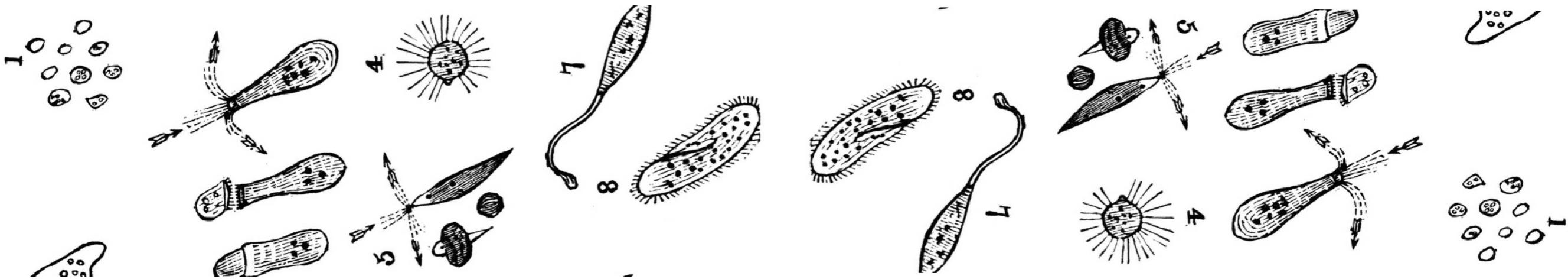




Homología, Evolución, y Bases de Datos

www.castrolab.org
www.cbib.cl

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



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

Nematodes (roundworms)

Lophotrochozoa (mollusks, segmented worms, brachiopods)

Deuterostomia (vertebrates, sea stars and urchins, certain worms)

Early diverging metazoa (cnidaria, comb jellies, sponges)

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

Fungi

Arthropods (insects, arachnids, crustaceans)

Scientists have identified about one million arthropods (tan); millions more remain undescribed

Experts expect that most new species to be discovered will be bacteria (orange) and archaea (magenta)

Archaea (single-celled micro-organisms that tolerate extreme conditions)

Bacteria

SARs[†] (diatoms, amoeboids, brown algae)

Early diverging archaeplastida (green algae, red algae)

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

[†]Stramenopiles, alveolates, Rhizaria

Evidencia sobre unidad de la vida

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

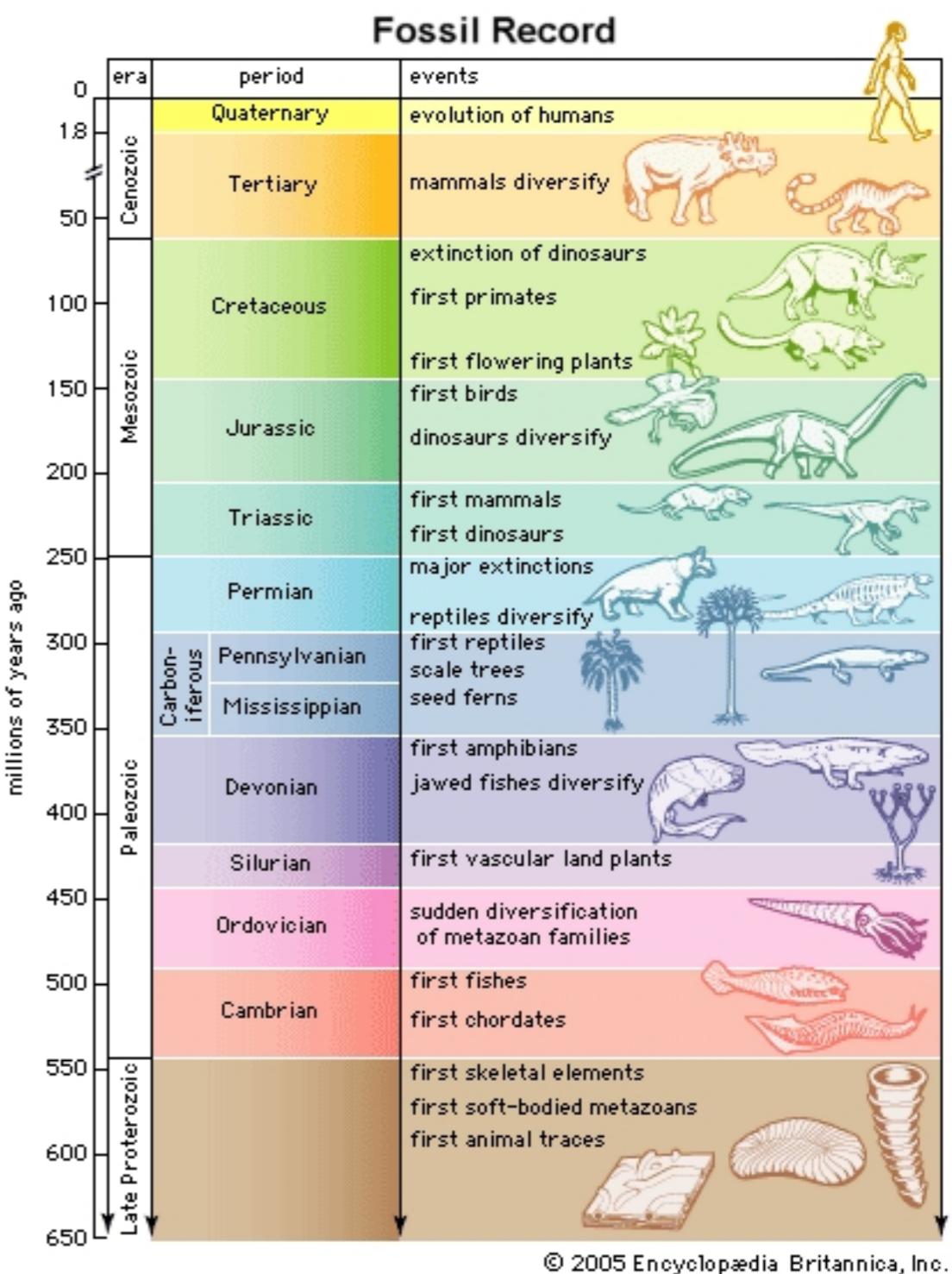
UUU [F] Phe	UCU [S] Ser	UAU [Y] Tyr	UGU [C] Cys
UUC [F] Phe	UCC [S] Ser	UAC [Y] Tyr	UGC [C] Cys
UUA [L] Leu	UCA [S] Ser	UAA [] Ter	UGA [] Ter
UUG [L] Leu	UCG [S] Ser	UAG [] Ter	UGG [W] Trp
CUU [L] Leu	CCU [P] Pro	CAU [H] His	CGU [R] Arg
CUC [L] Leu	CCC [P] Pro	CAC [H] His	CGC [R] Arg
CUA [L] Leu	CCA [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

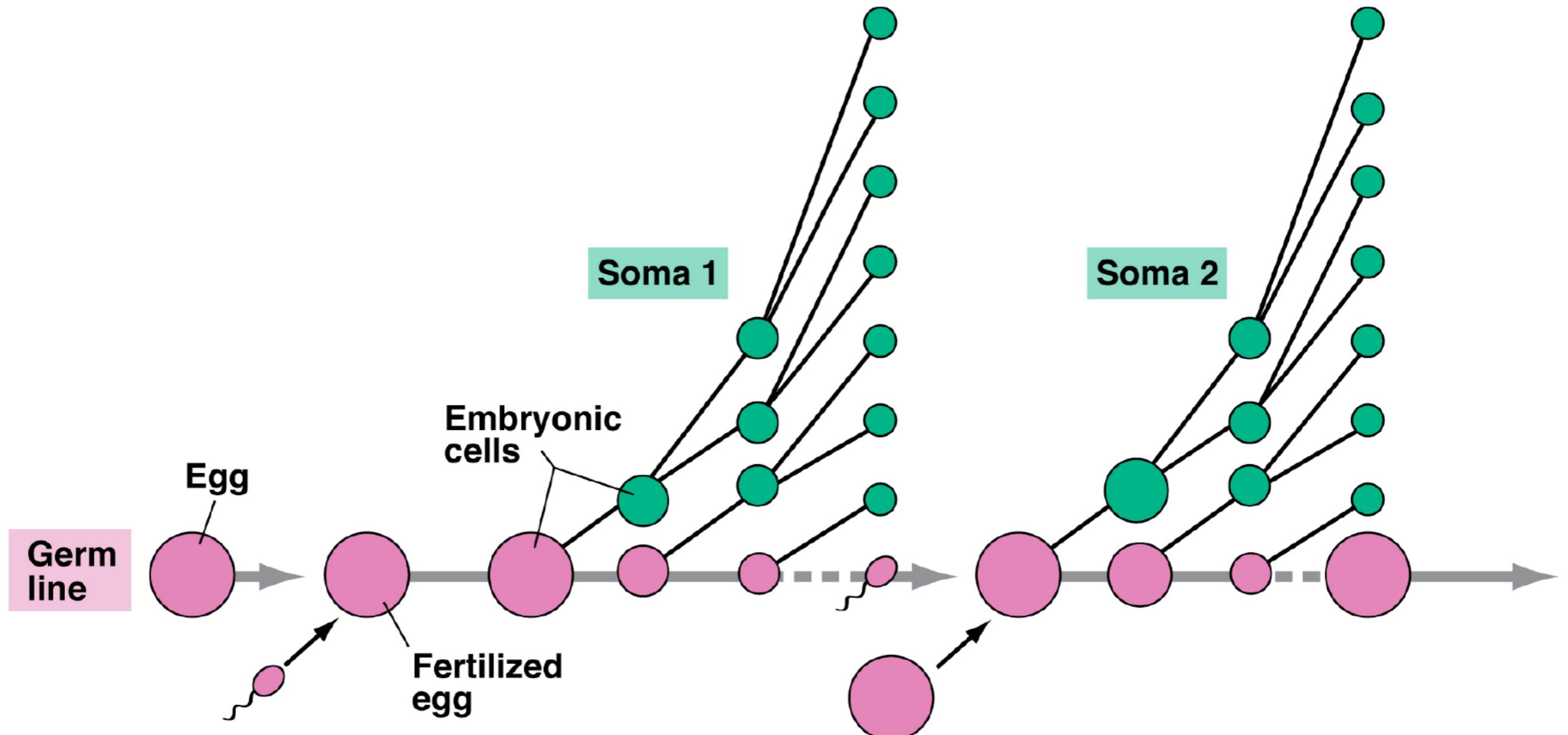
- Sistema común para todas las especies
 - Codones → aa
 - DNA para almacenar información genética
 - Síntesis de proteínas a través de ribosomas
 - Los mismos 20 aminoácidos

