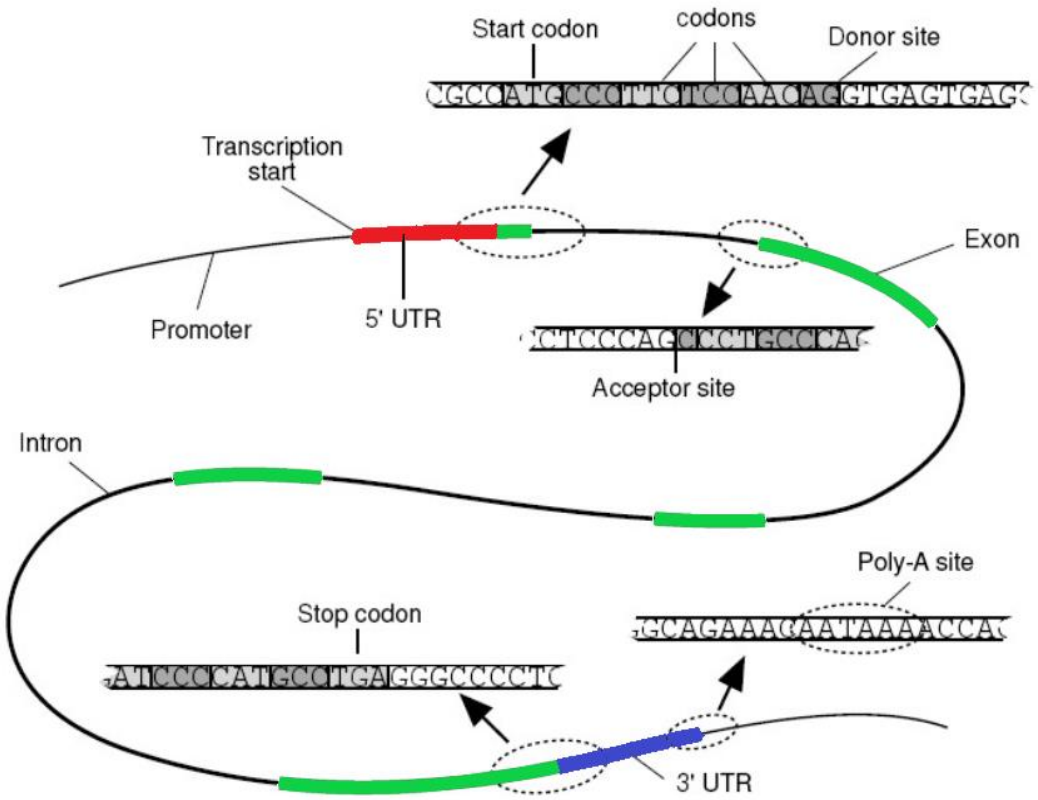
Genomic annotation: from sequence to predicted function

Raw sequence data: millions and millions of nucleotides

```
AAACACTTAGACAATCAATATAAAGATGAAGTGAACGC
TCTTAAAGAGAAGTTGGAAAACTTGCAGGAACAAATCA
AAGATCAAAAAAGGATAGAAGAACAAGAAAAACCACAA
ACACTTAGACAATCAATATAAAGATGAAGTGAACGCTC
TTAAAGAGAAGTTGGAAAACTTGCAGGAACAAATCAAA
GATCAAAAAAGGATAGAAGAACAAGAAAAACCACAAAC
ACTTAGACAATCAATATAAAGATGAAGTGAACGCTCTT
AAAGAGAAGTTGGAAAACTTGCAGGAACAAATCAAAGA
TCAAAAAAGGATAGAAGAACAAGAAAAACCACAAACAC
TTAGACAATCAATATAAAGATGAAGTGAACGCTCTTAA
AGAGAAGTTGGAAAACTTGCAGGAACAAATCAAAGATC
AAAAAAGGATAGAAGAACAAGAAAAACCACAAACACTT
AGACAATCAATATAAAGATGAAGTGAACGCTCTTAAAG
AGAAGTTGGAAAACTTGCAGGAACAAATCAAAGATCAA
AAAAGGATAGAAGAACAAGAAAAACCACAAACACTTAG
ACAATCAATATAAAGATGAAGTGAACGCTCTTAAAGAG
AAGTTGGAAAACTTGCAGGAACAAATCAAAGATCAAAA
AAGGATAGAAGAACAAGAAAAACCAC
```

