Gene Ontology

--Notes to PRO Annual Meeting--

Judith Blake

April 26, 2010

Protein Ontology Meeting

The English Language is hard to learn, even for computers.

Focus: creating the data structures and mining the biomedical literature to provide knowledge representations –

with the objective of using logical reasoning applications and predictive approaches to 'interrogate' very large data sets, generating new hypothesis for further experimental investigation

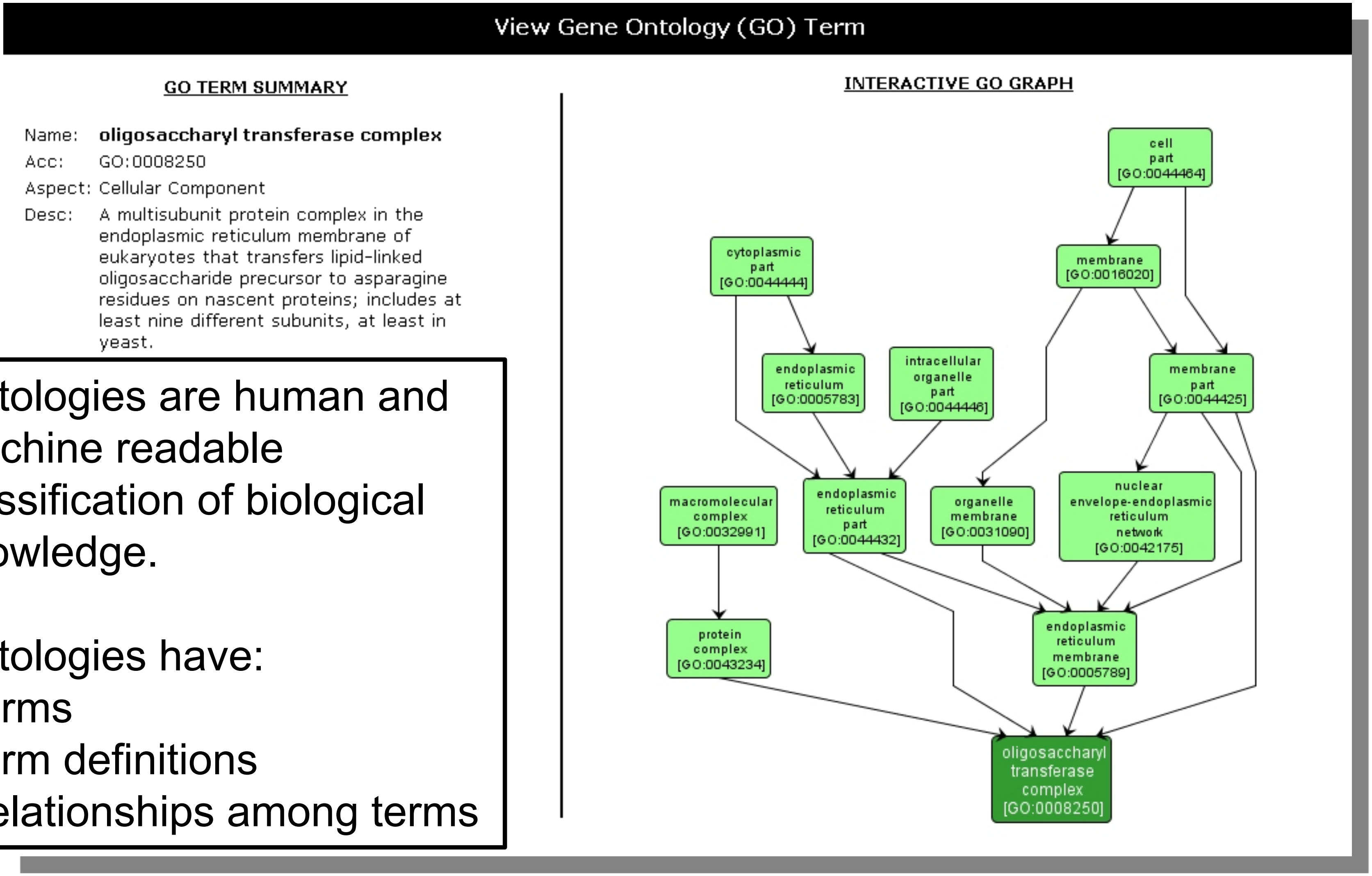
From Evidence to Inference

- Biologists often start investigations ‘species-blind’ and assemble their **initial** picture of