Tema 4. Bases de datos secundarias

Actualizado en: 13/02/2023

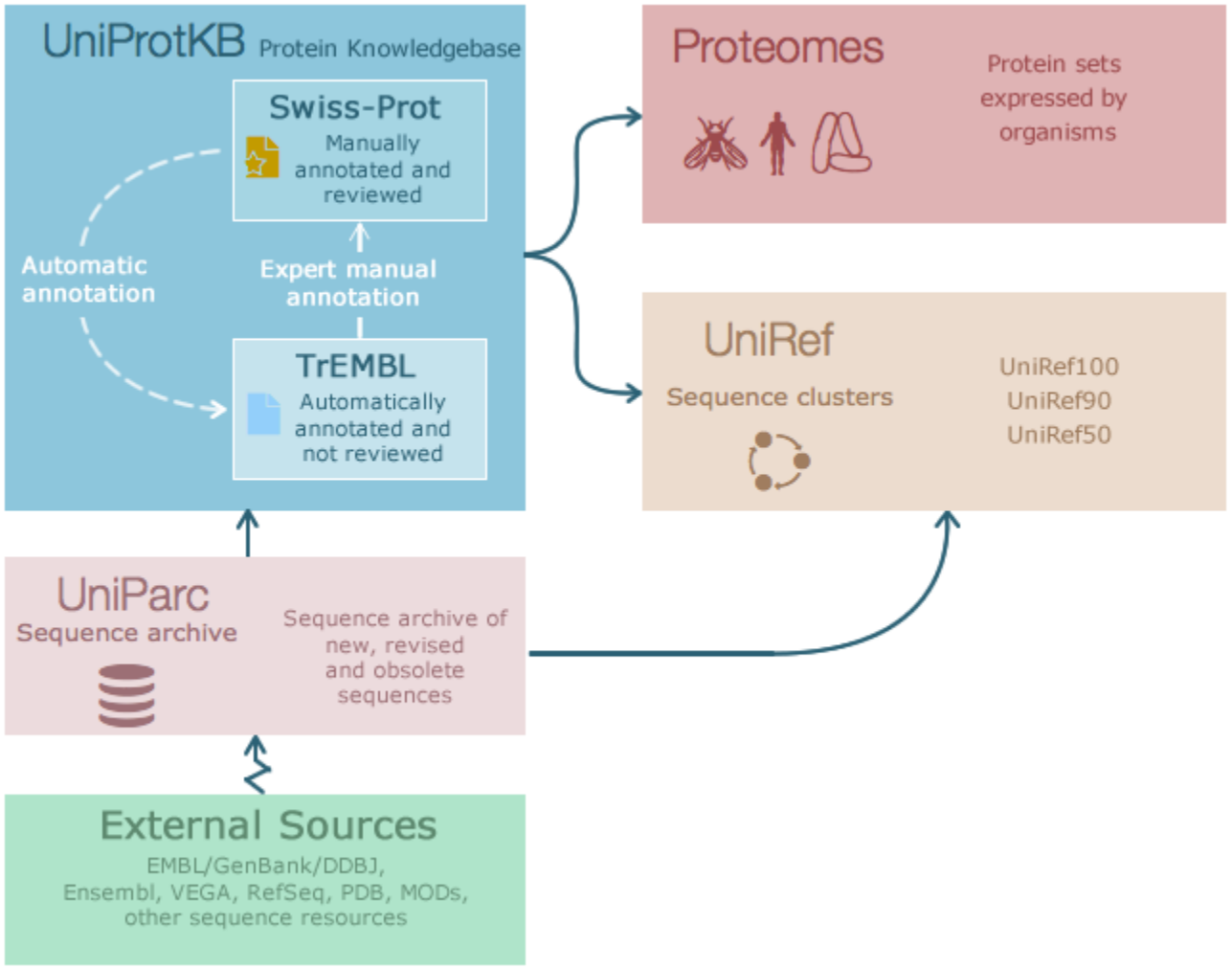
Objetivos

- UniProt
- Interpro
- ENSEMBL
- OrthoDB

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.”

