EJERCICIO DE EVOLUCIÓN MOLECULAR Y FILOGENIA

Ignacio Taguas Garzón
Curso 2017/2018

CONSTRUCCIÓN DE ÁRBOLES FILOGENÉTICOS

En base a los datos dados, se han construido dos árboles filogenéticos, uno en base a una estimación por *Maximum Likelihood* y otro basado en una estimación Bayesiana.

Máximum Likelihood es un método que se basa únicamente en los datos que se proporcionan para estimar los parámetros que maximizan la probabilidad. La estimación bayesiana, en cambio, utiliza unos datos *a priori*.

Maximum Likelihood

Para la construcción de este árbol filogenético se han empleado los programas jModelTest, Notepad ++ PhyML y FigTree.

jModelTest se ha empleado para la estimación del mejor modelo de sustitución nucleotídica del alineamiento. Los parámetros importantes que se han usado para este análisis han sido…

- *Number of substitution schemes*: se han seleccionado 203 esquemas de sustitución; se ha elegido esta opción porque al ser un alineamiento sencillo el tiempo de computación no va a ser excesivo, y al ir a realizar posteriormente el árbol con PhyML todos los modelos estarán disponibles
- Se han indicado: una frecuencia de bases no homogénea (+F), la presencia de sitios invariables (+I) y la existencia de un ratio de variación entre sitios (+G)
- El *nCat* (número de categorías de la distribución gamma para el ratio de variación entre sitios) se ha mantenido en 4, pues aumentar el número aumenta el tiempo de computación pero no asegura un mejor resultado
- Como árbol base se utiliza BIONJ, que calcula un árbol de Neighbor-Joining para cada modelo (para un árbol guía este método obtiene resultados suficientemente buenos y el tiempo de computación es bajo)
- *Base tree search*: indica el método de refinamiento del árbol; se utiliza el método NNI

Una vez terminado el análisis los resultados se ordenan en función del parámetro BIC. En la figura 1 se muestran los parámetros obtenidos para el mejor modelo, que en este caso es el modelo 012342+I+G+F, así como debajo los siguientes métodos que han obtenido un valor más bajo para BIC.[1]

[1] También se guardan los parámetros de GTR+I+G, pues se usarán posteriormente para el análisis en BEAST

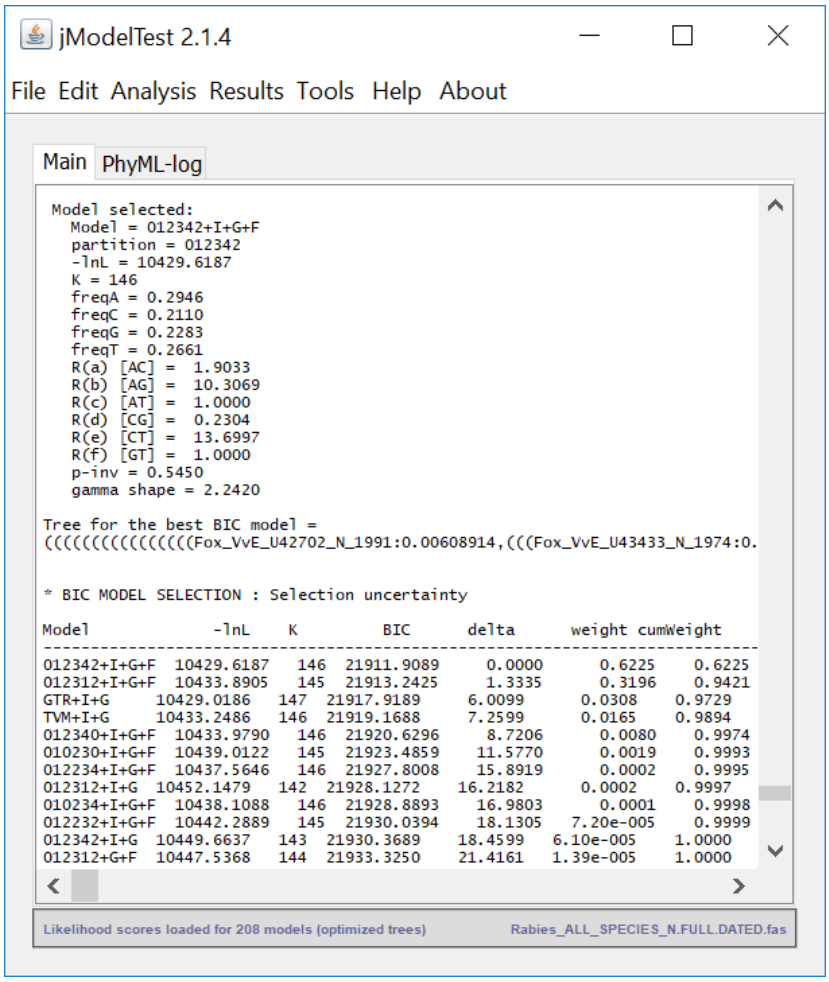
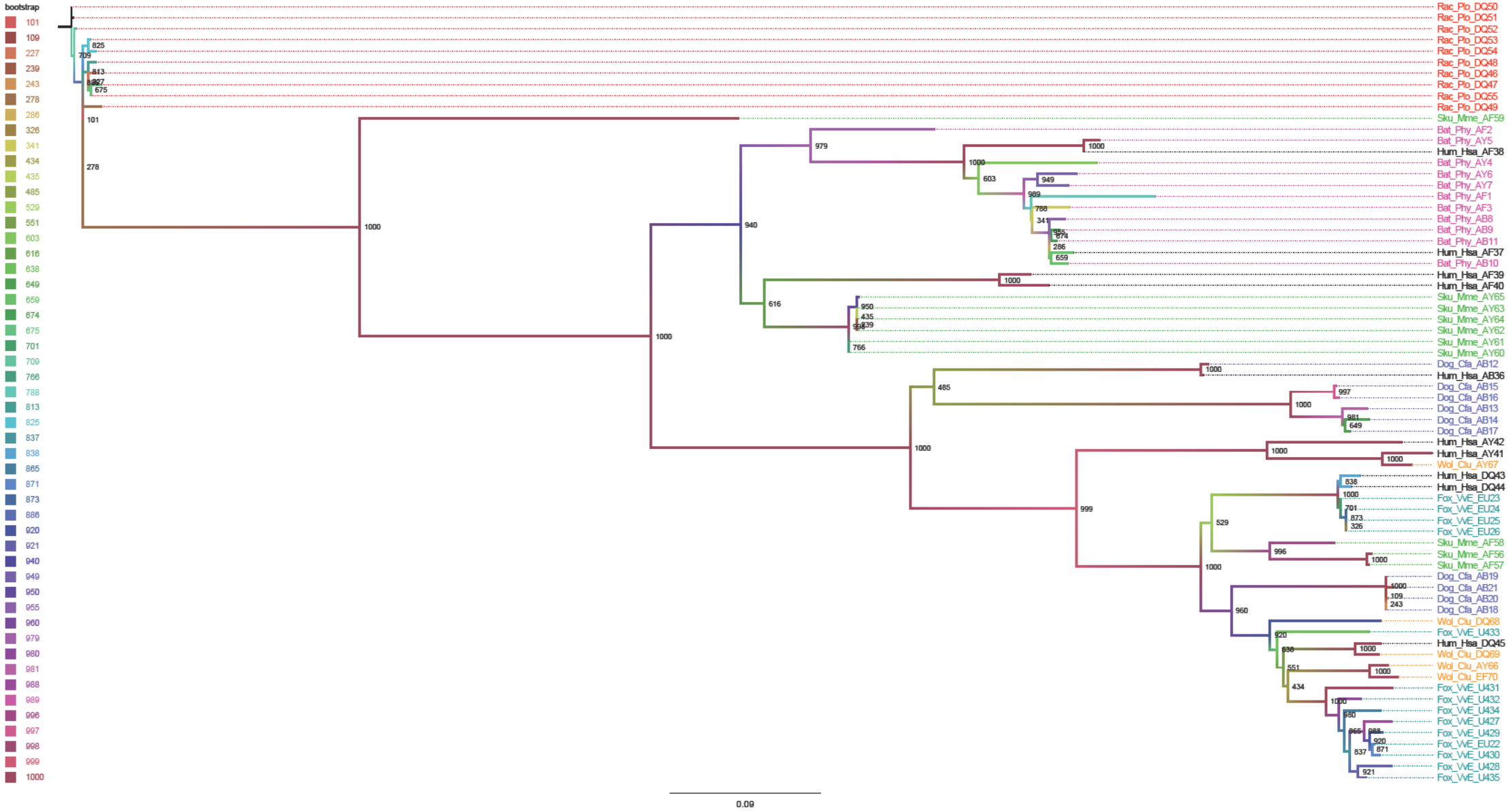


