

# Tema 4. Bases de datos secundarias

Actualizado en: 24/01/2023

# Objetivos

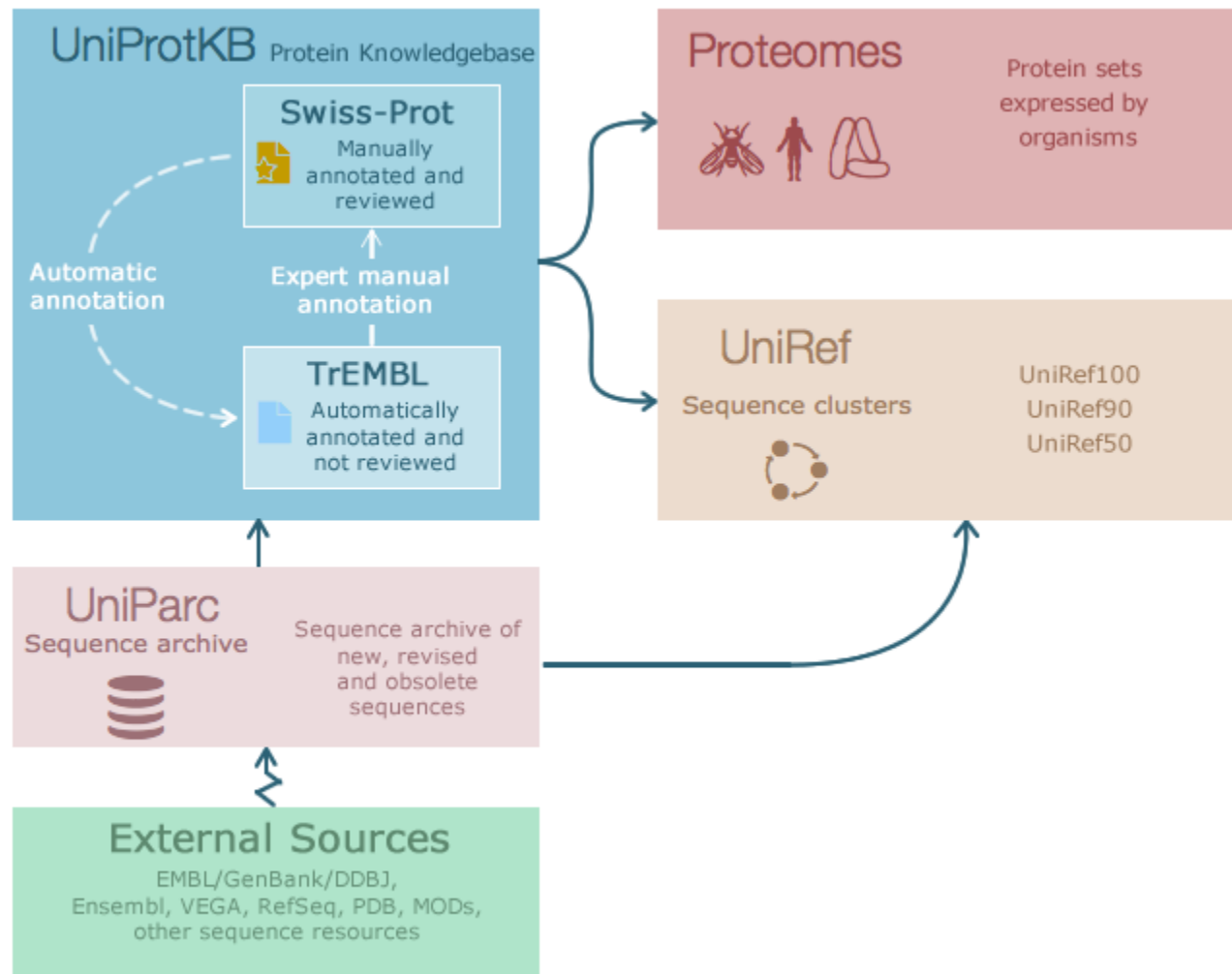
- UniProt
- Pfam
- Rfam
- ENSEMBL
- KEGG

# UniProt

---

“The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). The UniProt consortium and host institutions EMBL-EBI, SIB and PIR are committed to the long-term preservation of the UniProt databases.”

# UniProt



# UniProt. Cuestionario.

- ¿En qué sección de UniProt buscarías información sobre la función de una proteína?
- ¿Qué sección es menos *redundante*?
- ¿Por qué coexisten TrEMBL y Swiss-Prot?
- ¿De dónde salen los datos primarios?

# Pfam

“Pfam version 35.0 was produced at the European Bioinformatics Institute using a sequence database called Pfamseq, which is based on UniProt release 2021\_03.

Pfam is freely available under the Creative Commons Zero (“CC0”) licence.

Pfam is powered by the HMMER3 package written by Sean Eddy and his group at HHMI/ Harvard University, and built by the Xfam consortium.”

