

Métodos filogenéticos en la lingüística histórica

Las base de datos IELex y AJSP en el análisis y datación de lenguas

Javier Muñoz

Universidad de Valladolid

Cognado

- *Los cognados son palabras que se han desarrollado a partir de la misma raíz.*
- *Los cognados pueden tener un mismo origen etimológico, pero distinta evolución fonética y, a menudo, semántica (falsos amigos)*
- *Por ejemplo: lat. filius > fr. fils / it. figlio*
- *Al. Vater / ingl. father / su. far < germ. *faþer*
- *Podemos enunciar una regla por la cual cuanto más cerca están dos lenguas, mayor número de cognados tienen.*

Cognados

Deutsch	Althochdeutsch	Luxemburgisch	Niederländisch	Afrikaans	Altsächsisch	Altenglisch	Englisch	Altnordisch	Gotisch	Germanisch	Urindogermanisch
Vater	fater	.	vader	vader	fadar	fæder	father	faðir	fadar	*fader	*pētér
Mutter	muoter	.	moeder	moeder	modar	modor	mother	móðir	.	*mōðer	*mater
Bruder	bruoder	Brudder	broe(de)r	broer	broðar	broðor	brother	bróðir	broþar	*brōþer	*bhrater
Schwester	swester	Schwēster	zus(ter)	suster	swestar	sweostor	sister	systir	swistar	*swester	*suesor
Tochter	tohter	Duechter	dochter	dogter	dohtar	dohtar	daughter	dóttir	dauhtar	*duxter	*dhugeter
Sohn	sunu	.	zoon	seun	sunu	sunu	son	sunr	sunus	*sunuz	*suənu
Herz	herza	Häerz	hart	hart	herta	heorte	heart	hjarta	hairto	*χertōn	*kerd
Knie	knio	Knéi	knie	knie	knio	cneo	knee	kné	kniu	*knewa	*genu
Fuß	fuoz	Fouss	voet	voet	fot	fot	foot	fótr	fotus	*fōt-	*pod
Aue**	ouwi		ooi	ooi	ewwi	eowu	ewe	æw	aweþi	*awi	*owi
Kuh	kuo	Kou	koe	koei	ko	cu	cow	kýr	.	*k(w)ou	*gwou
Elch	elaho, eliho		eland		elaho	eolh, eolk	elk	elgr		*elhaz, *algiz	*h ₁ élk _{is} , *h ₁ ólk _{is}
Mähre	meriha		merrie		merge	mere, miere	mare	merr		*marhijō	*mark-
Schwein	swin	Schwäin	zwijn	swyn	swin	swin	swine	svín	swein	*swina	*sus/suino
Hund	hunt	Hond	hond	hond	hund	hund	° hound	hundr	hunds	*χundaz	*kuon
Wasser	wazzar	Waasser	water	water	watar	wæter	water	vatn	vato	*watōr	*wódr ₂
Feuer	fiur	Feier	vuur	vuur	fiur	fýr	fire	fúrr		*fōr, *fuīr	*péh ₂ ur
(Baum)			(boom)		trio	treo(w)	tree	tré	triu	*trevam	*deru
(Rad)			wiel	wiel		hweol	wheel	hvél		*hwehwla ₂	*k ^h ék ^h lo-
neu	niuwi	nei	nieuw	nuwe	niuwi	niwe	new	nýr	niujis	*neuja	*neujo

La lista Swadesh

- *Morris Swadesh (1909-1967) ['swɒdɛʃ]*
- *Lista de vocabulario básico resistente a préstamos, formado por palabras comunes existentes en cualquier lengua.*
- *Inicialmente compuesta por 200 términos, reducidos a 100 posteriormente.*
- *Según Swadesh el listado permite mediante la comparación establecer la relación entre dos lenguas.*
- *Constante glotocronológica. El ritmo del cambio es fijado en un promedio del 14% cada 1000 años (constante de 86%)*

La lista Swadesh

- Se podría expresar en forma de ecuación: siendo $P(t)$ el porcentaje de vocabulario retenido y t el periodo de tiempo

$$\frac{dP(t)}{dt} = \alpha P(t)$$

Al integrar

$$P(t) = 100e^{-\alpha t}$$

- De este modo, la separación temporal de lenguas (T) puede computarse contrastando el porcentaje de cognados comunes/retenidos (r), con la constante de retención (γ).

$$T = \frac{\log(r)}{\log(\gamma)}$$

- <http://bit.ly/2jAFV1i>

- Principios básicos:

- I. En el léxico de cualquier lengua puede localizarse un vocabulario básico/estable
- II. La tasa de retención es constante. Significa que un número de palabras del vocabulario básico persistirá en el tiempo.
- III. Partiendo del porcentaje de cognados compartidos entre dos lenguas podemos computar el tiempo transcurrido desde que se separaron.

Inglés	Español	Inglés	Español	Inglés	Español	Inglés	Español
I, me	(yo, me)	root	(raíz)	breasts	(senos)	stone	(piedra)
you	(tú)	bark	(corteza)	heart	(corazón)	sand	(arena)
we	(nosotros)	skin	(piel)	liver	(hígado)	earth	(tierra)
this	(esto)	flesh	(carne)	to eat	(comer)	cloud	(nube)
that	(eso)	blood	(sangre)	to drink	(beber)	smoke	(humo)
who	(quién)	bone	(hueso)	to bite	(morder)	fire	(fuego)
what	(qué)	grease	(grasa)	to see	(ver)	ash	(ceniza)
not	(no)	egg	(huevo)	to hear	(oír)	burn	(quemar)
all	(todo/s)	horn	(cuerno)	to know	(saber)	path	(camino)
many	(muchos)	tail	(rabo)	to sleep	(dormir)	mountain	(montaña)
one	(uno)	feather	(pluma)	to die	(morir)	red	(rojo)
two	(dos)	hair	(pelo)	to kill	(matar)	green	(verde)
big	(grande)	head	(cabeza)	to swim	(nadar)	yellow	(amarillo)
long	(largo)	ear	(oreja)	to fly	(volar)	white	(blanco)
small	(pequeño)	eye	(ojo)	to walk	(caminar)	black	(negro)
woman	(mujer)	nose	(nariz)	to lie	(recostarse)	night	(noche)
man	(hombre)	mouth	(boca)	to come	(venir)	hot	(caliente)
person	(persona)	tongue	(lengua)	to sit	(sentarse)	cold	(frío)
fish	(pez)	tooth	(diente)	to stand	(estar en pie)	full	(lleno)
bird	(pájaro)	claw	(garra)	to say	(decir)	new	(nuevo)
dog	(perro)	foot	(pie)	sun	(sol)	good	(bueno)
louse	(piojo)	knee	(rodilla)	moon	(luna)	round	(redondo)
tree	(árbol)	hand	(mano)	star	(estrella)	dry	(seco)
seed	(semilla)	belly	(panza)	water	(agua)	name	(nombre)
leaf	(hoja)	neck	(cuello)	rain	(lluvia)	to give	(dar)