Figura 1

El siguiente paso consiste en la introducción de los parámetros obtenidos en la figura 1, así como del alineamiento (al que se le han tenido que modificar los nombres en Notepad ++ para que le programa los distinga), en PhyML. Se lleva a cabo el análisis de bootstrap no paramétrico, con 1.000 replicados.

El resultado obtenido consiste en cuatro documentos de texto. El que tiene la terminación .phy_phyml_tree es el árbol final que se ha obtenido, con los valores de bootstrap obtenidos. Para la visualización y edición del árbol se emplea FigTree. El resultado se adjunta en la página 3.


bootstrap
101
109
227
239
243
278
286
326
341
434
435
485
529
551
603
616
638
649
659
674
675
701
709
766
788
813
825
837
838
865
871
873
886
920
921
940
949
950
955
960
979
980
981
988
989
996
997
998
999
1000
Rac_Plo_DQ50
Rac_Plo_DQ51
Rac_Plo_DQ52
Rac_Plo_DQ53
Rac_Plo_DQ54
Rac_Plo_DQ48
Rac_Plo_DQ46
Rac_Plo_DQ47
Rac_Plo_DQ55
Rac_Plo_DQ49
Sku_Mme_AF59
Bat_Phy_AF2
Bat_Phy_AY5
Hum_Hsa_AF38
Bat_Phy_AY4
Bat_Phy_AY6
Bat_Phy_AY7
Bat_Phy_AF1
Bat_Phy_AF3
Bat_Phy_AB8
Bat_Phy_AB9
Bat_Phy_AB11
Hum_Hsa_AF37
Bat_Phy_AB10
Hum_Hsa_AF39
Hum_Hsa_AF40
Sku_Mme_AY65
Sku_Mme_AY63
Sku_Mme_AY64
Sku_Mme_AY62
Sku_Mme_AY61
Sku_Mme_AY60
Dog_Cfa_AB12
Hum_Hsa_AB36
Dog_Cfa_AB15
Dog_Cfa_AB16
Dog_Cfa_AB13
Dog_Cfa_AB14
Dog_Cfa_AB17
Hum_Hsa_AY42
Hum_Hsa_AY41
Wol_Clu_AY67
Hum_Hsa_DQ43
Hum_Hsa_DQ44
Fox_VvE_EU23
Fox_VvE_EU24
Fox_VvE_EU25
Fox_VvE_EU26
Sku_Mme_AF58
Sku_Mme_AF56
Sku_Mme_AF57
Dog_Cfa_AB19
Dog_Cfa_AB21
Dog_Cfa_AB20
Dog_Cfa_AB18
Wol_Clu_DQ68
Fox_VvE_U433
Hum_Hsa_DQ45
Wol_Clu_DQ69
Wol_Clu_AY66
Wol_Clu_EF70
Fox_VvE_U431
Fox_VvE_U432
Fox_VvE_U434
Fox_VvE_U427
Fox_VvE_U429
Fox_VvE_EU22
Fox_VvE_U430
Fox_VvE_U428
Fox_VvE_U435
0.09

Bayesian

Para la construcción de este árbol filogenético se han empleado los programas BEAUti2, BEAST2, Tracer, EmEditor, TreeAnnotator, y FigTree[2].

El primer paso es cargar en BEAUti2 el archivo formato nex; el objetivo de este programa es generar un archivo que pueda ser leído por BEAST2. Se han modificado los siguientes parámetros…

- *Use tip dates*: se indica qué pieza de información contenida en el nombre muestra la fecha
- *Site Model*: de los modelos que ofrece BEAUti2 se utiliza GTR (que, como se ve en la figura 1) es el primero de los 4 que aparece); introducimos los parámetros obtenidos
- *Clock Model*: permite determinar el cambio de las mutaciones en el tiempo; se elige *Relaxed Clock Log Normal*, que es un modelo que permite que cada rama del árbol tenga su propia tasa
- *Priors*: se elige el *Coalescent Extended Bayesian Skyline*, y el resto de parámetros de esta pestaña se mantienen
- *MCMC*: se indica que la *Chain Length* debe de ser de 200.000.000, que serán los pasos que se den en la construcción de las cadenas de Markov; así mismo, se guarda un árbol cada 1.000 pasos (de forma que el número total de árboles obtenidos al final será de 200.000)

Se genera un archivo .xml que se introduce en BEAST2. Se generan dos archivos, uno .log y otro .trees, que se van actualizando a medida se realiza el análisis. Se puede ir comprobando cómo avanza el análisis añadiendo el archivo .log en Tracer: en caso de que todos los parámetros tengan un valor de ESS mayor que 200 (es decir, que tengan más de 200.000 eslabones utilizados para calcular el parámetro), se podría detener el análisis.

Una vez finalizado el análisis se introduce el archivo .log en Tracer; los resultados, mostrados en la figura 2, son bastante peores de lo que en principio se podría esperar; para entender por qué pasa esto, en *posterior* se analiza la gráfica del recorrido que se ha llevado a cabo (figura 3).

