CS 175: Introduction to Bioinformatics for Genomics and Transcriptomics

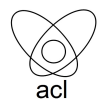
Lecture Slides

Annotations

Jan Michael C. Yap

Algorithms and Complexity Laboratory
Department of Computer Science
University of the Philippines, Diliman
janmichaelyap@gmail.com

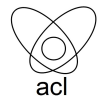
Lesson 7



Annotations

Ontologies

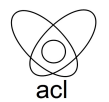
Annotations for Bioinformatics
- Gene Ontology
- Sequence Ontology
- GO Term Enrichment Analysis



Annotations

Ontologies



What is an Ontology?

"Ontology (the "science of being") is ... part of metaphysics that specifies the most fundamental categories of existence, the elementary substances or structures out of which the world is made. Ontology will thus analyze the most general and abstract concepts or distinctions that underlay every more specific description of any phenomenon in the world, e.g. time, space, matter, process, cause and effect, system."

---

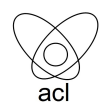
http://www.obitko.com/tutorials/ontologies-semantic-web/philosophical-roots.html

What is an Ontology?

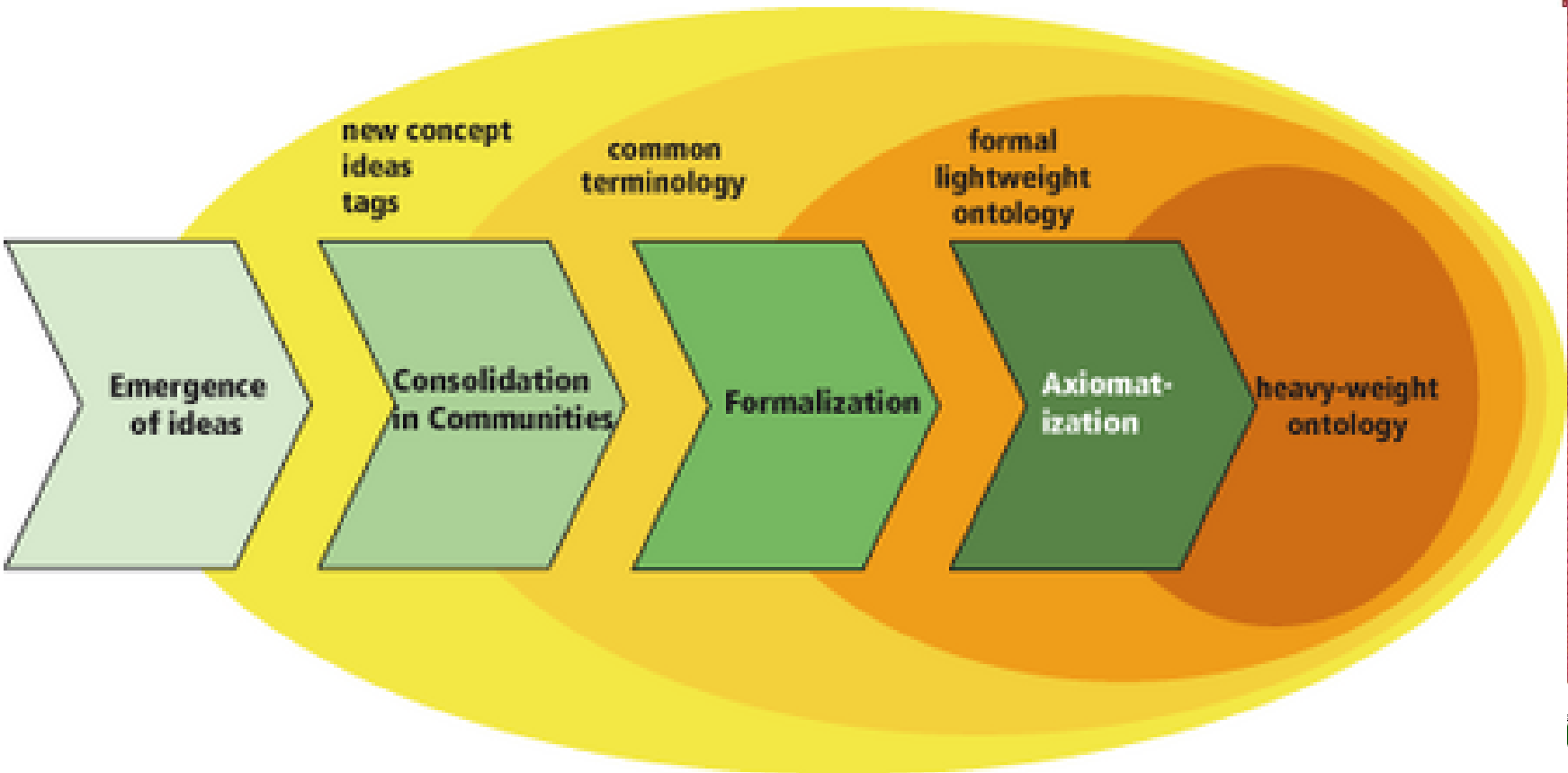
"Recently, the term of "ontology" has been up taken by researchers in Artificial Intelligence, who use it to designate the building blocks out of which models of the world are made. An agent (e.g., an autonomous robot) using a particular model will only be able to perceive that part of the world that his ontology is able to represent. In this sense, only the things in his ontology can exist for that agent. In that way, an ontology becomes the basic level of a knowledge representation scheme."

