Chapter 4. Dataset and Text Retrieval

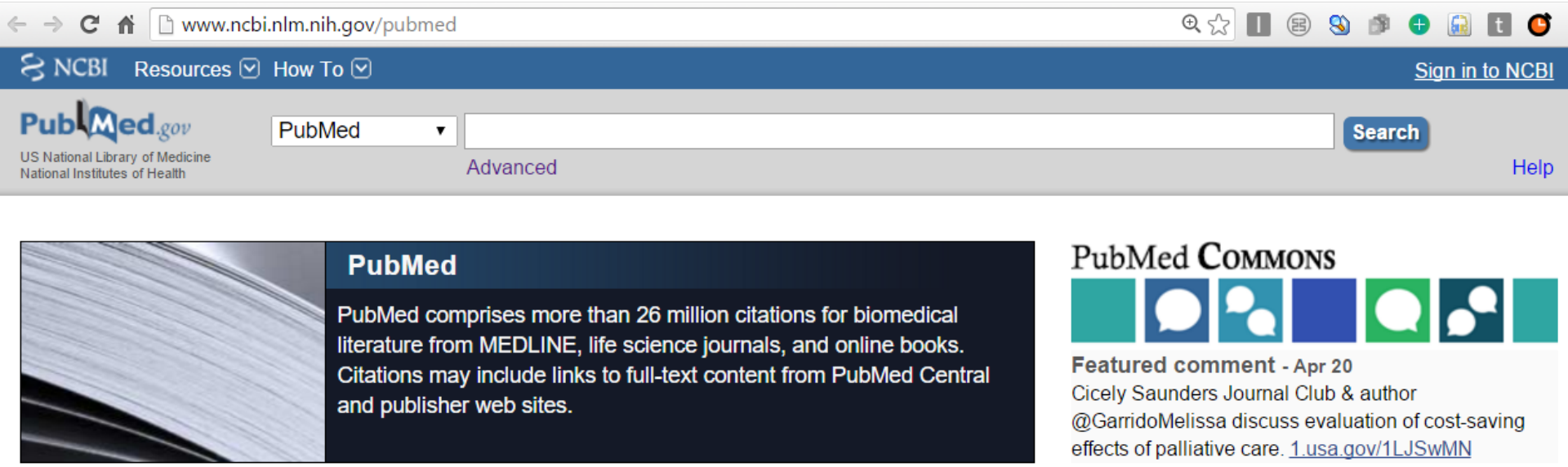
Jingbo Xia
College of Informatics, HZAU

Outline

- NCBI PubMED Text Retrieval
- Why Ontology
- Gene Ontology
- Some Other Ontologies, like UMLS(MetaMAP) and ICD (EHR)

