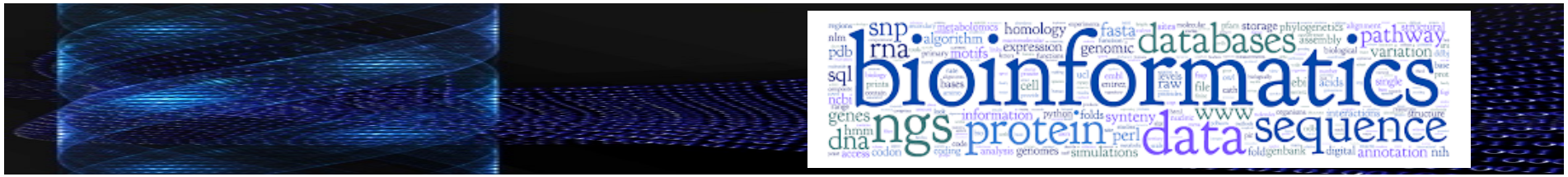


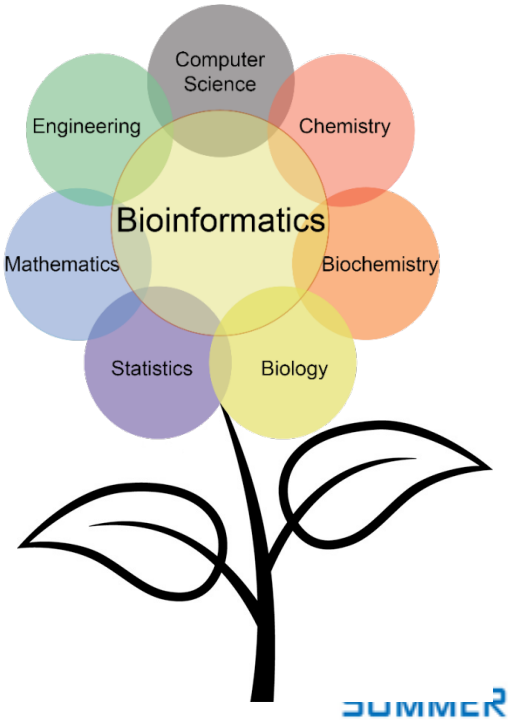
Session - Ontologies and biology databases

Benjamin Barré
Julie Bourbeillon
Claudine Landès

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Team

- Benjamin Barré
  - Associate Professor – University of Angers
  - Biochemistry, Molecular and cellular Biology
  - benjamin.barre@univ-angers.fr
- Julie Bourbeillon
  - Associate Professor – AgroCampus Ouest
  - Computer Sciences
  - julie.bourbeillon@agrocampus-ouest.fr
- Claudine Landès
  - Full Professor – University of Angers
  - Bioinformatics
  - claudine.landes@univ-angers.fr

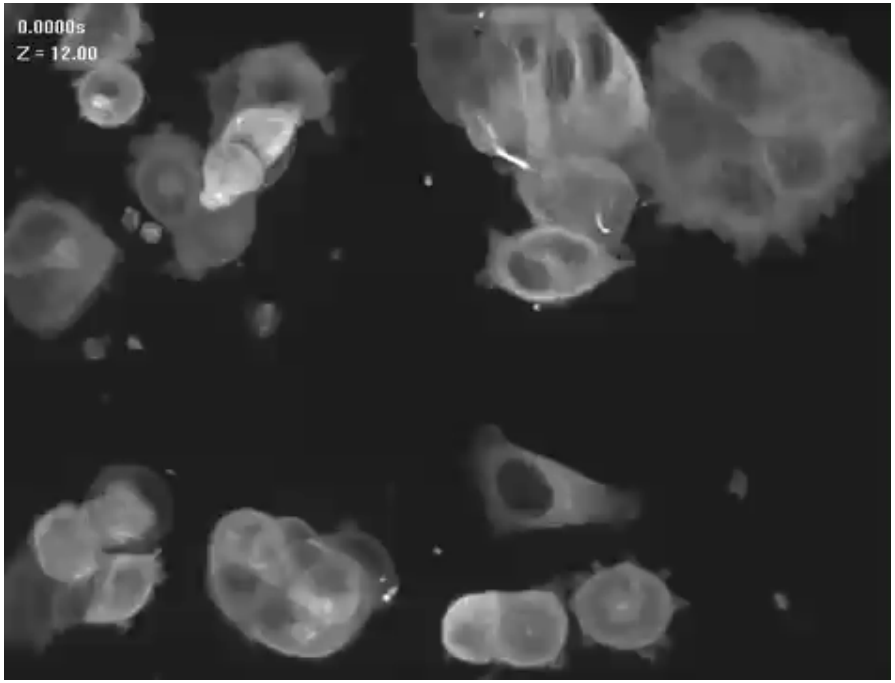


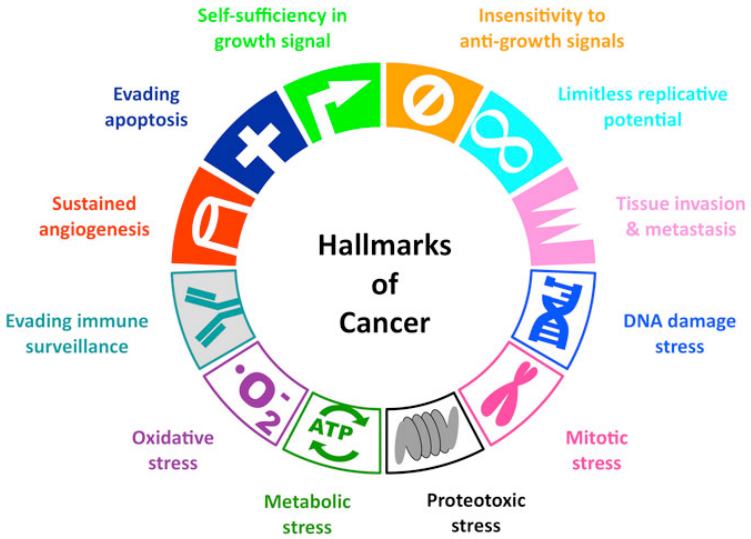
Aim of the session

- **Explore some biological databases to answer biological questions based on examples**
- **Approach the data integration issue**
- **Understand that storing data is not enough**
- **Realise there is a need for a common vocabulary → ontologies**

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

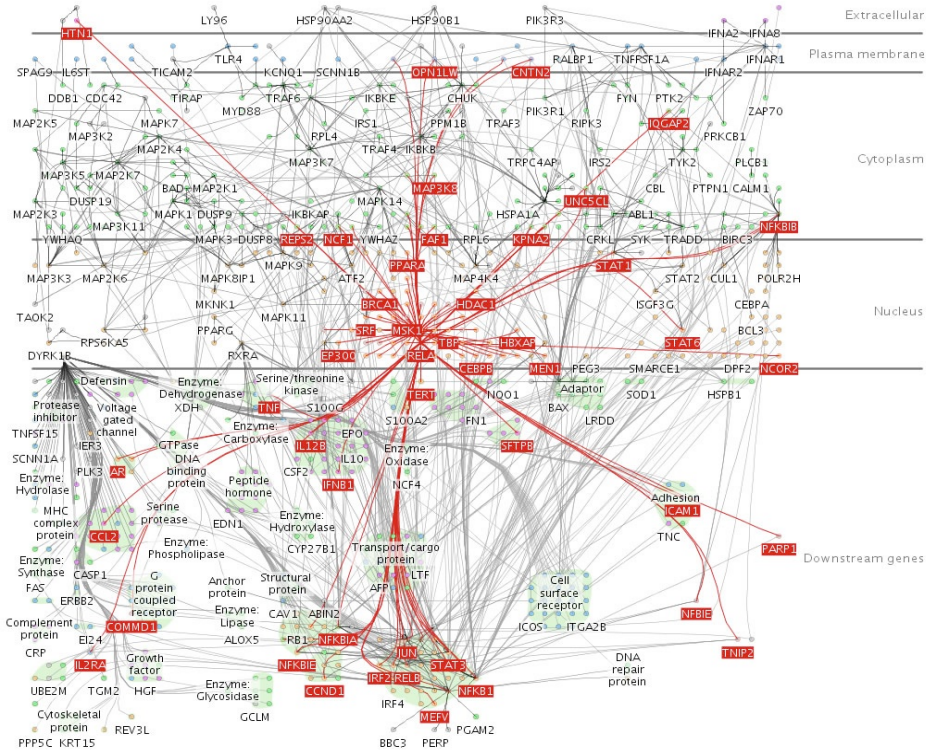
What is cancer ??





What's up in cancer ??

- Biology of systems



- 2 points :
  - ID of cancer (markers)
  - Functions of markers

For example in the lab

- Starting from Mass spectometry

PROTEOME
“the PROTein complement of a genOME”
Wasinger et al 1995 Electrophoresis: 16:1090

PROTEOMICS
“...the identification of all the proteins encoded in the human genome...”
Human Proteome Organisation (www.hupo.org)

One genome …
But different proteomes

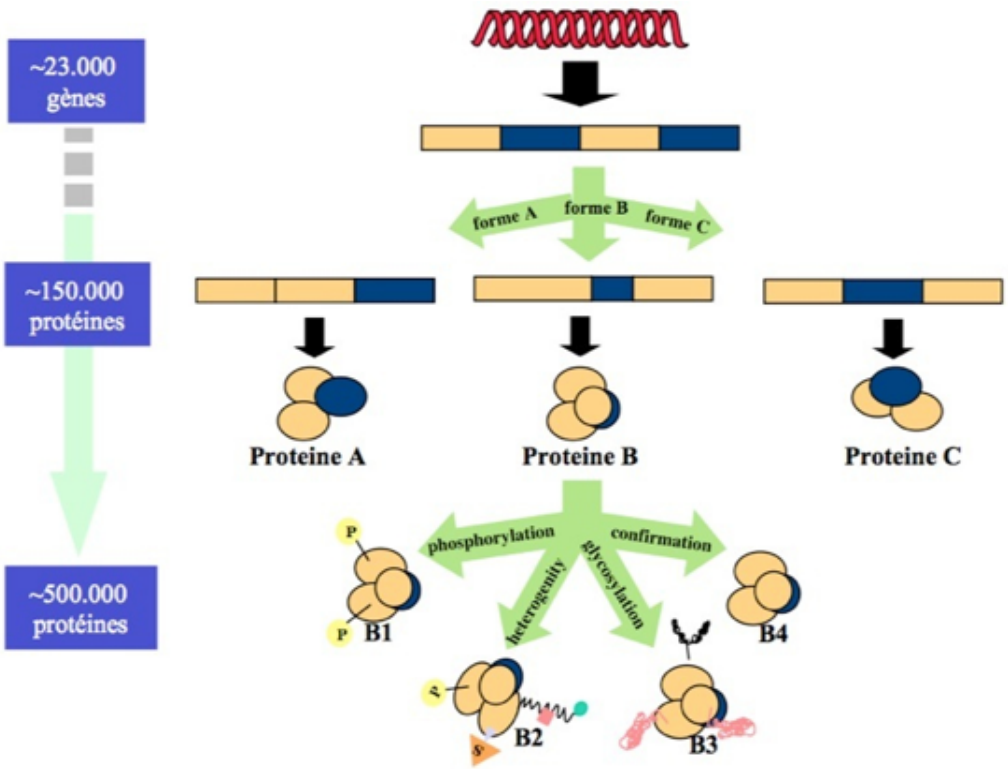
An example

A glamour example

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

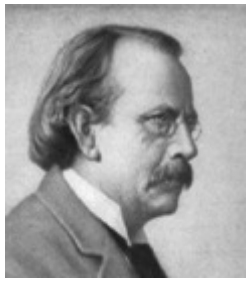
For example in the lab

- Starting from Mass spectometry
- Complexity of the proteome



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Mass spectrometry concept



Sir Joseph John Thomson
(1856 - 1940)
Cambridge University

Francis William Aston
(1877 - 1945)
Cambridge University

Protein and Polymer Analyses up to *m/z* 100 000 by Laser Ionization Time-of-flight Mass Spectrometry

Koichi Tanaka[†], Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida and Tamio Yoshida
Shimadzu Corporation, Nishinokyo-Kuwabaracho, Nakagyo-ku, Kyoto 604, Japan

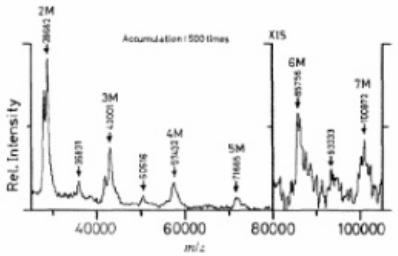


Figure 2. Laser ionization mass spectrum of lysozyme from chicken egg white, mol. wt 14306 Da.

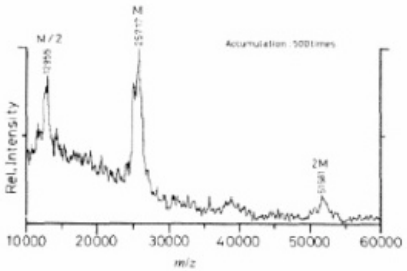


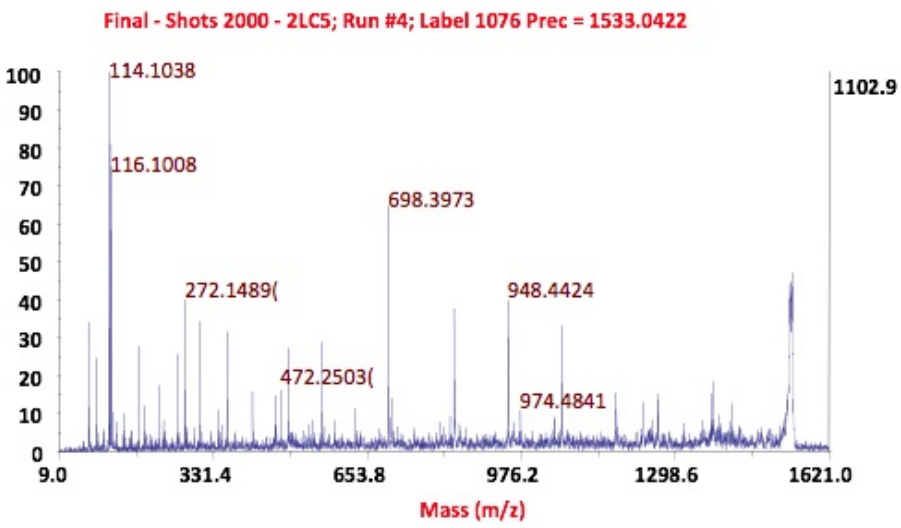
Figure 3. Laser ionization mass spectrum of chymotrypsinogen, mol. wt 25717 Da.

1917-2010
Prix Nobel 2002

1959- ...
Prix Nobel 2002

=> Session of Friday 5th July…

Mass spectrometry concept



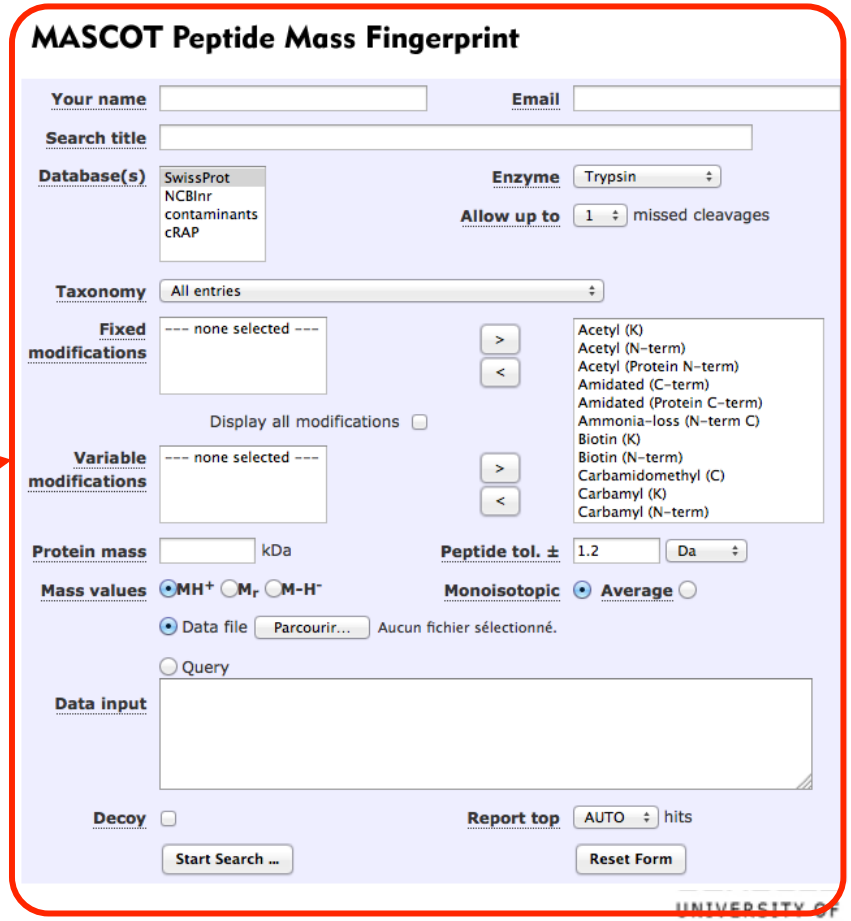
4148.1061
3221.5113
3392.5756
2993.3522
2468.2415
2325.0207
2144.9770
1948.9585
2005.9800
1906.8817
1883.0497
1721.7765
1618.7577
1732.8006
1508.7889
1288.7259
1134.6742
1008.5360
1008.5295
1065.5509
950.4499
1007.4714
880.4159
832.3658
792.3709
759.4247
715.4348

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Mass spectrometry analysis

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

http://www.matrixscience.com/search_form_select.html



UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Mass spectrometry analysis

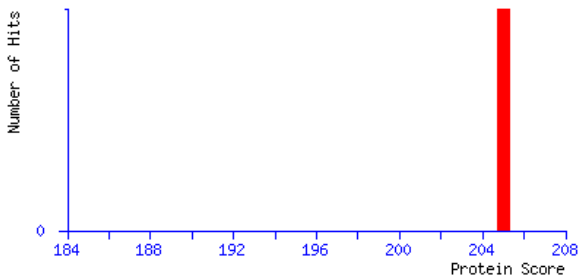
SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

MATRIX SCIENCE Mascot Search Results

```
User            : Ben
Email           : Benjamin.barre@univ-angers.fr
Search title    : AGR2
Database        : SwissProt 2016_06 (551385 sequences; 196948568 residues)
Timestamp       : 5 Jul 2016 at 07:59:21 GMT
Top Score       : 205 for AGR2_HUMAN, Anterior gradient protein 2 homolog OS=Homo sapiens GN=AGR2 PE=1 SV=1
```

Mascot Score Histogram

Protein score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 70 are significant (p<0.05).



Concise Protein Summary Report

Format As | Concise Protein Summary | Help

Significance threshold p< 0.05 | Max. number of hits AUTO

Preferred taxonomy | All entries

Re-Search All | Search Unmatched

1. AGR2_HUMAN Mass: 19967 Score: 205 Expect: 1.7e-15 Matches: 14
Anterior gradient protein 2 homolog OS=Homo sapiens GN=AGR2 PE=1 SV=1

Protein sequence coverage: 88%

Matched peptides shown in ***bold red***.

```
  1 MEKIPVSAFL LLVALSYTLA RDTTVKPGAK KDTKDSRPKL PQTLSRGWGD
 51 QLIWTQTYEE ALYKSKTSNK PLMIIHHLDE CPHSQALKKV FAENKEIQKL
101 AEQFVLLNLV YETTDKHLSP DGQYVPRIMF VDPSLTVRAD ITGRYSNRLY
151 AYEPADTALL LDNMKKALKL LKTEL
```

Unformatted sequence string: 175 residues (for pasting into other applications).

