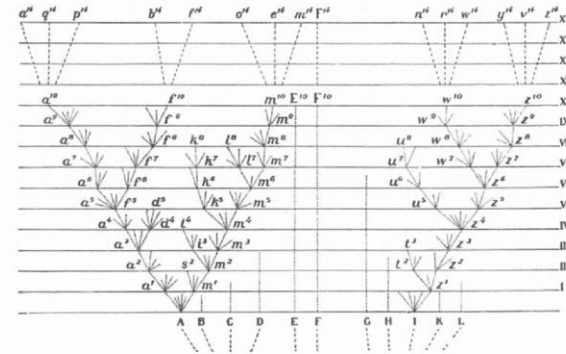
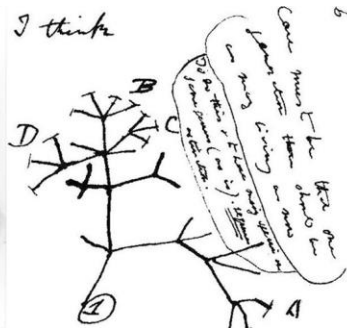
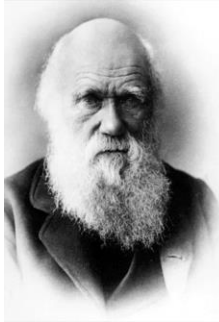


Filogenias Moleculares

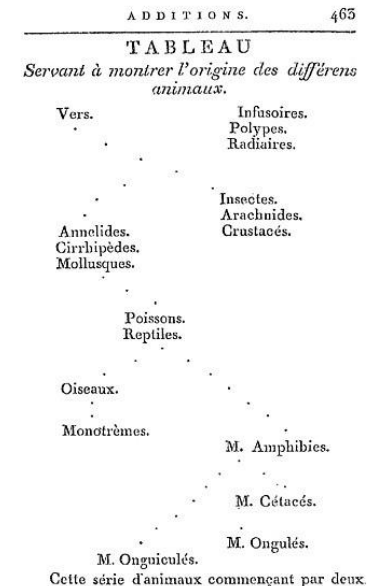
Los primeros árboles filogenéticos

El primer árbol filogenético moderno tal y como los entendemos actualmente fue publicado como única figura en el origen de las especies de Darwin.



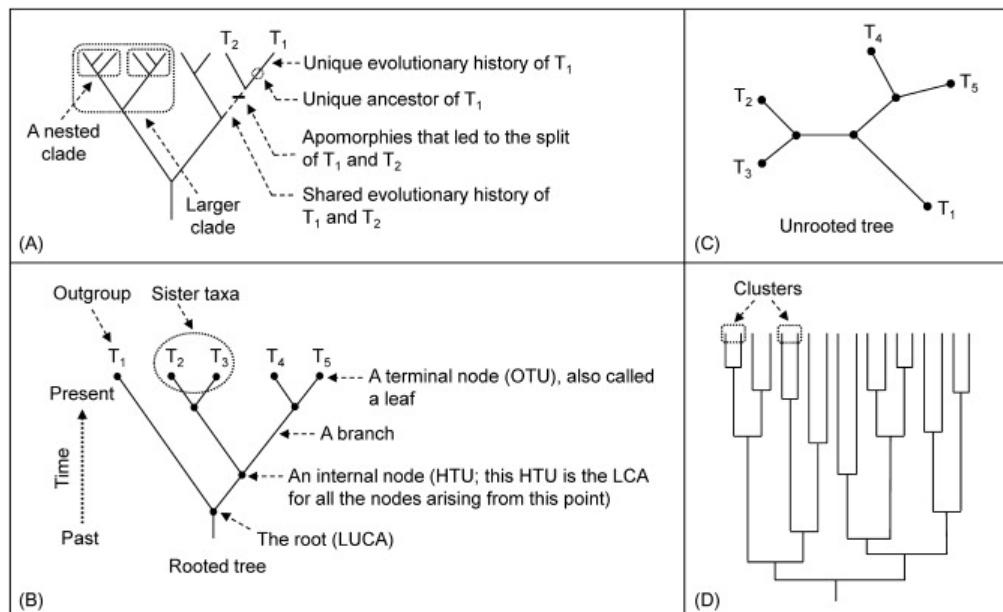
All living things have much in common, in their chemical composition, their germinal vesicles, their cellular structure, and their laws of growth and reproduction. ... Therefore, I should infer ... that probably **all the organic beings which have ever lived on this earth have descended from some primordial form**, into which life was first breathed. On the Origin of Species, (1859, page 484)

Antes que él, Lamarck ya publicó un árbol filogenético, pero con grupos taxonómicos jerarquizados lo que suponía la existencia de transformaciones entre los grupos taxonómicos. Es decir, unos grupos taxonómicos provenían directamente de otros tras ciertas transformaciones.

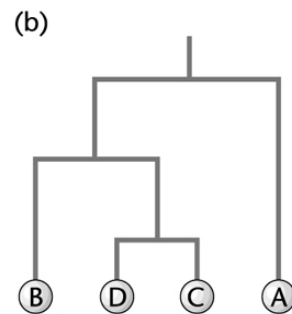
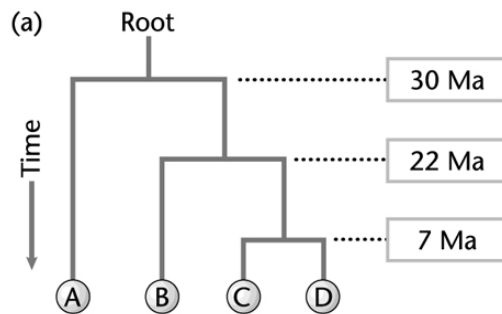


Los árboles filogenéticos

- La filogenética pretende conocer la relación de ancestría y descendencia de los OTUs a diferentes niveles taxonómicos, haciendo una reconstrucción de esta relación con base en diversos caracteres adquiridos por descendencia directa.
- Los **OTUs (Operational taxonomic unit)** son las unidades básicas utilizadas en la taxonomía numérica. Estas unidades pueden hacer referencia a individuos, especies, géneros o clases.
- Los OTUs se representan en lo que conocemos como árboles filogenéticos.