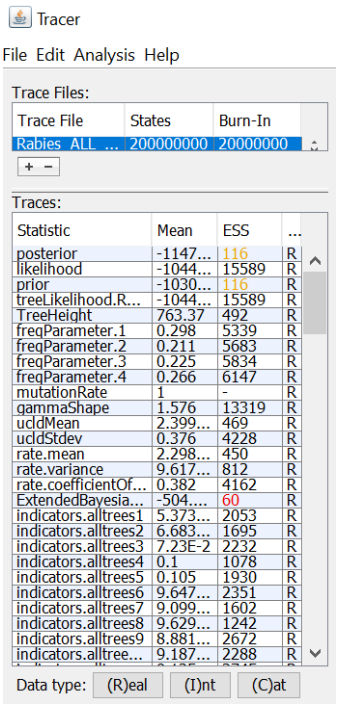


Figura 2

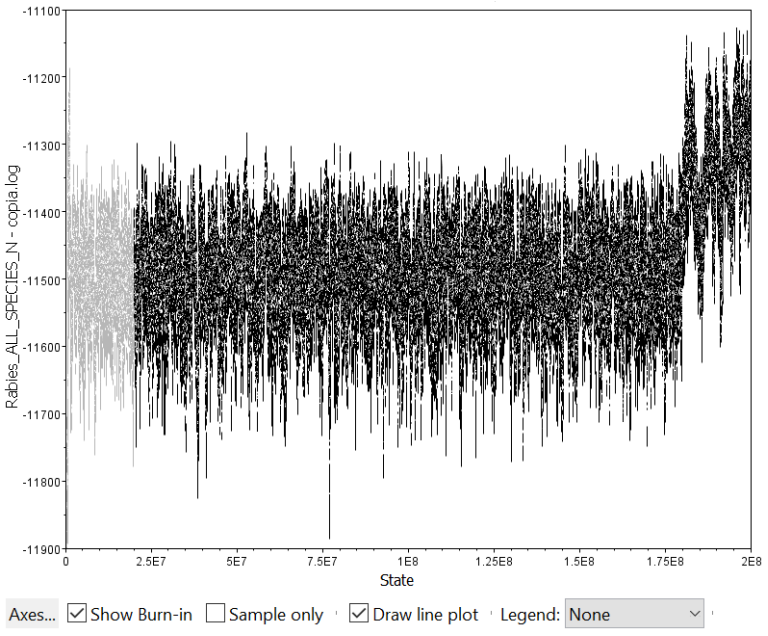


Figura 3

[2] He tenido que usar BEAST2 y BEAUti2 porque no he logrado que BEAST1 funcionase, y en BEAST2 no se pueden introducir archivos de BEAUti1.

En la gráfica se puede ver que el recorrido se ha mantenido estable durante prácticamente todo el análisis y al final, durante los últimos 20.000.000 de pasos aproximadamente, comienza a oscilar. La razón por la que podría haber pasado esto podría ser un fenómeno de metaestabilidad: al llevar estable mucho tiempo, produce un cambio, que no ha dado tiempo a volver a estabilizar porque termina el análisis. En caso de ser esta la causa, la solución hubiera sido ir comprobando el archivo .log como se ha indicado previamente y detener el análisis una vez se hubieran obtenido nuevos valores.

Para solucionar este problema se abren los archivos obtenidos con EmEditor (pues son demasiado grandes como para abrirlos con Notepad ++) y se eliminan manualmente las últimas secuencias. En este caso, se han eliminado 28.514.000 pasos, y los correspondientes árboles.

Se introduce el archivo modificado en Tracer; se puede ver que los valores obtenidos son ahora mucho mejores. Tracer aplica por defecto un *Burn-In* del 10%; sin embargo, como se ve en la figura 3, en este caso no hace falta un *Burn-In* tan alto, pues ha alcanzado la estabilidad muy rápido. Por tanto, se cambia el *Burn-In* a 1.300.000 pasos únicamente (es decir, no se consideran los primeros 1.300.000 pasos). Los valores obtenidos, así como la gráfica de los valores *a posteriori*, se muestran en la figura 4.

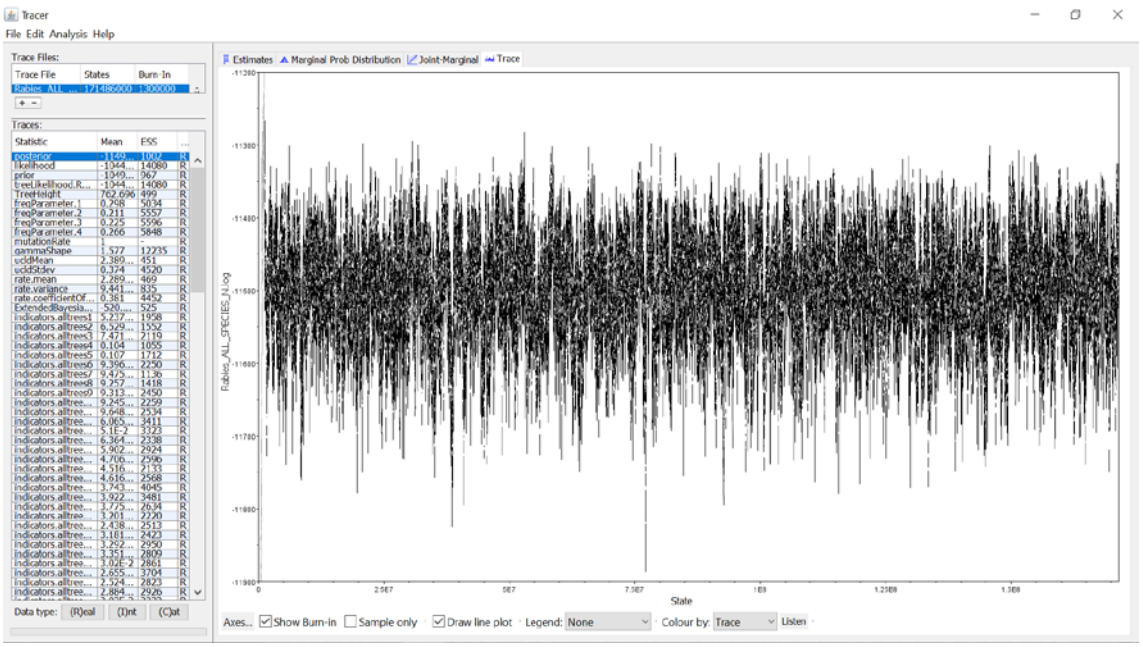
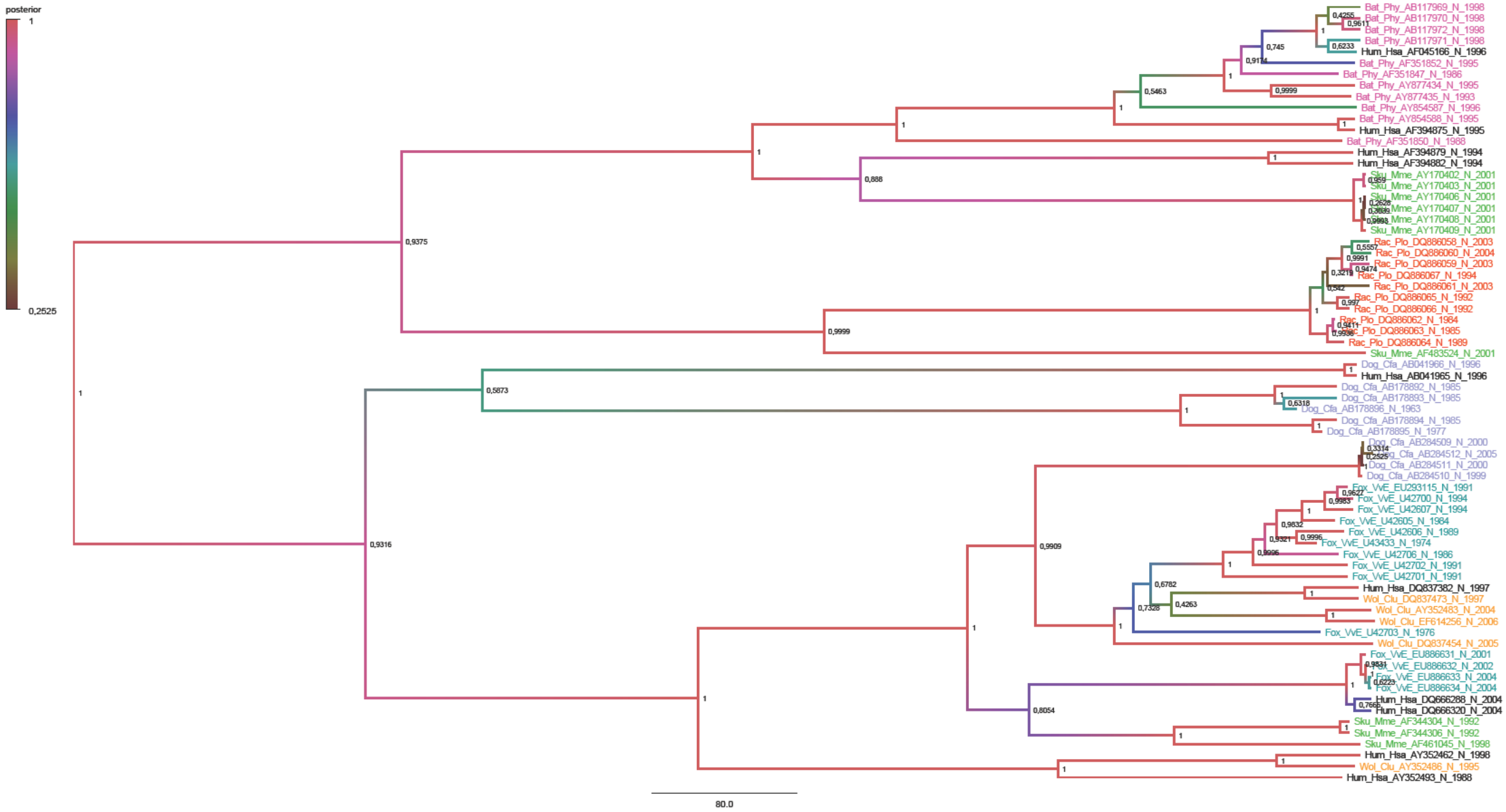