Outline

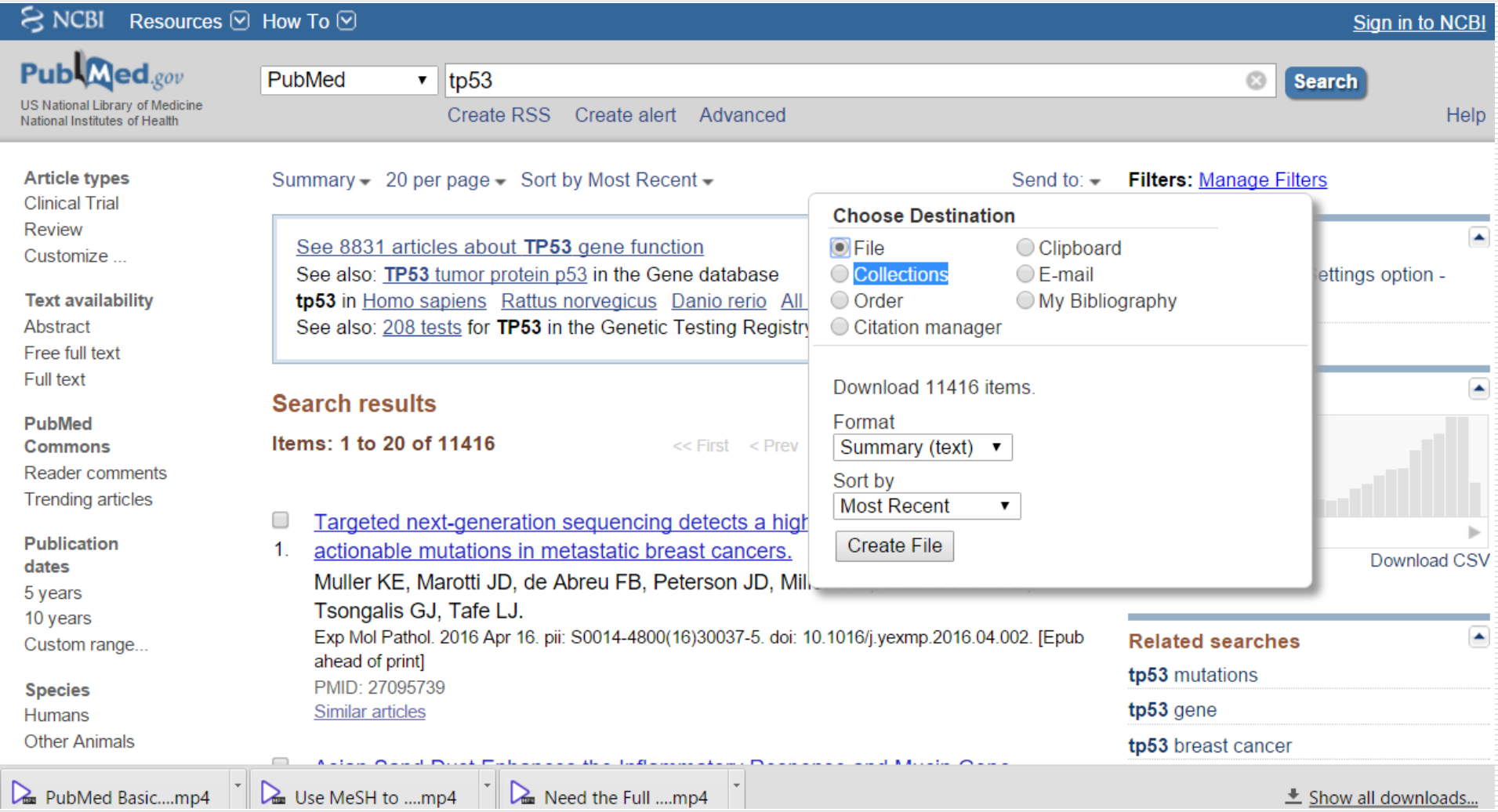
- NCBI PubMED Text Retrieval
- Why Ontology
- Gene Ontology
- Some Other Ontologies, like UMLS(MetaMAP) and ICD (EHR)


www.ncbi.nlm.nih.gov/pubmed
NCBI
Resources
How To
Sign in to NCBI
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Search
Advanced
Help
PubMed
PubMed comprises more than 26 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.
PubMed Commons
Featured comment - Apr 20
Cicely Saunders Journal Club & author @GarridoMelissa discuss evaluation of cost-saving effects of palliative care. 1.usa.gov/1LJSwMN
Using PubMed
PubMed Quick Start Guide
Full Text Articles
PubMed FAQs
PubMed Tutorials
New and Noteworthy
PubMed Tools
PubMed Mobile
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Topic-Specific Queries
More Resources
MeSH Database
Journals in NCBI Databases
Clinical Trials
E-Utilities (API)
LinkOut

