Using Gene Ontology Databases—Sequentially and Concurrently

Angela B. Shiflet[1], George W. Shiflet[1], Daniel S. Couch[1],
Pietro Hiram Guzzi[2], Mario Cannataro[2]
[1]*Wofford College, Spartanburg, SC, USA*
[2]*University "Magna Græcia" of Catanzaro, Catanzaro, Italy*
shifletab@wofford.edu, shifletgw@wofford.edu, couchds@email.wofford.edu, hguzzi@unicz.it, cannataro@unicz.it

Introduction

"Doctors have always recognized that every patient is unique, and doctors have always tried to tailor their treatments as best they can to individuals. You can match a blood transfusion to a blood type — that was an important discovery. What if matching a cancer cure to our genetic code was just as easy, just as standard? What if figuring out the right dose of medicine was as simple as taking our temperature?"

- President Barack Obama, January 30, 2015

A white female, born in 1911 had a life expectancy of about 54 years. Such a female, born in 2011, would have a life expectancy of about 81years (infoplease 2016). The additional 27 years of life were largely attributable to improved medical treatments that became available during that hundred intervening years. Such things as control of infectious disease with antibiotics, vaccines and various public health interventions, development of medical devices (pacemakers, dialysis, artificial heart valves, endoscopes, scanning devices, etc.), organ transplants, and many other developments all contributed to extending the extent and quality of human life.

In January of 2015, President Obama revealed a new program, the Precision Medicine Initiative, which in the long run aims to improve treatment for virtually all types of human disease (Bloss et al 2011). Many patients may not realize that medical treatments are generally prescribed to treat a statistically average patient. Unfortunately, patients are not necessarily average, and the prescribed medicine or treatment may be ineffective or even harmful to many patients. Up to now, physicians and other healthcare providers have made efforts to make sure their patients get the most appropriate, and presumably best, treatment possible. However, they may often be guided to prescribe remedies that are based on scientific evidence (Evidence Based Medicine (EBM)), which is implemented as 'best practice.' EBM uses data from populations, which is then applied to optimize healthcare for average patients (Sacristán 2013).

What the Precision Medicine Initiative is intended to do is to transform disease prevention and treatment so that doctors focus on individual patients. This approach will incorporate information from individual patient's genes, environment, and lifestyle. Practicing healthcare teams will be able to implement an arsenal of devices that will reveal the underlying factors that determine the nature of disease and the health of a patient. With this improved understanding, these teams can determine the most effective course of treatment or management. Already progress has been made in treatments for certain types of cancers, and clinicians, using genetic information about the cancer and the patient, are better able to select the optimal therapeutic remedy.

In April, 2003, the International Human Genome Sequencing Consortium announced an "essentially finished version of the the human genome sequence" (NHGRI 2010). Since that time, significant advances have been made in gene technologies, contributing to tremendous progress in genetics and genomics. Much of our progress in individualized disease prevention and management has been and will continue to be dependent upon the work done in the field of **genomics**. An individual's genome is defined as all of the genes found in the cells of that individual. The field of genomics, then, is the study of all of those genes and their interactions with each other and the environment. Applying what we have learned from genomics has given rise to what we term "genomic medicine," where genomic information is incorporated into clinical practice (Bloss et al 2011).

Gene Ontology Database

Genomic research is producing enormous amounts of data. A biologist typically stores such data, along with their complex interactions, in a **database**, or an integrated collections of files, instead of multiple independent files. Much data appears in the database only once. However, methods can interrelate and retrieve the information in many different forms.

Not only are scientists experiencing challenges in storing such quantities of data, but, more importantly, they must find fast ways to access and innovative ways to interpret a flood of data from multiple databases. A biologist might infer gene function in one organism from a similar gene in another organism, and two different databases might store the data for the two organisms. Unfortunately, usually the databases employ different formats, different record structures, and even different characterizing terms. For example, what might be termed "membrane component" in one database might be designated "membrane part" in another.

To facilitate comparisons by computers and people, the **Gene Ontology** (**GO**) project, a major bioinformatics initiative, seeks to standardize biological descriptions among participating databases. Thus, these collaborating databases employ the same vocabulary and structured relationships among terms involving genes, which enable queries across databases. Based on experiments published in over 100,000 peer-reviewed scientific papers, GO represents over 40,000 biological concepts with structured terminology (GO 2016, Gene Ontology Consortium 2015). Each GO term represents how a gene encodes a biological function in exactly one of the following levels:

- **Molecular Function** (**MF**), Molecular-level activities of gene products, including bindings and catalyzers
- **Biological Process** (**BP**), Operations related to the functions of cells, tissues, organs, and organisms
- **Cellular Component** (**CC**), Parts of the cell or the environment outside the cell

Besides describing aspects of genes and gene product functionality, each ontology gives **is_a**, **part_of**, **has_part**, or **regulates** relationships between the terms.