Figura 4

A continuación se introduce el archivo de extensión .trees (modificado, habiendo eliminado los últimos árboles) en TreeAnnotator; aquí se generará un árbol consenso entre todos los árboles. Se indica el *Burn-In* especificado previamente (1.300.000).

De nuevo, el resultado obtenido lo introducimos y modificamos en FigTree; en este caso añadimos los valores *posterior* en los nodos. El resultado se adjunta en la página 6.


posterior
1
0,2525
Bat_Phy_AB117969_N_1998
Bat_Phy_AB117970_N_1998
Bat_Phy_AB117972_N_1998
Bat_Phy_AB117971_N_1998
Hum_Hsa_AF045166_N_1996
Bat_Phy_AF351852_N_1995
Bat_Phy_AF351847_N_1986
Bat_Phy_AY877434_N_1995
Bat_Phy_AY877435_N_1993
Bat_Phy_AY854587_N_1996
Bat_Phy_AY854588_N_1995
Hum_Hsa_AF394875_N_1995
Bat_Phy_AF351850_N_1988
Hum_Hsa_AF394879_N_1994
Hum_Hsa_AF394882_N_1994
Sku_Mme_AY170402_N_2001
Sku_Mme_AY170403_N_2001
Sku_Mme_AY170406_N_2001
Sku_Mme_AY170407_N_2001
Sku_Mme_AY170408_N_2001
Sku_Mme_AY170409_N_2001
Rac_Plo_DQ886058_N_2003
Rac_Plo_DQ886060_N_2004
Rac_Plo_DQ886059_N_2003
Rac_Plo_DQ886067_N_1994
Rac_Plo_DQ886061_N_2003
Rac_Plo_DQ886065_N_1992
Rac_Plo_DQ886066_N_1992
Rac_Plo_DQ886062_N_1984
Rac_Plo_DQ886063_N_1985
Rac_Plo_DQ886064_N_1989
Sku_Mme_AF483524_N_2001
Dog_Cfa_AB041966_N_1996
Hum_Hsa_AB041965_N_1996
Dog_Cfa_AB178892_N_1985
Dog_Cfa_AB178893_N_1985
Dog_Cfa_AB178896_N_1963
Dog_Cfa_AB178894_N_1985
Dog_Cfa_AB178895_N_1977
Dog_Cfa_AB284509_N_2000
Dog_Cfa_AB284512_N_2005
Dog_Cfa_AB284511_N_2000
Dog_Cfa_AB284510_N_1999
Fox_VvE_EU293115_N_1991
Fox_VvE_U42700_N_1994
Fox_VvE_U42607_N_1994
Fox_VvE_U42605_N_1984
Fox_VvE_U42606_N_1989
Fox_VvE_U43433_N_1974
Fox_VvE_U42706_N_1986
Fox_VvE_U42702_N_1991
Fox_VvE_U42701_N_1991
Hum_Hsa_DQ837382_N_1997
Wol_Clu_DQ837473_N_1997
Wol_Clu_AY352483_N_2004
Wol_Clu_EF614256_N_2006
Fox_VvE_U42703_N_1976
Wol_Clu_DQ837454_N_2005
Fox_VvE_EU886631_N_2001
Fox_VvE_EU886632_N_2002
Fox_VvE_EU886633_N_2004
Fox_VvE_EU886634_N_2004
Hum_Hsa_DQ666288_N_2004
Hum_Hsa_DQ666320_N_2004
Sku_Mme_AF344304_N_1992
Sku_Mme_AF344306_N_1992
Sku_Mme_AF461045_N_1998
Hum_Hsa_AY352462_N_1998
Wol_Clu_AY352486_N_1995
Hum_Hsa_AY352493_N_1988
80.0

En la figura 5 se muestra un esquema de cada árbol, en el que se han incluido aquellos grupos formados por más de dos secuencias con un mismo huésped (razón por la que no se han incluido secuencias cuyo huésped sea el ser humano, pues en ambos árboles están muy dispersas)

