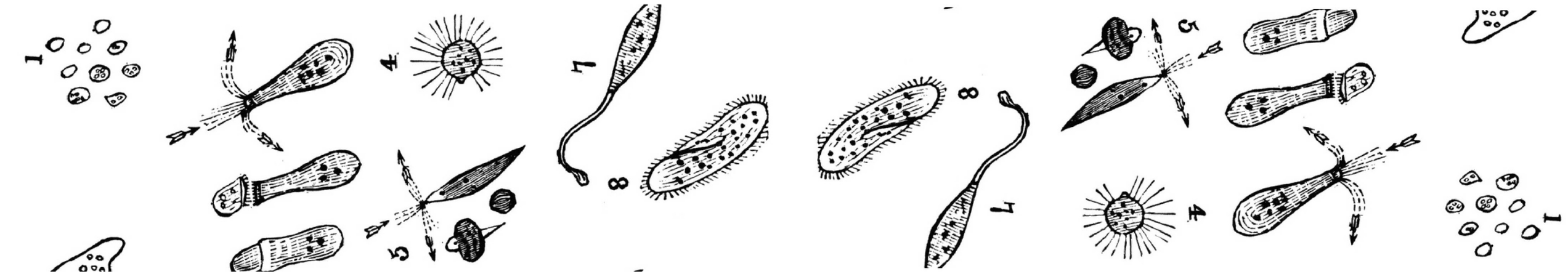
# Homología, Evolución, y Bases de Datos

Eduardo Castro, PhD
Universidad Andrés Bello
3 de agosto de 2018

www.castrolab.org
www.cbib.cl



https://github.com/bioinf-biotec/clases_bioinf

# El equipo

- Dr. Danilo González (fernando.gonzalez@unab.cl)
- Dr. Eduardo Castro (eduardo.castro@unab.cl)
- Jonathan Canan (jonathancanan@gmail.com)
- Katterinne Méndez (kat.mendez@uandresbello.edu)

# Evaluaciones

- La clase se divide en genómica y modelamiento de proteínas

- 3 pruebas (60%) & controles e informes (40%) + examen

- Programa en UNAB Virtual y en sitio web del curso

# El programa

| **Planificación de actividades** |
| --- |
| Homología y Evolución + Bases de Datos Biológicas y de Literatura + Búsqueda en Bases de Datos |
| Alineamiento de Pares de Secuencias, Múltiple y Perfiles (HMM’s) + Diseño de Partidores + BLAST |
| Ensamblaje de Genomas + Predicción de Genes |
| Modelos de Sustitución Nucleotídica y Proteica + Filogenética Molecular |
| **Solemne I** |
| Metagenómica, Metatranscriptómica y Microbioma humano |
| Visualización, Comparación y Clasificación de Estructura de Proteínas |
| Predicción de Estructura Secundaria y Terciaria de Proteínas |
| **Solemne II** |
| Búsqueda de proteínas homólogas + Redes de similitud + Modelado por homología |
| Bioinformática de enzimas |
| Introducción a la Simulación Molecular |
| Aplicación de la Simulación Molecular en ingeniería de proteínas |
| **Solemne III** |
| **Examen** |

# Bioinformática y Biología Computacional

- **Bioinformática** —> organización, almacenaje, clasificación de información biológica. Desarrollo de métodos, algoritmos, y software para analizar información biológica.

- **Biología Computacional** —> Aplicación de métodos analíticos y teóricos, modelamiento matemático y simulación computacional para el estudio de sistemas biológicos.

# So you want to be a computational biologist?

**Nick Loman & Mick Watson**

Loman, N., & Watson, M. (2013). So you want to be a computational biologist?. *Nature biotechnology*, *31*(11), 996-998.

# Para hoy

- Evolución y unidad de las formas de vida
- Homología, paralogía, xenología, etc.

- Organización del conocimiento científico y biológico

- Práctica: uso de bases de datos

# Evolución y unidad de las formas de vida

# Observación

- Alta biodiversidad
- Sin embargo todos exhiben elementos que los unifican —> organismos comparten ancestros comunes

## How to Read the Circle of Life

Primordial life begins at the center and branches out in all directions, leading to the groups of species that exist today (*colored rings*)

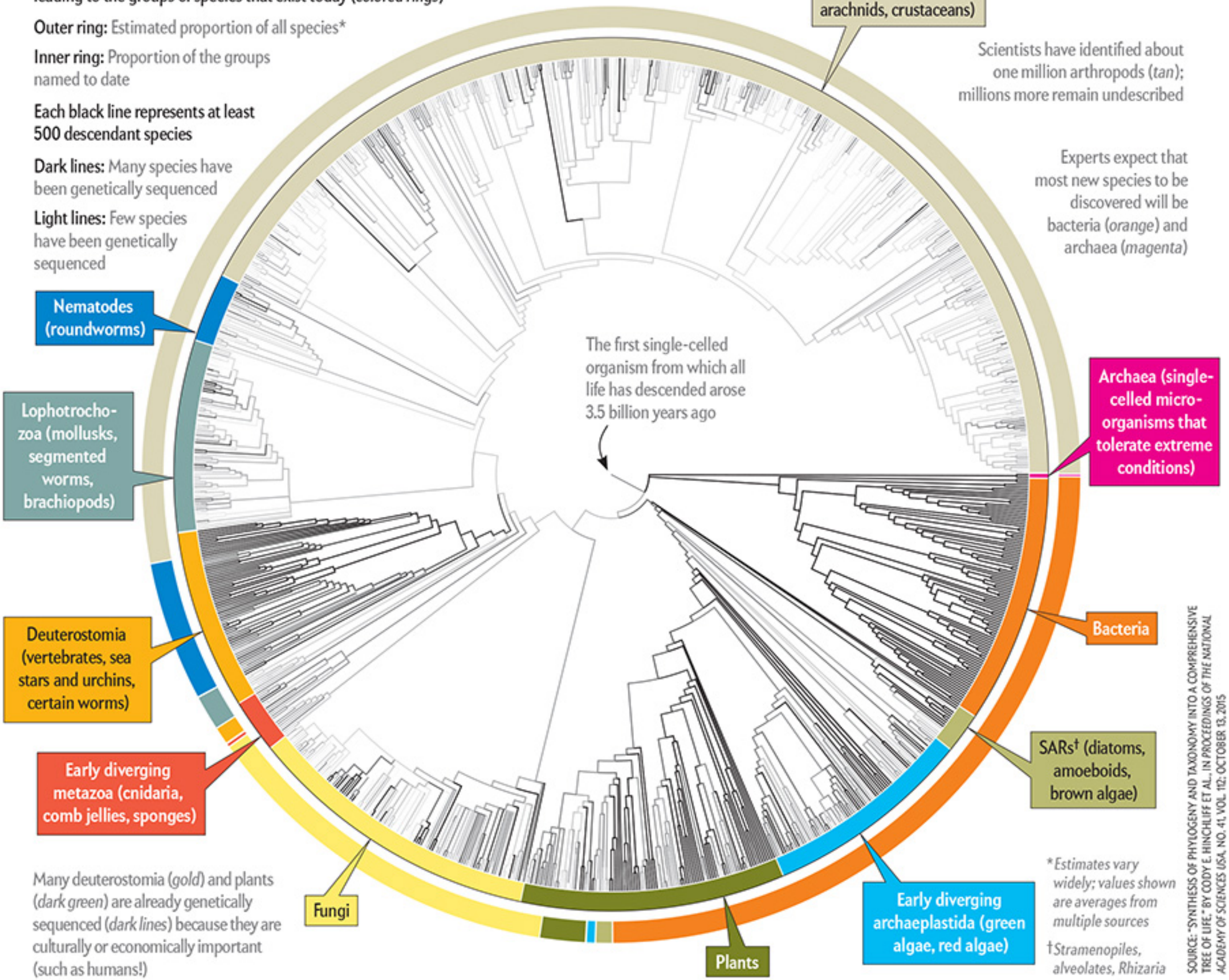
**Outer ring:** Estimated proportion of all species*

**Inner ring:** Proportion of the groups named to date

**Each black line represents at least 500 descendant species**

**Dark lines:** Many species have been genetically sequenced

**Light lines:** Few species have been genetically sequenced




Many deuterostomia (*gold*) and plants (*dark green*) are already genetically sequenced (*dark lines*) because they are culturally or economically important (such as humans!)

**Estimates vary widely; values shown are averages from multiple sources*

*†Stramenopiles, alveolates, Rhizaria*

SOURCE: "SYNTHESIS OF PHYLOGENY AND TAXONOMY INTO A COMPREHENSIVE TREE OF LIFE," BY CODY E. HINCHLIFF ET AL., IN *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES USA*, NO. 41, VOL. 112: OCTOBER 13, 2015

# Evidencia sobre unidad de la vida

- Código genético
- Registro fósil
- Función y estructura celular —> linea germinal
- Rasgos vestigiales
- Distribución de especies relacionadas