Sort peptides by | Residue Number | Increasing Mass | Decreasing Mass

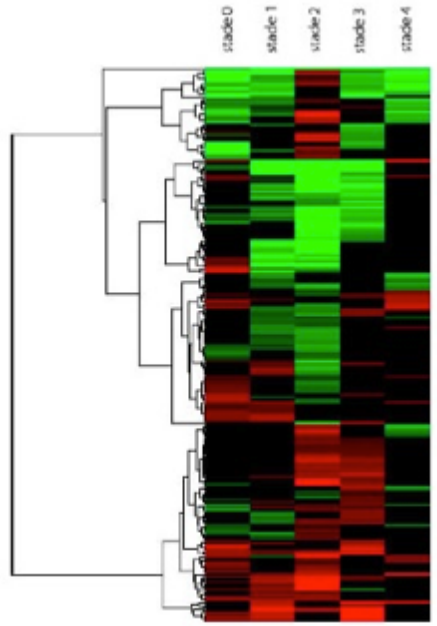
Show predicted peptides also

Start – End	Observed	Mr(expt)	Mr(calc)	Delta	M	Peptide
4 – 21	1947.1677	1946.1604	1946.1604	-0.0000	0	K.IPVSAFLLLVALSYTLAR.D
22 – 30	916.5098	915.5025	915.5025	-0.0000	0	R.DTTVKPGAK.K
35 – 39	602.3256	601.3183	601.3184	-0.0000	0	K.DSRPK.L
40 – 46	814.4781	813.4708	813.4708	-0.0000	0	K.LPQTLSR.G
47 – 64	2201.0549	2200.0476	2200.0477	-0.0000	0	R.GWGDQLIWTQTYEEALYK.S
67 – 88	2512.2799	2511.2726	2511.2726	0.0000	0	K.TSNKPLMIIHHLDECPHSQALK.K
90 – 95	707.3722	706.3649	706.3650	-0.0001	0	K.VFAENK.E
96 – 99	517.2980	516.2907	516.2907	-0.0000	0	K.EIQK.L
100 – 116	1996.0637	1995.0564	1995.0564	-0.0000	0	K.LAEQFVLLNLVYETTDK.H
117 – 127	1268.6382	1267.6309	1267.6309	-0.0000	0	K.HLSPDGQYVPR.I
128 – 138	1277.6922	1276.6849	1276.6850	-0.0000	0	R.IMFVDPSLTVR.A
139 – 144	632.3362	631.3289	631.3289	-0.0000	0	R.ADITGR.Y
145 – 148	539.2572	538.2499	538.2499	-0.0000	0	R.YSNR.L
149 – 165	1940.9673	1939.9600	1939.9601	-0.0000	0	R.LYAYEPADTALLLDNMK.K

BIOINFORMATICS

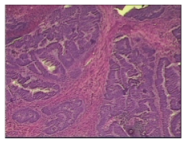
The question

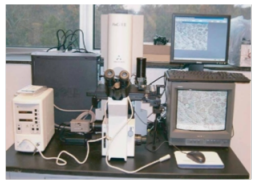
- How could we understand cell transformation and cancer ??



Tumor samples

Préparation des échantillons



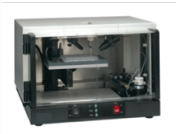


Pixcell II (Arcturus)

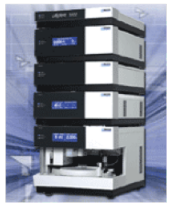
iTRAQ

3100 OFFGEL (Agilent)

MALDI-TOF/TOF 5800 (Applied Biosystems)



Probot™ Micro Fraction Collector (Dionex)



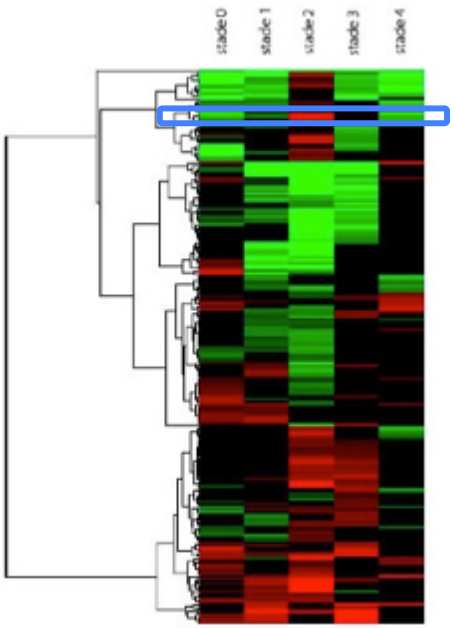
UltiMate 3000 nanoLC (Dionex)

ID of the tumor
(first point)

Breast Cancer Res. 2015 Mar 20;17:43.
Mol Cell Proteomics. 2015 Nov;14(11):2936-46.
Mol Cell Proteomics. 2011 Dec;10(12):M111.009712.

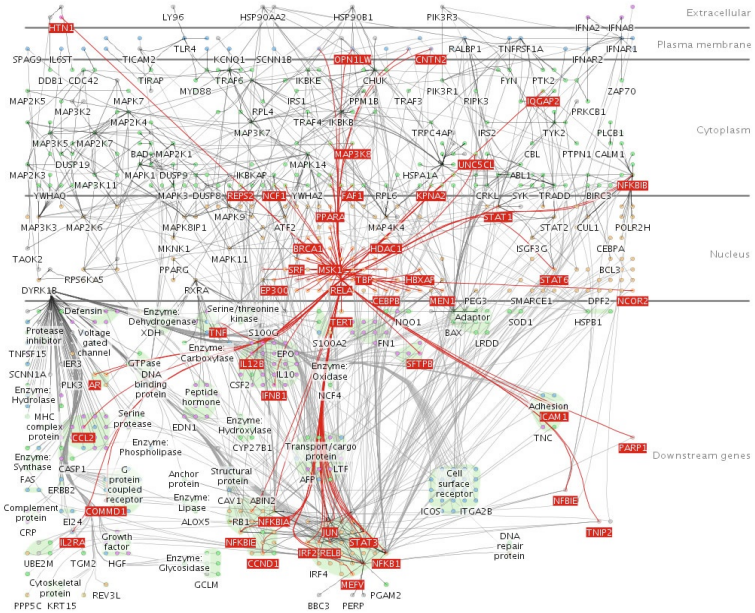
The question

- How could we understand cell transformation and cancer ??



ID of the tumor
(first point)

Define function of each marker to construct a map of biology of systems
(second point)



Question

- What do you need to understand the functions of the partners ?
  - Excel sheet containing the proteins list

(Partners.xlsx)

Question

- What do you need to understand the functions of the partners ?
  - Excel sheet containing the proteins list

(Partners.xlsx)

- Hint
  - What do you search ?
  - Where do you search ?
  - What do you expect to retrieve ?

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Question

- What do you need to understand the functions of the partners ?
  - Excel sheet containing the proteins list

(Partners.xlsx)

- Possible solution
  - Gene database
    - https://www.ncbi.nlm.nih.gov/gene

Game on !

Explore NCBI Databases

- The goal
  - Find informations about a list of genes
  - Input : accession numbers
  - Databases to search from NCBI
  - Output :
    - accession number of protein
    - accession number of gene
    - protein name, isoform
    - genomic location
    - homologous gene in *M.musculus*

NCBI search

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

NCBI Resources How To

NCBI
National Center for Biotechnology Information

All Databases A2RUR9 Sear

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

Popular Resources

PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog

Applications open for th
NCBI-NLM Bioinformati

From August 14th – 16t
involvement from sever

PubMed is now availabl

Search NCBI databases

Help

A2RUR9 Search

Results found in 3 databases for "A2RUR9"

RecName: Full=Coiled-coil domain-containing protein 144A
1,427 aa protein.
Accession: **A2RUR9**.1 GI: 162416041
GenPept FASTA Graphics Identical Proteins Gene

Literature

Books	0	books and reports
MeSH	0	ontology used for PubMed indexing
NLM Catalog	0	books, journals and more in the NLM Collections
PubMed	0	scientific & medical abstracts/citations
PubMed Central	0	full-text journal articles

Health

ClinVar	0	human variations of clinical significance
dbGaP	0	genotype/phenotype interaction studies
GTR	0	genetic testing registry
MedGen	0	medical genetics literature and links
OMIM	0	online mendelian inheritance in man
PubMed Health	0	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly information
BioCollections	0	museum, herbaria, and other biorepository collections
BioProject	0	biological projects providing data to NCBI
BioSample	0	descriptions of biological source materials
Clone	0	genomic and cDNA clones
dbVar	0	genome structural variation studies
Genome	0	genome sequencing projects by organism
GSS	0	genome survey sequences
Nucleotide	7	DNA and RNA sequences
Probe	0	sequence-based probes and primers

Genes

EST	0	expressed sequence tag sequences
Gene	1	collected information about gene loci
GEO DataSets	0	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	0	homologous gene sets for selected organisms
PopSet	0	sequence sets from phylogenetic and population studies
UniGene	0	clusters of expressed transcripts

Proteins

Conserved Domains	0	conserved protein domains
Protein	7	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	0	experimentally-determined biomolecular structures

Chemicals

BioSystems	0	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	0	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	0	deposited substance and chemical information

The Gene file !!!

NCBI Resources How To Sign in to NCBI

Gene | Gene | A2RUR9 | Search

Create RSS Create alert Advanced Help

Full Report Send to: Hide sidebar >>

Showing Current, Replaced and Discontinued items. Show only Current (1).

CCDC144A coiled-coil domain containing 144A [*Homo sapiens* (human)]

Gene ID: 9720, updated on 4-Jun-2017

Summary

Official Symbol	CCDC144A provided by HGNC
Official Full Name	coiled-coil domain containing 144A provided by HGNC
Primary source	HGNC:HGNC:29072
See related	Ensembl:ENSG00000170160 Ensembl:ENSG00000266302 Vega:OTTHUMG00000059178; Vega:OTTHUMG00000179236
Gene type	protein coding
RefSeq status	VALIDATED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Orthologs	all

Genomic context

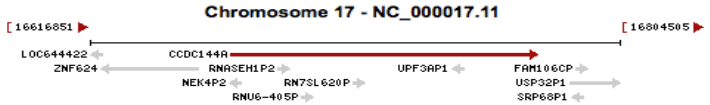
See CCDC144A in Genome Data Viewer Map Viewer

Location: 17p11.2

Exon count: 22

Annotation release	Status	Assembly	Chr	Location
108	current	GRCh38.p7 (GCF_000001405.33)	17	NC_000017.11 (16666768..16774999)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	17	NC_000017.10 (16593549..16678311)

Chromosome 17 - NC_000017.11



Genomic regions, transcripts, and products

Go to reference sequence details

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Expression
- Bibliography
- Phenotypes
- Variation
- Interactions
- General gene information
  - Markers, Related pseudogene(s), Clone Names, Homology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

Genome Browsers

- Genome Data Viewer
- Map Viewer
- Variation Viewer (GRCh37.p13)
- Variation Viewer (GRCh38)
- 1000 Genomes Browser (GRCh37.p13)
- Ensembl
- Ensembl
- UCSC

Related information

SUMMER SCHOOLS UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

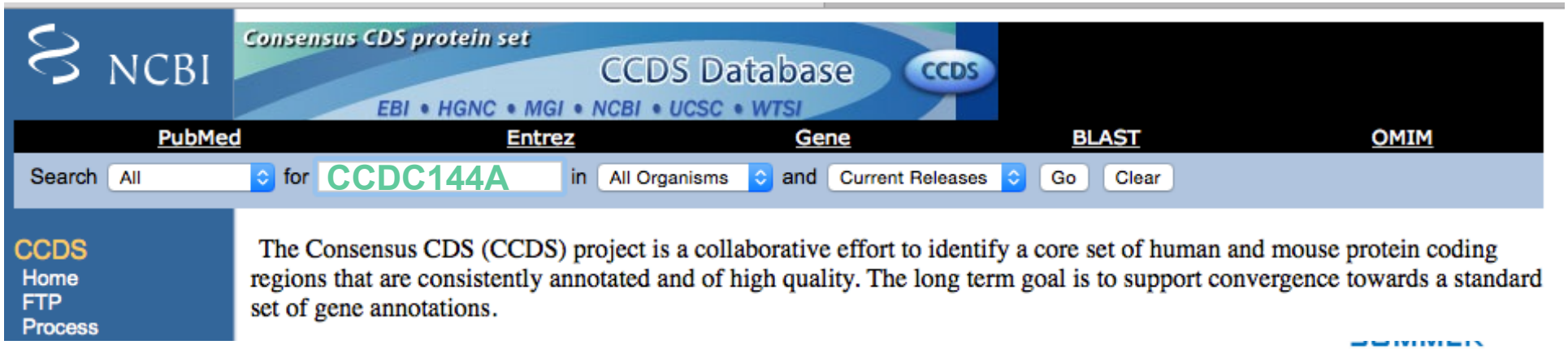
Try and Guess

- The accession number of the protein
- The accession number of the gene
- The protein name
- The genomic position of the gene
- The structure of the gene (exons/introns)
- The number of the isoform if exists
- The homologous genes in *M.musculus* (if exists)

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Results

- protein accession number : A2RUR9 UniprotKB
- gene AN : CCDC144A CCDS database
- protein name : Coiled-coil domain-containing protein 144A



Location details

Genomic context

Location: 17p11.2

Exon count: 22

See CCDC144A in Genome Data Viewer Map Viewer

Annotation release	Status	Assembly	Chr	Location
108	current	GRCh38.p7 (GCF_000001405.33)	17	NC_000017.11 (16666768..16774999)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	17	NC_000017.10 (16593549..16678311)

Chromosome 17 - NC_000017.11

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Results

- protein accession number : A2RUR9 UniprotKB
- gene AN : CCDC144A CCDS database
- protein name : Coiled-coil domain-containing protein 144A
- position : NC_000017.11 (Chr 17p11.2) (16666768..16774999)
- gene structure : a total of 22 exons
- isoform : in 8 isoforms

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Gene structure details

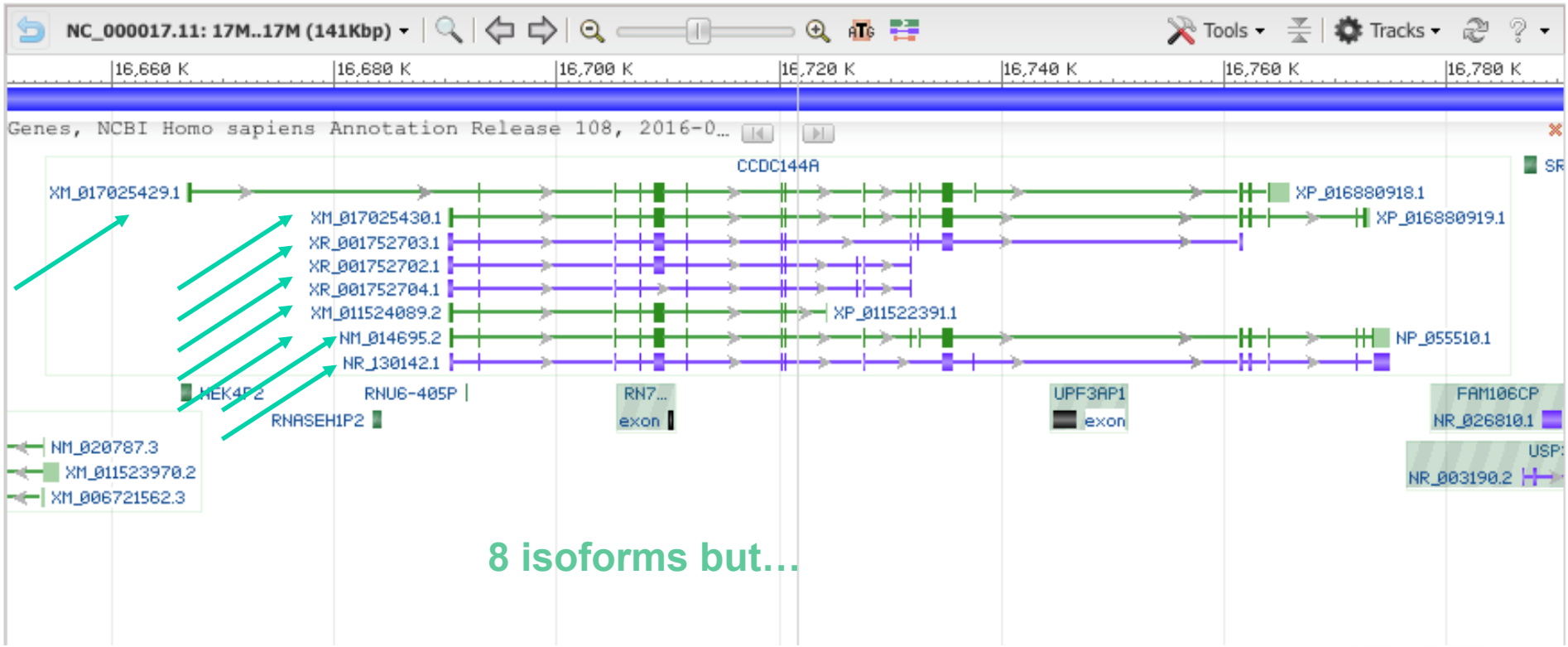
SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence: NC_000017.11 Chromosome 17 Reference GRCh38.p7 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank



8 isoforms but…

ANGERS - FRANCE
BIOINFORMATICS

Gene structure details

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

mRNA and Protein(s)

mRNA Protein

1. NM_014695.2 → NP_055510.1 **coiled-coil domain-containing protein 144A**

See identical proteins and their annotated locations for NP_055510.1

Status: VALIDATED

Description	Transcript Variant: This variant (1) represents the longer transcript and encodes the protein.
Source sequence(s)	BC133019, BC144225
Consensus CDS	CCDS45621.1
UniProtKB/Swiss-Prot	A2RUR9
UniProtKB/TrEMBL	B7ZM27
Related	ENSP00000353717.8, OTTHUMP00000260378, ENST00000360524.12, OTTHUMT00000444093

Conserved Domains (2) summary

pfam12001 Location:1244 → 1354	DUF3496; Domain of unknown function (DUF3496)
pfam14915 Location:822 → 1123	CCDC144C; CCDC144C protein coiled-coil region

RNA

RNA

1. NR_130142.1 **RNA Sequence**

Status: VALIDATED

Description	Transcript Variant: This variant (2) lacks three internal exons and contains an alternate internal exon, compared to variant 1. This variant is represented as non-coding because the use of the 5'-most expected translational start codon renders the transcript a candidate for nonsense-mediated mRNA decay (NMD).
Source sequence(s)	BC144224, BC144225
Related	ENST00000360495.9, OTTHUMT00000444094

2 validated isoforms and

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Gene structure details

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

Reference GRCh38.p7 Primary Assembly

Genomic

1. **NC_000017.11 Reference GRCh38.p7 Primary Assembly**

Range	16666768..16774999
Download	GenBank, FASTA, Sequence Viewer (Graphics)

mRNA and Protein(s)

1. XM_017025429.1 → XP_016880918.1 **coiled-coil domain-containing protein 144A isoform X1**
2. XM_017025430.1 → XP_016880919.1 **coiled-coil domain-containing protein 144A isoform X3**
3. XM_011524089.2 → XP_011522391.1 **coiled-coil domain-containing protein 144A isoform X2**

mRNA Protein

RNA

1. XR_001752703.1 **RNA Sequence**
2. XR_001752702.1 **RNA Sequence**

Related	ENST00000456009.4

3. XR_001752704.1 **RNA Sequence**

RNA

PREDICTED: Homo sapiens coiled-coil domain containing 144A (CCDC144A) transcript variant X1, mRNA

NCBI Reference Sequence: XM_017025429.1

FASTA Graphics

Go to:

```
LOCUS       XM_017025429            6106 bp    mRNA    linear   PRI 06-JUN-2016
DEFINITION  PREDICTED: Homo sapiens coiled-coil domain containing 144A
            (CCDC144A), transcript variant X1, mRNA.
ACCESSION   XM_017025429
VERSION     XM_017025429.1
DBLINK      BioProject: PRJNA168
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
COMMENT     MODEL REFSEQ:  This record is predicted by automated computational
            analysis. This record is derived from a genomic sequence
            (NC_000017.11) annotated using gene prediction method: Gnomon,
            supported by EST evidence.
            Also see:
                Documentation of NCBI's Annotation Process

            ##Genome-Annotation-Data-START##
            Annotation Provider         :: NCBI
            Annotation Status           :: Full annotation
            Annotation Version          :: Homo sapiens Annotation Release 108
            Annotation Pipeline         :: NCBI eukaryotic genome annotation
                                           pipeline
            Annotation Software Version :: 7.0
            Annotation Method           :: Best-placed RefSeq; Gnomon
            Features Annotated          :: Gene; mRNA; CDS; ncRNA
            ##Genome-Annotation-Data-END##
```

6 predicted isoforms

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Results details

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

✦ homologous gene in *M.musculus:*

Variation

Interactions

Products	Interactant	Other Gene	Complex	Source	Pubs	Description
BioGRID:115069	BioGRID:115158	TSC22D2		BioGRID	PubMed	Two-hybrid

General gene information

⊞ Markers

⊟ Related pseudogene(s)

2 found Review record(s) in Gene

⊟ Homology

Orthologs from Annotation Pipeline: 8 organisms have orthologs with human gene CCDC144A
The Hierarchical Catalog of Orthologs

⊟ Clone Names

FLJ43983, KIAA0565, MGC164650

General protein information

Preferred Names
coiled-coil domain-containing protein 144A

NCBI Resources How To Sign in to N

Gene

Gene ortholog_gene_9720[group] Search

Create RSS Create alert Advanced

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Protein-coding

Sequence content
CCDS
Ensembl
RefSeq

Status clear
✓ Current
✓ Discontinued
✓ Replaced

Chromosome locations
more...

Clear all

Show additional filters

Tabular 20 per page Sort by Relevance Send to:

Hide sidebar

Search results

no homolog in *M.musculus*

Items: 9

Showing Current, Replaced and Discontinued items. Show only Current (9).