---

http://www.obitko.com/tutorials/ontologies-semantic-web/philosophical-roots.html

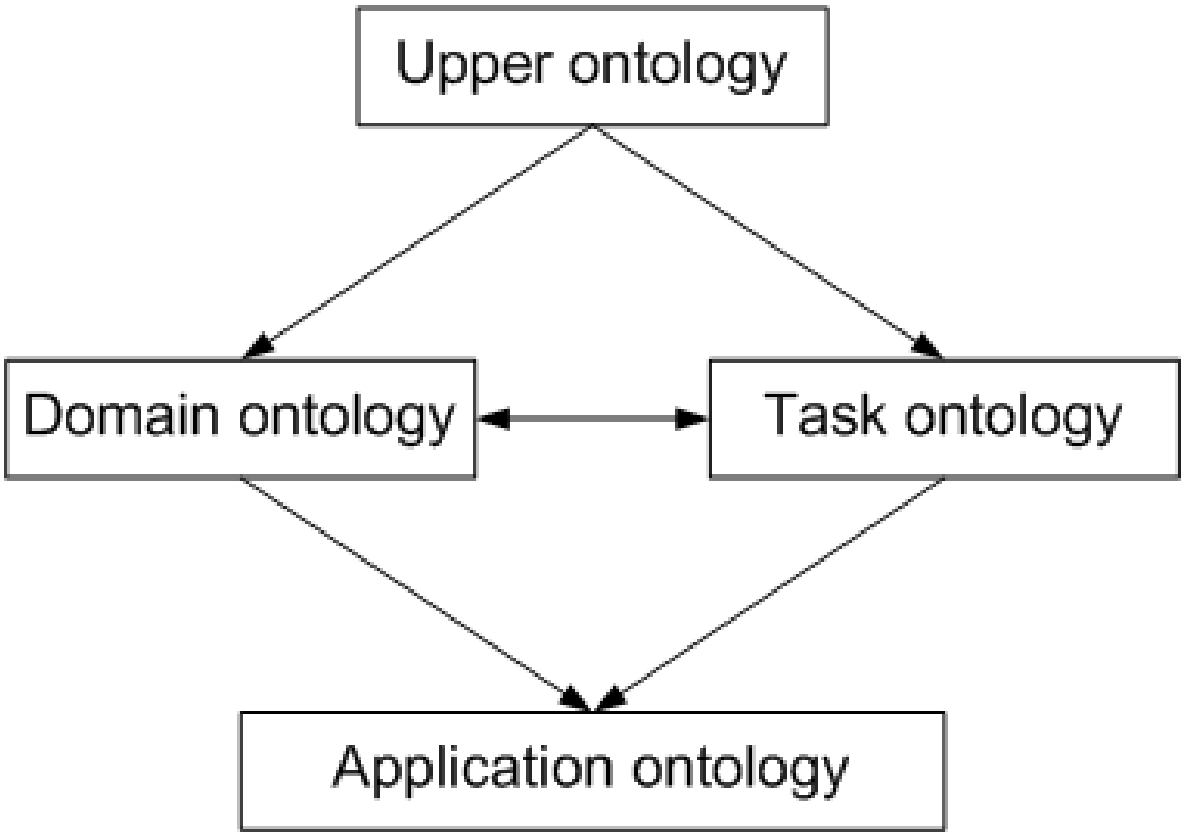


