The Gene Ontology

Slides obtained from the following presentations:

- *David Hill* (MGI), Harvard University, USA *What Is Ontology?*
- *Judith Blake* (MGI), Harvard University, USA *Gene Ontology Overview and Perspective*

The original slides are downloadable from: www.geneontology.org

What is Ontology?

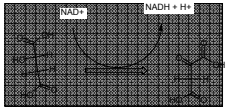
1606

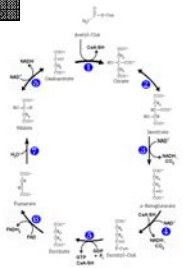
1700s

- Dictionary:A branch of metaphysics concerned with the nature and relations of being.
- Barry Smith:The science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality.

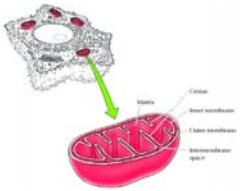
The Gene Ontology (GO) is actually three Ontologies

Molecular Function
GO term: Malate dehydrogenase.
GO id: GO:0030060
(S)-malate + NAD(+) = oxaloacetate + NADH.





Biological Process
GO term: tricarboxylic acid cycle
Synonym: Krebs cycle
Synonym: citric acid cycle
GO id: GO:0006099

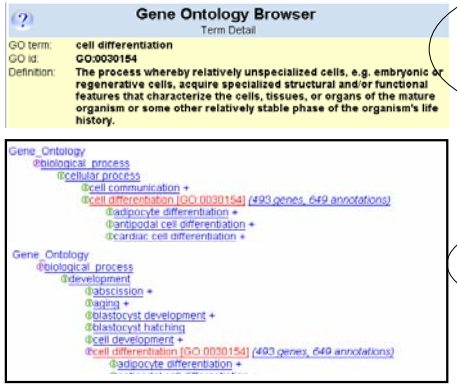


Cellular Component
GO term: mitochondrion
GO id: GO:0005739
Definition: A semiautonomous, self replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration.

Seven Healthy Habits of Highly Effective Ontology Construction

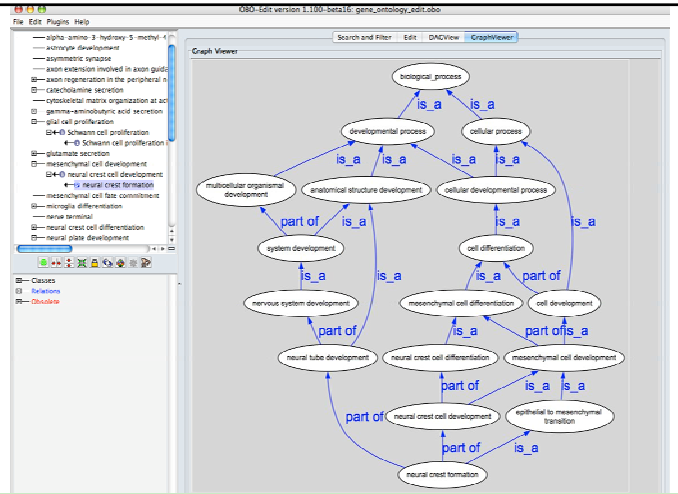
- Univocity
- Positivity
- Objectivity
- Single Inheritance
- Create Good Definitions
- Distinguish Between Types & Instances
- Basis in Reality

GO Definitions: Each GO term has 2 Definitions



A definition written by a biologist: *necessary & sufficient conditions* ***written definition*** (not computable)

Graph structure: *necessary conditions* **formal** (computable)



Terms are defined graphically relative to other terms

Appropriate Relationships to Parents

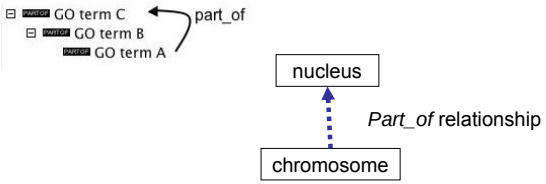
- GO currently has 2 relationship types
  - Is_a
    - An is_a child of a parent means that the child is a complete type of its parent, but can be discriminated in some way from other children of the parent.
  - Part_of
    - A part_of child of a parent means that the child is always a constituent of the parent that in combination with other constituents of the parent make up the parent.

Placement in the Graph: Selecting Parents

- To make the most precise definitions, new terms should be placed as children of the parent that is closest in meaning to the term.
- To make the most complete definitions, terms should have all of the parents that are appropriate.
- In an ontology as complicated as the GO this is not as easy as it seems.