Name/Gene ID	Description	Location	Aliases	MIM
CCDC144A ID: 9720	coiled-coil domain containing 144A [*Homo sapiens* (human)]	Chromosome 17, NC_000017.11 (16666768..16774999)		
CCDC144A ID: 742910	coiled-coil domain containing 144A [*Pan troglodytes* (chimpanzee)]	Chromosome 17, NC_006484.4 (37945047..38040303, complement)	CCDC144B	
CCDC144A ID: 105509140	coiled-coil domain containing 144A [*Colobus angolensis palliatus*]			
CCDC144A ID: 103242436	coiled-coil domain containing 144A [*Chlorocebus sabaeus* (green monkey)]	Chromosome 16, NC_023657.1 (14975398..15007531, complement)		
CCDC144A ID: 101139298	coiled-coil domain containing 144A [*Gorilla gorilla* (western gorilla)]	Chromosome 5, NC_018429.2 (61994943..62008736, complement)		
CCDC144A ID: 101051288	coiled-coil domain containing 144A [*Saimiri boliviensis* (Bolivian squirrel monkey)]			
CCDC144A ID: 101025815	coiled-coil domain containing 144A [*Papio anubis* (olive baboon)]	Chromosome 16, NC_018167.1 (15394960..15471954)		
CCDC144A ID: 100586810	coiled-coil domain containing 144A [*Nomascus leucogenys* (northern white-cheeked gibbon)]	Chromosome 19, NC_019834.1 (66487166..66563774)		
CCDC144A ID: 100443043	coiled-coil domain containing 144A [*Pongo abelii* (Sumatran orangutan)]			

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Homo sapiens *(1)*
Pan troglodytes *(1)*
Gorilla gorilla *(1)*
Pongo abelii *(1)*
Nomascus leucogenys *(1)*
Papio anubis *(1)*
Chlorocebus sabaeus *(1)*
Colobus angolensis palliatus *(1)*
Saimiri boliviensis *(1)*
Less...

Find related data

Database: Select

Find items

Search details

```
ortholog_gene_9720[group] AND
(alive[prop] OR replaced[Propertie
OR discontinued[Properties])
```

Search See m

Recent activity

Turn Off C

ortholog_gene_9720[group] AND ((alive[p
OR replaced[Propertie... (9)

BIOINFORMATICS

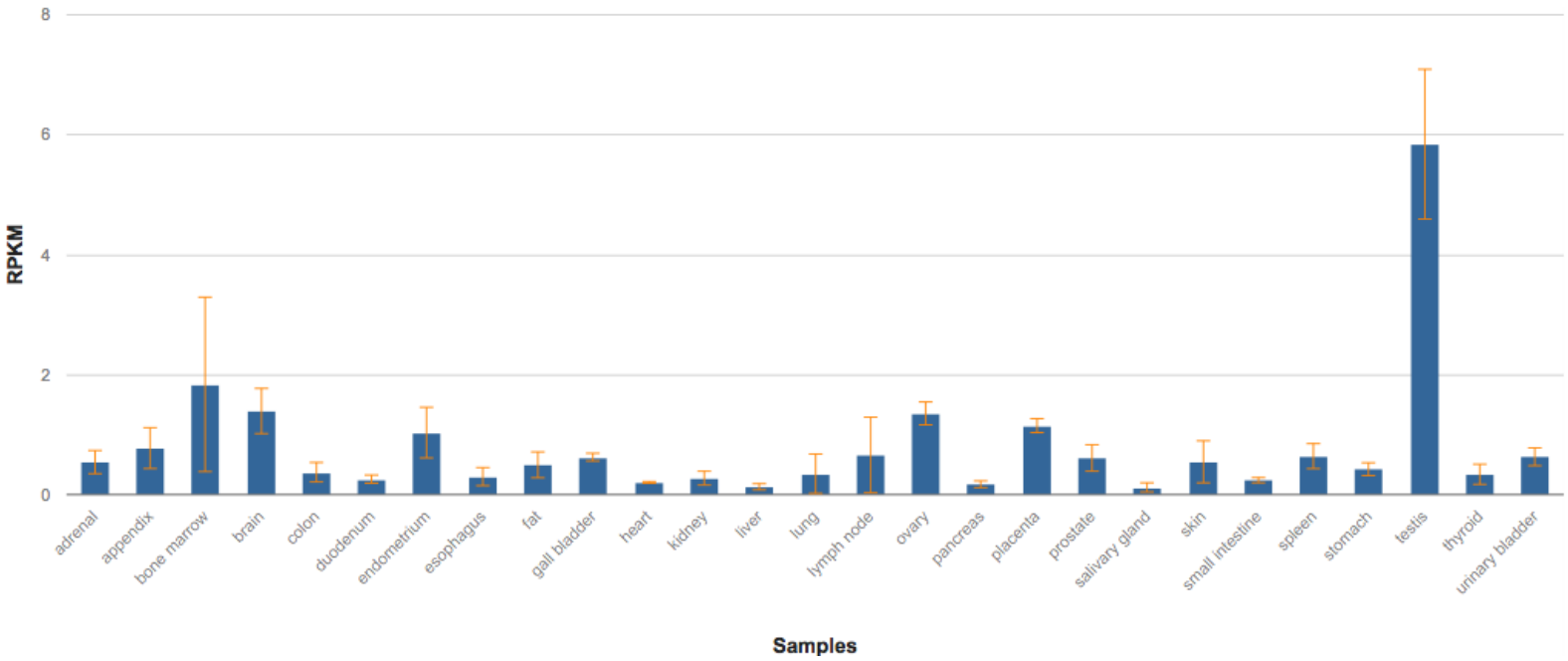
Results details

Expression

See details

HPA RNA-seq normal tissues

- Project title: HPA RNA-seq normal tissues
- Description: RNA-seq was performed of tissue samples from 95 human individuals representing 27 different tissues in order to determine tissue-specificity of all protein-coding genes
- BioProject: PRJEB4337
- Publication: PMID 24309898
- Analysis date: Wed Jun 15 11:32:44 2016



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Results details

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

Bibliography

Related articles in PubMed

1. Structure and evolution of the Smith-Magenis syndrome repeat gene clusters, SMS-REPs.
   Park SS, *et al.* Genome Res, 2002 May. PMID 11997339, **Free PMC Article**
2. Yeast two-hybrid screening identified WDR77 as a novel interacting partner of TSC22D2.
   Li Q, *et al.* Tumour Biol, 2016 Sep. PMID 27337956
3. Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro.
   Nagase T, *et al.* DNA Res, 1998 Feb 28. PMID 9628581
4. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences.
   Strausberg RL, *et al.* Proc Natl Acad Sci U S A, 2002 Dec 24. PMID 12477932, **Free PMC Article**

GeneRIFs: Gene References Into Functions What's a GeneRIF?

Submit: New GeneRIF Correction

Phenotypes

Find tests for this gene in the NIH Genetic Testing Registry (GTR)

Review eQTL and phenotype association data in this region using PheGenI

Variation

See variants in ClinVar

See studies and variants in dbVar

See Variation Viewer (GRCh37.p13)

See Variation Viewer (GRCh38)

Genotypes

See SNP Geneview Report
See 1000 Genomes Browser (GRCh37.p13)

UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Results details

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

Interactions

Products	Interactant	Other Gene	Complex	Source	Pubs	Description
BioGRID:115706	BioGRID:106848	APP		BioGRID	PubMed	Reconstituted Complex
BioGRID:115706	BioGRID:121263	ARNTL2		BioGRID	PubMed	Affinity Capture-MS
BioGRID:115706	BioGRID:106936	ASPH		BioGRID	PubMed	Affinity Capture-MS
BioGRID:115706	BioGRID:221186	Cep152		BioGRID	PubMed	Affinity Capture-MS
BioGRID:115706	BioGRID:118061	DISC1		BioGRID	PubMed	Two-hybrid
BioGRID:115706	BioGRID:120859	EDEM2		BioGRID	PubMed	Affinity Capture-MS
BioGRID:115706	BioGRID:108276	EGFR		BioGRID	PubMed	PCA
BioGRID:115706	BioGRID:122815	MANEA		BioGRID	PubMed	Affinity Capture-MS
BioGRID:115706	BioGRID:125007	OLFM2		BioGRID	PubMed	Affinity Capture-MS
BioGRID:115706	BioGRID:117746	PCOLCE2		BioGRID	PubMed	Affinity Capture-MS
BioGRID:115706	BioGRID:117577	TRAF3IP1		BioGRID	PubMed	Two-hybrid
BioGRID:115706	BioGRID:113129	TUBA3C		BioGRID	PubMed	Affinity Capture-MS
BioGRID:115706	BioGRID:119253	TXNDC11		BioGRID	PubMed	Affinity Capture-MS

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Results details

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

General gene information

Markers

Homology

Homologs of the OLFM1 gene: The OLFM1 gene is conserved in chimpanzee, Rhesus monkey, dog, mouse, rat, chicken, and frog.
Orthologs from Annotation Pipeline: 190 organisms have orthologs with human gene OLFM1
The Hierarchical Catalog of Orthologs

Gene Ontology Provided by GOA

Function	Evidence Code	Pubs
protein binding	IPI	PubMed

Process	Evidence Code	Pubs
atrioventricular valve formation	ISS	
cardiac epithelial to mesenchymal transition	ISS	
negative regulation of gene expression	ISS	
nervous system development	TAS	PubMed
neuronal signal transduction	ISS	
positive regulation of epithelial to mesenchymal transition	ISS	
positive regulation of gene expression	ISS	
regulation of axon extension	ISS	

Component	Evidence Code	Pubs
axon	ISS	
axonal growth cone	ISS	
cell	ISS	
cell junction	IEA	
endoplasmic reticulum	ISS	
extracellular space	ISS	
neuronal cell body	ISS	
perikaryon	IEA	
synapse	IEA	

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Practice

- Find information regarding the partner proteins
  - accession number of protein
  - accession number of gene
  - protein name, isoform
  - genomic location
  - homologous gene in M.musculus
- Excel sheet containing the proteins list (Partners.xlsx)
- Gene database
  - https://www.ncbi.nlm.nih.gov/gene

Game on !

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Results

- So ??....

Drowning in databases

- Exhaustive databases
  - Nucleic sequences: EMBL, Genbank, DDBJ
  - Proteins : SwissProt/Uniprot, Genpept, trEMBL
  - Structures : PDB
- Specialized databases (depending on point of view)
  - Non exhaustive list : organisms, proteomics, transcriptomics, metabolomics, comparative genomics, phylogenomics, diseases…
  - Examples :
    - GeneCards (integrated human genes database)
    - Protein Atlas (a tissue-based map of human proteome)
    - COMICS (catalogue of somatic mutations in cancer)
    - UCSC (human or mouse genome browser)

For a review

The 2018 *Nucleic Acids Research* database issue and the online molecular biology database collection

Daniel J Rigden, Xosé M Fernández

Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D1–D7,
https://doi.org/10.1093/nar/gkx1235
Published: 18 December 2017 **Article history**

This year's issue is again divided into eight sections that deal with (i) nucleic acid sequence and structure; (ii) protein sequence and structure; (iii) metabolic and signaling pathways; (iv) viruses, bacteria, protozoa and fungi; (v) genomes of human and model organisms; (vi) human diseases and drugs; (vii) plants and (viii) other topics, including mitochondrial databases and databases of chemical compounds.

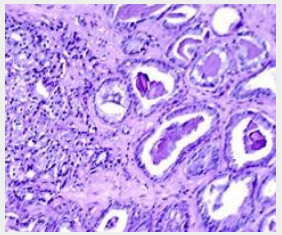
The underlying principle

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

Patient

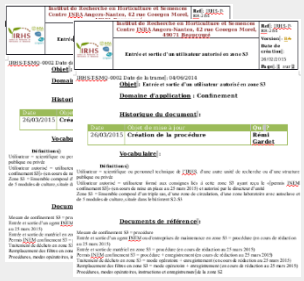
Sample

Image

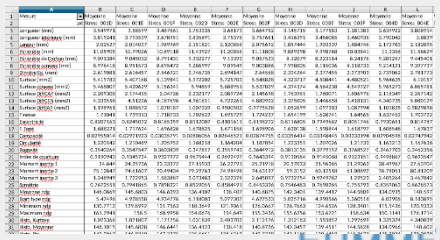


Experiment

Protocoles



Descriptors



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

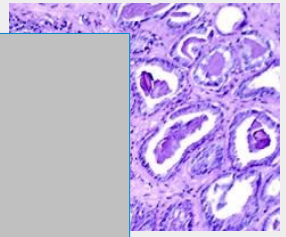
Needs

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

Patient

Sample

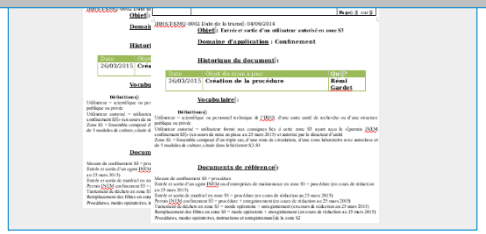
Image



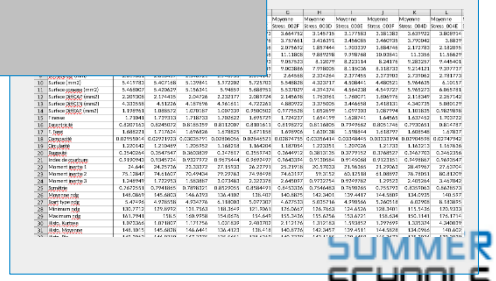
Expériment

Store information
Materialize the links

To analyze
To share
To reuse

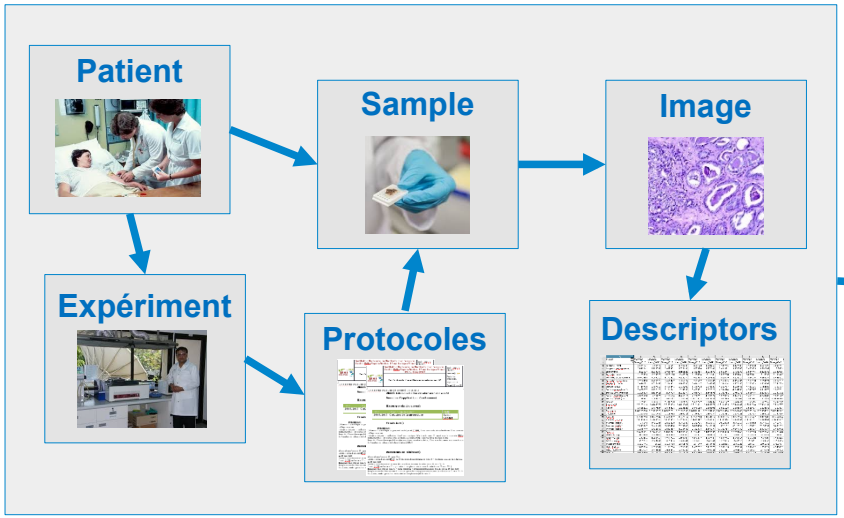


ptors

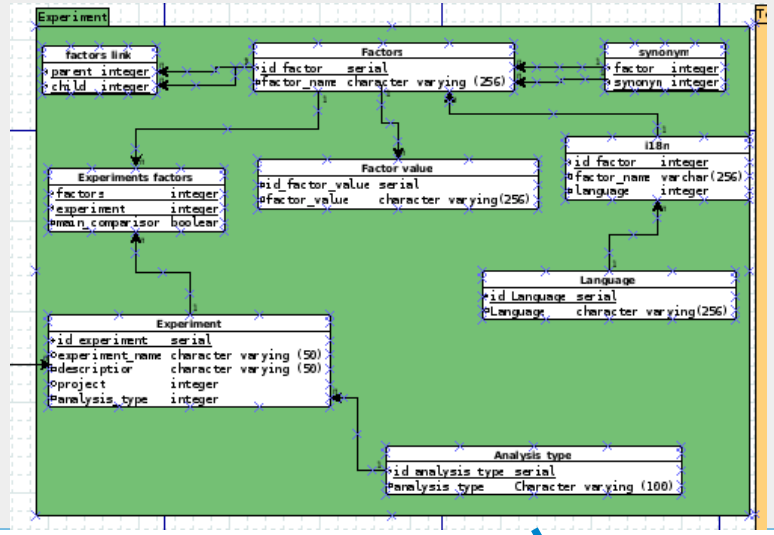


SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Classical approach

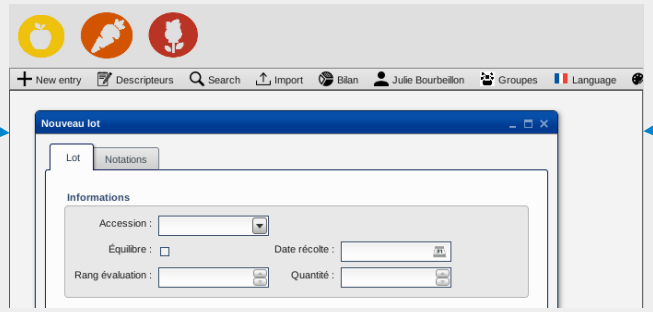


Model of the world



User

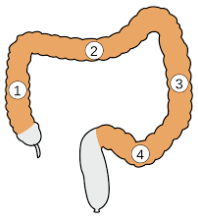
Information system



First limit : vocabulary

Samples description

Organ



Colon

colon

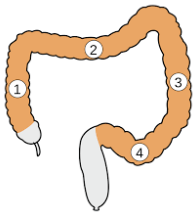
Large intestine

Cooln

Find all colon datasets ?

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Solution : controlled vocabulary



Colon

colon

Large intestine

Cooln

Controlled Vocabulary

Oesophagus

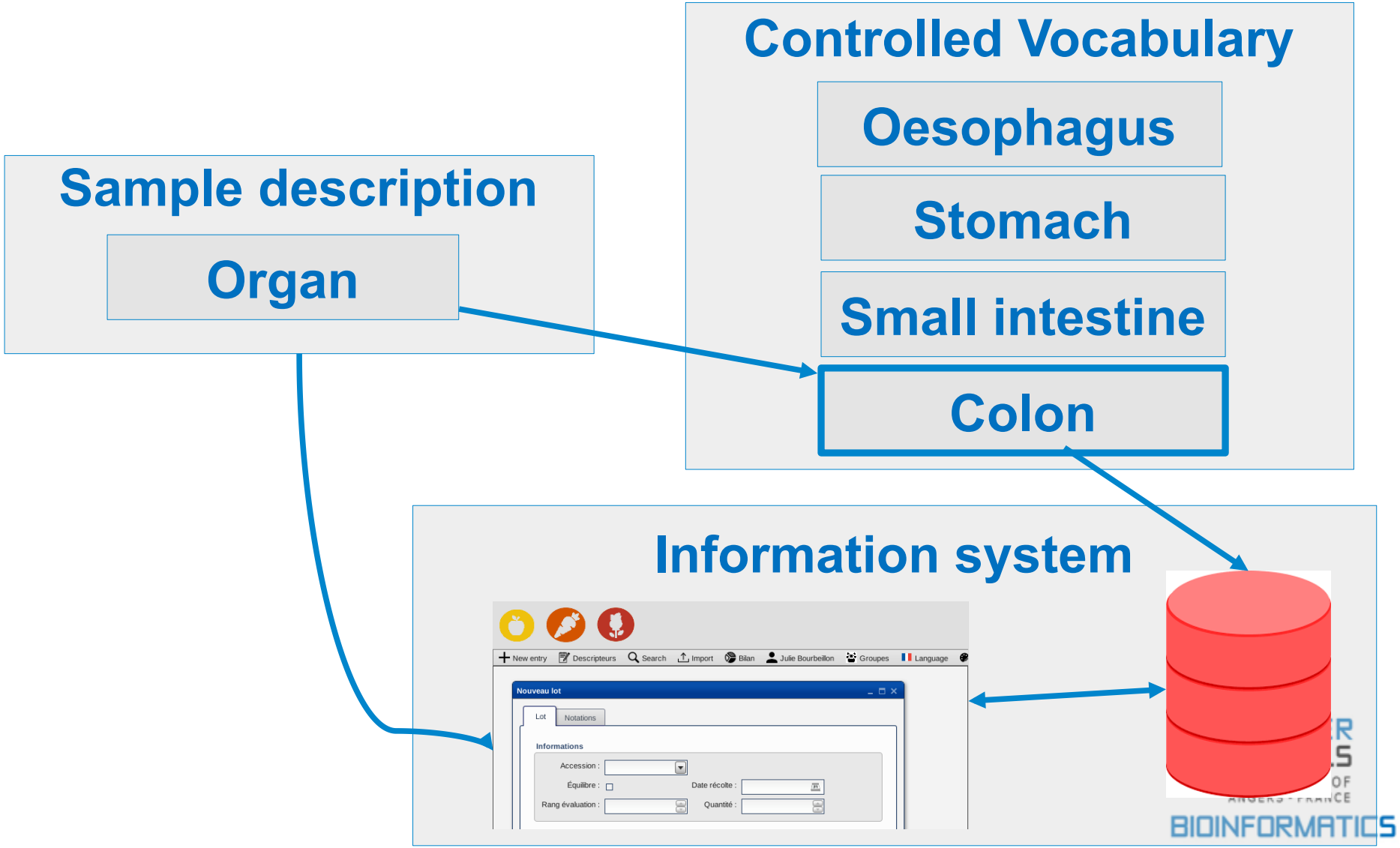
Stomach

Small intestine

Colon

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Solution : controlled vocabulary

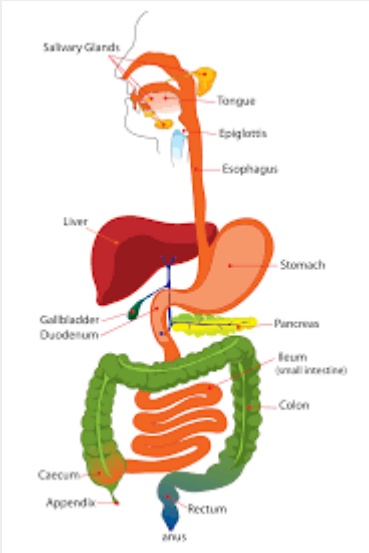


Limits of the controlled vocabulary

Controlled Vocabulary

- Oesophagus
- Stomach
- Small intertine
- Colon

Find all digestive system datasets ?