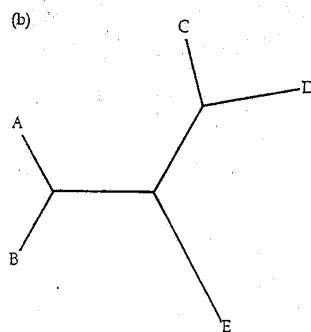
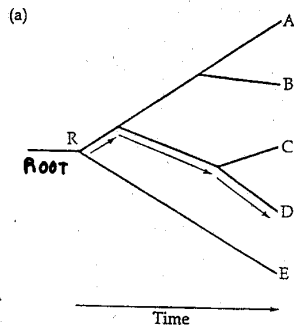
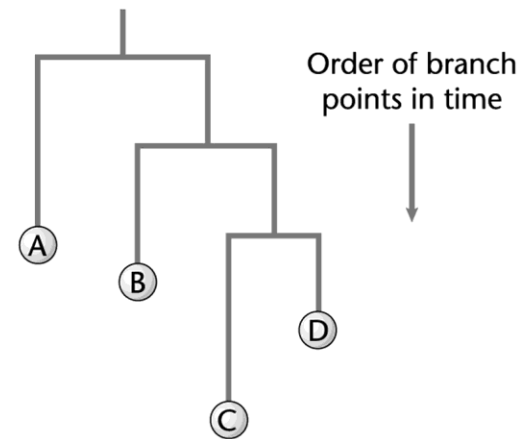


Los árboles filogenéticos



Un árbol filogenético enraizado puede **rotar sus ramas** horizontales y seguir siendo el mismo

Las ramas verticales de un árbol filogenético pueden **escalarsen** en función de la **cantidad de cambios evolutivos**, reflejando esa información en su longitud



A diferencia de un árbol sin raíz, un árbol **enraizado es un árbol polarizado**. Es decir, un árbol en donde conocemos la dirección de los cambios (qué linajes divergieron primero y cuales posteriormente)

Los árboles filogenéticos

How to root a tree

How to root a tree

- Outgroup – place root between distant sequence and rest group
- Midpoint – place root at midpoint of longest path (sum of branches between any two OTUs)

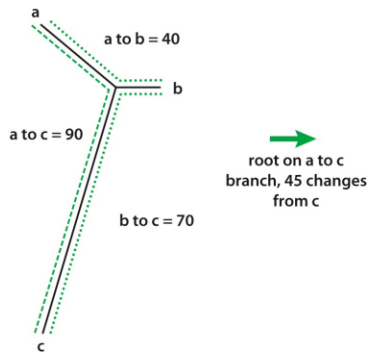
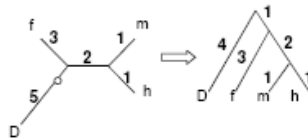
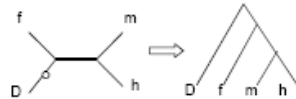
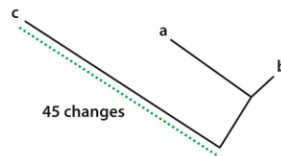


Figure 8.3 Phylogenetics: A Primer (© Garland Science 2013)



Para orientar la **dirección del cambio** es necesario incluir una secuencia distante a las secuencias problemas ("**outgroup**")

La **raíz** de nuestro árbol se ubicará en el **punto medio** de la distancia entre los OTUs más alejados

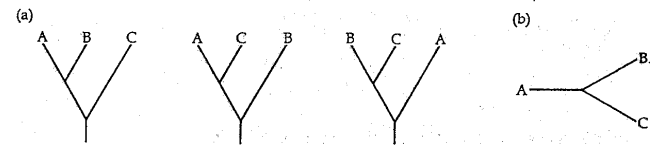


Figure 3. From three OTUs it is possible to construct three different rooted trees (a), but only one unrooted tree (b).

Combinatoric explosion

Combinatoric explosion

Los árboles filogenéticos deben contener **solo bifurcaciones**.

Las trifurcaciones o multifurcaciones se deben generalmente a falta de resolución del método, deben resolverse con más datos

# sequences	# unrooted trees	# rooted trees
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425

Grupos filogenéticos

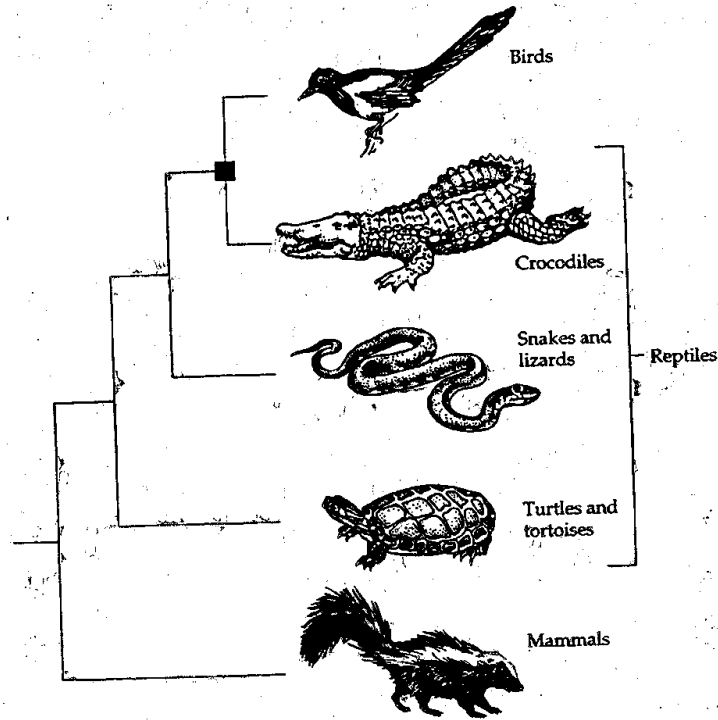


Figure 5.5 Phylogenetic tree of birds, reptiles, and mammals. The reptiles do not constitute a natural clade since they share ancestors with the birds, which are not included in the Reptilia. Birds and crocodiles, on the other hand, constitute a natural clade (Archosauria) since they share a common ancestor (black box) not shared by any other organism. From Li and Graur (1991).

- Un **grupo monofilético** es aquel que comprende una especie ancestral y a todos sus descendientes. Por ejemplo, las aves son un grupo monofilético (incluye a todos los descendientes de la primera ave). Los mamíferos también lo son.
- Un **grupo polifilético** es aquel que comprende una especie ancestral pero no a todos sus descendientes. Un ejemplo son los reptiles, ya que no incluyen a todos los descendientes del primer reptil que existió (las aves quedan fuera del grupo).