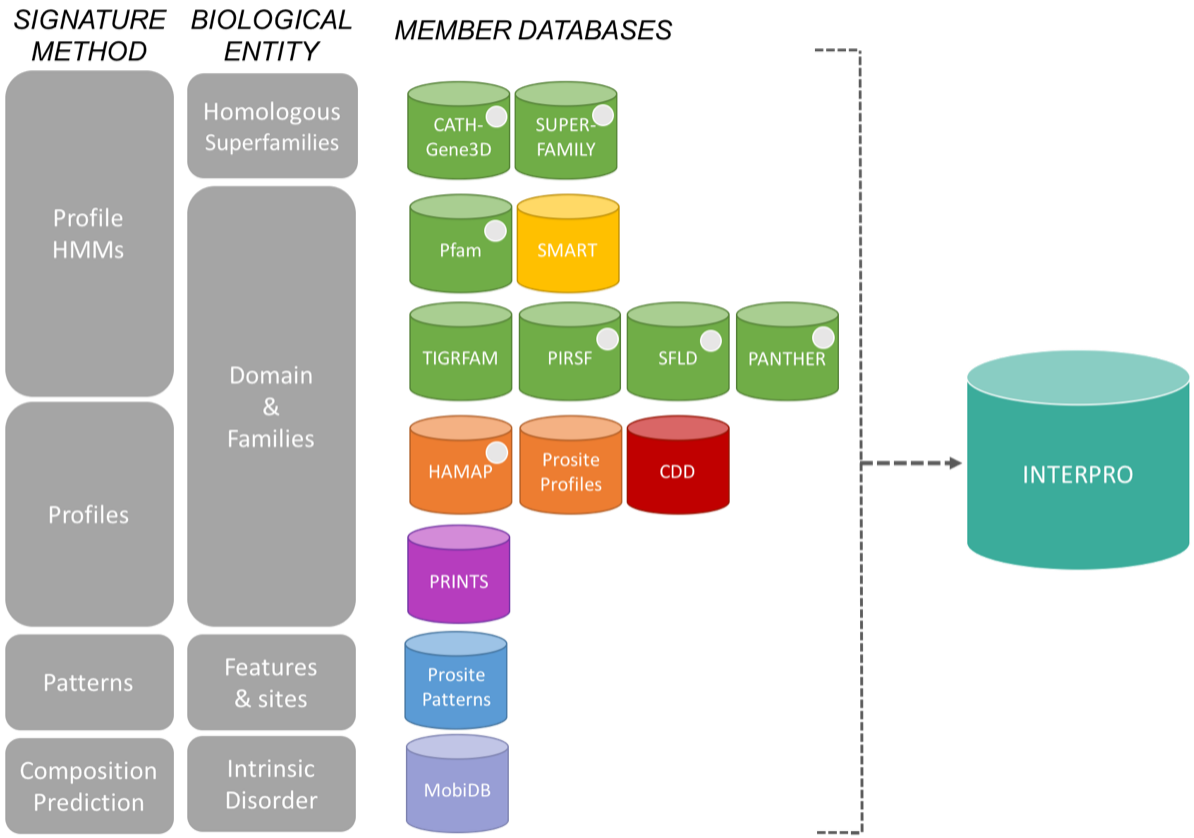
UniProt


UniProtKB Protein Knowledgebase
Swiss-Prot
Manually annotated and reviewed
Automatic annotation
Expert manual annotation
TrEMBL
Automatically annotated and not reviewed
Proteomes
Protein sets expressed by organisms
UniRef
Sequence clusters
UniRef100
UniRef90
UniRef50
UniParc
Sequence archive
Sequence archive of new, revised and obsolete sequences
External Sources
EMBL/GenBank/DDBJ, Ensembl, VEGA, RefSeq, PDB, MODs, other sequence resources

UniProt. Cuestionario.

