

Introducción a la Bioinformática

Ontologías

Fernán Agüero

Instituto de Investigaciones Biotecnológicas
Universidad Nacional de General San Martín

► Qué significa ontología?

■ Webster's Revised Unabridged Dictionary

► Ontology: the things which exist

- The department of the science of metaphysics which investigates and explains the nature and essential properties and relations of all beings, ...

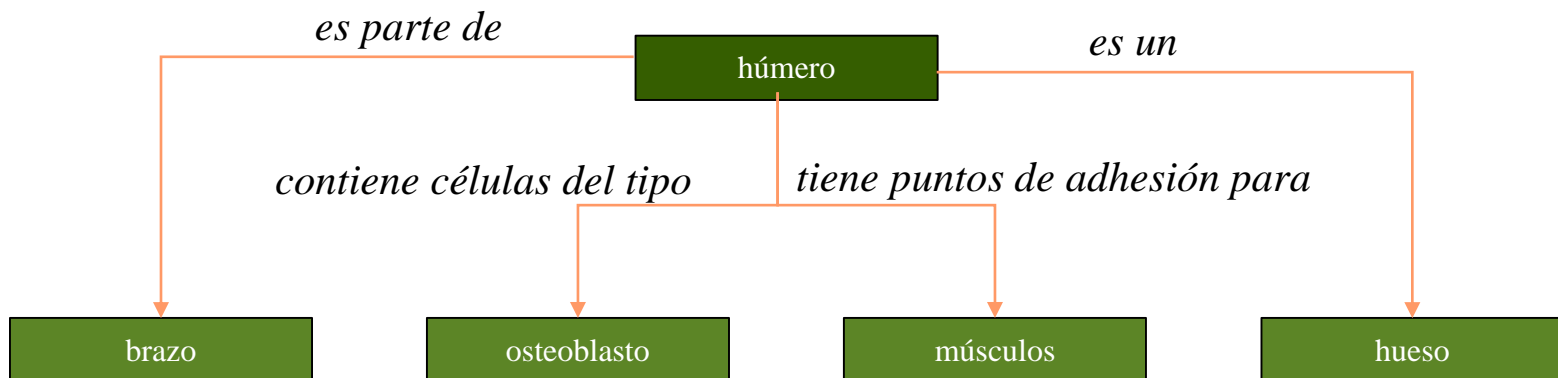
■ The Free On-Line Dictionary of Computing

► Ontology

- **Phylosophy:** a systematic account of experience
- **Artificial Intelligence:** an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them. [...]
- **Information Science:** the hierarchical structuring of knowledge about things by subcategorizing them according to their essential (or at least relevant and/or cognitive) qualities.

Otras definiciones y ejemplos

- ▶ Una ontología es un área del conocimiento que ha sido formalizada
 - Términos (conceptos) individuales
 - Afirmaciones que conectan términos entre sí
- ▶ Ejemplo: una ontología anatómica
 - Términos: húmero, brazo, osteoblasto, músculo, hueso
 - Conexiones (rules): *es parte de, contiene células del tipo, tiene puntos de adhesión para, es un*



Otros componentes

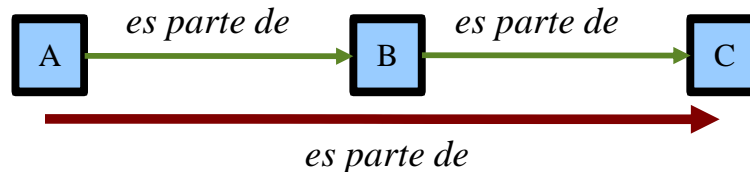
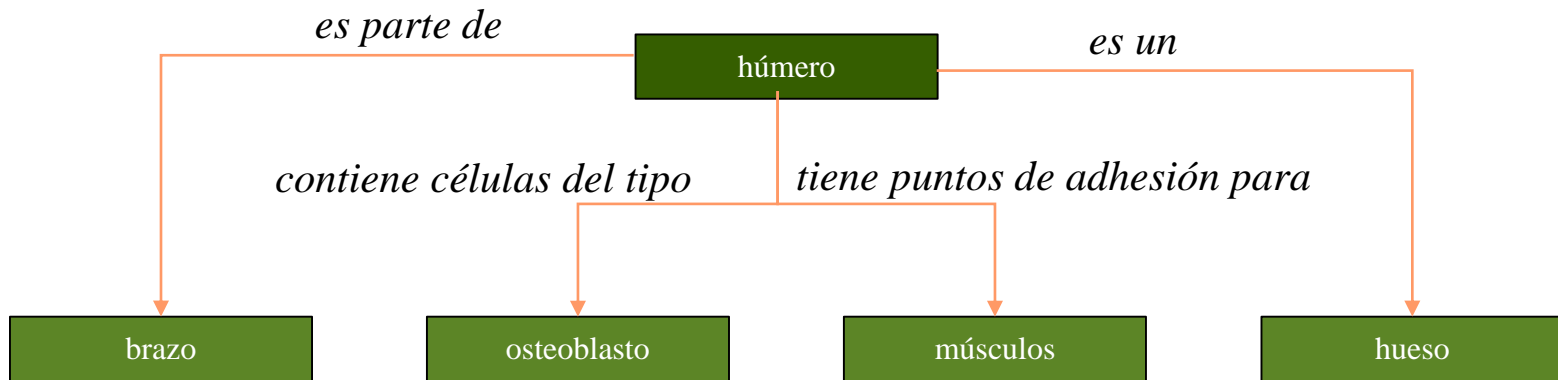
- ▶ Cada término en una ontología está asociado a:
 - un identificador único: **GO:0019505**
 - Un nombre: **resorcinol metabolism**
 - Una definición: *"the chemical reactions and physical changes involving resorcinol ($C_6H_4(OH)_2$), a benzene derivative with many applications (including dyes, explosives, resins and as an antiseptic)"*
 - Sinónimos: **1,3-benzenediol metabolism; 1,3-dihydroxybenzene metabolism**