Need the Full Text Article-.mp4

PubMed Basic Searching Tutorial.mp4

Use MeSH to Build a Better PubMed Query.mp4


NCBI Resources How To
Sign in to NCBI
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
tp53
Search
Create RSS Create alert Advanced
Help
Article types
Clinical Trial
Review
Customize ...
Text availability
Abstract
Free full text
Full text
PubMed Commons
Reader comments
Trending articles
Publication dates
5 years
10 years
Custom range...
Species
Humans
Other Animals
Summary 20 per page Sort by Most Recent
Send to: Filters: Manage Filters
See 8831 articles about TP53 gene function
See also: TP53 tumor protein p53 in the Gene database
tp53 in Homo sapiens Rattus norvegicus Danio rerio All
See also: 208 tests for TP53 in the Genetic Testing Registry
Choose Destination
File
Clipboard
Collections
E-mail
Order
My Bibliography
Citation manager
Download 11416 items.
Format
Summary (text)
Sort by
Most Recent
Create File
Search results
Items: 1 to 20 of 11416
<< First < Prev
1. Targeted next-generation sequencing detects a high actionable mutations in metastatic breast cancers.
Muller KE, Marotti JD, de Abreu FB, Peterson JD, Mil
Tsongalis GJ, Tafe LJ.
Exp Mol Pathol. 2016 Apr 16. pii: S0014-4800(16)30037-5. doi: 10.1016/j.yexmp.2016.04.002. [Epub ahead of print]
PMID: 27095739
Similar articles
Download CSV
Related searches
tp53 mutations
tp53 gene
tp53 breast cancer
PubMed Basic....mp4
Use MeSH to mp4
Need the Full mp4
Show all downloads...

Outline

- NCBI PubMED Text Retrieval
- **Why Ontology**
- Gene Ontology
- Some Other Ontologies, like UMLS(MetaMAP) and ICD (EHR)

What's in a name?

The problem:

- Same name for different concepts
- Different names for the same concept
- Vast amounts of biological data from different sources

→ Cross-species or cross-database comparison is difficult

Outline

- NCBI PubMED Text Retrieval
- Why Ontology
- **Gene Ontology**
- Some Other Ontologies, like UMLS(MetaMAP) and ICD (EHR)

What is the Gene Ontology?

- A (part of the) solution:
  - The Gene Ontology: "a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing"
- A controlled vocabulary to describe gene products - proteins and RNA - in any organism.

http://www.geneontology.org/

What is the Gene Ontology?

- One of the Open Biological Ontologies
- Standard, species-neutral way of representing biology
- Three structured networks of defined terms to describe gene product attributes
- More like a phrase book than a biology text book

How does GO work?

What information might we want to capture about a gene product?

- What does the gene product do?
- Where and when does it act?
- Why does it perform these activities?

No GO Areas

- GO covers ‘normal’ functions and processes
  - No pathological processes
  - No experimental conditions
- NO evolutionary relationships
- NO gene products
- NOT a system of nomenclature

The vocabularies

- Molecular function — elemental activity or task
- Biological process — broad objective or goal
- Cellular component — location or complex

The vocabularies

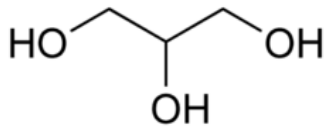
- Molecular function — elemental activity or task
  - nuclease, DNA binding, microtubule motor
- Biological process — broad objective or goal
- Cellular component — location or complex

The vocabularies

- Molecular function — elemental activity or task
  - nuclease, DNA binding, microtubule motor
- Biological process — broad objective or goal
  - mitosis, signal transduction, metabolism
- Cellular component — location or complex

The vocabularies

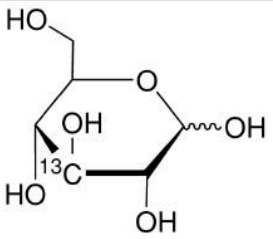
- Molecular function — elemental activity or task
  - nuclease, DNA binding, microtubule motor
- Biological process — broad objective or goal
  - mitosis, signal transduction, metabolism
- Cellular component — location or complex
  - nucleus, ribosome