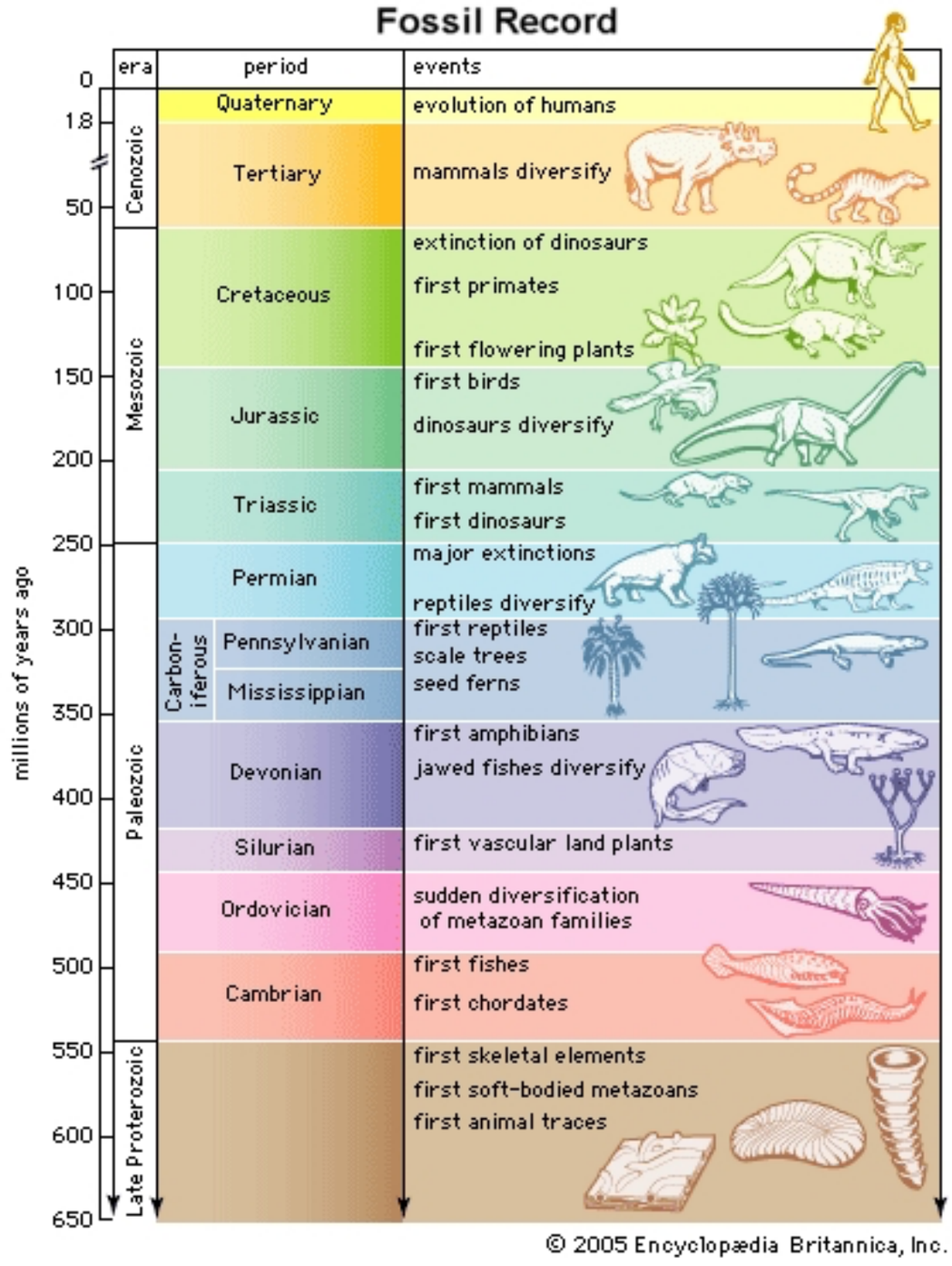
| | | | |
|---|---|---|---|
| UUU [F] Phe | UCU [S] Ser | UAU [Y] Tyr | UGU [C] Cys |
| UUC [F] Phe | UCC [S] Ser | UAC [Y] Tyr | UGC [C] Cys |
| UUA [L] Leu | UCA [S] Ser | UAA [ ] Ter | UGA [ ] Ter |
| UUG [L] Leu | UCG [S] Ser | UAG [ ] Ter | UGG [W] Trp |
| CUU [L] Leu | CCU [P] Pro | CAU [H] His | CGU [R] Arg |
| CUC [L] Leu | CCC [P] Pro | CAC [H] His | CGC [R] Arg |
| CUA [L] Leu | CCA [P] Pro | CAA [Q] Gln | CGA [R] Arg |
| CUG [L] Leu | CCG [P] Pro | CAG [Q] Gln | CGG [R] Arg |
| AUU [I] Ile | ACU [T] Thr | AAU [N] Asn | AGU [S] Ser |
| AUC [I] Ile | ACC [T] Thr | AAC [N] Asn | AGC [S] Ser |
| AUA [I] Ile | ACA [T] Thr | AAA [K] Lys | AGA [R] Arg |
| AUG [M] Met | ACG [T] Thr | AAG [K] Lys | AGG [R] Arg |
| GUU [V] Val | GCU [A] Ala | GAU [D] Asp | GGU [G] Gly |
| GUC [V] Val | GCC [A] Ala | GAC [D] Asp | GGC [G] Gly |
| GUA [V] Val | GCA [A] Ala | GAA [E] Glu | GGA [G] Gly |
| GUG [V] Val | GCG [A] Ala | GAG [E] Glu | GGG [G] Gly |

# Código genético

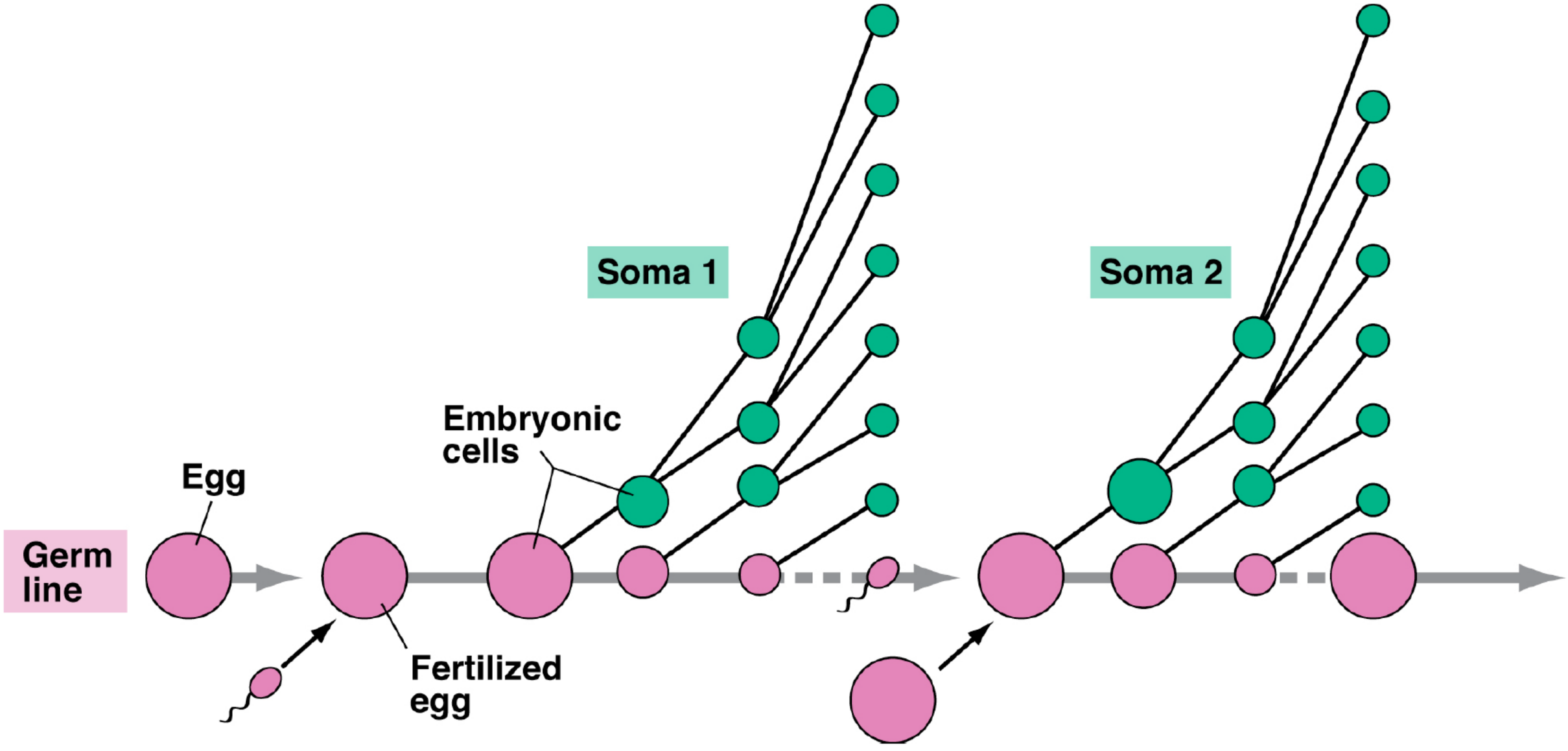
- Sistema común para todas las especies
  - Codones —> aa
  - DNA para almacenar información genética
  - Síntesis de proteínas a través de ribosomas
  - Los mismos 20 aminoácidos

# Registro fósil

- Registra la historia de la vida a través del tiempo
- 4 mil millones de años
- Documentan relaciones entre ancestros y descendientes