Ontologías vs anotaciones

- ▶ **Anotación: descripción textual de un objeto**
- ▶ **Las ontologías contienen reglas y afirmaciones que componen una 'descripción lógica' del área que abarcan**
 - **Se puede utilizar esta 'descripción lógica' de los objetos para:**
 - ▶ realizar consultas a distintos niveles de un set de datos
 - ▶ realizar consultas a través de distintos sets de datos

Propiedades de las reglas


- En este ejemplo las conexiones tienen dirección
 - El **húmero es parte del brazo, pero no viceversa**




Transitividad





Mouse anatomy – Gene expression



Outside image



Zoom   Deselect

Transverse Frontal Sagittal

Find:

alimentary system.gut

|-> gut+

|-> foregut

|-> gland

|-> thyroid primordium

|-> associated mesenchyme

|-> endoderm

|-> pharyngeal region

|-> associated mesenchyme

|-> epithelium

|-> lumen

|-> vascular element

Gene Expression Data

Query Results -- Summary

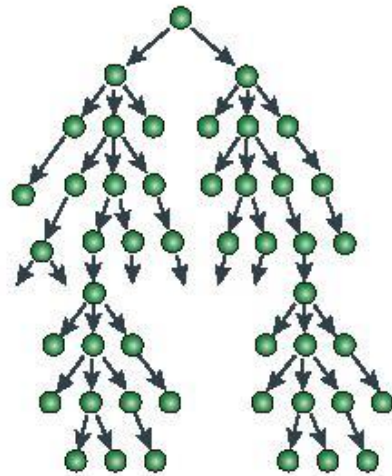
62 matching assay results displayed

Gene	Assay Type	Assay	Age	Structure	Detected?
Acta1	Immunohistochemistry	MGI:1927891	E9.25	TS14: gut	no
Actc1	Immunohistochemistry	MGI:1927890	E9.25	TS14: gut	no
Bmp4	RNA In Situ	MGI:1276404	E9.0	TS14: gut	yes
Fgf8	RNA In Situ	MGI:1328412	E8.5	TS14: foregut	yes
Foxa1	RNA In Situ	MGI:1339061	E9	TS14: gut	yes
Foxa1	RNA In Situ	MGI:1339061	E9	TS14: gut	yes

Representación y reglas en una ontología

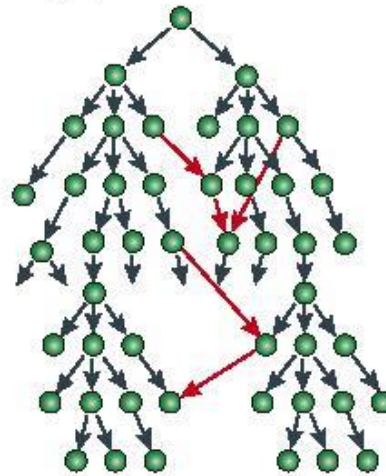
- Las afirmaciones (conexiones) y las reglas que definen una ontología pueden utilizarse para realizar inferencias lógicas acerca de los términos y sus propiedades asociadas

a Simple hierarchy



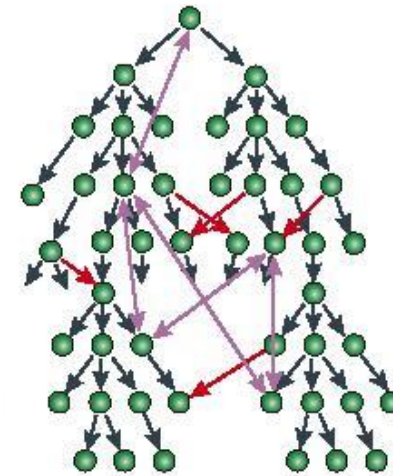
→ Rule: *is instance of*
Directed rule:
1 parent