a gene and its function without regard to taxonomic origin of gene in a particular experiment

- **“Tell me everything about this gene”**

- How can we coordinate the knowledge from large-scale genomic analysis and functional annotation for gene products for different organisms to help us understand biological systems?

Biomedical Ontologies



View Gene Ontology (GO) Term

GO TERM SUMMARY

Name: **oligosaccharyl transferase complex**
Acc: GO:0008250
Aspect: Cellular Component
Desc: A multisubunit protein complex in the endoplasmic reticulum membrane of eukaryotes that transfers lipid-linked oligosaccharide precursor to asparagine residues on nascent proteins; includes at least nine different subunits, at least in yeast.

INTERACTIVE GO GRAPH

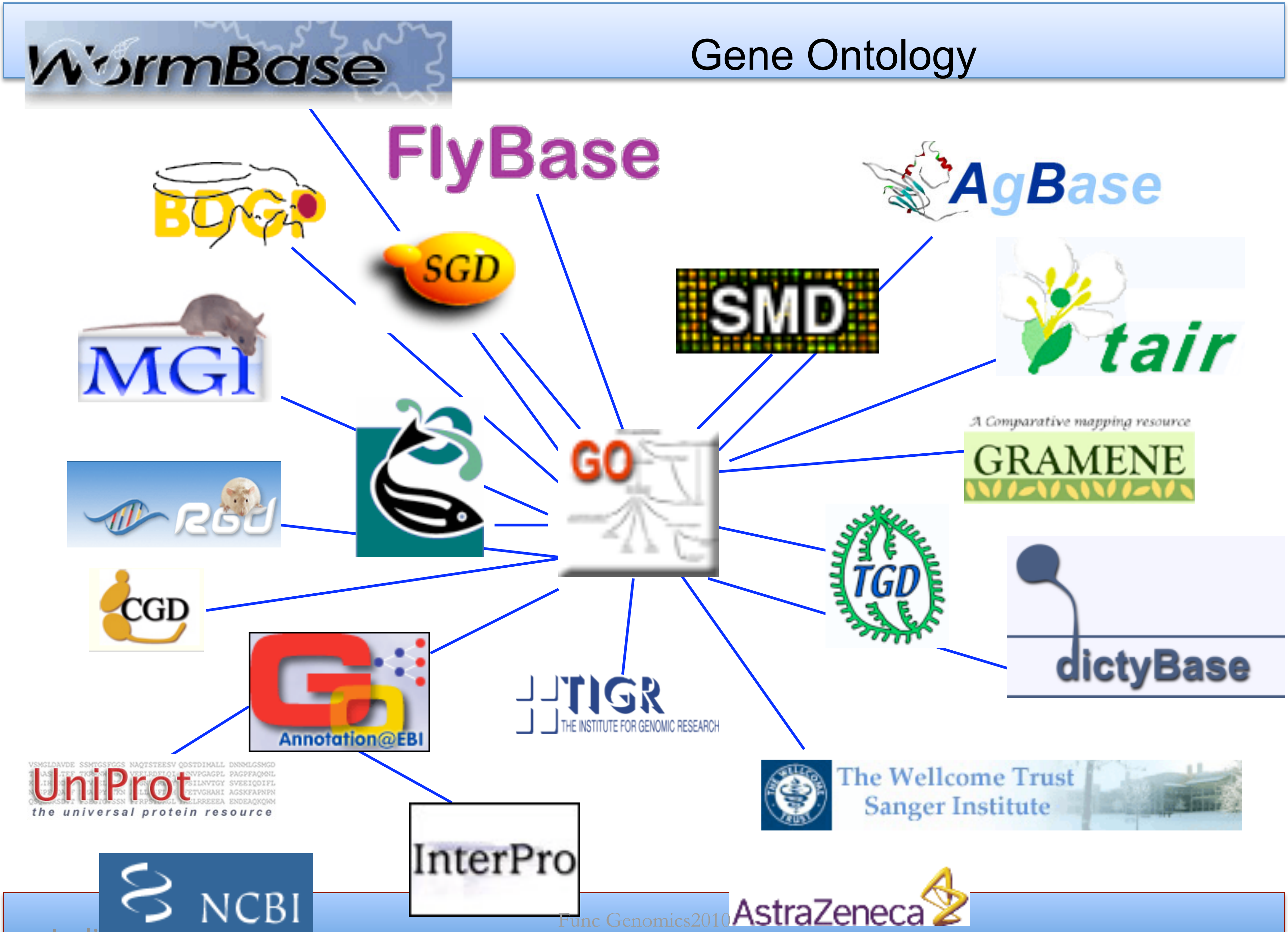
cell part [GO:0044464]
cytoplasmic part [GO:0044444]
membrane [GO:0016020]
endoplasmic reticulum [GO:0005783]
intracellular organelle part [GO:0044446]
membrane part [GO:0044425]
macromolecular complex [GO:0032991]
endoplasmic reticulum part [GO:0044432]
organelle membrane [GO:0031090]
nuclear envelope-endoplasmic reticulum network [GO:0042175]
protein complex [GO:0043234]
endoplasmic reticulum membrane [GO:0005789]
oligosaccharyl transferase complex [GO:0008250]