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

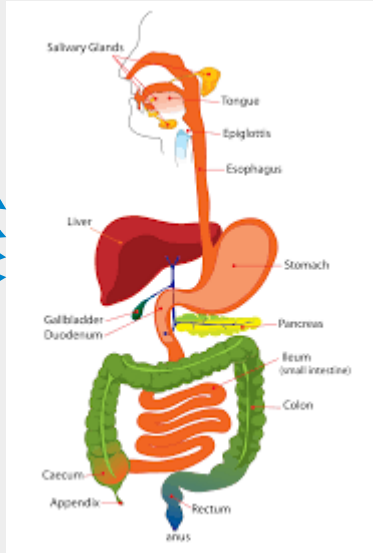
Limits of the controlled vocabulary

Controlled Vocabulary

- Oesophagus
- Stomach
- Small intestine
- Colon

Part of

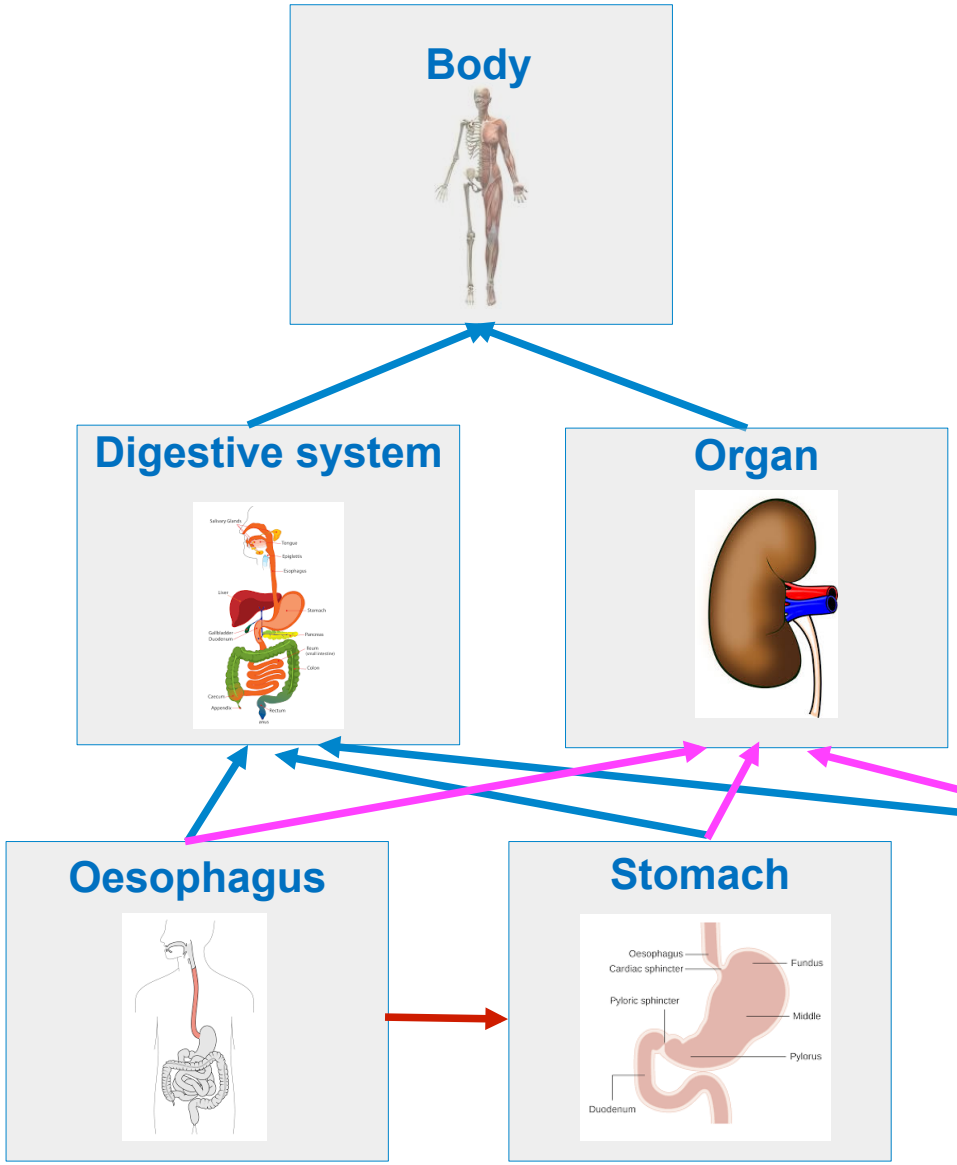
Find all digestive system datasets ?



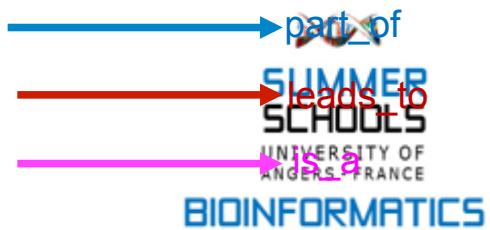
Human knowledge

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

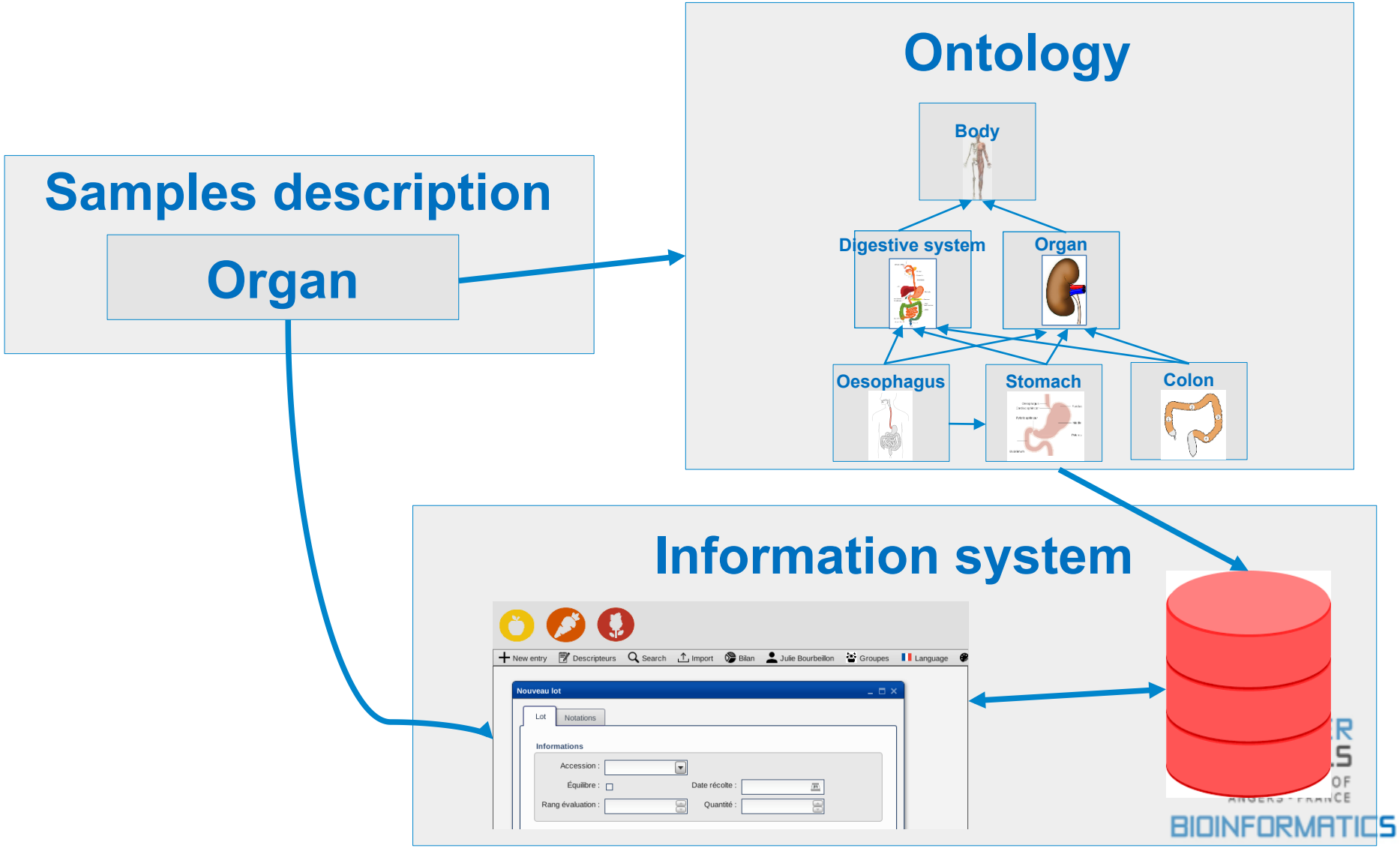
Ontology



- Directed graph
  - Nodes = concepts
    - Ex : Stomach, Organ
  - Arrows = relations
    - Ex : part_of, is_a, leads_to
- Langages
  - Ex : RDF+RDFS, OWL

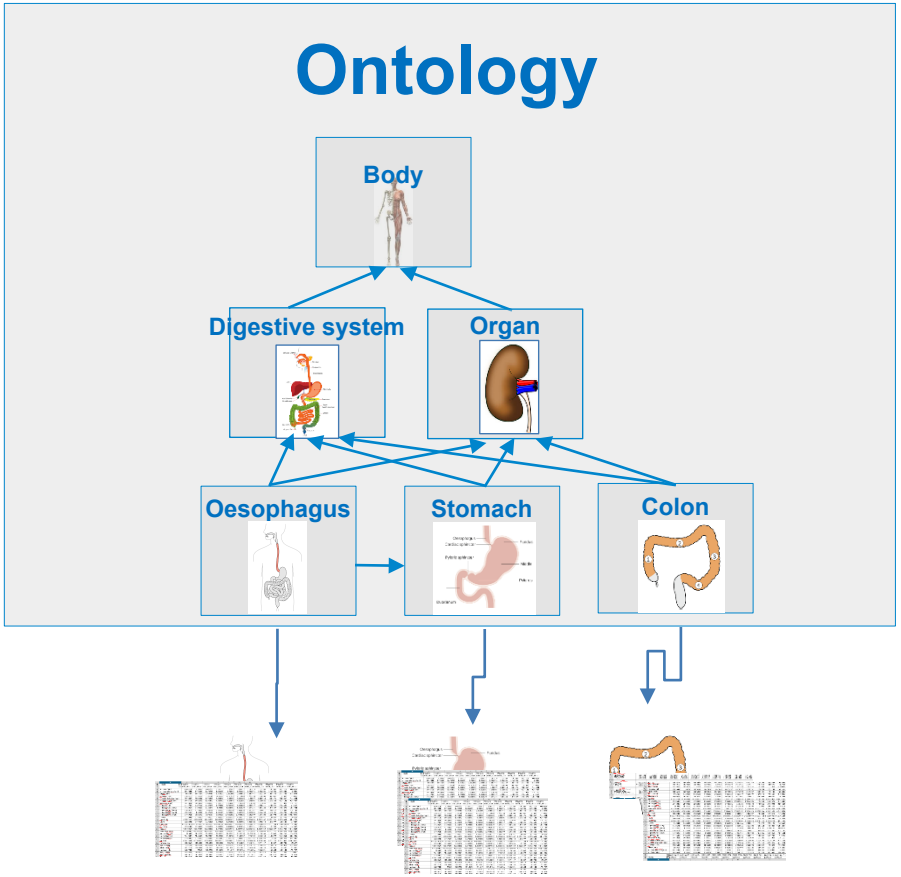


Ontology – Usage example



Ontology – Usage example

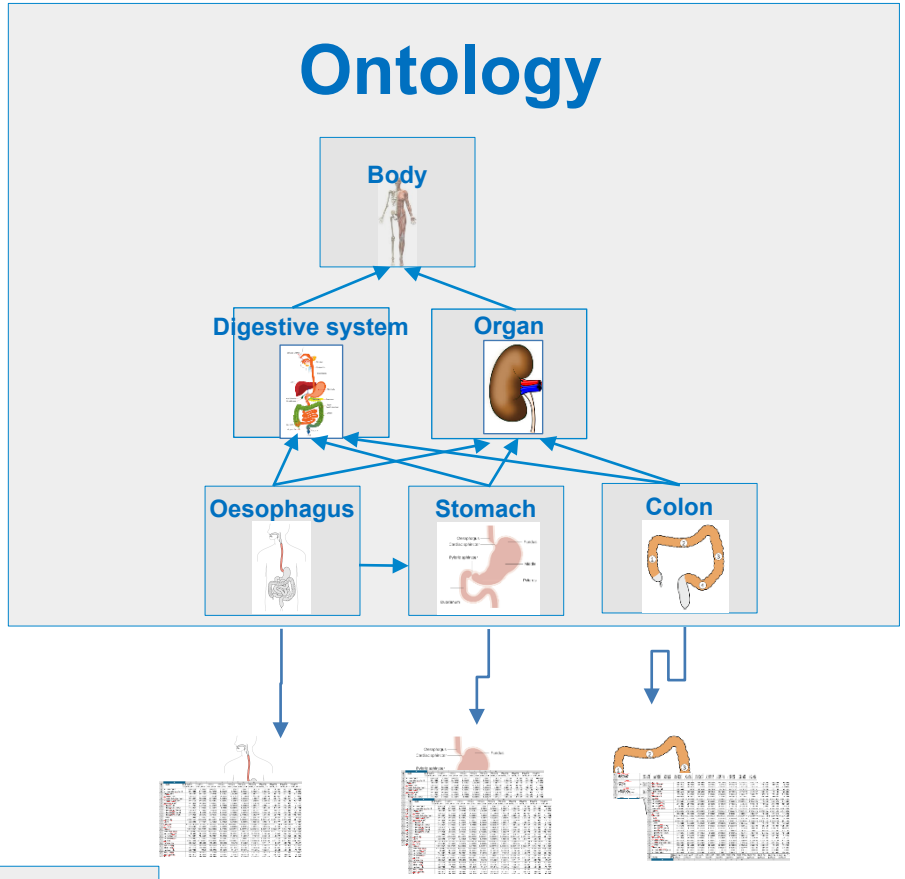
Find all digestive system datasets ?



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Ontology – Usage example

Find all digestive system datasets ?



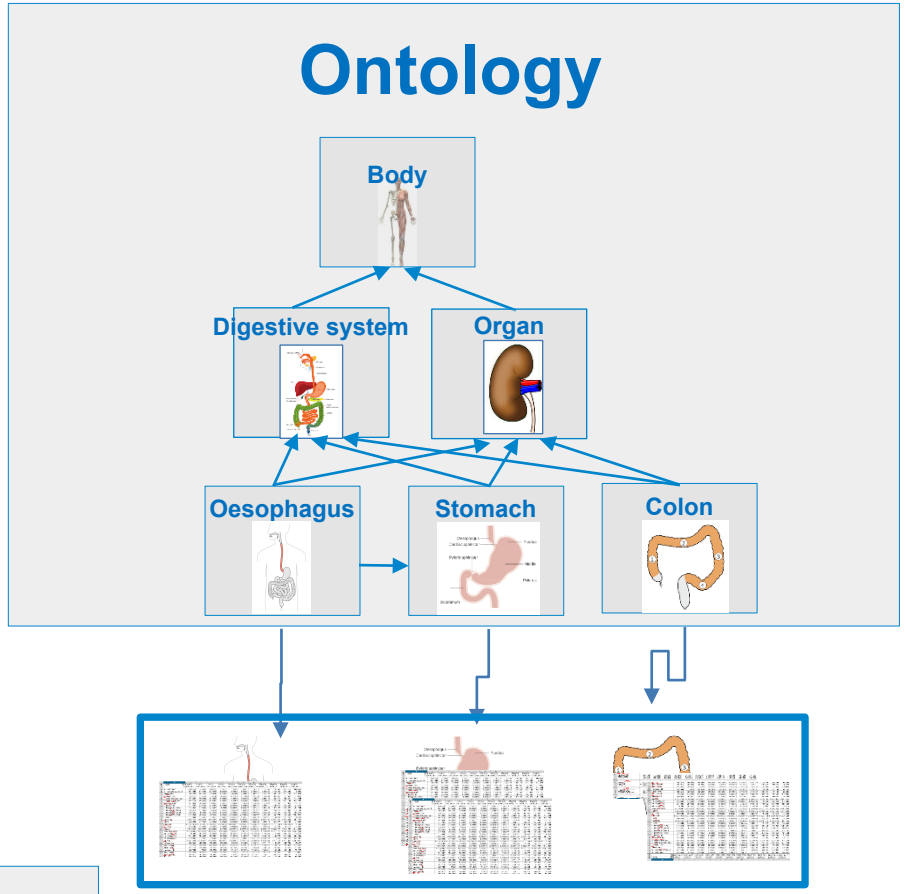
IT tool

Reasoner / Inference engine

Exploits the subsomption relationship

Ontology – Usage example

Find all digestive system datasets ?



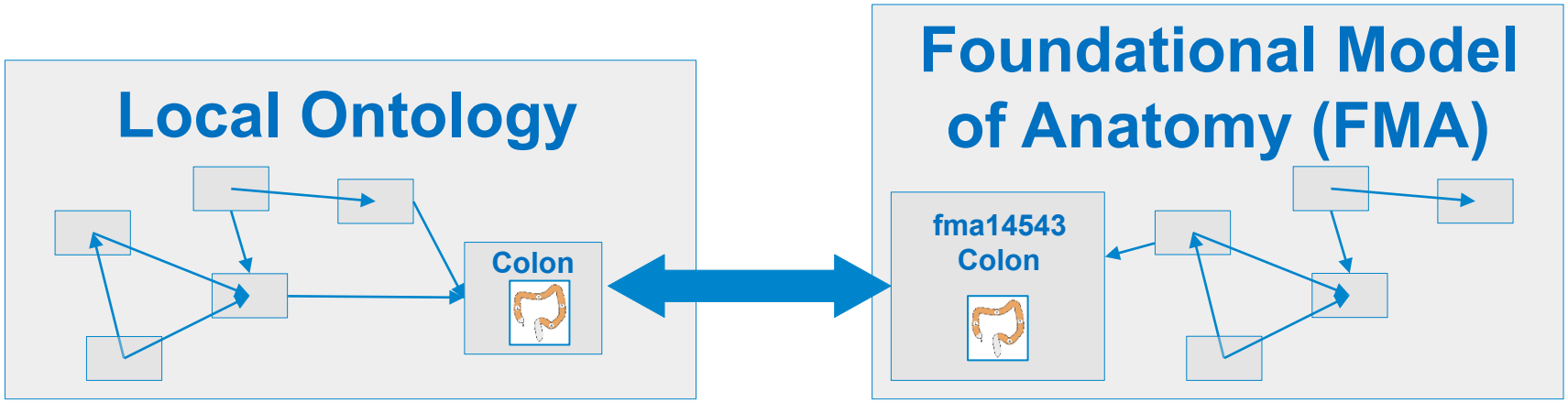
IT tool

Reasoner / Inference engine

Exploits the subsomption relationship

Ontologies typology

- Different types of Ontologies
  - Local construction
    - Close to user needs but specific
  - Reference ontologies → International efforts
    - Examples : Gene Ontology, Plant Ontology, etc.
    - Ontologies repositories : OBOFoundry, Bioportal (biomedical), Planteome (Plant)
- Links between the two



  - Ontologies interoperability
  - Associated data sharing

Use of ontologies in biological databases

- Items varying according to the database
  - Identifiers
  - Names + plenty of aliases for the same thing
  - Etc.
- → It's a mess

GeneCard data
For MAPT

Aliases for MAPT Gene

Aliases for MAPT Gene

Microtubule Associated Protein Tau [2 3 5]
G Protein Beta1/Gamma2 Subunit-Interacting Factor 1 [2 3]
Microtubule-Associated Protein Tau [2 3]
Neurofibrillary Tangle Protein [3 4]
Paired Helical Filament-Tau [3 4]
PHF-Tau [3 4]
MAPTL [3 4]
MTBT1 [3 4]
TAU [3 4]
Protein Phosphatase 1, Regulatory Subunit 103 [3]
Regulatory Subunit 103 [2]
Protein Phosphatase 1 [2]
Isoform 4 [2]
PPP1R103 [3]
FTDP-17 [3]
MTBT2 [3]
DDPAC [3]
MSTD [3]
PPND [3]

External Ids for MAPT Gene

HGNC: 6893 Entrez Gene: 4137 Ensembl: ENSG00000186868 OMIM: 157140 UniProtKB: P10636

Previous HGNC Symbols for MAPT Gene

DDPAC, MAPTL

Previous GeneCards Identifiers for MAPT Gene

GC17P043566, GC17P046217, GC17P043982, GC17P044447, GC17P041327, GC17P039635, GC17P043971

Export aliases for MAPT gene to outside databases

Use of ontologies in biological databases

- Items which are identical across databases
  - GO terms
  - NCBI taxonomy
  - Etc...