b Directed acyclic graph = DAG



→ Rule: *signals to*
Directed rule:
>1 parent

c Graph



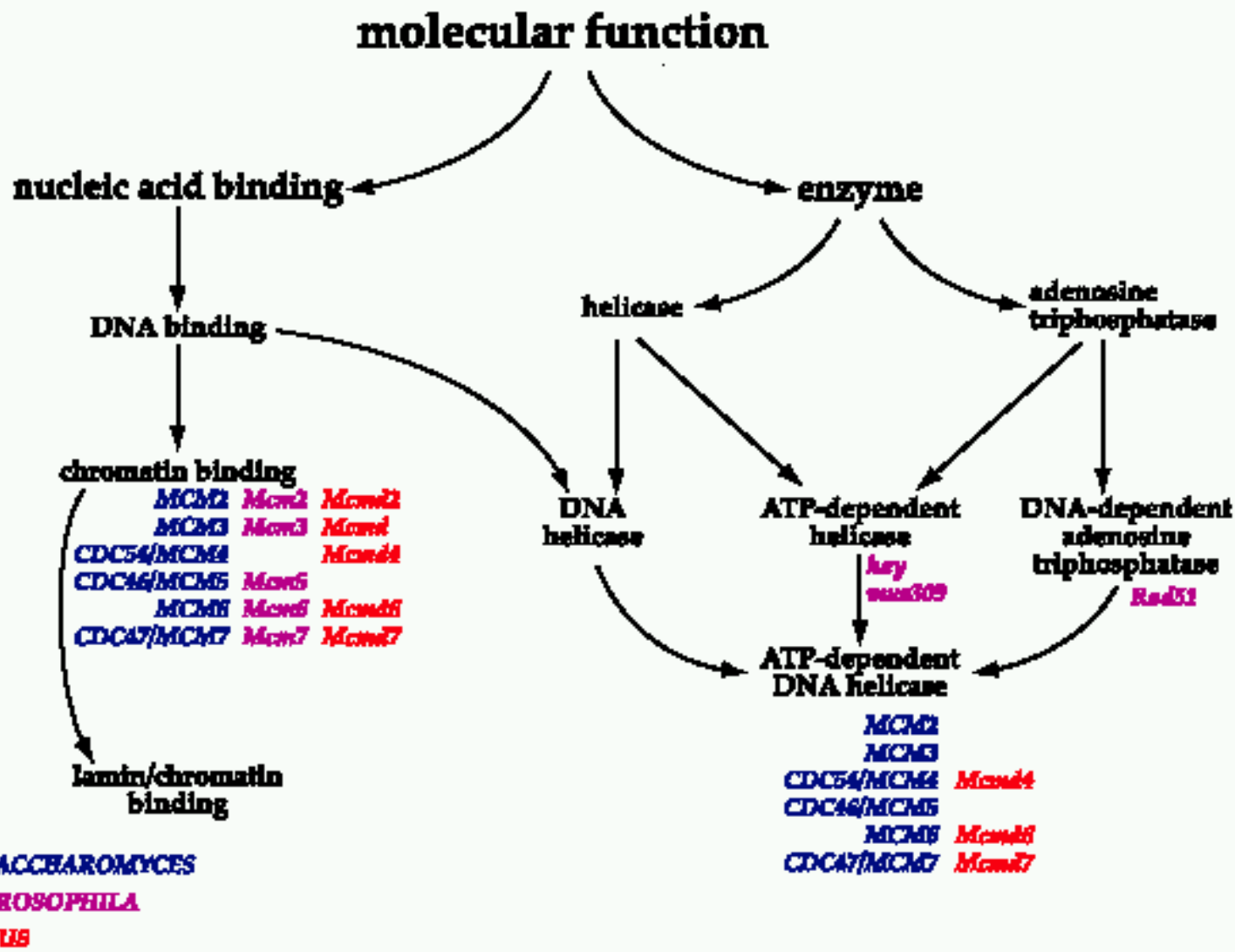
↔ Rule: *is next to*
Undirected rule:
parents are equivalent
to children

Gene Ontology (GO)