Ontologies are human and machine readable classification of biological knowledge.

Ontologies have:
- Terms
- Term definitions
- Relationships among terms

Annotation of genes and proteins using **ontologies** are key to data integration

Gene Ontology


WormBase
FlyBase
AgBase
BDGP
SGD
SMD
tair
MGI
A Comparative mapping resource
GRAMENE
RGD
GO
TGD
CGD
dictyBase
GO Annotation@EBI
TIGR
THE INSTITUTE FOR GENOMIC RESEARCH
UniProt
the universal protein resource
The Wellcome Trust Sanger Institute
InterPro
NCBI
AstraZeneca

The Gene Ontology

- Mid-size
  - ~18,000 terms in all 3 (4, 5) ontologies
    - Molecular function, biological process, cellular component, sequence ontology, cell ontology
  - ~2n,nnn links (is_a, part_of, *has_part, regulation of*)
- Each term represents a type
  - Terms also have alternate labels (synonyms)
    - These synonyms may not represent distinct types
    - Humans use different labels to refer to the same biological pattern
      - E.g: endoplasmic reticulum vs ER

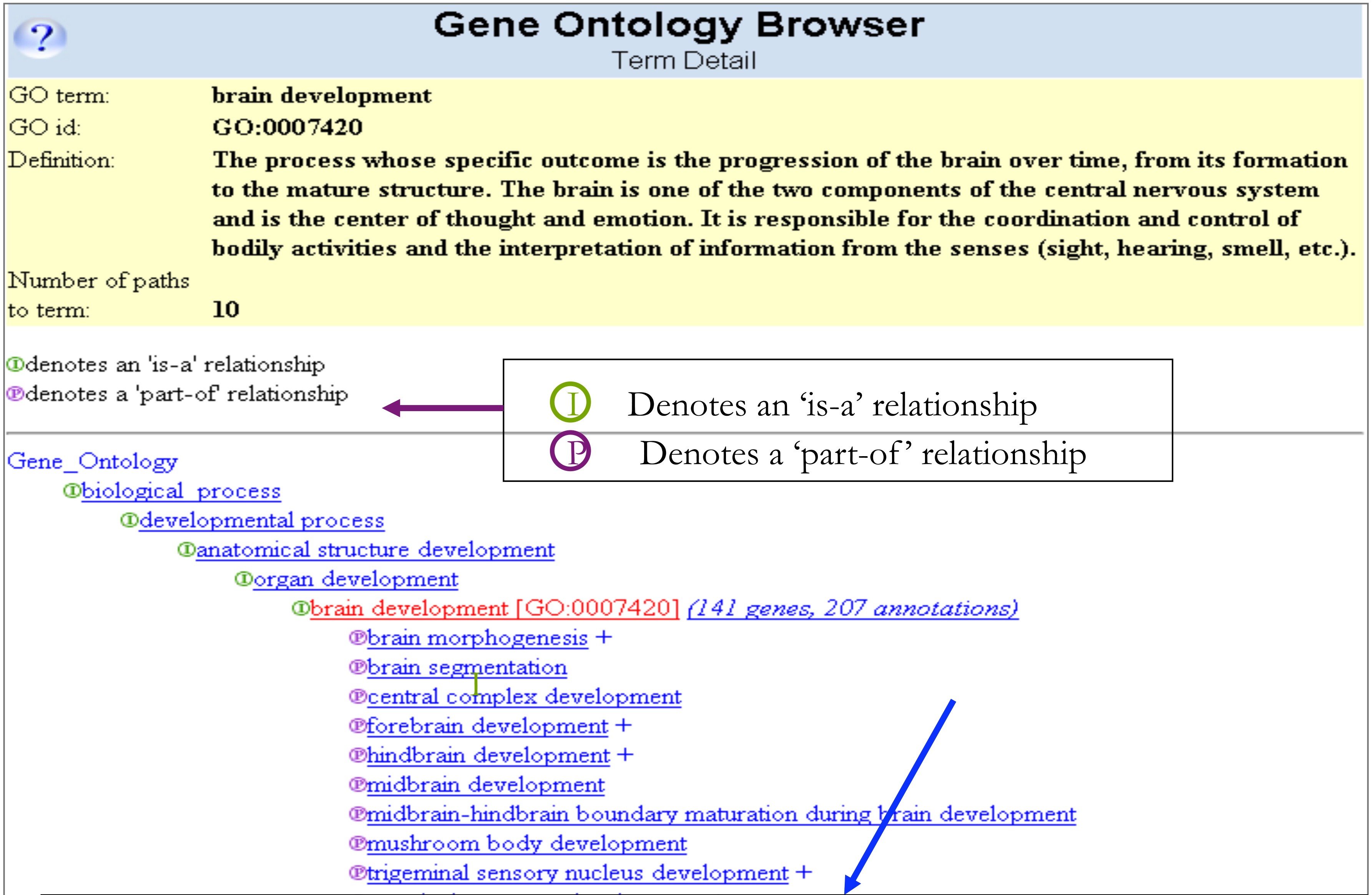
General notes: GO grant – units

- MOD curators/engineers supported
  - SGD, MGI, Dicty, TAIR, WormBase, RGD, EBI-GOA
  - Stanford/Berkeley (Jax/EBI) engineering

- Many major contributors not supported
  - Val Wood (*S. pombe*)
  - Michelle Gwinn Giglio (microbial)
  - Doug Howe (*Danio*)
  - Paul Thomas (Panther)
  - Jim Hu (*E. coli*)
  - Fiona McCarthy (AgBase)
  - Peter D'Eustachio (Reactome)

- New contributors (not supported) coming on
  - Swiss-Prot-Geneva (30+ curators trained by GOA)