Ontology Development Process



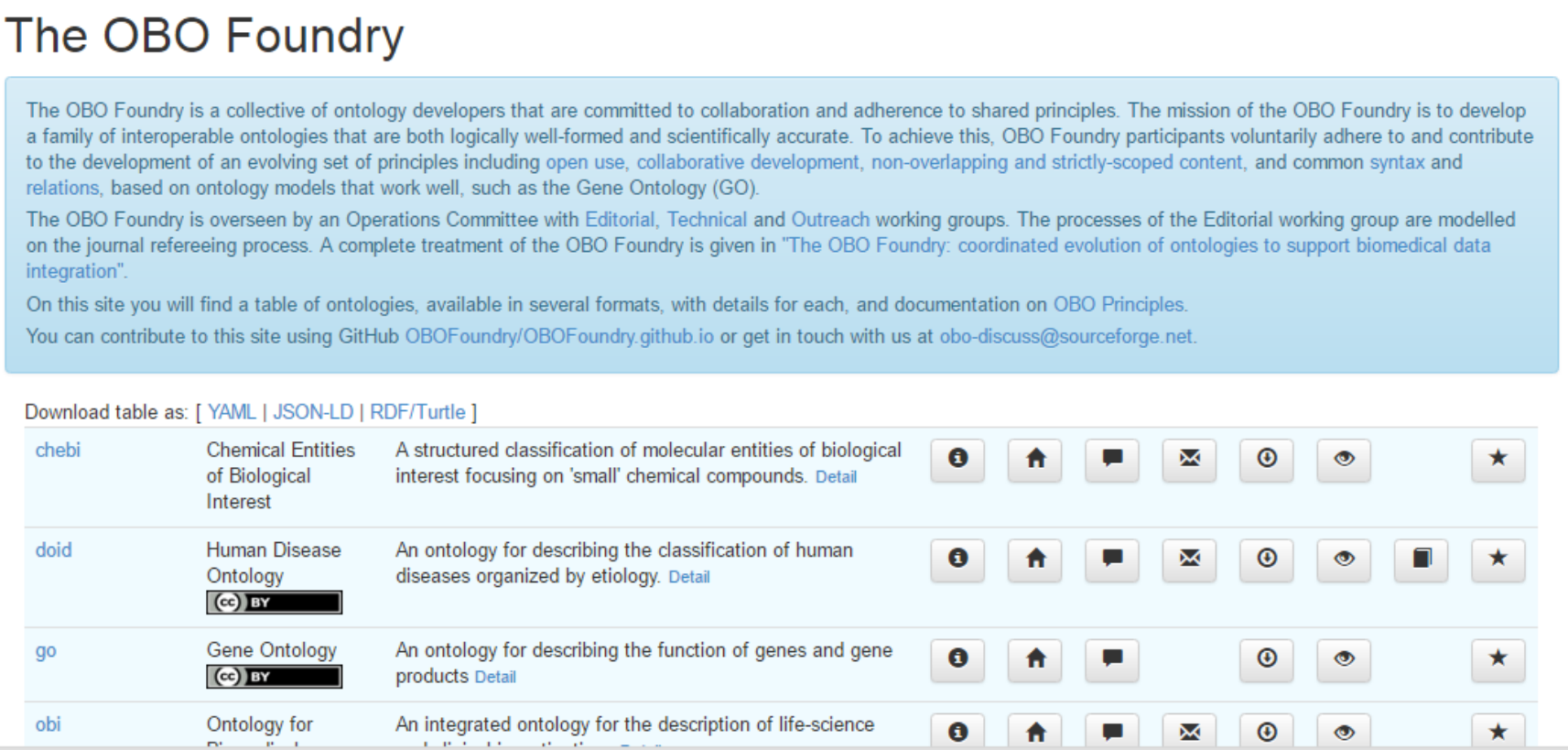
http://hlwiki.slais.ubc.ca/images/thumb/8/8d/Ontology.png/450px-Ontology.png