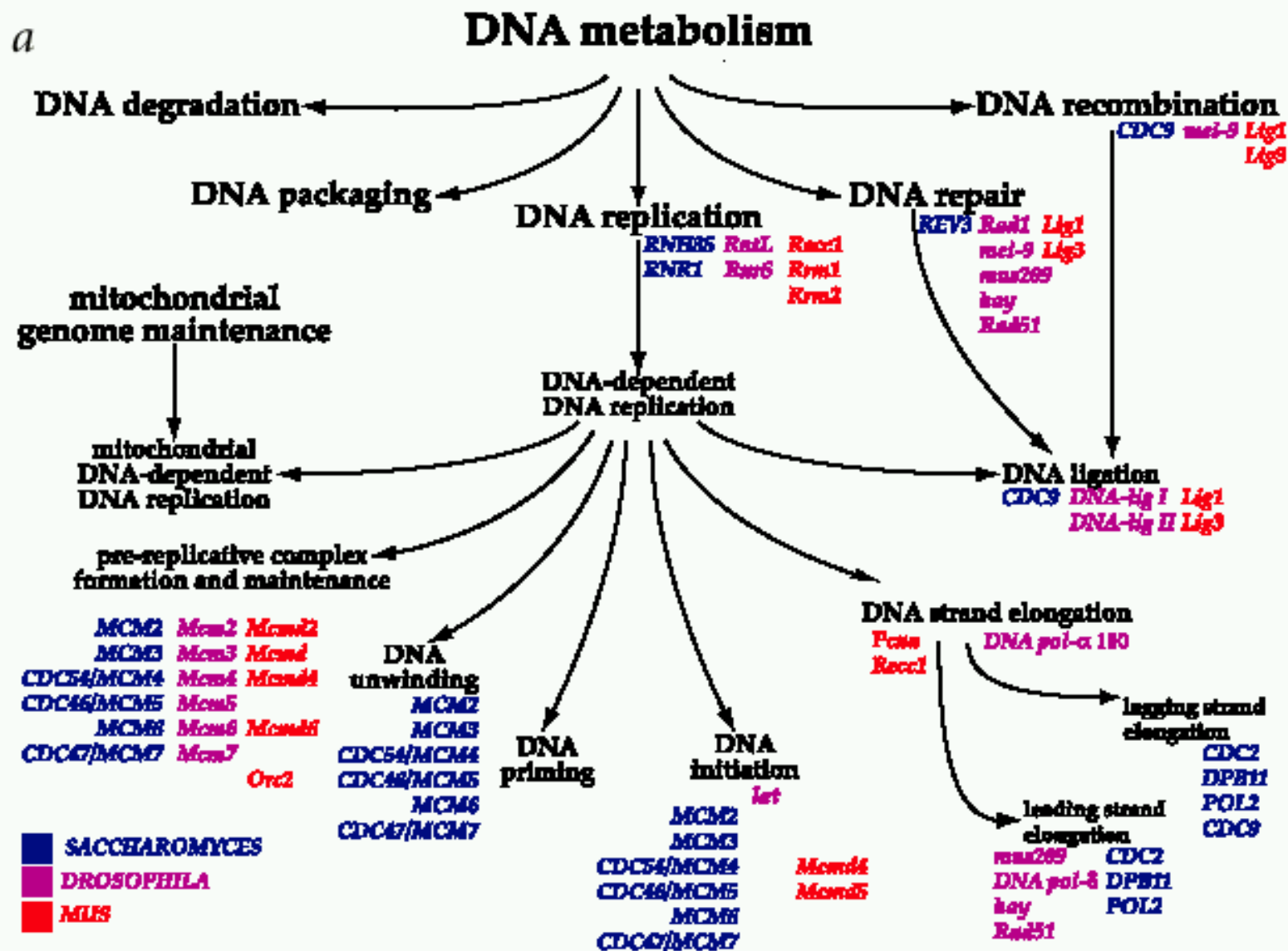
- ▶ **Describe tres ontologías independientes**
 - **Molecular function:** la actividad o función que cumple el producto de un gen. Ejemplos: transcription factor, DNA helicase.
 - **Biological process:** procesos en un sentido amplio, como "mitosis" o "metabolismo de purinas", que son llevados a cabo por conjuntos ordenados de funciones moleculares.
 - **Cellular component:** estructuras subcelulares, localizaciones, complejos macromoleculares. Ejemplos: núcleo, telómero, origin recognition complex
- ▶ **Cualquier gen puede ser mapeado en estas ontologías. O dicho de otra forma: el producto de un gen individual tiene una **función molecular**, es parte de algún **proceso biológico** y ocurre en algún **componente celular**.**

