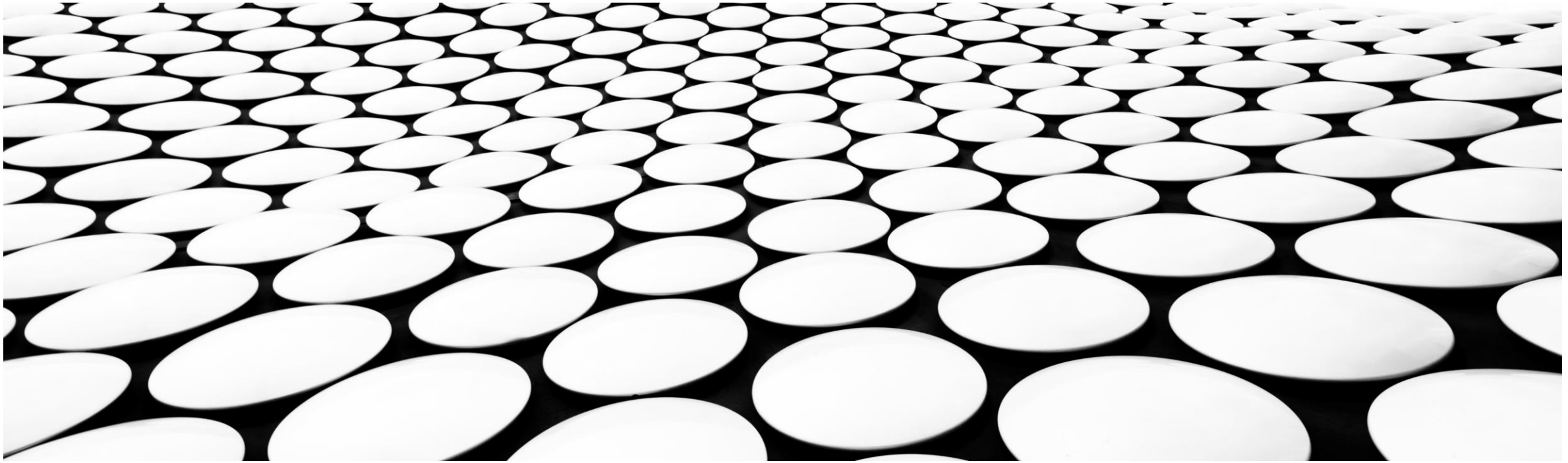
ARBOLES FILOGENÉTICOS

DRA. GUADALUPE DEL ROSARIO QUISPE SAJI



ARBOLES FILOGENETICOS – ETAPAS DE CONSTRUCCION

1. Adquision de secuencias
2. Alineamiento multiplo
3. Modelo de sustitucion de aminoacidos o nucleótidos
4. Metodos de construccion de arboles
5. Evaluacion del arbol

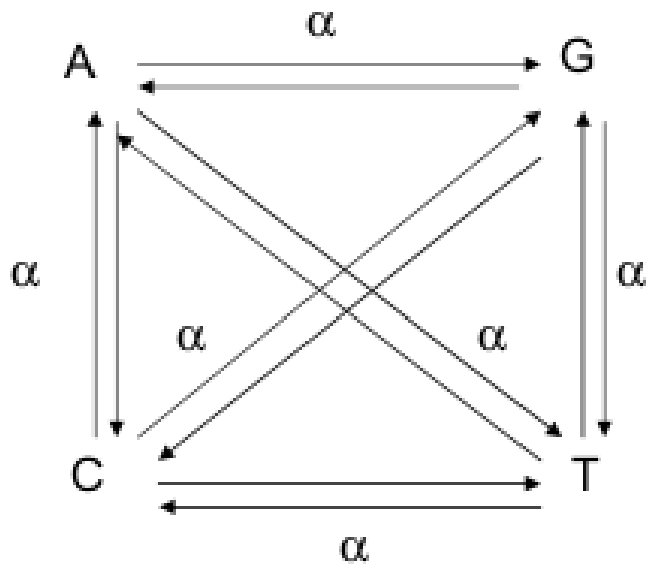
ARBOLES FILOGENETICOS – MODELOS DE SUSTITUCION

3. Modelos de sustitucion de aminoácidos y nucleótidos

La distancia evolutiva más sencilla es la basada en el modelo de Jukes y Cantor (JC69)

- asume que todas las sustituciones ocurren con la misma tasa α

- asume que los nts tienen la misma frecuencia 0.25



· El único parámetro del modelo JC69 es α

(tasa de sustitución instantánea)

ARBOLES FILOGENETICOS – MODELOS DE SUSTITUCION

3. Modelos de substicion de aminoácidos y nucleótidos

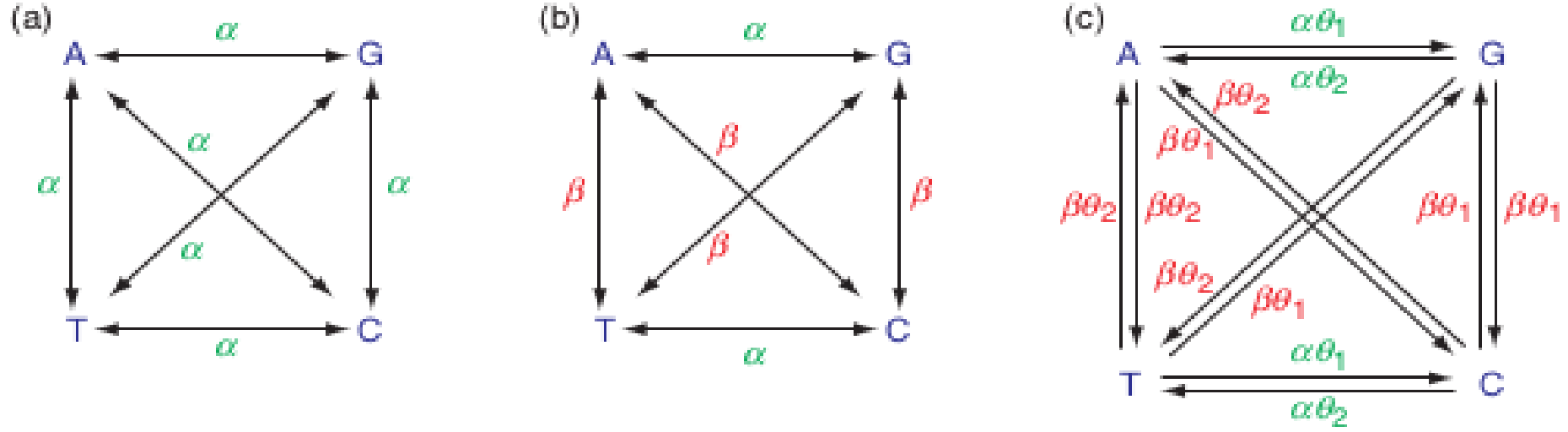


FIGURE 7.20 Models of nucleotide substitution. (a) The Jukes–Cantor model of evolution corrects for superimposed changes in an alignment. The model assumes that each nucleotide residue is equally likely to change to any of the other three residues and that the four bases are present in equal proportions. The rate of transitions α equals the rate of transversions β. (b) In the Kimura two-parameter model, $\alpha \neq \beta$. Typically, transversions are given more weight. (c) Tamura's model, which accounts for variations in GC content. This is an example of a more complex model of nucleotide substitution. Note that there are distinct parameters for nucleotide substitutions, and that many of these parameters are directional (e.g., the rate of changing from nucleotides T to C differs from the rate for C to T).

METODOS PARA RECONSTRUCCION DE ARBOLES FILOGENETICOS

- **Máxima Parsimonia** : Busca el árbol que implica el menor número de eventos de mutación.