Structural gene annotation:
Gene localization based on sequences

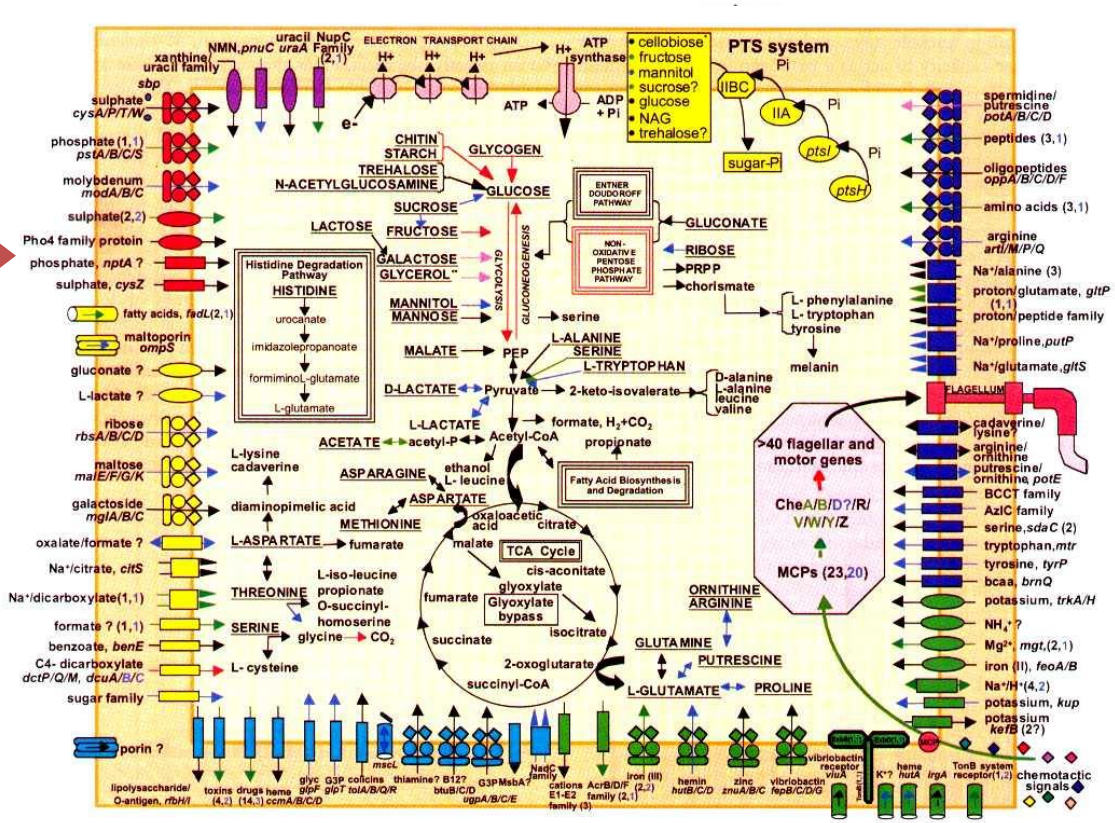


Genomic annotation: from sequence to predicted function

Raw sequence data: millions and millions of nucleotides

```
AAACACTTAGACAATCAATATAAAGATGAAGTGAACGC
TCTTAAAGAGAAGTTGGAAAACTTGCAGGAACAAATCA
AAGATCAAAAAAGGATAGAAGAACAAGAAAAACCACAA
ACACTTAGACAATCAATATAAAGATGAAGTGAACGCTC
TTAAAGAGAAGTTGGAAAACTTGCAGGAACAAATCAAA
GATCAAAAAAGGATAGAAGAACAAGAAAAACCACAAAC
ACTTAGACAATCAATATAAAGATGAAGTGAACGCTCTT
AAAGAGAAGTTGGAAAACTTGCAGGAACAAATCAAAGA
TCAAAAAAGGATAGAAGAACAAGAAAAACCACAAACAC
TTAGACAATCAATATAAAGATGAAGTGAACGCTCTTAA
AGAGAAGTTGGAAAACTTGCAGGAACAAATCAAAGATC
AAAAAAGGATAGAAGAACAAGAAAAACCACAAACACTT
AGACAATCAATATAAAGATGAAGTGAACGCTCTTAAAG
AGAAGTTGGAAAACTTGCAGGAACAAATCAAAGATCAA
AAAAGGATAGAAGAACAAGAAAAACCACAAACACTTAG
ACAATCAATATAAAGATGAAGTGAACGCTCTTAAAGAG
AAGTTGGAAAACTTGCAGGAACAAATCAAAGATCAAAA
AAGGATAGAAGAACAAGAAAAACCAC
```

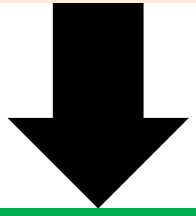
A virtual cell:
overview of predicted pathways



Genome annotation

Structural gene annotation

Structural genome annotation is the process of identifying the exact location of **genes** and all of the coding and non-coding regions in a **genome** and determining intron, exons and other elements



Functional gene annotation

Functional gene annotation means the description of the biochemical and biological function of proteins. Gene products, proteins and the description of protein domains as well as asigning Gene Ontology Annotations (GO terms)

Structural and Functional Annotation

Structural annotation

Identification of genomic elements.

- **ORFs predicted during genome assembly**
- **Location of ORFs**
- **Gene structure**
- **Coding regions**
- **Location of regulatory motifs etc**

Functional annotation

Attaching biological information to genomic elements.

- **Biochemical function**
- **Biological function**
- **Involved regulation and interactions**
- **Expression etc**

These steps may involve both biological experiments and *in silico* analysis.

Functional annotation

Attaching biological information to genomic elements.

- **Biochemical function**
- **Biological function**
- **Involved regulation and interactions**
- **Expression etc**

These steps may involve both biological experiments and *in silico* analysis.

- **predicted ORFs have no functional literature and GO annotation relies on computational methods**
- **Functional literature exists for many genes/proteins prior to genome sequencing (slow but provide high quality annotations)**

GO is a controlled vocabulary of terms split into three related ontologies covering basic areas of molecular biology:

- molecular function
- biological process
- cellular component

Types of Functional annotation

Based in direct experimental evidence of function

Biological assays

- Enzyme assays
- Binding experiments
- Pathway analysis
- Synthetic lethals
- Functional complementation
- Gene mutations
- RNAi
- 2-hybrid interactions etc

Genome project

Too many genes

Homology search

Indirect Evidence of function

- Expression analysis
- Structure analysis
- Sequence analysis

Functional Annotation

Problem:

- Many genes/proteins have no annotation
- Some have unknown functions

Challenge:

- We want to get the maximum functional annotation for modeling our data

Solution:

- Read papers (Pubmed etc)
- Search for homologs/orthologs of known function
- Homologs and orthologs help assign function….