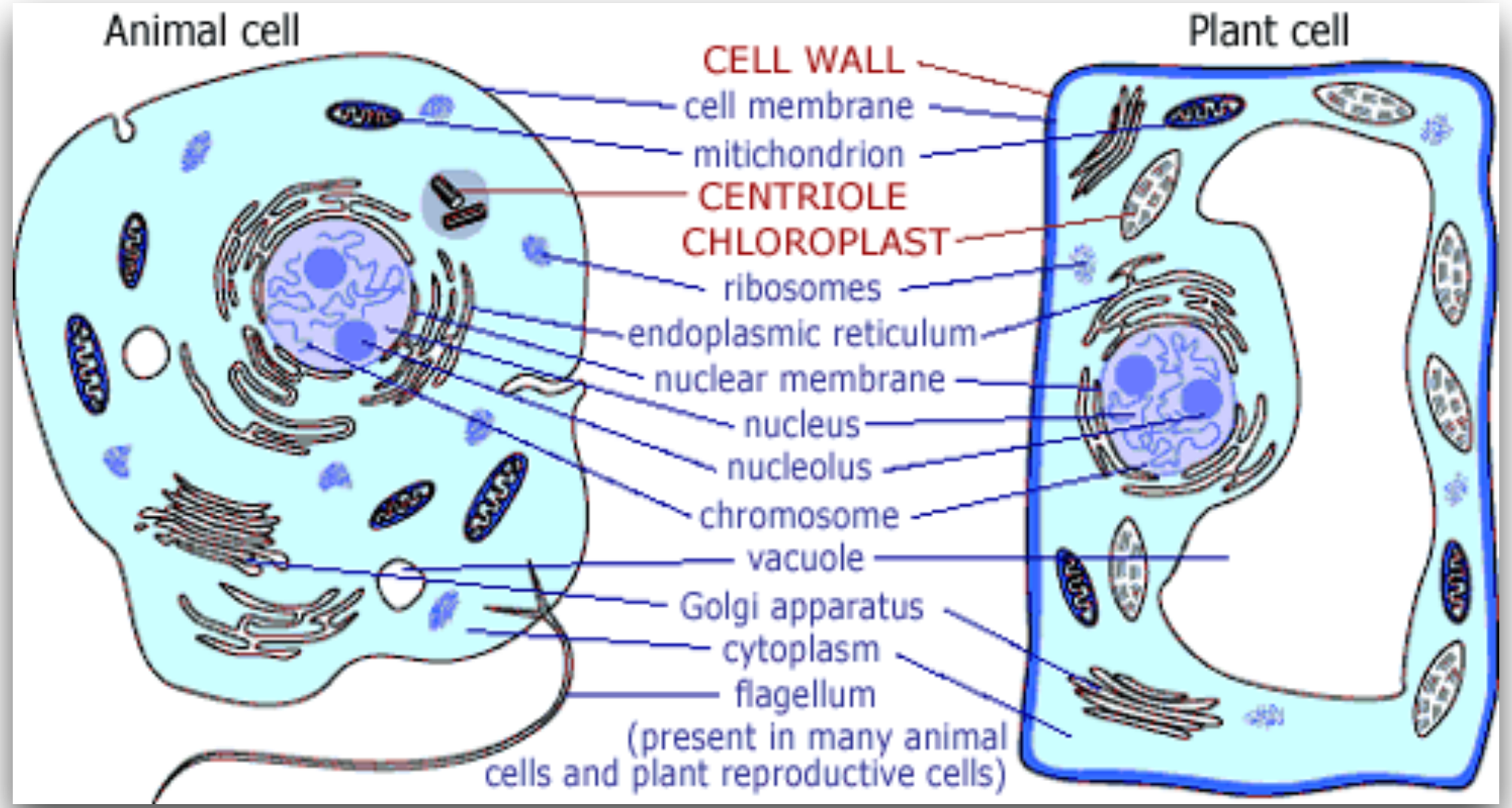


# Linea germinal




La linea germinal forma un continuo a través de individuos y generaciones

# Estructura y función



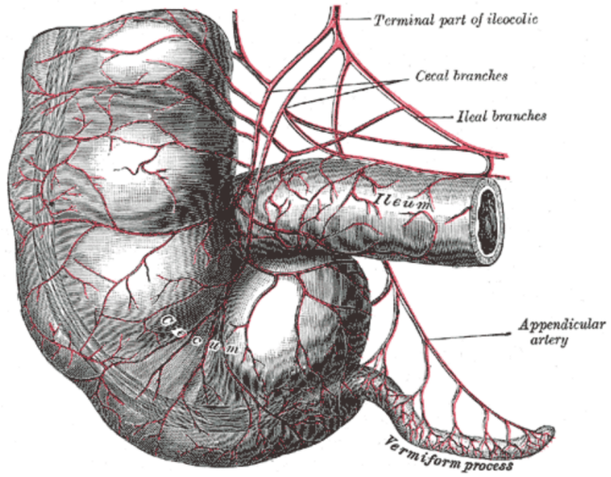

Células de distintos organismos poseen estructuras funcionales similares

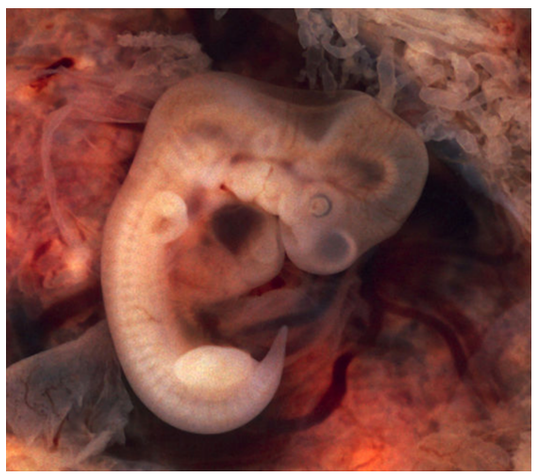
# Rasgos vestigiales

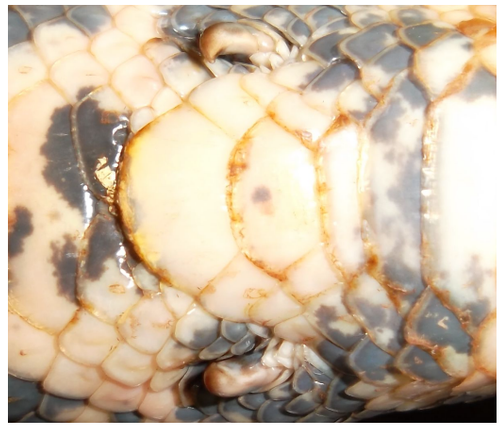
- Estructuras determinadas genéticamente que han sido perdido su función ancestral en una especie pero han sido retenidas durante el proceso evolutivo