- **Máxima probabilidad**: similar al método de máxima parsimonia, pero usa modelos evolutivos que calculan diferentes posibilidades de mutación dependiendo de la base que mutó

- **Métodos de distancia** utilizan comparaciones par a par entre las distancias (número de mutaciones) de las diferentes secuencias estudiadas para construir un árbol que reproduzca esta distancia del modo mas fiel posible.

ARBOLES FILOGENETICOS – MODELOS DE SUSTITUCION

4. Metodos de construccion de arboles

Metodos de distancia

METODOS DE NEIGHBOR JOINING

METODO UPGMA

Metodos de Máxima parsimonia

METODOS DE MAXIMA PARSIMONIA

Metodos Máxima probabilidad

METODOS DE MAXIMA VEROSIMILITUD

METODOS DE NEIGHBOR JOINING : CALCULO DEL PARAMETRO Q

	A	B	C	D	E
A	—	22	39	39	41
B	—	—	41	41	43
C	—	—	—	18	20
D	—	—	—	—	10
E	—	—	—	—	—

1. Se calcula un factor **u**, cuyo valor es igual a la suma de las distancias del nodo de interés (terminal) en relación a los otros nodos o puntos dividido por el numero de puntos (número de puntos-2)

- u_a=(22+39+39+41)/(5-2)=47
- u_b=(22+41+41+43)/(5-2)=49
- u_c=(39+41+18+20)/(5-2)=39.3
- u_d=(39+41+18+10)/(5-2)=36
- u_e=(41+43+20+10)/(5-2)=38

METODOS DE NEIGHBOR JOINING : CALCULO DEL PARAMETRO Q

	A	B	C	D	E
A	—	22	39	39	41
B	—	—	41	41	43
C	—	—	—	18	20
D	—	—	—	—	10
E	—	—	—	—	—

Para cada par de nodos terminales i,j se calcula Q definido por: $D_{i,j} - u_i - u_j$
Luego es seleccionado el menor valor

AB-> 22-47-49= -74

AC-> 39-47-39.3= -47.3

AD-> 39-47-36= -44

AE-> 41-47-38= -44

BC-> 41-49-39.3= -47.3

BD-> 41-49-36= -44

BE-> 43-49-38 = -44

CD-> 18-39.3-36=-57.3

CE-> 20-39.3-38= -57.3

DE-> 10-36-38= -64

METODOS DE NEIGHBOR JOINING : CALCULO DEL PARAMETRO Q

	A	B	C	D	E
A	—	22	39	39	41
B	—	—	41	41	43
C	—	—	—	18	20
D	—	—	—	—	10
E	—	—	—	—	—

Para cada par de nodos terminales i,j se calcula Q definido por: $D_{i,j} - u_i - u_j$
Luego es seleccionado el menor valor

AB-> 22-47-49= -74

AC-> 39-47-39.3= -47.3

AD-> 39-47-36= -44

AE-> 41-47-38= -44

BC-> 41-49-39.3= -47.3

BD-> 41-49-36= -44

BE-> 43-49-38 = -44

CD-> 18-39.3-36=-57.3

CE-> 20-39.3-38= -57.3

DE-> 10-36-38= -64

METODOS DE NEIGHBOR JOINING : CALCULO DEL NUEVO NODO

La distancia de los puntos A y B al nuevo punto o nodo AB se calculada como:

$v_a = ½ D_{AB} + ½ (u_a - u_b)$ $\qquad$ $v_b = ½ DAB + ½ (u_b - u_a)$

$v_a = 11 + ½ (47-49) = 10$ $\qquad$ $v_b = 11 + ½ (49-47) = 12$

- Se reemplazan los nodos terminales A y B por el nodo (AB) que se considera como un nuevo punto
- Las distancias de este nodo a los otros nodos terminales (k) se calculan con la siguiente fórmula: $D_{(AB),k} = (D_{ak} + D_{bk} - D_{ab})/2$

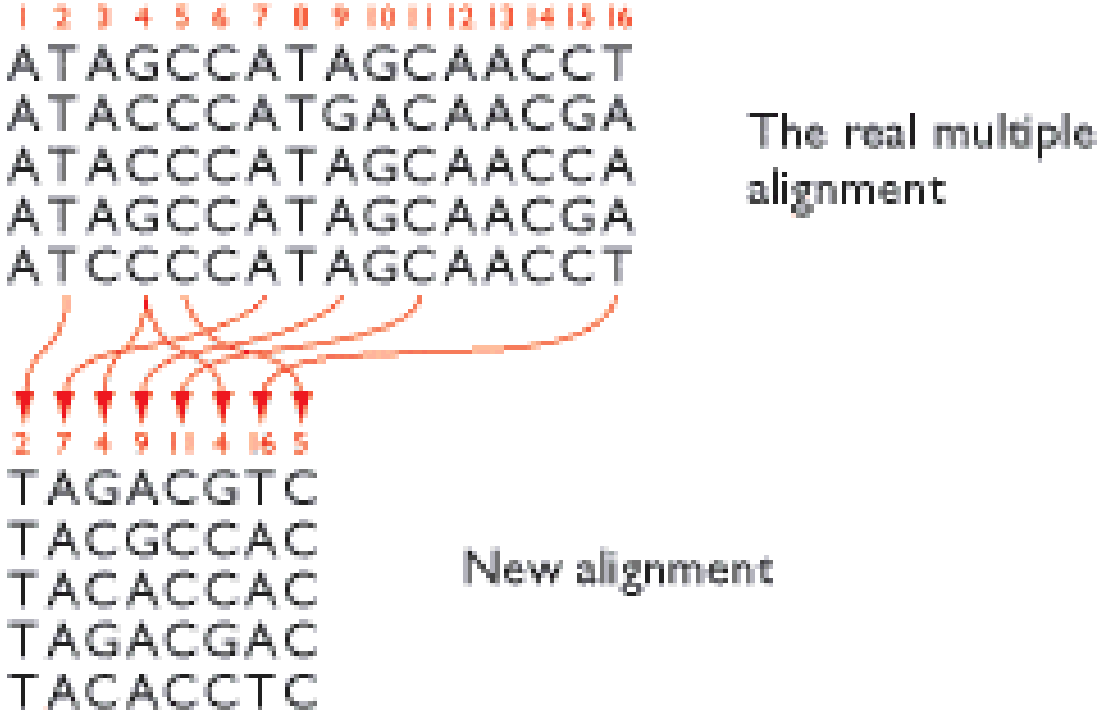
A	(AB)	C	D	E
(AB)	—	29	29	31
C	—	—	18	20
D	—	—	—	10
E	—	—	—	—

METODOS DE NEIGHBOR JOINING : CALCULO DEL *BOOTSTRAP*

El cálculo de *bootstrap* para árboles filogenéticos permite tener un parámetro que refleja la robustez del análisis filogenético producido.