Ontology Development Process



http://www.obitko.com/tutorials/ontologies-semantic-web/modularization-of-ontologies.html

Open Biomedical Ontologies



The OBO Foundry

The OBO Foundry is a collective of ontology developers that are committed to collaboration and adherence to shared principles. The mission of the OBO Foundry is to develop a family of interoperable ontologies that are both logically well-formed and scientifically accurate. To achieve this, OBO Foundry participants voluntarily adhere to and contribute to the development of an evolving set of principles including open use, collaborative development, non-overlapping and strictly-scoped content, and common syntax and relations, based on ontology models that work well, such as the Gene Ontology (GO).

The OBO Foundry is overseen by an Operations Committee with Editorial, Technical and Outreach working groups. The processes of the Editorial working group are modelled on the journal refereeing process. A complete treatment of the OBO Foundry is given in "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration".

On this site you will find a table of ontologies, available in several formats, with details for each, and documentation on OBO Principles.

You can contribute to this site using GitHub OBOFoundry/OBOFoundry.github.io or get in touch with us at obo-discuss@sourceforge.net.

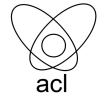
Download table as: [YAML | JSON-LD | RDF/Turtle]

chebi	Chemical Entities of Biological Interest	A structured classification of molecular entities of biological interest focusing on 'small' chemical compounds. Detail	
doid	Human Disease Ontology CC BY	An ontology for describing the classification of human diseases organized by etiology. Detail	
go	Gene Ontology CC BY	An ontology for describing the function of genes and gene products Detail	
obi	Ontology for	An integrated ontology for the description of life-science	

http://www.obofoundry.org/

Annotations

Annotations for Bioinformatics
- Gene Ontology
- Sequence Ontology
- GO Term Enrichment Analysis



About GO

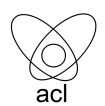
- "The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products across databases."
- "The use of GO terms by collaborating databases facilitates uniform queries across all of them. Controlled vocabularies are structured so they can be queried at different levels..."

http://geneontology.org/page/documentation

Sample GO Term

Term Information

Accession	GO:0052160
Name	modulation by symbiont of host systemic acquired resistance
Ontology	biological_process
Synonyms	modulation by symbiont of systemic acquired resistance in host
Definition	Any process in which an organism modulates the frequency, rate or extent of systemic acquired resistance in the host organism; systemic acquired resistance is response that confers broad spectrum systemic resistance. The host is defined as the larger of the organisms involved in a symbiotic interaction. *Source*: GOC:
Comment	None
History	See term history for GO:0052160 at QuickGO
Subset	gosubset_prok
Community	GN Add usage comments for this term on the GONUTS wiki.
Related	Link to all **genes and gene products** annotated to modulation by symbiont of host systemic acquired resistance.
	Link to all direct and indirect **annotations** to modulation by symbiont of host systemic acquired resistance.
	Link to all direct and indirect **annotations download** (limited to first 10,000) for modulation by symbiont of host systemic acquired resistance.
Feedback	Contact the GO Helpdesk if you find mistakes or have concerns about the data you find here.