La lista Swadesh

✦ *Críticas:*

- 1. Coseriu afirma que no existe un léxico universal (lista pensada para el inglés).*
- 2. Puede tratarse de préstamos y no cognados reales*
- 3. Los cambios fónicos pueden afectar al reconocimiento*
- 4. Ejemplo: Si comparamos los cognados del español y el hindi obtenemos un 23% de cognados comunes. Si comparamos el español y el pastún obtenemos un 14%. Pastún e hindi pertenecen a la familia indoiraniana, el español a la itálica, de modo que la distancia temporal debería ser semejante.*

IELex

- ✿ *Indo-European Lexical Cognacy Database*
- ✿ *Accesible en:*
<https://zenodo.org/record/5556801#.Y0z0EuxlhAc>
- ✿ *Basado en la base de datos de Kruskal-Dylen*
- ✿ *Dirigido por Michael Dunn en el Instituto Max Planck de Psicolingüística de Nimega.*

- ✿ *Contiene:*

Lenguas	163
Significados	225
Palabras	34619
Sets de cognados	5013
Caracteres codificados	32651

IELex

- *Lista Swadesh con 207 conceptos para 157 lenguas indoeuropeas*
- *En parte ortografía y/o transcripción fonética consideradas (parcialmente)*
- *Cada entrada es asignada a una clase de cognado*

Clases de Cognados

- *Concepto “montaña”:*
 - *Clase A: armenio sar, serbio/ruso/polaco gora, checo slowak, ucraniano hora...*
 - *Clase B: alemán Berg, frisón berch, danés bjerg ...*
 - *Clase C: albanés mal*
 - *Clase D: armenio ler*
 - *Clase E: panjabi par, nepali parbat, marathi parwat...*
 - *Clase F: inglés mountain, francés montagne, italiano monte, bretón menez...*
- *Elementos de la misma clase de cognados son cognados entre ellos. Elementos de distinta clase, no son cognados.*
- *La clasificación es realizada por lingüistas, no automática*

Cognados como caracteres filogenéticos

- *A través de cambios lingüísticos puede cambiar la combinación Lengua/Concepto/Clase de cognado*
- *Por ejemplo: ahd. bein (Clase B) al. Bein ingl. bone*
- *pero nhd. Knochen (Clase G)*
- *Cambios comparables a una mutación genética*

Cognados como caracteres filogenéticos

- *Las clases de cognados pasan a ser tratadas como caracteres biológicos.*
- *Binarización:*
 - *Cada clase de cognado es un carácter.*
 - *2 posibles estados 0/1*
 - *0 la lengua no usa un concepto de la clase de cognado para el concepto.*
 - *1 usa un elemento de la clase de cognado*
- *Cambios como ahd ubil → nhd schlecht se corresponden con 2 mutaciones*

Cognados como caracteres filogenéticos

- *La matriz que se crea se representa con un archivo Nexus*
- *.nex o .nxs son los archivos usados en bioinformática.*
- *Analizados con PAUP*, MrBayes, Mesquite etc.*
- *[https://de.wikipedia.org/wiki/Nexus_\(Bioinformatik\)](https://de.wikipedia.org/wiki/Nexus_(Bioinformatik))*

✿ *Archivo Nexus disponible en:*

<https://zenodo.org/record/5556801#.Y0z0EuxlhAc>

✿ *Programa para analizar los datos PAUP**

✿ *Descarga:*

http://people.sc.fsu.edu/~dswofford/paup_test/

Análisis con PAUP*

- *Ejecuta en PAUP* el fichero nexus (es conveniente que estén en el mismo directorio)*

```
> execute lelex_binarizedFull.nex
```

```
> Hsearch
```

```
Do you want to increase 'Maxtrees'? (Y/n): Y
```

```
Enter new value for 'Maxtrees' (100): 10000
```

```
Action if limit is hit:
```

```
(1) Prompt for new value
```

```
(2) Automatically increase by 100 (= AUTOINC)
```

```
(3) Leave unchanged, and don't prompt: 2
```

```
> SaveTrees file='ielexFull_MP.tree' format=Newick brlens=yes
```

```
> q
```

- ✿ *El formato resultante es un archivo Newick*
- ✿ https://en.wikipedia.org/wiki/Newick_format
- ✿ *Podemos representarlo con Dendroscope o Splitstree*
- ✿ <http://ab.inf.uni-tuebingen.de/software/dendroscope/>
- ✿ <http://www.splitstree.org/>

Métodos de análisis filogenéticos

- *MP: Máxima Parsimonia*
- *ML: Máxima Verosimilitud (Maximum Likelihood)*
- *IB: Inferencia Bayesiana*

MP

- *El análisis realizado anteriormente con PAUP* estaba basado en MP.*
- *MP parte de la premisa de que el árbol más sencillo es el que refleja la realidad*
- *Parte del principio de que las mutaciones son poco probables*
- *El árbol resultante implica pocos cambios evolutivos, suele denominarse árbol de consenso y proporciona el resultado más probable entre múltiples árboles generados.*
- *Serían representaciones compactas de gran número de árboles.*

ML

- ✿ *Con esta metodología asumimos que las mutaciones no son tan raras.*

- ✿ *Ejemplo con Paup**

- > execute ielex_binarizedFull.nex
- > set criterion=likelihood
- > set storebrlens
- > Hsearch
- > SaveTrees file='ielexFull_ML.tree'
- format=Newick brlens=yes
- > q

IB

- *Inferencia Bayesiana*

- *Teorema de Bayes: El teorema de Bayes parte de una situación en la que es posible conocer las probabilidades de que ocurran una serie de sucesos a partir de otro.*

- Queremos conocer la probabilidad de desarrollar una enfermedad de hígado en alcohólicos.