El *Bootstrap* se genera mediante la creación de múltiples conjuntos de secuencias en las que las columnas de alineación múltiple se eligen al azar

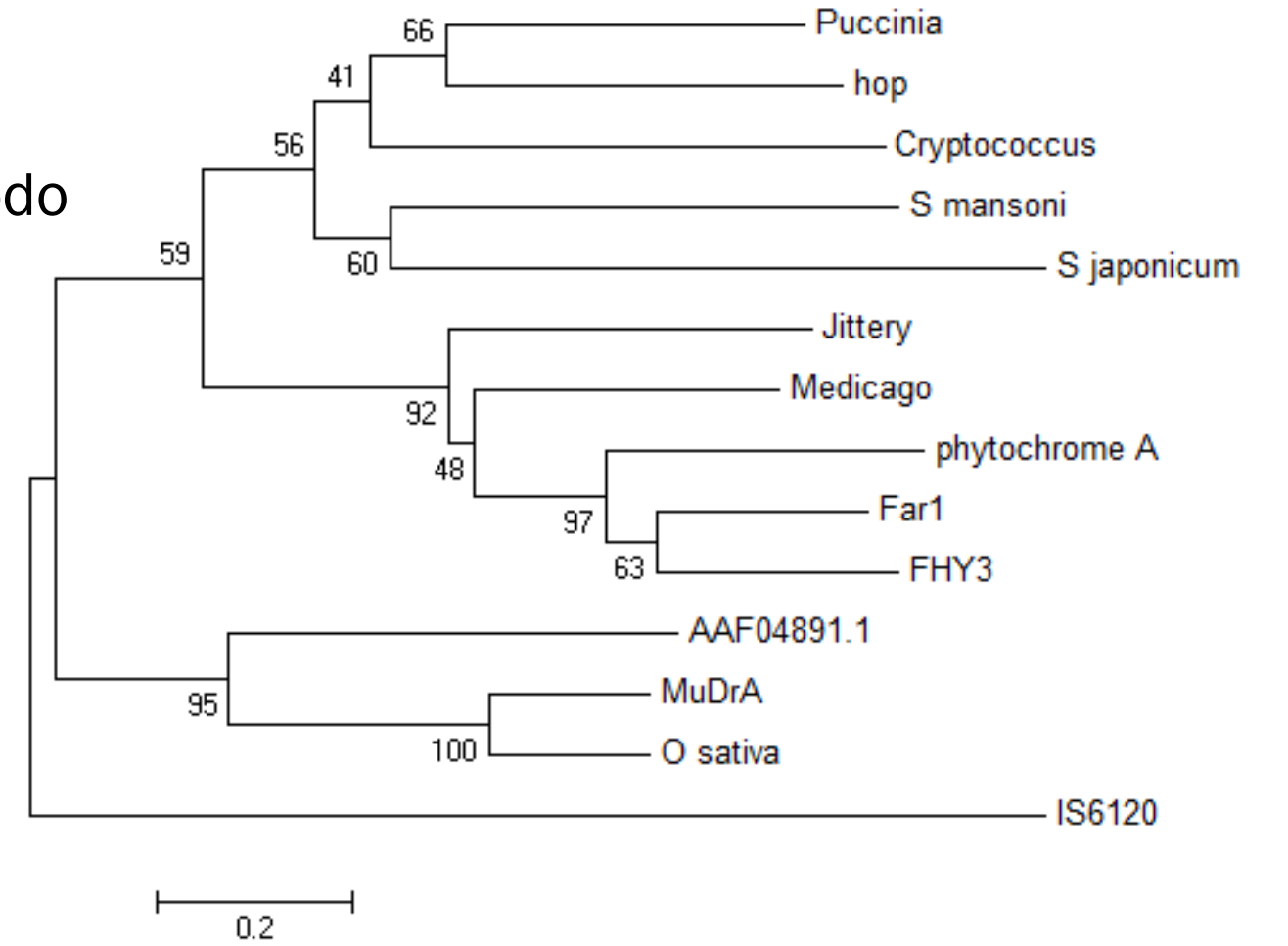
Se generan nuevos análisis para cada uno de los nuevos conjuntos generados.



METODOS DE NEIGHBOR JOINING : CALCULO DEL *BOOTSTRAP*

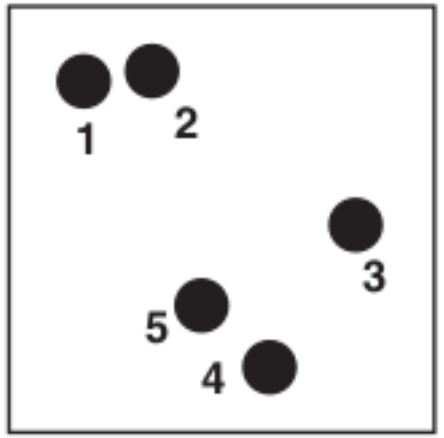
Los valores en los nodos internos representan los valores del Bootstrap, es decir el numero de apariciones de este nodo interno en 100 réplicas

El valor bootstrap indica que en 95/100 muestreos las secuencias AAF04891.1. MuDrA y O sativa fueron agrupadas en una rama que contiene solo las tres secuencias.



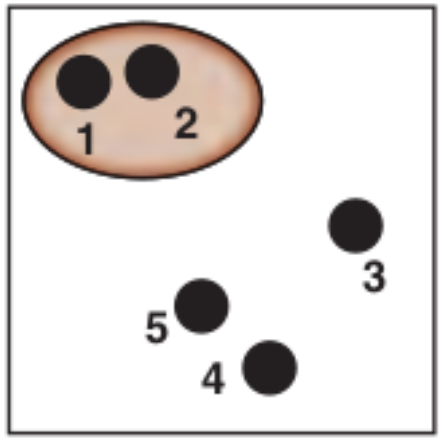
METODO UPGMA

(a)

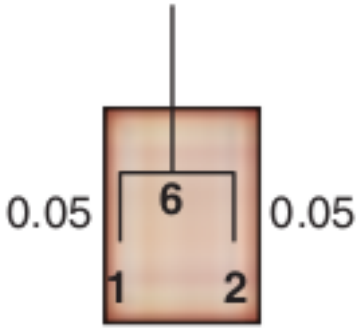


	1	2	3	4	5
1	—				
2	0.1	—			
3	0.8	0.8	—		
4	0.8	1	0.3	—	
5	0.9	0.9	0.3	0.2	—

(b)



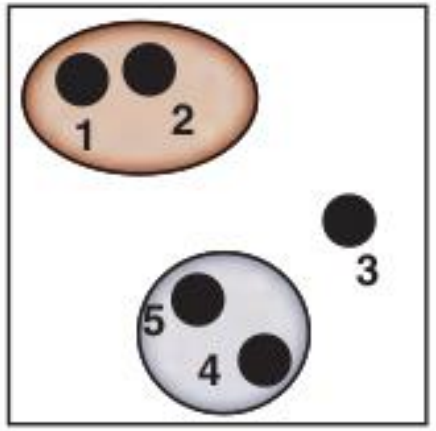
	(1,2)	3	4	5
(1,2)	—			
3	0.8	—		
4	0.9	0.3	—	
5	0.9	0.3	0.2	—



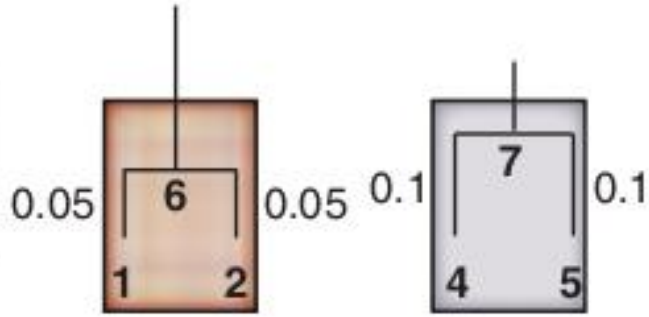
(a) Each sequence is assigned to its own cluster. A distance matrix, based on some metric, quantitates the distance between each object. The circles in the figure represent these sequences. (b) The taxa with the closest distance (sequences 1 and 2) are identified and connected. This allows us to name an internal node (right, node 6 in (b)). The distance matrix is reconstructed counting taxa 1 and 2 as a group. We can also identify the next closest sequences (4 and 5; distance is in red).