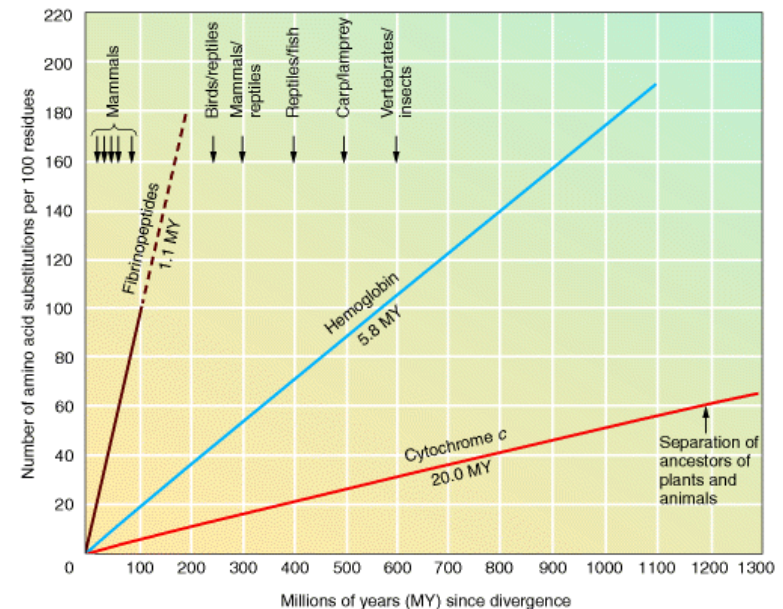
Reloj Molecular

El **reloj molecular** es una técnica para datar la divergencia de dos especies. Estima el tiempo de divergencia entre dos especies a partir del número de diferencias entre dos secuencias de ADN o proteínas.

Zuckerkandl y Pauling (1962): diferencias en los **aminoácidos de las hemoglobinas** entre linajes encajaba con la **divergencia basada en la evidencia fósil**. Se generalizó afirmando que la tasa de cambio evolutivo de cualquier proteína específica era aproximadamente constante a lo largo del tiempo y en diferentes linajes.

Kimura (1968): la **acumulación de mutaciones neutras en el ADN**, debido a errores durante la replicación, también podía usarse para estimar la divergencia evolutiva.

Sin embargo, en posteriores estudios se observó que **diferentes proteínas y genes divergían en su tasa de cambio**. Por lo tanto, en función de las tasas de cambio podemos usar unas secuencias para estudiar las filogenias entre especies más cercanas (**relojes rápidos**) o más lejanas (**relojes lento**).



Reloj Molecular

En la siguiente tabla podemos encontrar las tasas de cambio de sustitución de aminoácidos para diferentes proteínas.

La unidad de **tiempo evolutivo** (T_u) es el tiempo medio requerido para que haya 1 sustitución por cada 100 posiciones.

A partir de la tabla:

$$T_u = 1 \times 10^9 / \text{Rate} \times 100$$

50

CHANGE OF AMINO ACID SEQUENCES

Table 4.4 Rates of amino acid substitutions per amino acid site per 10^9 years ($\lambda \times 10^9$) in various proteins. Modified from Dayhoff (1978).

Protein	Rate	Protein	Rate
Fibrinopeptides	9.0	Thyrotropin beta chain	0.74
Growth hormone	3.7	Parathyrin	0.73
Ig kappa chain C region	3.7	Parvalbumin	0.70
Kappa casein	3.3	Protease inhibitors, BP1 type	0.62
Ig gamma chain C region	3.1	Trypsin	0.59
Lutropin beta chain	3.0	Melanotropin beta	0.56
Ig lambda chain C region	2.7	Alpha crystallin A chain	0.50
Complement C3a anaphylatoxin	2.7	Endorphin	0.48
Lactalbumin	2.7	Cytochrome b ₅	0.45
Epidermal growth factor	2.6	Insulin (exc. guinea pig and coypu)	0.44
Somatotropin	2.5	Calcitonin	0.43
Pancreatic ribonuclease	2.1	Neurophysin 2	0.36
Lipotropin beta	2.1	Plastocyanin	0.35
Haptoglobin alpha chain	2.0	Lactate dehydrogenase	0.34
Serum albumin	1.9	Adenylate kinase	0.32
Phospholipase A ₂	1.9	Triosephosphate isomerase	0.28
Protease inhibitor, PST1 type	1.8	Vasoactive intestinal peptide	0.26
Prolactin	1.7	Corticotropin	0.25
Pancreatic hormone	1.7	Glyceraldehyde 3-PO ₄ dehydrogenase	0.22
Carbonic anhydrase C	1.6	Cytochrome c	0.22
Lutropin alpha chain	1.6	Plant ferredoxin	0.19
Hemoglobin alpha chain	1.2	Collagen (exc. nonrepetitive ends)	0.17
Hemoglobin beta chain	1.2	Troponin C, skeletal muscle	0.15
Lipid-binding protein A-II	1.0	Alpha crystallin B chain	0.15
Gastrin	0.98	Glucagon	0.12
Animal lysozyme	0.98	Glutamate dehydrogenase	0.09
Myoglobin	0.89	Histone H2B	0.09
Amyloid AA	0.8	Histone H2A	0.05
Nerve growth factor	0.85	Histone H3	0.014
Acid proteases	0.84	Ubiquitin	0.010
Myelin basic protein	0.74	Histone H4	0.010

Thus, T_u for fibrinopeptides is 1.1×10^6 years, whereas T_u for histone H4 is 10^9 years.

Reloj Molecular

El **reloj molecular** entonces se basa en la tasa de acumulación de mutaciones neutras. Sin embargo, **¿Qué ocurre con las mutaciones adaptativas?** ¿Cómo influyen en los relojes moleculares?

Por ejemplo, la hemoglobina en cocodrilos ha evolucionado perdiendo su habilidad de responder a la concentración CO_2 , pero ganando la capacidad de disminuir su afinidad al O_2 mediante la unión de iones bicarbonato. Los cocodrilos pueden pasar largos periodos bajo el agua, aumentando la hipoxia y el CO_2 resultante al disolverse en la sangre produce bicarbonato. Por lo tanto, esta adaptación permite liberar eficientemente el O_2 en las células que más lo necesitan y a la vez eliminar de manera eficiente este bicarbonato.