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

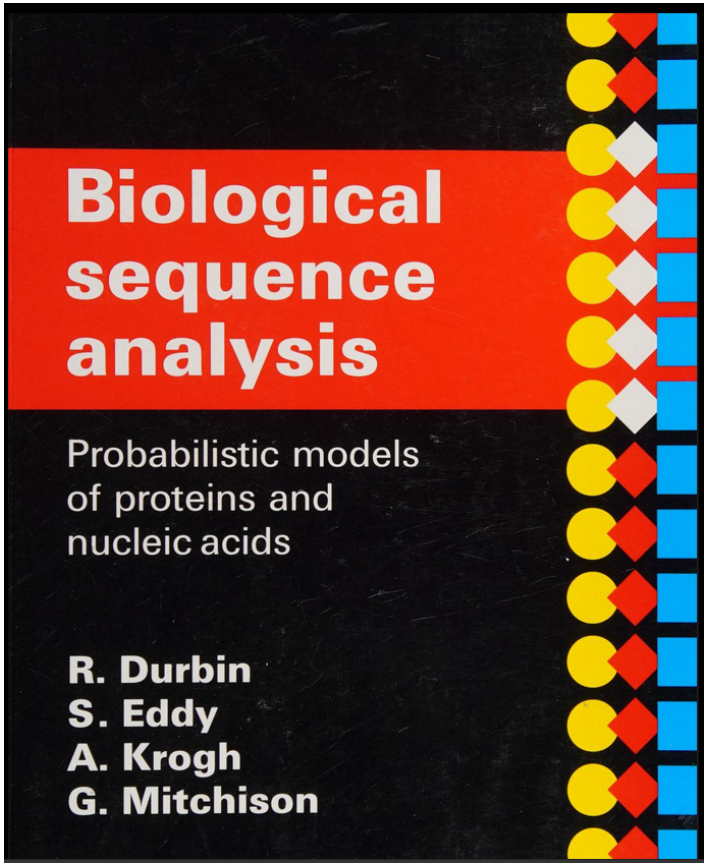
Interpro

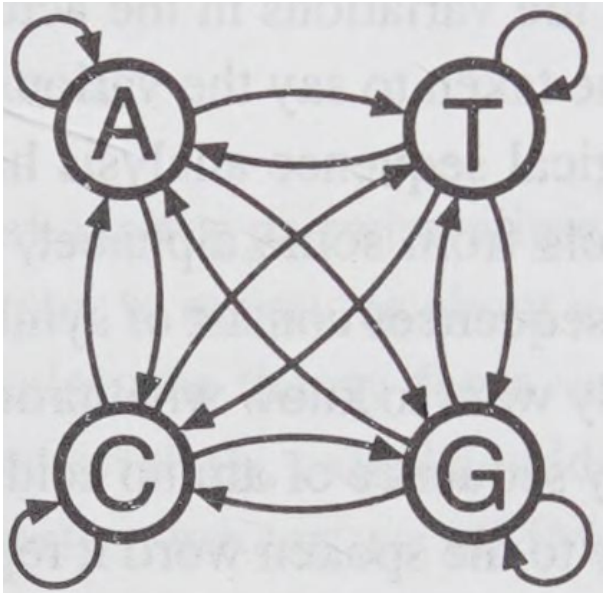

SIGNATURE METHOD
BIOLOGICAL ENTITY
MEMBER DATABASES
Profile HMMs
Profiles
Patterns
Composition Prediction
Homologous Superfamilies
Domain & Families
Features & sites
Intrinsic Disorder
CATH-Gene3D
SUPER-FAMILY
Pfam
SMART
TIGRFAM
PIRSF
SFLD
PANTHER
HAMAP
Prosite Profiles
CDD
PRINTS
Prosite Patterns
MobiDB
INTERPRO

Interpro. Tipos de registros

- **Familia**: Proteínas homólogas con secuencias, estructuras y funciones relacionadas.
- **Dominio**: Unidad funcional y estructural de las proteínas.
- **Sitio**: Motivos cortos: sitios de unión, sitios activos…
- **Repetición**: Secuencia de <50 aa. repetida muchas veces en una proteína.
- **Superfamilia**: Proteínas homólogas, aunque las secuencias no se parezcan.
- **No integrados**: Registros de las bases de datos participantes, no incluídos en Interpro.