METODO UPGMA

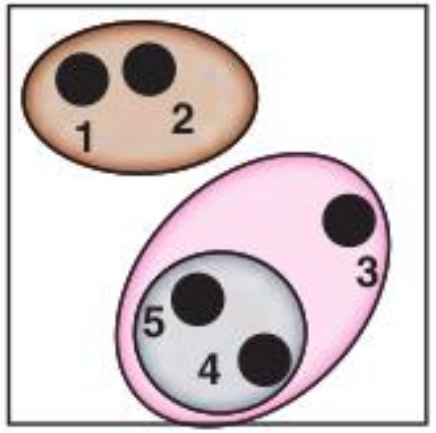
(c)



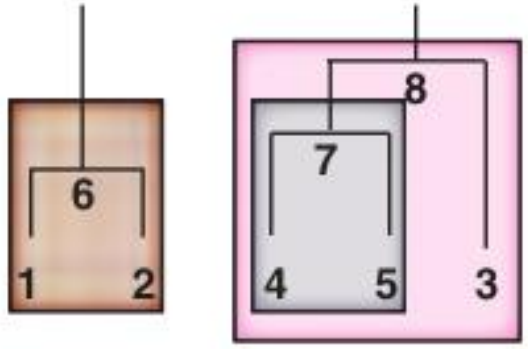
	(1,2)	3	(4,5)
(1,2)	—		
3	0.8	—	
(4,5)	0.9	0.3	—



(d)

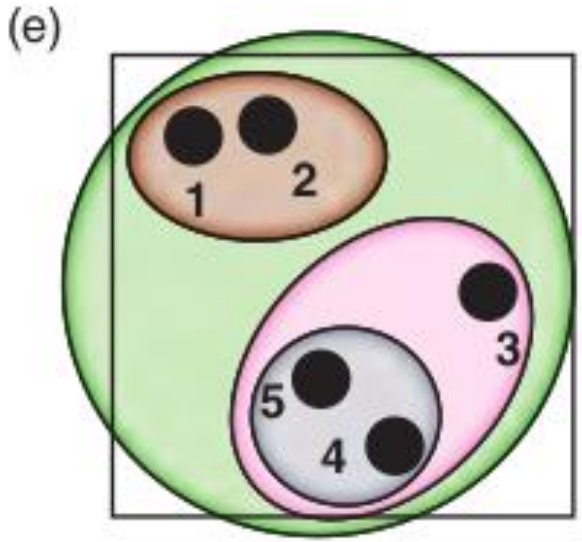


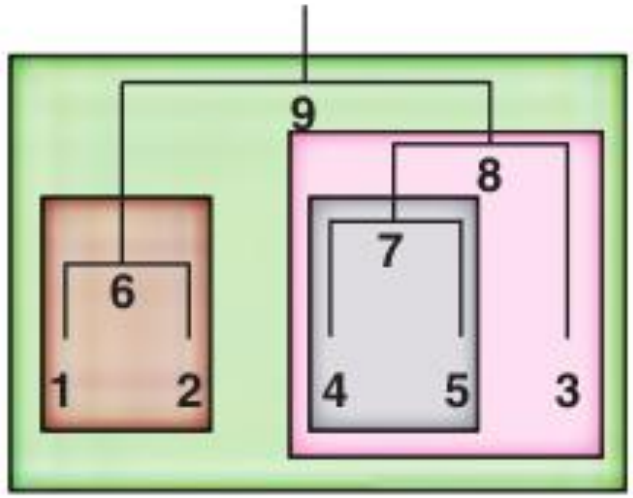
	(1,2)	[3,(4,5)]
(1,2)	—	
[3,(4,5)]	0.85	—



(c) These next closest sequences (4 and 5) are combined into a cluster, and the matrix is again redrawn. In the tree (right side) taxa 4 and 5 are now connected by a new node, 7. We can further identify the next smallest distance (value 0.3, red font) corresponding to the union of taxon 3 to cluster (4, 5). (d) The newly formed group (cluster 4, 5 joined with sequence 3) is represented on the emerging tree with new node 8.

METODO UPGMA



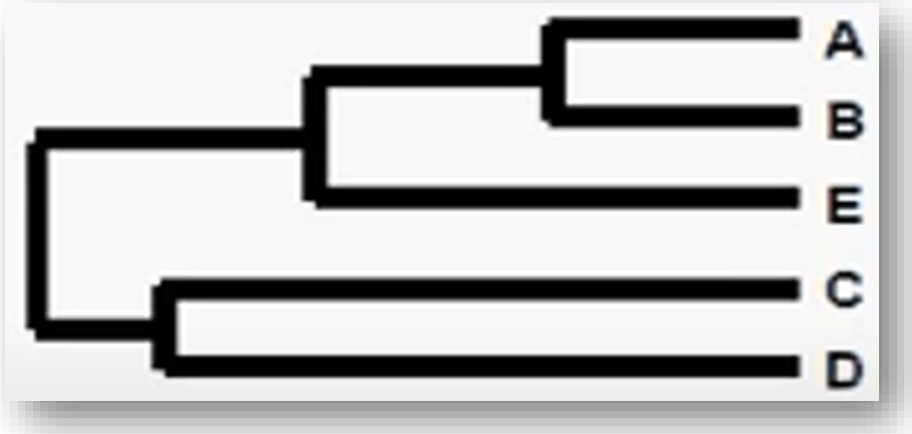


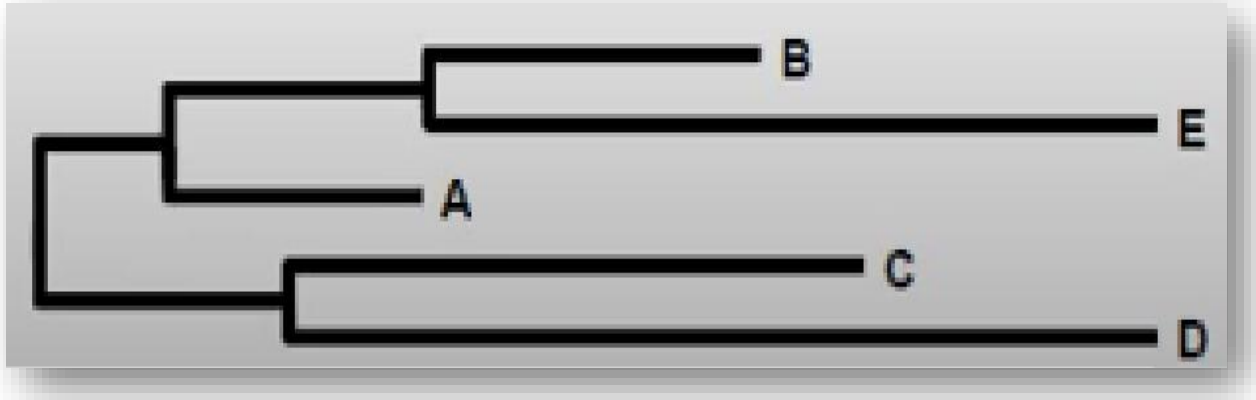
(e) all sequences are connected in a rooted tree.