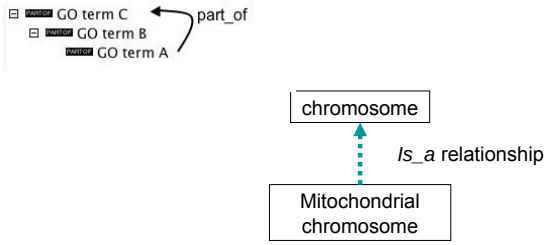
True Path Violations Create Incorrect Definitions

.."the pathway from a child term all the way up to its top-level parent(s) must always be true".



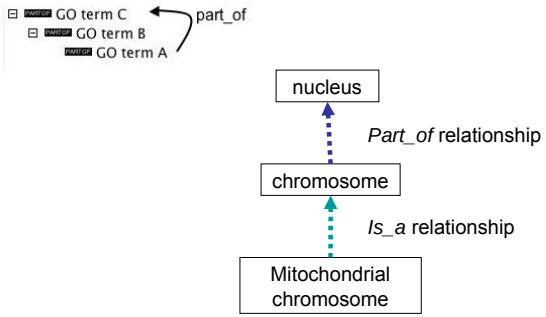
True Path Violations

.."the pathway from a child term all the way up to its top-level parent(s) must always be true".



True Path Violations

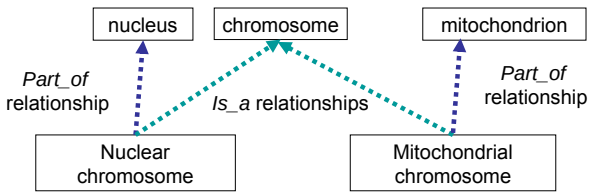
. "the pathway from a child term all the way up to its top-level parent(s) must always be true".



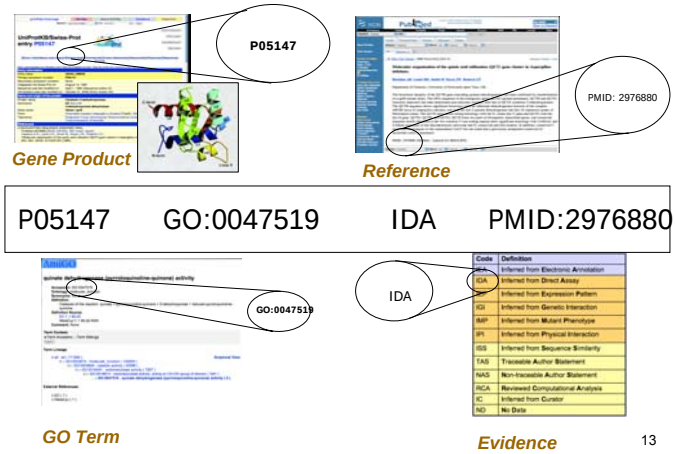
A mitochondrial chromosome is not part of a nucleus!

True Path Violations

.."the pathway from a child term all the way up to its top-level parent(s) must always be true".



Annotating Gene Products using GO



P05147 GO:0047519 IDA PMID:2976880

13

Annotations are assertions

- There is *evidence* that this gene product can be best classified using this term
- The *source* of the evidence and other information is included
- *There is agreement on the meaning of the term*

14

Annotations are assertions

Biological Process	axon cargo transport	IMP	MGI:2136847	J:73257
Biological Process	NOT axon cargo transport	IMP	MGI:2136847	J:98609
Biological Process	axon midline choice point recognition	IMP	MGI:2154522	J:21937
Biological Process	axonogenesis	IMP	MGI:2136847	J:101975
Biological Process				
Biological Process				
Biological Process				
Biological Process				
Biological Process	dendrite development	IGI	MGI:88047	J:98108
Biological Process	dendrite development	IMP	MGI:2136847	J:101975, J:53024
Biological Process	endocytosis	IEA	SP_KW:KW-0254	J:60000
Biological Process	extracellular matrix organization and biogenesis	IGI	MGI:2154545|MGI:2137246	J:93306
Biological Process	forebrain development	IMP	MGI:2154522	J:21937
Biological Process	forebrain development	IMP	MGI:2154522|MGI:2154535	J:54584
Biological Process	G2 phase of mitotic cell cycle	IMP	MGI:2154535	J:95286

Annotations are the connections between genomic information and the GO.
Experiments provide the data that enables us to annotate gene products with terms from the ontologies.