GO terms are used for functional annotations



Gene Ontology Browser

Term Detail

GO term:	**brain development**
GO id:	**GO:0007420**
Definition:	**The process whose specific outcome is the progression of the brain over time, from its formation to the mature structure. The brain is one of the two components of the central nervous system and is the center of thought and emotion. It is responsible for the coordination and control of bodily activities and the interpretation of information from the senses (sight, hearing, smell, etc.).**
Number of paths to term:	**10**

Ⓘdenotes an 'is-a' relationship
Ⓟdenotes a 'part-of' relationship

Gene_Ontology
- Ⓘbiological_process
  - Ⓘdevelopmental process
    - Ⓘanatomical structure development
      - Ⓘorgan development
        - Ⓘbrain development [GO:0007420] *(141 genes, 207 annotations)*
          - Ⓟbrain morphogenesis +
          - Ⓟbrain segmentation
          - Ⓟcentral complex development
          - Ⓟforebrain development +
          - Ⓟhindbrain development +
          - Ⓟmidbrain development
          - Ⓟmidbrain-hindbrain boundary maturation during brain development
          - Ⓟmushroom body development
          - Ⓟtrigeminal sensory nucleus development +

Ⓘ Denotes an 'is-a' relationship
Ⓟ Denotes a 'part-of' relationship

Ⓘ Brain development [GO:0007420] (141 genes, 207 annotations)

GOC – comprehensive and accurate ontologies

- Ontology Development
  - By Process; by Meeting; by Request
- OBO_Edit and Reasoner
- Edges between MF and BP
- New relationships
  - Regulates (up and down)
  - Has_Part
- Cross-Products
  - Internal, ChEBI, CL, PRO, Uberon

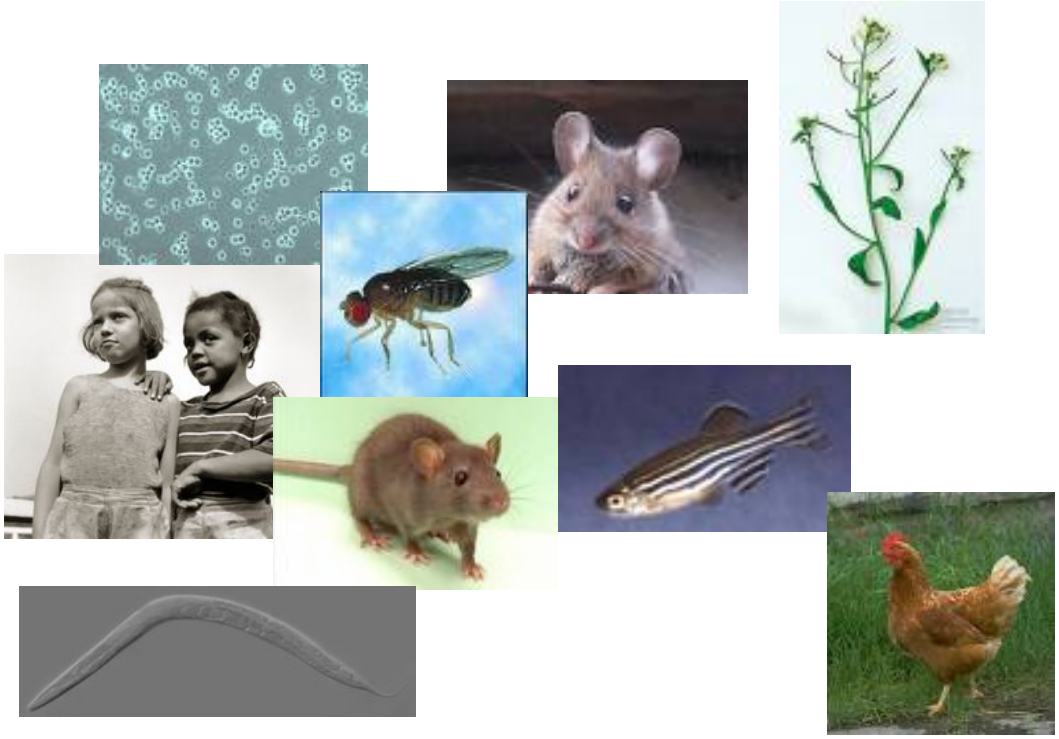
Ontologies and annotation

- Ontologies are of little practical use without *annotation*
  - GO has ~6 million annotations linking genes and gene products to GO terms
  - Mostly (but not all) MOD & Human
  - Same terms are shared across species
- All annotation statements have provenance
  - Source/publication
  - Evidence & evidence *codes*