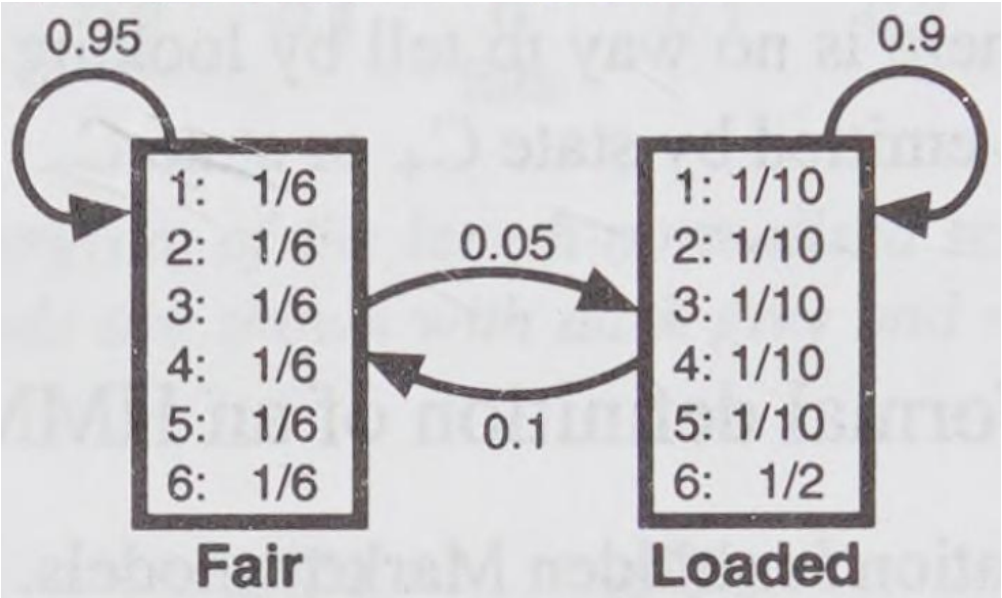
Modelo de Markov





$$a_{st} = P(x_i = t | x_{i-1} = s)$$

Modelo de Markov Oculto (HMM)



Ejemplo del casino con dos dados, uno justo y otro *cargado*, con probabilidades de *emisión* diferentes. No sabemos en cada momento con qué dado se está jugando.

Modelo de Markov Oculto (HMM)