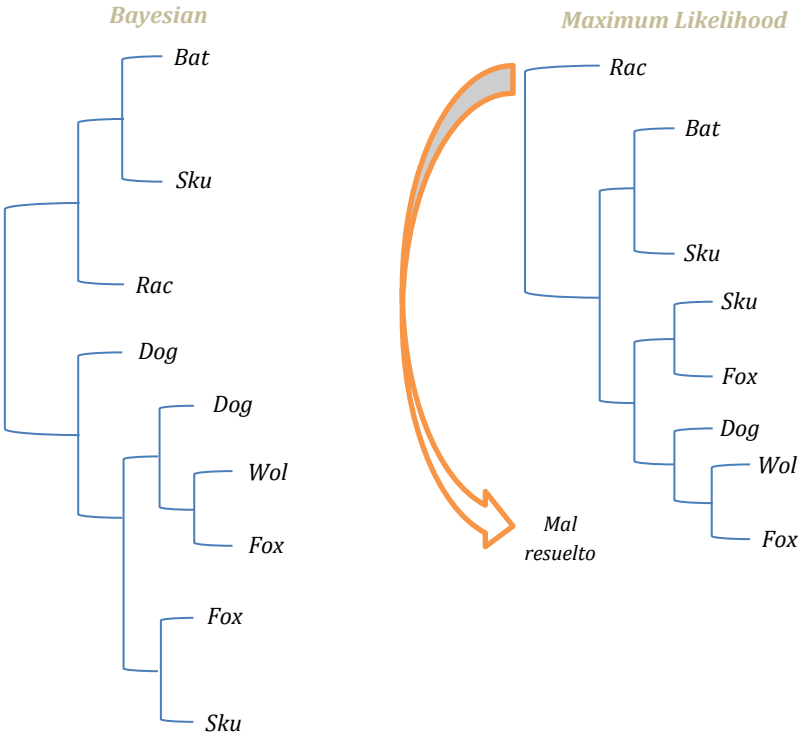


Figura 5

Podemos ver que en ambos casos se obtienen resultados similares; la principal diferencia consiste en que en el árbol obtenido por el método *Maximum Likelihood* presenta una zona (el aquella de la que parten los valores de Rac) con valores de bootstrap muy bajos.

PREGUNTAS

¿En qué huésped tiene el virus mayor diversidad genética?

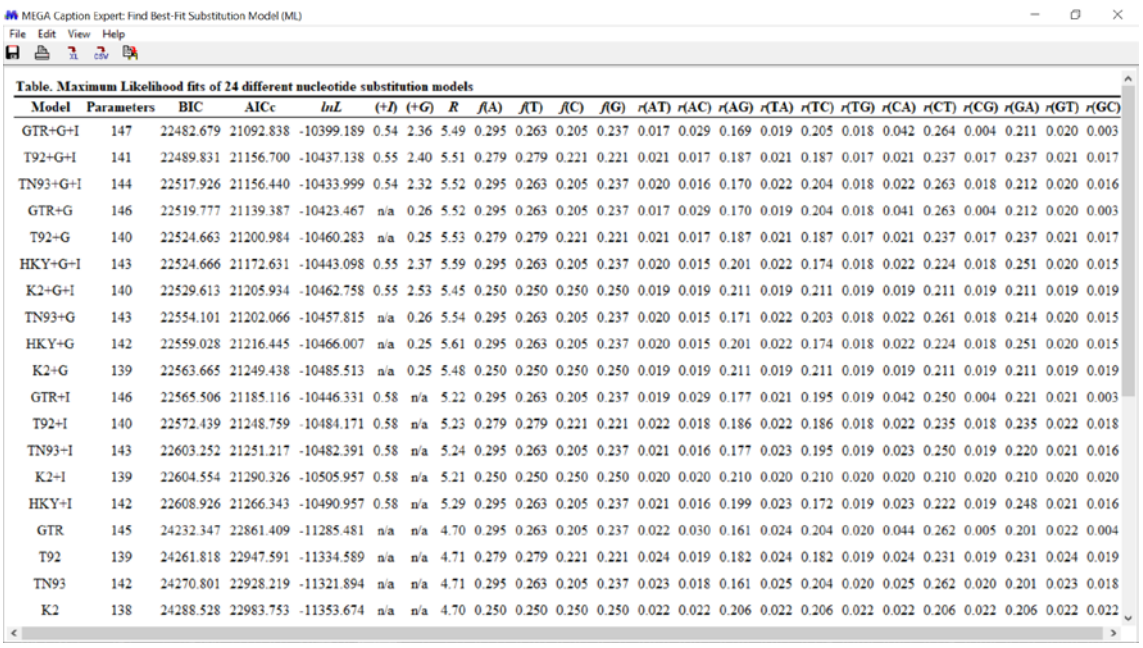
Para contestar a esta pregunta, debemos acudir a MEGA, que nos permite realizar análisis evolutivos como estimar las distancias genéticas en base a distintos modelos.

Para ello, lo primero que hay que hacer es estimar el mejor modelo de sustitución nucleotídica. Simplemente hay que introducir el alineamiento en el programa y accedemos a *Find Best DNA/Protein Models (ML)*.

Los parámetros relevantes que usamos para este análisis son…

- Como árbol guía se utiliza *Neighbor-joining*
- Como método estadístico se utiliza *Maximum likelihood*, que lleva más tiempo que *Neighbor-joining* pero ofrece mejores resultados
- En el alineamiento no hay gaps, por lo que la forma de tratarlos es irrelevante
- El refinamiento del árbol guía se deja en *Moderate*, pues ofrece unos resultados aceptables y no requiere un tiempo excesivo

En la figura 6 se muestran las primeras líneas del resultado obtenido.


MEGA Caption Expert: Find Best-Fit Substitution Model (ML)