Anatomy of a GO term

id: GO:0006094	unique GO ID
name: gluconeogenesis	term name
namespace: process	ontology
def: The formation of glucose from noncarbohydrate precursors, such as pyruvate, amino acids and glycerol. [http://cancerweb.ncl.ac.uk/omd/index.html]	definition
exact_synonym: glucose biosynthesis	synonym
xref_analog: MetaCyc:GLUCONEO-PWY	database ref
is_a: GO:0006006 is_a: GO:0006092	parentage

Anatomy of a GO term



- GO synonyms include alternative wordings, spellings, and related concepts
  - Broader, narrower, exact or related
  - Useful search aid

name: glucose transport

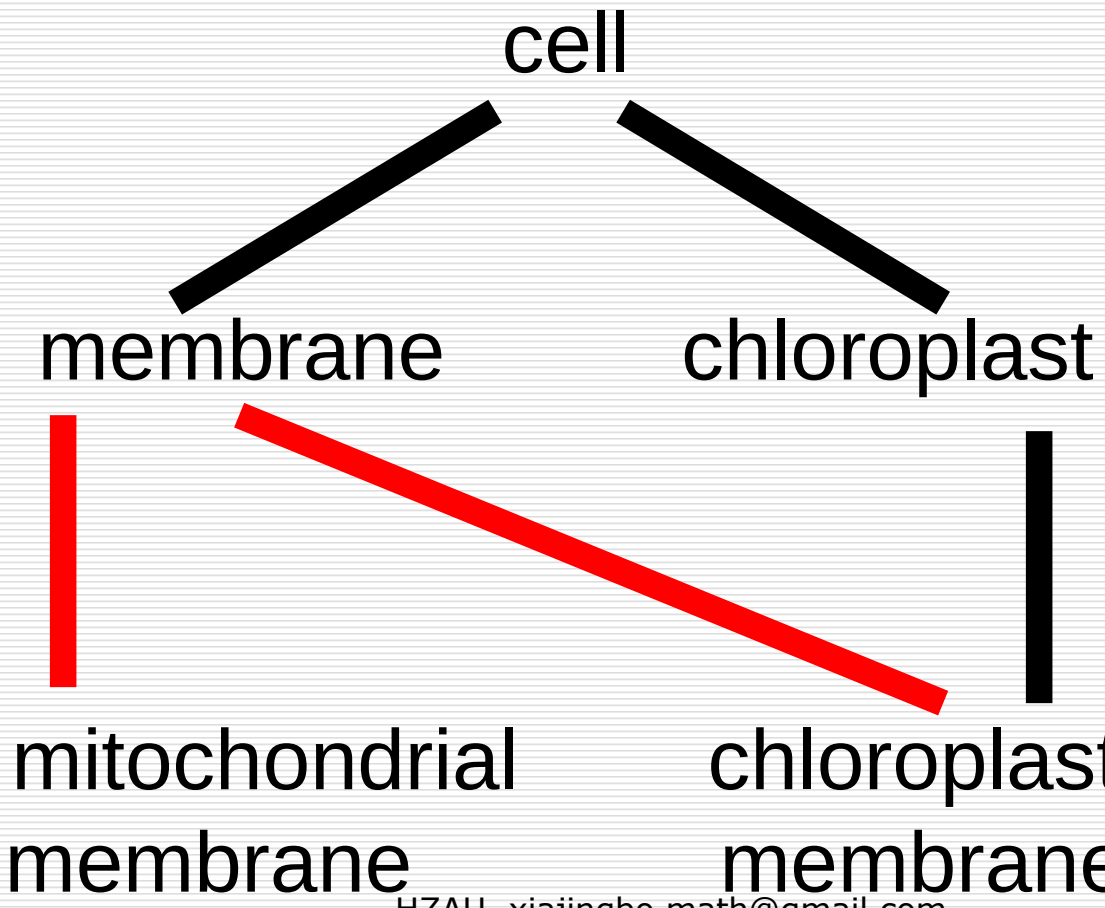
exact_synonym: gluco-hexose transport

narrow_synonym: glucose shuttling

Ontology Structure

- Ontologies are structured as a hierarchical directed acyclic graph
- Terms can have more than one parent and zero, one or more children
- Terms are linked by two relationships
  - is-a ⓘ
  - part-of Ⓟ

Ontology Structure



is-a

part-of

True Path Rule

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

cell

Ⓟ nucleus

Ⓟ chromosome

But what about bacteria?