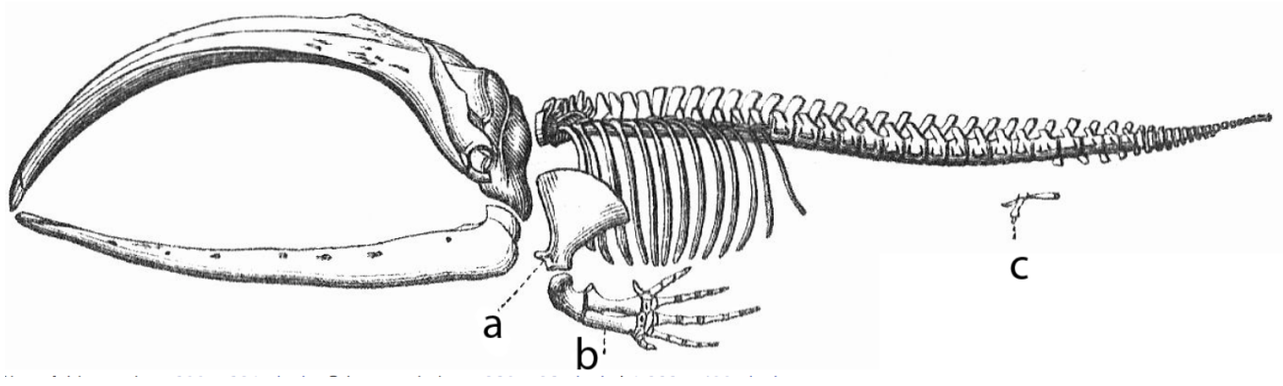
- Apéndice, cola, reflejo ante frio

- Patas traseras en boa constrictor, ballena



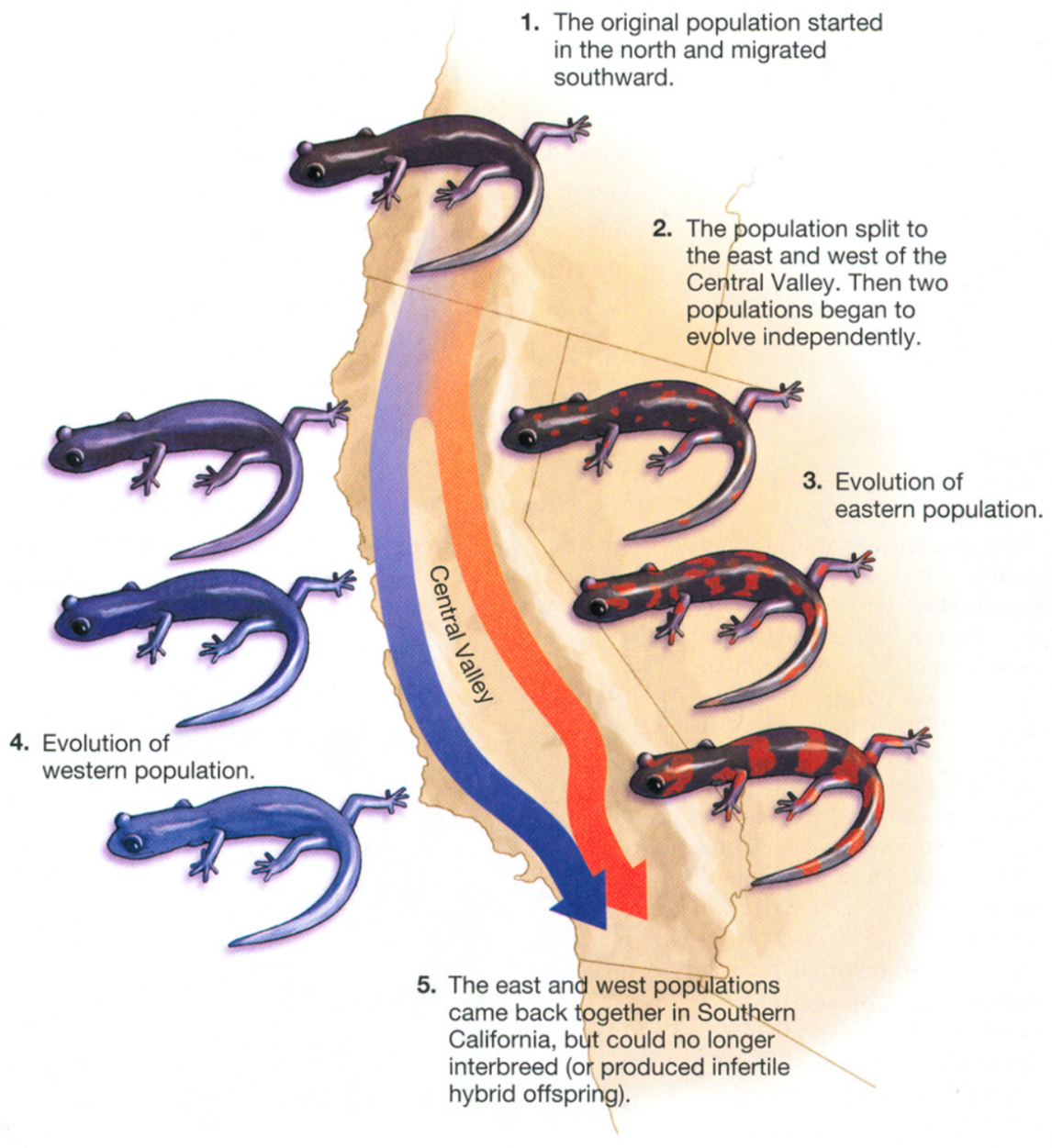




# Distribución de especies relacionadas

- Geografía aisla poblaciones y genera especies con origen común

- Especies relacionadas tienden a estar correlacionadas geograficamente

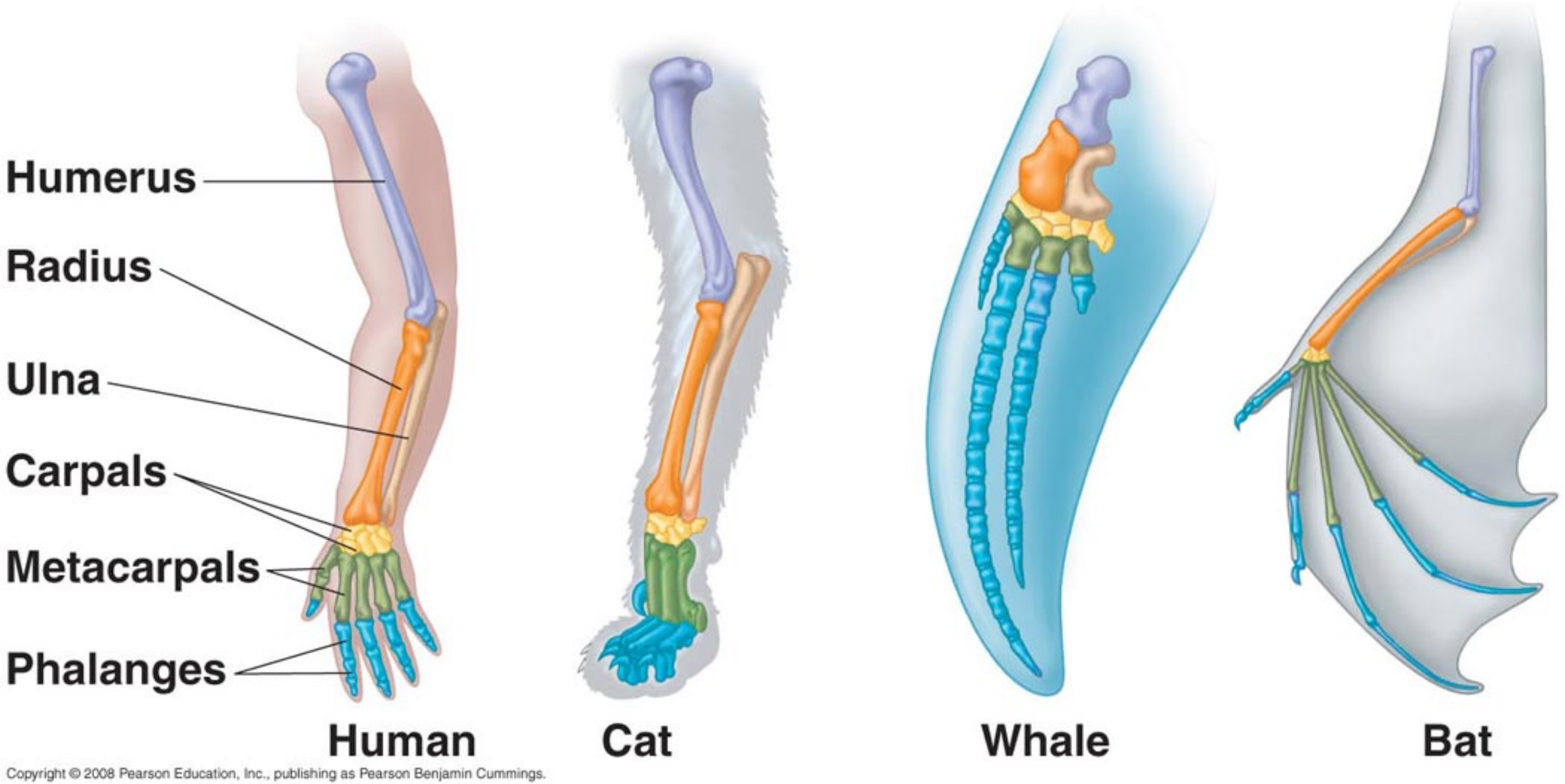


# Homología

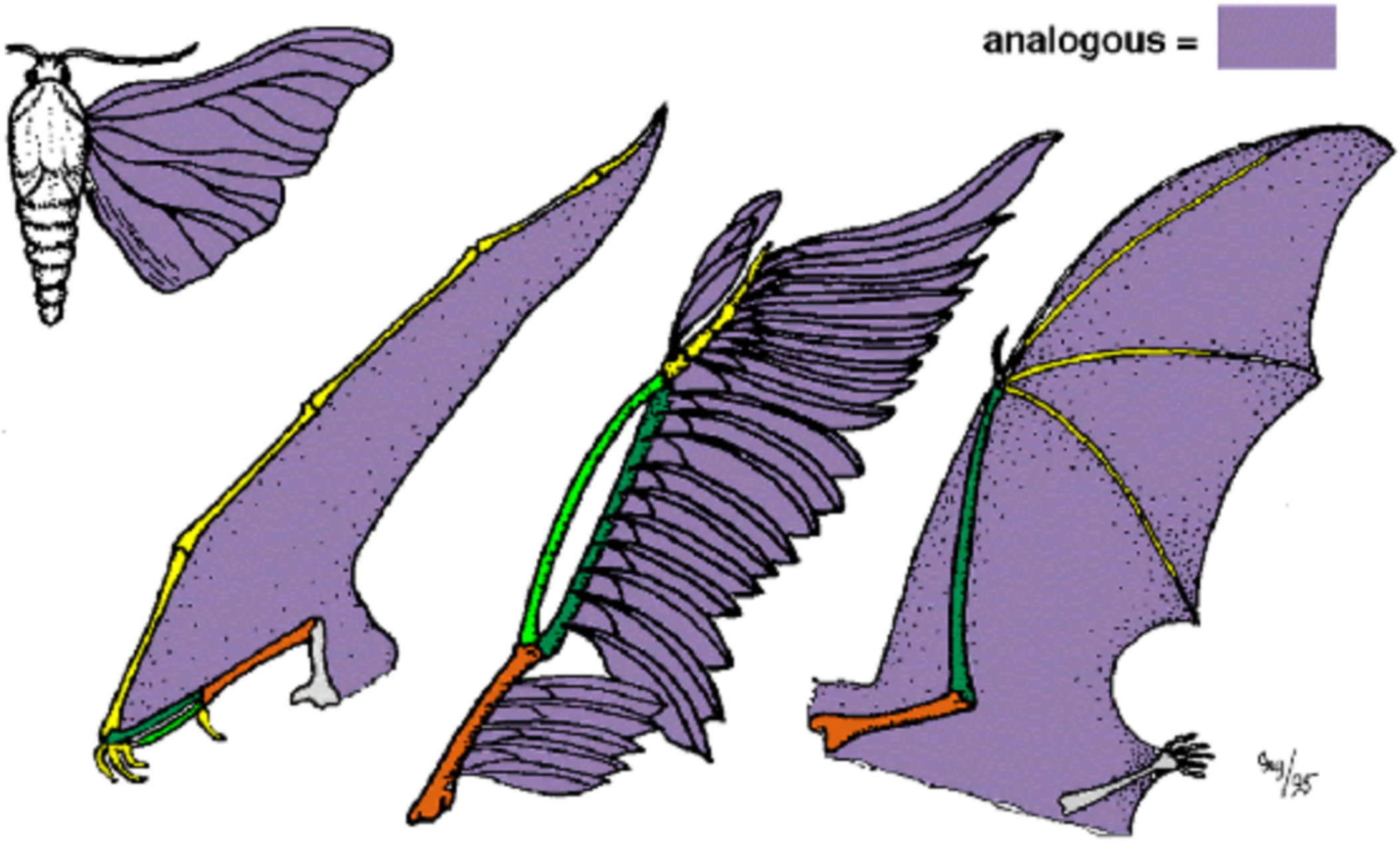
# ¿Qué es homología?

- **homología morfológica** —> especies que pertenecen al mismo grupo taxonómico muestran similitudes anatómicas debido a que comparten un ancestro común (no por chance)

# Estructuras homólogas



# Estructuras análogas



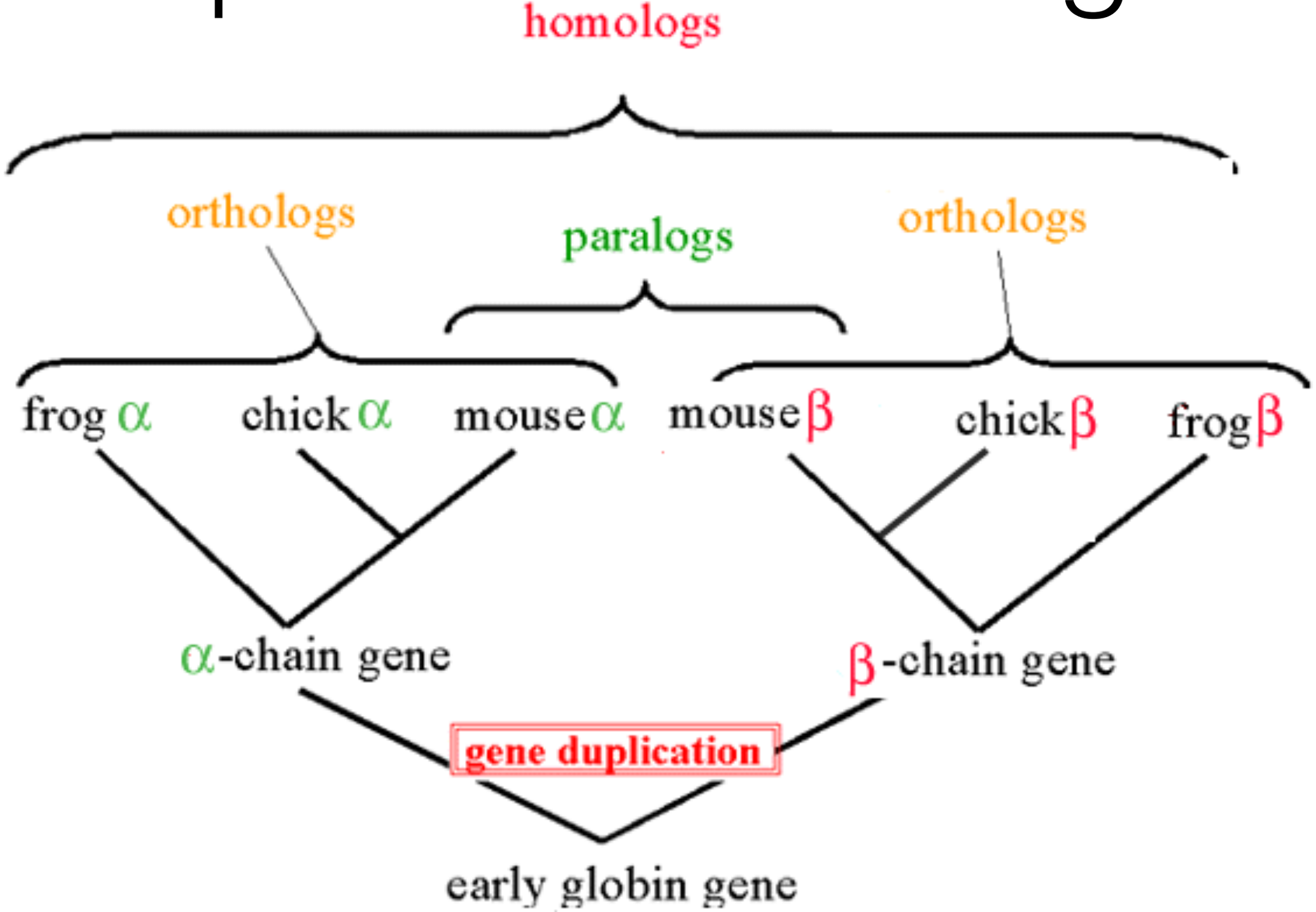
# ¿Qué es homología?

- **homología molecular**—> especies que pertenecen al mismo grupo taxonómico muestran similitudes en DNA, RNA y proteínas debido a que comparten un ancestro común (no por chance)