[HBB_SQUAC/7-137](#)
[HBB_HETPO/7-136](#)
[HBB_RANCA/1-135](#)
[HBB1_XENBO/6-140](#)
[HBB_LEPPA/7-142](#)
[HBB1_CYGMA/7-141](#)
[HBB_ALLMI/7-141](#)
[HBBN_AMMLE/2-136](#)
[HBB0_MOUSE/7-141](#)
[HBB2_XENLA/7-141](#)
[HBA_SQUAC/6-136](#)
[HBA1_TORMA/6-136](#)
[HBA_HETPO/13-143](#)
[HBA_LEPPA/7-138](#)
[HBA_LIOMI/6-136](#)
[HBA0_PASMO/6-136](#)

```
LTVAYNHI....D.DVKPHFVELSKK.HYEEL...HVDPENFKLLANCL
LGVAVTHL....G.DVKSQFTDLSKK.HAEEL...HVDVESFKLLAKCF
IEEGLKHE....Z.BLKAYYAKLSEK.HSGEL...HVDPANFYRLGNVL
TGEAIKHM....D.DIKGYAQLSKY.HSETL...HVDPCNFKRFGGCL
IIECTRHE....G.NIKCHLANLSHL.HSEKL...HVDPHNFRVLGQCL
LDRGLKNM....D.NIVDAYAELSTL.HSEKL...HVDPDNFKLLSDCI
FGEAVKHL....D.NIKGHFANLSKL.HCEKF...HVDPENFKLLGDII
FSBGLKHL....B.BLKGAFAFSLSZL.HCBKL...HVBFPZBFRLIGBVL
LGLAVKNM....D.NLKEFHAHLSEL.HCDKL...HADPENFKLLGNML
VGNAISHI....D.SVKSSLQQLSKI.HATEL...FVDPENFKRFGGVL
VADATDHL....D.NVAGHLDLAVL.HGTTT...CVDPHNFPFLTQCI
LVDTNKHL....D.DLPHHLNKLAEK.HGKGL...LVDPHNFKLFSDCI
LAKACDHL....D.DLKTHLHKLATF.HGSEL...KVDPANFYRLSYCL
TGGVQVHL....H.DLSSCLHTLSEK.HAREL...MVDPCNFKQYLIEAT
LDSALKHL....D.DIRGTLSQLSDL.HAYNL...RVDPVNFQQLSKCI
LSNAIKNL....D.NLSQALSELSNL.HAYNL...RVDPVNFKFLSQCL
```

Al comparar el alineamiento de HBB entre el humano y el cocodrilo, sólo la sustitución de **5 aas dirigen este cambio** adaptativo. Mientras que en total se observan **123 sustituciones neutras** entre las globinas de ambas especies.

Las mutaciones adaptativas son **mucho menos frecuentes** y **no presentan una tasa constante de cambio**, por lo que no influyen en la estimación de la distancia evolutiva.

Consideraciones para un análisis filogenético

Selección de las secuencias de partida:

- Ortólogas: ancestro común
- Que estén presentes en todas las especies del grupo de interés
- Ritmo evolutivo adecuado: ni muy alto ni muy bajo

	720	*	740	*	760	*	780	*
Mouse	: ACUAAAGCACUAAGGAGGAUUUAGUAGUAAUUAAGAAUAGAGAGC	UUUAUGAAUUGAGCAAUGAAGUACGCACACAC	: 788					
GuineaPig	: AUUCAAGAUUAAGGAGGAUUUAGUAGUAAUUAAGAAUAGAGAGC	UUAUGAAUAUAGGCCAUUAAGCACGCACACAC	: 789					
T.belanger	: ACUAUGAGUCCAAGGAGGAUUUAGCAGUAAAGUCGGAAUAGAGAGC	UCCGACUGAAUUAAGGCCAUAAAGCACGCACACAC	: 788					
T.tana	: AACUAUGCGUGUAAGGAGGAUUUAGCAGUAAACCCGAAUAGAGAGC	UCCACUGAAUUAAGGCCAUAAAGCACGCACACAC	: 787					
Bushbaby	: ACAAUUGCCAAAGGCGGAUUUAGUAGUAAUUAAGAAUAGAGAGC	UUUAUGAAUUGAGGCCAUUAAGCACGCACACAC	: 790					
Lemur	: ACUAAAAGCCAAAGGAGGAUUUAGCAGUAAAUUAAGAAUAGAGAGC	UUUAUGAAUUGAGGCCAUUAAGCACGCACACAC	: 789					
W.Tarsier	: AAUAAAAGUUAAGGAGGAUUUAGCAGUAAACCAAGAAUAGAGAGC	UUUAUGAAUUAAGGCCAUUAAGCACGCACACAC	: 787					
Ph.Tarsier	: AAUAAAGAAUUAAGGAGGAUUUAGUAGUAAUUAAGAAUAGAGAGC	UUAUGAAUUAAGGCCAUUAAGCACGCACACAC	: 789					
SakiMonk	: ACUAAAAGCCAAAGGUGGAUUUAGCAGUAAAUCCAGAAUAGAGAGC	UUAUGAAUACAGGCCAUUAAGCACGCACACAC	: 788					
HowlerMonk	: AUUAAAGGUC	CAAGGUGGAUUUAGCAGUAAUUAAGAAUAGAGAGC	UUAUGAAUACAGGCCAUUAAGCACGCACACAC	: 790				
SpiderMonk	: AUUAAAGGCCCCAAGGUGGAUUUAGCAGUAAUUAAGAAUAGAGAGC	UUAUGAAUACAGGCCAUUAAGCACGCACACAC	: 790					
Capuchin	: AUUAAAGGCCCCAAGGUGGAUUUAGCAGUAAUUAAGAAUAGAGAGC	UUAUGAAUACAGGCCAUUAAGCACGCACACAC	: 789					
Marmoset	: AUUAAAGGCCCCAAGGUGGAUUUAGCAGUAAACCAAGAAUAGAGAGC	UUAUGAAUACAGGCCAUUAAGCACGCACACAC	: 786					
Tamarin	: AUUAAAGAGUC	CAAGGUGGAUUUAGCAGUAAACCAAGAAUAGAGAGC	UUAUGAAUACAGGCCAUUAAGCGGGCACACAC	: 785				
Proboscis	: ACUAAAGCCCCAAGGAGGAUUUAGUAGUAAUUAAGAAUAGAGAGC	UUUAUGAAUUAAGGCCAUUAAGCACGCACACAC	: 788					
Baboon	: AUUAGGACCCCCAAGGAGGAUUUAGCAGUAAAUUAAGAAUAGAGAGC	UUUAUGAAUUAAGGCCAUUAAGCACGCACACAC	: 786					
RhesusMonk	: AUUAAAGGUC	CAAGGAGGAUUUAGCAGUAAUUAAGAAUAGAGAGC	UUUAUGAAUACAGGCCAUUAAGCGGGCACACAC	: 788				
Gibbon	: ACUGAGGUC	CAAGGAGGAUUUAGCAGUAAUUAAGAAUAGAGAGC	UUUAUGAAUACAGGCCAUUAAGCGGGCACACAC	: 790				
Orangutan	: AUUGAAGGUC	CAAGGUGGAUUUAGCAGUAAAUUAAGAGUAGAGAGC	UUUAUGAAUUAAGGCCAUUAAGCGGGCACACAC	: 789				
Gorilla	: ACUAAAGGUC	CAAGGUGGAUUUAGCAGUAAUUAAGAGUAGAGAGC	UUUAUGAAUUAAGGCCAUUAAGCGGGCACACAC	: 789				
Human	: ACUAAAGGUC	CAAGGUGGAUUUAGCAGUAAUUAAGAGUAGAGAGC	UUUAUGAAUUAAGGCCAUUAAGCGGGCACACAC	: 790				
Chimpanzee	: ACUAAAGGUC	CAAGGUGGAUUUAGCAGUAAUUAAGAGUAGAGAGC	UUUAUGAAUUAAGGCCAUUAAGCGGGCACACAC	: 789				
PygmyChimp	: ACUAAAGGUC	CAAGGUGGAUUUAGCAGUAAUUAAGAGUAGAGAGC	UUUAUGAAUUAAGGCCAUUAAGCGGGCACACAC	: 786				
	A uaa q c AAGG	GGAUUUAGcAGUAAa	aaqAaUAGAg	Gcuu	uUGAa	qGCC	U AAGc cG ACACAC	