File Edit View Help

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	*(+I)*	*(+G)*	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
GTR+G+I	147	22482.679	21092.838	-10399.189	0.54	2.36	5.49	0.295	0.263	0.205	0.237	0.017	0.029	0.169	0.019	0.205	0.018	0.042	0.264	0.004	0.211	0.020	0.003
T92+G+I	141	22489.831	21156.700	-10437.138	0.55	2.40	5.51	0.279	0.279	0.221	0.221	0.021	0.017	0.187	0.021	0.187	0.017	0.021	0.237	0.017	0.237	0.021	0.017
TN93+G+I	144	22517.926	21156.440	-10433.999	0.54	2.32	5.52	0.295	0.263	0.205	0.237	0.020	0.016	0.170	0.022	0.204	0.018	0.022	0.263	0.018	0.212	0.020	0.016
GTR+G	146	22519.777	21139.387	-10423.467	n/a	0.26	5.52	0.295	0.263	0.205	0.237	0.017	0.029	0.170	0.019	0.204	0.018	0.041	0.263	0.004	0.212	0.020	0.003
T92+G	140	22524.663	21200.984	-10460.283	n/a	0.25	5.53	0.279	0.279	0.221	0.221	0.021	0.017	0.187	0.021	0.187	0.017	0.021	0.237	0.017	0.237	0.021	0.017
HKY+G+I	143	22524.666	21172.631	-10443.098	0.55	2.37	5.59	0.295	0.263	0.205	0.237	0.020	0.015	0.201	0.022	0.174	0.018	0.022	0.224	0.018	0.251	0.020	0.015
K2+G+I	140	22529.613	21205.934	-10462.758	0.55	2.53	5.45	0.250	0.250	0.250	0.250	0.019	0.019	0.211	0.019	0.211	0.019	0.019	0.211	0.019	0.211	0.019	0.019
TN93+G	143	22554.101	21202.066	-10457.815	n/a	0.26	5.54	0.295	0.263	0.205	0.237	0.020	0.015	0.171	0.022	0.203	0.018	0.022	0.261	0.018	0.214	0.020	0.015
HKY+G	142	22559.028	21216.445	-10466.007	n/a	0.25	5.61	0.295	0.263	0.205	0.237	0.020	0.015	0.201	0.022	0.174	0.018	0.022	0.224	0.018	0.251	0.020	0.015
K2+G	139	22563.665	21249.438	-10485.513	n/a	0.25	5.48	0.250	0.250	0.250	0.250	0.019	0.019	0.211	0.019	0.211	0.019	0.019	0.211	0.019	0.211	0.019	0.019
GTR+I	146	22565.506	21185.116	-10446.331	0.58	n/a	5.22	0.295	0.263	0.205	0.237	0.019	0.029	0.177	0.021	0.195	0.019	0.042	0.250	0.004	0.221	0.021	0.003
T92+I	140	22572.439	21248.759	-10484.171	0.58	n/a	5.23	0.279	0.279	0.221	0.221	0.022	0.018	0.186	0.022	0.186	0.018	0.022	0.235	0.018	0.235	0.022	0.018
TN93+I	143	22603.252	21251.217	-10482.391	0.58	n/a	5.24	0.295	0.263	0.205	0.237	0.021	0.016	0.177	0.023	0.195	0.019	0.023	0.250	0.019	0.220	0.021	0.016
K2+I	139	22604.554	21290.326	-10505.957	0.58	n/a	5.21	0.250	0.250	0.250	0.250	0.020	0.020	0.210	0.020	0.210	0.020	0.020	0.210	0.020	0.210	0.020	0.020
HKY+I	142	22608.926	21266.343	-10490.957	0.58	n/a	5.29	0.295	0.263	0.205	0.237	0.021	0.016	0.199	0.023	0.172	0.019	0.023	0.222	0.019	0.248	0.021	0.016
GTR	145	24232.347	22861.409	-11285.481	n/a	n/a	4.70	0.295	0.263	0.205	0.237	0.022	0.030	0.161	0.024	0.204	0.020	0.044	0.262	0.005	0.201	0.022	0.004
T92	139	24261.818	22947.591	-11334.589	n/a	n/a	4.71	0.279	0.279	0.221	0.221	0.024	0.019	0.182	0.024	0.182	0.019	0.024	0.231	0.019	0.231	0.024	0.019
TN93	142	24270.801	22928.219	-11321.894	n/a	n/a	4.71	0.295	0.263	0.205	0.237	0.023	0.018	0.161	0.025	0.204	0.020	0.025	0.262	0.020	0.201	0.023	0.018
K2	138	24288.528	22983.753	-11353.674	n/a	n/a	4.70	0.250	0.250	0.250	0.250	0.022	0.022	0.206	0.022	0.206	0.022	0.022	0.206	0.022	0.206	0.022	0.022

Figura 6

Esta estimación se ha realizado porque para hacer los cálculos de distancia y divergencia genéticas es necesario emplear un modelo de sustitución nucleotídica. En la figura 6 aparecen los modelos ordenados en función del criterio de información BIC[3], de valor menor a mayor; según estos parámetros el mejor modelo es GTR+G+I[4]. No obstante, MEGA no permite realizar cálculos con este modelo, por lo que deberemos escoger Tamura 3-parámetros (T92).

[3] No se han demostrado grandes diferencias entre BIC, AICc e InL, por lo que no es muy relevante cuál de los tres criterios se utiliza

[4] + I indica la presencia de sitios invariantes, mientras que + G indica que consideramos heterogeneidad entre sitios

A continuación se divide el alineamiento en 7 grupos, en función del huésped (*Bat -murciélago-, Dog -perro-, Fox -zorro-, Hum -ser humano-, Rac -mapache-, Sku -mofeta-* y *Wol -lobo-*), pues se nos pide cuál es el grupo que tiene mayor divergencia genética en función del huésped.

En el enunciado se pide calcular el grupo con mayor diversidad genética; sin embargo, MEGA no permite calcular la diversidad genética dentro cada grupo. Sí permite, no obstante, calcular la distancia genética dentro de cada grupo.

Para calcular todos los valores de distancia y divergencia usaremos los siguientes parámetros…

- Modelo de sustitución nucleotídica: como ya se ha explicado, se utilizará el *Tamura 3-parameter model*
- El valor de Gamma obtenido (2'4)
- Para la estimación de la varianza utilizamos el método *Bootstrap*, con 1.000 replicaciones.
- El patrón entre linajes lo situamos en *Different (heterogeneous)*, pues las distintas secuencias pueden tener distintas velocidades de cambio

Distancia y diversidad coinciden cuando cada genotipo esté representado por una única secuencia. Para comprobar si son lo mismo, calculamos…

- *Overall Mean Distance*: calcula la media de distancias entre todas las secuencias
- *Mean Diversity in Entire Population*: calcula la media de las diversidades entre todas las secuencias

En ambos casos el valor es de 0'153; además la desviación estándar es la misma (0'008), lo que indica que la distancia entre grupos es igual que la diversidad entre grupos.

Por otro lado, también calculamos…

- *Within Mean Group Distance* (figura 7): calcula las medias de la distancia dentro de cada grupo
- *Mean Diversity Within Subpopulation*: calcula la media de las divergencias dentro de cada grupo

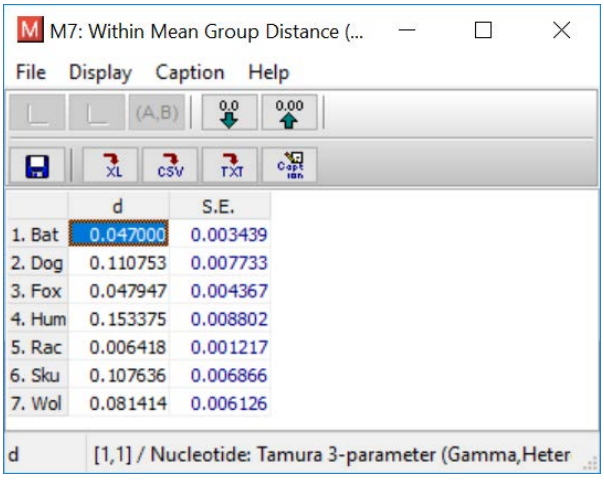


Figura 7

Si hacemos la media de las distancias mostradas en la figura 7 obtenemos un valor de 0'079220, que es exactamente el que obtenemos de calcular la media de las divergencias dentro de cada grupo.

Por tanto, siendo ambos parámetros iguales, se puede concluir que distancia es igual que divergencia en este caso, y por tanto se pueden tomar las medias de la distancia genética dentro de cada grupo (figura 7) como las medias de la diversidad genética dentro de cada grupo. Se puede observar que **el grupo que presenta una mayor diversidad genética es el que tiene como huésped a la raza humana** (0'153 de media).

La explicación biológica podría ser que el virus de la rabia no se puede transmitir entre seres humanos, pues éstos no son su huésped natural; por ello no hay una única cepa del virus que llegue a seres humanos, sino que llegan de distintos organismos; esto podría explicar la gran variación genética.

¿Se estructura genéticamente la población del virus en función del huésped?