Reference Genomes project

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

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



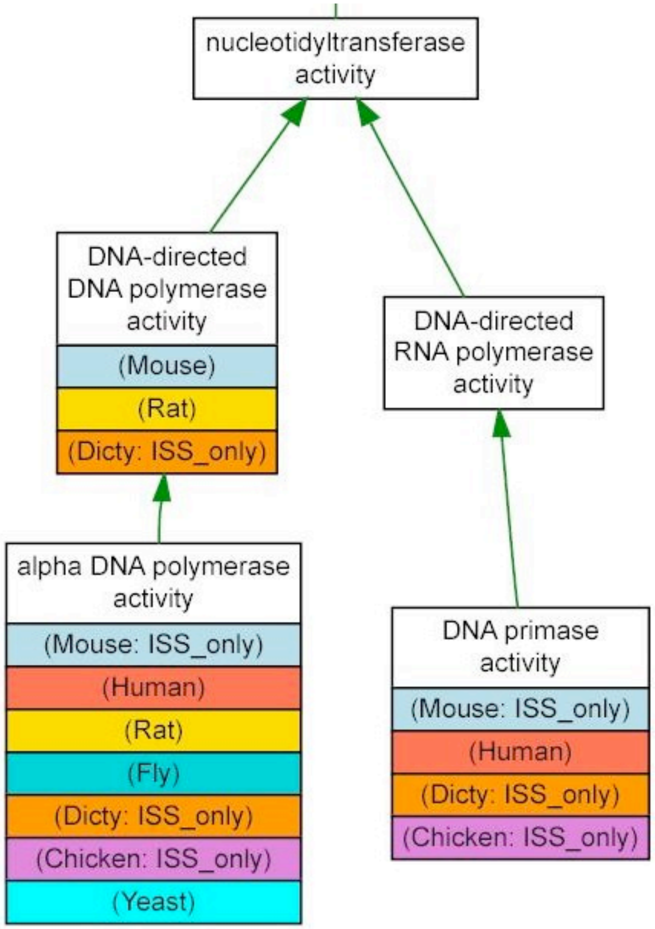
Reference Genome Annotation Project

- Select priority gene set
  1. Implicated in human diseases
  2. Current 'hot' genes
  3. Biochemical or signaling pathways
  4. Highly conserved across multiple species

- For these genes, comprehensively curate biomedical literature

- Determine orthologs (homologies) in other reference genomes using evolutionarily-based methods