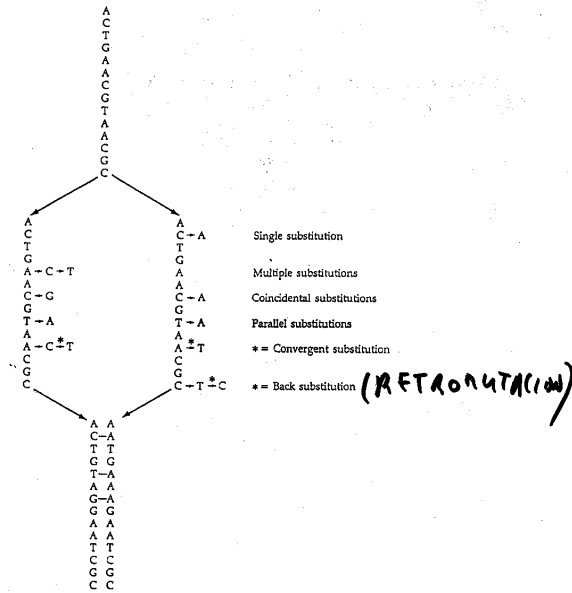
A uaa g c AAGG GGAUUUAGCAGUAAa aagAaUAGAG Gcuu uUGAa gGcC U AAGc cG ACACAC

Fracción del alineamiento de la subunidad pequeña del ARNr entre primates, musarañas y roedores

Los huecos del alineamiento no se consideran en el cómputo de la distancia evolutiva

Cálculo de distancias evolutivas

A la hora de transformar similitudes de secuencias a distancias evolutivas, el número de mutaciones encontradas no necesariamente refleja el número de cambios reales que han ocurrido en la secuencia.



En el ejemplo, en realidad se han producido 12 mutaciones pero comparando ambas secuencias sólo somos capaces de identificar 3.

Las mutaciones simples podemos cuantificarlas bien. Sin embargo, **mutaciones múltiples, coincidentes, paralelas, convergentes y retromutaciones** hacen que infraestimemos el número de cambios que se han producido.

En principio, la **distancia evolutiva cruda** (p) se calcula como el número de posiciones diferentes divididas por la longitud de emparejamientos (descartando los huecos).

$$p = n_d / n$$
$$q = n_i / n$$
$$p+q=1$$

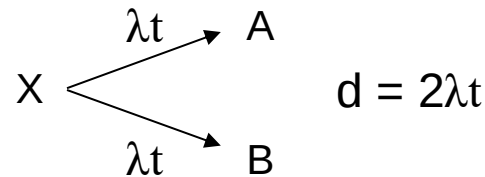
Si existen pocas mutaciones en la secuencia, la distancia evolutiva cruda (p) será muy similar al número real de mutaciones que han ocurrido. Pero cuando se observan muchas mutaciones es posible que infraestimemos la tasa de cambio real.

n_d = posiciones diferentes
 n_i = posiciones idénticas
 n = emparejamientos

Cálculo de distancias evolutivas

Por lo tanto, se debe buscar un **modelo matemático para ajustar la tasa de cambio**, de tal manera que secuencias con numerosos cambios pueda ser estimada correctamente. Para ello sería necesario una distribución que modele eventos raros (Distribución de Poisson).

λ = Tasa de sustitución nts/aas por año y sitio
 t = Tiempo en años



La probabilidad de que en la secuencia ancestral X ocurran r sustituciones en un sitio en t años viene dada por la distribución de Poisson:

$$P_r(t) = e^{-\lambda t} * (\lambda t)^r / r!$$

La probabilidad de que durante t años no haya ocurrido ninguna sustitución, es decir, si $r = 0$:

$$P_0(t) = e^{-\lambda t} * (\lambda t)^0 / 0! = e^{-\lambda t}$$

Así pues, la probabilidad (q) de que no haya ocurrido ninguna sustitución en ninguna de las dos secuencias (sucesos independientes) será:

$$q = e^{-\lambda t} * e^{-\lambda t} = e^{-2\lambda t}$$

Aplicando logaritmos y dado que $d = 2\lambda t$:

$$\ln(q) = -2\lambda t \quad d = -\ln(q)$$

Distancia Evolutiva (PC)
Corrección de Poisson

Cálculo de distancias evolutivas

Mediante la corrección de las distancias evolutivas mediante la distribución de Poisson se consigue **evitar la infraestimación de distancias evolutivas en secuencias con muchas mutaciones.**

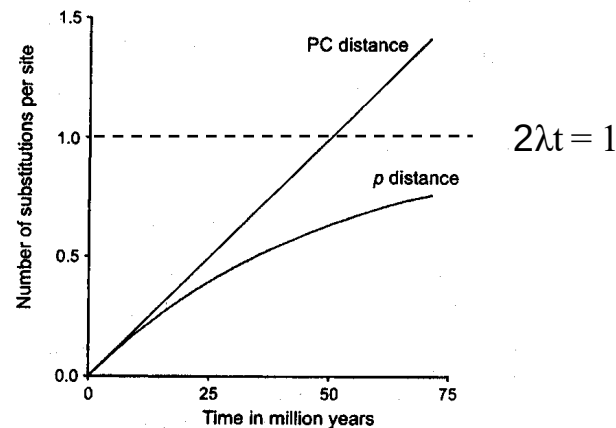
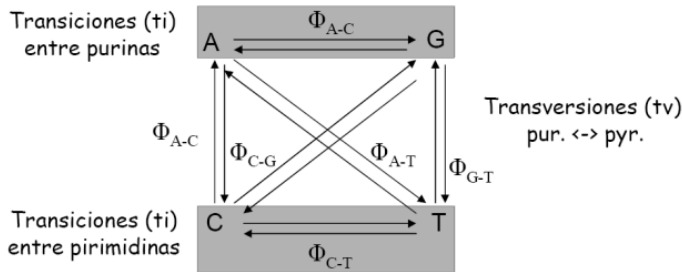


FIGURE 2.2. Relationships of the p distance and the Poisson correction (PC) distance with time. The rate of amino acid substitution (r) is assumed to be 10^{-8} per site per year.