With nodes containing the terms and edges representing the relationships, Figure 1 contains **directed acyclic graphs** (**DAG**s) for terms “plasma membrane” and “integral component of membrane,” having Gene Ontology representations GO: 0005886 and GO: 0016021, respectively. The graphs are **directed** because the edges are arrows. Also, each graph is **acyclic** because no cycles exist; it is impossible to start at a node and travel in a circle following the directions of the edges back to the starting node. In Figure 1, solid edges indicate is_a relationships, and dashed arrows represent part_of relationships. The **root** of each tree is the term, “cellular component,” which annotates each gene in the Cellular Component database of GO. All the other terms in these graphs and in GO are **descendants** of this root, which is the most general term. In Figure 1a, the **children** of “cellular component” are the more specific terms, “cell,” “cell part,” and “membrane.” The term “plasma membrane” has **parents** “cell periphery,” “cell part,” and “membrane,” while the term “integral component of membrane” in Figure 1b has parent “intrinsic component of membrane.” The **ancestors** of “plasma membrane” are the five terms that are on **paths** starting with the parents through the root. We employ the notation $\boldsymbol{DAG_A}$ = (A, T_A, E_A) for each graph, where A is the principle term under consideration, T_A is the set of terms in the graph, and E_A is the set of edges. Thus, $DAG_{\text{“plasma membrane”}}$ = (“plasma membrane”, T_A, E_A), where the number of nodes in T_A is 6 and the number of edges in E_A is 8.

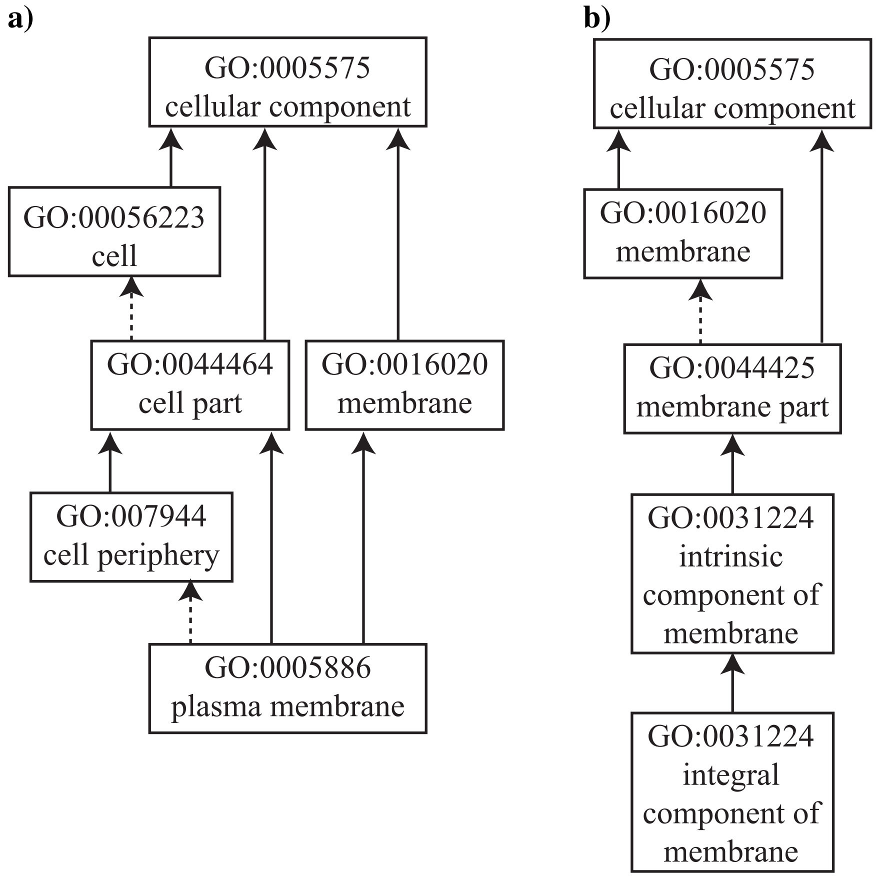

Figure 1 Graphs for terms “plasma membrane” (GO:0005886) and “integral component of membrane” (GO:0016021) in GO (AmiGO 2016) with solid edges indicating is_a relationships and dashed arrows representing part_of relationships

Answers to Quick Review Questions appear at the end of the module.

