problems (computational)
- too many combinations to consider
- ask Hai to clarify!
Methods of using pairwise comparisons

Method 1: pairwise max

- too exclusive
- doesn’t account for level of similarity
- doesn’t account for differences

Method 2: pairwise average

- too inclusive
- compares unrelated terms