# Pfam. HMMER3

Start with a multiple  
sequence alignment

↓

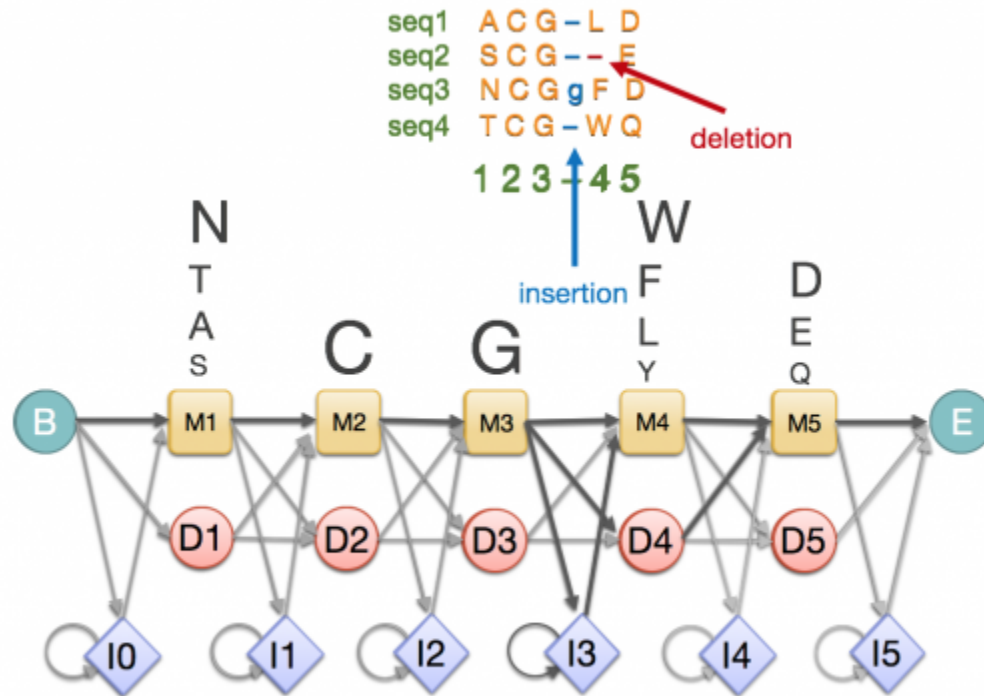
Insertions / deletions can  
be modelled

↓

Occupancy and amino acid  
frequency at each position in  
the alignment are encoded

↓

Profile created



# Pfam. Para qué sirve

- Buscar proteína o DNA contra los models (perfiles HMM).
- Explorar familias de proteínas y *clanes*.
- Obtener anotación funcional de familias o dominios.
- Ver alineamientos múltiples de una familia o clan.
- Ver las relaciones entre las familias de un clan.
- Ver información de estructura tridimensional de una familia.
- Ver familias de acuerdo a su distribución taxonómica.
- Buscar la base de datos por palabras clave.



# Pfam. Categorías de los registros

- Dominio.
- Familia.
- Repetición (*repeat*).
- Motivo.
- *Coiled-coil*
- Desordenada.

# Pfam. Clanes

Registros en Pfam relacionados por semejanza de: - secuencia, - estructura, - función, - o perfil.

## Pfam 33.1 (May 2020, 18259 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

### QUICK LINKS

[SEQUENCE SEARCH](#)

[VIEW A PFAM ENTRY](#)

[VIEW A CLAN](#)

[VIEW A SEQUENCE](#)

[VIEW A STRUCTURE](#)

[KEYWORD SEARCH](#)

[JUMP TO](#)

### YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

Analyze your protein sequence for Pfam matches

View Pfam annotation and alignments

See groups of related entries

Look at the domain organisation of a protein sequence

Find the domains on a PDB structure

Query Pfam by keywords

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

# Pfam. Resultados

## Sequence search results

[Show](#) the detailed description of this results page.

We found **9** Pfam-A matches to your search sequence (**6** significant and **3** insignificant)



[Show](#) the search options and sequence that you submitted.

[Return](#) to the search form to look for Pfam domains on a new sequence.