METODO UPGMA VS NEIGHBOUR-JOINING

	A	B	C	D	E	F
A	-	17	21	31	23	92
B	-48	-	30	34	21	102
C	-49	-43	-	28	39	118
D	-45	-45	-57	-	43	136
E	-50	-55	-42	-44	-	136

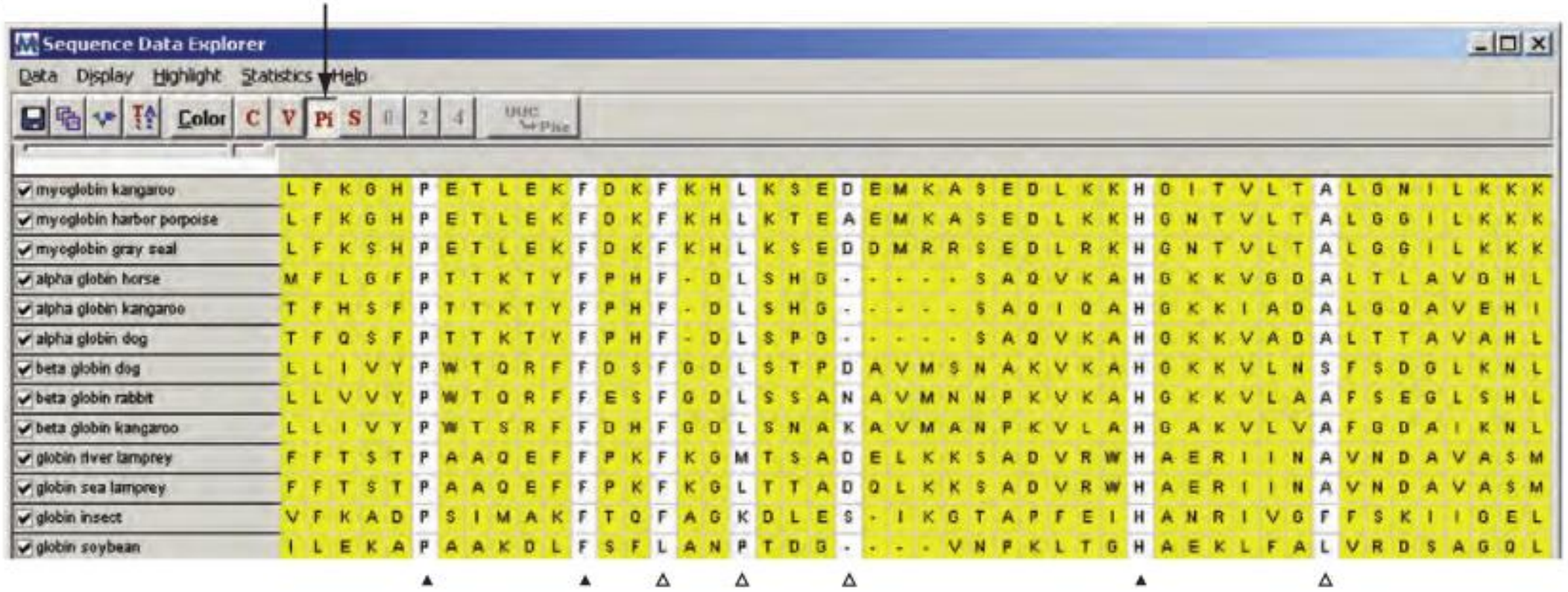
Triangulo superior: matriz de distancias
Triangulo inferior: distancias de ratio corregido





METODOS DE MAXIMA PARSIMONIA

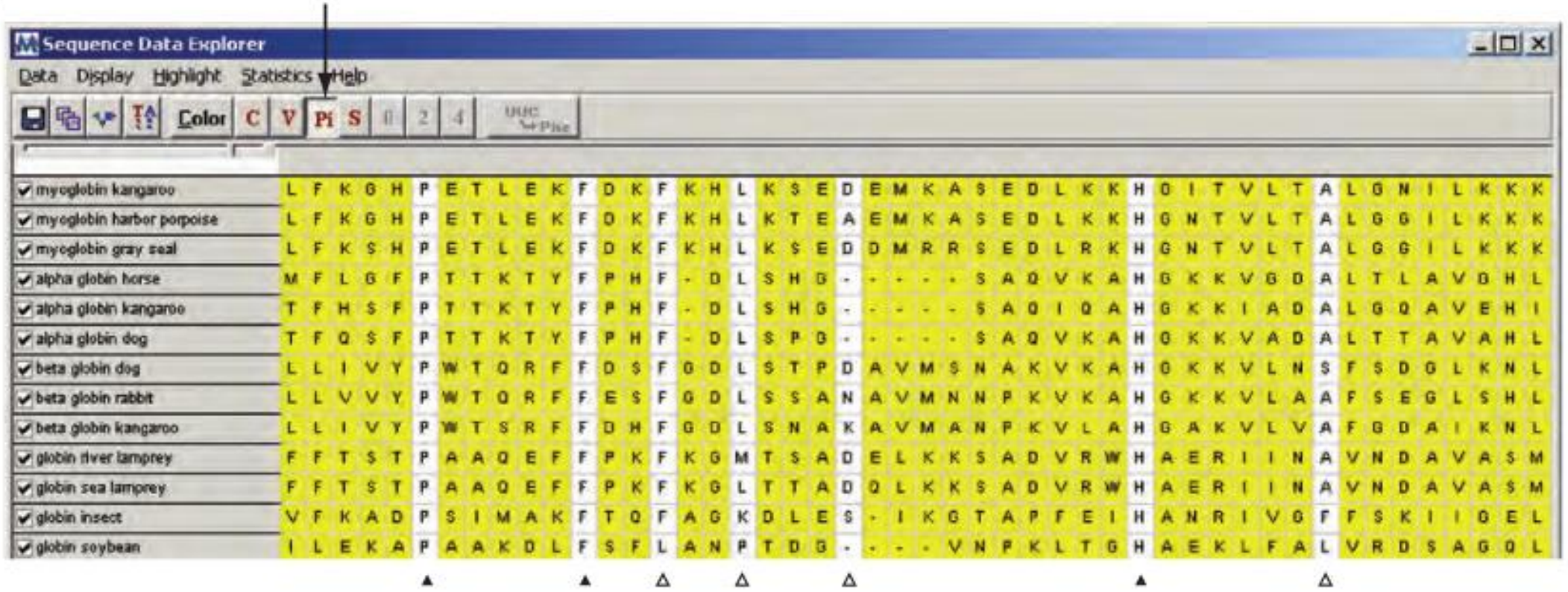
(a)



a) Many columns of aligned residues are informative for parsimony analysis. However, columns with entirely conserved residues (filled arrowheads) are not informative, nor are columns in which there are at least two different residues that occur at least two times (open arrowheads). This alignment of 13 globin proteins was viewed in MEGA software, with the option to display parsimony-informative characters selected (see arrow); other options include viewing conserved or variable position s.

METODOS DE MAXIMA PARSIMONIA

(a)



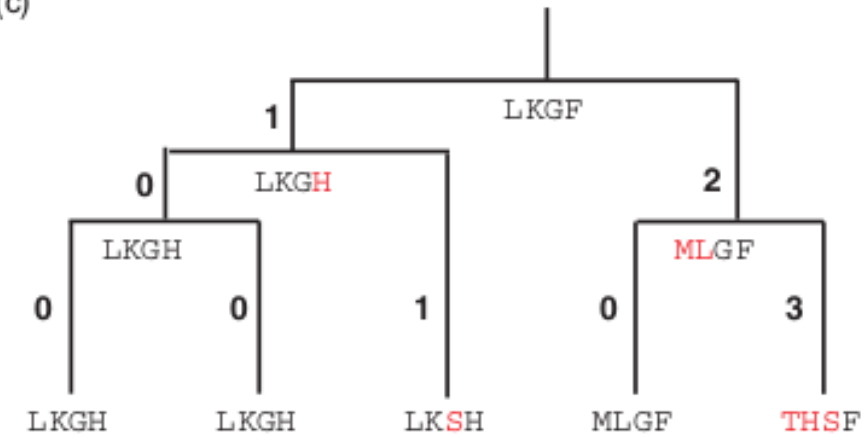
a) Many columns of aligned residues are informative for parsimony analysis. However, columns with entirely conserved residues (filled arrowheads) are not informative, nor are columns in which there are at least two different residues that occur at least two times (open arrowheads). This alignment of 13 globin proteins was viewed in MEGA software, with the option to display parsimony-informative characters selected (see arrow); other options include viewing conserved or variable position s.