Sin embargo, debemos seguir teniendo en cuenta la **longitud de las secuencias**. Ya que al ser limitadas si la tasa de cambios en la secuencia de estudio es muy rápida el número de sustituciones puede ser igual a la longitud de la secuencia, no siendo posible estimar correctamente la distancia evolutiva.

Modelos de evolución de secuencias de ADN



Gene Structure, Genetic Codes, and Mutation

TABLE 1.5 Pattern of nucleotide substitution in pseudogenes^a

From	To				Row totals
	A	T	C	G	
A	—	4.7±1.3 (5.3±1.4)	5.0±0.7 (5.6±0.8)	9.4±1.3 (10.3±1.4)	19.1 (21.2)
T	4.4±1.1 (4.8±1.1)	—	8.2±1.3 (9.2±1.3)	3.2±1.2 (3.6±1.3)	15.9 (17.6)
C	6.5±1.1 (7.1±1.3)	21.0±2.1 (18.2±2.3)	—	4.2±0.5 (4.2±0.6)	31.7 (29.5)
G	20.7±2.2 (18.6±1.9)	7.2±1.1 (7.7±1.3)	5.3±1.0 (5.5±1.3)	—	33.2 (31.8)
Column totals	31.6 (30.5)	32.9 (31.2)	18.5 (20.3)	16.9 (18.1)	

From Gojobori et al. (1982b) and Li et al. (1984).

^aTable entries are the inferred percentages (f_{ij}) of base changes from i to j based on 13 mammalian pseudogene sequences. Values in parentheses were obtained by excluding all CG dinucleotides from comparison.

En la conversión de la similitud de secuencia de ADN a distancia evolutiva debería tenerse en cuenta tanto la **proporción de nucleótidos** en la secuencia como la **tasa de sustitución esperada** entre los diferentes nucleótidos.

En secuencias no sometidas a selección como los pseudogenes, las **transiciones** ocurren casi dos veces más que lo esperado por azar (~60% obs. vs 33% esp.)

Las **mutaciones C>T / G>A** son las más frecuentes (>40% vs 16%), nótese la diferencia cuando no se incluyen los dinucleótidos CpG (entre paréntesis).

Los **cambios hacia CG > AT** son más frecuentes (~65% vs 50%), lo que explica el mayor % de AT en el genoma.

Las tasas de sustitución en genes funcionales cambian, **C>T menos frecuente** y las **mutaciones C>T / G>A** bajan casi a lo esperado por azar (52.4%). Los genes son ricos en GC, a diferencia del resto del genoma.

Table 5.5 Relative substitution rates among the four nucleotides A, T, C, and G in functional globin and ACTH genes. From Gojobori et al. (1982a).

Mutant nucleotide	Original nucleotide				
	A	T	C	G	
A	—	5.0±1.2	8.1±0.6	20.9±3.1	34
T	4.2±2.0	—	9.4±3.3	4.8±0.8	18.4
C	6.3±2.8	3.7±1.6	—	11.5±3.8	21.5
G	11.5±1.4	1.7±1.4	13.0±1.5	—	26.2
	22	10.4	30.5	37.2	

Modelos de evolución de secuencias de ADN

Jukes-Cantor: frecuencias y tasas de sustitución iguales.

Kimura 2 parámetros: frecuencias iguales y tasas diferentes para t_v y t_i .

Felsenstein: diferentes frecuencias y tasas iguales.

Tamura-Nei: diferentes frecuencias y 3 tasas de sustitución.

Modelo general tiempo reversible: diferentes frecuencias y 6 tasas de sustitución.

Modelo general estocástico: diferentes frecuencias y 12 tasas de sustitución.

Table 3.2 Models of nucleotide substitution.

	A	T	C	G		A	T	C	G
(A) Jukes-Cantor model					(E) HKY model				
A	-	α	α	α	-	βg_T	βg_C	αg_G	
T	α	-	α	α	βg_A	-	αg_C	βg_G	
C	α	α	-	α	βg_A	αg_T	-	βg_G	
G	α	α	α	-	αg_A	βg_T	βg_C	-	
(B) Kimura model					(F) Tamura-Nei model				
A	-	β	β	α	-	βg_T	βg_C	$\alpha_1 g_G$	
T	β	-	α	β	βg_A	-	$\alpha_2 g_C$	βg_G	
C	β	α	-	β	βg_A	$\alpha_2 g_T$	-	βg_G	
G	α	β	β	-	$\alpha_1 g_A$	βg_T	βg_C	-	
(C) Equal-input model					(G) General reversible model				
A	-	αg_T	αg_C	αg_G	-	ag_T	bg_C	cg_G	
T	αg_A	-	αg_C	αg_G	ag_A	-	dg_C	eg_G	
C	αg_A	αg_T	-	αg_G	bg_A	dg_T	-	fg_G	
G	αg_A	αg_T	αg_C	-	cg_A	eg_T	fg_C	-	
(D) Tamura model					(H) Unrestricted model				
A	-	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$	-	a_{12}	a_{13}	a_{14}	
T	$\beta\theta_2$	-	$\alpha\theta_1$	$\beta\theta_1$	a_{21}	-	a_{23}	a_{24}	
C	$\beta\theta_2$	$\alpha\theta_2$	-	$\beta\theta_1$	a_{31}	a_{32}	-	a_{34}	
G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	-	a_{41}	a_{42}	a_{43}	-	

Note: An element (a_{ij}) of the above substitution matrices stands for the substitution rate from the nucleotide in the i -th row to the nucleotide in the j -th column. g_A , g_T , g_C , and g_G are the nucleotide frequencies. $\theta_1 = g_C + g_G$; $\theta_2 = g_A + g_T$.

Los modelos más complejos no tienen que ser mejores, **más parámetros necesitan más valores** para una buena estimación y las secuencias son finitas. Además, a más parámetros las fórmulas son más complejas y a veces inaplicables.

La importancia del modelo depende del grado de divergencia

Valores bajos (Pos. 2), poco divergentes la corrección no es importante (8.5 a 9.3)

Valores altos (Pos. 3), muy divergentes la corrección es muy importante (36.8 a 87.9)

Table 3.4 Estimates ($\hat{d}$) of the number of nucleotide substitutions per site between the human and Rhesus monkey mitochondrial cytochrome *b* genes for the first, second, and third codon positions ($\hat{d} \times 100$).

Position in Codon	p	Jukes-Cantor	Kimura	Tajima-Nei	Tamura-Nei
First	15.5 $\pm$ 1.9	17.3 $\pm$ 2.4	17.8 $\pm$ 2.5	18.0 $\pm$ 2.6	17.9 $\pm$ 2.5
Second	8.5 $\pm$ 1.4	9.1 $\pm$ 1.6	9.2 $\pm$ 1.7	9.2 $\pm$ 1.7	9.3 $\pm$ 1.7
Third	36.8 $\pm$ 2.5	50.6 $\pm$ 4.9	52.3 $\pm$ 5.4	66.5 $\pm$ 9.4	87.9 $\pm$ 39.0

Modelos de evolución de secuencias de ADN

El tipo de sustituciones varían enormemente entre diferentes genes, siendo siempre más propenso a ocurrir mutaciones sinónimas (no modifican el aminoácido).