Se puede decir que el huésped del virus afecta a la estructura genética del mismo cuando…

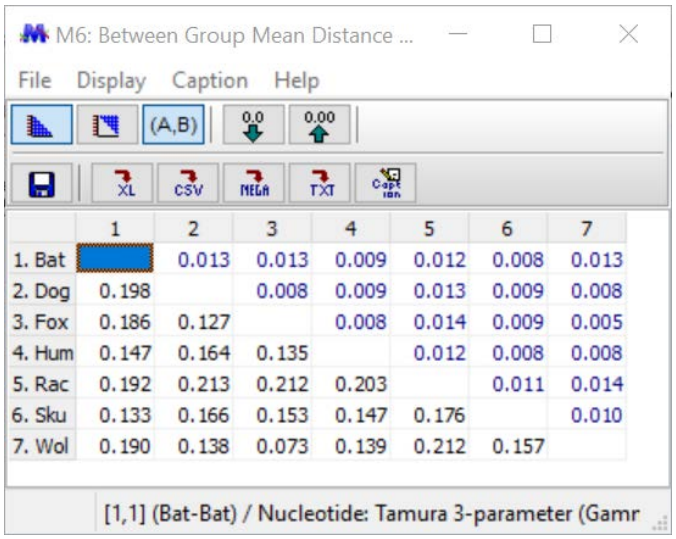
- La distancia genética dentro de los virus que comparten un mismo huésped es pequeña
- La distancia genética entre los virus con distinto huésped es grande

Mega da la opción de calcular el coeficiente de diferenciación, que se calcula como la media de la diversidad entre los grupos (*Mean Interpopulation Diversity*) entre la media de las diversidades dentro de cada grupo (*Mean Diversity in Entire Population*)…

$$Coeficiente\ de\ diferenciación = \frac{0'074}{0'153} = 0'482$$

Cuanto más cercano esté el valor a 1 mayor será la influencia del huésped sobre la estructura genética; este valor, sin embargo, no indica que exista una influencia significativa, pues la diversidad entre grupos y de cada grupo son prácticamente iguales (0'074 y 0'079, respectivamente).

Para tener una mayor certeza de que el huésped, efectivamente, no influye en la estructura genética, se puede comparar cada grupo por separado Para ello comparamos la diversidad dentro de cada grupo (figura 7) con la media de las distancias de ese grupo con el resto, que se calculan con la *Between Group Mean Distance* (figura 8); podemos ver la comparación en la figura 9.



	1	2	3	4	5	6	7
1. Bat		0.013	0.013	0.009	0.012	0.008	0.013
2. Dog	0.198		0.008	0.009	0.013	0.009	0.008
3. Fox	0.186	0.127		0.008	0.014	0.009	0.005
4. Hum	0.147	0.164	0.135		0.012	0.008	0.008
5. Rac	0.192	0.213	0.212	0.203		0.011	0.014
6. Sku	0.133	0.166	0.153	0.147	0.176		0.010
7. Wol	0.190	0.138	0.073	0.139	0.212	0.157	

Figura 8

En la figura 9 se muestra cómo, efectivamente, **algunos grupos sí parece que podrían estar influenciados por el huésped** (el caso más evidente es el del mapache).

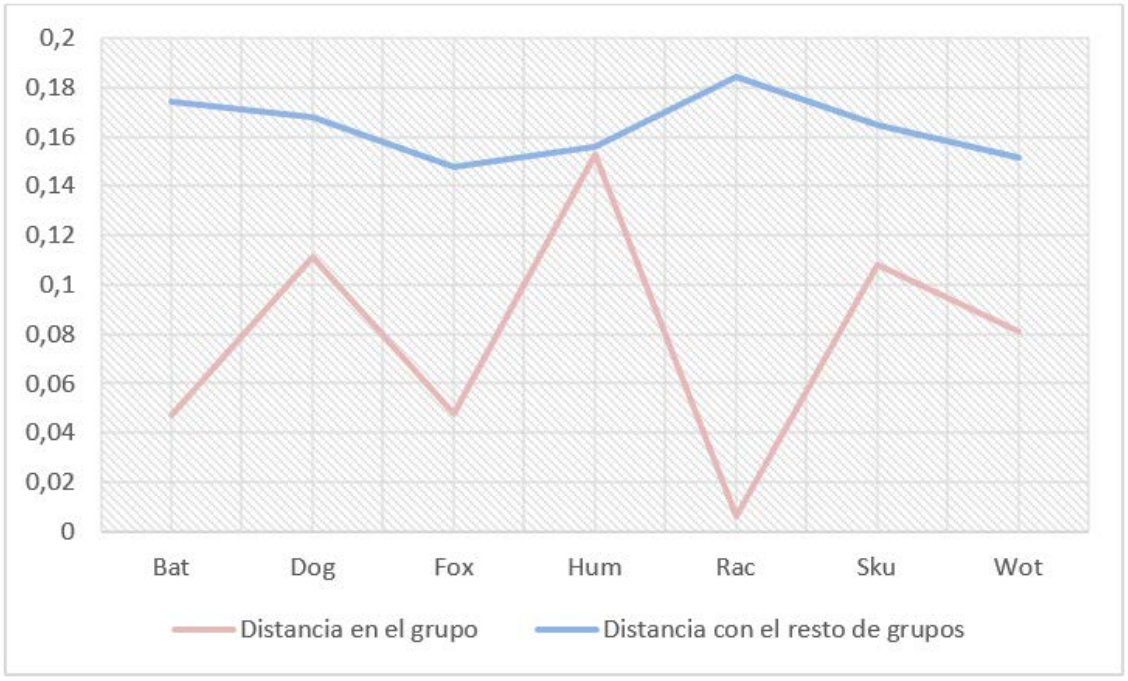


Figura 9

Este resultado se confirma observando los árboles filogenéticos obtenidos; ilustrando dos ejemplos…

- En el árbol obtenido por estadística bayesiana (página 6), se puede ver cómo todas las secuencias cuyo huésped es el mapache se encuentran juntas (además con un valor *a posteriori* de 1)
- En ambos árboles se observa una dispersión de las secuencias con huésped en la raza humana, que en la figura s9 e puede ver que es la que presenta una menor relación entre la estructura genética y el huésped

¿De qué huésped son las secuencias que se parecen más al virus de la rabia que infecta a los murciélagos?

Para responder a esta pregunta se puede, de nuevo, acudir tanto a Mega como a los árboles filogenéticos.

En Mega se calcula la *Between Group Mean Distance* (figura 8), y se observa que **el grupo que más se parece genéticamente al de murciélagos es el de los mapaches** (0'133), seguido del de seres humanos (0'147).

Efectivamente, en los árboles filogenéticos se puede sacar la misma conclusión (se puede ver rápidamente en la figura 5).

¿Podrías decir cuál es el tiempo del ancestro común más reciente, y la tasa de sustitución nucleotídica del virus de la rabia?

Se utilizan los archivos del árbol realizado con BEAST2, pues en el caso del árbol realizado con PhyML los primeros nodos no están bien definidos.

Para calcular estos parámetros se puede acceder de nuevo a Tracer. La edad del ancestro común más reciente se calcula en *TreeHeight*…

$$Media: 763'6961\ años$$

$$Mediana: 716'1233\ años$$

$$Intervalo\ HPD\ al\ 95\%: [398'907,1227'0951]$$

El *Effective Sample Size* está en casi 500, por lo que es un valor bueno; Además, se puede ver en la figura 10 que la representación gráfica se asemeja a una distribución normal, por lo que la media estimada es buena.