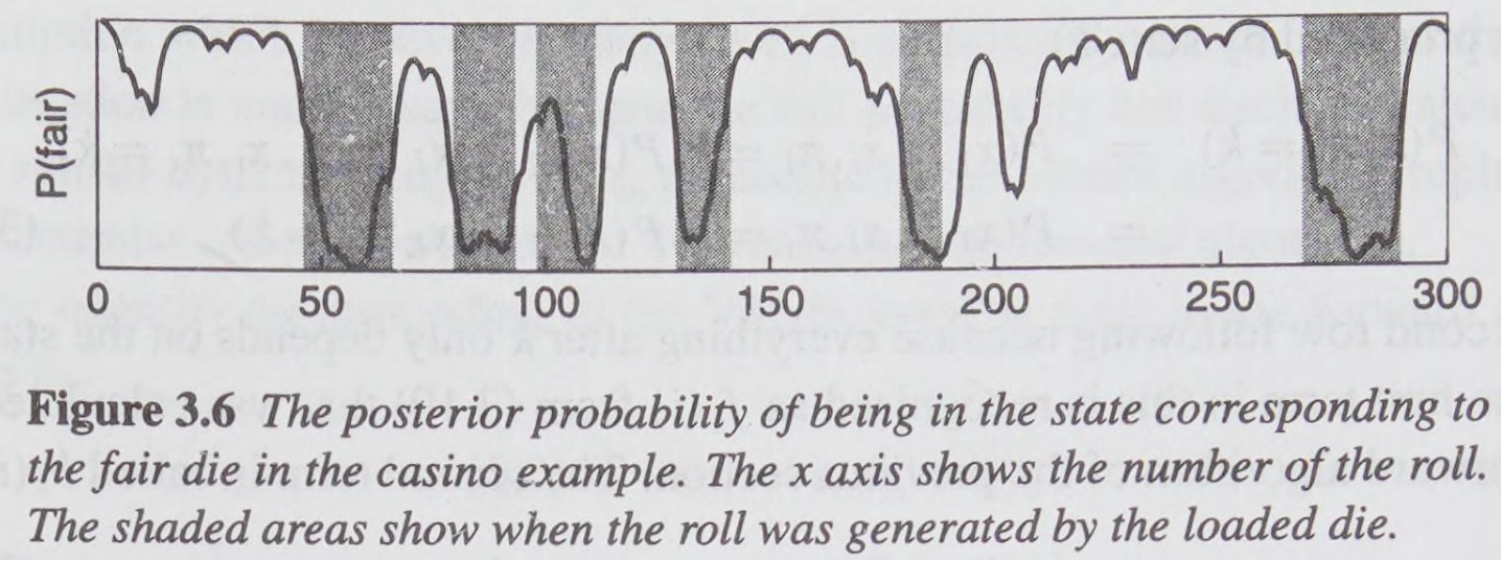


Figure 3.6 *The posterior probability of being in the state corresponding to the fair die in the casino example. The x axis shows the number of the roll. The shaded areas show when the roll was generated by the loaded die.*

A partir de los datos visibles, podemos ajustar los parámetros del modelo de Markov y predecir los estados ocultos a lo largo de la cadena de eventos.

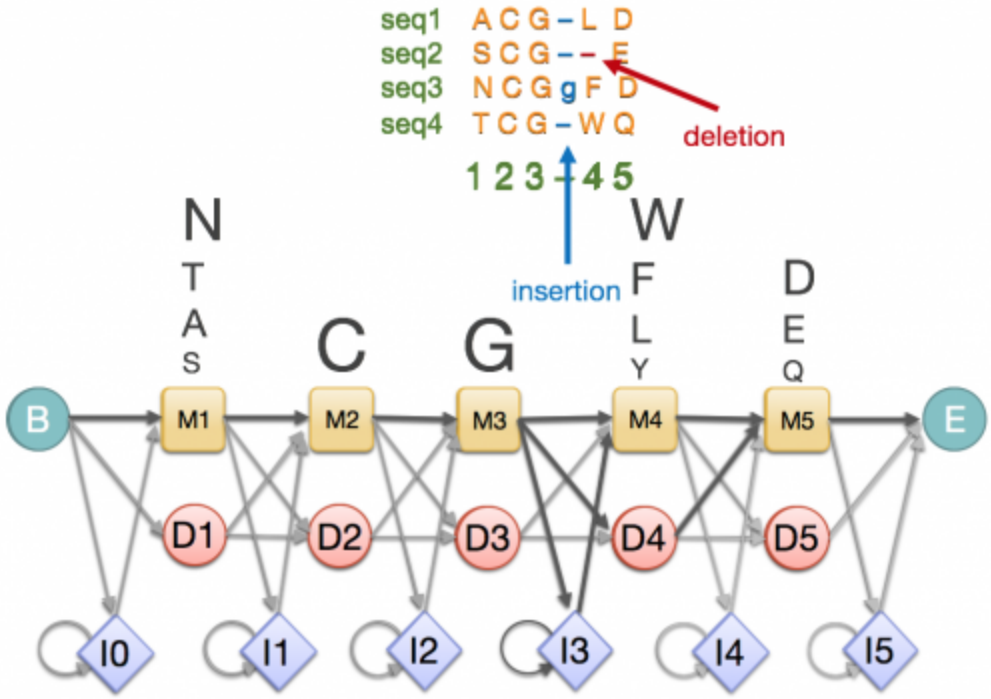
Modelo de Markov Ocultos (HMM)