Functional annotation

Concepts: orthologs and homologs

Homolog
Is a relationship between genes separated by the event of speciation or genetic duplication

Ortholog
Orthologs are homologous genes in different species that evolved from a common ancestor gene by speciation. Normally (not always), orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes.

Paralog
Paralogs are homologous genes related by duplication within a genome. Paralogs evolve new functions, even if these are related to the original one.

Why can I annotate gene by homology search?

“Based on the assumption that orthologs of a gene have the same function in the course of evolution”

Systems for Functional Annotation

1. **C**lusters of **O**rthologous **G**roups (**COGs**)

   → Prokaryotes

2. eu**K**aryote **O**rthologous **G**roups (**KOGs**)

   → Eukaryotes

3. **G**ene **O**ntology (**GO**)

- Both are based on orthology.
- Genes are assigned to broad categories (A-Z)
- Each category corresponds to an ancient conserved domain

  - COGs - prokaryotes
  - KOGs - eukaryotes

Clusters of Orthologous Groups (COGs)

http://www.ncbi.nlm.nih.gov/COG/

COGs has 25 functional categories (A – Z) in four broad groups

1. Information storage and processing
2. Cellular processes and signaling
3. Metabolism
4. Poorly characterized

COGs Categories

INFORMATION STORAGE AND PROCESSING

[J] Translation, ribosomal structure and biogenesis
[A] RNA processing and modification
[K] Transcription
[L] Replication, recombination and repair
[B] Chromatin structure and dynamics

Clusters of Orthologous Groups (COGs)

http://www.ncbi.nlm.nih.gov/COG/

COGs Categories

CELLULAR PROCESSES AND SIGNALING

[D] Cell cycle control, cell division, chromosome partitioning
[Y] Nuclear structure
[V] Defense mechanisms
[T] Signal transduction mechanisms
[M] Cell wall/membrane/envelope biogenesis
[N] Cell motility
[Z] Cytoskeleton
[W] Extracellular structures
[U] Intracellular trafficking, secretion, and vesicular transport
[O] Posttranslational modification, protein turnover, chaperones

Clusters of Orthologous Groups (COGs)

http://www.ncbi.nlm.nih.gov/COG/

COGs Categories

METABOLISM

[C] Energy production and conversion
[G] Carbohydrate transport and metabolism
[E] Amino acid transport and metabolism
[F] Nucleotide transport and metabolism
[H] Coenzyme transport and metabolism
[I] Lipid transport and metabolism
[P] Inorganic ion transport and metabolism
[Q] Secondary metabolites biosynthesis, transport and catabolism

POORLY CHARACTERIZED

[R] General function prediction only
[S] Function unknown

ftp://ftp.ncbi.nih.gov/pub/COG/COG/fun.txt

Gene Ontology overview

An ontology is a formal representation of a body of knowledge within a given domain. Ontologies usually consist of a set of classes (or terms or concepts) with relations that operate between them. The Gene Ontology (GO) describes our knowledge of the biological domain with respect to three aspects:

Molecular Function	Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as "catalysis" or "transport". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products (*i.e.* a protein or RNA), but some activities are performed by molecular complexes composed of multiple gene products. Examples of broad functional terms are *catalytic activity* and *transporter activity*; examples of narrower functional terms are *adenylate cyclase activity* or *Toll-like receptor binding*. To avoid confusion between gene product names and their molecular functions, GO molecular functions are often appended with the word "activity" (a *protein kinase* would have the GO molecular function *protein kinase activity*).
Cellular Component	The locations relative to cellular structures in which a gene product performs a function, either cellular compartments (*e.g.*, *mitochondrion*), or stable macromolecular complexes of which they are parts (*e.g.*, the *ribosome*). Unlike the other aspects of GO, cellular component classes refer not to processes but rather a cellular anatomy.
Biological Process	The larger processes, or 'biological programs' accomplished by multiple molecular activities. Examples of broad biological process terms are *DNA repair* or *signal transduction*. Examples of more specific terms are *pyrimidine nucleobase biosynthetic process* or *glucose transmembrane transport*. Note that a biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

The Gene Ontology (GO)

- The Gene Ontology (GO) is the *de facto* Standard for functional annotation
- GO functional annotation is based on orthology AND direct experimental evidence
- GO terms allow much more detailed functional analysis (> 24,000 terms) than COGs & KOGs (25 broad terms)
- GO is a controlled vocabulary of terms split into three related ontologies covering basic areas of molecular biology:

1.	Biological process terms	28890
2.	Molecular function terms	11178
3.	Cellular component terms	4196

2020-10

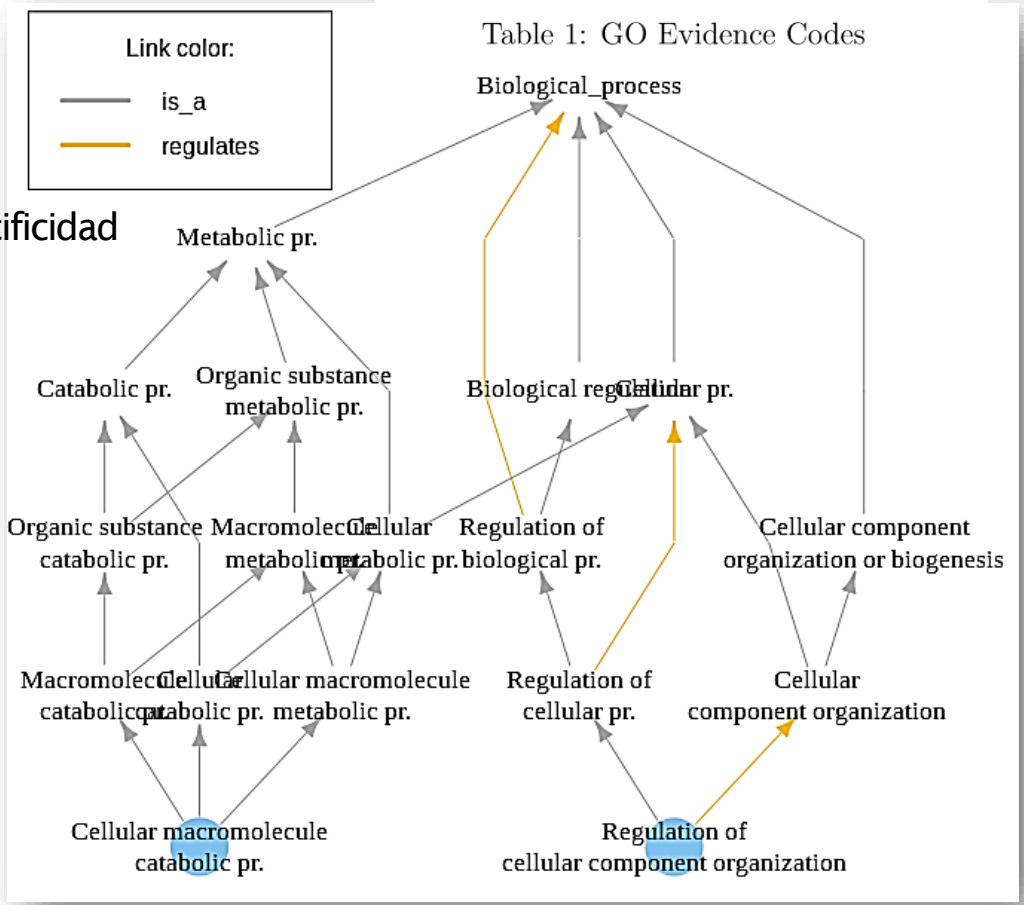
http://geneontology.org/stats.html

Ontologías GO

IMP	inferred from mutant phenotype
IGI	inferred from genetic interaction
IPI	inferred from physical interaction
ISS	inferred from sequence similarity
IDA	inferred from direct assay
IEP	inferred from expression pattern
IEA	inferred from electronic annotation
TAS	traceable author statement
NAS	non-traceable author statement
ND	no biological data available
IC	inferred by curator

Table 1: GO Evidence Codes

- Modelo de la realidad
  - Vocabulario controlado
  - Definiciones detérminos
  - Definición deconexiones
- Grafo con términos y relaciones
  - Nuevos niveles de complejidad
  - Estructura → Terminos hijos amplian especificidad
- Lenguaje común
  - Unificar notación / nomenclatura
  - Formalización del lenguaje
  - Comunicación entre científicos
  - Búsquedas automáticas



Use GO for…….

- Modeling function in high-throughput datasets (arrays!) started by Fly, Yeast, Mouse (Ashburner et al 2000, 2001)
- Grouping gene products by biological function
- Determining which classes of gene products are over-represented or under-represented
- Focusing on particular biological pathways and functions (*hypothesis-driven*)
- Relating a protein's location to its function

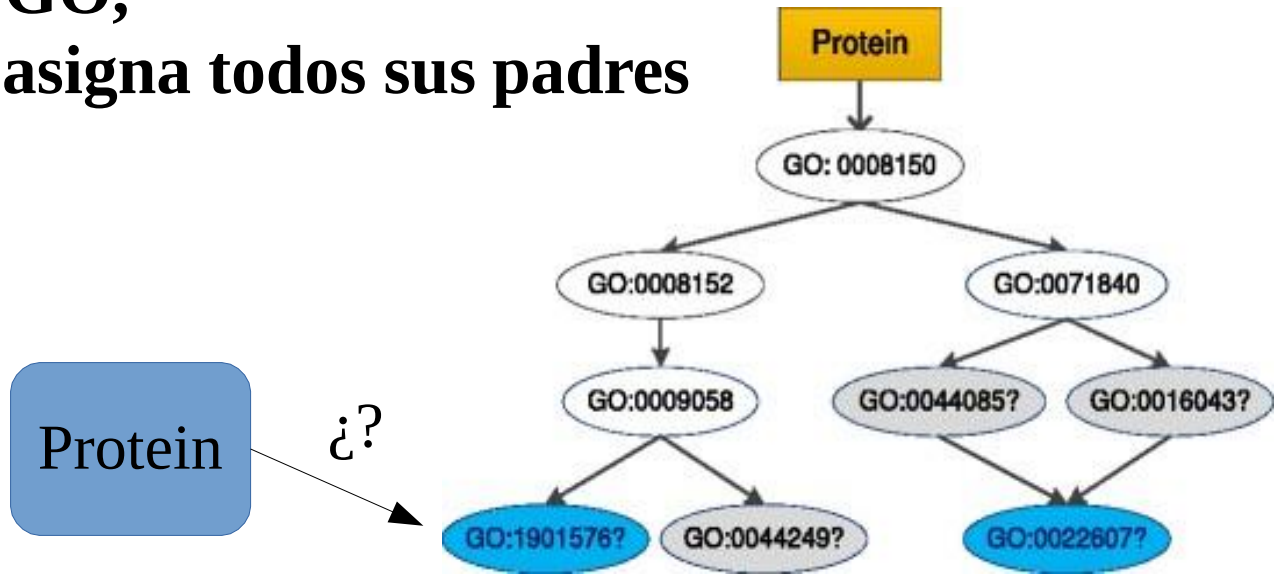
How to find functional annotation for your species

How to search for Orthology?