GO: molecular function

b

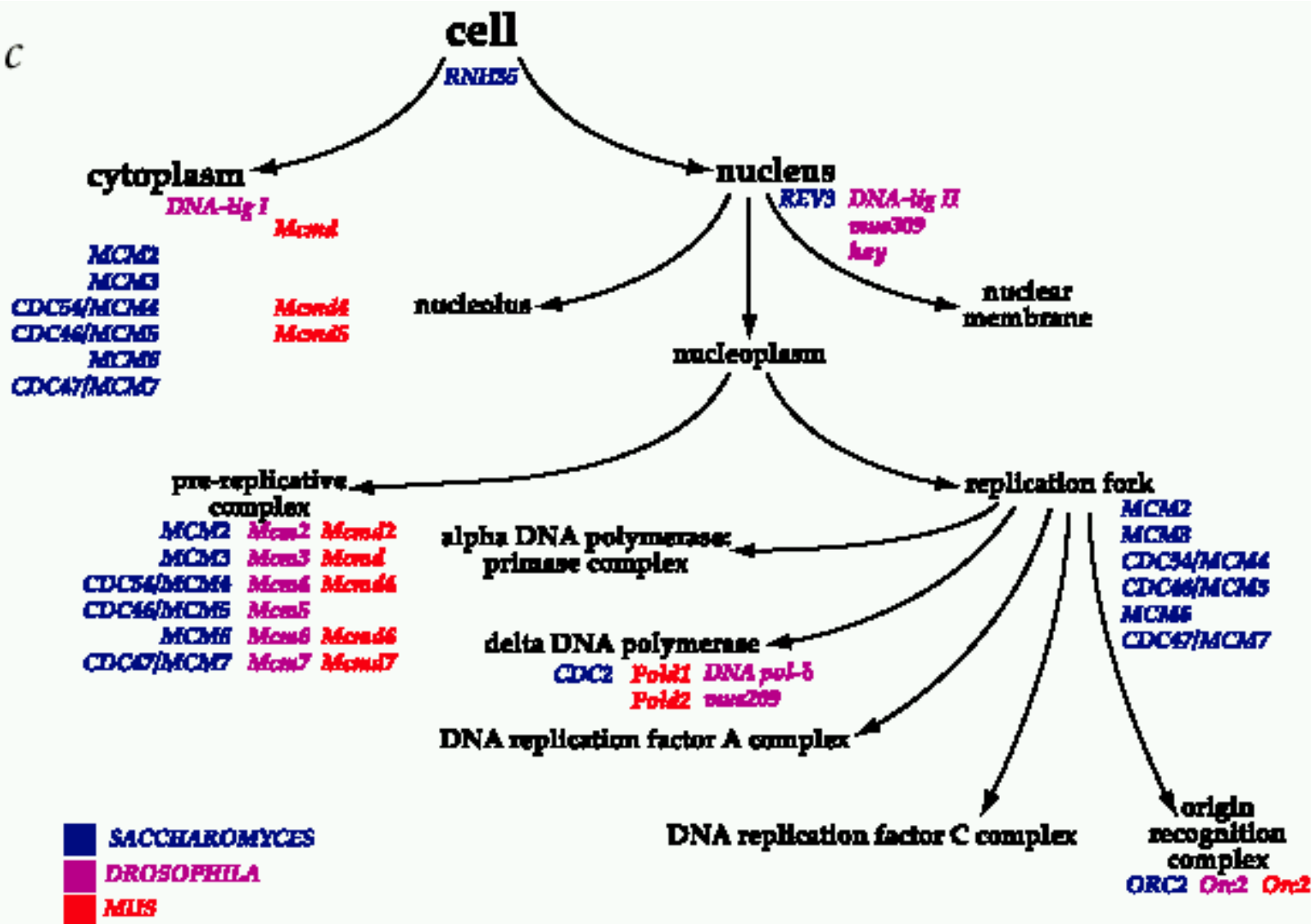


GO: biological process



GO: cellular component

c



- ▶ **Ontology statistics (Aug.2012)**
- ▶ **37,928 términos**
- ▶ **Los términos están asociados (linkeados) a una base de datos de más de 597,000 genes de cerca de 50 organismos**
 - **Cada proteína está asociada a uno o más GO Ids**
 - **Se pueden buscar las proteínas asociadas a un determinado término**
 - **O todos los términos asociados con una proteína**

AmiGO: <http://amigo.geneontology.org>

► **Simple**

- **Permite buscar términos en GO asociados a productos génicos**
- **O viceversa**

► **Links a varias bases de datos: de secuencia, organismo específicas, etc.**





The image shows a screenshot of the AmiGO web interface. The interface is divided into several sections:

- Top Section:** Contains the "AmiGO" logo. An arrow points to it with the label "Tope de la jerarquía".
- Search Section:** Contains a "Search GO" input field, radio buttons for "Exact Match", "Terms", and "Gene Products", and a "Submit" button. An arrow points to the "Submit" button with the label "Buscar".
- Search Filters Section:** Contains three dropdown menus: "Species" (with options "All", "A. aeolicus", "A. fulgidus"), "Datasource" (with options "All", "FlyBase", "SGD"), and "Evidence Code" (with options "All Curator Approved", "IMP", "IGI"). An arrow points to the "Species" dropdown with the label "Aplicar filtros a Los productos génicos". Below the filters is another "Submit" button and a link to "Advanced Query".
- Results Section:** Displays a hierarchical list of Gene Ontology terms. The top term is "GO:0003673 : Gene_Ontology (146200)". Below it are three sub-terms: "GO:0008150 : biological_process (96312)", "GO:0005575 : cellular_component (79199)", and "GO:0003674 : molecular_function (97507)". An arrow points from the "Gene_Ontology" term to the label "Navegar".

Annotations with arrows:

- From the "AmiGO" logo to "Tope de la jerarquía".
- From the "Submit" button in the Search section to "Buscar".
- From the "Species" dropdown in the Search Filters section to "Aplicar filtros a Los productos génicos".
- From the "GO:0003673 : Gene_Ontology (146200)" term to "Navegar".

AmiGO (cont.)

[-] **GO:0003673 : Gene_Ontology (146200)** 
[+]  GO:0008150 : biological_process (96312)
[+]  GO:0005575 : cellular_component (79199)
[+]  GO:0003674 : molecular_function (97507)

Tipo de relación

P: part of














I: is a

IDs

nombres

Nro de genes mapeados

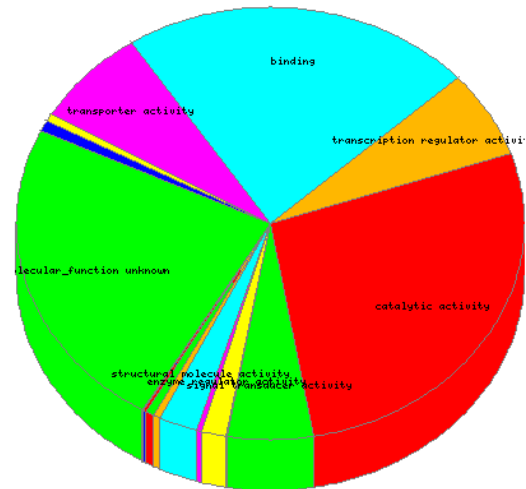
AmiGO: navegación

- [-] GO:0003673 : Gene_Ontology (146200) 
- [-]  GO:0008150 : biological_process (96312) 
- [-]  GO:0007610 : behavior (2293)
- [-]  GO:0000004 : biological_process unknown (26924)
- [-]  GO:0009987 : cellular process (31905)
- [-]  GO:0007275 : development (14496)
- [-]  GO:0008371 : obsolete biological process (90)
- [-]  GO:0007582 : physiological process (60310)
- [-]  GO:0050789 : regulation of biological process (2533)
- [-]  GO:0016032 : viral life cycle (252)
- [-]  GO:0005575 : cellular_component (79199)
- [-]  GO:0003674 : molecular_function (97507)