True Path Rule

Resolved component ontology structure:

cell

Ⓟ cytoplasm

Ⓟ chromosome

ⓘ nuclear chromosome

Ⓟ nucleus

Ⓟ nuclear chromosome

GO Annotation

- Using GO terms to represent the activities and localizations of a gene product
- Annotations contributed by members of the GO Consortium
  - model organism databases
  - cross-species databases, eg. UniProt
- Annotations freely available from GO website

GO Annotation

- Database object
  - gene or gene product
- GO term ID
  - e.g. GO:0003677
- Reference for annotation
  - e.g. PubMed paper, BLAST results
- Evidence code
  - from evidence code ontology

GO Annotation

- Electronic annotation
  - from mappings files
    - e.g. UniProt keyword2go
  - High quantity but low quality
    - Annotations to low level terms
    - Not checked by curators
- Manual annotation
  - From literature curation
  - Time consuming but high quality

GO Annotation

ISS	Inferred from Sequence/Structural Similarity
IDA	Inferred from Direct Assay
IPI	Inferred from Physical Interaction
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
IMP	Inferred from Mutant Phenotype
IGI	Inferred from Genetic Interaction
IEP	Inferred from Expression Pattern
IC	Inferred by Curator
NDNo Data available	

IEA Inferred from electronic annotation

HZAU, xieming@mail.hzau.edu.cn

GO Annotate (An example)

In this study, we report the isolation and molecular characterization of the B. napus PERK1 cDNA, that is predicted to encode a novel receptor-like kinase. We have shown that like other plant RLKs, the kinase domain of PERK1 has serine/threonine kinase activity. In addition, the location of a PERK1-GTP fusion protein to the plasma membrane supports the prediction that PERK1 is an integral membrane protein…these kinases have been implicated in early stages of wound response…

GO Annotate (An example)

In this study, we report the isolation and molecular characterization of the B. napus PERK1 cDNA, that is predicted to encode a novel **receptor-like kinase**. We have shown that like other plant RLKs, the kinase domain of PERK1 has **serine/threonine kinase activity**. In addition, the location of a PERK1-GTP fusion protein to the plasma membrane supports the prediction that PERK1 is an **integral membrane protein**…these kinases have been implicated in early stages of **wound response**…

Function: protein serine/threonine kinase activity ; GO:0004674 (IDA)