## Significant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
<a href="#">PI-PLC-X</a>	Phosphatidylinositol-specific phospholip ...	Domain	<a href="#">CL0384</a>	322	465	322	465	1	145	145	218.3	2.9e-65	n/a	<a href="#">Show</a>
<a href="#">SH2</a>	SH2 domain	Domain	<a href="#">CL0541</a>	550	639	550	639	1	77	77	85.3	2.2e-24	n/a	<a href="#">Show</a>
<a href="#">SH2</a>	SH2 domain	Domain	<a href="#">CL0541</a>	668	741	668	741	1	77	77	72.5	2.3e-20	n/a	<a href="#">Show</a>
<a href="#">PI-PLC-Y</a>	Phosphatidylinositol-specific phospholip ...	Domain	<a href="#">CL0384</a>	953	1068	953	1067	1	<b>114</b>	115	133.7	3.6e-39	n/a	<a href="#">Show</a>
<a href="#">SH3_1</a>	SH3 domain	Domain	<a href="#">CL0010</a>	797	843	797	843	1	48	48	54.2	7.7e-15	n/a	<a href="#">Show</a>
<a href="#">C2</a>	C2 domain	Domain	<a href="#">CL0154</a>	1088	1193	1090	1187	<b>3</b>	<b>95</b>	103	53.3	2.7e-14	n/a	<a href="#">Show</a>

## Insignificant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					

# Pfam. Acceso a los datos

- Descarga de alineamientos, etc.
- Descarga de tablas que componen la base de datos entera (ftp).
- Acceso programático con API **RESTful**.

# Rfam

[HOME](#)[SEARCH](#)[BROWSE](#)[FTP](#)[BLOG](#)[HELP](#)[COVID-19](#)[Q Search](#)

## Rfam 14.4 (December 2020, 3941 families)

The Rfam database is a collection of RNA families, each represented by **multiple sequence alignments**, **consensus secondary structures** and **covariance models**

[Learn more about Rfam →](#)

## Q Search and browse Rfam

[Q Search](#)

Examples: [SAM](#), [Homo sapiens](#), [snoRNA](#), [author:"Weinberg"](#)

Secuencias genómicas



Predicción automática de genes



Base de datos MySQL



Análisis y visualización

# ENSEMBL

- Servicio del EMBL-EBI.
- Originalmente para genoma humano.
- Incluye genomas de 311 especies.
- La base de datos y sus herramientas estan disponibles.



# ENSEMBL. Acceso a los datos

- Consulta y visualización.
- Opciones de descarga de resultados de búsqueda.
- Descarga de tablas por ftp.
- Perl API.
- REST API (independiente de lenguaje).

**Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog Login/Register

Human (GRCh38.p13) Search all species...

Location: 17:63,992,802-64,038,237

**Location-based displays**

- Whole genome
- Chromosome summary
- Region overview
- Region in detail**
  - Comparative Genomics
    - Synteny
    - Alignments (image)
    - Alignments (text)
    - Region Comparison
  - Genetic Variation
    - Variant table
    - Resequencing
    - Linkage Data
  - Markers
  - Other genome browsers
    - UCSC
    - NCBI
    - Ensembl GRCh37

**Configure this page**

**Custom tracks**

**Export data**

**Share this page**

**Bookmark this page**

## Chromosome 17: 63,992,802-64,038,237

Assembly exceptions

Chr. 17

Assembly exceptions

### Region in detail

Scroll: ◀ ▶ Track height: ⏏ Drag/Select: ↔

Forward strand

Chromosome bands

Contigs

Genes (Comprehensive set from GENCODE 36)

Regulatory Build

Gene Legend

Regulation Legend

Location: 17:63992802-64038237 Go Gene: Go

Chromosome bands

45.44 kb

64.00Mb 64.01Mb 64.02Mb 64.03Mb

q23.3

Forward strand

# KEGG

“KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.”



KEGG ▾

Search

[Help](#)

[» Japanese](#)

## KEGG Home

[Release notes](#)  
[Current statistics](#)

## KEGG Database

[KEGG overview](#)  
[Searching KEGG](#)  
[KEGG mapping](#)  
[Color codes](#)

## KEGG Objects

[Pathway maps](#)  
[Brite hierarchies](#)  
[KEGG DB links](#)

## KEGG Software

[KEGG API](#)  
[KGML](#)

## KEGG FTP

[Subscription](#)  
[Background info](#)

[GenomeNet](#)

## KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See [Release notes](#) (January 1, 2021) for new and updated features.

**New article** [KEGG: integrating viruses and cellular organisms](#)

### Main entry point to the KEGG web service

**KEGG2** [KEGG Table of Contents](#) [\[Update notes\]](#) [\[Release history\]](#)

### Data-oriented entry points

**KEGG PATHWAY** [KEGG pathway maps](#)  
**KEGG BRITE** [BRITE hierarchies and tables](#)  
**KEGG MODULE** [KEGG modules](#)  
**KEGG ORTHOLOGY** [KO functional orthologs](#) [\[Annotation\]](#)  
**KEGG GENOME** [Genomes](#) [\[Pathogen\]](#) [\[Virus\]](#) [\[Plant\]](#)  
**KEGG GENES** [Genes and proteins](#) [\[SeqData\]](#)  
**KEGG COMPOUND** [Small molecules](#)  
**KEGG GLYCAN** [Glycans](#)

### Classification

[Pathway](#)  
[Brite](#)  
[Brite table](#)  
[Module](#)  
[KO \(Function\)](#)  
[Organism](#)  
[Virus](#)  
[Compound](#)