UniProt
For MAPT

Organism [i]	Homo sapiens (Human)
Taxonomic identifier [i]	9606 [NCBI]

GeneCards
For MAPT

Gene Ontology (GO) - Molecular Function for MAPT Gene

Filter: (2 results)

GO ID	Qualified GO term
GO:0015631	tubulin binding
GO:0019904	protein domain specific binding

NCBI Gene
For MAPT

Official Symbol	MAPT provided by HGNC
Official Full Name	microtubule associated protein tau provided by HGNC
Primary source	HGNC:HGNC:6893
See related	Ensembl:ENSG00000186868 HPRD:01142; MIM:157140; Vega:OTTHUMG00000168833
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; E
Also known as	TAU; MSTD; PPND; DDPAC; MAPTL; MTBT1; MTBT2; FTDP-17; PPP1R103

InterPro
For MAPT

Molecular Function

GO:0008017 microtubule binding
GO:0015631 tubulin binding

SUMMER SCHOOLS UNIVERSITY OF ANGERS - FRANCE
ANGERS - FRANCE BIOINFORMATICS

Use of ontologies in biological databases

- Items which are identical across databases
  - GO terms
  - NCBI taxonomy
  - Etc...

UniProt
For MAPT

Based on Ontologies = Annotation

GeneCards
For MAPT

ntology (GO) - Molecular Function for MAPT Gene

er: (2 results)

O ID	Qualified GO term
D:0015631	tubulin binding
D:0019904	protein domain specific binding

NCBI Gene
For MAPT

Official Symbol	MAPT provided by HGNC
Official Full Name	microtubule associated protein tau provided by HGNC
Primary source	HGNC:HGNC:6893
See related	Ensembl:ENSG00000186868 HPRD:01142; MIM:157140; Vega:OTTHUMG00000168833
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; E
Also known as	TAU; MSTD; PPND; DDPAC; MAPTL; MTBT1; MTBT2; FTDP-17; PPP1R103

InterPro
For MAPT

Molecular Function

GO:0008017 microtubule binding
GO:0015631 tubulin binding

SUMMER SCHOOLS UNIVERSITY OF ANGERS - FRANCE
ANGERS - FRANCE BIOINFORMATICS

Gene ontology

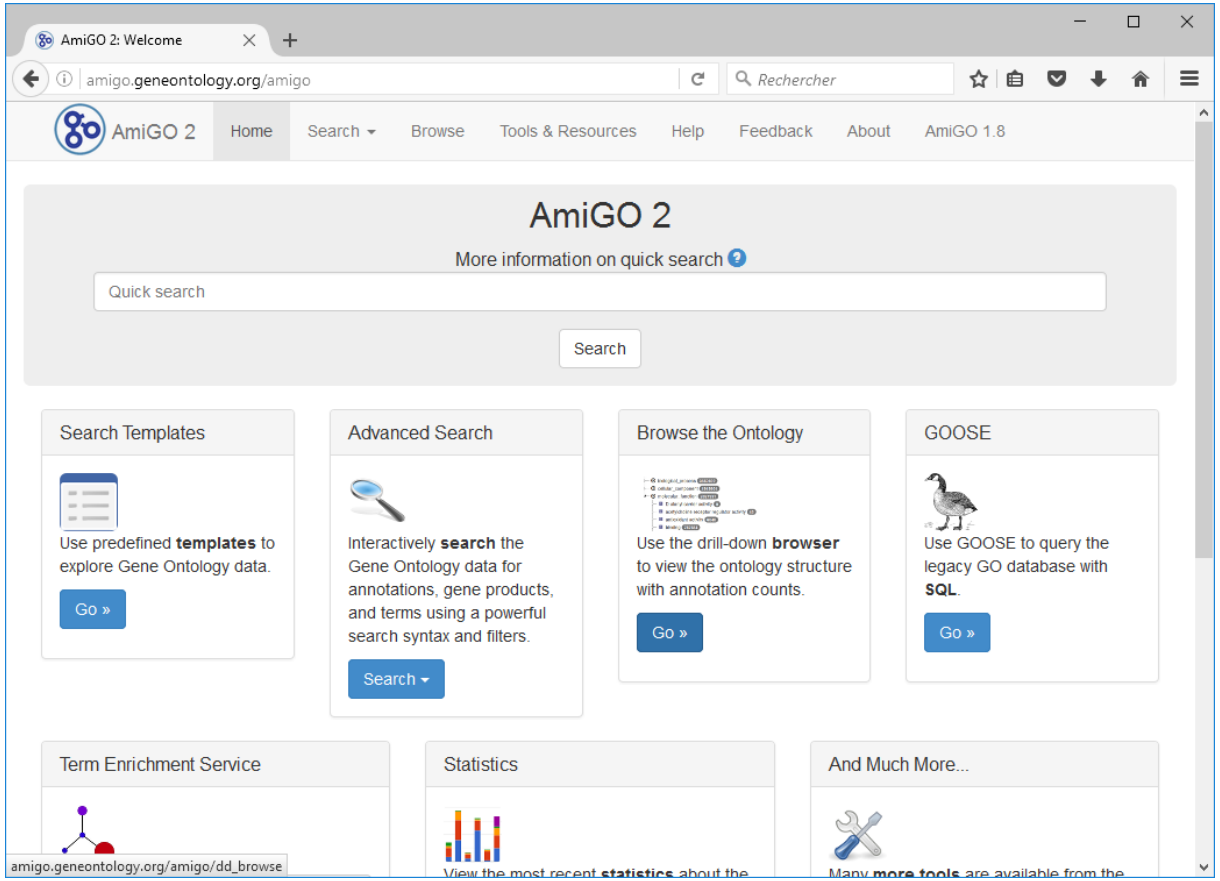
- GO : Gene ontology
  - http://geneontology.org
  - The biological ontology the most widely used
  - Developed since 2000
  - Goal : generic ontology to describe the role of genes and proteins

Practice

- Explore The Gene Ontology

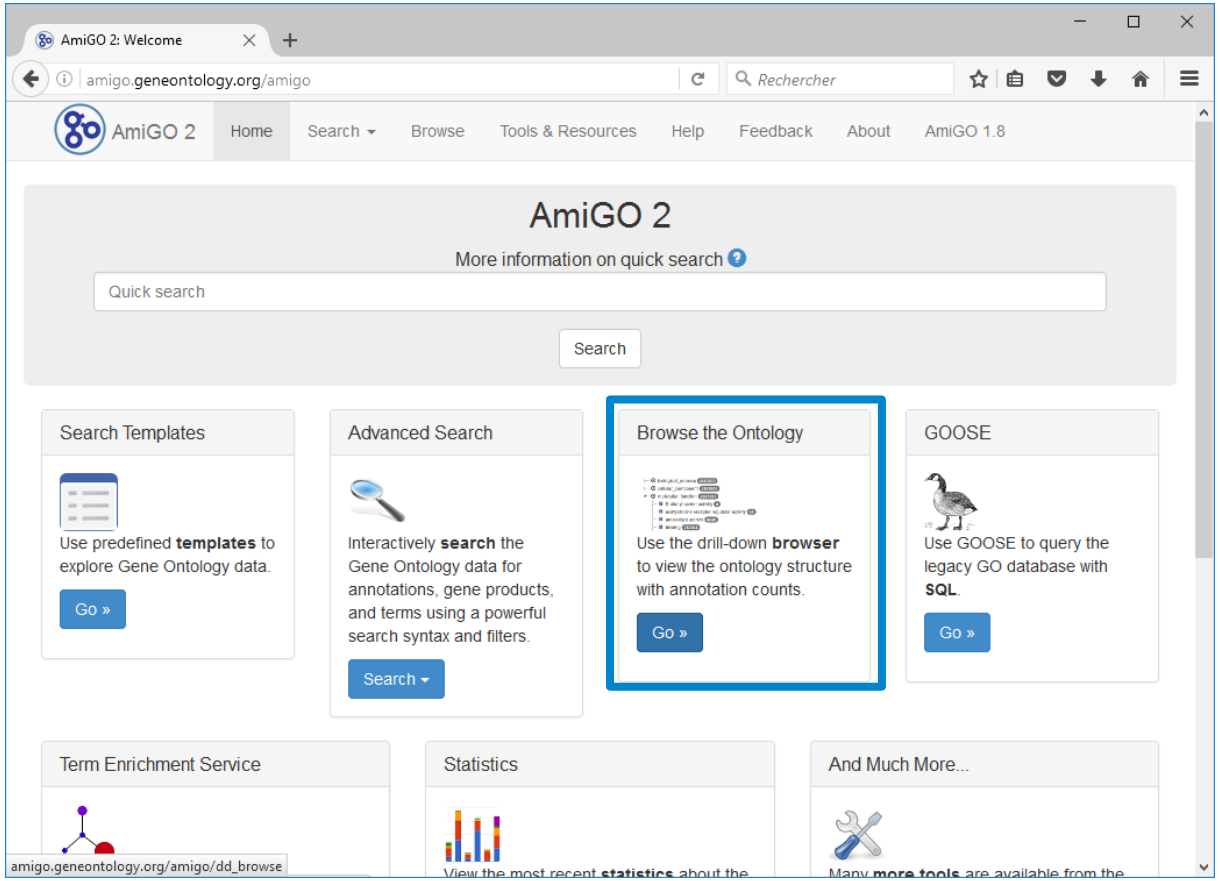
GO Tools

- http://amigo.geneontology.org



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Explore the GO



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

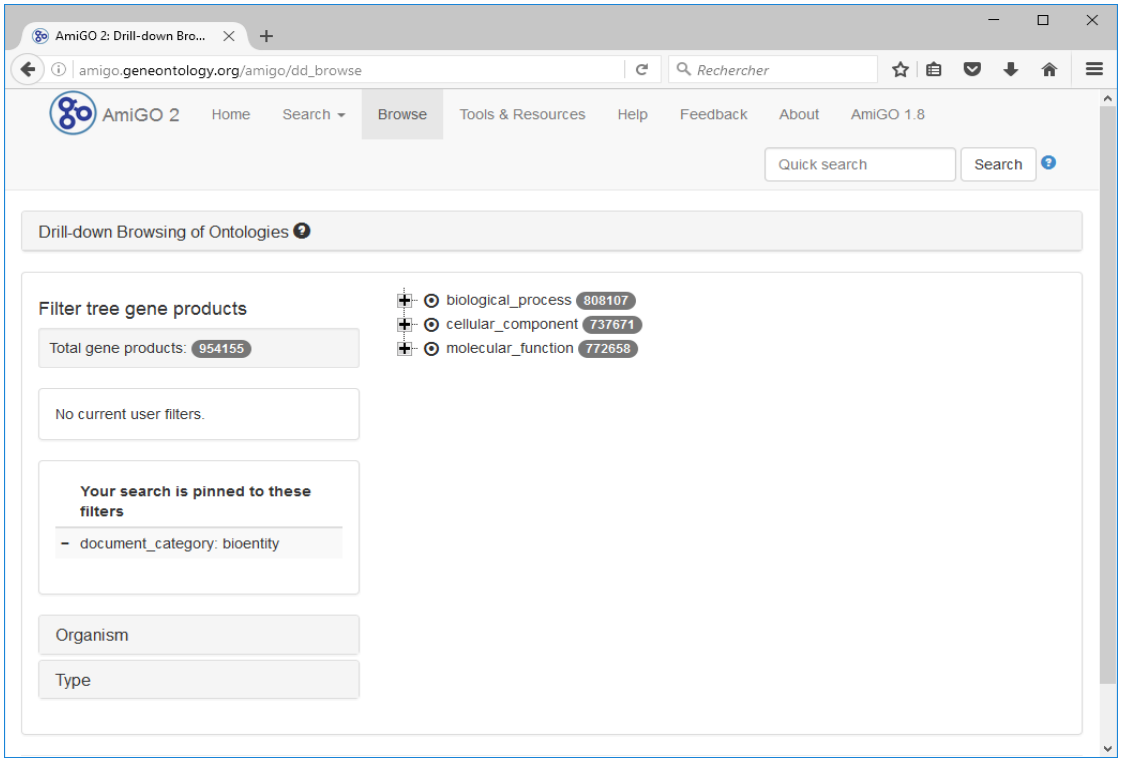
Explore the GO

- How many top level items are there ?
- What are their names ?
- What do they represent ?

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Explore the GO

- How many to level concepts are there ?
  - 3
- What are their names ?



Explore the GO

- What do they represent ?
  - Three subontologies (sections)
    - biological process : P (the more difficult to predict)
    - molecular function : F (biochemistry function)
    - cellular component : C (subcellular location)

GO : biological process

- Definition
  - biological objective to which a gene or gene product contributes
- Complementary information
  - achieved by one or more ordered assemblies of molecular functions
  - involves a collection of molecular events with a defined beginning and end
  - involves generally a chemical or physical transformation at the level of granularity of the cell or organism

GO : molecular function

- Definition
  - Biochemical activity of a gene product
- Concepts
  - enzymatic reaction, or other activities such as the specific binding to a ligand or macromolecular structures
- Limits
  - Only a potential of the gene product to perform an activity, that would occur in the correct context
  - Only what is done at a molecular level of granularity,
  - Not where or when the event actually occurs, or what purpose it serves the organism

GO: cellular component

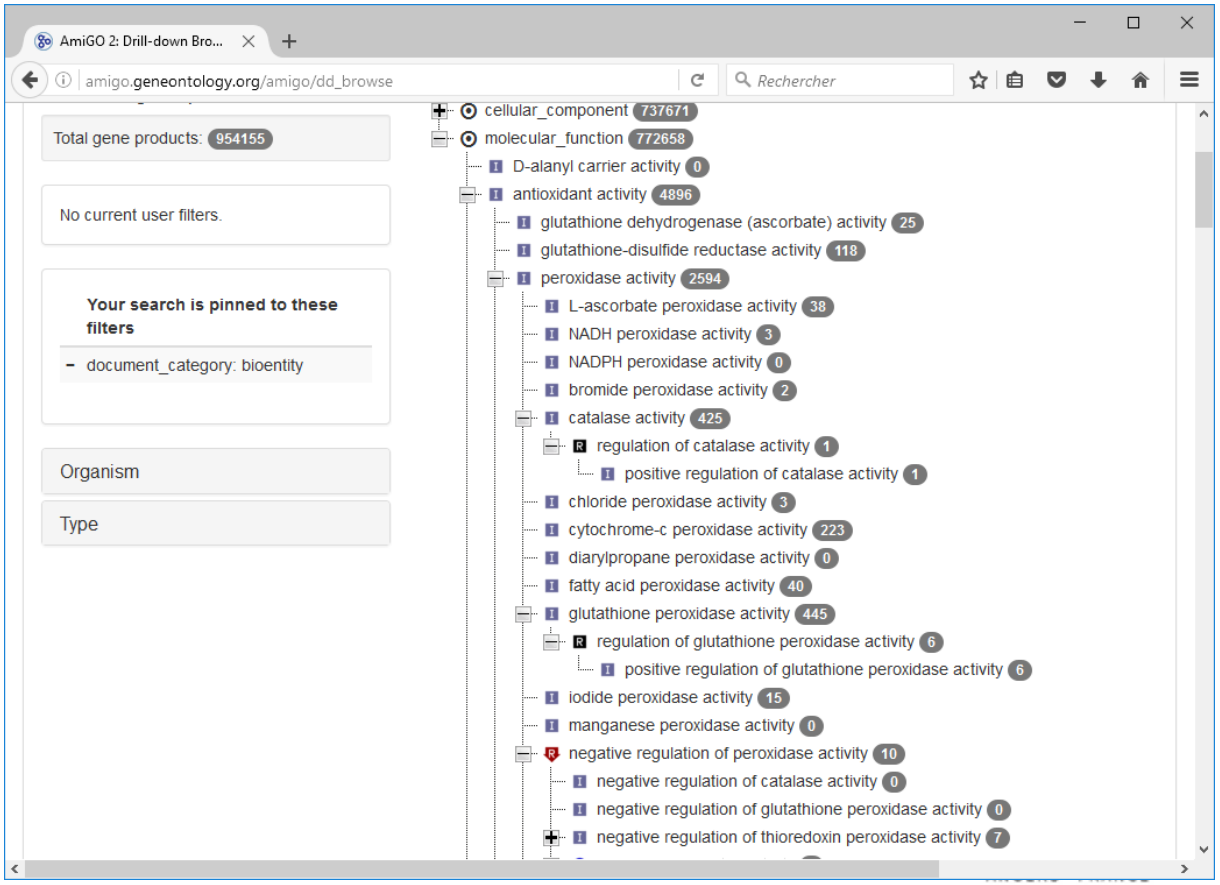
- Definition
  - Location inside or outside of the cell where a gene product is active
- Concepts
  - cytoplasm, cellular organelles such as the nucleus or mitochondrion, for protein complexes such as the proteasome, and for the extracellular region

Explore the GO

- Drill down the ontology
- How many levels did you go ?

Explore the GO

- How many levels did you go ?
  - Variable
  - → Problem



Explore the GO

- Click on one item
- What do you see ?

Explore the GO

- Click on one item
- What do you see ?

Term	Golgi apparatus part
Definition	Any constituent part of the Golgi apparatus, a compound membranous cytoplasmic organelle of eukaryotic cells, consisting of flattened, ribosome-free vesicles arranged in a more or less regular stack.
Ontology source	cellular_component
Ontology ID space	GO
Synonyms	Golgi component, Golgi subcomponent
Alt ID	n/a
Gene products	**retrieve gene products annotated to this term for this filter set**

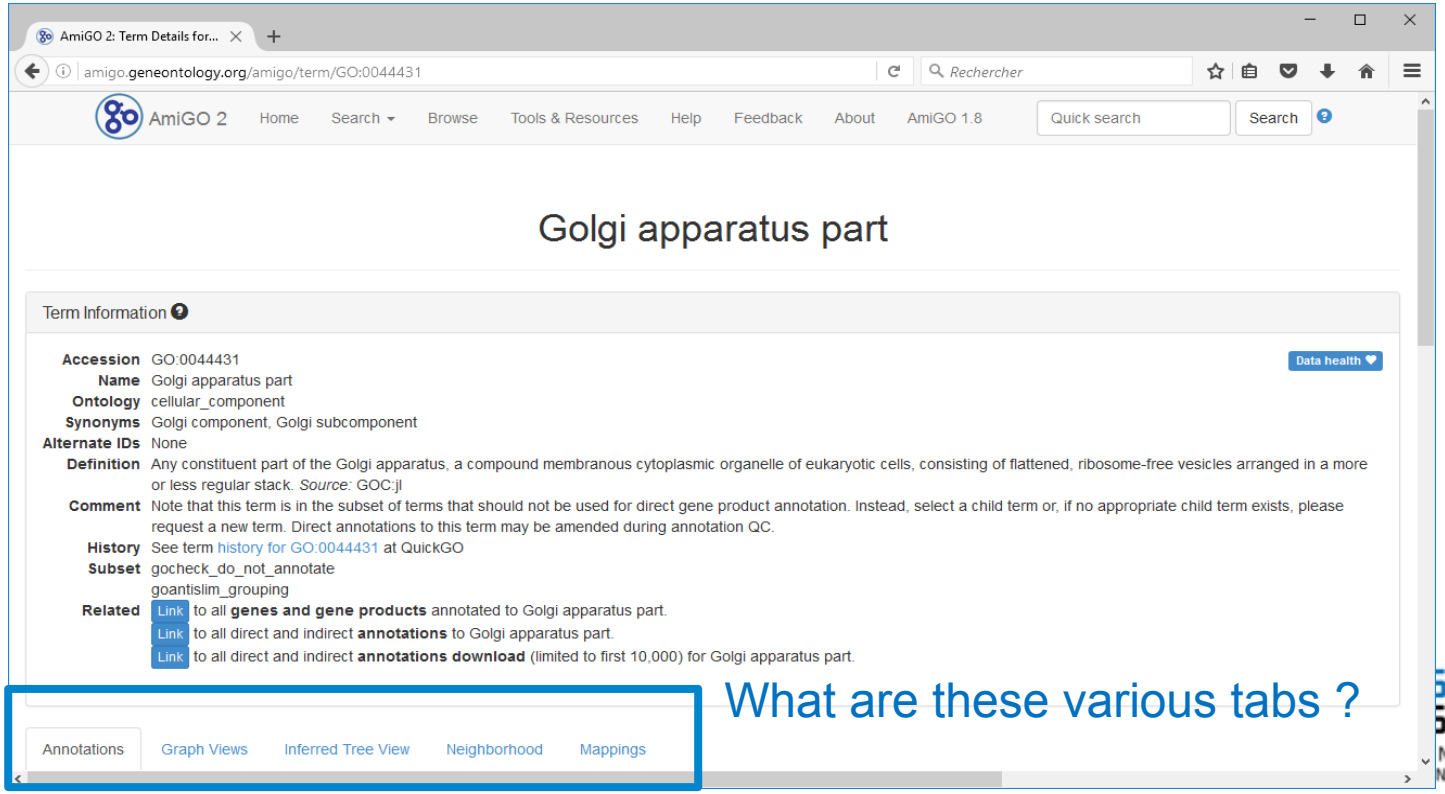
SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Explore the GO

- Click on the term name
- What do you get ?