Anotación → Asociar secuencia con termino GO

Se intenta anotar la proteína en el GO mas especifico que sea correcto

Al asignar un GO,
También se le asigna todos sus padres



How to search for Orthology?

BLAST : http://www.ncbi.nlm.nih.gov/BLAST/

- Sequence alignment search tool
- Utilizes heuristic algorithm

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

- Sequence comparison tool
- Implement Smith & Waterman algorithm
- Utilizes exhaustive algorithm

Domain analysis: http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

- Analysis of regions of sequence homology among sets of proteins that are not all full-length homologs.
- Homology domains often, but not always, correspond to recognizable protein folding domains

Protein family databases (e.g. COGs & KOGs)

- **Superfamily:** Complete set of proteins having sequence homology over essentially their full length.
- **Subfamilies:** Incomplete set of homologous proteins which yet encompass proteins of diverse function

Enzime databases (KEGGs)

Anotación usando dominios

- Utilización de dominios:

- Familia → Grupo de proteínas con una función común Dominio
  → Unidad evolutiva básica
- La función de una proteína es el resultado de las funciones de sus dominios
- Proteínas homólogas pueden tener diferente organización de dominios
- Búsquedas basadas en regiones conservadas de proteinas

  - Pequeñas zonas conservadas
  - Caracteres funcionales
    - Centros Activos
    - Sitios de unión de ligandos

Anotación por similitud

¿Como alineamos solo en las regiones correspondientes a los dominios?

InterPro2GO mapping

InterPro is an integrated resource of protein families, domains and sites which are combined from a number of different protein signature databases, including. Gene3D, Panther, PRSF, Pfam, PRINTS, ProSite, ProDom, SMART, SUPERFAMILY and TIGRFAMs.

Anotación usando dominios

Bases de datos deperfiles:

- *InterProScan*
  - Secuencias consenso
  - Patrones
  - Perfiles simples yHMM
  - Organiza información de conjunto de programas
    - PFAM
    - TIGRFAM
    - TMHMM
    - signalP

    …


UniProt
Universal Protein Resource
proSite
Fasta
Similarity and Homology Searching
WU-Blast2
Basic Local Alignment Search Tool
IntEnz
SeqVISTA
Pfam
PRINTS
Protein Fingerprints Database
InterPro
GENOMES
ClustalW
Multiple Sequence Alignment Tool
emboss
CAST
ProDom
SMART
Alternative Splicing Database
Integr8
CSA
Catalytic Site Atlas
PDBj
Protein Data Bank Japan
TIGR
tigr fams
PANTHER
Classification System
caBIG
SeqHound
PIRSF
Gene3D
BioLayout
NCBI
LION
Superfamily
SCOP
BALI
GeneQuiz
PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
caBIO
MOUNT SINAI HOSPITAL
CATH
MSD
Ensembl
Parasite Blast
DDBJ
DNA Data Bank of Japan
wellcome trust
sanger
institute
SWISS-MODEL
MODBASE
GENOME REVIEWS
IUPHAR
KEGG
Kyoto Encyclopedia of Genes and Genomes
PathPort
The Pathogen Portal Web Project

The InterPro Consortium

The following databases make up the InterPro Consortium:

CATH-Gene3D database describes protein families and domain architectures in complete genomes. Protein families are formed using a Markov clustering algorithm, followed by multi-linkage clustering according to sequence identity. Mapping of predicted structure and sequence domains is undertaken using hidden Markov models libraries representing CATH and Pfam domains. CATH-Gene3D is based at University College, London, UK.

CDD is a protein annotation resource that consists of a collection of annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domain models, which use 3D-structure information to explicitly define domain boundaries and provide insights into sequence/structure/function relationships, as well as domain models imported from a number of external source databases.

MobiDB offers a centralized resource for annotations of intrinsic protein disorder. The database features three levels of annotation: manually curated, indirect and predicted. The different sources present a clear tradeoff between quality and coverage. By combining them all into a consensus annotation, MobiDB aims at giving the best possible picture of the "disorder landscape" of a given protein of interest.

HAMAP stands for High-quality Automated and Manual Annotation of Proteins. HAMAP profiles are manually created by expert curators. They identify proteins that are part of well-conserved proteins families or subfamilies. HAMAP is based at the SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.

PANTHER is a large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise. These subfamilies model the divergence of specific functions within protein families, allowing more accurate association with function, as well as inference of amino acids important for functional specificity. Hidden Markov models (HMMs) are built for each family and subfamily for classifying additional protein sequences. PANTHER is based at at University of Southern California, CA, US.

Pfam	Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. Pfam is based at EMBL-EBI, Hinxton, UK.
PIRSF	PIRSF protein classification system is a network with multiple levels of sequence diversity from superfamilies to subfamilies that reflects the evolutionary relationship of full-length proteins and domains. PIRSF is based at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, US.
PRINTS	PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family or domain. PRINTS is based at the University of Manchester, UK.
ProDom	ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches. ProDom is based at PRABI Villeurbanne, France.
prosite	PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family a new sequence belongs. PROSITE is base at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland.
SFLD	SFLD (Structure-Function Linkage Database) is a hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities.
SMART	SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. SMART is based at at EMBL, Heidelberg, Germany.
Superfamily	SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent the entire SCOP superfamily that the domain belongs to. SUPERFAMILY is based at the University of Bristol, UK.
	TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. TIGRFAMs is based at the J. Craig Venter Institute, Rockville, MD, US.