http://amigo.geneontology.org/amigo/term/GO:0052160

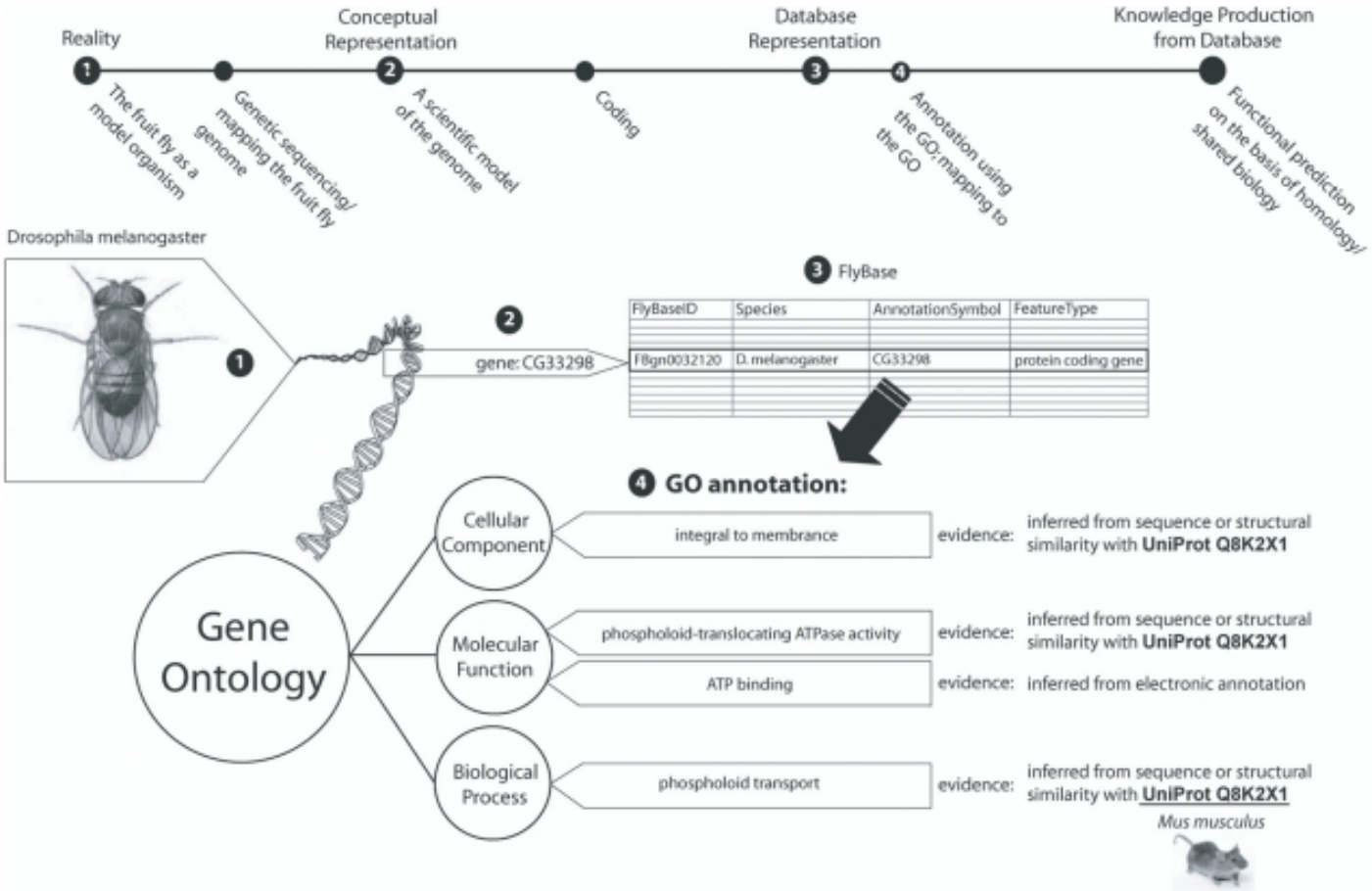
Sample GO Term

Ancestors of modulation by symbiont of host systemic acquired resistance (GO:0052160)