Explore the GO

- Click on the term name
- What do you get ?



What are these various tabs ?

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Explore the GO

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

✦ Gene products annotated with this term

AmiGO 2: Term Details for...

amigo.geneontology.org/amigo/term/GO:0044431

Filter results

Total annotations: 22620

No current user filters.

Your search is pinned to these filters

- document_category: annotation
- regulates_closure: GO:0044431

Ontology (aspect)

Organism

Evidence

GO class

GO class (direct)

Annotation qualifier

Total annotations: 22620; showing: 1-10

Results count 10

«First <Prev Next> Last» Download

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with
AgaP_AGAP002175			integral component of Golgi membrane		GO_Central	Anopheles gambiae	IBA	PANTHER:PTN000363
AgaP_AGAP007654	AGAP007654-PA		trans-Golgi network		GO_Central	Anopheles gambiae	IBA	PANTHER:PTN000636
PITG_03683			Golgi-associated vesicle membrane		GO_Central	Phytophthora infestans T30-4	IBA	PANTHER:PTN000256
AgaP_AGAP007605			trans-Golgi network		GO_Central	Anopheles gambiae	IBA	PANTHER:PTN000361
AgaP_AGAP007586	AGAP007586-PA		Golgi transport complex		GO_Central	Anopheles gambiae	IBA	PANTHER:PTN000487
SYS1	Protein SYS1 homolog isoform a		trans-Golgi network		GO_Central	Macaca mulatta	IBA	PANTHER:PTN000310
SYS1	Protein SYS1 homolog isoform a		integral component of Golgi membrane		GO_Central	Macaca mulatta	IBA	PANTHER:PTN000310

BIOINFORMATICS

Explore the GO

✦ Filter by organism

Ontology (aspect)

Organism

- (24160) Eukaryota
- (14239) Metazoa
- (10720) Vertebrata
- (8941) Mammalia
- (4058) Fungi
- (3900) Viridiplantae
- (3127) Homo sapiens
- (1209) Rattus norvegicus
- (778) Arabidopsis thaliana
- (759) Bos taurus
- (726) Mus musculus
- (664) Glycine max
- (648) Canis lupus

			NRRL 181		
ACLA_007410	Golgi apparatus part	AspGD	Aspergillus clavatus NRRL 1	IEA	AspGD:ASPL00000067
ACLA_007410	Golgi cis cisterna	AspGD	Aspergillus clavatus NRRL 1	IEA	AspGD:ASPL00000067
Afu1g09090	Golgi lumen	AspGD	Aspergillus fumigatus	IEA	SGD:S000002168
kexA	trans-Golgi network	AspGD	Aspergillus fumigatus	IEA	SGD:S000003171
Afu1g07740	trans-Golgi network	AspGD	Aspergillus fumigatus	IEA	SGD:S000000690
Afu1g07420	COPII-coated ER to Golgi transport vesicle	AspGD	Aspergillus fumigatus	IEA	SGD:S000001266
AN2752	Golgi membrane	AspGD	Aspergillus nidulans	IEA	PomBase:SPBC19C7.1

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Explore the GO

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

✦ Filter by « evidence »

- regulates_closure: GO:0044431

Ontology (aspect)

Organism

Evidence

- (2253) traceable author statement used in manual assertion
- (391) evidence used in automatic assertion
- (288) experimental evidence
- (263) direct assay evidence used in manual assertion

B3GAT1	Galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase 1	Golgi membrane	UniProt	Homo sapiens	ISS	UniP
B3GAT1	Galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase 1	Golgi membrane	GO_Central	Homo sapiens	IBA	PAN
B3GAT1	Galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase 1	Golgi membrane	Reactome	Homo sapiens	TAS	
TRAPPC2L	Trafficking protein particle complex subunit 2-like protein	Golgi membrane	GOC	Homo sapiens	IEA	GO:
TRAPPC2L	Trafficking protein particle complex subunit 2-like protein	COPII vesicle coating	Reactome	Homo sapiens	TAS	

MER OLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Explore the GO

✦ Results

Filter results

Total annotations: 3127

User filters

+ taxon_subset_closure_label: Homo sapiens

Your search is pinned to these filters

- document_category: annotation
- regulates_closure: GO:0044431

Ontology (aspect)

Total annotations: 3127; showing: 1-10
Results count 10

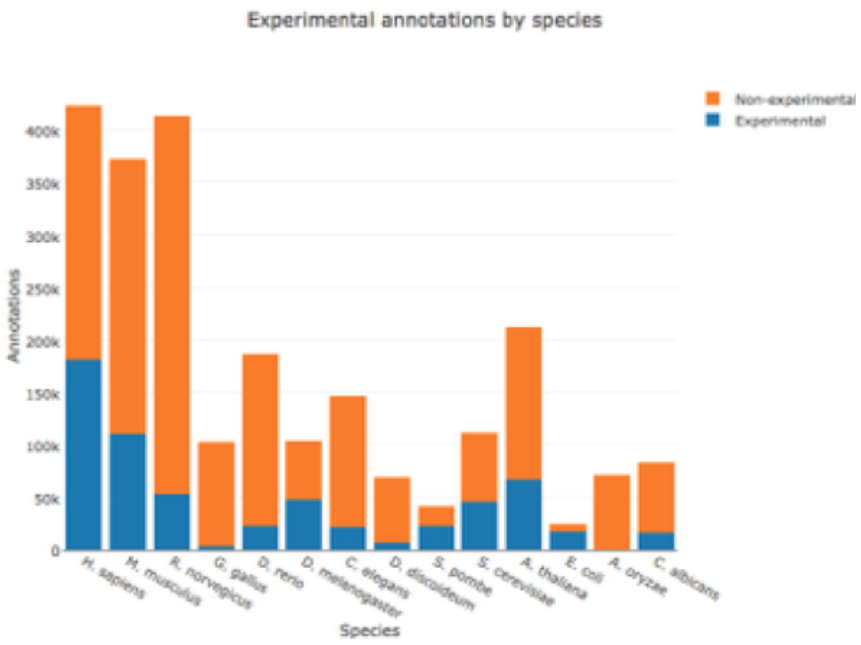
«First <Prev Next> Last» Download

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence
CUX2	Homeobox protein cut-like 2		Golgi membrane		GO_Central	Homo sapiens	IBA
SYBU	Syntabulin		Golgi membrane		UniProt	Homo sapiens	IEA
TMEM132A	Transmembrane protein 132A		Golgi membrane		UniProt	Homo sapiens	IEA
PDCD10	Programmed cell death protein 10		Golgi membrane		UniProt	Homo sapiens	IEA
STX18	Syntaxin-18		Golgi membrane		UniProt	Homo sapiens	IEA
B3GAT1	Galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase 1		Golgi membrane		UniProt	Homo sapiens	ISS
B3GAT1	Galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase 1		Golgi membrane		GO_Central	Homo sapiens	IBA

Explore the GO

✦ Why filter by « evidence » ??

http://geneontology.org/page/guide-go-evidence-codes



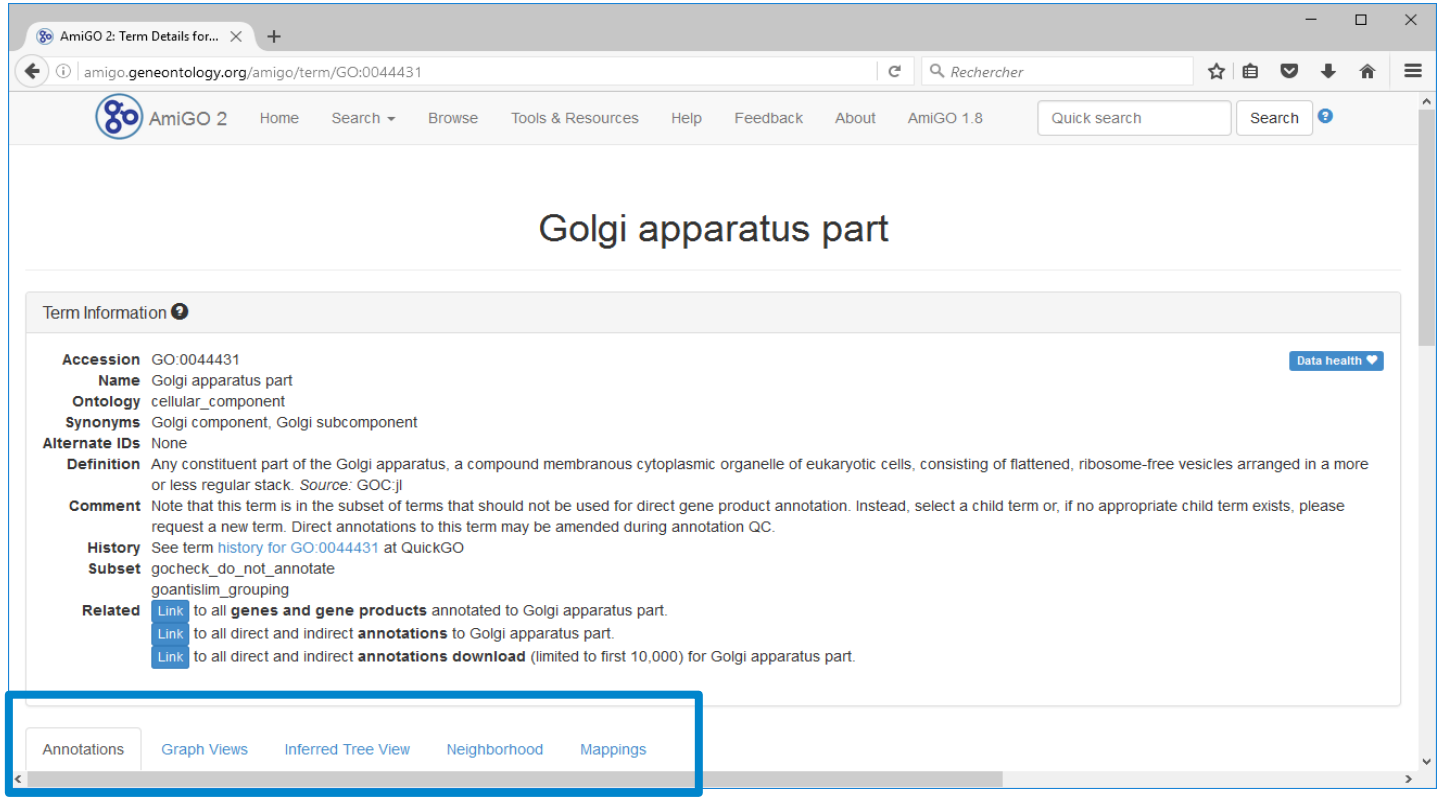
- Inferred from Experiment (EXP)
- Inferred from Direct Assay (IDA)
- Inferred from Physical Interaction (IPI)
- Inferred from Mutant Phenotype (IMP)
- Inferred from Genetic Interaction (IGI)
- Inferred from Expression Pattern (IEP)

Use of the computational analysis evidence codes indicates tha data as described in the cited reference. The evidence codes in *Analysis evidence codes* are:

- Inferred from Sequence or structural Similarity (ISS)
- Inferred from Sequence Orthology (ISO)
- Inferred from Sequence Alignment (ISA)
- Inferred from Sequence Model (ISM)
- Inferred from Genomic Context (IGC)
- Inferred from Biological aspect of Ancestor (IBA)
- Inferred from Biological aspect of Descendant (IBD)
- Inferred from Key Residues (IKR)
- Inferred from Rapid Divergence(IRD)
- Inferred from Reviewed Computational Analysis (RCA)

Explore the GO

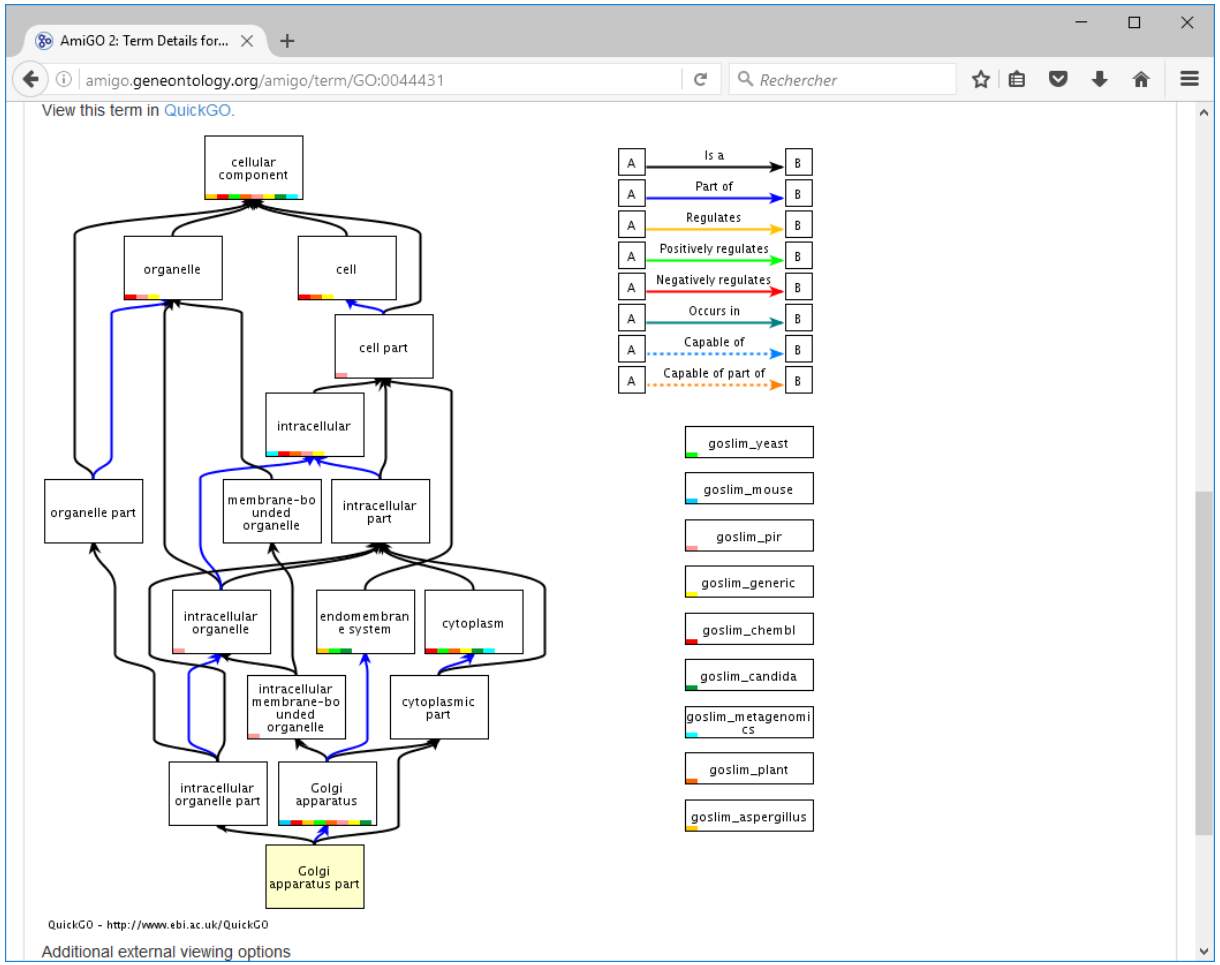
- Click on the term name
- What do you get ?



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Explore the GO

- Ontology graph centered on the term



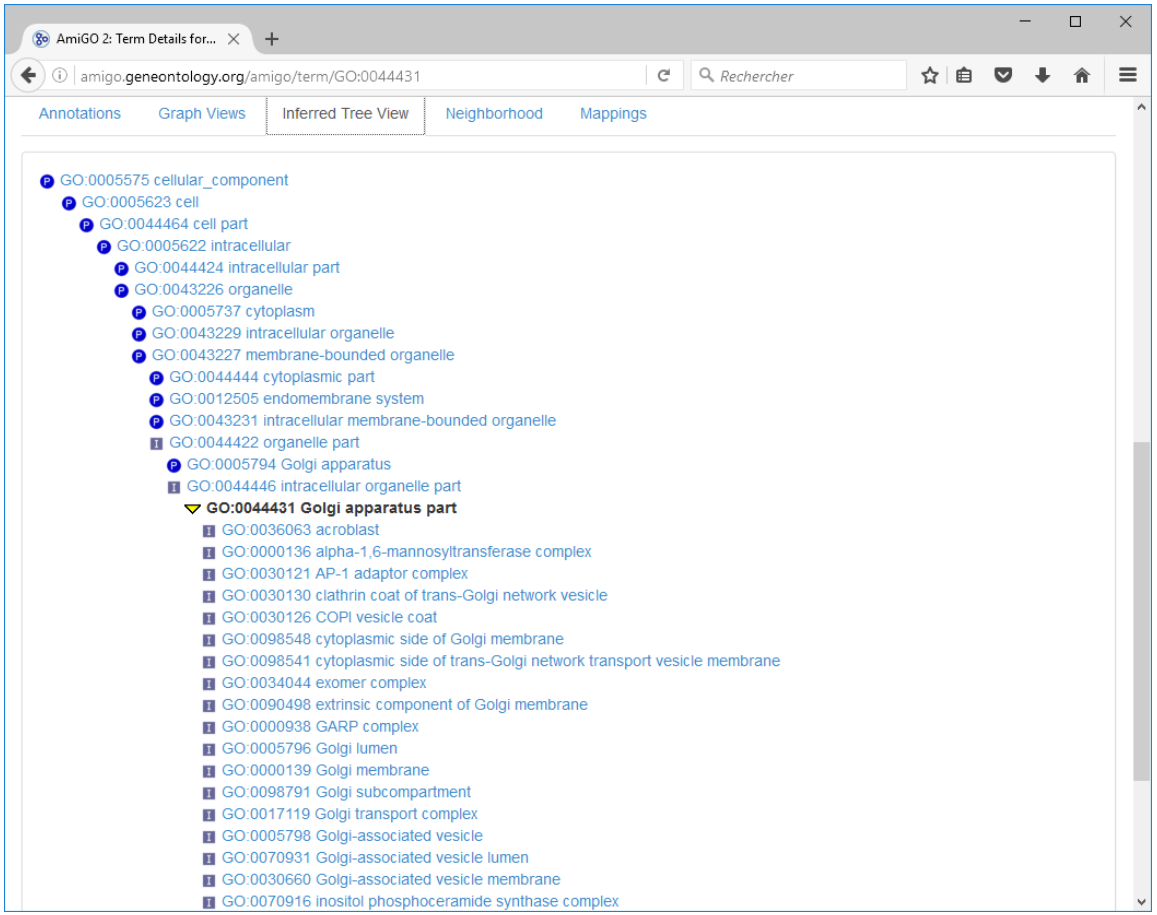
SUMMER SCHOOLS UNIVERSITY OF ANGERS - FRANCE

SUMMER SCHOOLS UNIVERSITY OF ANGERS - FRANCE BIOINFORMATICS

Explore the GO

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

✦ The term in the ontology tree



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Explore the GO

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE

✦ The neighbours of the term in the graph

AmiGO 2: Term Details for...

amigo.geneontology.org/amigo/term/GO:0044431

Rechercher

Annotations | Graph Views | Inferred Tree View | Neighborhood | Mappings

Parents of Golgi apparatus part (GO:0044431)

subject	relation	object
Golgi apparatus part	is_a	cellular_component (GO:0005575)
Golgi apparatus part	part_of	Golgi apparatus (GO:0005794)
Golgi apparatus part	is_a	intracellular organelle part (GO:0044446)
Golgi apparatus part	is_a	cytoplasmic part (GO:0044444)