- A significa que el paciente está enfermo del hígado (p.ej. El 10% de los pacientes)

- B el paciente es alcohólico (p. ej. El 5%)

- El 7% de los enfermos de hígado son alcohólicos.

- **$P(A|B) = (0.07 * 0.1)/0.05 = 0.14$**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- *Si el paciente es alcohólico, las posibilidades de tener una enfermedad de hígado son del 14%*

- *Similar en resultados a ML*

Paup*

- http://paup.scs.fsu.edu/Command_ref_v2.pdf

ASJP

- *Automated Similarity Judgment Program*
- *S. Wichmann , Instituto Max Planck de Leipzig*
- *Base de datos con 40 términos de la lista Swadesh*
- *Seleccionados los más estables y menos propicios a cambios*
- *Actualmente incluye 6895 lenguas de todos los continentes (lenguas, dialectos, lenguas muertas, reconstruidas)*

ASJP

Körperteile

Auge
Ohr
Nase
Zunge
Zahn
Hand
Knie
Blut
Knochen
Brust (der Frau)
Leber
Haut

Tiere und Pflanzen

Laus
Hund
Fisch
Horn (von Tieren)
Baum
Blatt

Menschen

Mensch
Name

Natur

Sonne
Stern
Wasser
Feuer
Stein
Pfad
Berg
Nacht

Verben und Adjektive

Trinken
Sterben
Sehen
Hören
Kommen
Neu
Voll

Ordnungszahlen und Pronomen

Eins
Zwei
Ich
Du

ASJP

- *Distancia Levenshtein (DL)*
- *DL entre dos cadenas sería la cifra mínima de operaciones de edición que existe entre dos elementos.*
- *Puede haber eliminación, inserción o sustitución*

h	o	r	n	
k	o	r	n	u

ASJP

- Se introduce el concepto de distancia Levenshtein normalizada.
- Se divide la distancia obtenida por la longitud total de la cadena.
- En el ejemplo de horn tendríamos una DL normalizada de 0,4 (2/5)
- El objetivo es realizar una calibración de la distancia analizando los pares de lenguas en una matriz de datos
- Ejemplo del inglés y del sueco

	Ei	yu	wi	w3n	tu	fiS	...
yog	1	$\frac{2}{3}$	1	1	1	1	
du	1	$\frac{1}{2}$	1	1	$\frac{1}{2}$	1	
vi	$\frac{1}{2}$	1	$\frac{1}{2}$	1	1	$\frac{2}{3}$	
et	1	1	1	1	1	1	
tvo	1	1	1	1	$\frac{2}{3}$	1	
fisk	$\frac{3}{4}$	1	$\frac{3}{4}$	1	1	$\frac{1}{2}$	
⋮							

- *La tabla resultante es de 40x40*
- *La diagonal es la DL normalizada*
- *Con el resto de elementos se usa un alineamiento PMI (Pointwise Mutual Information)*
- *Basado en un algoritmo matemático (Needleman-Wunsch)*