Quick Review Question 1

a. What do we call an integrated collections of files that minimizes repetition of data?

b. Give the name and acronym for the project that seeks to standardize biological descriptions among participating databases.

c. Give the four types of relationships between terms in these databases

Quick Review Question 2

a. A collection of nodes with edges connecting nodes is a ____.

b. A graph in which the edges are arrows is called ____.

c. A graph in which it is impossible to start at a node and travel in a circle following the edges back to the starting node is called ____.

d. Give the root of the graph in Figure 2b.

e. Give the children of “reproduction” in Figure 2b.

f. Give the parent(s) “anatomical structure morphogenesis” in Figure 2a.

g. Give the ancestors of “developmental process involved in reproduction” in Figure 2b.

h. Give the descendants of “reproduction” in Figure 2b.

i. Give the common ancestors of “anatomical structure morphogenesis” and “developmental process involved in reproduction” in Figure 2.

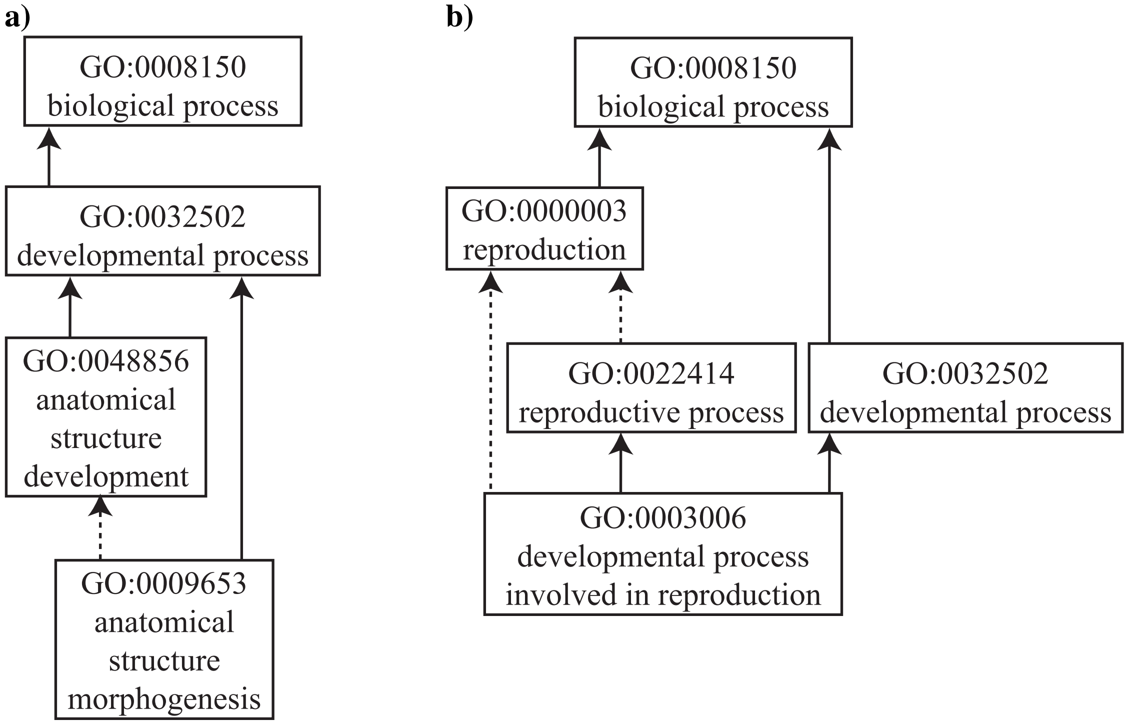

Figure 2 Graphs for terms “anatomical structure morphogenesis” (GO:0009653) and “developmental process involved in reproduction” (GO:0003006) in GO (AmiGO 2016) with solid edges indicating is_a relationships and dashed arrows representing part_of relationships