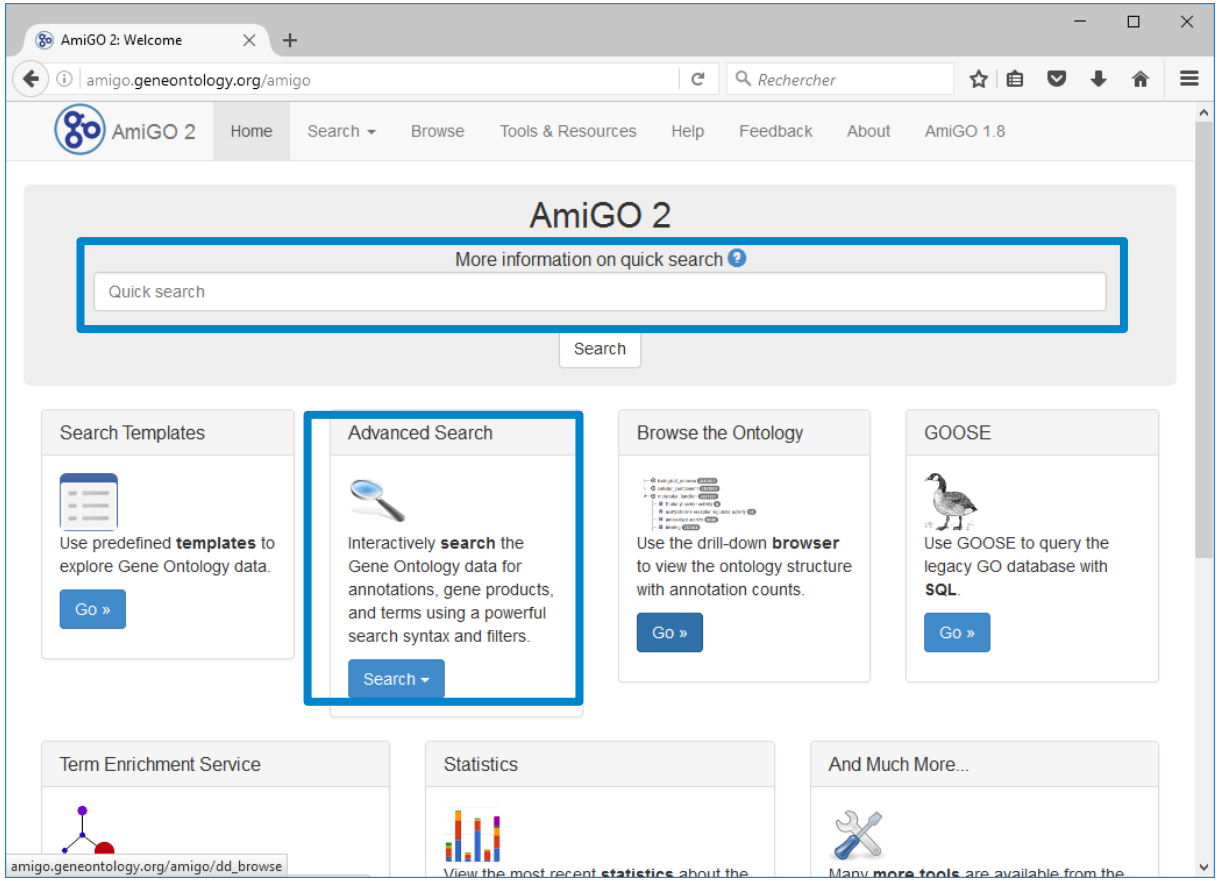
Children of Golgi apparatus part (GO:0044431)

subject	relation	object
intrinsic component of Golgi membrane (GO:0031228)	is_a	Golgi apparatus part
Golgi-associated vesicle membrane (GO:0030660)	is_a	Golgi apparatus part
cytoplasmic side of trans-Golgi network transport vesicle membrane (GO:0098541)	is_a	Golgi apparatus part
acroblast (GO:0036063)	is_a	Golgi apparatus part
AP-1 adaptor complex (GO:0030121)	is_a	Golgi apparatus part
Golgi lumen (GO:0005796)	is_a	Golgi apparatus part
RIC1-RGP1 guanyl-nucleotide exchange factor complex (GO:0034066)	is_a	Golgi apparatus part
Golgi subcompartment (GO:0098791)	is_a	Golgi apparatus part
clathrin coat of trans-Golgi network vesicle (GO:0030130)	is_a	Golgi apparatus part

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Search the GO

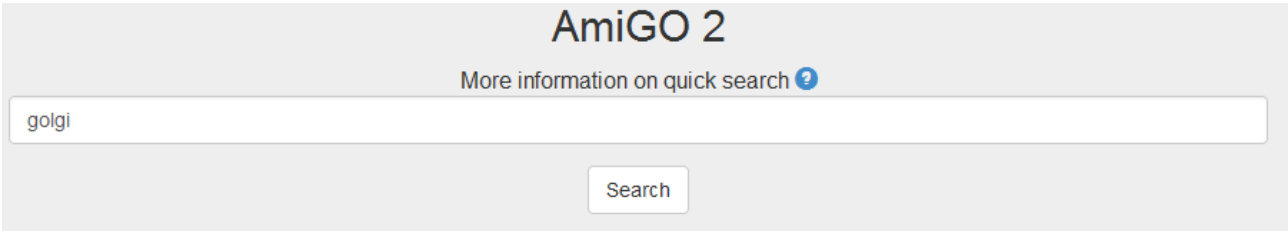
✦ Browsing = cumbersome → search



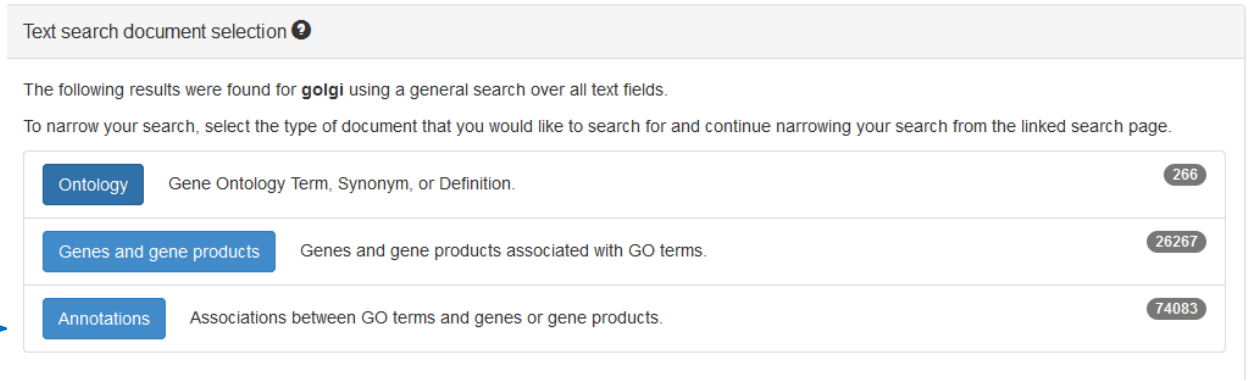
SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Search the GO

- Do a quick search



- Select ontology results



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Search the GO

✦ Ontology search results

AmiGO 2: Search

amigo.geneontology.org/amigo/search/ontology?q=golgi

AmiGO 2 | Home | Search | Browse | Tools & Resources | Help | Feedback | About | AmiGO 1.8

Quick search | Search

Information about **Ontology** search

Filter results

Total term(s): 242

golgi

User filters

+ idspace: GO

+ is_obsolete: false

Your search is pinned to these filters

– document_category: ontology_class

Ontology source

Total term(s): 242; showing: 1-10

Results count 10

«First | <Prev | Next> | Last»

Custom DL (up to 100000) | Bookmark

Term	Definition	Ontology source	Ontology ID space	Synonyms	Alt ID
Golgi membrane fusion	The joining of two lipid bilayers that surround the Golgi apparatus to form a single Golgi membrane.	biological_process	GO	membrane fusion involved in Golgi reassembly post-mitotic fusion of Golgi membranes	
release of sequestered calcium ion into cytosol by Golgi	The directed movement of calcium ions (Ca2+) out of the Golgi apparatus into the cytosol.	biological_process	GO	Golgi calcium ion export	
retrograte transporter complex, Golgi to ER	Transporter complex that recognises, binds and returns endoplasmic reticulum (ER) resident proteins more...	cellular_component	GO	retrograte receptor complex, Golgi to ER retrograte receptor complex, Golgi	

SUMMER SCHOOLS UNIVERSITY OF ANGERS - FRANCE

BIOINFORMATICS

So far regarding ontology use

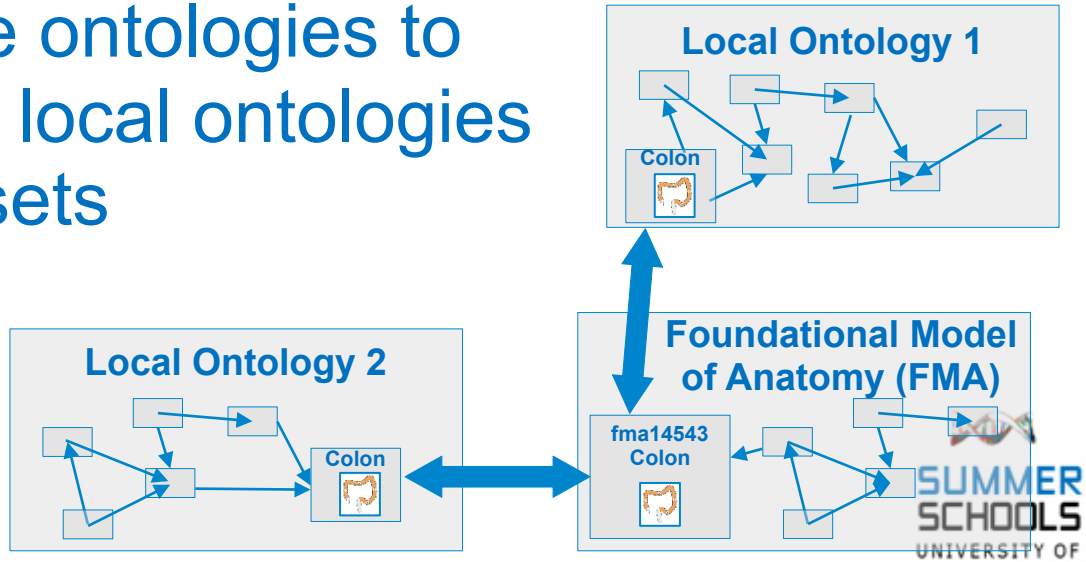
- Annotation in databases
  - Control vocabulary being used in database fields

InterPro
For MAPT

Molecular Function

GO:0008017 microtubule binding
GO:0015631 tubulin binding

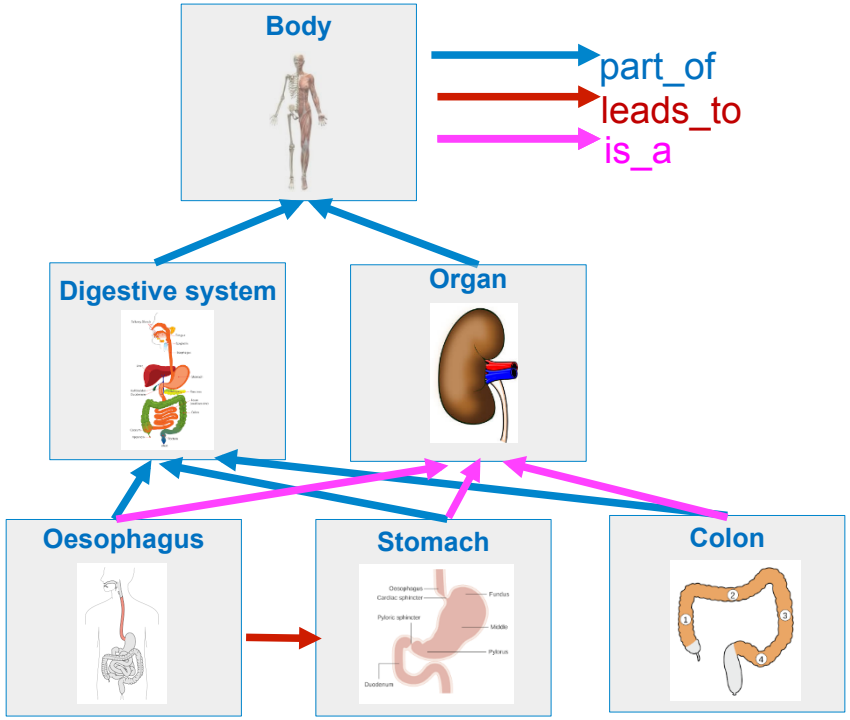
- Mapping between databases
  - Use reference ontologies to map between local ontologies and link datasets
- And also…



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Extracting knowledge from text

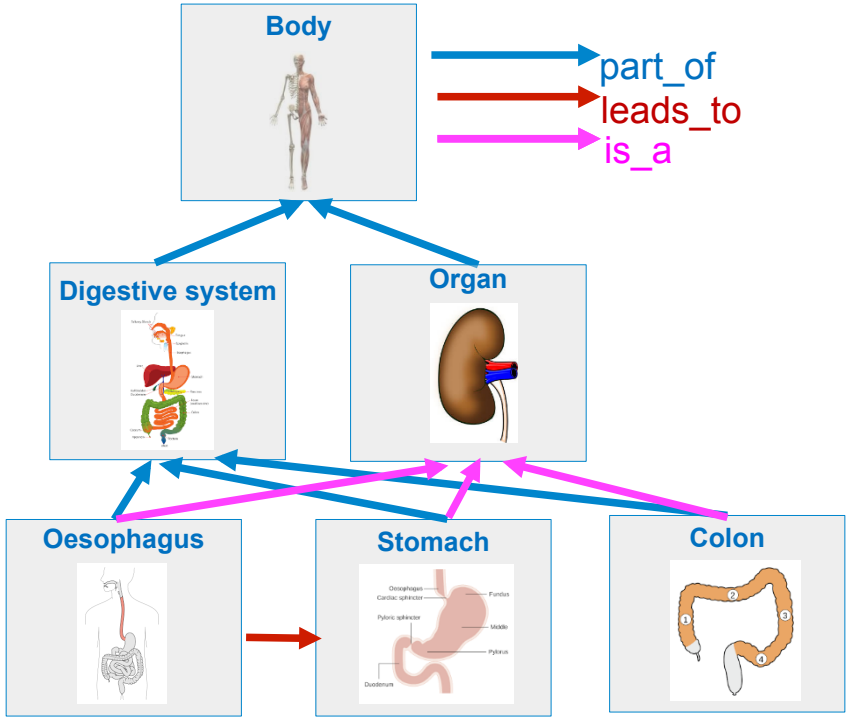
- Text
  - Food goes down the oeasophagus.
- Ontology knowledge
  - Oesophagus leads_to Stomach.
- Infered knowledge
  - Food goes down the stomach.



SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Reasoning - Principle

- Add rules
  - Rules composed of antecedent and consequence
  - Ex: partOf Digestive system→ hasFunction Digestion
- Rules expresses in a special langages
  - SWRL for knowledge bases in OWL
- Apply rules
  - Rules applied by a reasoner
- Result
  - New Knowledge, etc.
  - Ex: Stomach part_of Digestiove System → Stomach has_function Digestion



Reasoning - Use

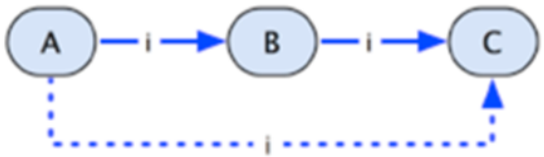
- Classification
  - Ex: The stomach is an organ. Organs are part of the body → The stomach is part of the body.
- Model checking
  - Ex: Parts of the digestive system lead to and/or from other parts of the digestive system. The stomach is part of the digestive system. The stomach leads to the lungs. → model violation
- Finding “holes”
  - Ex: Parts of the digestive system are organs. If one part of the digestive system is not defined as organ then something is missing
- Finding new relationships
  - Ex: Cats are carnivores. Carnivores eat meat.→ Cats eat meat

Reasoning – GO example

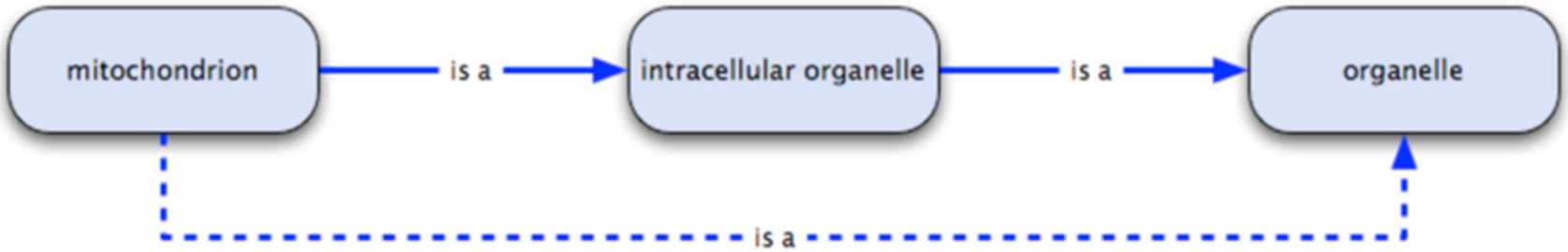
Reasoning over *is a*

is a ∘ *is a* → *is a*

The *is a* relation is *transitive*, which means that if A *is a* B, and B *is a* C, we can infer that A *is a* C.



For example:



mitochondrion *is an* intracellular organelle and intracellular organelle *is an* organelle therefore mitochondrion *is an* organelle.

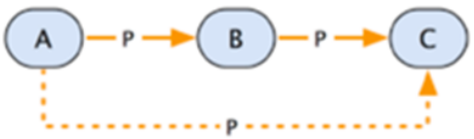
Stated more formally, this would be: *is a* ∘ *is a* → *is a*

Reasoning – GO example

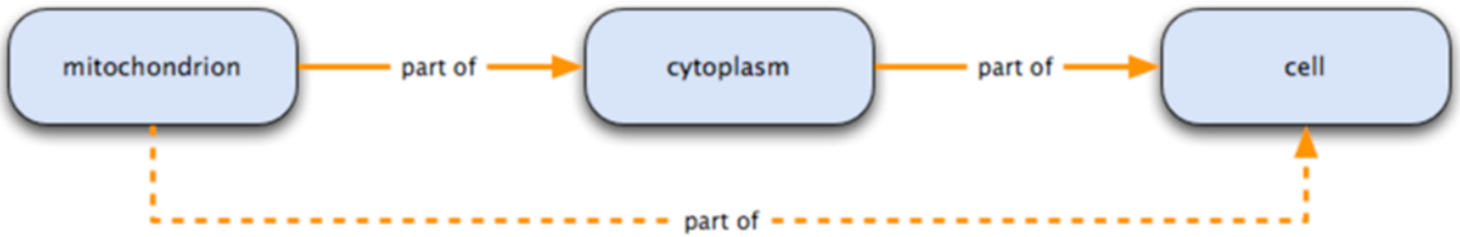
Reasoning over *part of*

part of ∘ *part of* → *part of*

Like *is a*, *part of* is transitive: *if* A *part of* B *part of* C *then* A *part of* C

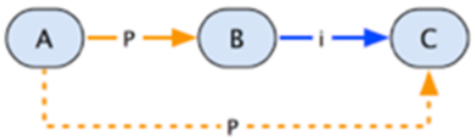


For example, mitochondrion is *part of* cytoplasm and cytoplasm is *part of* cell therefore mitochondrion is *part of* cell



part of ∘ *is a* → *part of*

If a *part of* relation is followed by an *is a* relation, it is equivalent to a *part of* relation; if A is *part of* B, and B *is a* C, we can infer that A is *part of* C.