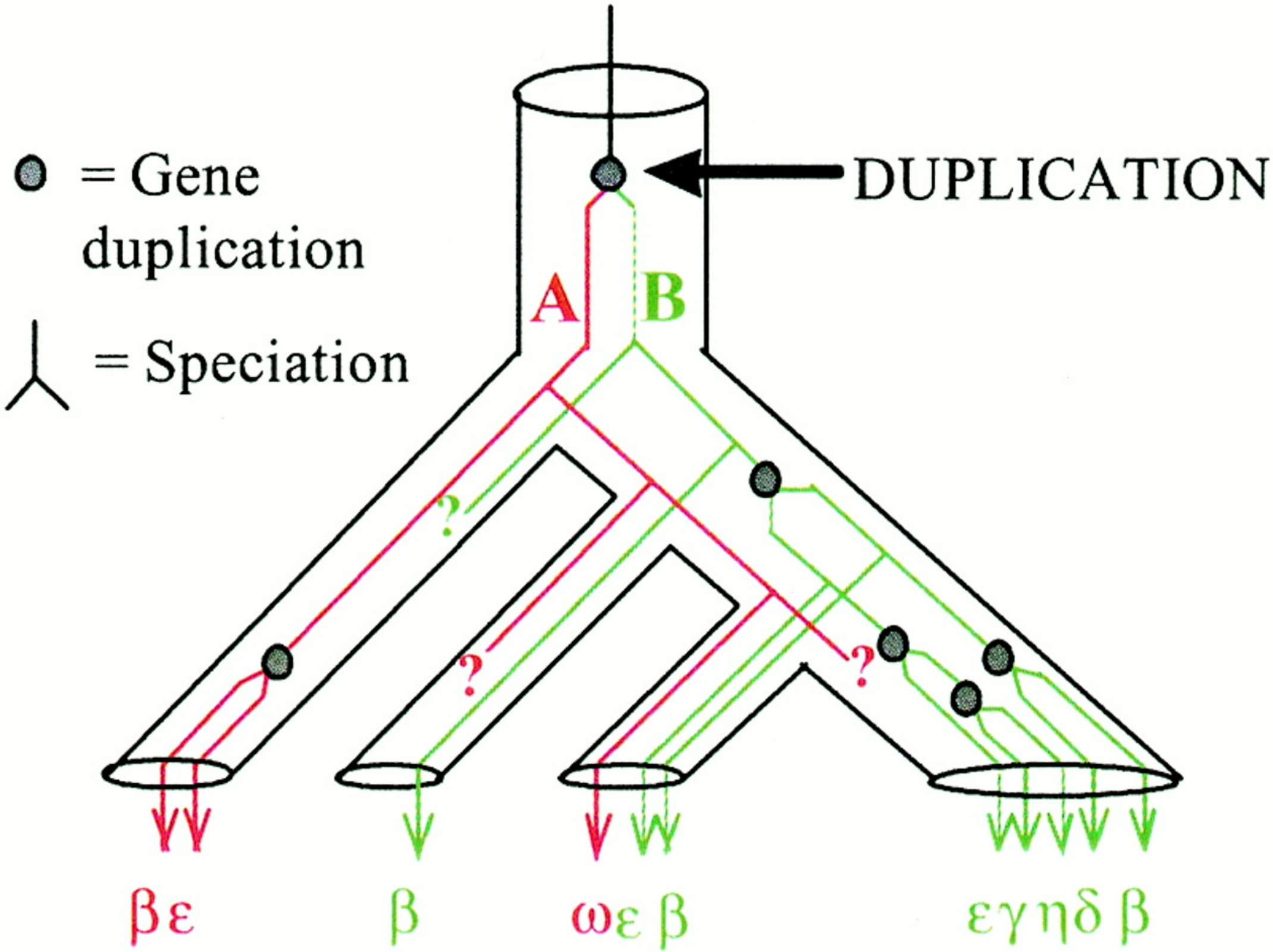
# Tipos de homología

- Ortólogos
- Paralógos
- Xenólogos

# Tipos de homología



# Ancestral β-globin gene

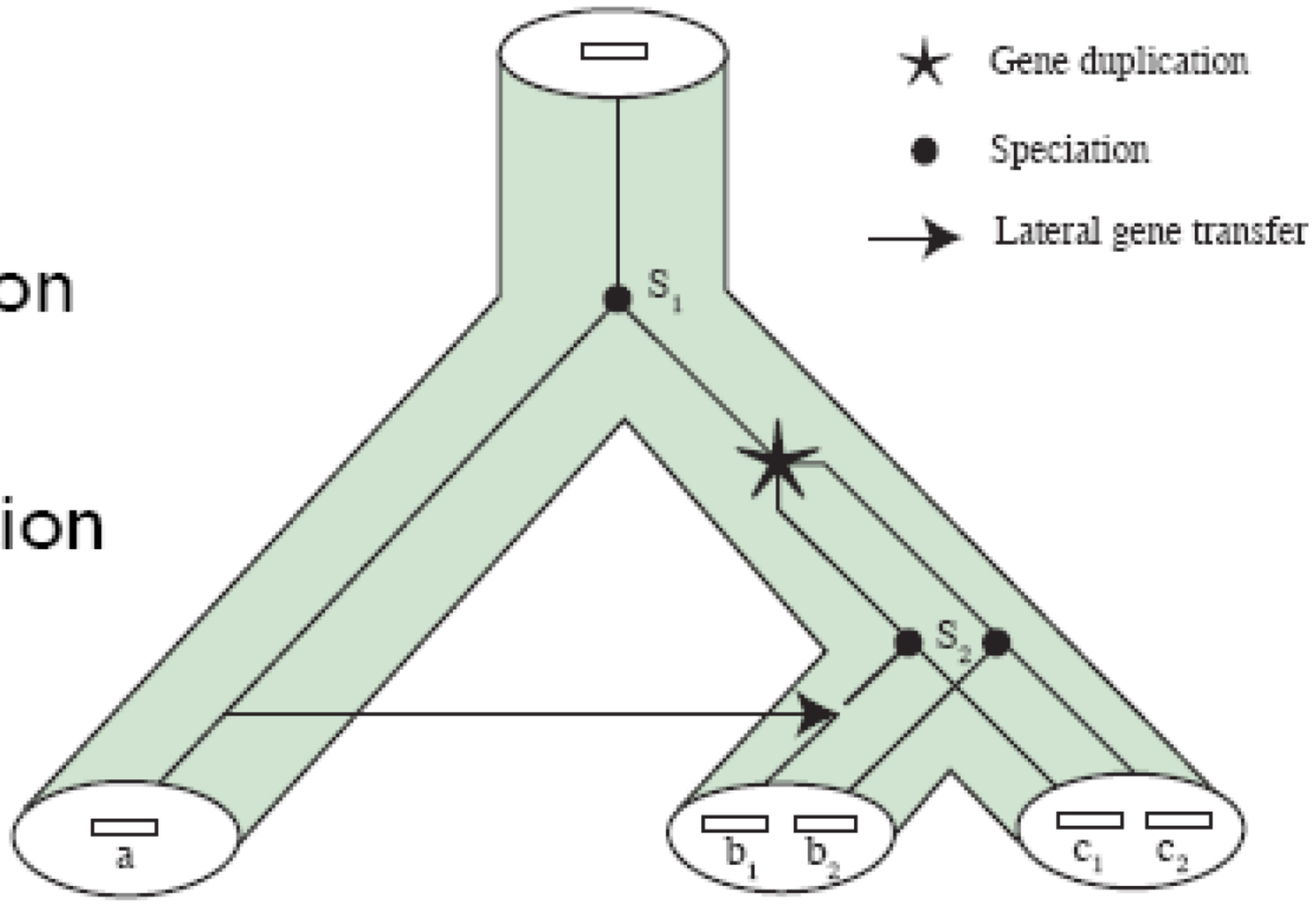


# Tipos de homología

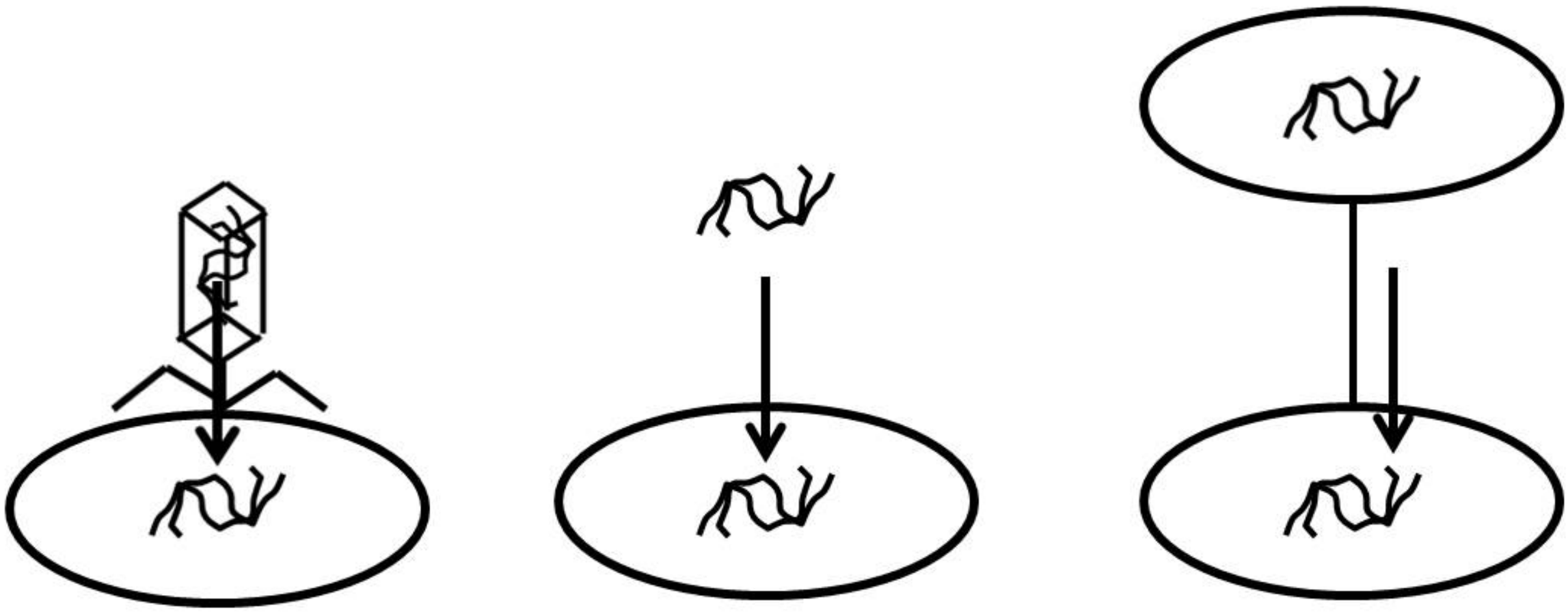
Two genes (or characters) are *homologs* if they have a common ancestor.

Main Subtypes

- *Orthologs*: through speciation
- *Paralogs*: through duplication
- *Xenologs*: through lateral transfer



# Xenólogos en bacteria



**Transduction** **Transformation** **Conjugation**

# Bases de datos

# ¿Por qué necesitamos bases de datos?

- Biología es una ciencia cuantitativa
- Experimentos de alto rendimiento como secuenciamiento masivo producen vastas cantidades de datos
- Necesitamos clasificar, almacenar, buscar información biológica para transformar información en conocimiento

# ¿Qué pueden hacer las bases de datos?

- Hacer datos disponibles para científicos alrededor del mundo (human-readable)

- Hacer datos disponibles para que puedan ser analizados por algoritmos (machine-readable)

# Tipos de bases de datos

- Primarias: Resultados experimentales directos.
  Secundarias: Resultados de análisis

- Nucleótidos y proteínas

- Dominios y motivos

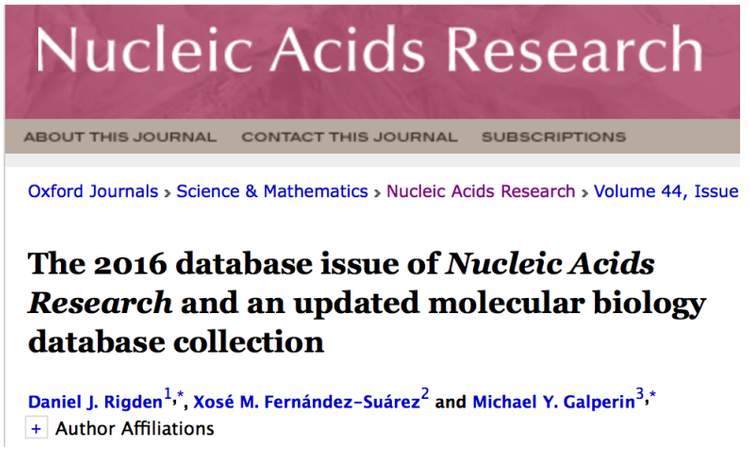
- Estructuras 3D

- Expresión génica

- Rutas metabólicas

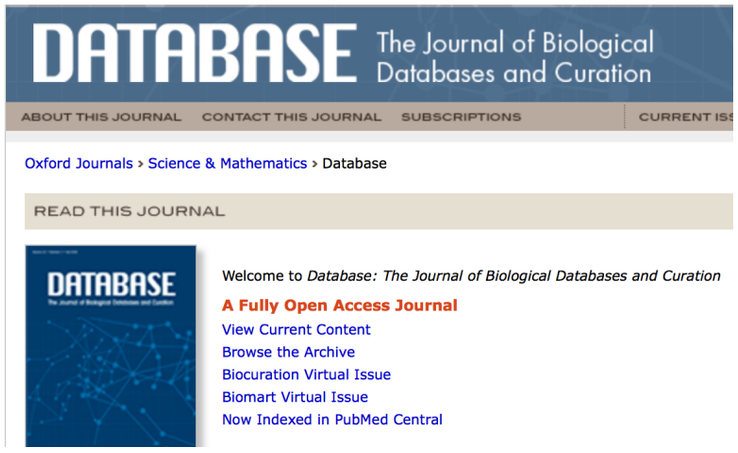
# ¿Cómo encuentro la base de datos que necesito?