Ranking Genes Using Ontologies

The **GoD** (**Gene ranking based On Diseases**) package, written in the computer language R, ranks a given set of genes based on ontology annotations (Cannataro et al 2015). The process of obtaining data for GoD begins with an experiment using a microarray with selected probes and a subject’s DNA. A **probe** is a genetically engineered RNA fragment or single-stranded DNA specifically designed to search for a particular gene or sequence of DNA. With a complementary sequence of bases to the target sequence, the probe attaches to the target through base pairing. Thus, the microarray experiment, which has a number of probes, yields a set of probes with genes that are candidates for involvement in a particular disease. For each candidate gene, a list of annotations representing terms is obtained from the following databases: Gene Ontology (GO), Human Phenotype Ontology (HPO), Online Mendelian Inheritance in Man (OMIM), and Disease Ontology (DO). A program using GoD specifies a list of terms of particular interest and a semantic similarity metric, which we discuss shortly. Based on this metric, GoD determines a ranking of the genes' relationship to the disease(s).

The following is a typical annotation record that might appear in an input file for a bioinformatics program using GoD, with boldface separating the different types of data, as indicated in the description below the record:

AM_10396 116085 **SLC22A12** GO:0015849 GO:0055085
GO:0005886 GO:0016021 GO:0015143 GO:0005886 GO:0016324
GO:0042493 GO:0070062 GO:0019725 GO:0031526 GO:0015747
GO:0015747 GO:0030165 GO:0046415 GO:0005886 GO:0055085
GO:0015849 GO:0031526 **HP:0000791 HP:0001919 HP:0000007**
HP:0003537 OMIM:220150 **HP:0000791 HP:0012611**
HP:0001919 HP:0003537 HP:0000007 DOID:1920 DOID:13189
DOID:1074 DOID:13189 DOID:1920 DOID:225

The entry has information for **Probe**, ID, **GENE**, GO_ANNOTATIONS, **GENE_HPO_ANNOTATIONS**, ID_OMIM, **HPO_ANNOTATIONS**, and DO_ANNOTATIONS. The following are descriptions of the various databases employed:

- GO – Gene Ontology – The GO designation contains accession, name, ontology, synonyms, definition, comment, history, subset, community, related, and feedback (GO 2016)
- HPO – Human Phenotype Ontology –The HPO "aims to provide a standardized vocabulary of phenotypic abnormalities encountered in human disease." (HP 2016)
- OMIM® – Online Mendelian Inheritance in Man® – "OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes…." (OMIM 2016)
- DO – Disease Ontology – "The Disease Ontology has been developed as a standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts…." (DOID 2016)

Explanations for terms in the record, as indicated in Google searches, follow:

AM_10396 - Array Design Name[DMET_Plus] Affymetrix human DMET Plus array

116085 – ID for Homo sapiens solute carrier family 22 (organic anion/urate transporter) member 12 (SLC22A12) transcript variant 1 mRNA

SLC22A12 – gene solute carrier family 22 (organic anion/urate transporter), member 12 [Homo sapiens (human)]

GO:0015849 – "organic acid transport" in BP

GO:0055085 – "transmembrane transport" in BP

GO:0005886 – "plasma membrane" in CC

GO:0016021 – "integral component of membrane" in CC

GO:0015143 – "urate transmembrane transporter activity" in MF

GO:0005886 – “plasma membrane” in CC
GO:0016324 – “apical plasma membrane” in CC
GO:0042493 – “response to drug” in BP
GO:0070062 – “extracellular exosome” in CC
GO:0019725 – “cellular homeostasis” in BP
GO:0031526 – “brush border membrane” in CC
GO:0015747 – “urate transport” in BP
GO:0015747 – “urate transport” in BP
GO:0030165 – “PDZ domain binding” in MF
GO:0046415 – “urate metabolic process” in BP
GO:0005886 – “plasma membrane” in CC
GO:0055085 – “transmembrane transport” in BP
GO:0015849 – “organic acid transport” in BP
GO:0031526 – “brush border membrane” in CC

HP:0000791 – “Uric acid nephrolithiasis”
HP:0001919 – AKA “Acute renal failure”
HP:0000007 – “Autosomal recessive inheritance”
HP:0003537 – “Hypouricemia”

OMIM:220150 – “HYPOURICEMIA, RENAL, 1; RHUC1”

HP:0000791 – “Uric acid nephrolithiasis”
HP:0012611 – “AKA: Acute renal failure”
HP:0001919 – AKA “Acute renal failure”
HP:0003537 – “Hypouricemia”
HP:0000007 – “Autosomal recessive inheritance”

DOID:1920 – “hyperuricemia”
DOID:13189 – “gout”
DOID:1074 – “kidney failure”
DOID:13189 – “gout”
DOID:1920 – “hyperuricemia”
DOID:225 – “syndrome”

The programmer indicates a set of probes, such as {AM_10430, AM_10198, AM_10192, AM_10229, AM_10013, AM_10339, AM_10028, AM_10182, AM_10175}, and a set of terms, such as {GO:0015849, GO:0055085, GO:0005886, GO:0016021, GO:0015143, GO:0005886, GO:0016324, GO:0042493}, of particular interest along with a metric to rank the probes.