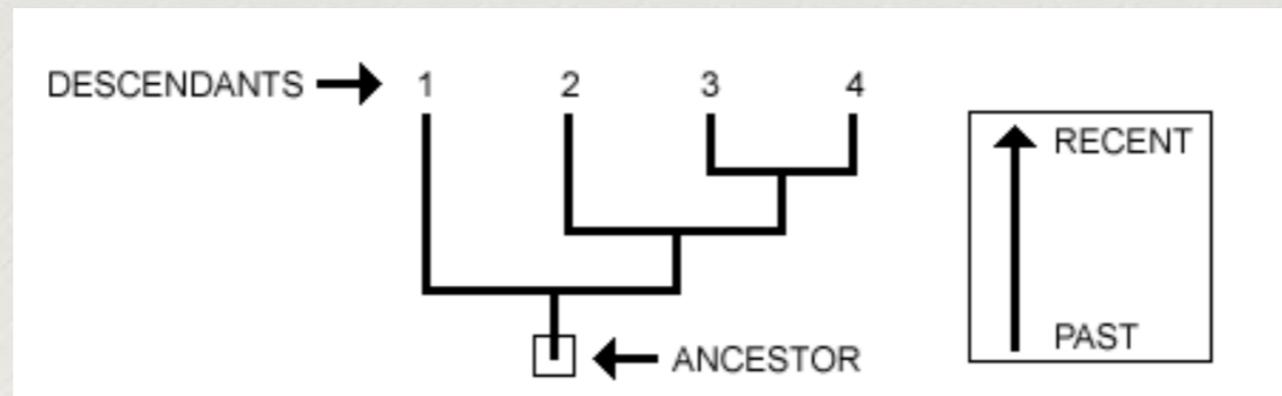
Subfamily: West Germanic

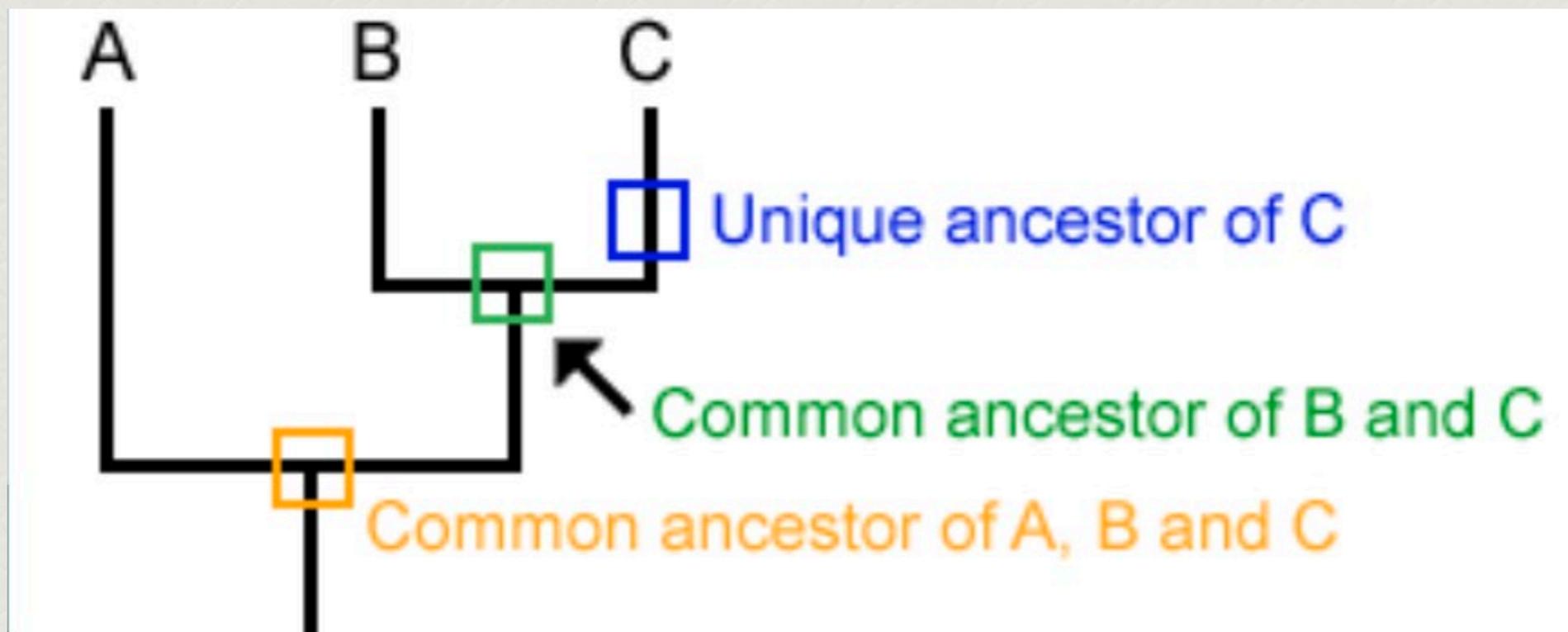
- <http://glottolog.org/resource/languoid/id/west2793>

Introducción a la filogenia de lenguas

LAS FILOGENIAS DESCRIBEN LAS RELACIONES ENTRE ANCESTROS Y DESCENDIENTES DE LOS ORGANISMOS BASÁNDOSE EN LA HOMOLOGÍA

ESTAS RELACIONES EVOLUTIVAS SE REPRESENTAN MEDIANTE DIAGRAMAS LLAMADOS CLADOGRAMAS (DIAGRAMAS DE RAMIFICACIÓN QUE ORGANIZAN LAS RELACIONES)