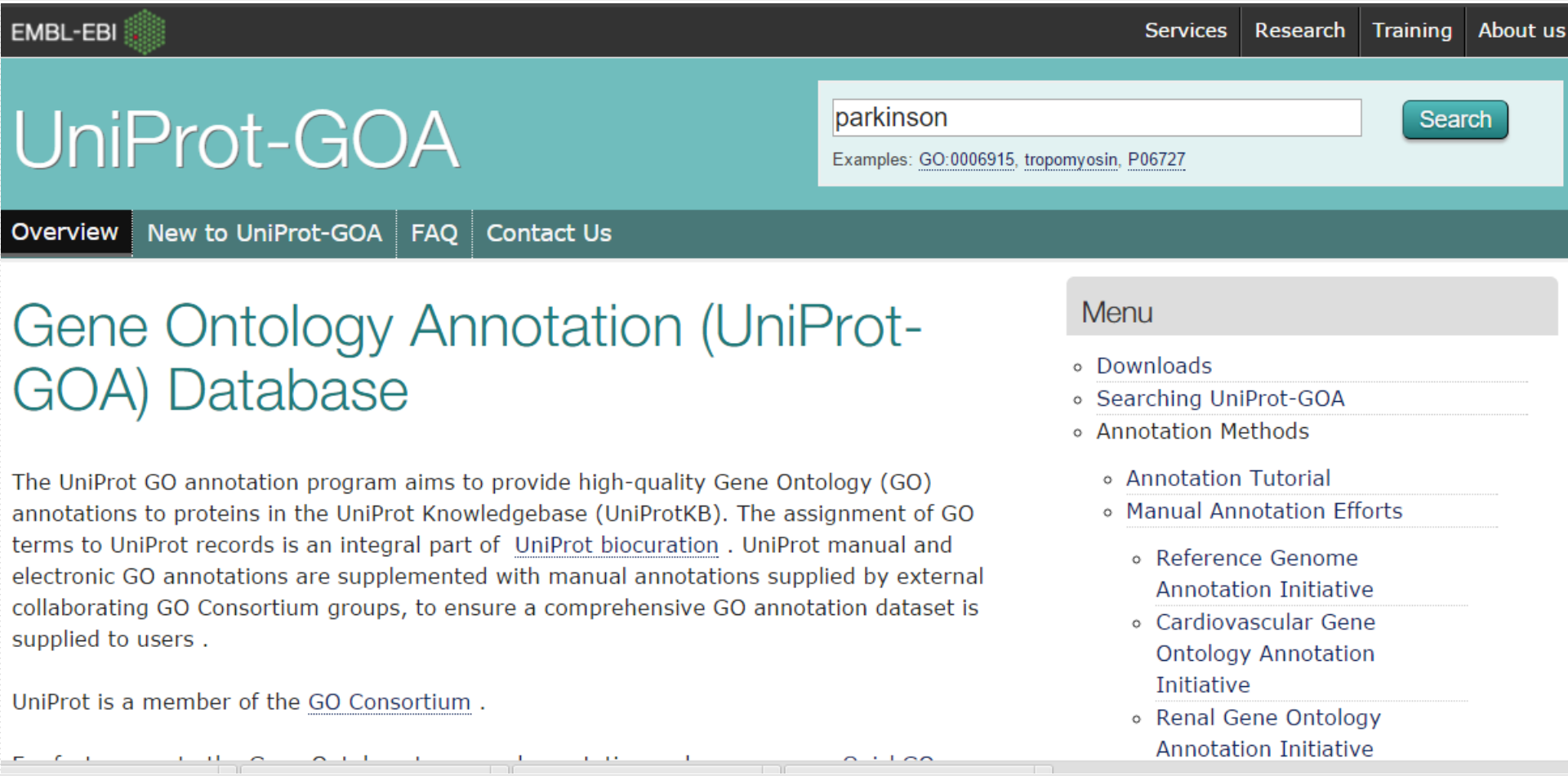
Component: integral to plasma membrane ; GO:0005887 (IDA)

Process: response to wounding ; GO:0009611 (NAS)