Information Content

The basis of many of the metrics used in GoD and other ontology analyses is the **information content** (**IC**) of a term. The IC of a term, c, is

$$\boldsymbol{IC(c) = -\log p(c)}$$

where $p(c)$ is the probability of the term, which is usually the number of concepts, such as genes or gene products, annotated with the term and its descendants, divided by the total number of concepts annotated with the term in the ontology. A probability is always between 0 and 1. Moreover, $\log(1) = 0$ and $\log(p(c))$ becomes more and more negative as $p(c)$ approaches zero. Thus, $IC(c)$ is 0 for the root term, which annotates all genes and has a probability of 1, and very large for a term that annotates few genes. For instance, in the Cellular Component database of GO, the root term "cellular component" has *IC*("cellular component") = -log(*p*("cellular component")) = -log(1) = 0. However, as of this writing, the term "membrane" (GO:0016020) and its descendants are associated with 12,815,684 gene products, while the total number of gene products in the total number of gene products in the database is 20,475,119 (AmiGO 2016). Thus *p*("membrane") = 12,815,684 / 20,475,119 = 0.6259, and *IC*("membrane") = -log(0.6259) = 0.2035.

Quick Review Question 3 This question refers to terms in Figure 2. Suppose term "biological process" annotates 30,791,906 genes or gene products, while "developmental process" annotates 588,047, "anatomical structure morphogenesis" annotates 350,203, and "developmental process involved in reproduction" annotates 50,827, such concepts, respectively.

a. Determine *IC*("biological process").

b. Determine *IC*("developmental process").

c. Determine *IC*("anatomical structure morphogenesis").

d. Determine *IC*("developmental process involved in reproduction").

Resnik and Lin Semantic Similarity Scores

The **Resnik similarity score** of two ontology terms, c_1 and c_2, $\boldsymbol{sim_{res}(c_1, c_2)}$, is the information content of the most informative common ancestor of the two terms:

$$\boldsymbol{sim_{res}(c_1,c_2) = \max_{c\in CommonAncestors(c_1,c_2)} IC(c)}$$

where $CommonAncestors(c_1, c_2)$ is the set of common ancestors of c_1 and c_2. Thus, tracing the graphs in Figure 1 from "plasma membrane" and "integral component of membrane," we find the lowest common ancestor to be "membrane." Thus, according to the calculations of the information content of "membrane" in the last section, sim_{res}("plasma membrane", "integral component of membrane") = 0.2035.

One disadvantage of this approach is its inconsistency with the graphical hierarchy of the terms. For instance, looking at Figure 1, sim_{res}("plasma membrane", "integral component of membrane") has the same value as "plasma membrane" paired with terms between "integral component of membrane" and "membrane." That is, sim_{res}("plasma membrane", "intrinsic component of membrane") = sim_{res}("plasma membrane", "membrane part") = 0.2035, even though "intrinsic component of membrane" and "membrane part" are more general terms than "integral component of membrane."

Another disadvantage of the Resnik similarity score is that the values are not between 0 and 1. To overcome this objection, Lin introduced a normalized form, $\boldsymbol{sim_{Lin}(c_1,}$

$\boldsymbol{c_2)}$, that doubles $sim_{res}(c_1, c_2)$ and divides the result by the sum of the information content values for the terms, as follows:

$$\boldsymbol{sim_{Lin}(c_1,c_2) = \frac{2sim_{res}(c_1,c_2)}{IC(c_1) + IC(c_2)}}$$

Thus, with sim_{res}(“plasma membrane”, “integral component of membrane”) = 0.2035, *IC*(“plasma membrane”) = 1.002, and *IC*(“integral component of membrane”) = 0.2563, sim_{Lin}(“plasma membrane”, “integral component of membrane”) = 2 · 0.2035 / (1.00171 + 1.002) = 0.2031.