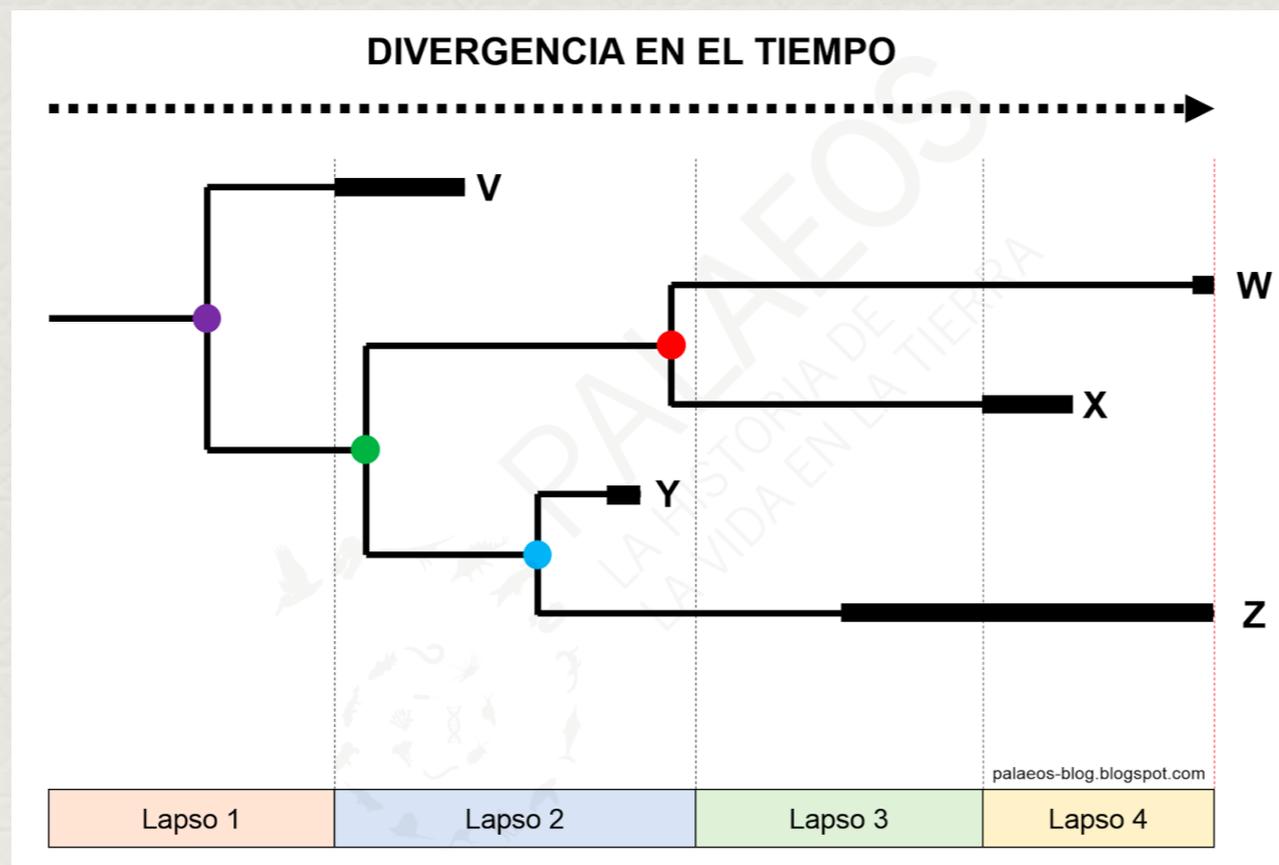




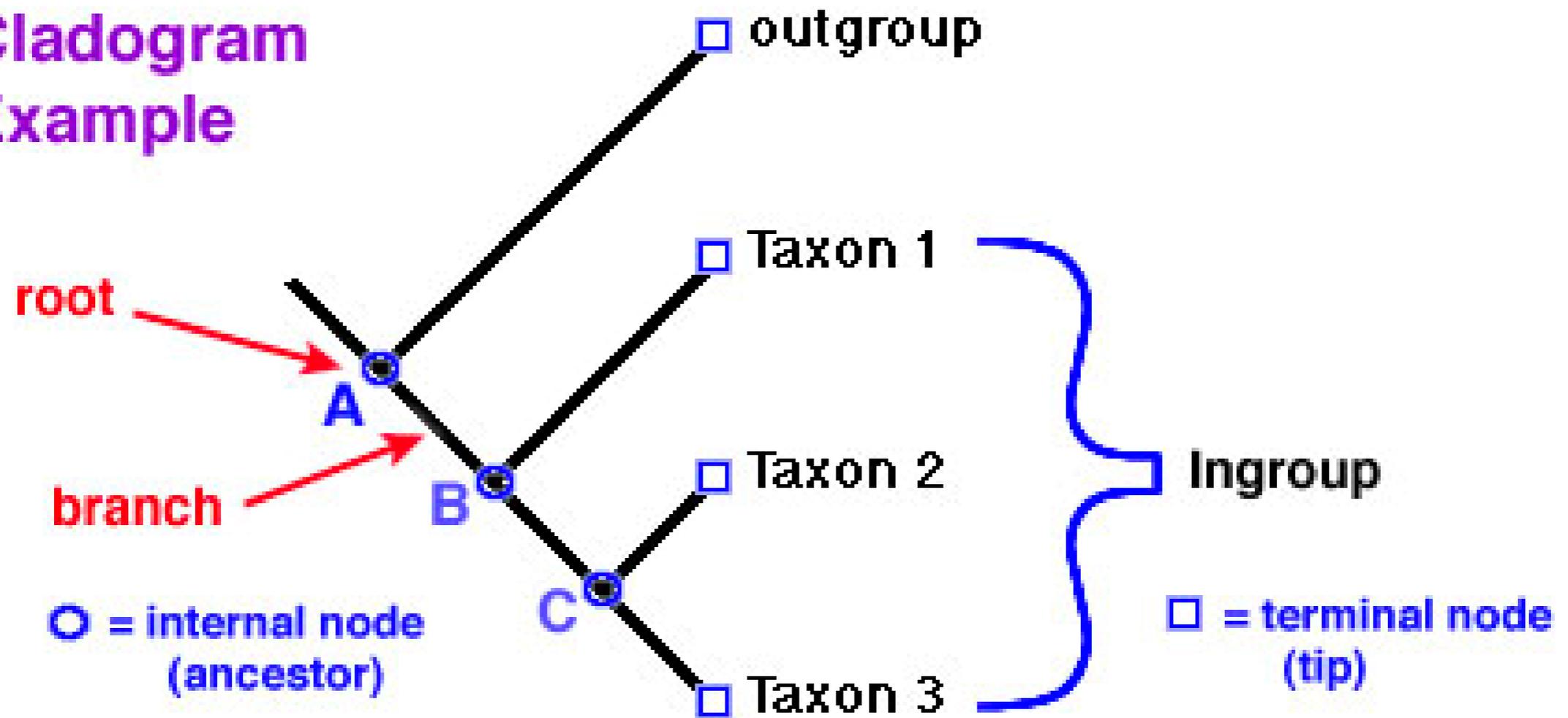
Cuando un linaje ancestral se divide: se indica la división debido a la aparición de algún rasgo nuevo. Cada linaje tiene rasgos únicos para sí mismo y rasgos compartidos con otros linajes. Cada linaje tiene ancestros que son únicos para ese linaje y ancestros que son compartidos con otros linajes - ancestros comunes.



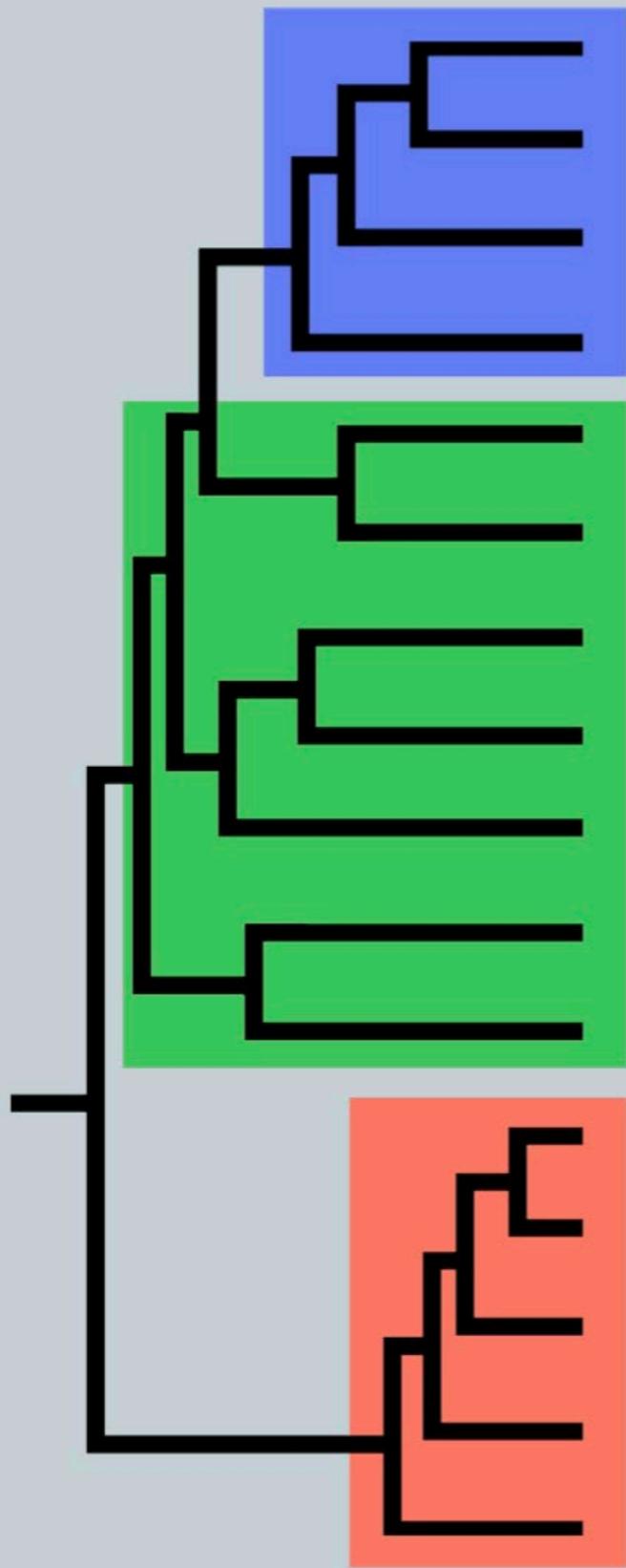
Algunos cladogramas pueden estar calibrados temporalmente: no todos los cladogramas horizontales están calibrados. Si no tiene línea temporal, no está calibrado. Un cladograma calibrado puede tomar cualquier forma. Además, la longitud de las ramas, no siempre indica temporalidad. A veces, la longitud de las ramas indica acumulación de cambios (ramas largas: muchos cambios, ramas cortas: pocos cambios).