To go further with GO

- GO mapping with others databases
- GO slims
  - Cut-down versions of the GO ontologies
  - Subset of the terms in the whole GO
    - broad overview of the ontology content
    - without the detail of the specific fine grained terms
  - http://geneontology.org/page/go-slim-and-subset-guide
- GO tools
  - AMIGO : current official web-based set of tools for searching and browsing the Gene Ontology database
  - GO Enrichment Analysis

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

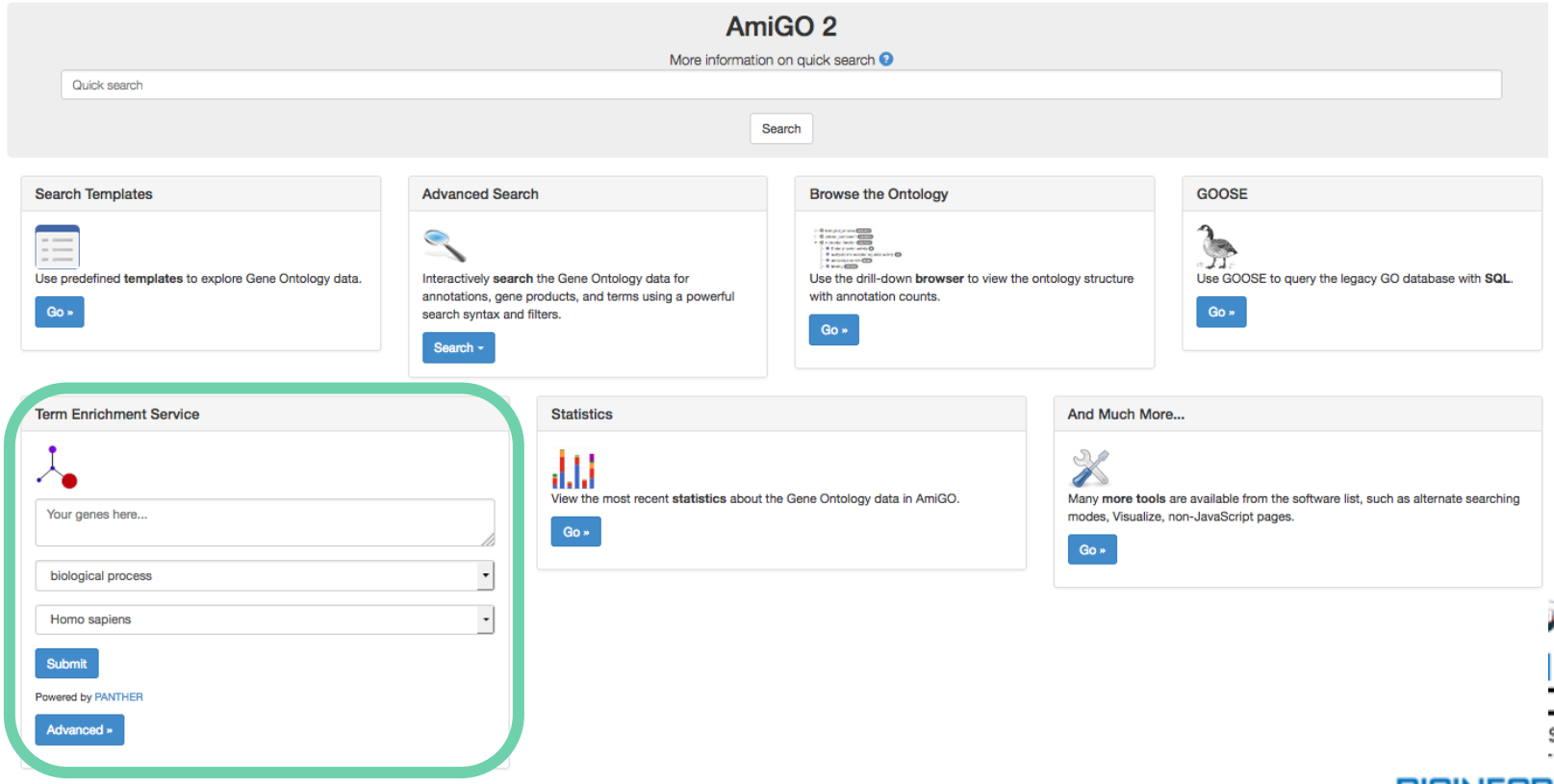
Genes Set Enrichment Analysis

- Principle
  - Input
    - A set of genes
    - Ex: gene that are up-regulated under certain conditions
  - Analysis output
    - GO terms are over-represented (or under-represented) using annotations for that gene set
    - P_value Fisher's exact test, bonferroni correction
    - Analysis using complete GO or GO-SLIM

SUMMER SCHOOLS
UNIVERSITY OF ANGERS - FRANCE
BIOINFORMATICS

Practice

- Perform the analysis for our set of genes



Practice GO-CC

Displaying only results with P<0.05; click here to display all results

GO cellular component complete	Homo sapiens (REF) #	upload_1 (∇ Hierarchy NEW!) #	expected	Fold Enrichment	+/-	P value
ficolin-1-rich granule lumen	124	13	.83	15.59	+	5.56E-09
↳ficolin-1-rich granule	185	13	1.24	10.45	+	7.13E-07
↳secretory granule	814	22	5.47	4.02	+	4.10E-05
↳secretory vesicle	930	22	6.25	3.52	+	4.02E-04
↳cytoplasmic vesicle	2177	33	14.64	2.25	+	8.43E-03
↳intracellular vesicle	2180	33	14.66	2.25	+	8.68E-03
↳vesicle	4221	94	28.38	3.31	+	9.42E-30
↳cytoplasmic vesicle part	1409	25	9.47	2.64	+	1.04E-02
↳intracellular organelle lumen	4997	79	33.60	2.35	+	3.37E-13
↳organelle lumen	4997	79	33.60	2.35	+	3.37E-13
↳membrane-enclosed lumen	4997	79	33.60	2.35	+	3.37E-13
cortical actin cytoskeleton	59	6	.40	15.13	+	4.62E-03
↳cortical cytoskeleton	87	13	.58	22.23	+	6.84E-11
↳cell cortex part	133	15	.89	16.77	+	4.10E-11
↳cell cortex	245	18	1.65	10.93	+	1.35E-10
↳cytoplasmic region	306	19	2.06	9.24	+	4.98E-10
focal adhesion	392	37	2.64	14.04	+	2.94E-28
↳cell-substrate adherens junction	395	37	2.66	13.93	+	3.84E-28
↳adherens junction	496	42	3.33	12.59	+	9.49E-31
↳anchoring junction	515	42	3.46	12.13	+	4.21E-30
↳cell junction	1210	46	8.14	5.65	+	1.82E-19
↳cell-substrate junction	400	37	2.69	13.76	+	5.96E-28
cell-cell adherens junction	88	8	.59	13.52	+	2.44E-04
↳cell-cell junction	420	16	2.82	5.67	+	3.98E-05

In cell-junction

Fisher's exact test

✦ Principle

Example : Biological process (BP)

actin filament organization GO:0007015

	genome	set
GO:0007015	236	16
Biological P	20996	142

Expected : 236/20996 = ?/142 => Enrichissment :

? = 1.596 = 1.6 16 vs1.6 = 10

✦ A video to understand Fisher's Exact Test

StatQuickie: Enrichment Analysis using Fisher's Exact Test and the Hypergeometric Distribution (5'14")

https://www.youtube.com/watch?v=udyAvvaMjfM

Practice GO-BP

Displaying only results with P<0.05; click here to display all results

	Homo sapiens (REF)	upload_1 (∇ Hierarchy NEW! ?)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
cell junction assembly	132	11	.89	12.39	+	1.81E-05
↳cell junction organization	190	11	1.28	8.61	+	7.16E-04
↳cellular component assembly	2254	44	15.15	2.90	+	3.28E-07
↳cellular component biogenesis	2508	48	16.86	2.85	+	5.79E-08
ephrin receptor signaling pathway	87	7	.58	11.97	+	2.09E-02
Fc-gamma receptor signaling pathway involved in phagocytosis	126	10	.85	11.80	+	1.57E-04
↳Fc receptor mediated stimulatory signaling pathway	128	10	.86	11.62	+	1.82E-04
↳positive regulation of biological process	5400	71	36.31	1.96	+	3.07E-06
↳immune system process	2512	39	16.89	2.31	+	3.17E-03
↳Fc-gamma receptor signaling pathway	130	10	.87	11.44	+	2.10E-04
↳immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	126	10	.85	11.80	+	1.57E-04
↳phagocytosis	277	11	1.86	5.91	+	2.76E-02
↳endocytosis	596	16	4.01	3.99	+	2.56E-02
↳vesicle-mediated transport	1748	40	11.75	3.40	+	2.81E-08
↳single-organism transport	2521	45	16.95	2.65	+	3.42E-06
↳single-organism localization	2646	48	17.79	2.70	+	3.85E-07
↳immune effector process	1003	27	6.74	4.00	+	6.13E-06
actin filament organization	195	15	1.31	11.44	+	5.80E-08
↳supramolecular fiber organization	374	20	2.51	7.95	+	1.10E-08
↳actin cytoskeleton organization	413	23	2.78	8.28	+	8.30E-11
↳single-organism organelle organization	1303	37	8.76	4.22	+	3.95E-10
↳cytoskeleton organization	941	35	6.33	5.53	+	8.00E-13
↳actin filament-based process	482	24	3.24	7.41	+	2.28E-10
mitochondrial membrane organization	121	8	.81	9.83	+	1.67E-02
↳membrane organization	982	28	6.60	4.24	+	7.69E-07
protein stabilization	152	9	1.02	8.81	+	9.28E-03
↳regulation of protein stability	243	11	1.63	6.73	+	7.93E-03
establishment or maintenance of cell polarity	155	9	1.04	8.64	+	1.09E-02
nuclear export	162	9	1.09	8.26	+	1.56E-02

Regulation of cytoskeleton organisation

Practice GO-MF

Displaying only results with P<0.05; click here to display all results

	Homo sapiens (REF)	upload_1 (Hierarchy NEW!)				
GO molecular function complete	#	#	expected	Fold Enrichment	+/-	P value
actin filament binding	151	19	1.02	18.72	+	4.11E-15
↳actin binding	409	31	2.75	11.27	+	5.41E-20
↳cytoskeletal protein binding	864	40	5.81	6.89	+	4.77E-19
structural constituent of cytoskeleton	106	13	.71	18.24	+	1.87E-09
↳structural molecule activity	730	29	4.91	5.91	+	3.74E-11
unfolded protein binding	113	13	.76	17.11	+	4.14E-09
cadherin binding	295	32	1.98	16.13	+	1.90E-25
↳cell adhesion molecule binding	447	35	3.01	11.65	+	1.71E-23
RNA polymerase II distal enhancer sequence-specific DNA binding	64	6	.43	13.94	+	1.71E-02
protein N-terminus binding	105	9	.71	12.75	+	1.57E-04
ion channel binding	113	7	.76	9.21	+	4.09E-02
ribonucleoprotein complex binding	113	7	.76	9.21	+	4.09E-02
calmodulin binding	187	10	1.26	7.95	+	2.09E-03
protein C-terminus binding	192	9	1.29	6.97	+	2.20E-02
ubiquitin protein ligase binding	297	12	2.00	6.01	+	2.95E-03
↳ubiquitin-like protein ligase binding	302	12	2.03	5.91	+	3.50E-03
↳enzyme binding	1855	44	12.47	3.53	+	1.64E-10
protein kinase binding	555	17	3.73	4.56	+	7.37E-04
↳kinase binding	626	21	4.21	4.99	+	5.09E-06
protein domain specific binding	662	17	4.45	3.82	+	8.05E-03

Link with structural molecule activity

Predicted function (has to be validated by experiments)

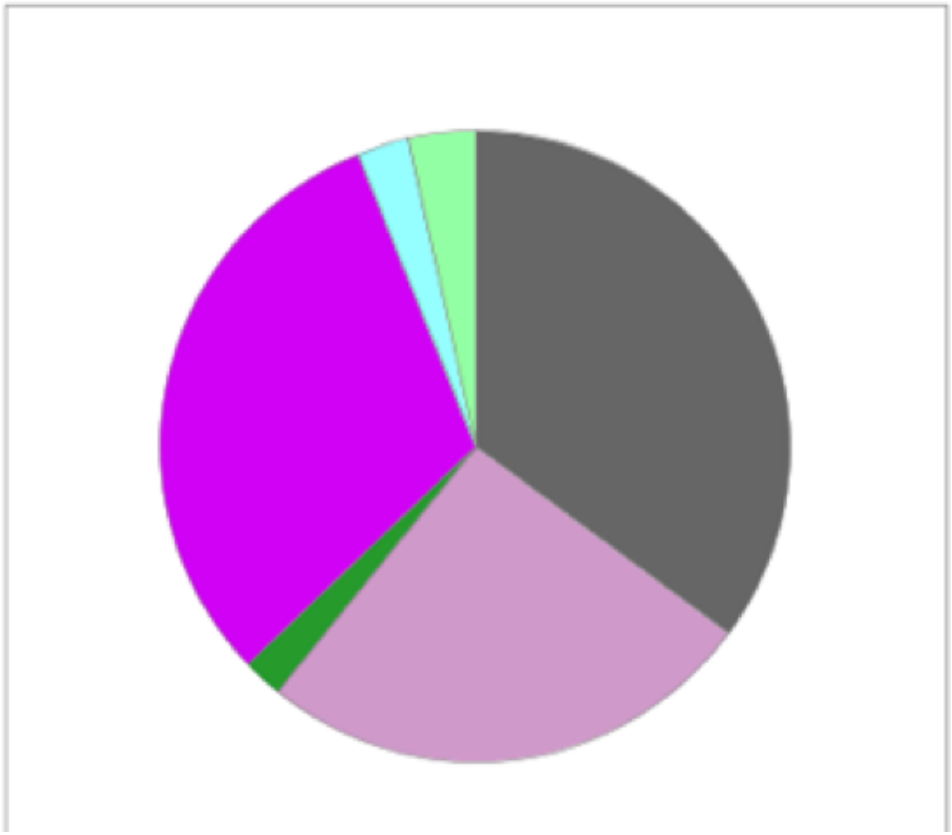
Another view :

http://pantherdb.org

Select Ontology: Molecular Function View: 100%

PANTHER GO-Slim Molecular Function

Total # Genes: 141 Total # function hits: 148



Click to get gene list for a category:

- binding (GO:0005488)
- catalytic activity (GO:0003824)
- receptor activity (GO:0004872)
- structural molecule activity (GO:0005198)
- translation regulator activity (GO:0045182)
- transporter activity (GO:0005215)

Color picker powered by Web Colors by VisiBone

Another view :

http://pantherdb.org

GENEONTOLOGY Unifying Biology

PANTHER Classification System

LOGIN REGISTER CONTACT US

Home | About | PANTHER Data | PANTHER Tools | Workspace | Downloads | Help/Tutorial

Now includes comprehensive GO annotations directly imported from the GO database

Search

Ontology terms

Go

Quick links

- Whole genome function views
- Genome statistics
- Data Version
- How to cite PANTHER
- NEW! Recent publication describing PANTHER

News

New functionalities of PANTHER Gene List Analysis Tools

Click for additional info.

Newsletter subscription

Enter your Email:

Subscribe

Gene List Analysis | Browse | Sequence Search | cSNP Scoring | Keyword Search

Please refer to our article in Nature Protocols for detailed instructions on how to use this page.

Help Tips

Steps:

1. Select list and list type to analyze
2. Select Organism
3. Select operation

1. Enter ids and or select file for batch upload. Else enter ids or select file or list from workspace for comparing to a reference list.

input

Enter IDs: Supported IDs

Q9NYL9
Q9Y2L6
Q9Y3Y2

separate IDs by a space or comma

Upload IDs: File format

Choisir le fichier aucun fichier sél.

Please login to be able to select lists from your workspace.

Select List Type:

- ID List
- Previously exported text search results
- Workspace list
- PANTHER Generic Mapping File
- VCF File Flanking region 20 Kb

2. Select organism.

- Homo sapiens
- Mus musculus
- Rattus norvegicus
- Gallus gallus
- Danio rerio

3. Select Analysis.

- Functional classification viewed in gene list
- Functional classification viewed in pie chart
- Statistical overrepresentation test Use default settings
- Statistical enrichment test Use default settings

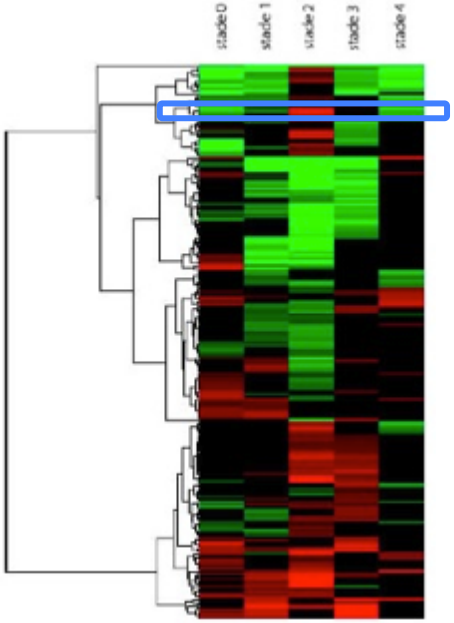
submit

SUMMER SCHOOLS UNIVERSITY OF ANGERS - FRANCE

BIOINFORMATICS

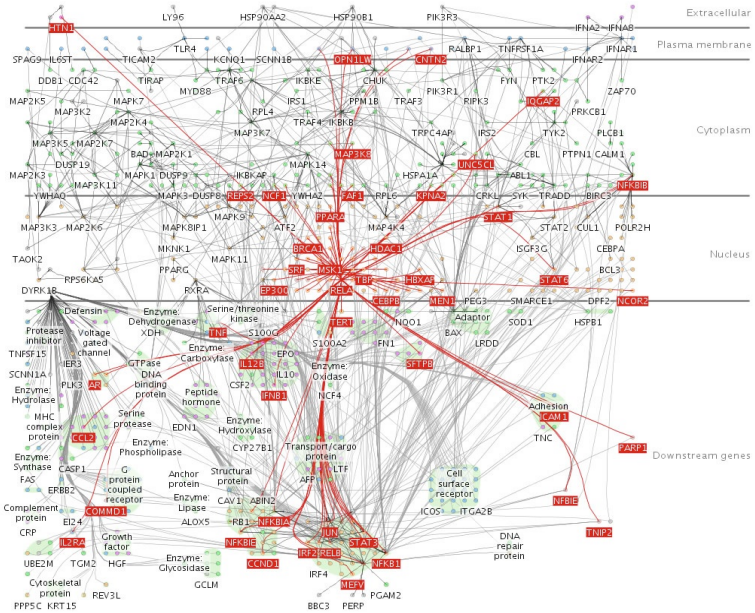
Biology conclusion ...

- How could we understand cell transformation and cancer ??

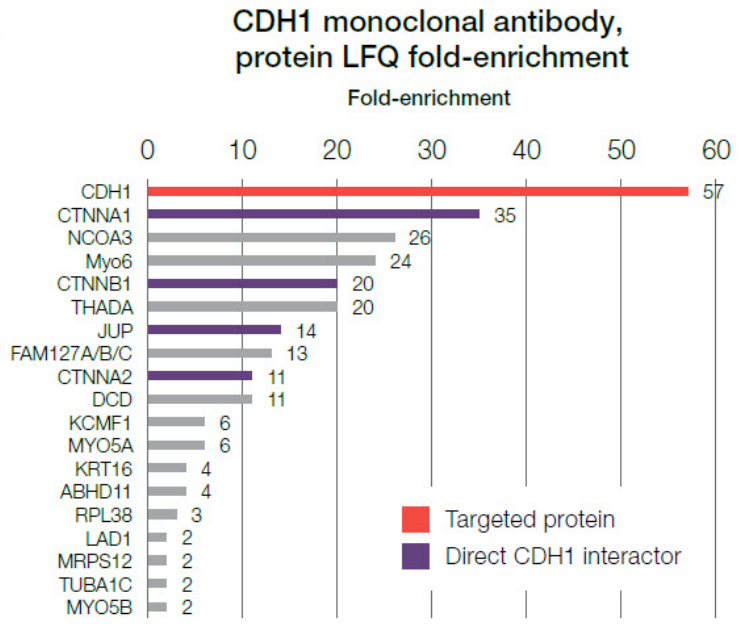


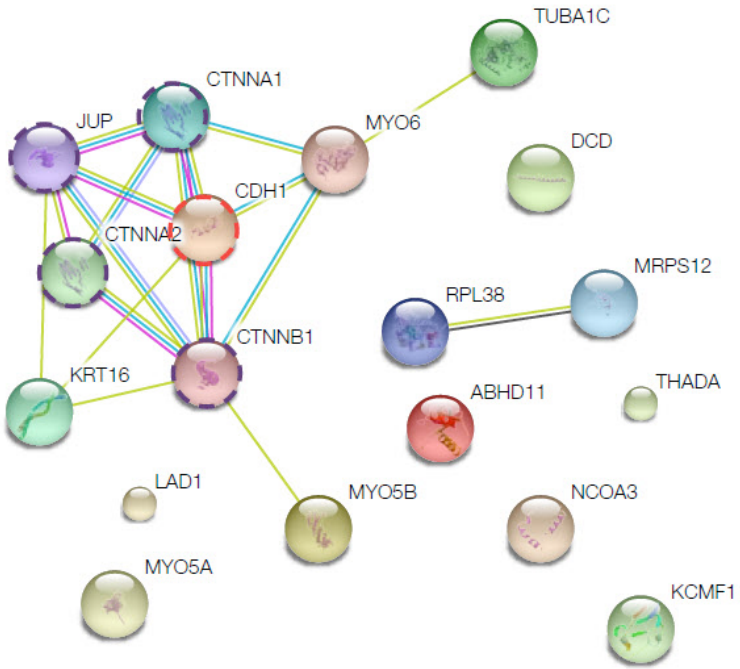
ID of the tumor
(first point)

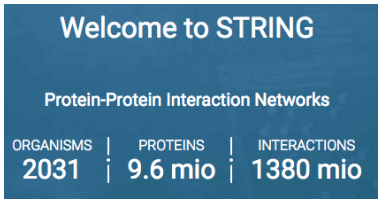
Define function of each marker to construct a map of biology of systems
(second point)



Biology conclusion ...

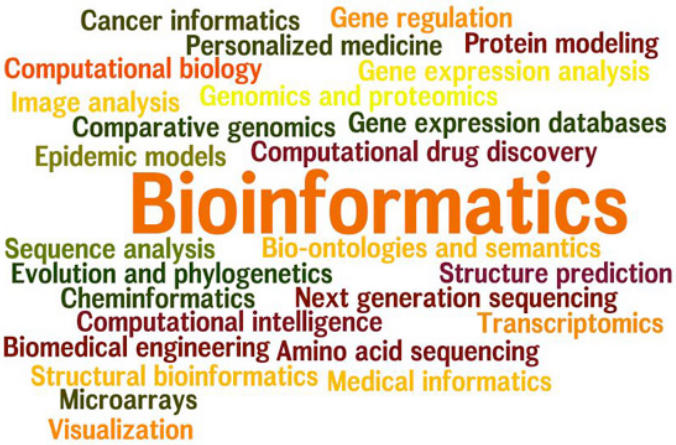






Define a function in a environment

Biology conclusion …



And in the example ?? New targeted therapy ??

Practice

- Design your own ontology
  - Biological topics considered during the session
  - Define
    - Concepts
    - Relationships between these concepts