Table 1. Rates of synonymous and nonsynonymous substitutions in various mammalian protein-coding genes.^a

Gene	L ^b	Nonsynonymous rate (× 10 ³)	Synonymous rate (× 10 ³)
HISTONES			
Histone 3	135	0.00 ± 0.00	6.38 ± 1.19
Histone 4	101	0.00 ± 0.00	6.12 ± 1.32
CONTRACTILE SYSTEM PROTEINS			
Actin α	376	0.01 ± 0.01	3.68 ± 0.43
Actin β	349	0.03 ± 0.02	3.13 ± 0.39
HORMONES, NEUROPEPTIDES, AND OTHER ACTIVE PEPTIDES			
Somatostatin-28	28	0.00 ± 0.00	3.97 ± 2.66
Insulin	51	0.13 ± 0.13	4.02 ± 2.29
Thyrotropin	118	0.33 ± 0.08	4.66 ± 1.12
Insulin-like growth factor II	179	0.52 ± 0.09	2.32 ± 0.40
Erythropoietin	191	0.72 ± 0.11	4.34 ± 0.65
Insulin C-peptide	35	0.91 ± 0.30	6.77 ± 3.49
Parathyroid hormone	90	0.94 ± 0.18	4.18 ± 0.98
Luteinizing hormone	141	1.02 ± 0.16	3.29 ± 0.60
Growth hormone	189	1.23 ± 0.15	4.95 ± 0.77
Urokinase-plasminogen activator	435	1.28 ± 0.10	3.92 ± 0.44
Interleukin I	265	1.42 ± 0.14	4.60 ± 0.65
Relaxin	54	2.51 ± 0.37	7.49 ± 6.10
HEMOGLOBINS AND MYOGLOBIN			
α-globin	141	0.55 ± 0.11	5.14 ± 0.90
Myoglobin	153	0.56 ± 0.10	4.44 ± 0.82
β-globin	144	0.80 ± 0.13	3.05 ± 0.56

(Continued on next page)

Gene	L ^b	Nonsynonymous rate (× 10 ³)	Synonymous rate (× 10 ³)
APOLIPOPROTEINS			
E	283	0.98 ± 0.10	4.04 ± 0.53
A-I	243	1.57 ± 0.16	4.47 ± 0.66
A-IV	371	1.58 ± 0.12	4.15 ± 0.47
IMMUNOGLOBULINS			
Ig V _H	100	1.07 ± 0.19	5.66 ± 1.36
Ig γ1	321	1.46 ± 0.13	5.11 ± 0.64
Ig κ	106	1.87 ± 0.26	5.90 ± 1.27
INTERFERONS			
α1	166	1.41 ± 0.13	3.53 ± 0.61
β1	159	2.21 ± 0.24	5.88 ± 1.08
γ	136	2.79 ± 0.31	8.59 ± 2.56
OTHER PROTEINS			
Aldolase A	363	0.07 ± 0.03	3.59 ± 0.51
Hydroxanthine phosphoribosyltransferase	217	0.13 ± 0.04	2.13 ± 0.35
Creatine kinase M	380	0.15 ± 0.03	3.08 ± 0.37
Glyceradehyde-3-phosphate dehydrogenase	331	0.20 ± 0.05	2.84 ± 0.37
Lactate dehydrogenase A	331	0.20 ± 0.04	5.03 ± 0.61
Acetylcholine receptor γ subunit	540	0.29 ± 0.04	3.23 ± 0.31
Fibrinogen γ	411	0.55 ± 0.06	5.82 ± 0.67
Albumin	590	0.91 ± 0.07	6.63 ± 0.61
Average ^c		0.85 (0.73)	4.61 (1.44)

^a All rates are based on comparisons between human and rodent genes and the time of divergence was set at 80 million years ago. Rates are in units of substitutions per site per 10³ years.

^b L = number of codons compared.

^c Average is the arithmetic mean, and values in parentheses are the standard deviations, computed over all genes.

Modelos de evolución de secuencias de ADN

Y por supuesto varía en función de la localización génica.

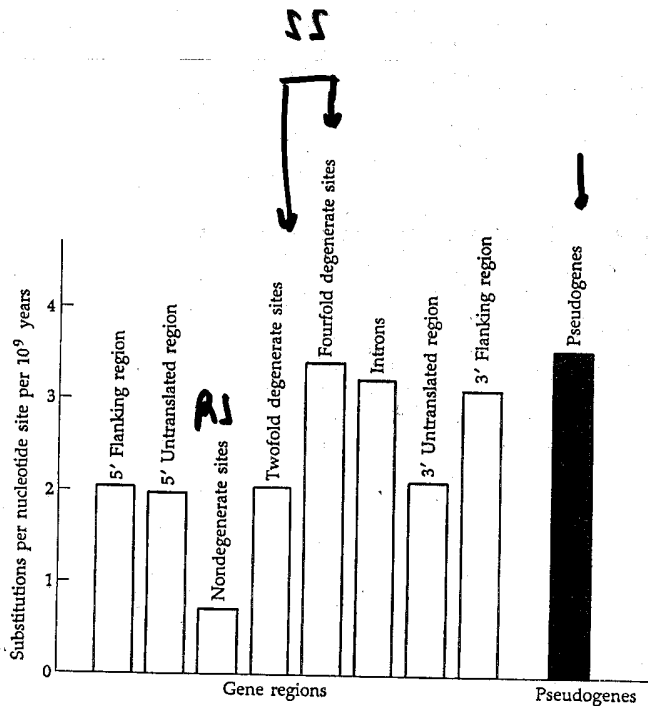


Figure 2. Average rates of substitution in different parts of genes and in pseudogenes.

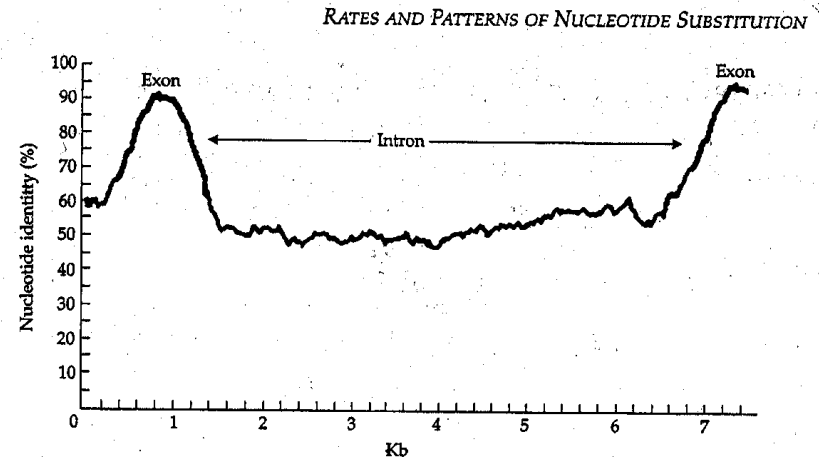


FIGURE 4.4 A similarity profile for two aligned DNA sequences. Conserved regions show up as peaks, which in protein-coding genes may allow us to identify the junctions between exons and introns.