METODO DE MAXIMA PARSIMONIA

(b)

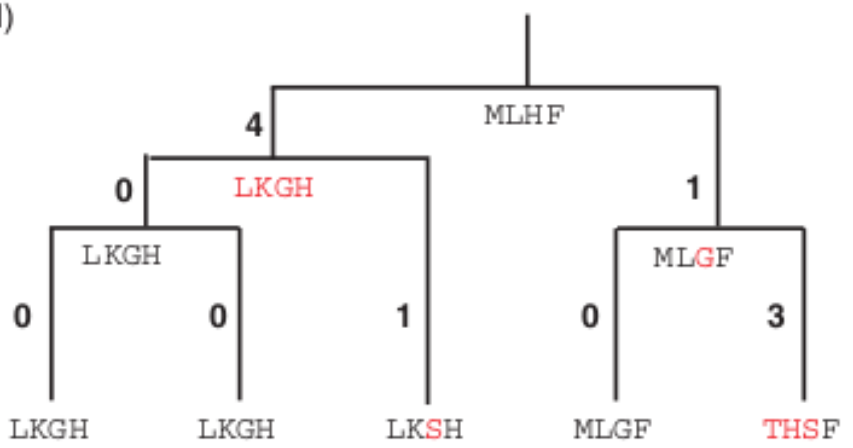
kangaroo	LKGH
porpoise	LKGH
gray seal	LKSH
horse α	MLGF
kangaroo α	THSF

(c)



Total cost: 7

(d)



Total cost: 9

(b) Example of four amino acid residues from five different species (taken from the top left of (a)). Maximum parsimony identifies the simplest (most parsimonious) evolutionary path by which those sequences might have evolved from ancestral sequences.
(c, d) Two trees showing possible ancestral sequences. The tree in (c) requires 7 changes from its common ancestor, while the tree in (d) requires 9 changes. Maximum parsimony would therefore select the tree in (c).

METODO DE MAXIMA VEROSIMILITUD

Evalúa la probabilidad de que el modelo de evolución elegido genere los datos observados: P(D|H)

Por ejemplo, todos los nucleótidos tienen la misma probabilidad

El programa prueba todos los nucleótidos posibles en cada nodo dentro del árbol y calcula la probabilidad de que estas elecciones generen los datos observados (las secuencias de las hojas)

METODO DE MAXIMA VEROSIMILITUD

Evalua la probabilidad de cada sustitucion de cada posible arbol

Todos los arboles posibles son considerados, y el nro de sustituciones que podria haber sucedido son calculados

Las probabilidades de todas las reconstrucciones posibles son sumadas para determinar la probabilidad de cada sitio

La probabilidad del árbol es el producto de las probabilidades para todas las posiciones de alineamiento

El arbol con la mas alta probabilidad es seleccionado a ser el arbol correcto.

METODO DE MAXIMA VEROSIMILITUD

Likelihood (L) = Probability ($data_{observed}$ | model)

Data : HHTHTH

Model 1 : fair coin	Prob(H) = 0.5, Prob(T) = 0.5
Model 2 : 2-head coin	Prob(H) = 1.0, Prob(T) = 0.0
Model 3 : 2-tail coin	Prob(H) = 0.0, Prob(T) = 1.0

L (Data|Model1)

= Prob(H|Model1) * Prob(H|Model1) * Prob(T|Model1) * Prob(H|Model1) * Prob(T|Model1) * Prob(H|Model1)

= 0.5 * 0.5 * 0.5 * 0.5 * 0.5 * 0.5 = 0.0156

L (Data|Model2) = 1.0 * 1.0 * 0.0 * 1.0 * 0.0 * 1.0 = 0.0

L (Data|Model3) = 0.0 * 0.0 * 1.0 * 0.0 * 1.0 * 0.0 = 0.0

METODO DE MAXIMA VEROSIMILITUD

Find the model that maximizes the likelihood of the observed data

Data : GGACGCCTGACGCCGCTCGG

Model 1: equal base composition - 0.25, 0.25, 0.25, 0.25 – A, C, G, T, respectively
Model 2: G+C bias - 0.1, 0.4, 0.4, 0.1 – A, C, G, T, respectively
Model 3: A+T bias - 0.4, 0.1, 0.1, 0.4 – A, C, G, T, respectively

L (Data|Model1) = Prob(G|Model1)*Prob(G|Model1)*Prob(A|Model1)*...* Prob(G|Model1) = $0.25^{20} = 9.1 \times 10^{-13}$

L (Data|Model2) = $0.4^{16} * 0.1^{4} = 4.3 \times 10^{-11}$ ← maximum likelihood