A disadvantage of both of these metrics is that they only use the information content of terms, ignoring the ontology’s structural information. However, the graphical structure is important because a GO term’s biological meanings, or **semantics**, inherits the semantics of its more general term ancestors.

Quick Review Question 4 This question refers to terms in Figure 2 and Quick Review Question 3.

a. Determine sim_{res}(“anatomical structure morphogenesis”, “developmental process involved in reproduction”).

b. According to this metric, are these two terms more or less similar than “plasma membrane” and “integral component of membrane”?

c. Determine sim_{Lin}(“anatomical structure morphogenesis”, “developmental process involved in reproduction”).

d. According to this metric, are these two terms more or less similar than “plasma membrane” and “integral component of membrane”?

Wang Semantic Similarity Score

In response, Wang designed a graph-based semantic similarity metric. First, we define the **S-value** of a term, *A*, under consideration to be $\boldsymbol{S_A(A)}$ = 1. For an ancestor’s sematic contribution to *A*, we employ weights between 0 and 1 to each of the edges. An is_a relationship is stronger and has a higher edge weight, say 0.8, while a part_of relationship might have an edge weight of 0.6. For each ancestor, *t*, of *A*, we obtain its S-value, $\boldsymbol{S_A(t)}$, by multiplying together the weights of the edges from *A* to *t*. If there is more than one path from *A* to *t*, we pick the maximum of the path calculations.

$$\begin{cases} \boldsymbol{S_A(A) = 1} \\ \boldsymbol{S_A(t) = \max\{weight_e * S_A(t') \mid t' \textbf{ is a child of } t\}, \textbf{ for } t \neq A} \end{cases}$$

where $weight_e$ is the weight, or semantic contribution, of the edge from *t*’ to *t*. Table 1 gives the S-values of the terms in the two graphs in Figure 1.

Term	**Calculation of S-Value**	**S-Value**
Figure 1a		
plasma membrane	1	1
cell periphery	0.6	0.6
cell part	max(0.6*0.8, 0.8)	0.8
membrane	0.8	0.8
cell	0.8*0.6	0.48
cellular component	max(0.48 *0.8, 0.8*0.8, 0.8*0.8)	0.64
Figure 1b		
integral component of membrane	1	1
intrinsic component of membrane	0.8	0.8
membrane part	0.8*0.8	0.64
membrane	0.64*0.6	0.384
cellular component	max(0.384*0.8, 0.64*0.8)	0.512

Table 1 S-values of terms in the graphs of Figure 1 using weights of 0.8 and 0.6 for is_a and part_of relations, respectively

The **sematic value** of a GO term, *A*, $\boldsymbol{SV(A)}$, is the sum of the $S_A(t)$ values for all terms *t* in *A*'s graph, DAG_A:

$$\boldsymbol{SV(A) = \sum_{t \in T_A} S_A(t)}$$

Thus, adding the S-values for each figure in Table 1, we find *SV*("plasma membrane") = 1 + 0.6 + 0.8 + 0.8 + 0.48 + 0.64 = 4.32 and *SV*("integral component of membrane") = 1 + 0.8 + 0.64 + 0.384 + 0.512 = 3.336.

For the **Wang semantic similarity** of terms *A* and *B*, $\boldsymbol{sim_{Wang}(A, B)}$ we take the sum of the S-values of terms that appear in both graphs DAG_A and DAG_B and divide the result by the sum of the SV-values of the terms under consideration, *SV*(*A*) and *SV*(*B*):

$$\boldsymbol{sim_{Wang}(A, B) = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{SV(A) + SV(B)}}$$

Thus, for the values in Table 1 with "membrane" and "cellular component" being the common terms for the two graphs of Figure 1, we calculate the Wang semantic similarity of "plasma membrane" and "integral component of membrane" as follows:

$$\frac{(0.8 + 0.384) + (0.64 + 0.512)}{4.32 + 3.336} = 0.305$$

Addressing the shortcomings of SimResnik and SimLin, the Wang semantic similarity of two terms is determined by the locations of the terms in the GO graph and their ancestor terms' semantic relations. A term, *t*, in T_A and T_B may have different S-values relative to *A* and *B* because because *A* and *B* have different positions in the entire GO graph (Wang et al 2007).