GO cross-links

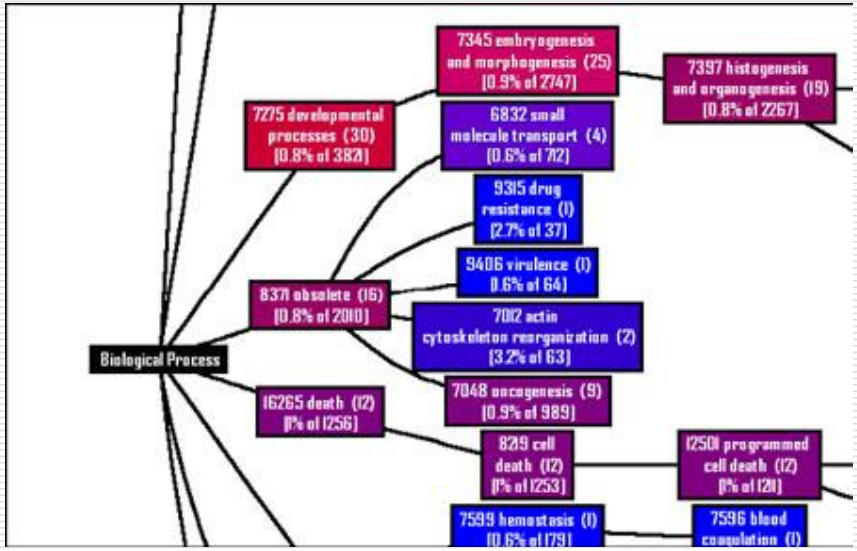
- Cross-references within GO
  - EC
  - Uniprot
- Mappings
  - SWISS-PROT keywords
- Links in other databases
  - UMLS/MeSH

http://www.ebi.ac.uk/GOA



GO tools

- Affymetrix also provide a Gene Ontology Mining Tool as part of their NetAffx™ Analysis Center which returns GO terms for probe sets



GO tools

- Many tools exist that use GO to find common biological functions from a list of genes:

ftp://ftp.geneontology.org/pub/go/www/GO.tools.annotation.shtml

GO tools

- Most of these tools work in a similar way:
  - input a gene list and a subset of 'interesting' genes
  - tool shows which GO categories have most interesting genes associated with them i.e. which categories are 'enriched' for interesting genes
  - tool provides a statistical measure to determine whether enrichment is significant

Outline

- NCBI PubMED Text Retrieval
- Why Ontology
- Gene Ontology
- **Some Other Ontologies, like UMLS(MetaMAP) and ICD (EHR)**

Case in using BioPortal:
https://bioportal.bioontology.org/annotator

Ontology used:

- Ontology of Clinical Research (OCRE)
- International Classification of Diseases, Version 10 - Clinical Modification (ICD10CM)
- Gene Ontology (GO)
- Human Phenotype Ontology (HP)
- The Drug Ontology (DRON)

Testing the following text:

To evaluate the role of oral ketamine as an adjuvant to oral morphine in cancer patients experiencing neuropathic pain, 9 cancer patients (5 men, 4 women) taking maximally tolerated doses of either morphine, amitriptyline, sodium valporate, or a combination of these drugs for intractable neuropathic pain, and reporting a pain score of >6 on a 0–10 scale, were studied prospectively to evaluate analgesia and adverse effects. Ketamine in the dose of 0.5 mg/kg body weight three times daily was added to the existing drug regimen. Patients were taught to maintain a pain diary wherein they daily recorded their pain, sedation, and vomiting scores, and other side effects. A decrease of more than 3 from the baseline in the average pain score, or a score of ≤3 was taken as a successful response. Seven patients exhibited a decrease of more than 3. Four patients experienced nausea, of which one had vomiting. Two developed loss of appetite. Eight patients reported drowsiness during the first two weeks of therapy ($P = 0.001$), and this gradually improved over the next two weeks in 5 of these 8 patients. Three patients withdrew from the study, two owing to excessive sedation and another due to a "feeling of unreality." None of the patients reported visual or auditory hallucinations. This experience suggests that low dose oral ketamine is beneficial and effective in the management of intractable neuropathic pain in patients with advanced cancer. However, its utility is limited in some patients by the adverse effects that accompany its use.

Assignment

- Imagine you are looking for drug & therapy information discovery based on the above text, choose 3 best ontologies from the above link, and enunciate the reason.

Reference

- Amelia Ireland, GO Curator, GO : the Gene Ontology. “because you know sometimes words have two meanings”. EBI, Cambridge, UK
- Jane Lomax, Gene Ontology Tutorial
- TEXT MINING, Bioinformatics and Computational Biology, Summer School – University Complutense of Madrid