L (Data|Model3) = $0.1^{16} * 0.4^{4} = 2.6 \times 10^{-16}$

METODO DE MAXIMA VEROSIMILITUD

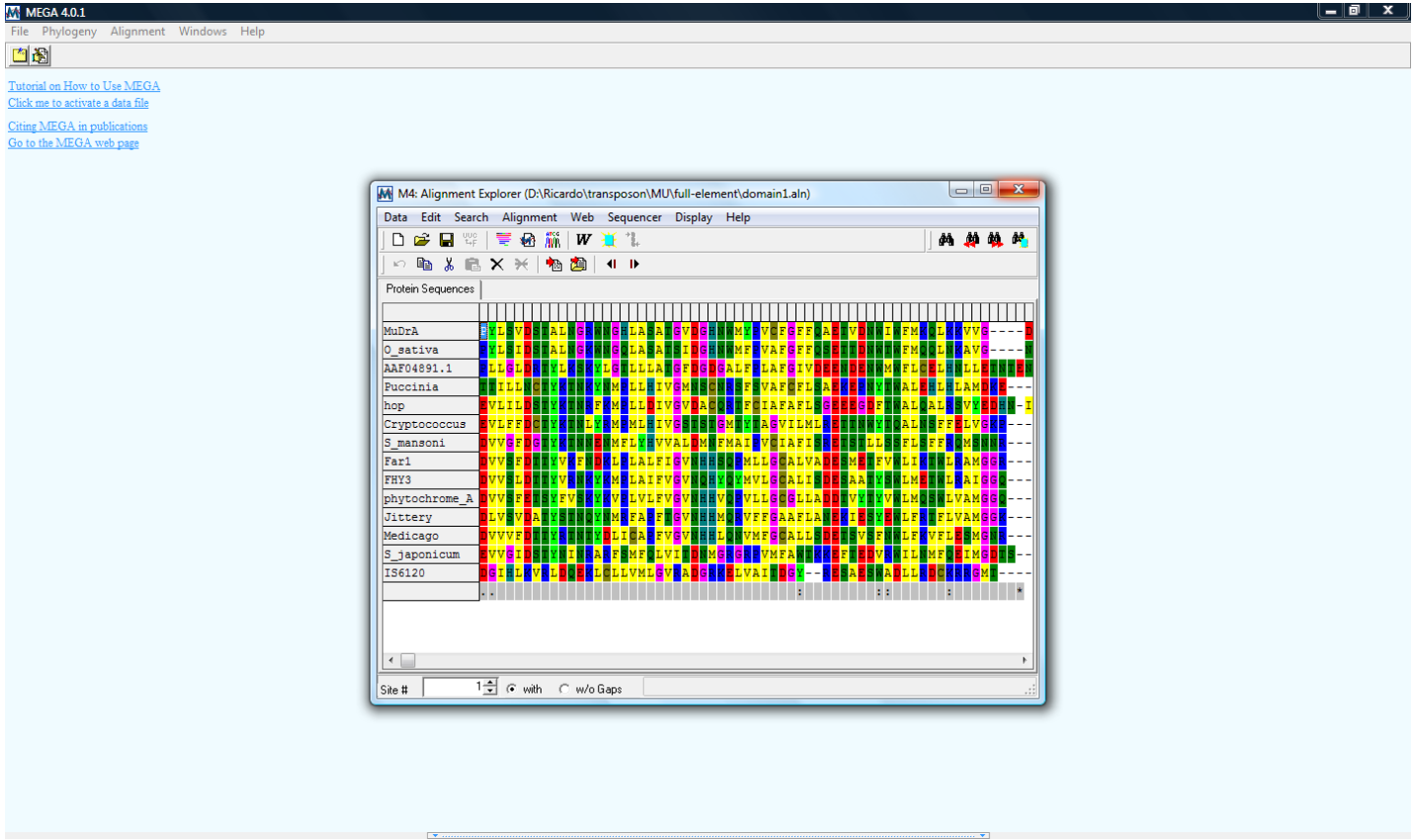
Programas

- RaXML
  - A. Stamatakis
  - http://www.exelixis-lab.org/
- phyML
  - O. Gascuel et al. Systematic Biology, 59(3):307-21, 2010
  - http://www.atgc-montpellier.fr/phyml/
- fastTree
  - Morgan N. Price in Adam Arkin's group
  - http://www.microbesonline.org/fasttree/
  - "FastTree can handle alignments with up to a million of sequences in a reasonable amount of time and memory"
- IQ-Tree
  - http://www.iqtree.org/
  - A fast and effective stochastic algorithm to infer phylogenetic trees by maximum likelihood. *IQ-TREE compares favorably to RAxML and PhyML* in terms of likelihoods with similar computing time (Nguyen et al., 2015)

MEGA

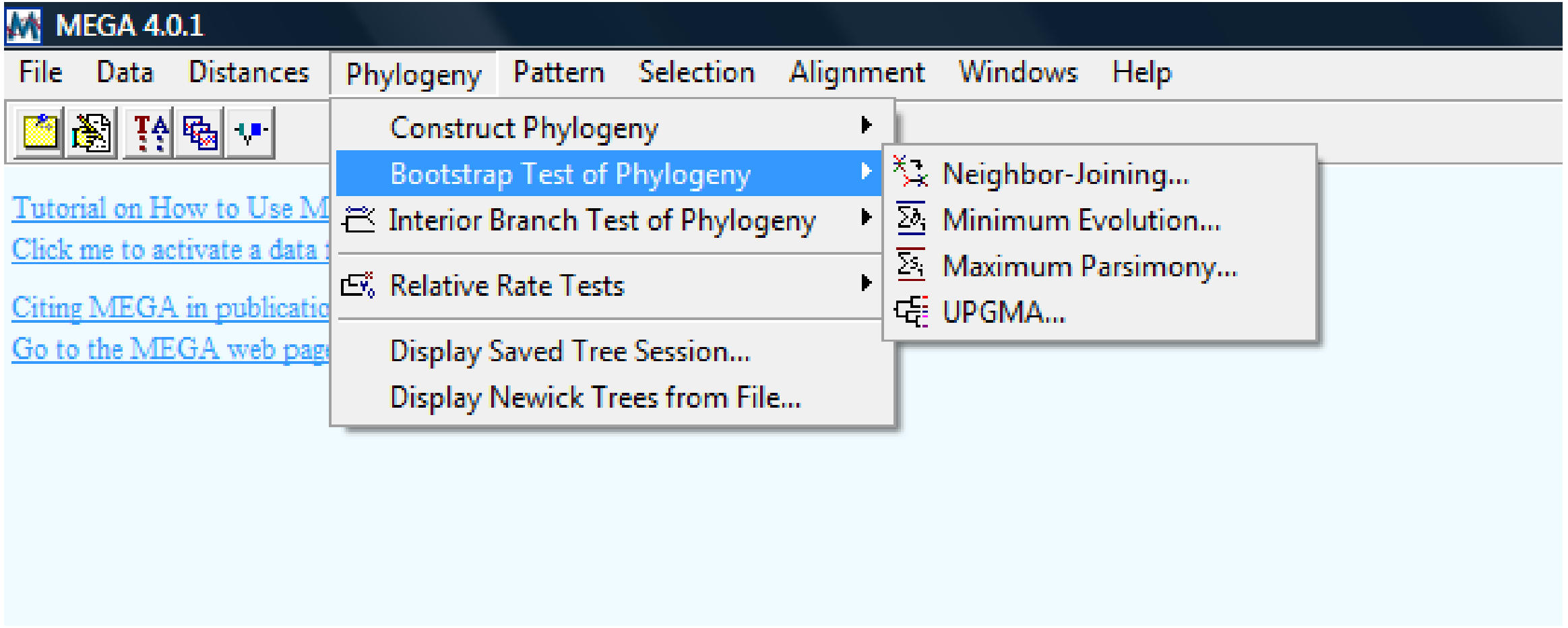
El programa MEGA realiza una serie de análisis evolutivos basados en múltiples alineamientos múltiples de secuencias.

Es posible importar secuencias de aln resultantes del alineamiento obtenido con el clustal