AmiGO: pie charts

Gene Products Annotated Below molecular_function

GO:0003674 : molecular_function



Term Name	Total Gene Products	Percent of All molecular_function
All molecular_function	97507	100.0 %
catalytic activity	32256	33.0
molecular_function unknown	27869	28.5
binding	26483	27.1
transporter activity	8671	8.89
transcription regulator activity	7695	7.89
signal transducer activity	6386	6.54
structural molecule activity	2898	2.97
enzyme regulator activity	1810	1.85
chaperone activity	883	0.90
obsolete molecular function	672	0.68
translation regulator activity	586	0.60
triplet codon-amino acid adaptor activity	553	0.56
motor activity	414	0.42
antioxidant activity	320	0.32
nutrient reservoir activity	36	0.03
chaperone regulator activity	13	0.01
molecular_function	0	0

AmiGO: gene search

Search GO

☐ Exact Match

☐ Terms

☒ Gene Products

Gene Product	Datasource	Associated Terms
<input type="checkbox"/> Bcat1	RGD	TAS branched-chain-amino-acid transaminase activity
<input type="checkbox"/> T27I1.8	TIGR_Ath1	ISS catalytic activity

Search Filters

Species

All
A. aeolicus
A. fulgidus
A. pernix

Datasource

All
FlyBase
SGD
MGI

Evidence Code

All Curator Approved
IMP
IGI
IPI

GO: evidence codes

Evidence codes

IC: inferred by curator

IDA: inferred from direct assay

IEA inferred from electronic annotation

IEA inferred from electronic annotation

IGI inferred from genetic interaction

IMP inferred from mutant phenotype

IPI inferred from physical interaction

ISS inferred from sequence or structural similarity

TAS traceable author statement

NAS non-traceable author statement

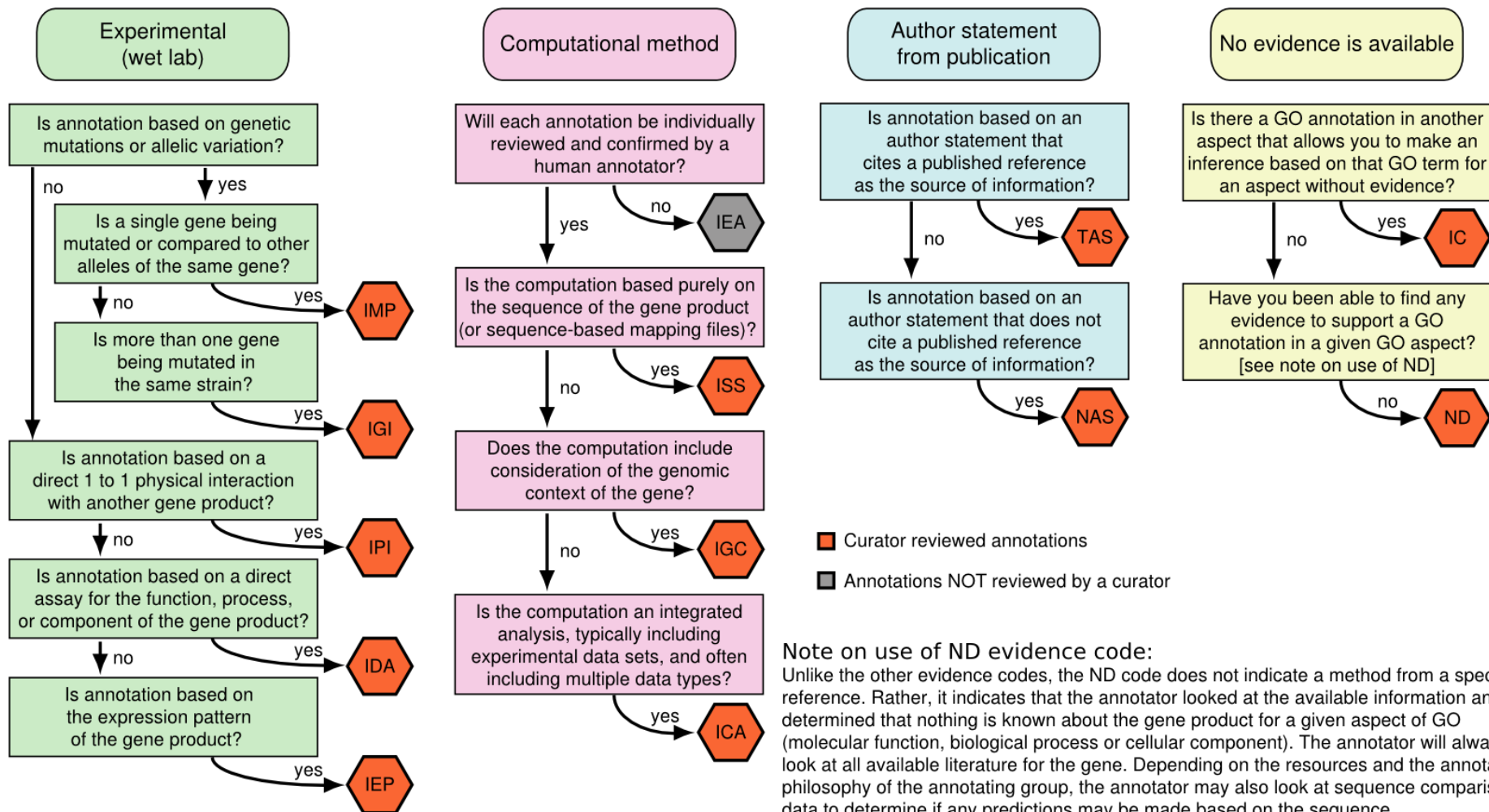
ND no biological data available

<http://www.geneontology.org/GO.evidence.shtml>

Evidence Codes, explained

GO Evidence Code Decision Tree

What type of evidence is the annotation based on?