Annotations for App**: amyloid beta (A4) precursor protein**

15

We use evidence codes to describe the basis of the annotation

- IDA: Inferred from direct assay
- IPI: Inferred from physical interaction
- IMP: Inferred from mutant phenotype
- IGI: Inferred from genetic interaction

Direct Experiment in organism

- IEP: Inferred from expression pattern
- IEA: Inferred from electronic annotation
- ISS: Inferred from sequence or structural similarity
- TAS: Traceable author statement
- NAS: Non-traceable author statement
- IC: Inferred by curator

NO Direct Experiment
Inferred from evidence

- RCA: Reviewed Computational Analysis
- ND: no data available

16

GO Annotation Stats:

GO Annotations

Total **manual** GO annotations - 388,633

Total proteins with manual annotations – 80,402

Contributing Groups (including MGI): - 19

Total Pub Med References – 346,002

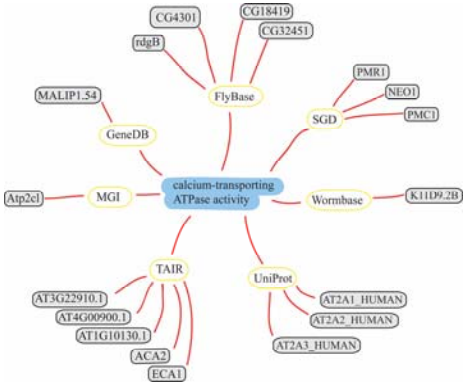
Total number **predicted** annotations – 17,029,553

Total number taxa – 129,318

Total number distinct proteins – 2,971,374

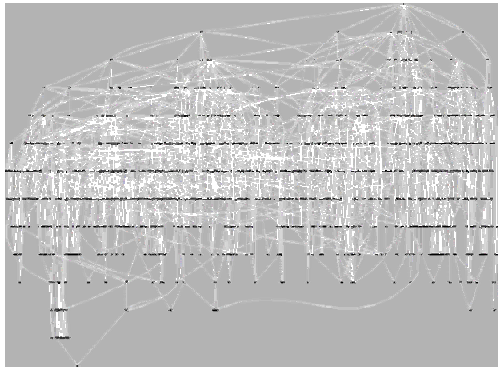
April 24, 2007

Annotations of gene products to GO are genome specific



Now we can query across all annotations based on shared biological activity.

S: cerevisiae: GO DAG of the BP ontology

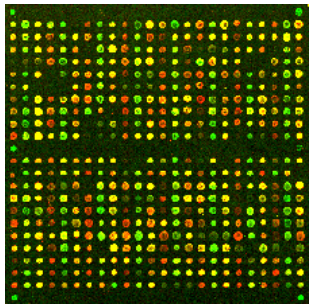


1074 GO classes (nodes) connected by 1804 edges (TAS evidence only)

GO is a functional annotation system of great utility in computational biology

GO enables genomic data analysis

- Microarrays allow biologists to record changes in gene function across entire genomes
- Result: *Vast* amounts of gene expression data desperately needing cataloging and tagging
- Many data analysis tools use GO graph structure to statistically evaluate clusters of co-expressed genes based on shared functional annotations
  - 680 pub (of 1517) on GO list
  - 46 microarray tools contributed



GO supports functional classifications

Table 3. Functional classification of CAN genes, with CaMP score to the right of each gene name. CAN genes were assigned to functional classes using Gene Ontology (GO) groups, INTERPRO domains, and available literature. Representative GO groups and INTERPRO domains are listed for each class.

Breast cancers						Colorectal cancers			
Cellular adhesion and motility (examples: cytoskeletal protein binding GO:0008092, cell adhesion GO:0007155, metallopeptidase activity GO:0008237)									
FDNC	3.4	*TMPRSS6*	2.0	*RAPH1*	1.4	*PKHD1*	3.5	*CNTN4*	1.6
MYH1	2.7	*COL11A1*	1.6	*PCDHB15*	1.4	*ADAMTSL3*	3.3	*CKL1*	1.3
SPTAN1	2.6	*DNAH9*	1.7	*CMYA1*	1.4	*OBSCN*	3.0	*HAPLN1*	1.2
DBN1	2.5	*OBSCN*	1.7	*MAGF1*	1.3	*ADAMTS18*	2.7	*MGC33407*	1.2
TECTA	2.4	*COL7A1*	1.5	*SYNE2*	1.3	*MMP2*	2.3	*MAP2*	1.0
ADAM12	2.3	*MAGEE1*	1.5	*NRCAM*	1.1	*TTLL3*	2.2		
GSN	2.2	*CDH10*	1.5	*COL19A1*	1.1	*EVL*	2.0		
CDH20	2.2	*SULF2*	1.5	*SEMA5B*	1.1	*ADAM29*	2.0		
BGN	2.1	*CNTN6*	1.4	*ITGA9*	1.1	*CSMD3*	1.9		
ICAM5	2.1	*THBS3*	1.4			*ADAMTS15*	1.8		
Signal transduction (examples: intracellular signaling cascade GO:0007242, receptor activity GO:0004872, GTPase regulator GO:0030695)									
VEPH1	2.1	*PFC*	1.5	*PRPF4B*	1.3	*APC*	>10	*PTPRD*	2.2
SBNO1	2.1	*GAB1*	1.5	*CENTG1*	1.3	*KRAS*	>10	*MCP*	2.1