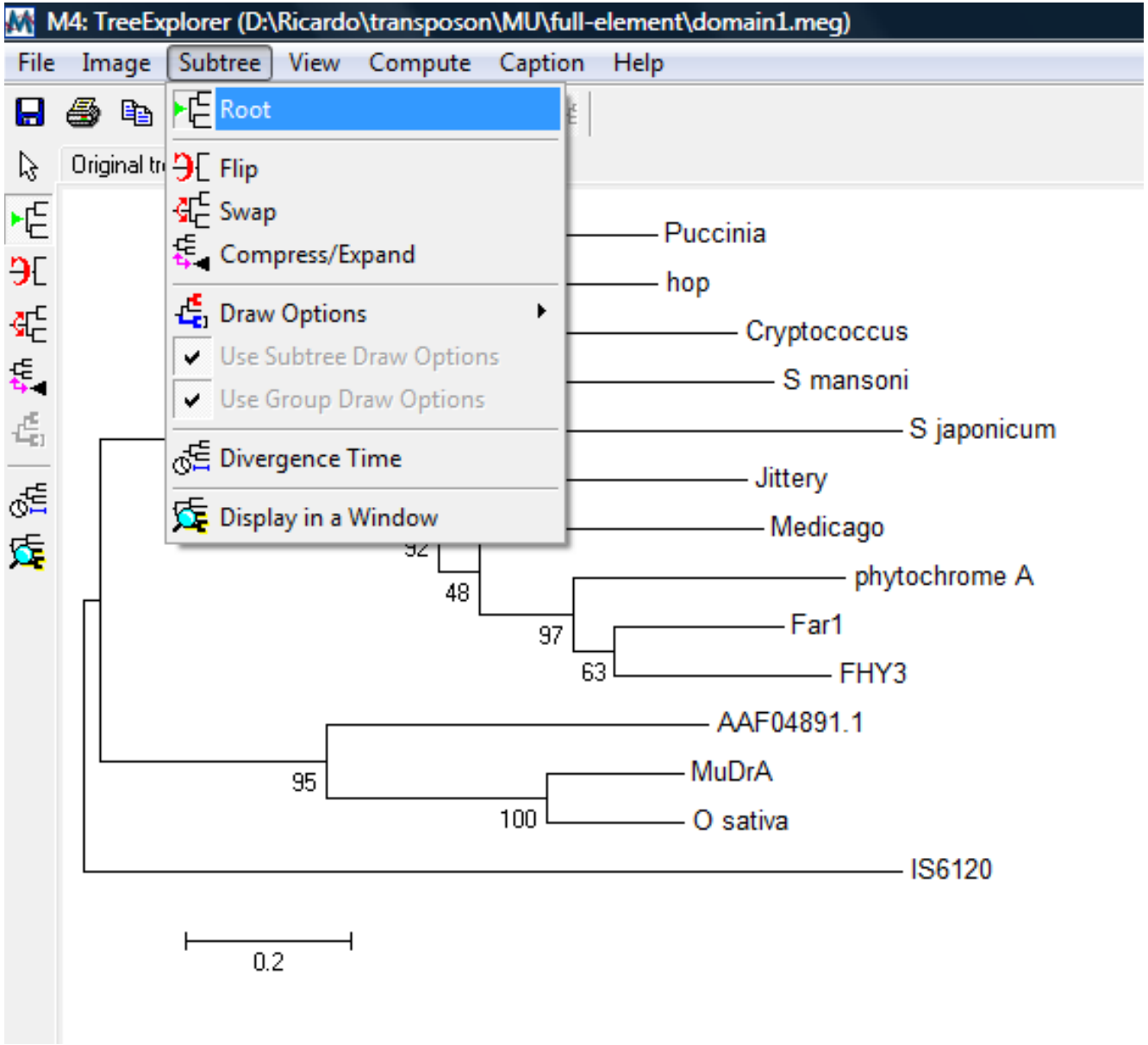
MEGA

El programa MEGA permite realizar el análisis filogenetico de alineamientos por diferentes metodos



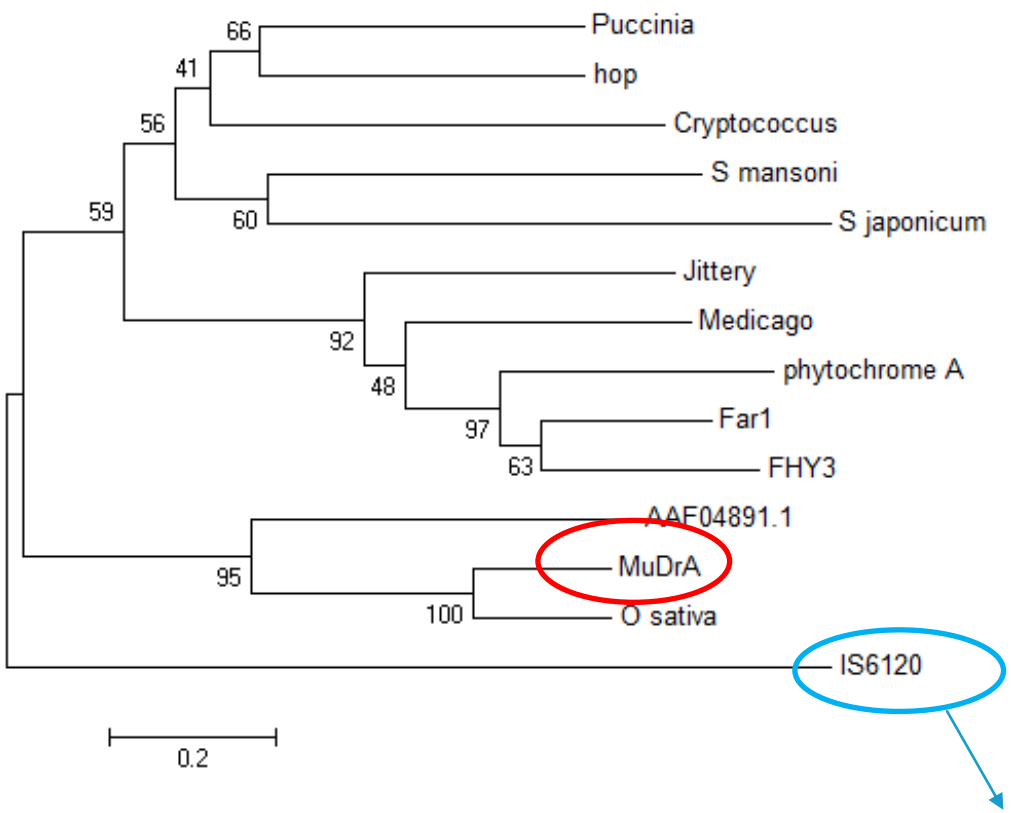
MEGA

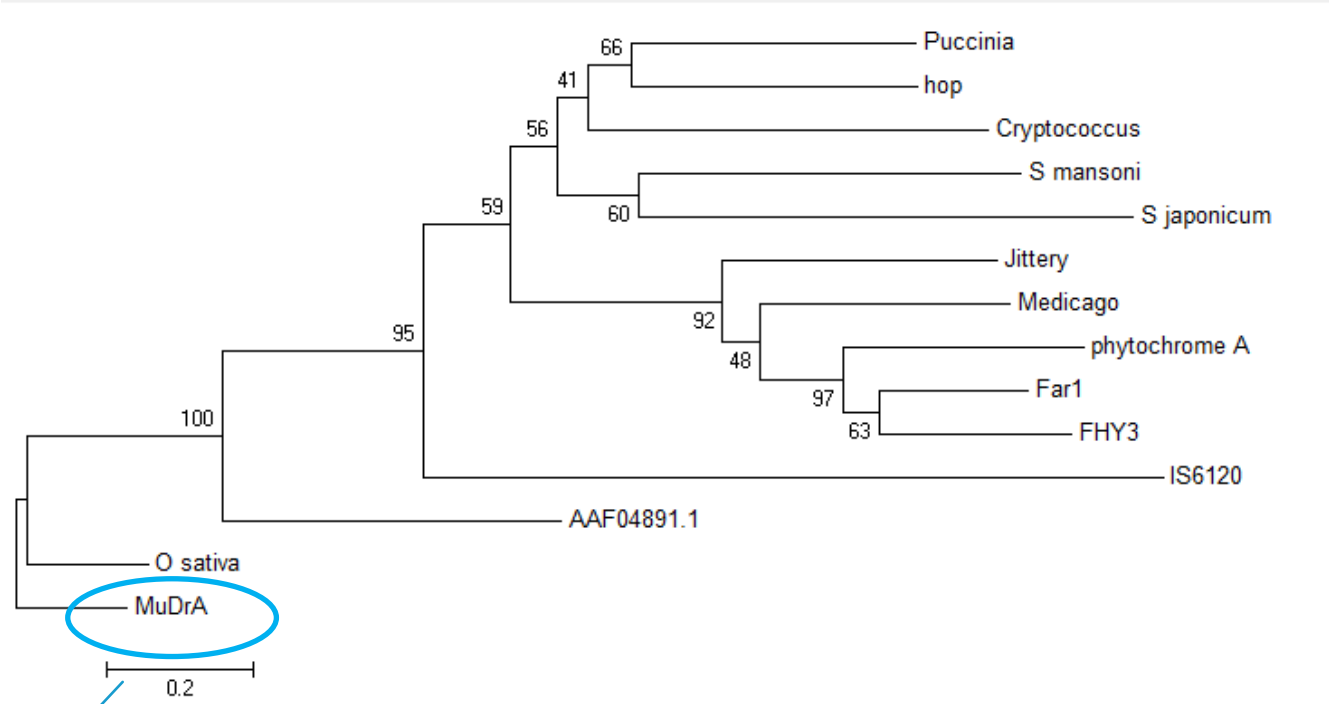
Permite escoger la rama a partir de la cual se desea enraizar el árbol usando la opción “root”



MEGA

Selección de la rama marcada en rojo





outgroup