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



ENSEMBL

Secuencias genómicas

Predicción automática de genes

Base de datos MySQL

Análisis y visualización

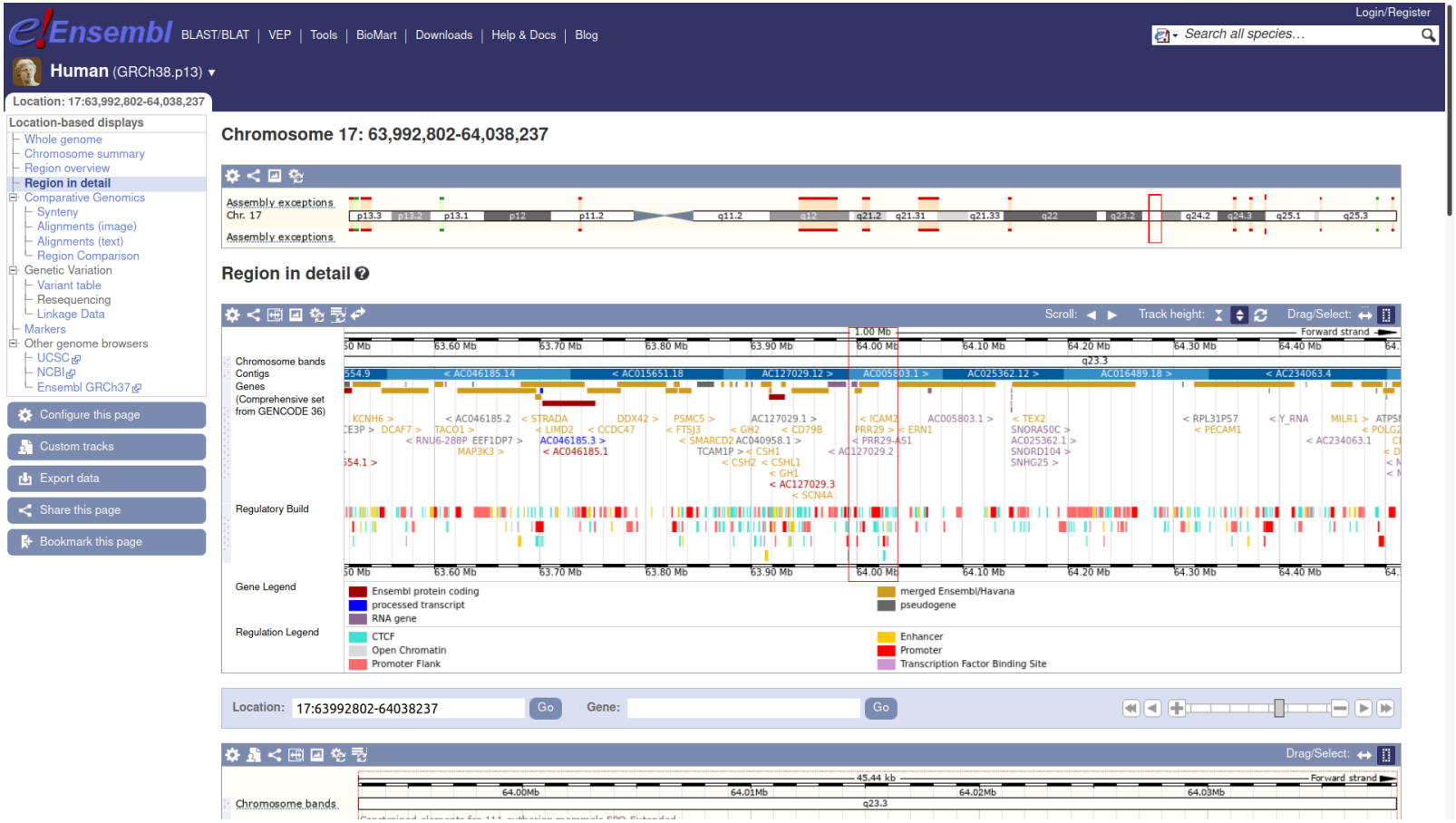
ENSEMBL

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

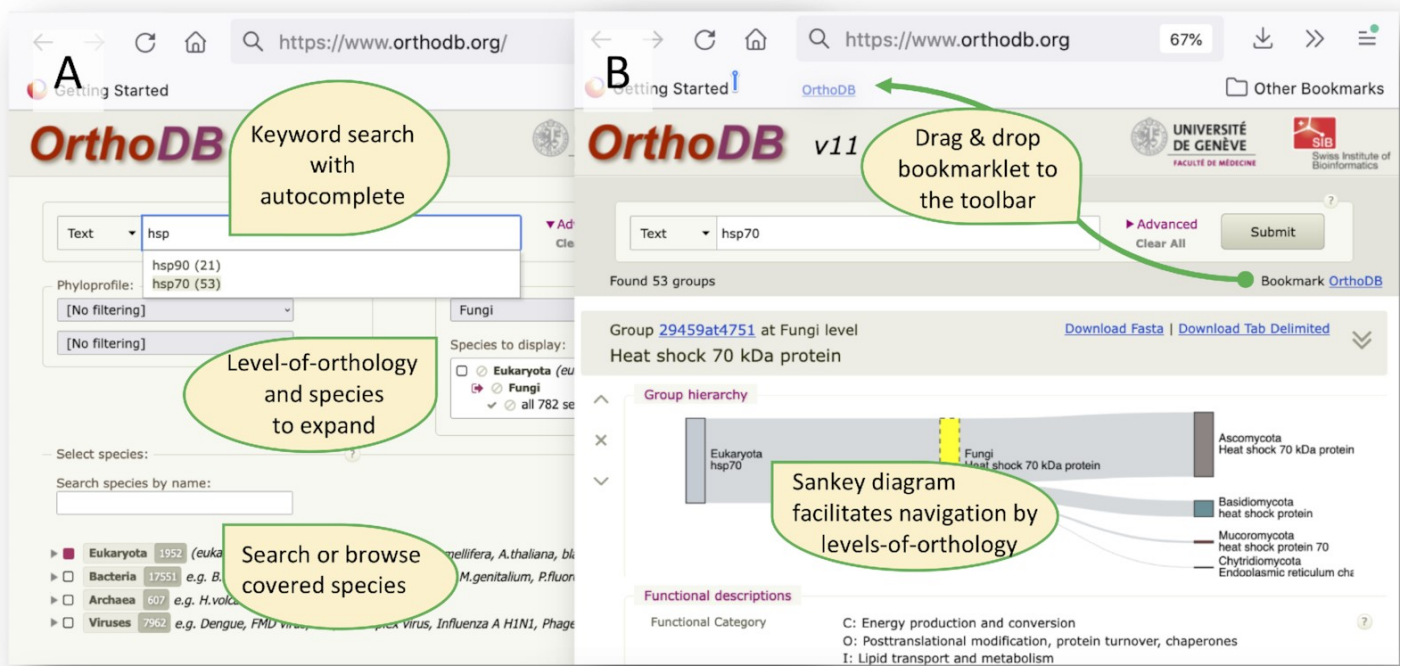
ENSEMBL. Acceso a los datos

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

ENSEMBL


Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Login/Register
Search all species…
Human (GRCh38.p13)
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
Ensembl protein coding
processed transcript
RNA gene
merged Ensembl/Havana
pseudogene
Regulation Legend
CTCF
Open Chromatin
Promoter Flank
Enhancer
Promoter
Transcription Factor Binding Site
Location: 17:63992802-64038237
Go
Gene:
Go
Drag/Select:
Chromosome bands