Note on use of ND evidence code:

Unlike the other evidence codes, the ND code does not indicate a method from a specific reference. Rather, it indicates that the annotator looked at the available information and determined that nothing is known about the gene product for a given aspect of GO (molecular function, biological process or cellular component). The annotator will always look at all available literature for the gene. Depending on the resources and the annotation philosophy of the annotating group, the annotator may also look at sequence comparison data to determine if any predictions may be made based on the sequence.

AmiGO: term search

Search GO

☐ Exact Match☒ Terms☐ Gene Products

Go Term

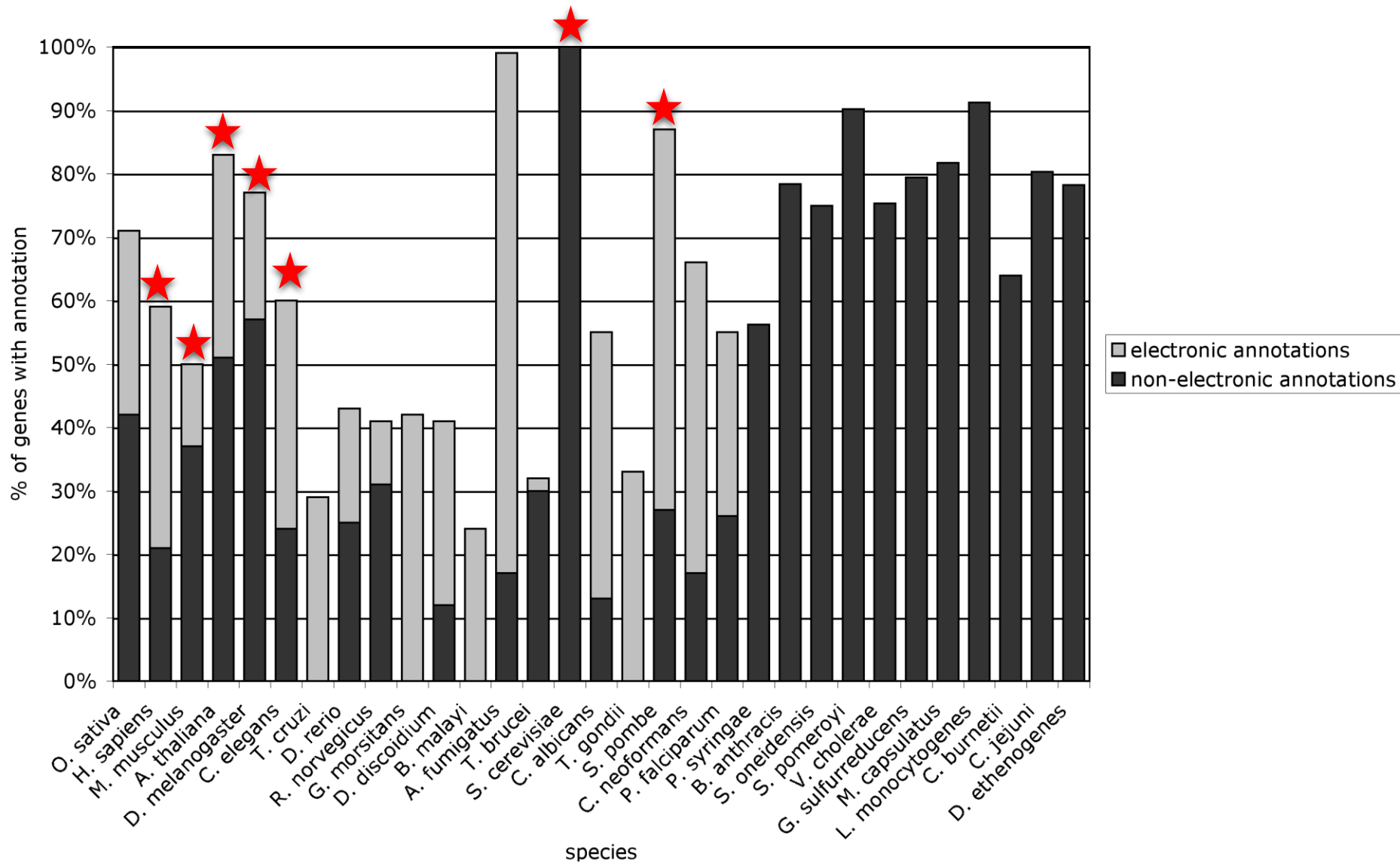
☐ [transcriptional activator activity](#)☐ [zinc-mediated transcriptional activator activity](#)

Definition

Any transcription factor required for initiation or upregulation of transcription.

Initiates or upregulates transcription in the presence of zinc.

Cobertura variable!