Registro fósil

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

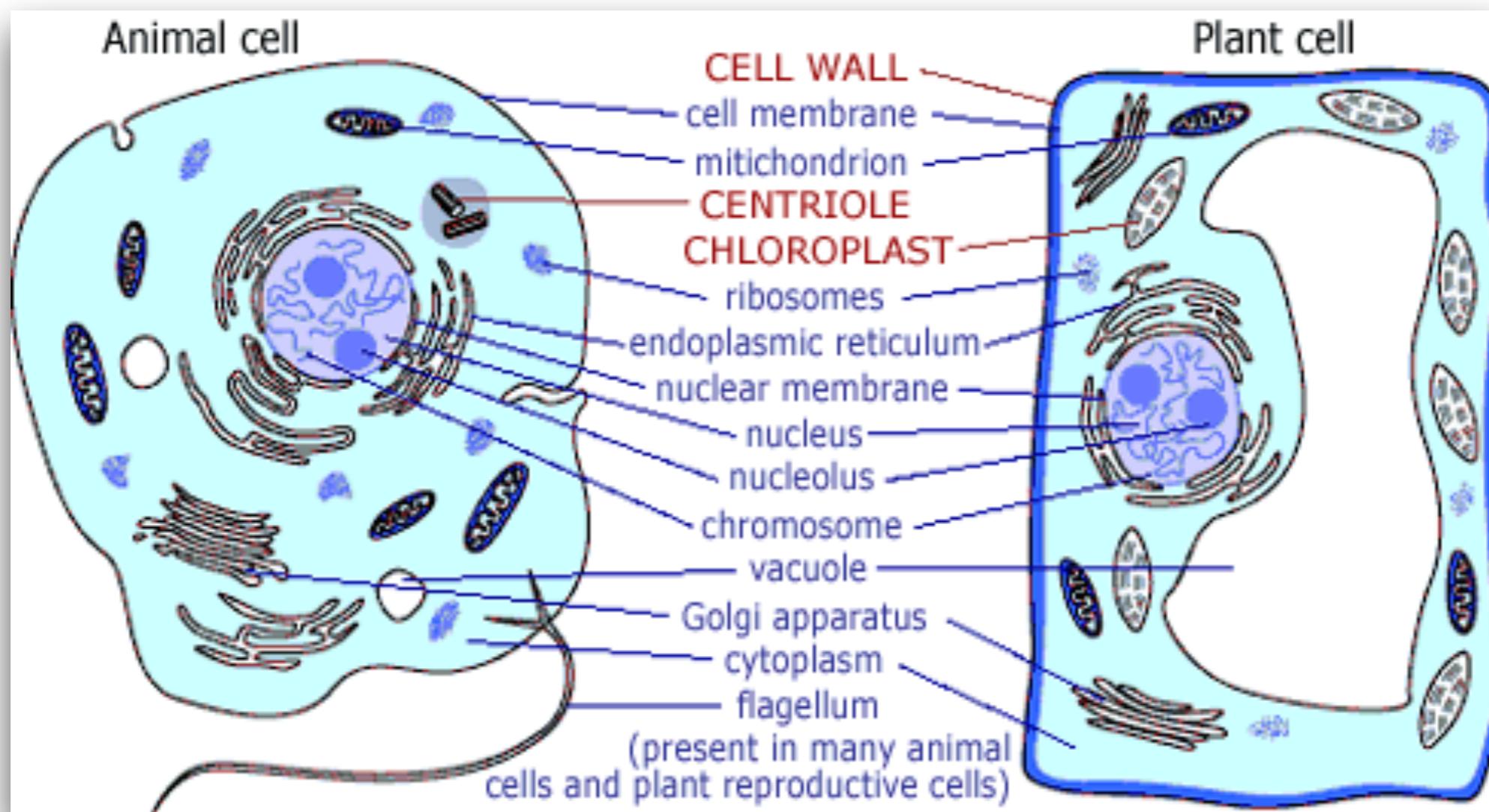


Linea germinal



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

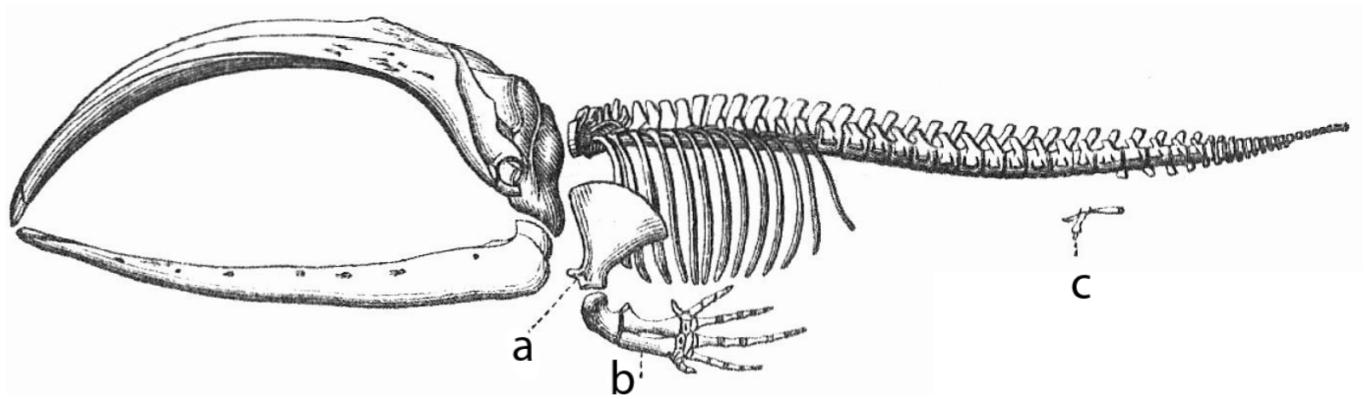
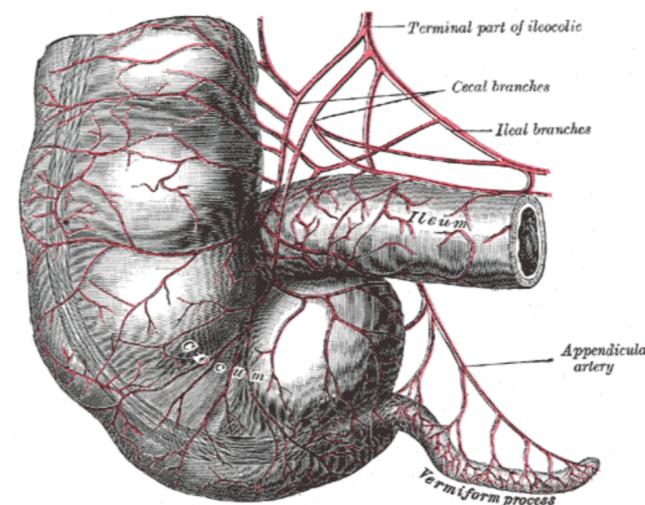
Estructura y función