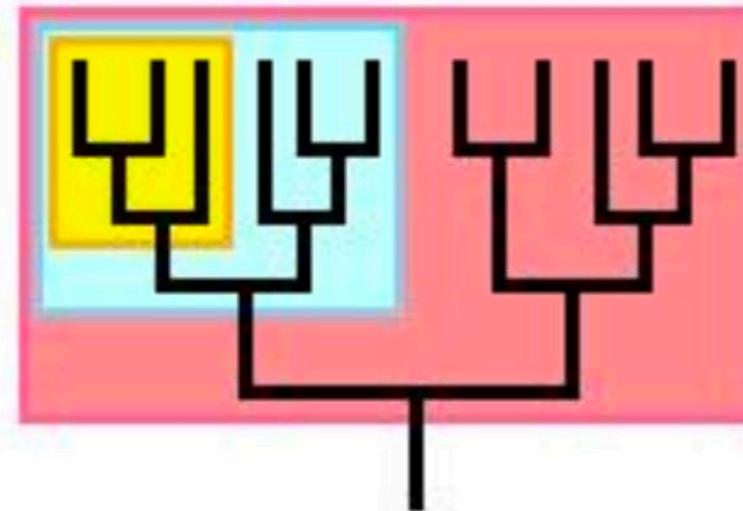
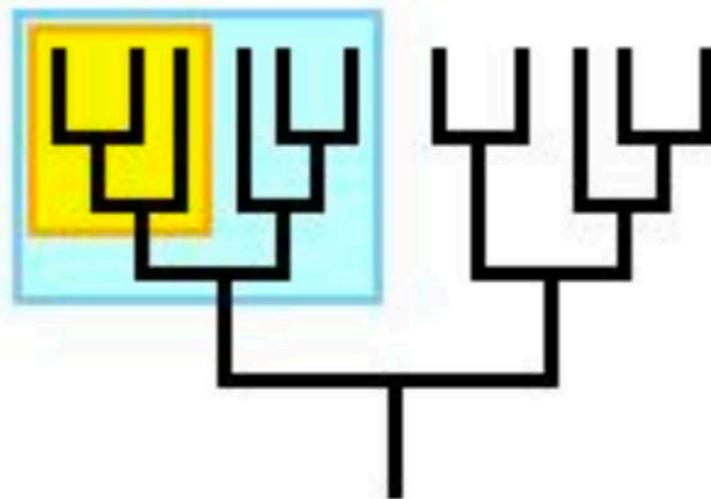
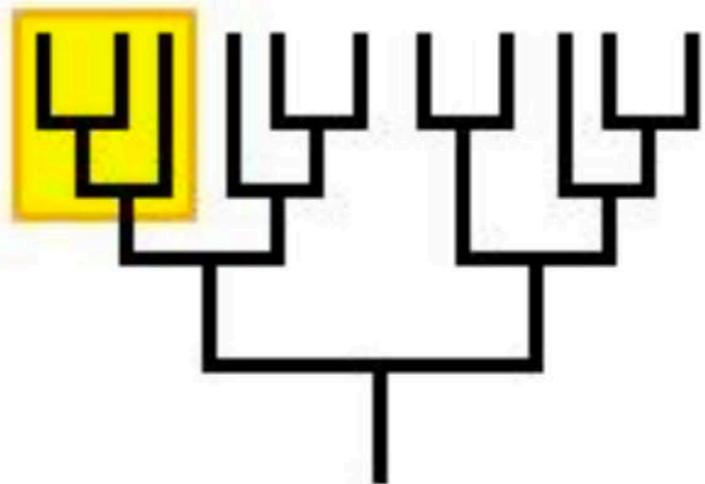
Cladogram Example