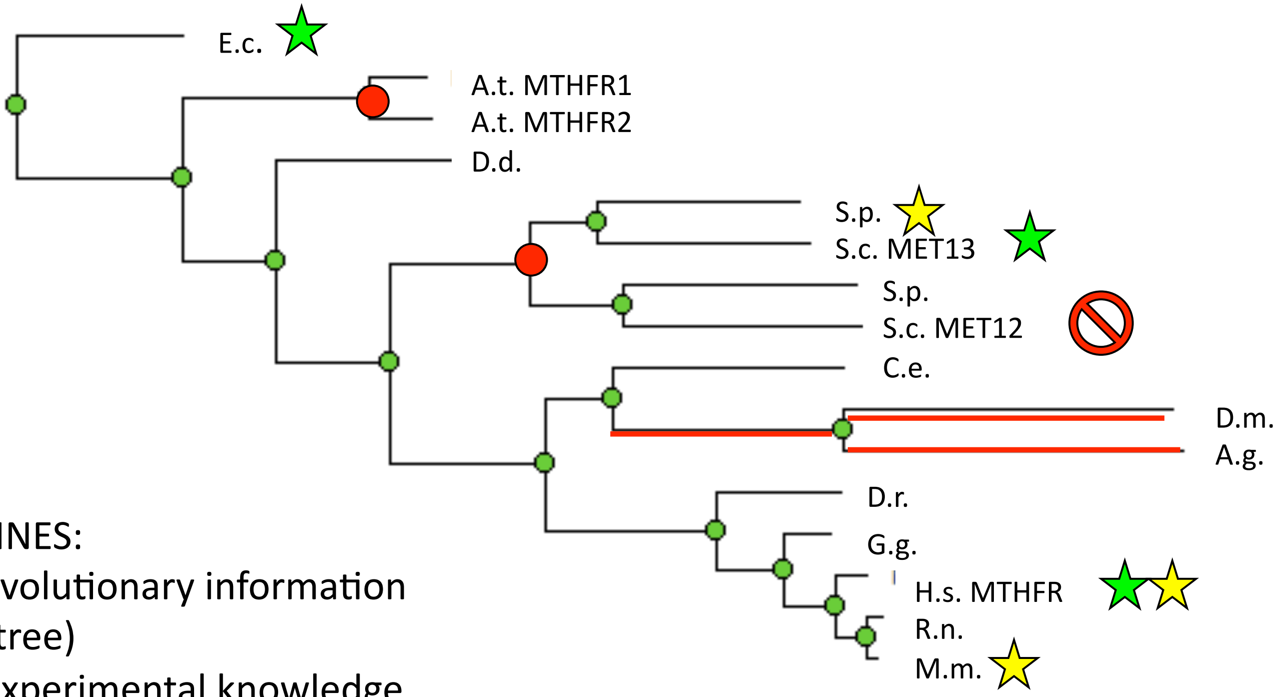
- Infer annotations via experimental literature in closely related group



GOC - comprehensive annotation of select genomes

- 12 Reference Genomes
  - [*PLoS Comput Biol. 2009 ; 5: e1000431*]
- Protein sets
  - Quest for Orthologs 51 species (going to 82)
  - [*Genome Biology* 2009, **10:**403]
- Generation of Family Sets
  - Panther/Princeton
  - 7,000 families
- Annotation Control
  - Inference by phylogenetic analysis
  - 400 families curated (7,000 genes)
- Integration into MODs
  - Test case of 10 families
- Ready to scale

Function inference from evolutionary context



COMBINES:

1. Evolutionary information (tree)
2. Experimental knowledge (GO annotations from literature)
3. Organism-specific biological knowledge (curators)

● = ‘gene’ duplication

★ = experimental annotation

☆ = inferential annotation

Uses of GO in studies of

- role of regulation of gene expression in axon guidance during development in Drosophila (PMID 17672901)
- prevention of ischemic damage to the retina in rats (PMID 17653046)
- immune system involvement in abdominal aortic aneurisms in humans (PMID 17634102)
- how the white spot syndrome virus affects cell function in shrimp (PMID 17506900)
- relationships between protein interaction networks involving the ash1 and ash2 genes in flies and in humans (PMID 17466076)

GO Bibliography: 3202 papers
Google Scholar: 6660 papers 'tool for the unification of biology'

Intersections for GO and PRO and others

- Defining classes of complexes
  - GO, PRO, IntACT
- Aligning evolutionary constructs
  - ProEvo, PantherSets, Pfam, PIRSF and more
- Tracking source data
  - Assays, Resources, Publications
- Recognizing Authorities- who for what
  - Again, PRO vs GO vs UniProtKB

BioOntologies (GO) enable science

GO (and other biomedical ontologies) allows a new kind of biological research, based on analysis and comparison of the massive quantities of annotations providing semantic intersections of information about gene products