- NAR database issue
- Database Journal



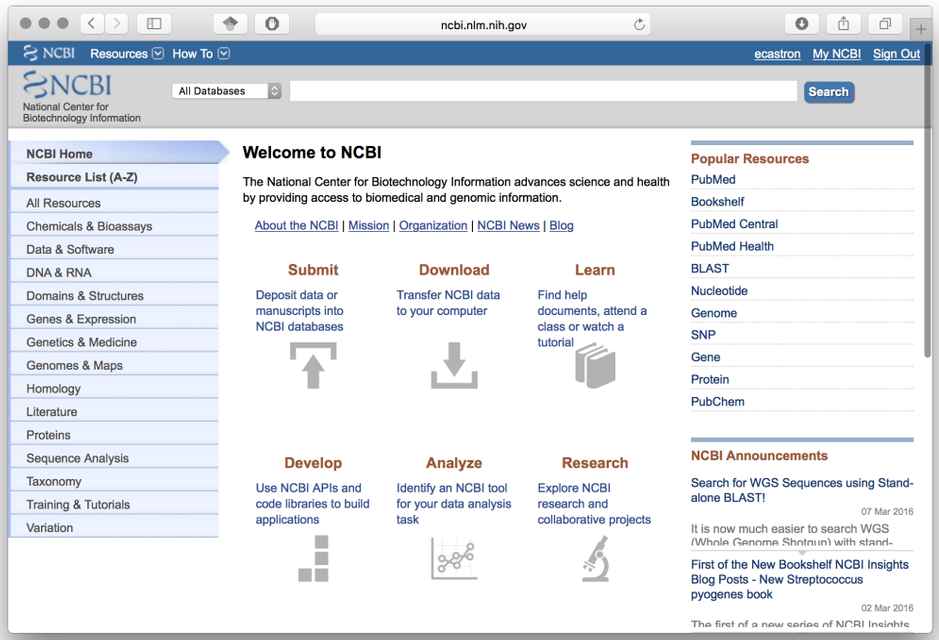

**NAR Database Summary Paper Category List**

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Cell biology

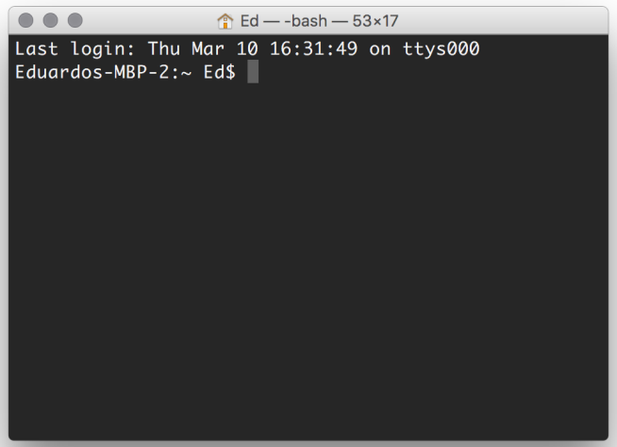


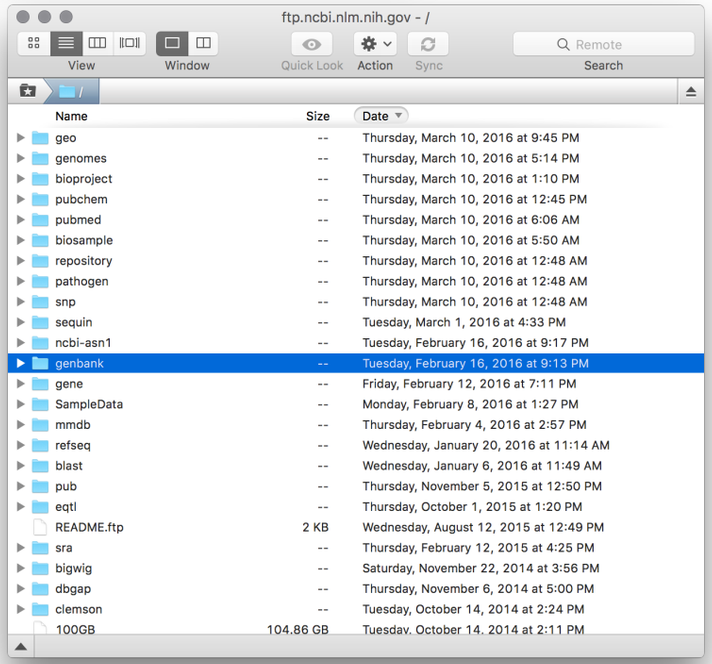
# ¿Cómo accedo a los datos contenidos en bases de datos?