Quick Review Question 5

a. Calculate the S-value for each term in Figure 2a.
b. Calculate the S-value for each term in Figure 2b.
c. Find the semantic value of "anatomical structure morphogenesis."
d. Find *SV*("developmental process involved in reproduction").
e. Find sim_{Wang}("developmental process involved in reproduction").
f. According to this metric, are these two terms more or less similar than "plasma membrane" and "integral component of membrane"?

GoD

The GoD (Gene ranking based On Diseases) package allows the user to specify a set of genes and a sematic similarity measure, such as SimLin or Wang, and returns a histogram of the probe-disease similarity, based on ontology annotations. The probe with the highest bar has the greatest semantic similarity value to the disease under consideration and implicates the corresponding gene most strongly to the disease.

Although GoD is predominantly in sequential R, we can see a speedup using a parallel version that splits the file so that different processes perform calculations on different probes concurrently. Because little communication is required, the concurrent algorithm is called **embarrassingly parallel**. The situation is analogous to giving each student in a class a different independent problem to solve. The results are obtained much faster than if one student is charged with the task of solving all of the problems.

Figure 1 shows the effect of parallelization on runtime. With each test occurring about 30 times, the standard deviation was less than 10 seconds (approximately 6 seconds with four processes, approximately 8 seconds with two, and about 9 seconds with one process). Eight probes were prioritized.

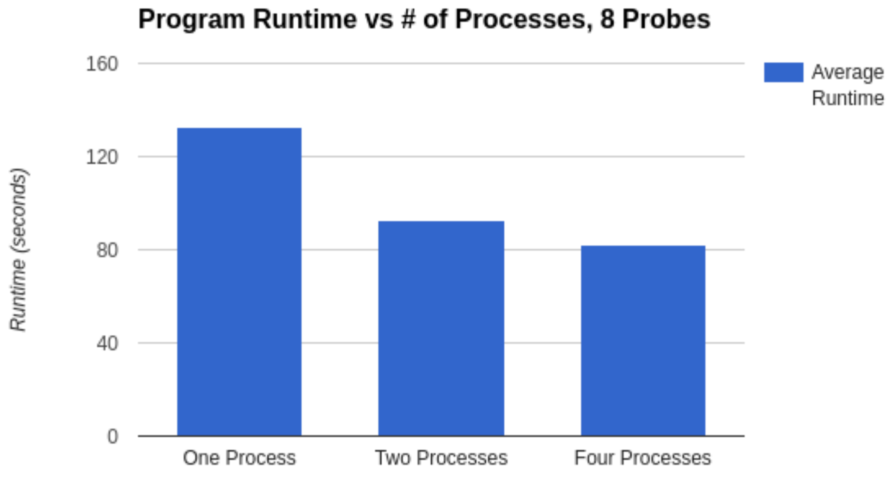

Figure 3 Program runtime with sequential and parallel versions

Projects

1. (Introduction to database usage) Using the R packages "GO.db" and "GOSemSim," write an R program that has the following input and output:
INPUT: A single GO accession ID
OUTPUT: The following information about the GO ID: 1) number of ancestors, 2) number of descendants, 3) general information about the GO term

2. (Introduction to database usage) Using the R packages "GO.db" and "GOSemSim," write an R program that has the following input and output:

INPUT: A single GO accession ID, a list containing GO IDs, a measure (e.g. "Resnik"), and an ontology

OUTPUT: The following information about the GO ID: 1) number of ancestors, 2) number of descendants, 3) general information about the GO term, 4) the computed semantic similarity between the single GO term and each of the GO terms in the list

3. Parallelize the semantic similarity measure part of Project 2.

4. Researchers in a lab purify a protein with the following amino acid sequence:

MIILGIESSCDETSIAVVKDGKEILSNNISSQIEIHKEYGGVVPEIASRQHIKNI
ATVLEESLEEAKITLDDVDYIAVTYAPGLIGALLVGVSFAKGLSYAKNIPIIP
VHHIKGHMYANFLEHDVELPCISLVVSGGHTNIIYIDENHNFINIGETLDDAV
GESCDKVARVLGLGYPGGPVIDKMYYKGDRDFLKITKPKVSRFDFSFSGIK

TAIINFDNNMKMKNQEYKKEDLAASFLGTVVDILCDKTLNAAVEKNVKTI
MLAGGVAANSLLRSQLTEKAAEKGIKVIYPSMKLCTDNAAMIAEAAYYKL
KNAKNEKDCFAGLDLNGVASLMVSDEKAI

Align this sequence using NCBI's BLAST tool. Then describe the protein using GO terms. That is, list its GO annotations in each ontology.