Anotación por similitud

Pasos para la transferencia de anotación a partir de similitud con secuencias de función conocida

- Blast2GO:
  1. Alineamiento Local (BLAST)
  2. Identificación de Dominios Funcionales
  3. Mapeo a GO terms
  4. Transferencia de anotación

Anotación por similitud

¿Cuales son los problemas de esta metodología?

¿Que errores podemoscometer?

Anotación por similitud

Problemas:

- Transferencia de anotacioneserroneas
  - Genes mal anotados presentes en bases de datos
- BLAST no alinea solo en regiones funcionales
  - Regiones repetitivas /UTRs

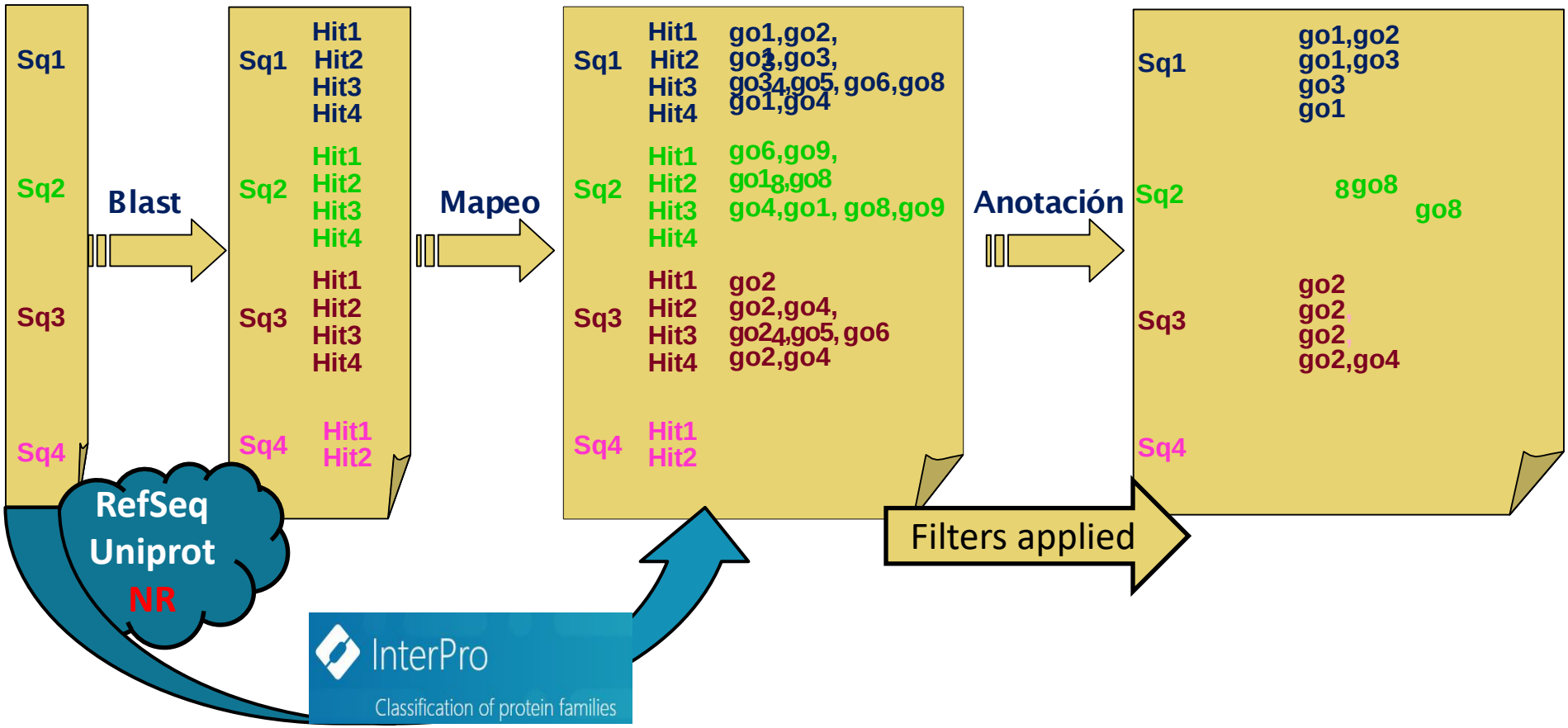
Soluciones:

- Usar bases de datos curadas → **UniProt/SwissProt – RefSeq**
- Revisar mas de un ***hit*** y sus anotaciones
- Alineamiento de regiones representativas → Dominios funcionales

– InterProScan

Anotación por similitud Blast2GO

How does Blast2GO work?



Anotación por similitud Blast2GO

% of similarity

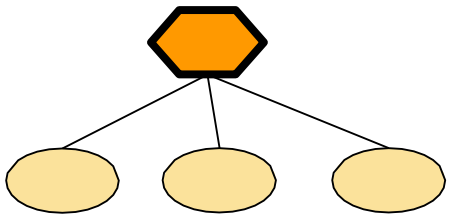
$$sim = \frac{\sum positives_{hsp}}{\sum alignmentlength_{hsp}}$$

Annotation source

EC	weight
IC	1
TAS	1
IDA	1
IMP	0.9
IGI	0.9
IPI	0.9
ISS	0.8
IEP	0.8
NAS	0.7
IEA	0.7
ND	0.5
NR	0.5
RCA	0.5

Evidence codes

Graph structure



The lowest term per branch (55)

$$\%sim \times ECw + (N_{GO} - 1) \times GOw \geqslant threshold$$

GO-Weight. This is the weight given to the contribution of mapped children terms to the annotation of a parent term (default=5).