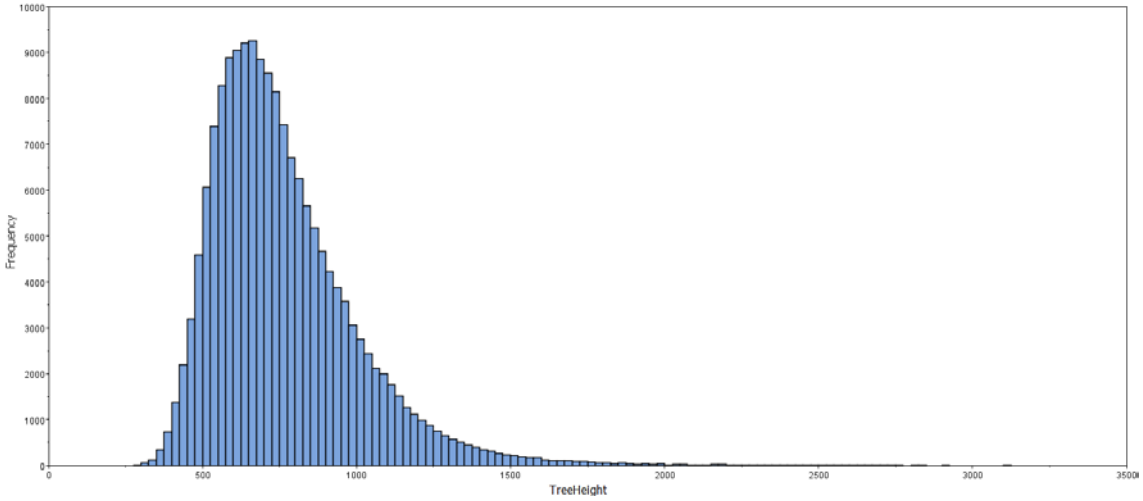


Figura 10

Otra opción para calcular la edad del nodo más antiguo en TempEst; en este caso se indica que **la edad del nodo más antiguo es de aproximadamente 716 años**, lo que coincide con la mediana obtenida en Tracer. Además se puede dibujar un árbol que contenga la edad de cada nodo (figura 11).

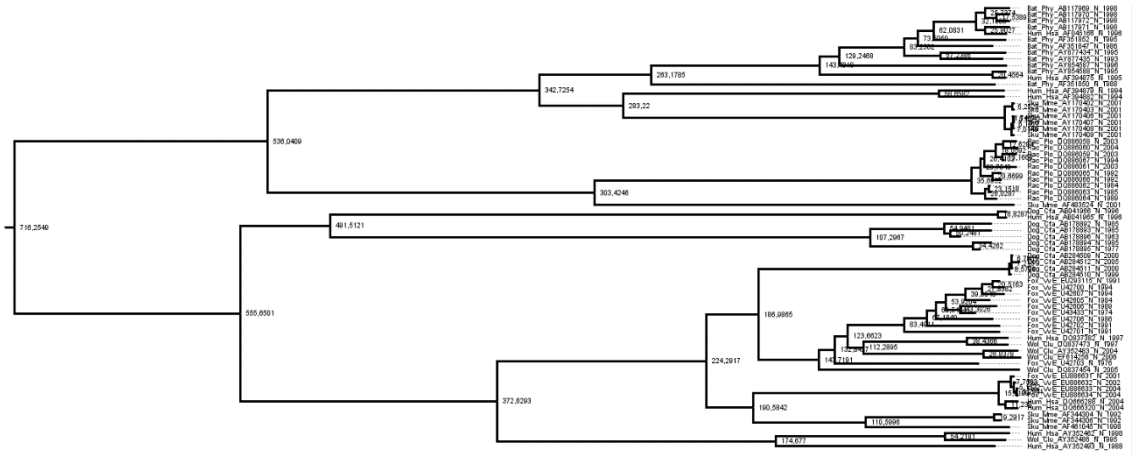


Figura 11

La tasa de sustitución nucleotídica también se calcula en Tracer, en la pestaña *rate.Mean*. Los valores obtenidos son…

$$\boldsymbol{Media}: \mathbf{2'2895 * 10^{-4}}$$

$$Mediana: 2'2636 * 10^{-4}$$

$$Intervalo\ HPD\ al\ 95\%: [1'2002 * 10^{-4}, 4609 * 10^{-4}]$$

El valor ESS es 469, y la representación gráfica (figura 12) se asemeja enormemente a una distribución normal, por lo que se puede asumir que la estimación es buena.

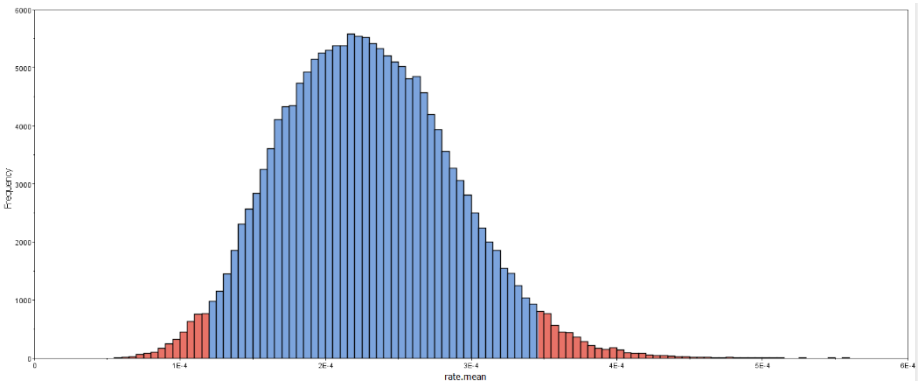


Figura 12

¿En qué huésped el virus está bajo una presión de selección negativa más fuerte?

Se divide el archivo que contiene las 70 secuencias en 7 archivos, cada uno con las secuencias que comparten huésped. A continuación se realizan los árboles filogenéticos en MEGA (*Maximum Likelihood*), pues el mejor modelo estimado es Tamura 3 parámetros (+G). Una vez se han obtenido los alineamientos, los introducimos (uno a uno) en HyPhy para calcular las presiones de selección.

Para ello se lleva a cabo un análisis de *Selection/Recombination*, y se escoge la opción *Quick Selection Detection*. Se introduce en *Custom* el código del modelo tamura 3-parámetros. Finalmente se introducen los alineamientos y sus árboles. Los valores de dN/dS obtenidos son…

Bat: 0'05341

Dog: 0'01645

Fox: 0'07084

Hum: 0'02564

Rac: 0'03742

Sku: 0'02861

Wol: 0'04654

Las secuencias que están sometidas a una mayor presión de selección negativa (valor más bajo de dN/dS) son aquellas cuyo huésped es el perro. Una explicación podría ser que, al estar actualmente la rabia en perros muy controlada (pues son animales de compañía que conviven con los seres humanos), de forma que la transmisión ha disminuido mucho y por tanto que una mutación perdure en el tiempo es complicado.