Células de distintos organismos poseen estructuras funcionales similares

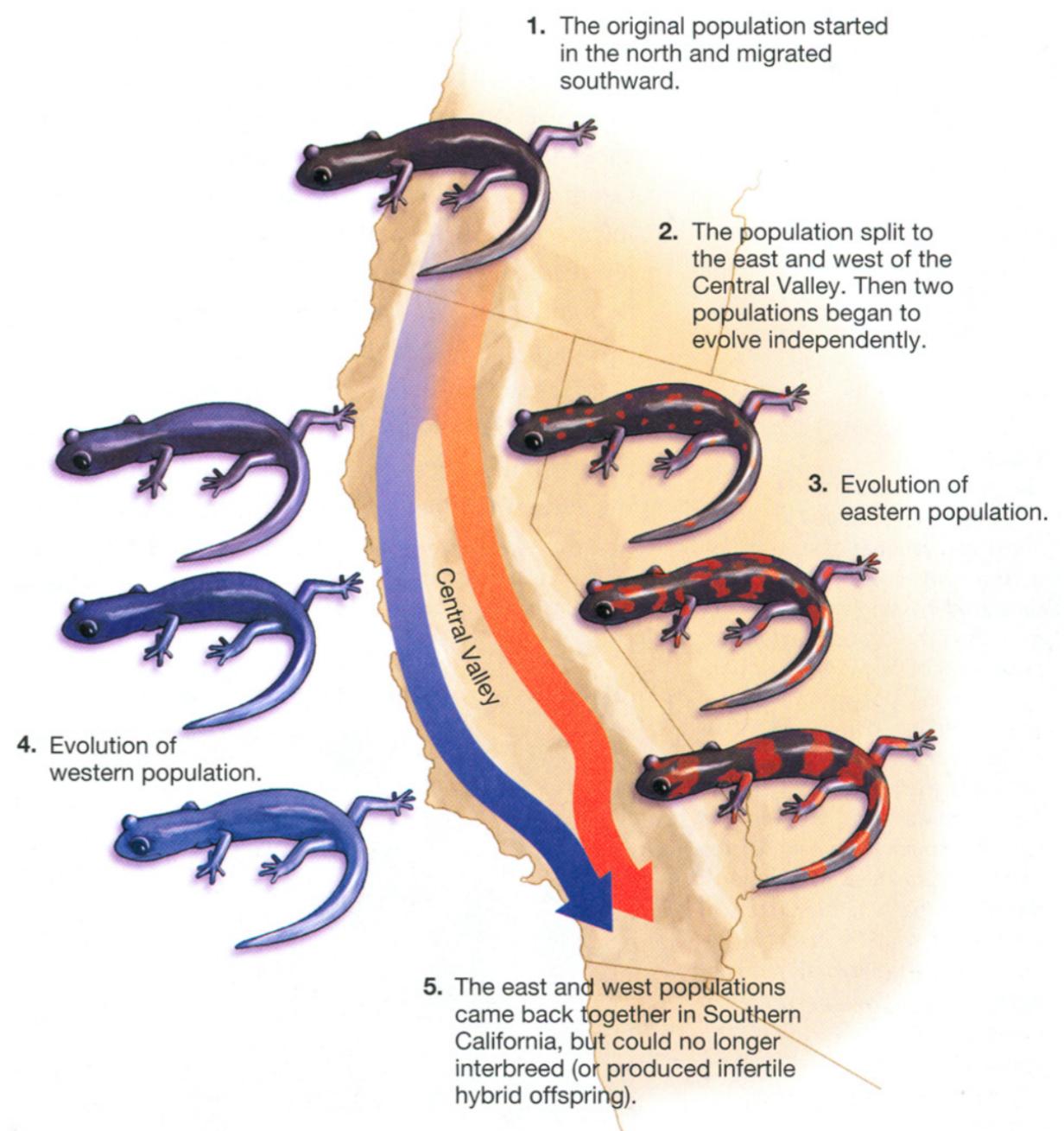
Rasgos vestigiales

- 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 geográficamente



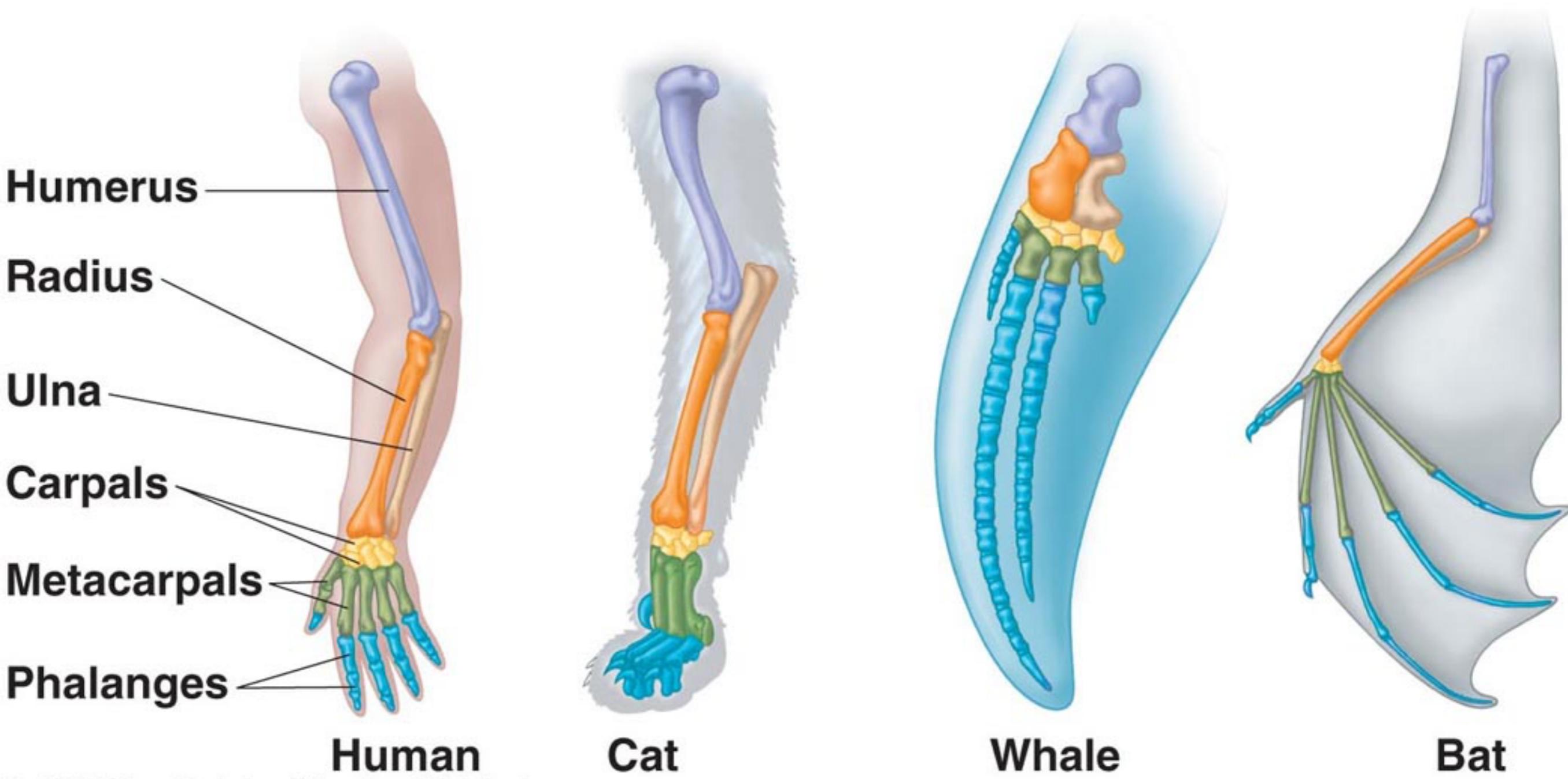
Homología

¿Qué es homología?

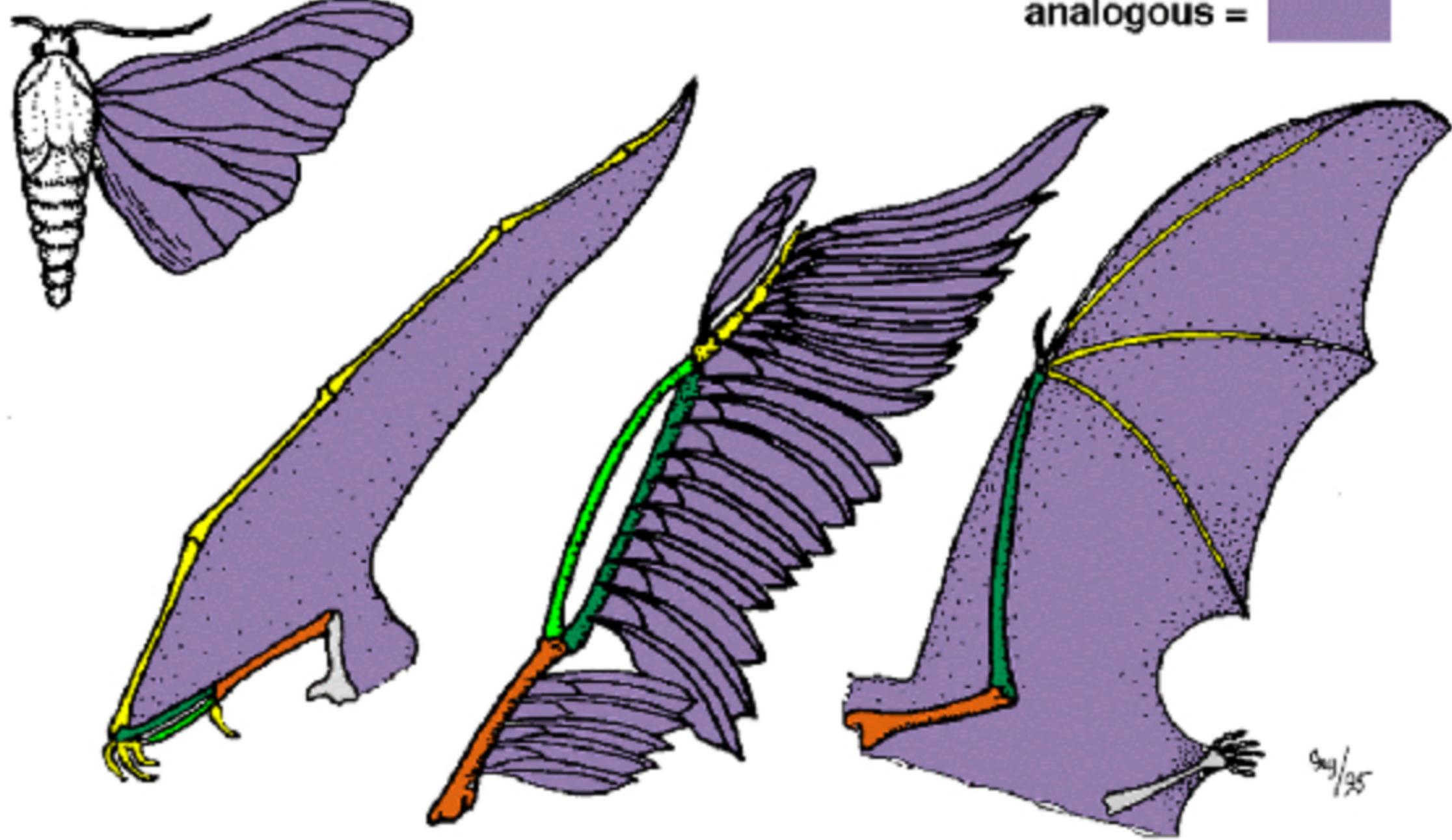
- **homología morfológica**
—> especies que pertenecen al mismo grupo taxonómico muestran similitudes anatómicas debido a que comparten un ancestro común (no por chance)



Estructuras homólogas



Estructuras análogas



99/95

¿Qué es homología?