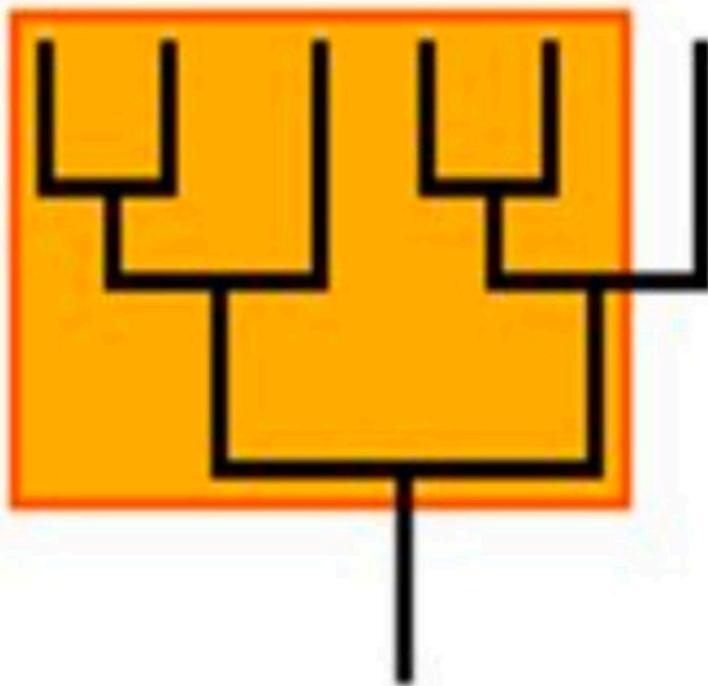
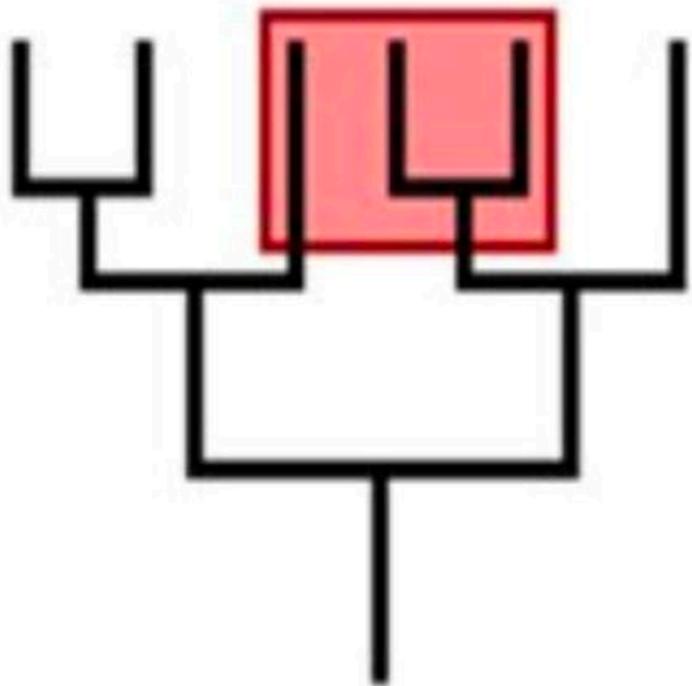
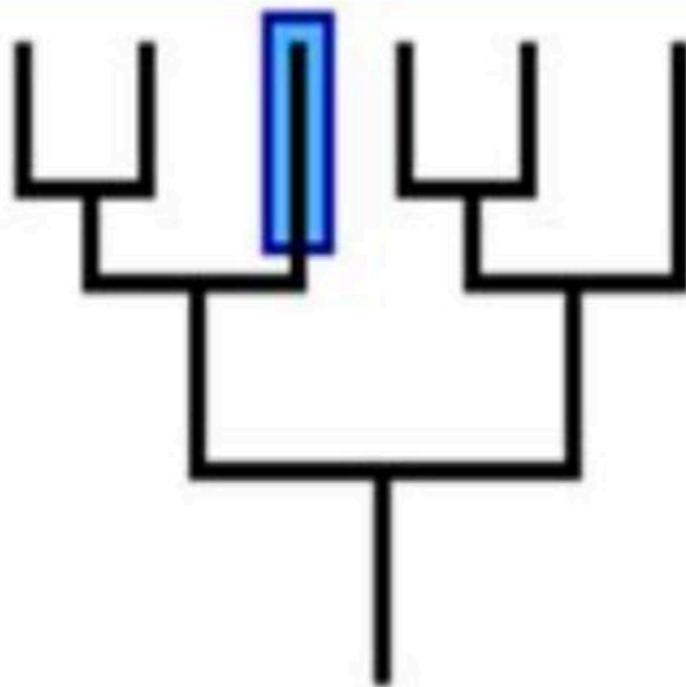
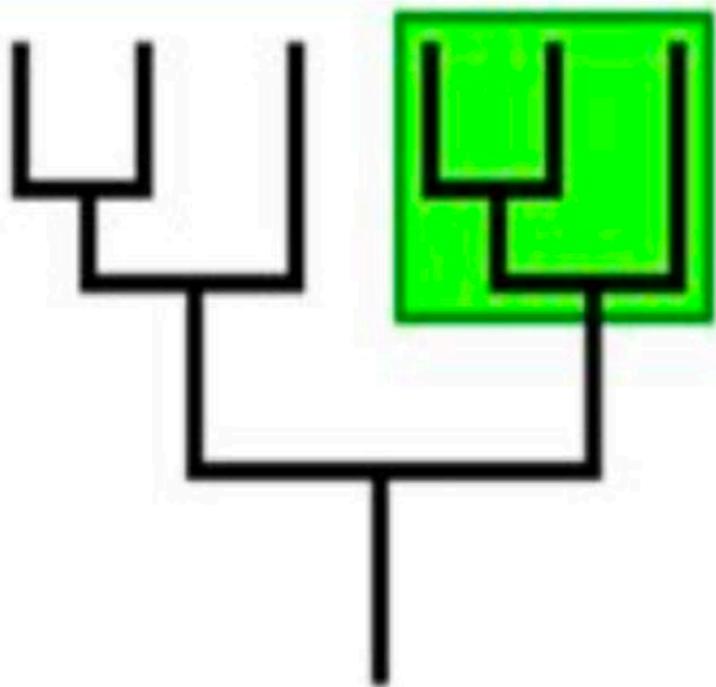