- Interface gráfica —>




- “Puerta de atrás” FTP, scripting —> ftp.ncbi.nlm.nih.gov

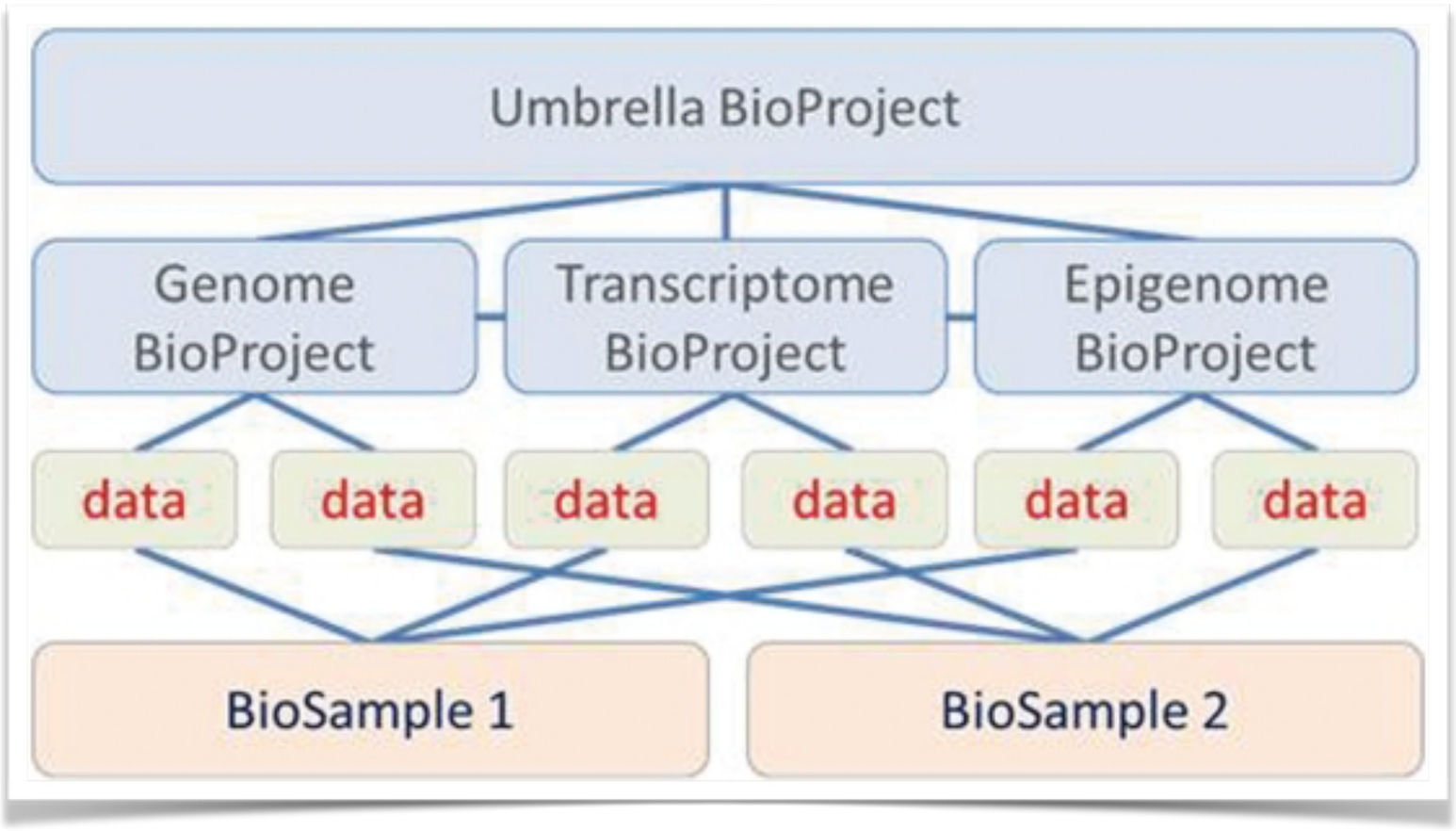


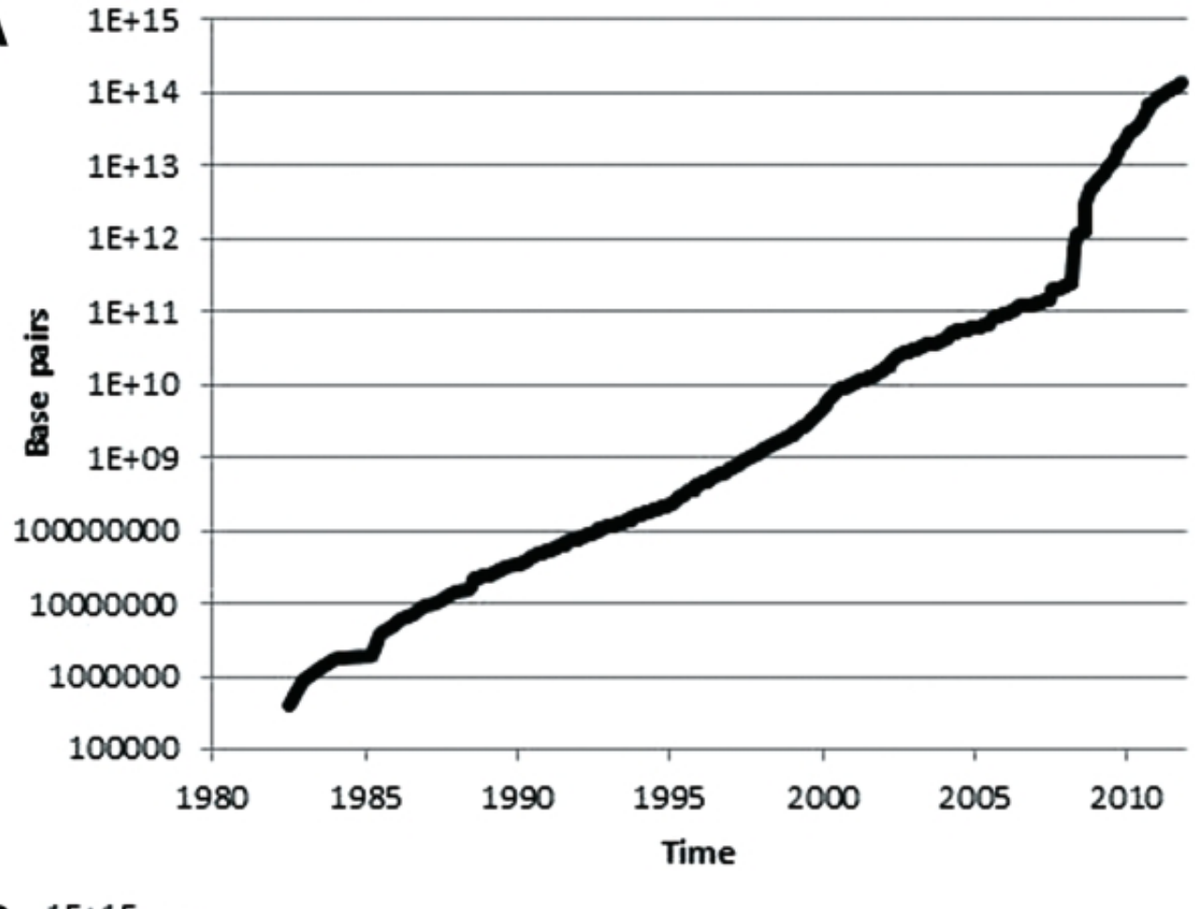
# Ejemplo más popular

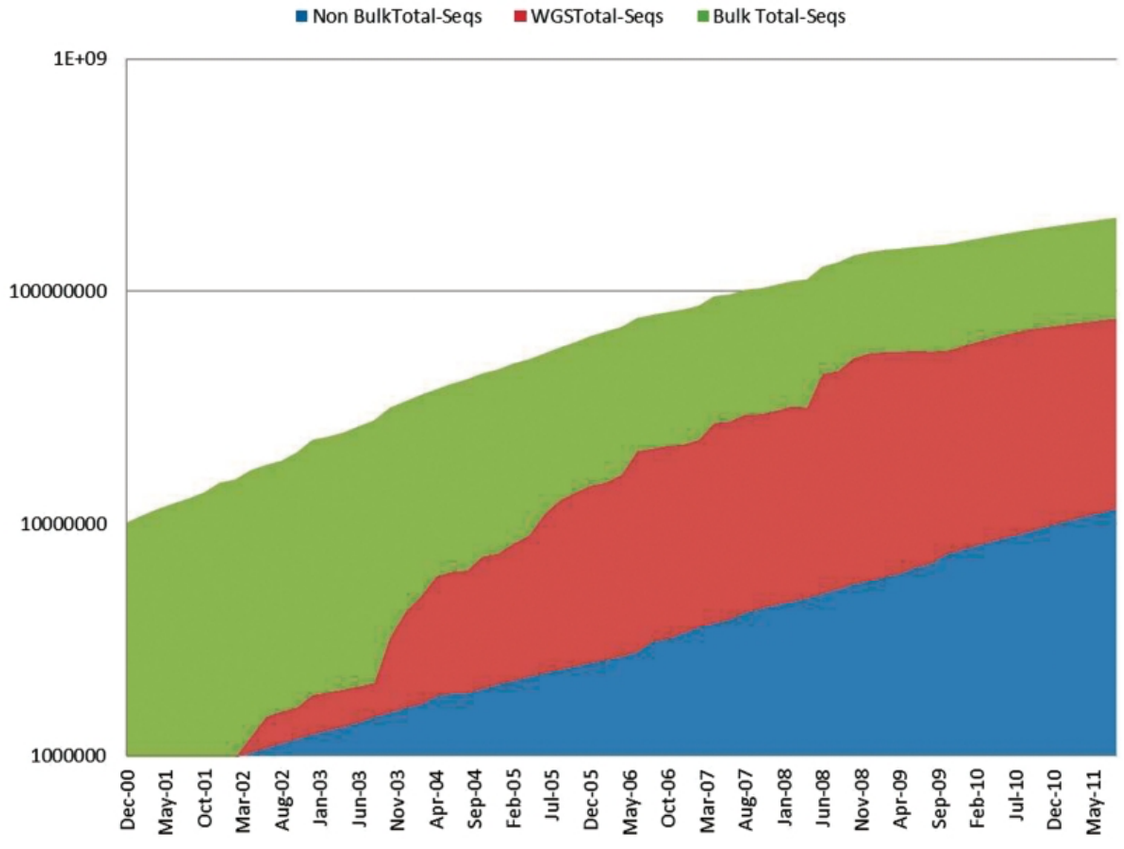
- International Nucleotide Sequence Database Collaboration (INSDC)

- 30 años de antigüedad

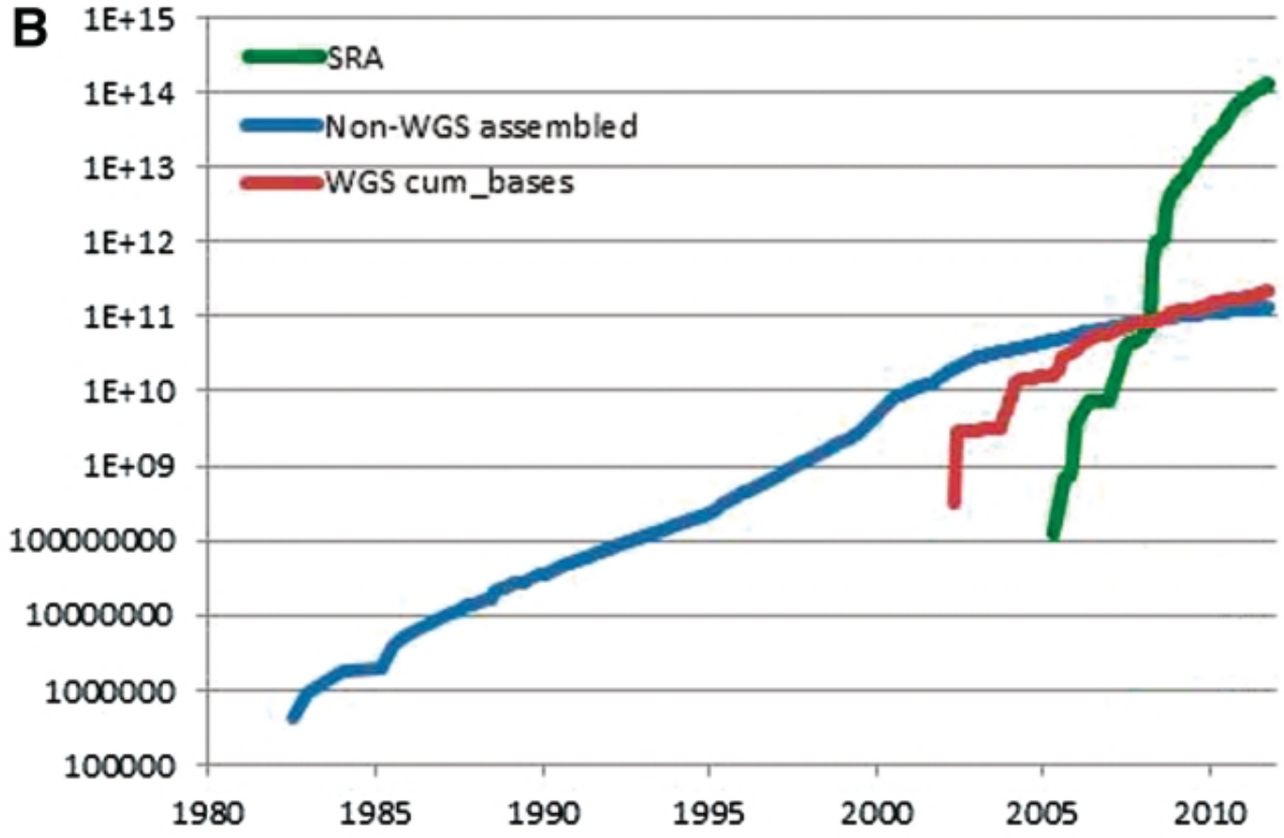


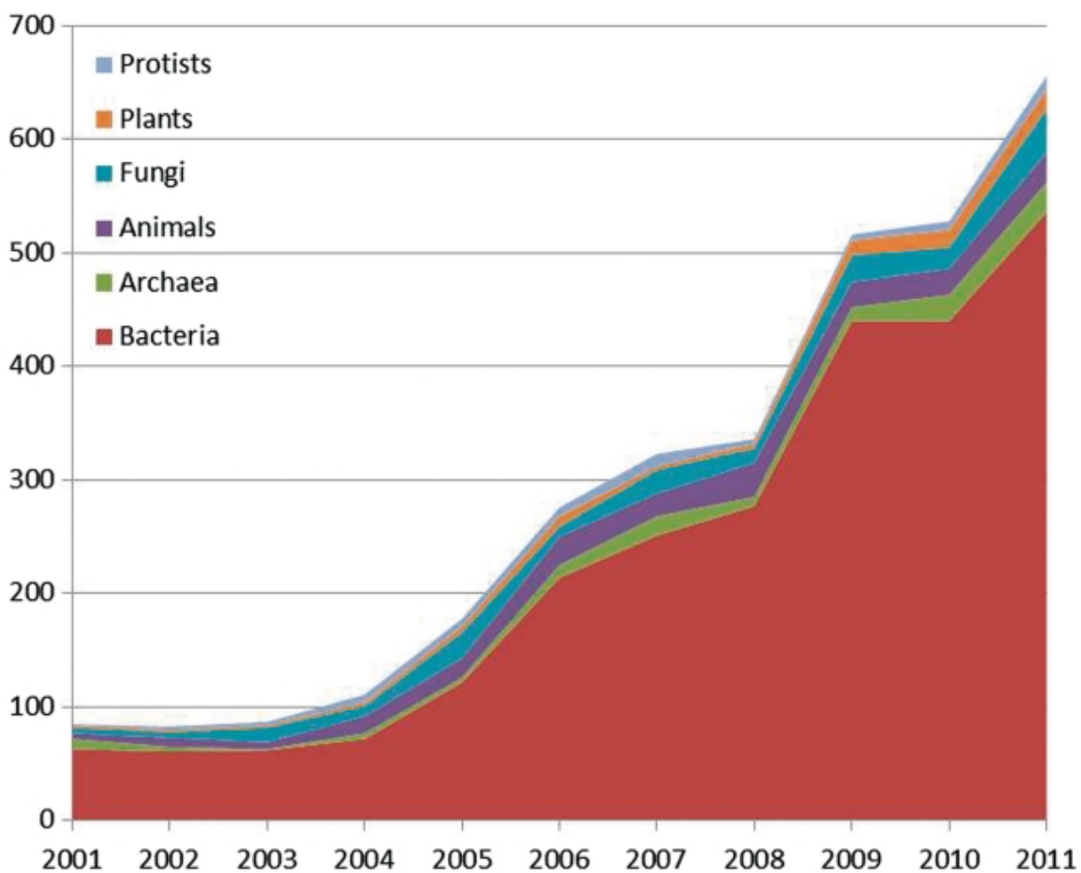
A


Base pairs
1E+15
1E+14
1E+13
1E+12
1E+11
1E+10
1E+09
100000000
10000000
1000000
100000
1980
1985
1990
1995
2000
2005
2010
Time


Non BulkTotal-Seqs
WGSTotal-Seqs
Bulk Total-Seqs
1E+09
100000000
10000000
1000000
Dec-00
May-01
Oct-01
Mar-02
Aug-02
Jan-03
Jun-03
Nov-03
Apr-04
Sep-04
Feb-05
Jul-05
Dec-05
May-06
Oct-06
Mar-07
Aug-07
Jan-08
Jun-08
Nov-08
Apr-09
Sep-09
Feb-10
Jul-10
Dec-10
May-11