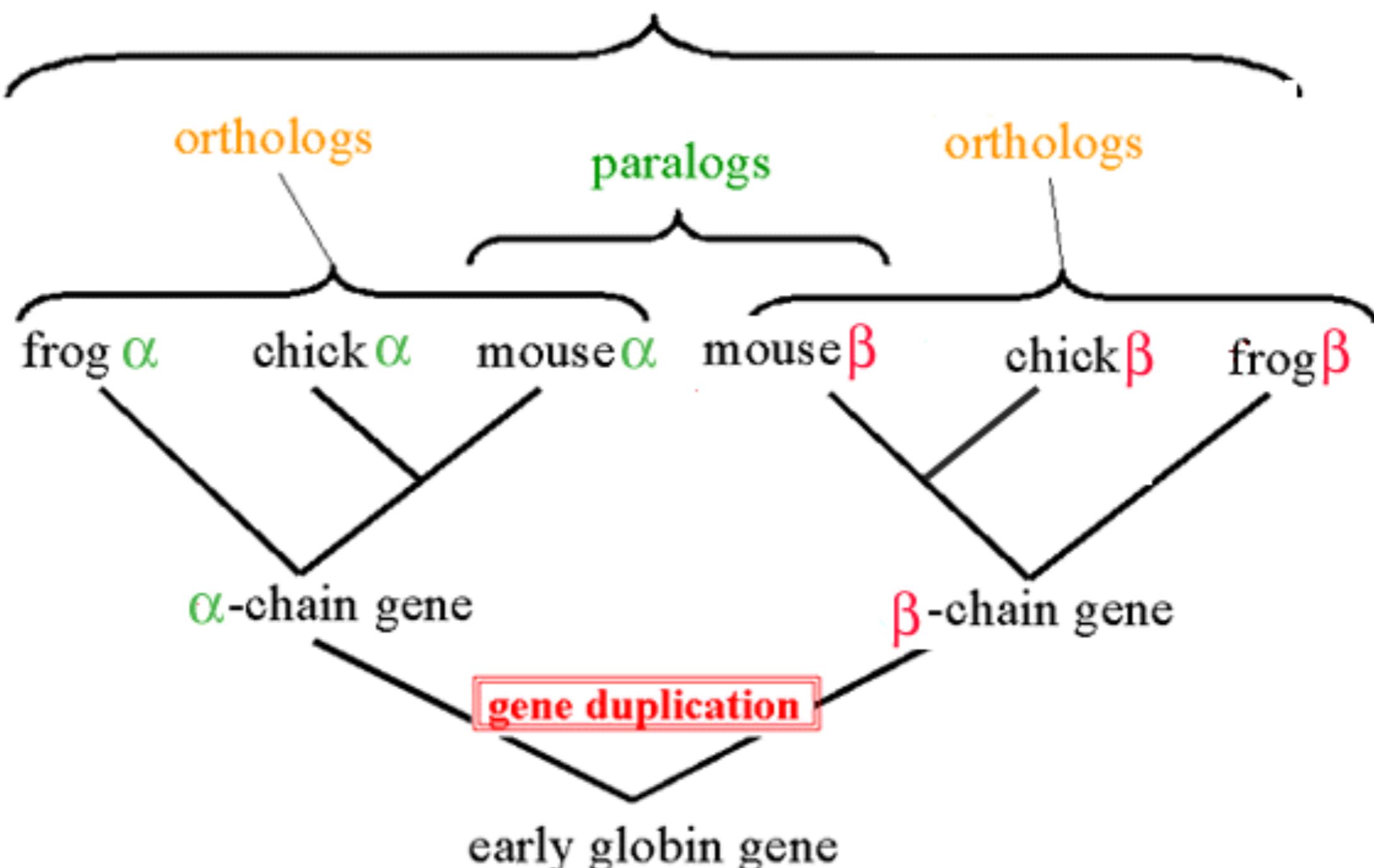
- **homología molecular**—> especies que pertenecen al mismo grupo taxonómico muestran similitudes en DNA, RNA y proteínas debido a que comparten un ancestro común (no por chance)

<i>Aquifex aeolicus</i>	MKVRSSVKK---	RCAKCKIIRRKGRVMVICE-IPSHKQKTG
<i>Bacillus subtilis</i>	MKVRPSVKP---	ICEKCKVIRRKGKVMVICE-NPKHKQKQG
<i>Campylobacter jejuni</i>	MKVRPSVKK---	MCDKCKVVRRKGVRRIICE-NPKHKQRQG
<i>Chlamydia trachomatis</i>	MRVSSSIKA---	PSKGDKLVRKGRLYVINKDPRNRKQRQA
<i>Escherichia coli</i>	MKVRASVKK---	LCRNCKIVKRDGVRVICSAEPHKQRQG
<i>Helicobacter pylori</i>	MKVRPSVKK---	MCDNCKIIRRGVRVIC-TPKHKQRQG
<i>Lactococcus lactis</i>	MKVRPSVKP---	ICEYCKVIRRNGRVMVICPANPKHKQRQG
<i>Mycobacterium leprae</i>	MKVNPSPVKP---	MCDKCRVIRRHRVMVICV-DPRHKQRQG
<i>Mycoplasma genitalium</i>	MKVRASVKP---	ICKDCKIIKRHRILRVICK-TKKHKQRQG
<i>Rickettsia prowazekii</i>	MKVSSLKSLKKRDKDQIVKRRGKIFVINKKNKFRAKQG	
<i>Synechocystis</i> sp.	MKVRASVKK---	MCDKCRVIRRGRVMVICSANPKHKQRQG
<i>Treponema pallidum</i>	MKIRTSVVKV---	ICDKCKLIKRFGIIRVIC-NPKHKQRQG
<i>Thermotoga maritima</i>	MKVQASVKK---	RCEHCKIIRRKKRVYVICKVNPKNQKQG
<i>Vibrio cholerae</i>	MKVRASVKK---	ICRNCKVIKRNGVVRVIC-SEPKHKQRQG
<i>Xylella fastidiosa</i>	MKVSSLKSAKTRHRDCKVIRRKGKIFVICKSNPRFKARQR	
Yeast	...	FKVRTSVKK---FCSDCYLVRRKGRTVYIYCKSNKKHKQRQG
Rice	...	MKIRASVRK---ICTKCRLLIRRGRIRVIC-SNPKHQQRQG
Fruit fly	...	FKVKGRLKR---RCKDCYIVVRQERGYVICPTHPRHKQMSM
Mouse	...	FTKKGVIKK---RCKDCYKVRRGRWFILCKTNPKHKQRQM
Human	...	FKNKTVLKK---RCKDCYLVKRRGRWYVYCKTHPRHKQRQM