Modelos de evolución de secuencias de proteínas

Al igual que ocurre con las secuencias de nucleótidos, la tasa de sustitución de aminoácidos también es diferente en función de la posición:

- La tasa de sustitución suele ser más **alta** en los sitios **menos importantes desde el punto de vista funcional**.
- Esta **variación en diferentes posiciones sigue una distribución gamma**, lo que permite a su vez corregir la distancia PC.

y el tipo de aminoácidos:

- Experimentalmente se observó que existe un **mayor número de sustituciones entre aminoácidos bioquímicamente similares** que entre aminoácidos de distintos grupos bioquímicos.
- Algunos aminoácidos como: **cisteína, glicina o triptófano cambian muy raramente**, mientras que otros lo hacen más a menudo. Esto también puede alterar la estima de la distancia.
- Margaret Dayhoff (1978) propuso por tanto otro método para calcular la distancia, basado en las llamadas matrices **PAM (Point Accepted Mutations)**.

Modelos de evolución de secuencias de proteínas

Las matrices PAM (Point Accepted Mutations) representan en cada celda la probabilidad de la sustitución de un aminoácido por otro. La matriz se calcula **observando diferencias en proteínas muy cercanas evolutivamente** (con, al menos, un 85% de similitud), las sustituciones en cuestión **no tienen efecto sobre la función de la proteína**, por lo que se trata de mutaciones aceptadas (de ahí su nombre) en el proceso evolutivo.

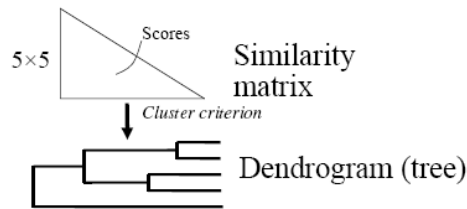
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg	R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn	N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp	D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys	C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln	Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu	E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly	G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His	H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile	I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu	L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys	K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met	M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe	F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro	P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser	S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr	T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp	W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr	Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val	V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Table 1: Margaret Dayhoff's 1978 mutation data matrix for 1 PAM. Frequencies have been multiplied by 10000. For a given column, the rows give the probabilities of change, so that each column adds up to 1.

El número de secuencias disponibles en aquellos años era limitado por lo que trabajos más recientes han **recalculado y proporcionado mejores estimaciones** en las matrices **BLOSUM**. Estas matrices se denominan seguidas de un número que **representa el % de similitud** entre las secuencias a partir de las que se estimaron las probabilidades de sustitución. BLOSUM80 (80% de identidad entre las secuencias) se usa para alineamientos menos divergentes, mientras que BLOSUM 45 se usa para alineamientos más divergentes. La más utilizada es la BLOSUM62.

Cálculo de la similitud de secuencias

Slide 19 Cluster analysis – Clustering criteria



Four different clustering criteria:

Single linkage - Nearest neighbour

Complete linkage – Furthest neighbour

Group averaging – UPGMA

Neighbour joining (global measure)

Note: these are all *agglomerative* cluster techniques; i.e. they proceed by merging clusters as opposed to techniques that are *divisive* and proceed by cutting clusters

- Una vez obtenidos los alineamientos múltiples entre las secuencias seleccionadas se calcula una matriz de similitud entre todos los pares.
 - Y en base a esta matriz de distancias se construye el dendrograma. Para ello se pueden utilizar diferentes **métodos de agrupamiento**.
-
- **UPGMA:** identificación de pares similares y cálculo de la media de las distancias entre ellas y el resto. Válido si se cumple la hipótesis del reloj molecular, donde todas las secuencias (ramas) evolucionan a la misma velocidad.
 - **Neighbor-joining:** secuencialmente busca vecinos que minimizan la longitud del árbol (evolución mínima). Ramas pueden tener diferentes longitudes (no necesario reloj molecular).
 - **Máxima parsimonia:** minimizan el número de cambios entre ramas (explicación más simple). Evalúa múltiples árboles, no siempre hay una única solución.
 - **Máxima verosimilitud:** busca el árbol más probable que haya generado los datos observados en base a un modelo de mutación *a priori*. Tienen una gran precisión, pero son computacionalmente demandantes.
 - **Métodos bayesianos:** calcula una probabilidad a posteriori para cada árbol posible dado un modelo de evolución y probabilidades a priori. Similar a los de máxima verosimilitud.

Estabilidad del árbol filogenético

Método de bootstrapping

Se barajan las columnas del alineamiento original y se toman al azar (con repetición) tantas como sea la longitud del mismo n veces.

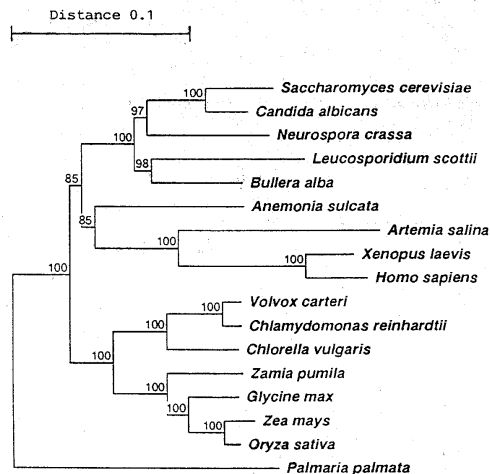
Original data:

Taxa	Sequence							
	1	2	3	4	5	6	7	8
A	G	C	A	G	T	A	C	T
B	G	T	A	G	T	A	C	T
C	A	C	A	A	T	A	C	C
D	A	C	A	A	C	A	C	T
E	G	C	G	G	C	A	T	T

A pseudosample:

Taxa	Sequence							
	6	1	6	5	2	1	1	5
A	A	G	A	T	C	G	G	T
B	A	G	A	T	T	G	G	T
C	A	A	A	T	C	A	A	T
D	A	A	A	C	C	A	A	C
E	A	G	A	C	C	G	G	C

x1000



Se cuantifica cuantas veces se mantienen los mismos clados en las sucesivas iteraciones, calculando el % del total nos da una estimación de los robusto que es cada clado del árbol.