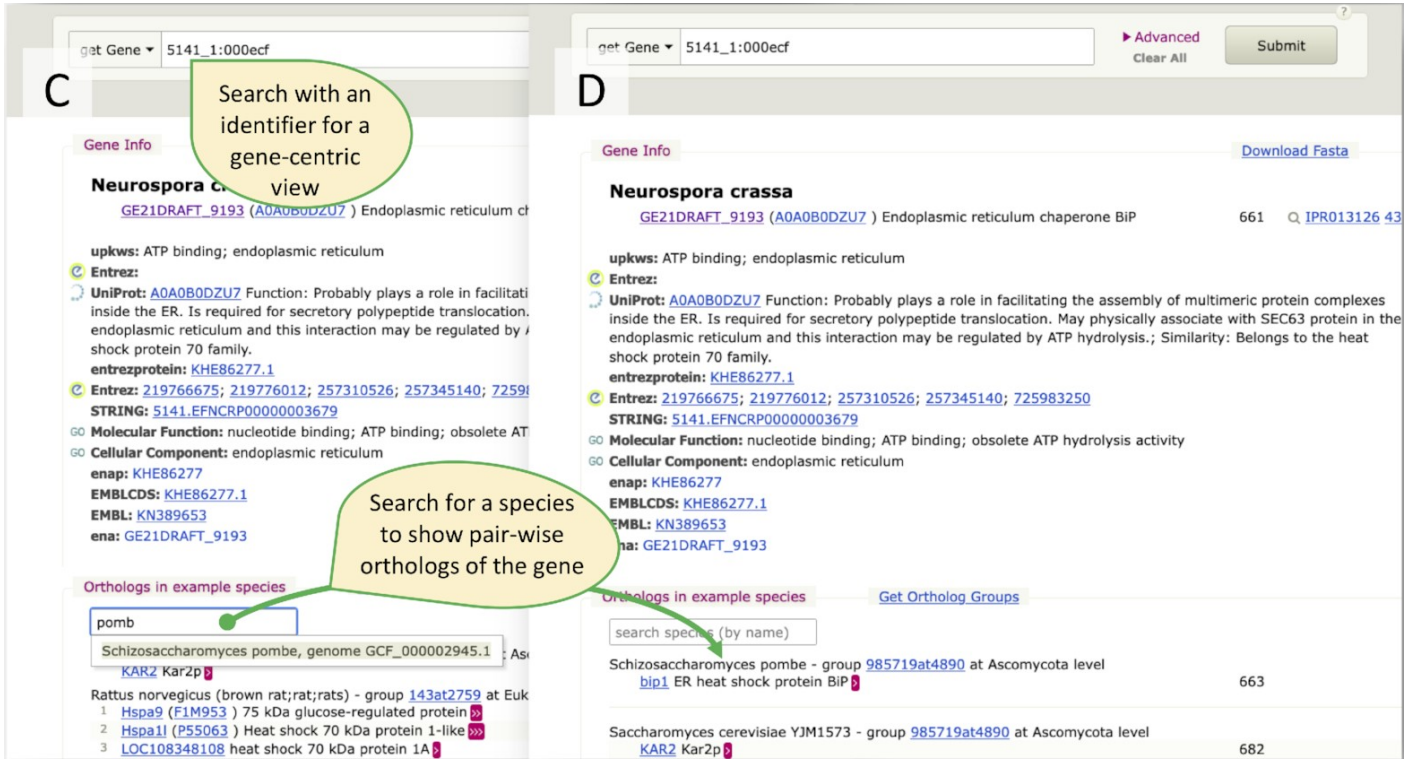
OrthoDB



Fuente: Kuznetsov, D. et al. 2023

A Se puede limitar la búsqueda por grupo taxonómico. **B** El grupo de ortología dispone de un diagrama interactivo.

OrthoDB



Fuente: Kuznetsov, D. et al. 2023

C La visualización del gen permite acceder a genes ortólogos. **D** Se puede buscar el ortólogo de una especie concreta.

OrthoDB

- Los datos primarios són las proteínas de UniProtKB.
- Utiliza el programa orthologer para agrupar proteínas ortólogas.
- El programa BUSCO utiliza a su vez OrthoDB para determinar qué genes suelen estar en copia única en la mayor parte de organismos de un grupo.