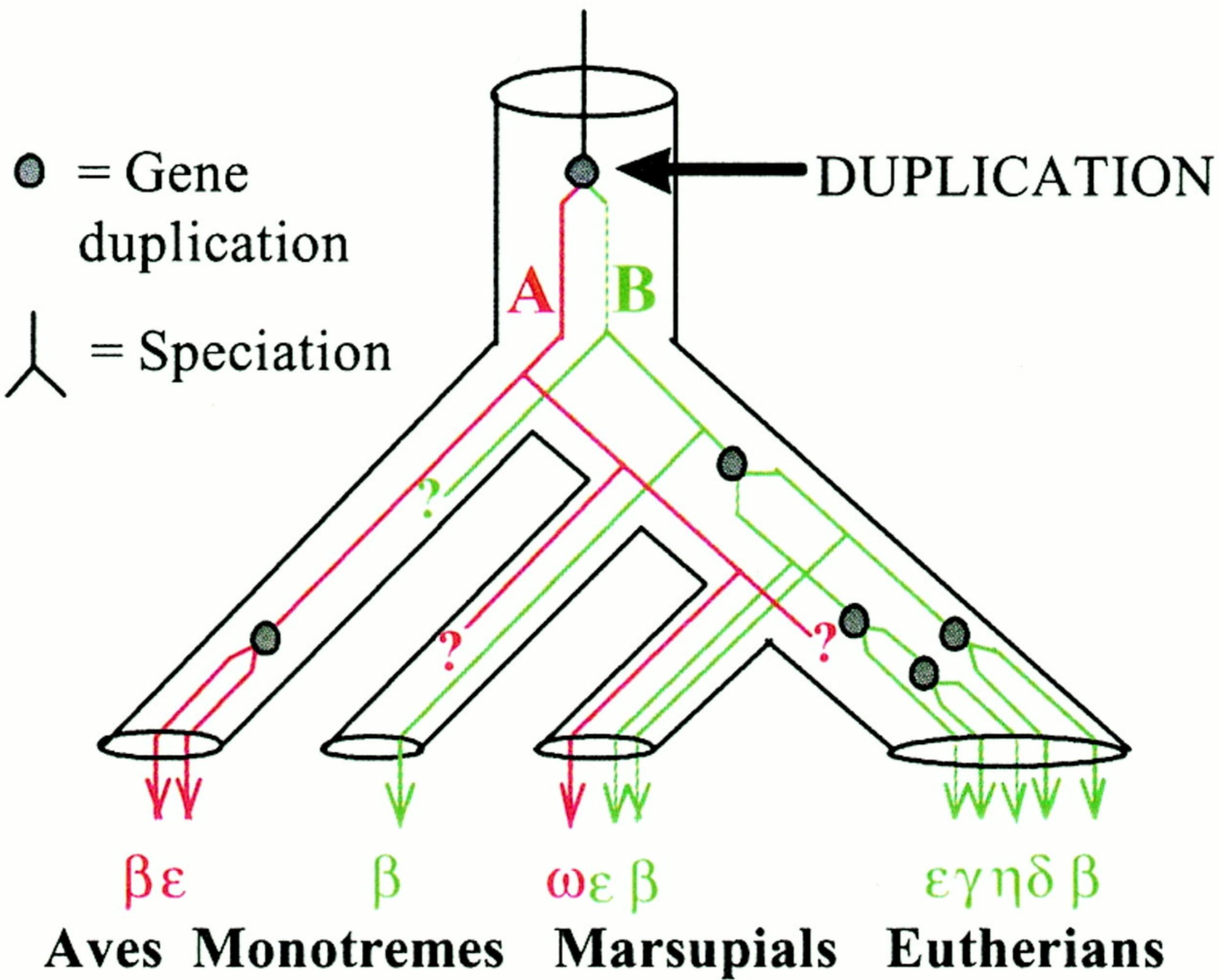
Tipos de homología

- Ortólogos
- Paralógos
- Xenólogos

Tipos de homología



Ancestral β -globin gene

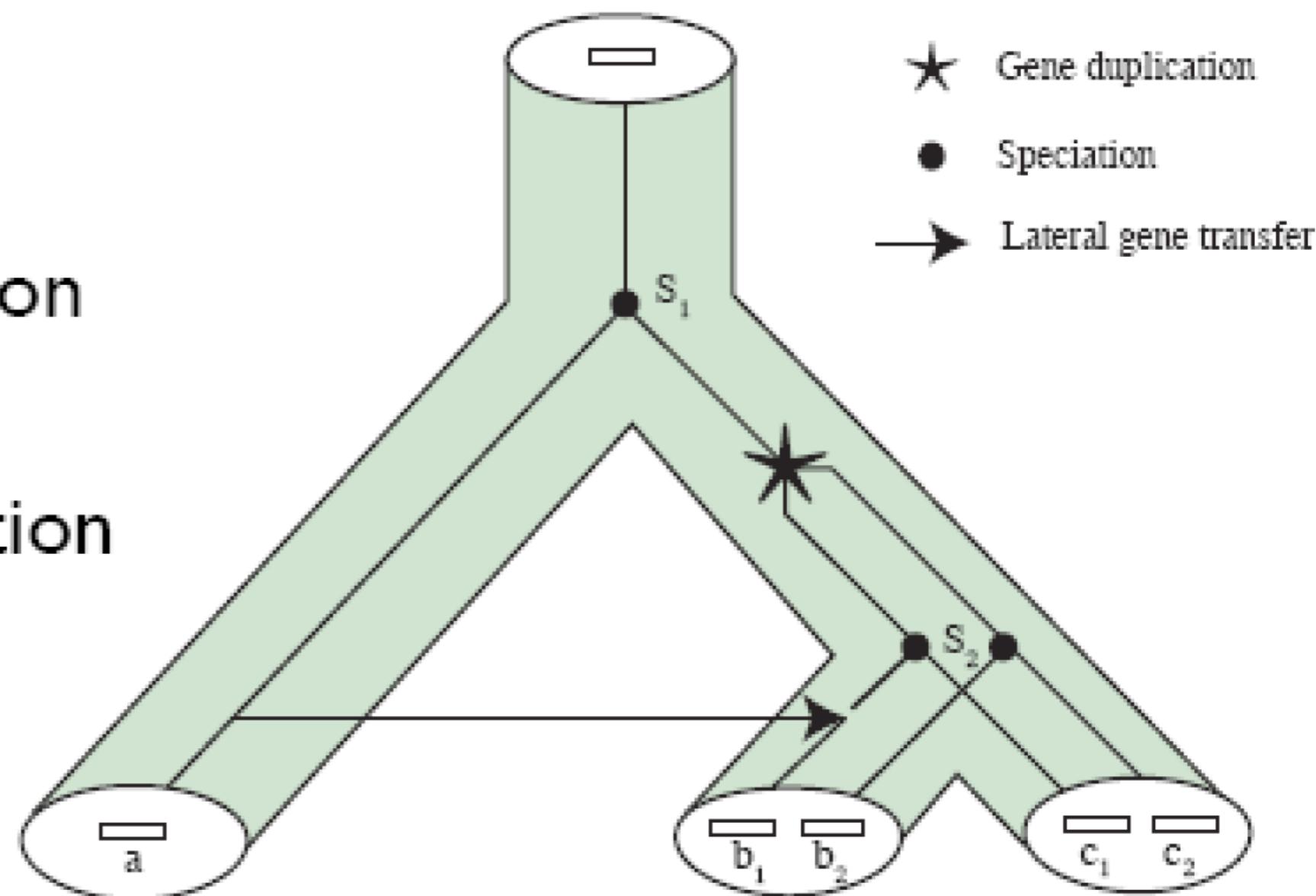


Tipos de homología

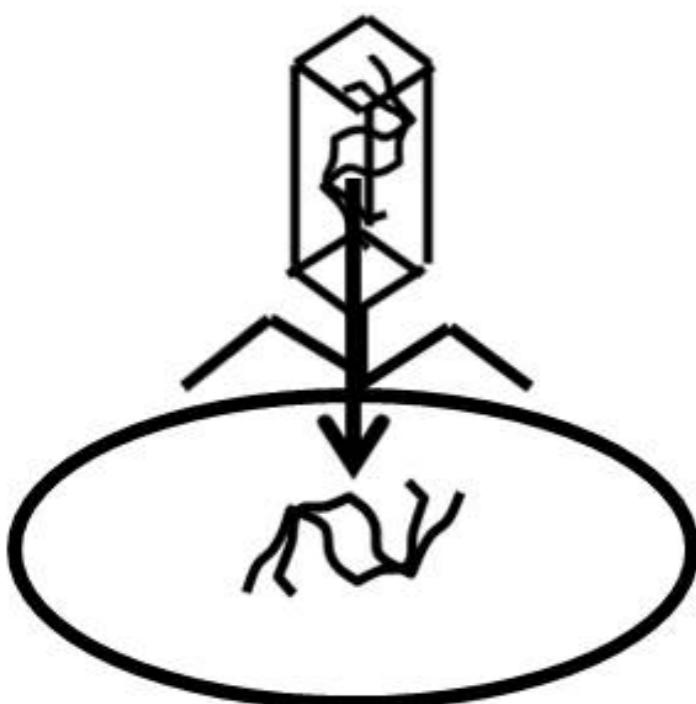
Two genes (or characters) are *homologs* if they have a common ancestor.

Main Subtypes

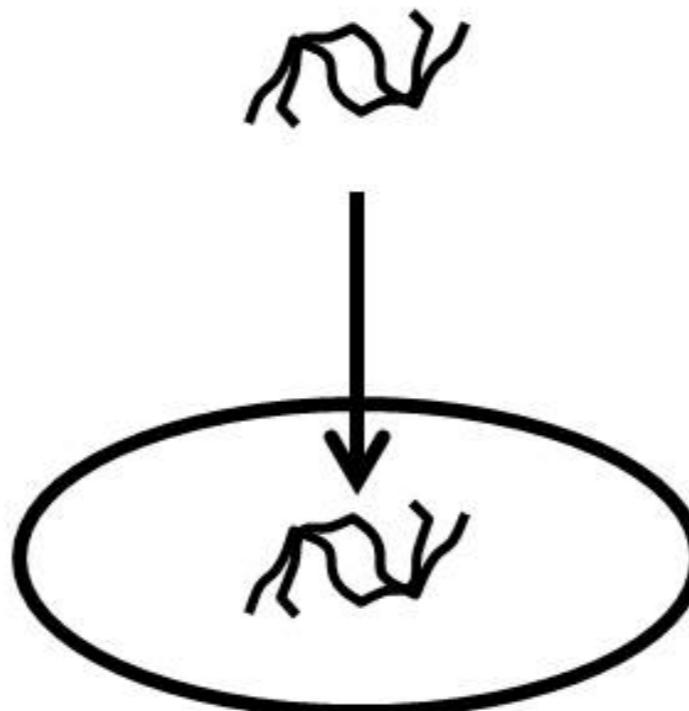
- *Orthologs*: through speciation
- *Paralogs*: through duplication
- *Xenologs*: through lateral transfer



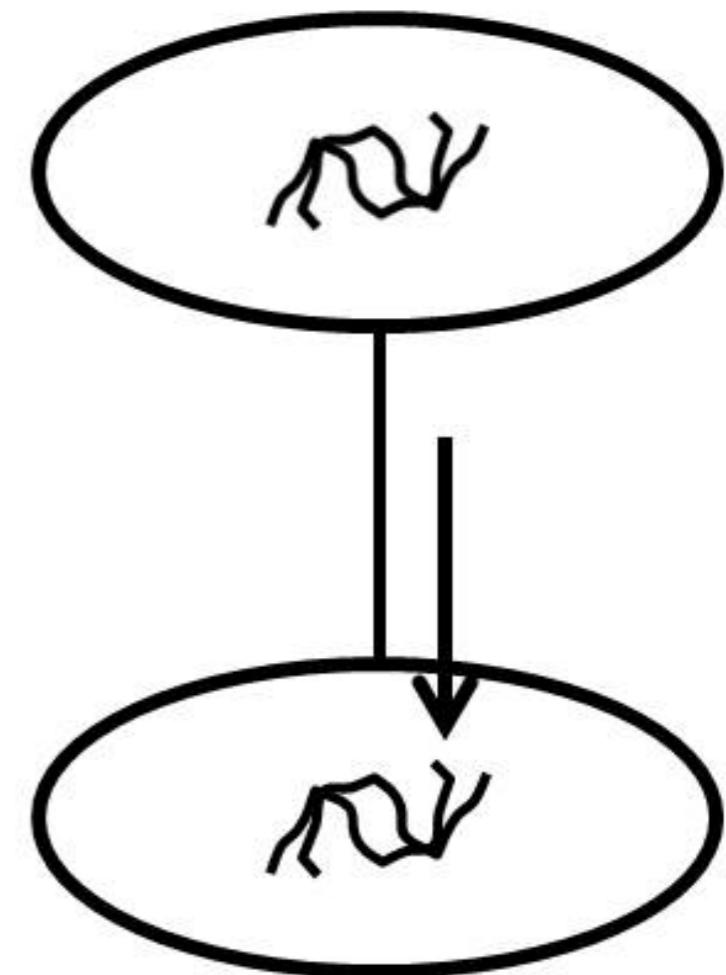
Xenólogos en bacteria



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

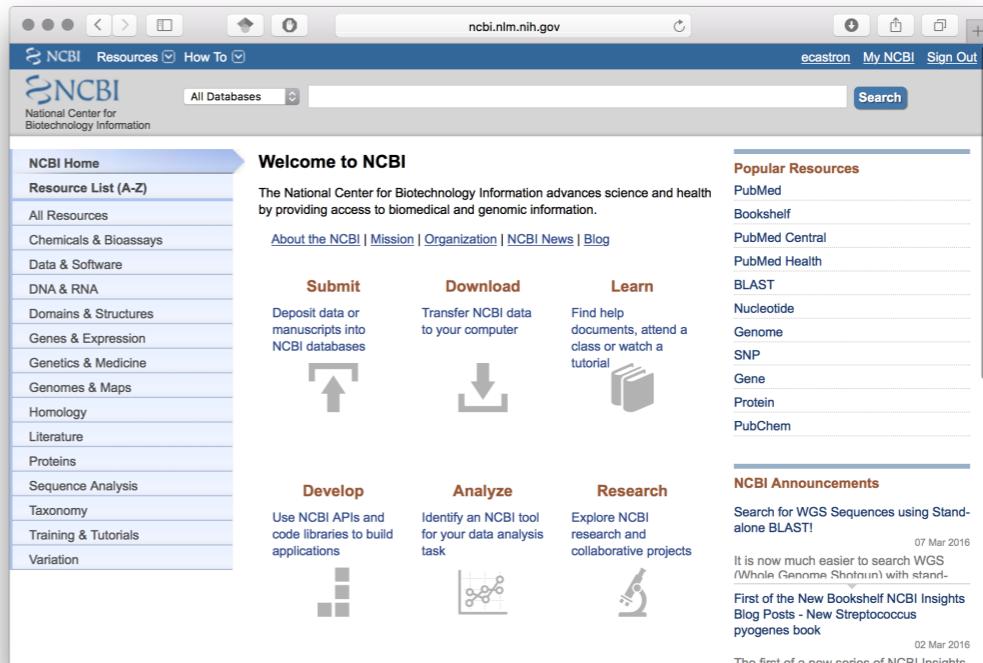
Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources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

The screenshot shows the homepage of Nucleic Acids Research. At the top, the journal title "Nucleic Acids Research" is displayed in white on a maroon background. Below the title are links for "ABOUT THIS JOURNAL", "CONTACT THIS JOURNAL", and "SUBSCRIPTIONS". A navigation bar at the bottom includes "Oxford Journals > Science & Mathematics > Nucleic Acids Research > Volume 44, Issue". A main heading reads "The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection". Below this, authors Daniel J. Rigden, Xosé M. Fernández-Suárez, and Michael Y. Galperin are listed, along with a link to "Author Affiliations".