Anotación por similitud

$$\%sim \times ECw + (N_{GO} - 1) \times GOw \geqslant threshold$$

ECw (IDA)=1; ECw(ISS) = 0.8; ECw(IEA) = 0.7; umbral = 55; **GOw = 0**

Ejemplo:

Para una secuencia con 3 hits con los siguientes términos GO asociados:

- Hit 1: **60%** similitud; Término GO: GO1 con EC = **IDA**
- Hit 2: **65%** similitud; Término GO: GO2 con EC = **ISS**
- Hit 3: **74%** similitud; Término GO: GO3 con EC = **IEA**
- Hit 4: **74%** similitud; Término GO: GO4 con EC = **IEA**
- GO2 y GO3 son hermanos que comparten el padre GO4

IDA = Inferido por un ensayo directo
ISS = Inferido por similitud de secuencia IEA = Anotaciónelectrónica

Anotación por similitud

Caso 1

$$\%sim \times ECw + (N_{GO} - 1) \times GOw \geqslant threshold$$

ECw (IDA)=1; ECw(ISS) = 0.8; ECw(IEA) = 0.7; umbral = 55; **GOw = 0**

AS(GO1) = (60 * 1) + (1-1) * **0** = 60 > 55 ➔ **GO1 asignado**
AS(GO2) = (65 * 0.8) + (1-1) * **0** = 52 < 55 ➔ GO2 no asignado
AS(GO3) = (74 * 0.7) + (1-1) * **0** = 52 < 55 ➔ GO3 no asignado
AS(GO4) = (74 * 0.7) + (2-1) * **0** = 54 < 55 ➔ GO4 no asignado

- Hit 1: **60%** similitud; Término GO: GO1 con EC = **IDA**
- Hit 2: **74%** similitud; Término GO: GO2 con EC = **ISS**
- Hit 3: **74%** similitud; Término GO: GO3 con EC = **IEA**
- Hit 4: **74%** similitud; Término GO: GO4 con EC = **IEA**
- GO2 y GO3 son hermanos que comparten el padre GO4

Anotación por similitud

Caso 1

$$\%sim \times ECw + (N_{GO} - 1) \times GOw \geqslant threshold$$

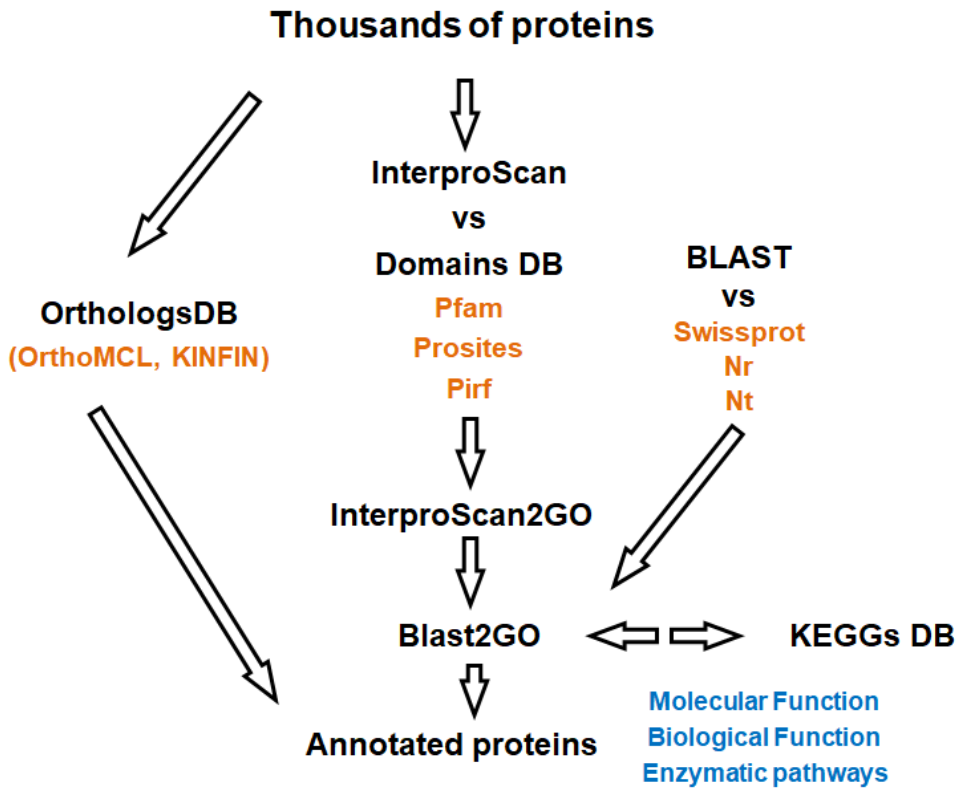
ECw (IDA)=1; ECw(ISS) = 0.8; ECw(IEA) = 0.7; umbral = 55; **GOw = 5**

AS(GO1) = (60 * 1) + (1-1) * **5** = 60 > 55 ➔ **GO1 asignado**
AS(GO2) = (65 * 0.8) + (1-1) * **5** = 52 < 55 ➔ GO2 no asignado
AS(GO3) = (74* 0.7) + (1-1) * **5** = 52 < 55 ➔ GO3 no asignado
AS(GO4) = (74* 0.7) + (2-1) * **5** = 57 > 55 ➔ **GO4 asignado**

- Hit 1: **60%** similitud; Término GO: GO1 con EC = **IDA**
- Hit 2: **74%** similitud; Término GO: GO2 con EC = **ISS**
- Hit 3: **74%** similitud; Término GO: GO3 con EC = **IEA**
- Hit 4: **74%** similitud; Término GO: GO4 con EC = **IEA**
- GO2 y GO3 son hermanos que comparten el padre GO4

How to search for Orthology?