Cancer Genome Projects

nature Vol 445|11 January 2007|doi:10.1038/nature05453

ARTICLES

Nature: January 2007

Genome-wide atlas of gene expression in the adult mouse brain

Ed S. Lein[1*], Michael J. Hawrylycz[1*], Nancy Ao[2], Mikael Ayres[1], Amy Bensinger[1], Amy Bernard[1], Andrew F. Boe[1], Mark S. Boguski[1†], Kevin S. Brockway[1], Emi J. Byrnes[1], Lin Chen[1], Li Chen[2], Tsuey-Ming Chen[2], Mei Chi Chin[1], Jimmy Chong[1], Brian E. Crook[1], Aneta Czaplinska[2], Chinh N. Dang[1], Suvro Datta[1], Nick R. Dee[1], Aimee L. Desaki[1], Tsega Desta[1], Ellen Diep[1], Tim A. Dolbeare[1], Matthew J. Donelan[1], Hong-Wei Dong[1], Jennifer G. Dougherty[1], Ben J. Dunc
Shanna R. F
Matthew R.
Reena Kaw

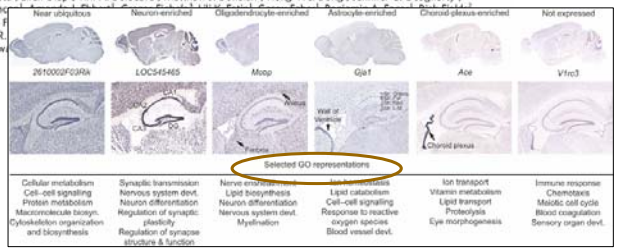
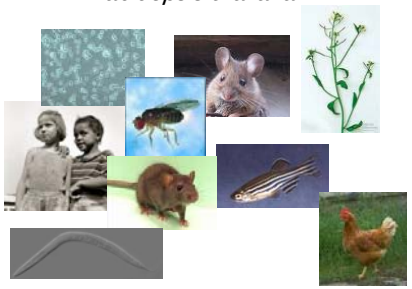


FIGURE 3. Representative cell-type-specific genes and corresponding molecular functions.

Comprehensively annotate Reference Genomes

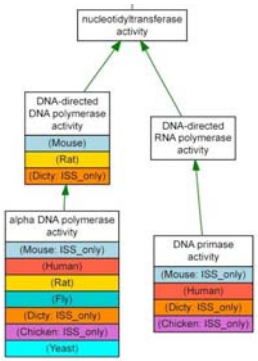
- Human
- Mouse
- Fly
- Rat
- Chicken
- Zebrafish
- Worm
- Dicty
- *E.coli*

- *Saccharomyces cerevisiae*
- *Schizosaccharomyces pombe*
- *Arabidopsis thaliana*



Reference Genome Annotation Project

- Priority genes: those implicated in human diseases
- Determine orthologs/homologs in reference genomes
- For these genes, comprehensively curate biomedical literature



Mary Dolan

25

Reference Genome Development Projects

- Shared annotation focus = Coordinated attention to ontology structure
- Orthology/homology set across primary model organisms
- Reference ID mappings including associations of sequences, gene/proteins, and human diseases
- Ultimately, transparent access to comprehensive information about genes among the primary data providers

26

Gene Ontology
www.geneontology.org

Mouse Genome Informatics
www.informatics.jax.org

GO Consortium is supported by NIH-NHGRI and by the European Union RTD Programme

MGI projects are supported by NIH [NHGRI, NICH, and NCI].

PRO is supported by NIGMS

Corpora is supported by NLM

27