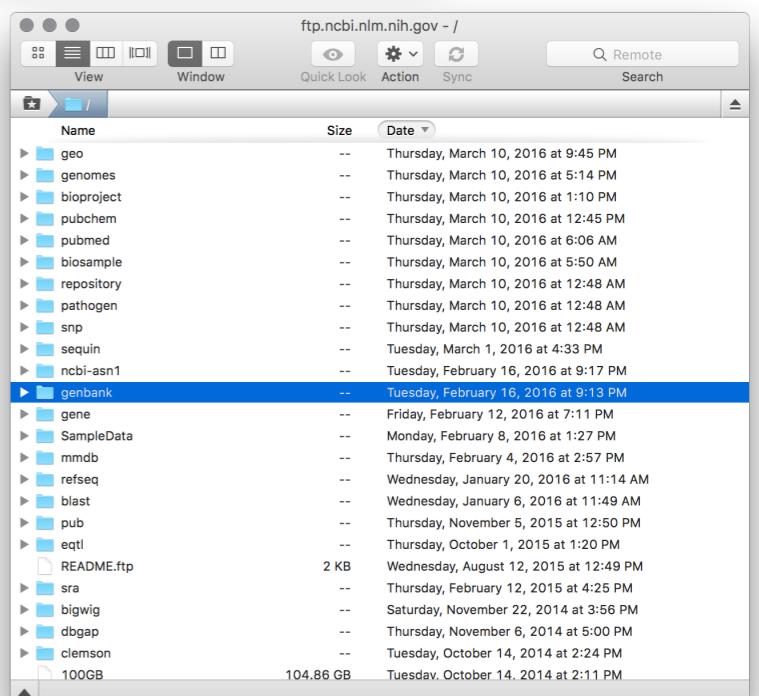
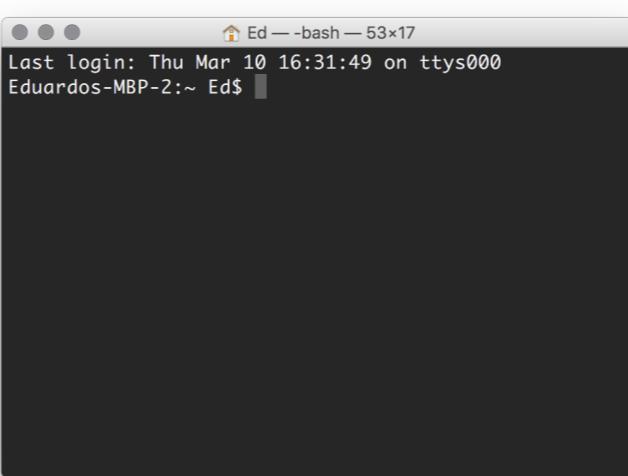
The screenshot shows the homepage of Database: The Journal of Biological Databases and Curation. The title "DATABASE" is prominently displayed in large white letters on a dark blue background. Below it, the subtitle "The Journal of Biological Databases and Curation" is shown. A navigation bar at the top includes "ABOUT THIS JOURNAL", "CONTACT THIS JOURNAL", "SUBSCRIPTIONS", and "CURRENT ISSUE". A navigation bar at the bottom includes "Oxford Journals > Science & Mathematics > Database". A "READ THIS JOURNAL" button is visible. On the right, there is a sidebar with the text "Welcome to Database: The Journal of Biological Databases and Curation", "A Fully Open Access Journal", and links to "View Current Content", "Browse the Archive", "Biocuration Virtual Issue", "Biomart Virtual Issue", and "Now Indexed in PubMed Central".