Ontologías anatómicas

- ▶ **Comprenden la descripción de estructuras físicas supracelulares que hacen a un determinado organismo**
 - **Reglas del tipo:**
 - ▶ Localización relativa: el ventrículo *es parte del* corazón
 - ▶ Linaje: el tubo digestivo *deriva del* endodermo
 - ▶ Clase: el sistema cardiovascular *es un* sistema orgánico

- ▶ **Distintos usuarios requieren distintas ontologías**
 - **Cirujano: ontología anatómica que incluya relaciones espaciales entre tejidos (*next to*)**
 - ▶ Galen: www.opengalen.org
 - ▶ Digital Anatomist:
depts.washington.edu/ventures/pfolio/fma.htm
 - **Biólogo estudiando desarrollo: ontología con relaciones estructurales (*part of*) o de linaje (*derived from*)**
 - ▶ Mouse Developmental Anatomy:
genex.hgu.mrc.ac.uk/Databases/Anatomy
 - ▶ Human Developmental Anatomy:
genex.hgu.mrc.ac.uk/Databases/HumanAnatomy

Ontologías cruzadas

- ▶ El uso de IDs para identificar los términos de una ontología facilita las referencias cruzadas entre distintas ontologías

- ID: CL:0000188, skeletal muscle cell

- ▶ Ejemplos:

- **Edinburgh Mouse Atlas Project (EMAP):**

- genex.hgu.mrc.ac.uk

- ▶ Secciones de estadíos tempranos del desarrollo del ratón con sus tejidos identificados y mapeados a IDs en EMAP

- **Mouse Gene Expression Database (GXD):**

- www.informatics.jax.org/searches/expression_form.shtml

- ▶ Tabla de todos los genes que se expresan en el/los tejidos identificados por el/los EMAP IDs
 - ▶ A su vez GXD asocia términos de GO. Se pueden hacer búsquedas del tipo

- Genes expresados en el corazón en desarrollo (EMAP:XXXXX) y que tengan actividad de factor de transcripción (GO:XXXXXX)

Anatomy & Gene Expression

emap

3D digital atlas | **TS14**

HOME 3D DIGITAL ATLAS EMAGE DATABASE RESOURCES CONTACT SITE SEARCH

☐ 3D Navigation ☒ Navigation Window ☒ Section Window ☒ Anatomy Window Theiler Stage **TS14**

m.nervous syst
ture brain

TS14

- embryo
 - branchial arch
 - 1st arch
 - branchial groove
 - branchial membrane
 - branchial pouch*
 - in GXD**
 - mandibular component
 - maxillary component
 - maxillary-mandibular groove
 - mesenchyme*
 - mesenchyme derived from
 - mesenchyme derived from
 - 2nd arch

< Contract Expand >

Zoom

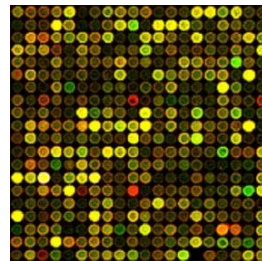
Transverse Frontal Sagittal

Deselect No component selected

Find :

Interpretar listas de genes

- Muchos experimentos de alta escala devuelven como resultado listas de genes
 - Transcriptómica
 - Proteómica
 - Metabolomics
 - Protein-protein interactions
 - CHIP-Seq (DNA-protein interactions)
 - Genetic association studies GWAS)



Ranking or
clustering

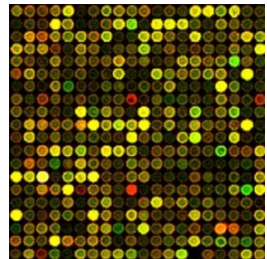


GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2



Aplicaciones: interpretar experimentos

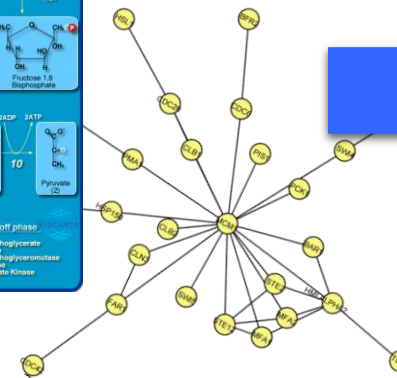
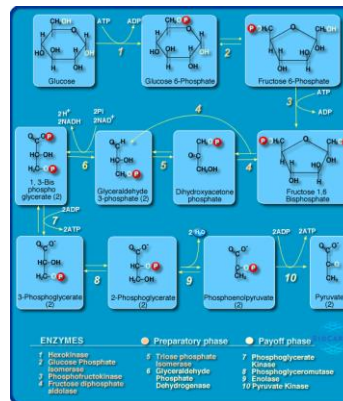
Al poner las listas en contexto biológico se puede analizar enriquecimiento en *términos* o *conceptos*



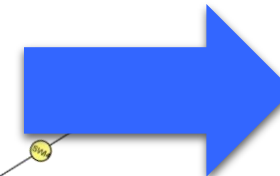
Ranking or clustering



GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2



Analysis tools



Eureka! New heart disease gene!

Referencias

Ontologies: formalising biological knowledge for bioinformatics. Bard J. Bioessays 25 (2003): 501-506.

Ontologies in Biology: design, applications and future challenges. Bard JBL, Rhee SY. Nature Reviews Genetics 5 (2004): 213-222

Obofoundry