subject	relation	object
modulation by symbiont of host systemic acquired resistance	R regulates (inferred)	biological_process (GO:0008150)
modulation by symbiont of host systemic acquired resistance	R regulates (inferred)	response to stimulus (GO:0050896)
modulation by symbiont of host systemic acquired resistance	I is_a (inferred)	biological regulation (GO:0065007)
modulation by symbiont of host systemic acquired resistance	R regulates (inferred)	multi-organism process (GO:0051704)
modulation by symbiont of host systemic acquired resistance	R regulates (inferred)	response to biotic stimulus (GO:0009607)
modulation by symbiont of host systemic acquired resistance	R regulates (inferred)	response to external stimulus (GO:0009605)
modulation by symbiont of host systemic acquired resistance	R regulates (inferred)	biological attribute (OBA:0000001)
modulation by symbiont of host systemic acquired resistance	R regulates (inferred)	immune system process (GO:0002376)
modulation by symbiont of host systemic acquired resistance	P part_of (inferred)	interspecies interaction between organisms (GO:00441
modulation by symbiont of host systemic acquired resistance	I is_a (inferred)	regulation of biological process (GO:0050789)
modulation by symbiont of host systemic acquired resistance	R regulates (inferred)	response to external biotic stimulus (GO:0043207)
modulation by symbiont of host systemic acquired resistance	R regulates (inferred)	response to stress (GO:0006950)

http://amigo.geneontology.org/amigo/term/GO:0052160

Formalization and Annotation Process



GO Annotation Policies: http://geneontology.org/page/go-annotation-policies

---

Schuurman N and Leszczynski A (2008). Ontologies for Bioinformatics. Bioinformatics and Biology Insights 2: 187-200

Not Covered by GO

- Gene products e.g. cytochrome c is not in the ontologies, but attributes of cytochrome c, such as oxidoreductase activity, are.
- Processes, functions or components that are unique to mutants or diseases e.g. oncogenesis is not a valid GO term, as "causing cancer" is the result of reprogrammed, not normal cells and thus it is not the normal function of a gene.
- Attributes of sequence such as "intron" or "exon" parameters belong in a separate sequence ontology
- Protein domains or structural features
- Protein-protein interactions
- Environment, evolution and expression
- Anatomical or histological features above the level of cellular components, including cell types

http://geneontology.org/page/documentation

About SO

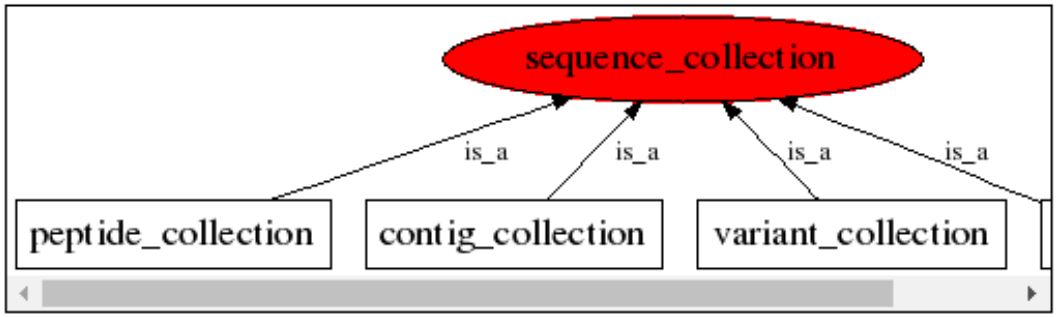
- "The Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequence."
- "SO includes different kinds of features which can be located on the sequence. Biological features are those which are defined by their disposition to be involved in a biological process. "

http://www.sequenceontology.org/index.html

Sample SO Term

sequence_collection (CURRENT_SVN)	
SO Accession:	SO:0001260 (SOWiki)
Definition:	A collection of discontinuous sequences.
Synonyms:	sequence collection
DB Xrefs:	SO: ke
Children:	peptide_collection (SO:0001501)
	contig_collection (SO:0001462)
	variant_collection (SO:0001507)
	genome (SO:0001026)

In the image below graph nodes link to the appropriate terms. Clicking the image background will toggle the image between large and small formats.



http://www.sequenceontology.org/browser/current_svn/term/SO:0001260

Annotation using Generic Feature Format version 3 (GFF3)