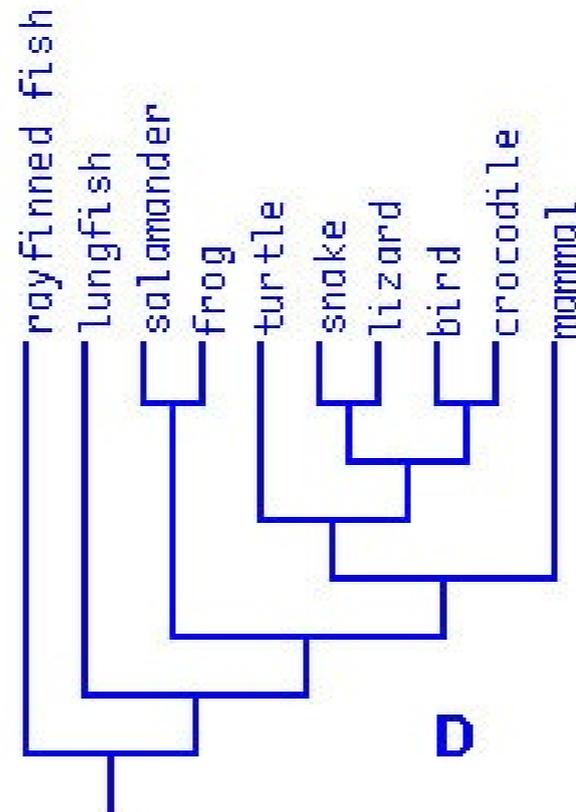
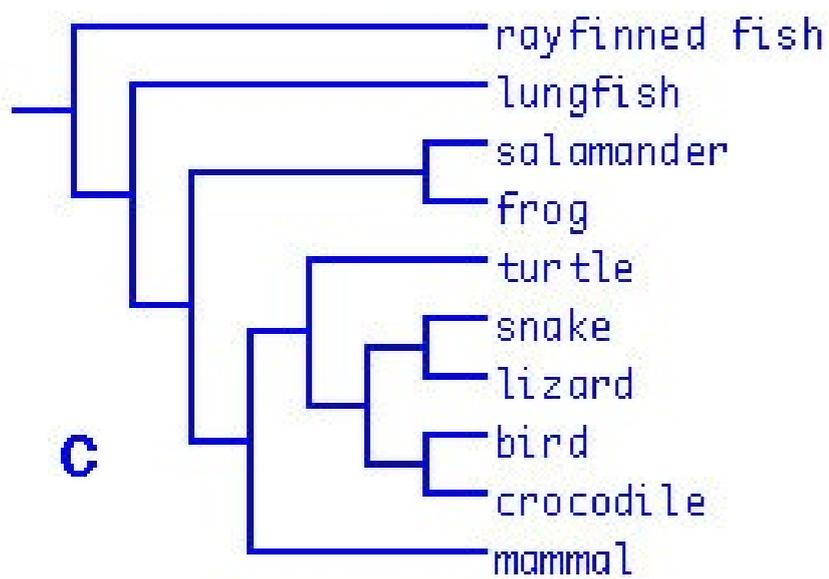
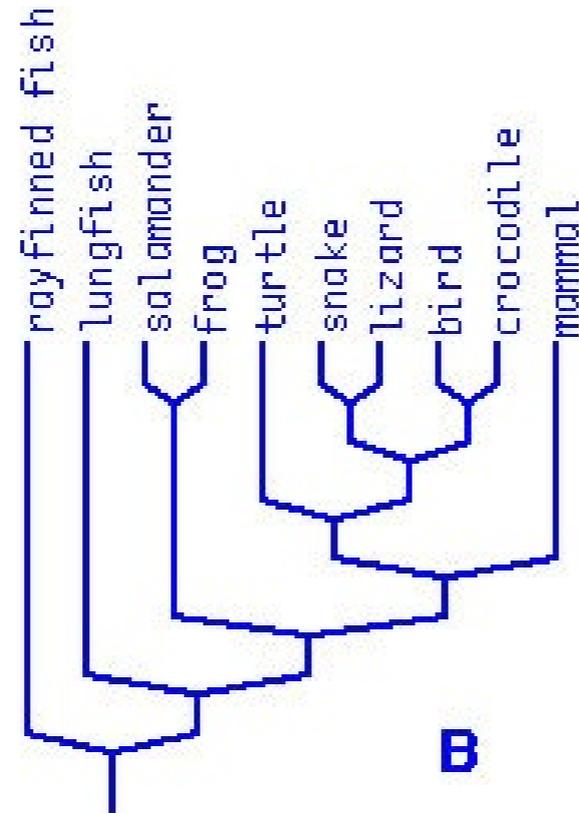
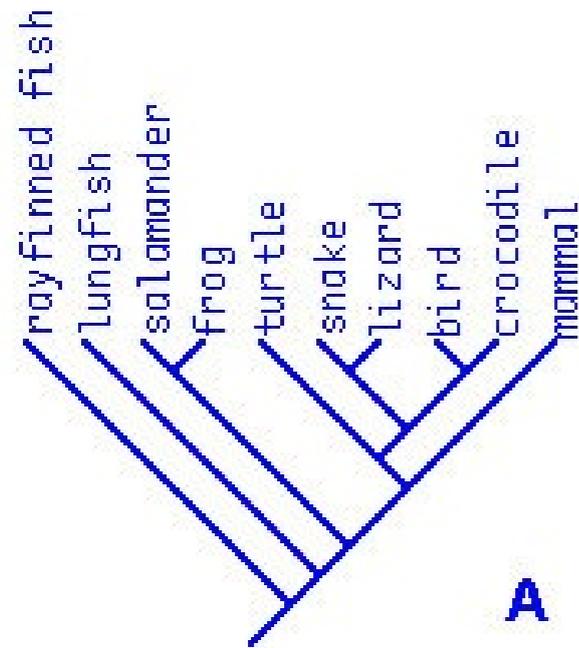
Un **nodo** corresponde a un hipotético ancestro. Un **nodo terminal** es la hipotética última población ancestral común de entrecruzamiento del **taxón** etiquetado en un extremo del cladograma. Un **nodo interno** es la hipotética última población ancestral común que se dividió para dar lugar a dos o más taxones hijos, que por tanto son taxones hermanos entre sí.



Con un cladograma es fácil saber si un grupo de linajes forma un clado. Los clados están anidados unos dentro de otros, forman una jerarquía anidada.







Los cladogramas y las redes son dos formas de representar las relaciones (genealogía) entre taxones. Normalmente llamamos a estos diagramas árboles. Los cladogramas son árboles enraizados.

Las diferencias no significan nada, porque el eje vertical no significa nada. Es decir, no importa la longitud de las ramas. Lo único que importa es qué taxones se unen en los nodos. Del mismo modo, el eje horizontal no significa nada. Puedes dibujar cladogramas tan gordos o tan flacos como quieras y no cambiará el cladograma, siempre que las relaciones entre taxones hermanos sean las mismas.



- *Muñoz-Acebes, J. (2018), De la glotocronología a la filogenética. Estado de la cuestión y los nuevos desarrollos en la metodología de clasificación lingüística, Revista de Investigación Lingüística (21) DOI: <https://doi.org/10.6018/ril.21.367611>*
- *Muñoz-Acebes, J. (2022), Cladística e Historia de la Lengua Alemana: introducción al uso de PAUP*, Revista de Humanidades Digitales (7) DOI: <https://doi.org/10.5944/rhd.vol.7.2022.336>*

Consenso estricto

- *Cuando dos o más hipótesis de cladograma completamente resueltas que compiten entre sí pueden estar igualmente bien apoyadas por la parsimonia. Esa incertidumbre se representa a veces mediante un árbol de consenso estricto.*
- *Específicamente, un consenso estricto conserva sólo los nodos internos comunes a dos o más hipótesis de cladograma en las que se basa.*