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

- Interface gráfica →

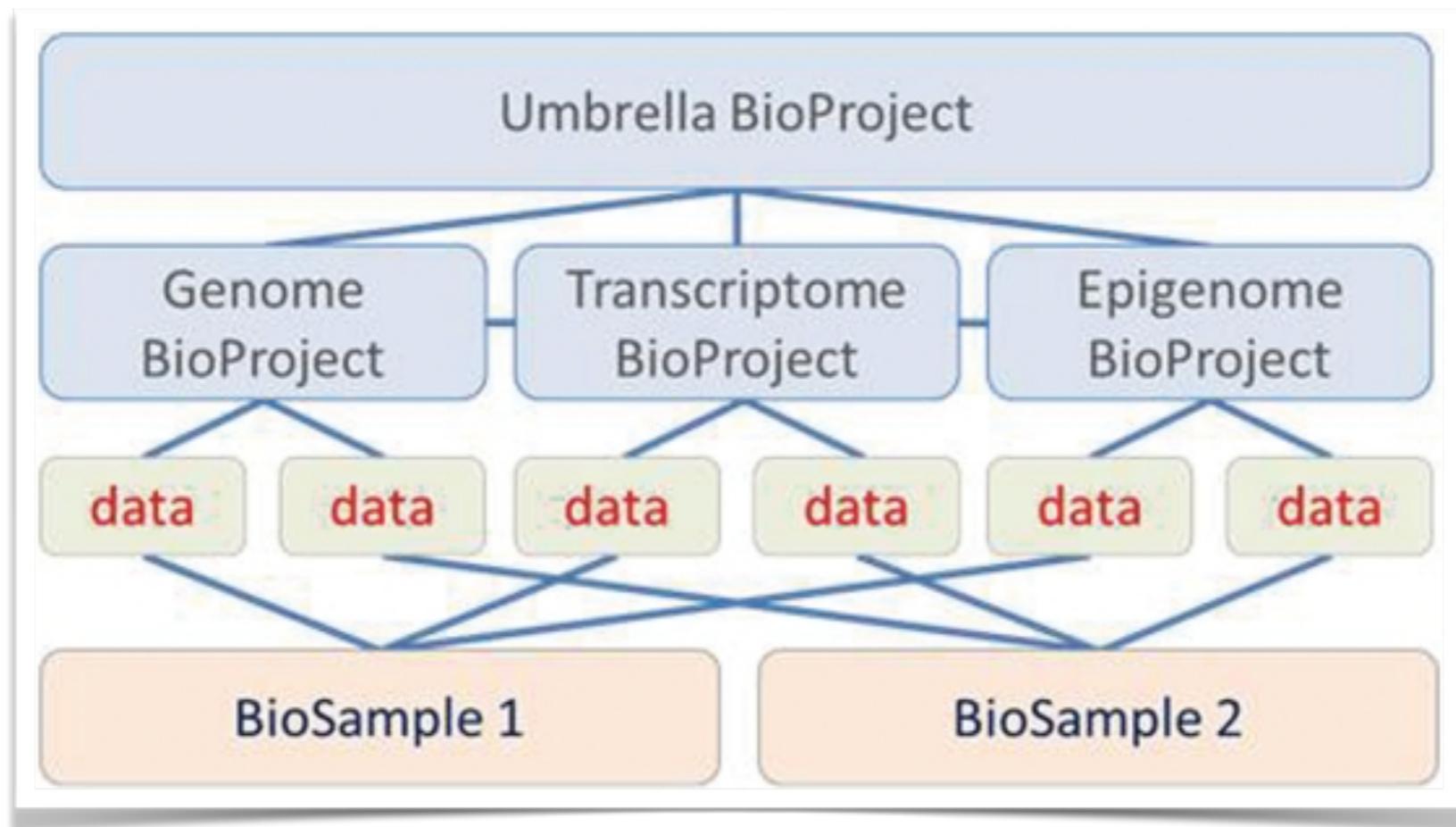


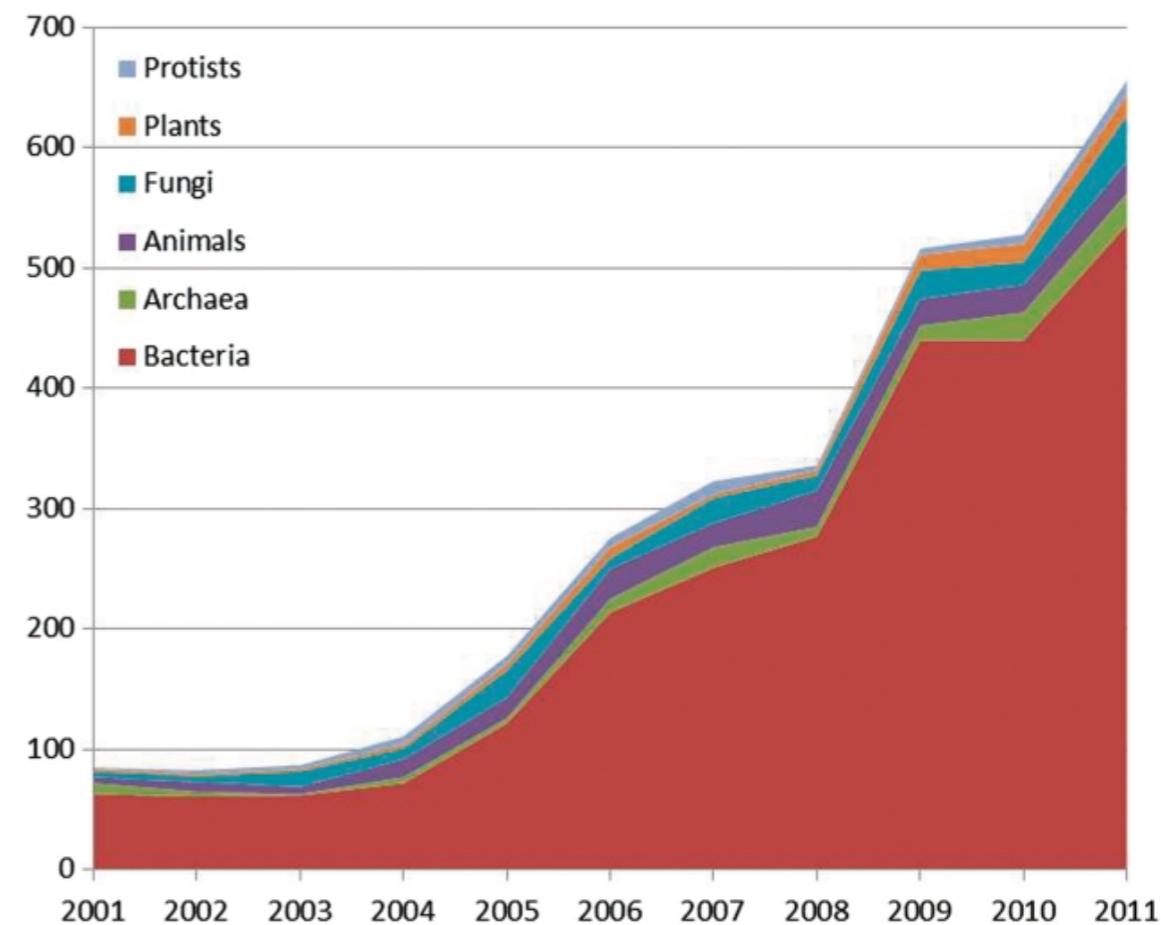
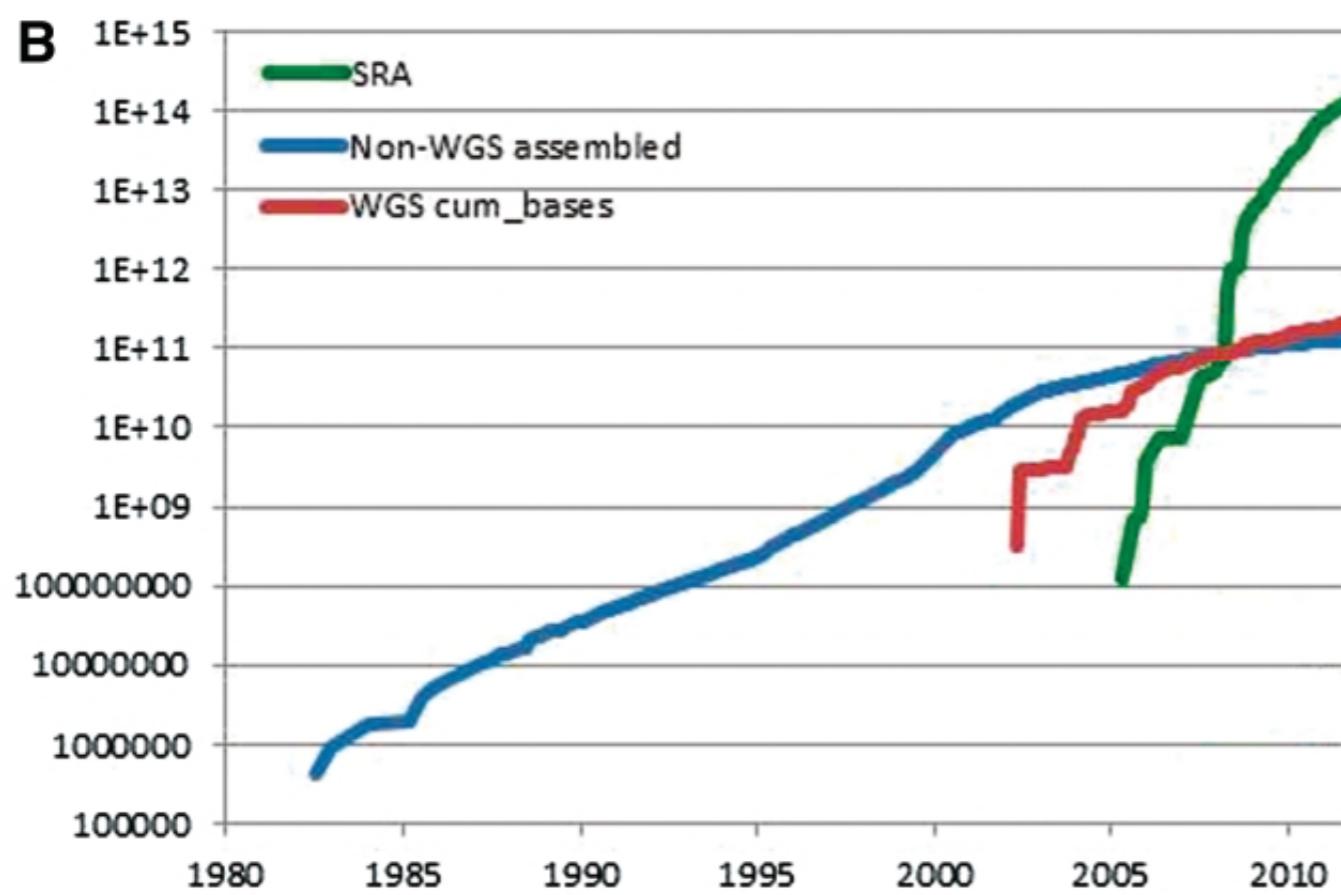
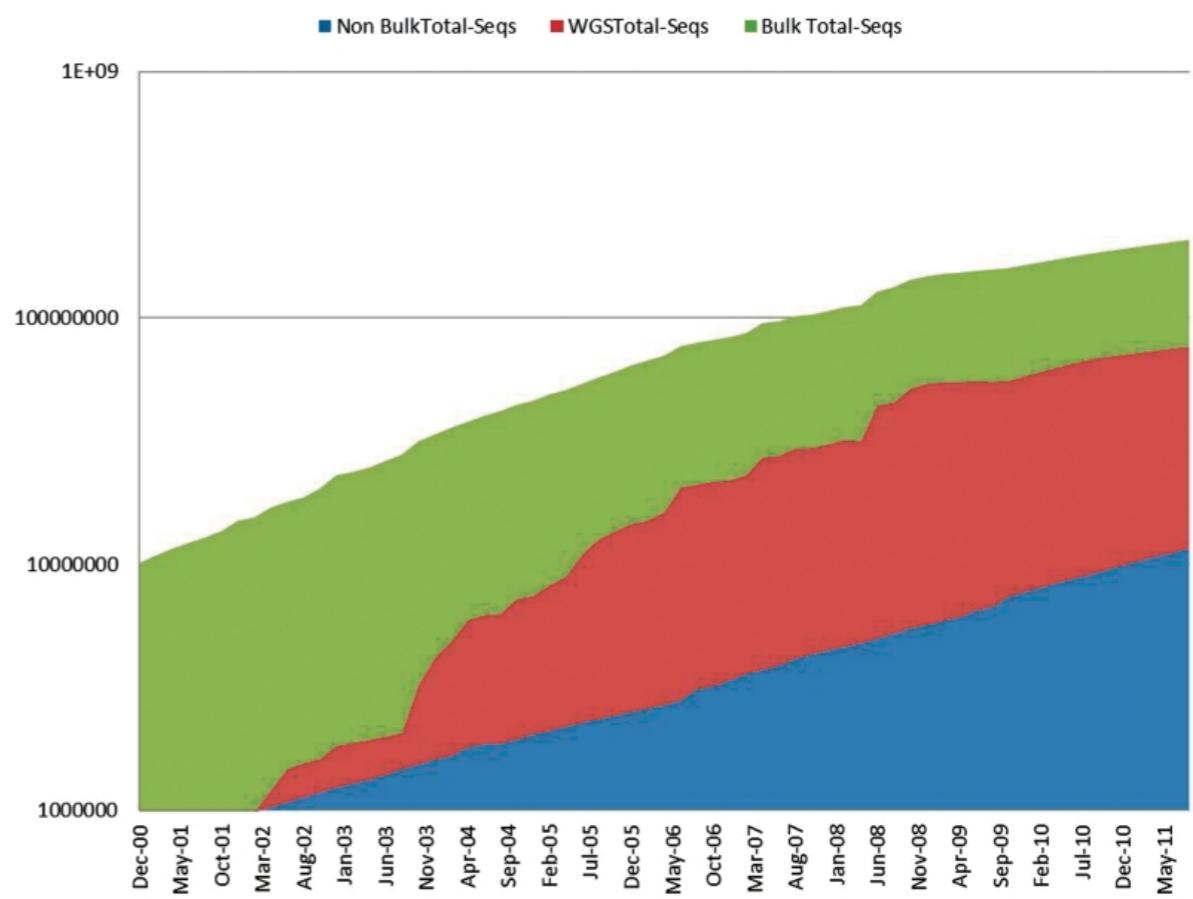
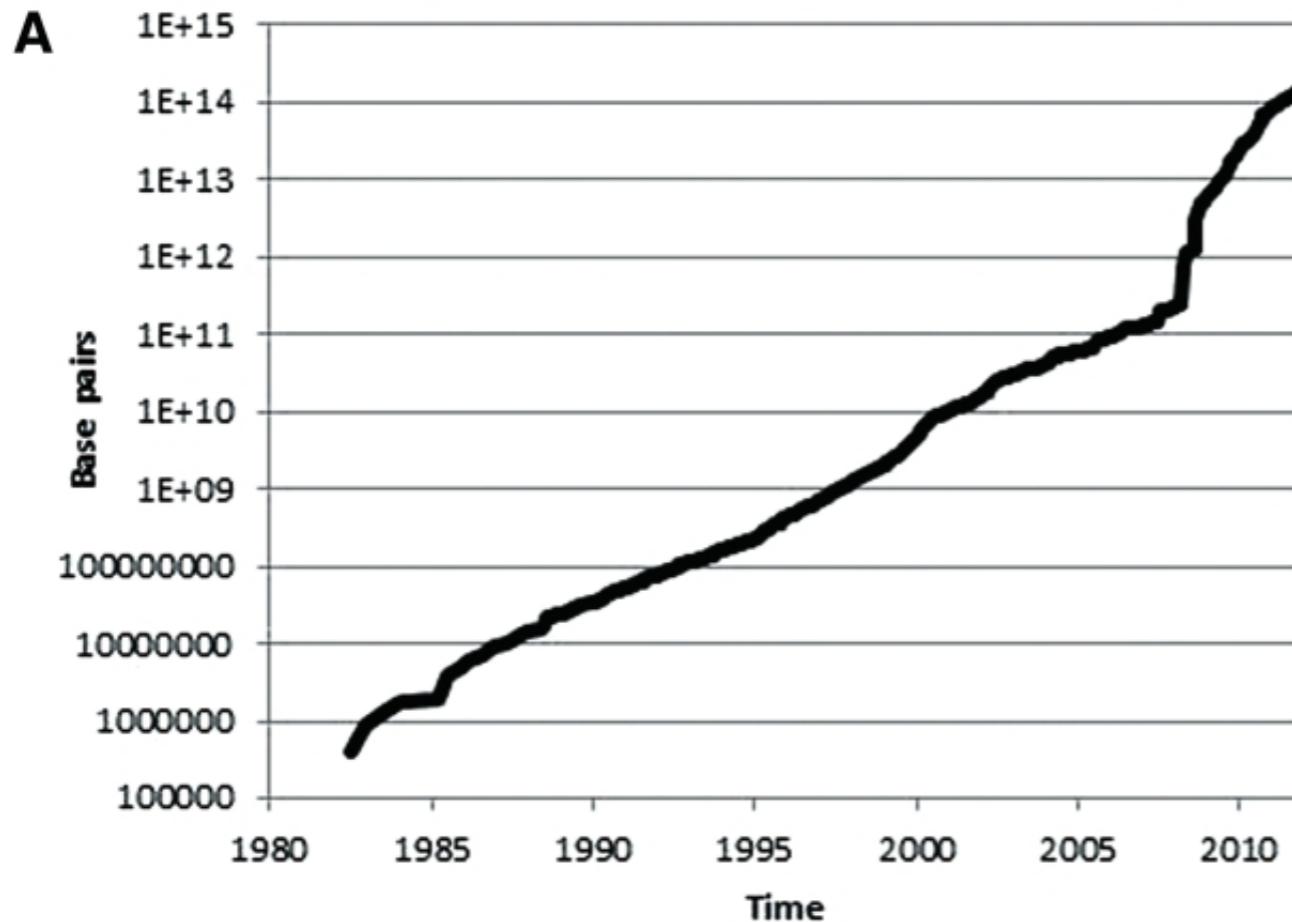
- “Puerta de atrás” FTP, scripting →
ftp.ncbi.nlm.nih.gov