References

AmiGO 2 version: 2.3.2 (amigo2b). 2016. http://amigo.geneontology.org (accessed February 5, 2016)

Bloss, Cinnamon S., Dilip V. Jeste, and Nicholas J. Schork. 2011. "Genomics for disease treatment and prevention." *Psychiatric Clinics of North America* 34, no. 1 (2011): 147-166

Carlson, M. *GO.db*: *A set of annotation maps describing the entire Gene Ontology*. R package version 3.2.2. 2016. https://www.bioconductor.org/packages/3.3/data/annotation/html/GO.db.html (accessed May 20, 2016)

Cannataro, Mario, Pietro H. Guzzi, and Marianna Milano. 2015. "GoD: An R-Package based on Ontologies for Prioritization of Genes with respect to Diseases," *J. of Computational Science*, Impact Factor: 1.23, DOI: 10.1016/j.jocs.2015.04.017

DOID (Disease Ontology). 2016. http://disease-ontology.org/ (accessed February 1, 2016)

GO (Gene Ontology). 2016. http://geneontology.org/ (accessed February 1, 2016)

The Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. *Nucl Acids Res* 43 Database issue D1049–D1056. Online at Nucleic Acids Research, http://nar.oxfordjournals.org/content/43/D1/D1049

HP (Human Phenotype Ontology). 2016. http://human-phenotype-ontology.github.io/about.html (accessed February 1, 2016)

infoplease. 2016. "Life Expectancy by Age, 1850–2011." http://www.infoplease.com/ipa/A0005140.html (accessed February 9, 2016)

NHGRI (National Human Genome Research Institute), NIH. 2010. "The Human Genome Project Completion: Frequently Asked Questions" https://www.genome.gov/11006943 (accessed February 9, 2016)

OMIM (Online Mendelian Inheritance in Man). 2016. NCBI. http://www.ncbi.nlm.nih.gov/omim (accessed February 1, 2016)

Sacristán, J. A. 2013. "Evidence based medicine and patient centered medicine: some thoughts on their integration." Revista Clínica Española (English Edition) 213, no. 9 (2013): 460-464

Wang, James Z., Zhidian Du, Rapeeporn Payattakool, Philip S. Yu and Chin-Fu Chen. 2007. "A new method to measure the semantic similarity of GO terms." *Bioinformatics*, Vol. 23 no. 10, pages 1274–1281 doi:10.1093/bioinformatics/btm087

Yu G., F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang. 2010. "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products." *Bioinformatics*, 26(7), pp. 976-978

Answers to Quick Review Questions

1.

- **a.** database
- **b.** Gene Ontology (GO)
- **c.** is_a, part_of, has_part, regulates

2.

- **a.** graph
- **b.** directed
- **c.** acyclic
- **d.** biological process
- **e.** “developmental process involved in reproduction” and “reproductive process”
- **f.** “anatomical structure development” and “developmental process”
- **g.** “reproductive process,” “developmental process,” “reproduction,” “biological process”
- **h.** “reproductive process” and “developmental process involved in reproduction”
- **i.** “developmental process” and “biological process”

3.

- **a.** 0
- **b.** 1.719 = -log(0.0190975)
- **c.** 1.944 = -log(0.0113732)
- **d.** 2.782 = -log(0.00165066)

4.

- **a.** 1.719 = *IC*(“developmental process”)
- **b.** more
- **c.** 0.7275 = 2 · 1.719 / (1.944 + 2.782)
- **d.** more

5.

a.

Term	**Calculation of S-Value**	**S-Value**
Figure 2a		
anatomical structure morphogenesis	1	1
anatomical structure development	0.6	0.6
developmental process	max(0.6*0.8, 0.8)	0.8
biological process	0.8*0.8	0.64

b.

Term	**Calculation of S-Value**	**S-Value**
Figure 2b		
developmental process involved in reproduction	1	1
reproductive process	max(0.6, 0.8)	0.8
developmental process	0.8	0.8
reproduction	max(0.6, 0.8*0.6)	0.6
biological process	max(0.6*0.8, 0.8*0.8, 0.8*0.8)	0.64

c. 3.04 = 1 + 0.6 + 0.8 + 0.64

d. 3.84 = 1 + 0.8 + 0.8 + 0.6 + 0.64

e. 0.418605 = ((0.8 + 0.8) + (0.64 + 0.64)) / (3.04 + 3.84)

f. more similar