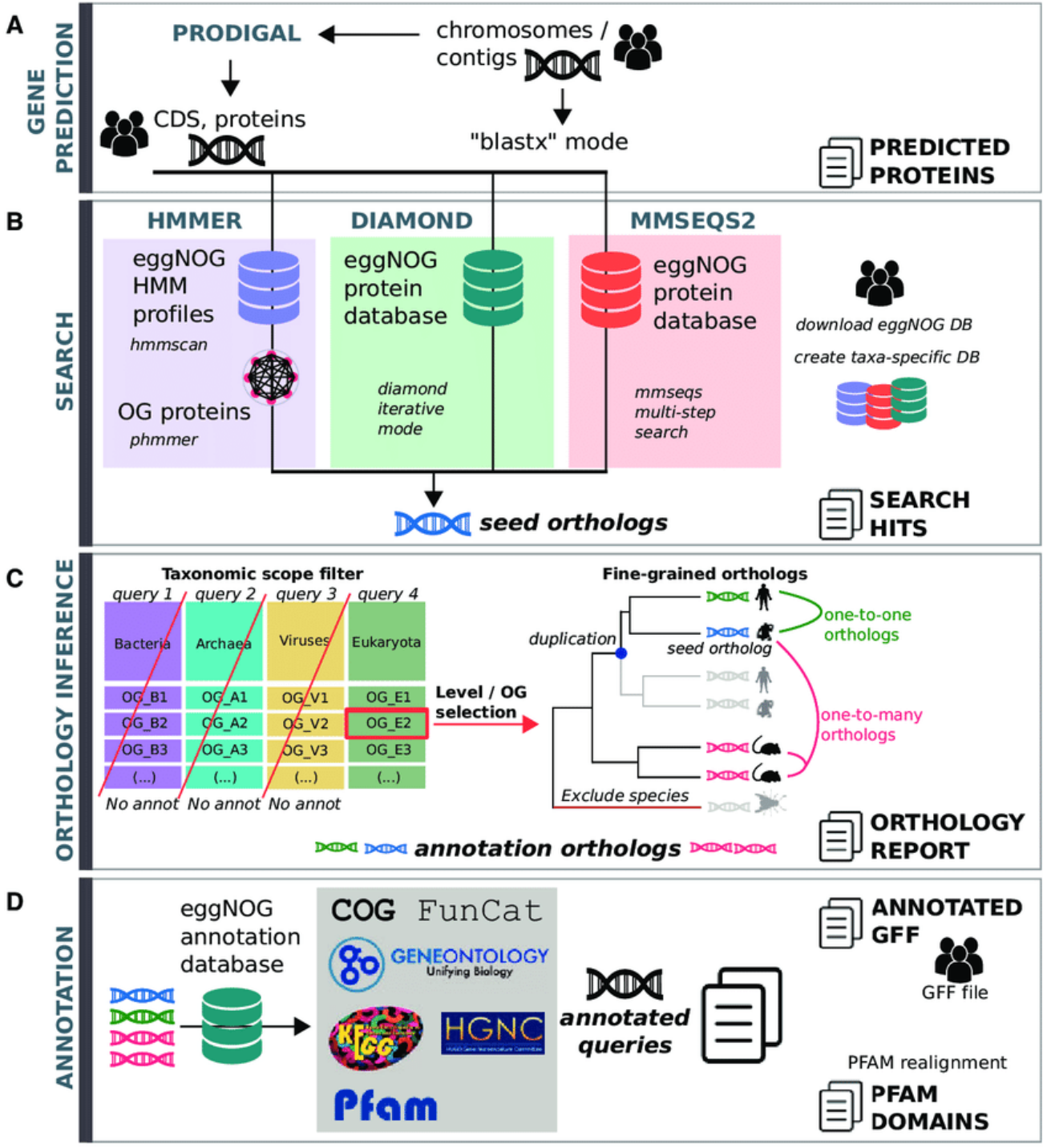
Integrative approach:



- Grouping gene products by biological function
- Determining which classes of gene products are over-represented or under-represented
- Focusing on particular biological pathways and functions (*hypothesis-driven*)
- Relating a protein's location to its function

What can I do with this information?

Only curated DB


A
GENE PREDICTION
PRODIGAL
chromosomes / contigs
CDS, proteins
"blastx" mode
PREDICTED PROTEINS
B
SEARCH
HMMER
DIAMOND
MMSEQS2
eggNOG HMM profiles
hmmscan
OG proteins
phmmer
eggNOG protein database
diamond iterative mode
eggNOG protein database
mmseqs multi-step search
download eggNOG DB
create taxa-specific DB
seed orthologs
SEARCH HITS
C
ORTHOLOGY INFERENCE
Taxonomic scope filter
query 1
query 2
query 3
query 4
Bacteria
Archaea
Viruses
Eukaryota
OG_B1
OG_B2
OG_B3
(...)
OG_A1
OG_A2
OG_A3
(...)
OG_V1
OG_V2
OG_V3
(...)
OG_E1
OG_E2
OG_E3
(...)
No annot
No annot
No annot
Level / OG selection
Fine-grained orthologs
duplication
seed ortholog
one-to-one orthologs
one-to-many orthologs
Exclude species
annotation orthologs
ORTHOLOGY REPORT
D
ANNOTATION
eggNOG annotation database
COG FunCat
GENEONTOLOGY
Unifying Biology
KEGG
HGNC
Pfam
annotated queries
ANNOTATED GFF
GFF file
PFAM realignment
PFAM DOMAINS