Ejemplo más popular

- International Nucleotide Sequence Database Collaboration (INSDC)
- 30 años de antigüedad





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

The image displays three side-by-side screenshots of biological databases:

- KEGG:** The screenshot shows the main KEGG homepage. It features a sidebar with links to KEGG Home, Database, Objects, Software, and more. The main content area is titled "KEGG: Kyoto Encyclopedia of Genes and Genomes" and provides an overview of the database's purpose and features.
- BRENDA:** The screenshot shows the BRENDA homepage. It features a banner announcing a new release and links to Word Maps, Pathway Maps, and a search bar. The main content area includes sections for Sequence Search, Genome Explorer, and Ontology Explorer.
- CAZymes:** The screenshot shows the CAZymes homepage. It features a sidebar with links to What's new, Definitions and Terminology, Help, Citing CAZy, PULDB, Enzyme & Glyco Resources, Commercial Providers, Scientific Meetings, About Us, and Position(s) available. The main content area is titled "Welcome to the Carbohydrate-Active enZYmes Database" and provides an overview of the database's purpose and features.

... de proteínas

- Protein Data Bank

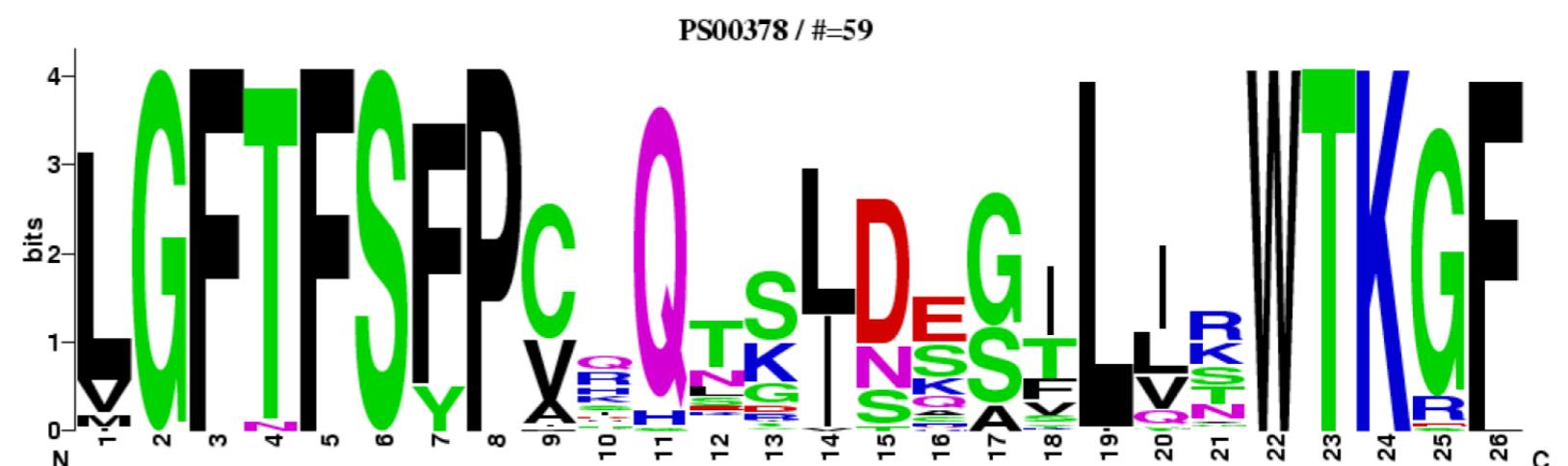
The screenshot shows the homepage of the RCSB Protein Data Bank (PDB) at rcsb.org. The top navigation bar includes links for Deposit, Search, Visualize, Analyze, Learn, More, and MyPDB Login. Below the header is the PDB logo and a banner stating "An Information Portal to 116816 Biological Macromolecular Structures". A search bar allows users to search by PDB ID, author, macromolecule, sequence, or ligands. Below the search bar are links for Advanced Search and Browse by Annotations. Logos for PDB-101, Worldwide PDB, EMDDataBank, NDB, and Structural Biology Knowledgebase are displayed, along with social media icons. The main content area features a "Welcome" section with a "Feature Highlight: Gene View" showing a screenshot of the software interface and a "Gene View Tutorial" video thumbnail. To the right, a "March Molecule of the Month" section highlights the "RAF Protein Kinases" in a 3D ribbon model, with red, blue, and purple components. A "Contact Us" button is located at the bottom right.

Bases de datos secundarias

- PROSITE
- Pfams
- Rfams
- PRINTS

PROSITE [HEXOKINASES PS00378]

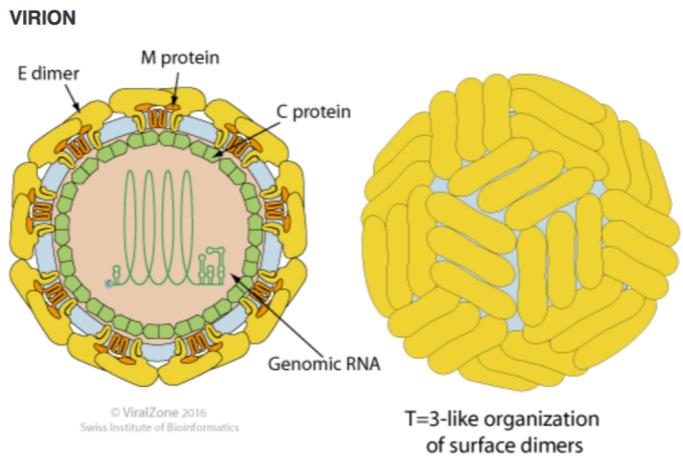
- Database of protein domains, families and functional sites
- Hexokinases signature: Pattern [LIVM]-G-F-[TN]-F-S-[FY]-P-x(5)-[LIVM]-[DNST]-x(3)-[LIVM]-x(2)-W-T-K-x-[LF].



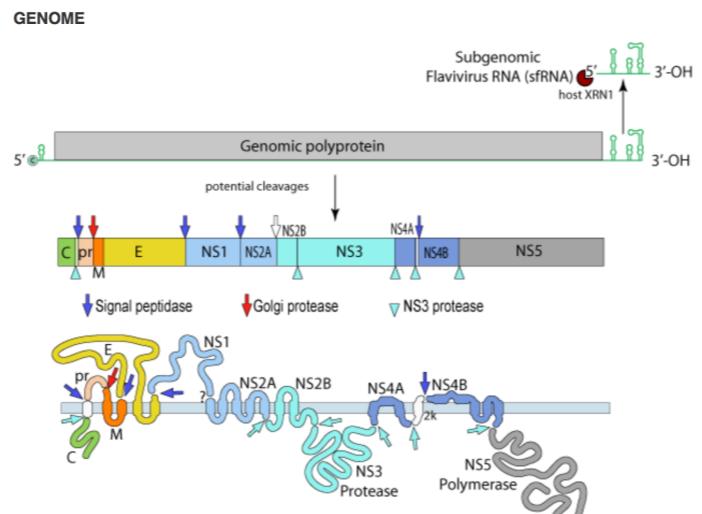
Zika virus y taxonomía

Zika virus (strain Mr 766)

- UniProt Taxonomy
- NCBI Taxonomy
- ViralZone



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



Monopartite, linear, ssRNA(+) genome 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](https://github.com/bioinf-biotec)