1	*seqid*	The landmark to which the coordinates are given.
2	*source*	The procedure that produced the feature. For example, the name of a piece of software or another database may be appropriate. Not all features have a source.
3	*type*	The type of feature using either a term name or accession number from the Sequence Ontology.
4	*begin*	The *begin* coordinate of the feature relative to the landmark given in column 2 where 1-based integer coordinates are used.
5	*end*	The *end* coordinate of the feature relative to the landmark given in column 1 where 1-based integer coordinates are used.
6	*score*	The score attributed to the feature if required.
7	*strand*	The direction of the annotation.
8	*phase*	The phase of the feature. Not all features have a phase.
9	*attributes*	The attributes of the feature are recorded as tag-value pairs and multiple attributes are separated by semi-colons. Lower case tags are unrestricted, but upper case tags are reserved for special meanings. There are several tags with predefined meanings: ID is the identifier for the feature and the value of this tag must be unique within the document. Name is the tag used for display purposes for the feature so it does not have to be unique. Another commonly used reserved tag is Parent, which is used to capture **part_of** relations. The value of this tag is the ID of the 'parent'.

Eilbeck K and Lewis SE (2004). Sequence Ontology Annotation Guide. Comparative and Functional Genomics 5(8): 642-647

Annotation using Generic Feature Format version 3 (GFF3)

##GFF-version3

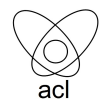
##sequence region 2L 19486843–19480420

##seqid	source	type	begin	end	score	strand	phase	attribute
2L	.	gene	19480420	19486843	.	–	.	ID=0001; Name=CG10188; has_genome_location =nuclear_gene;
2L	.	mRNA	19480420	19486843	.	–	.	ID=0002; Name=CG10188-RA; Parent=0001;
2L	.	five_prime_UTR	19486435	19486843	.	–	.	ID=0003; Parent=0002;
2L	.	five_prime_UTR	19486270	19486348	.	–	.	ID=0003; Parent=0002;
2L	.	CDS	19485573	19486269	.	–	.	ID=0004; Name=CG10188-cdsA; Parent=0002;
2L	.	CDS	19481212	19484444	.	–	.	ID=0004; Name=CG10188-cdsA; Parent=0002;
2L	.	three_prime_UTR	19480420	19481213	.	–	.	ID=0005; Parent=0002;
2L	.	mRNA	19480420	19486843	.	–	.	ID=0006; Name=CG10188-RB; Parent=0001;
2L	.	exon	19485573	19486843	.	–	.	ID=0007; Name=CG10188-1; Parent=0006;
2L	.	exon	19480420	19484444	.	–	.	ID=0008; Name=CG10188-2; Parent=0006;
2L	.	five_prime_UTR	19486270	19486843	.	–	.	ID=0009; Parent=0006;
2L	.	three_prime_UTR	19480420	19481213	.	–	.	ID=0005; Parent=0006;
2L	.	transposable_element	19484822	19485356	.	–	.	ID=0010; Name=Cr1a{}412-RA;

Eilbeck K and Lewis SE (2004). Sequence Ontology Annotation Guide. Comparative and Functional Genomics 5(8): 642-647

Motivation for GO Term Enrichment Analysis

- Suppose that you have a set of (annotated) genes that are significantly up-regulated in a differential expression test.
- We can make use of the annotations to qualitatively assess their biological significance.
- "[A]n enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations for that gene set."



http://geneontology.org/page/go-enrichment-analysis

GO Term Enrichment Analysis: An Illustration

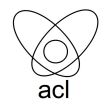
	Total	Annotated with "DNA Repair"
Entire Set	6,442	100
Input Set	10	5

- Background percentage: $100/6,442 \sim 0.016$
- Sample percentage: $5/10 = 0.5$
- "DNA Repair" is overrepresented ($\chi^2_{df=1}$, $\alpha = 0.05$) in the input set

Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Protocol 2013; 8(8): 1551-1566.

Further Notes

- Other distributions/tests:
  - Binomial distribution
  - Hypergeometric distribution
  - Fisher's exact test
- Since some GO terms are related, family-wise error rate control such as Bonferroni correction is used.
  - More "general" terms are given more weight



Tipney H and Hunter L (2010). An introduction to effective use of enrichment analysis software. Human Genomics 4(3): 202-206

Questions?

See you next meeting!