B


1E+15
1E+14
1E+13
1E+12
1E+11
1E+10
1E+09
100000000
10000000
1000000
100000
SRA
Non-WGS assembled
WGS cum_bases
1980
1985
1990
1995
2000
2005
2010

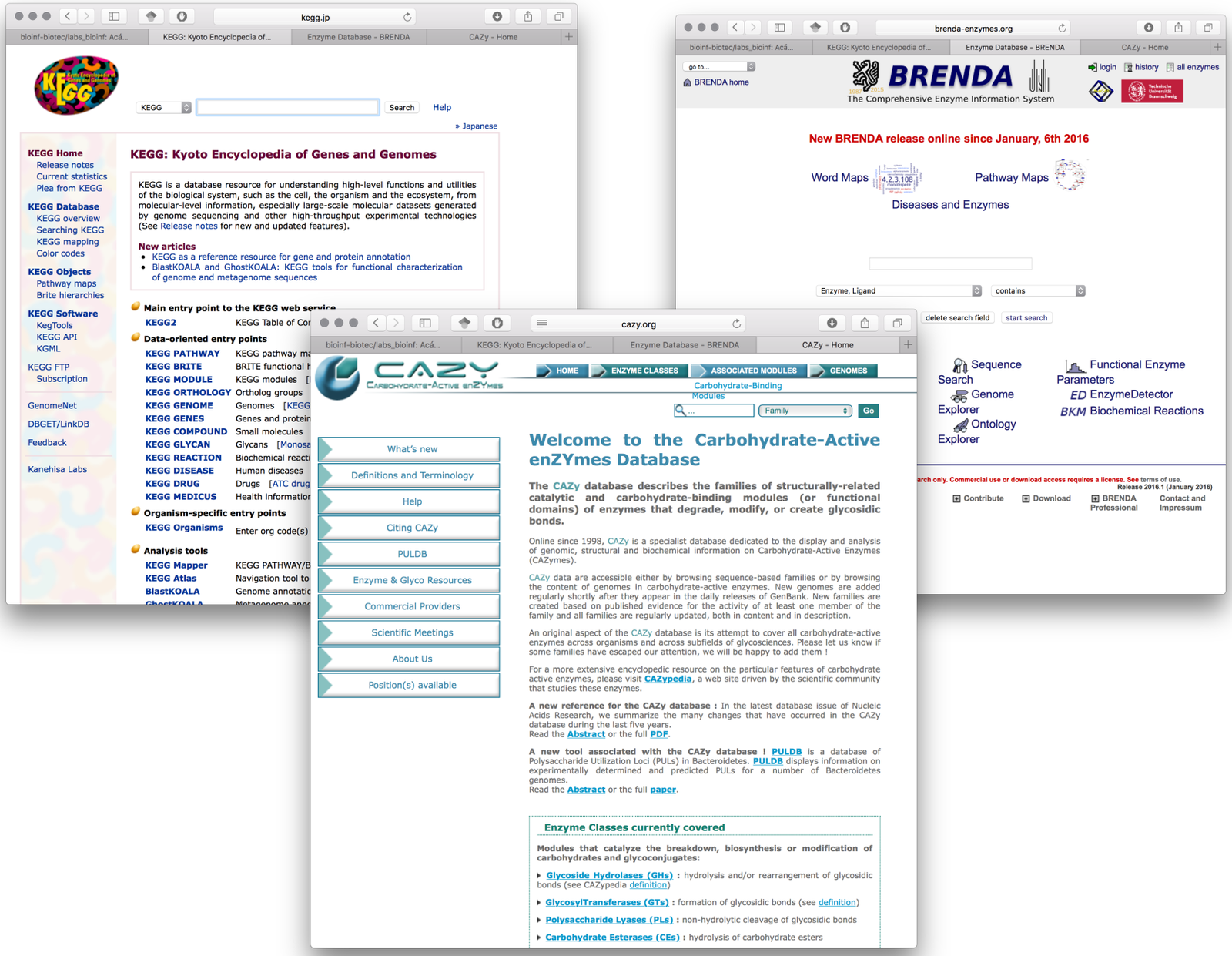

700
600
500
400
300
200
100
0
Protists
Plants
Fungi
Animals
Archaea
Bacteria
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011

# Clases especiales de bases de datos

- No curadas
  - NCBI nt, TrEMBL
- Curadas (con diferentes niveles de evidencia)
  - Swiss-Prot, PIR
- Especializadas
  - PeptidesAtlas —> espectros de masas de péptidos

# ... de enzimas

- KEGG
- BRENDA
- CAZymes



# … de proteínas

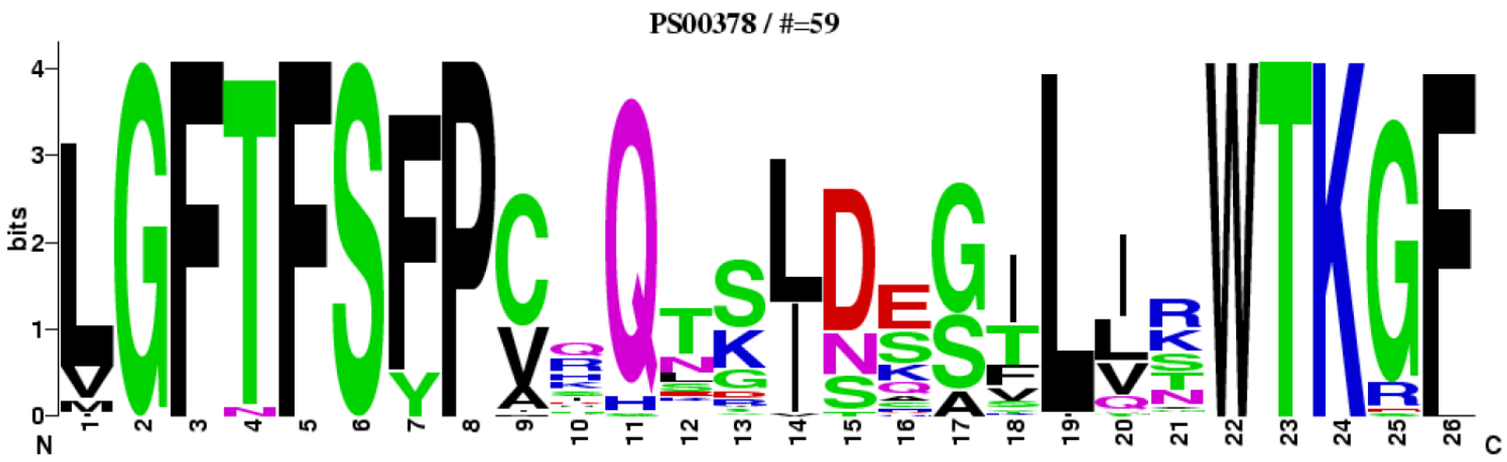
- Protein Data Bank

# Bases de datos secundarias

- PROSITE
- Pfams
- Rfams
- PRINTS

PROSITE [HEXOKINASES PS00378]

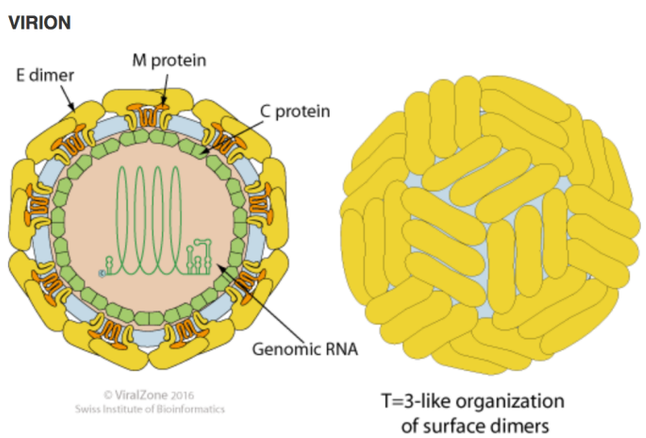
- Database of protein domains, families and functional sites
- Hexokinases signature: Pattern [LIVM]-G-F-[TN]-F-S-[FY]-P-x(5)-[LIVM]-[DNST]-x(3)-[LIVM]-x(2)-W-T-K-x- [LF].



# y taxonomia

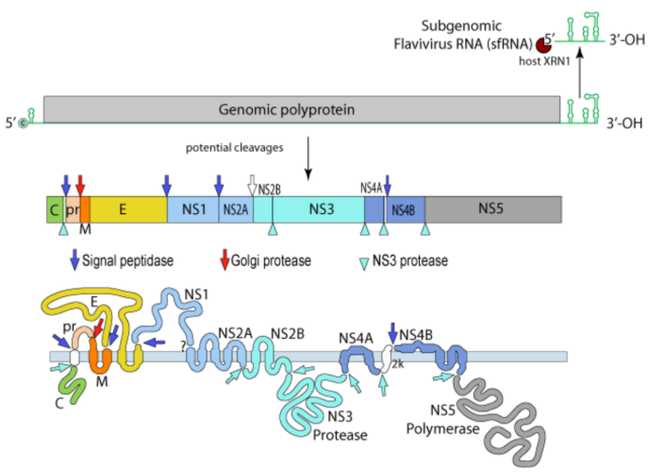
- UniProt Taxonomy
- NCBI Taxonomy
- ViralZone

**Zika virus (strain Mr 766)**




Enveloped, spherical, about 50 nm in diameter. The surface proteins are arranged in an icosahedral-like symmetry.

**GENOME**




Monopartite, linear, ssRNA(+) genome of of 10,794 bp. The genome 5′ end has a methylated nucleotide cap for canonical cellular translation. The 3′ terminus is not polyadenylated but forms a loop structure. This secondary structure leads to the formation of a subgenomic flavivirus RNA (sfRNA) through genomic RNA degradation by host XRN1. sfRNA is essential for pathogenicity, and may play a role in inhibiting host RIG-I antiviral activity as shown for Dengue virus

**GENE EXPRESSION**

The virion RNA is infectious and serves as both the genome and the viral messenger RNA. The whole genome is translated in a polyprotein 3,419 aa long, which is processed co- and post-translationally by host and viral proteases.

**REPLICATION**

**CYTOPLASMIC** in mammals, **NUCLEAR** in insects?

1. Attachement of the viral envelope protein E to host receptors mediates internalization into the host cell by apoptotic mimicry
2. Fusion of virus membrane with host endosomal membrane. RNA genome is released into the cytoplasm.
3. The positive-sense genomic ssRNA is translated into a polyprotein, which is cleaved into all structural and non structural proteins (to yield the replication proteins).
4. Replication takes place at the surface of endoplasmic reticulum in cytoplasmic viral factories. A dsRNA genome is synthesized from the genomic ssRNA(+).
5. The dsRNA genome is transcribed/replicated thereby providing viral mRNAs/new ssRNA(+) genomes.
6. Virus assembly occurs at the endoplasmic reticulum. The virion buds via the host ESCRT complexes at the endoplasmic reticulum, is transported to the Golgi apparatus.
7. The prM protein is cleaved in the Golgi, thereby maturing the virion which is fusion competent.
8. Release of new virions by exocytosis.

# Ahora, el lab…

https://github.com/bioinf-biotec