

Pablo Vinuesa (vinuesa[at]ccg{.}unam{.}mx)

Programa de Ingeniería Genómica, CCG-UNAM, México
<http://www.ccg.unam.mx/~vinuesa/>

material del curso (presentaciones, tutoriales y datos de secuencias) lo encontraron en
<https://github.com/vinuesa/TIB-filoinfo>

Tutorial sobre uso de modeltest3.7 y jmodeltest2

github.com/ddarriba/jmodeltest2

Basic frequencies

Frequency	Rate variation
A	1.0
C	1.0
G	1.0
T	1.0

Rate variation

Rate variation	Frequency
A	0.25
C	0.25
G	0.25
T	0.25

Para ello **necesita que PAUP* o PhyML calculen los -lnL scores** de un subconjunto (56) de todos los posibles modelos de la familia GTR (203). Estos scores de -lnL se calculan corriendo un "batch file" de comandos PAUP*. Lo primero que se estima es un árbol (rápido) NJ-JC69. Se usa la topología resultante para evaluar los distintos modelos y obtener estimas de ML de los parámetros correspondientes.

The figure displays five screenshots of the iModelTest 2.1.7 software interface:

- Top-left (About):** Shows the version (2.1.7, 20161120) and authors: C. G. Taboada, D. Doallo, and D. Posada. It lists affiliations: Department of Biochemistry, Genetics and Biophysics, University of Vigo; Department of Electronics and Systems, University of Girona; and a personal email address.
- Top-right (Bayesian Information Criterion (BIC) Settings):** A window for configuring BIC calculations. It includes options for 'Calculate parameter Importances' (checked), 'Model averaging' (unchecked), and 'Write PAUP-Block' (unchecked). It also shows a 'Confidence Interval' of 100%.
- Bottom-left (File Explorer):** A screenshot of a file explorer showing the directory structure of the iModelTest 2.1.7 application, including folders like 'Carpets' and 'Archivos'.
- Bottom-right (Progress):** A window showing the progress of the calculation. It indicates that 56 models were completed. Below this, it shows the 'BIC MODEL SELECTION' results for 56 models, including a table of model selection results.

#=====

```
phym1 -i primate-mtDNA-interleaved.phy -d nt -n 1 -b 0 -m 010010 -f m -c 4 -a e --no_memory_check -o lrt
```

```
# genera el un directorio de trabajo para este ejercicio
mkdir practica_jmodeltest && cd practica_jmodeltest

# haz liga simbolica al set de datos primate-mtDNA-interleaved1.phy en este directorio de trabajo
ln -s /home/vinuesa/intro2bioinfo/seq_data/primate-mtDNA-interleaved1.phy .

# explora el archivo
less primate-mtDNA-interleaved1.phy

# corre jmodeltest2 con los parametros abajo indicados
# -d datos
# -i usa modelos que asumen una proporcion de sitios invariantes
# -f usa modelos que asumen diferentes frecuencias de bases
# -g usa distribucion gamma con 4 clases discretas de tasas para modelar
#      la heterogeneidad de tasas de sustitucion intrinsecos
# -AIC usa criterio de informacion de Akaike para la seleccion de modelos

jmodeltest -d primate-mtDNA-interleaved1.phy -i -f -g 4 -AIC

# Ahora corremos phyml bajo el mejor modelo seleccionado
phyml -i primate-mtDNA-interleaved1.phy -d nt -m 010010 -b -4 -f e -c 4 -a e --no_memory_check -o tlr -s
BEST
```

- interpretación de la salida de modeltest: **1. hLRTs**

* ----- HIERARCHICAL LIKELIHOOD RATIO TESTS (hLRTs) ----- *			
Confidence level = 0.01			
Equal base frequencies			
Null model = JC	-lnL0 = 6424.2026		
Alternative model = F81	-lnL1 = 6284.9956		
2(lnL1-lnL0) = 278.4141		df = 3	
P-value = <0.000001			
Ti=Tv			
Null model = F81	-lnL0 = 6284.9956		
Alternative model = HKY	-lnL1 = 5981.7202		
2(lnL1-lnL0) = 606.5508		df = 1	
P-value = <0.000001			
Equal Ti rates			
Null model = HKY	-lnL0 = 5981.7202		
Alternative model = TrN	-lnL1 = 5978.8550		
2(lnL1-lnL0) = 5.7305		df = 1	
P-value = 0.016673			
Equal Tv rates			
Null model = HKY	-lnL0 = 5981.7202		
Alternative model = K81uf	-lnL1 = 5973.2393		
2(lnL1-lnL0) = 16.9619		df = 1	
P-value = 0.000038			

(continúa en la siguiente página)

- interpretación de la salida de modeltest: **1. hLRTs** (Continuación)

```
Only two Tv rates
Null model = K81uf          -lnL0 = 5973.2393
Alternative model = TVM      -lnL1 = 5938.5615
2(lnL1-lnL0) = 69.3555      df = 2
P-value = <0.000001

Equal rates among sites
Null model = TVM            -lnL0 = 5938.5615
Alternative model = TVM+G    -lnL1 = 5709.6323
2(lnL1-lnL0) = 457.8584     df = 1
Using mixed chi-square distribution
P-value = <0.000001

No Invariable sites
Null model = TVM+G          -lnL0 = 5709.6323
Alternative model = TVM+I+G  -lnL1 = 5709.6323
2(lnL1-lnL0) = 0.0000       df = 1
Using mixed chi-square distribution

P-value = >0.999999 es decir, no rechazo la H0 !!! El modelo seleccionado es TVM+G
```

- interpretación de la salida de modeltest: **1. hLRTs** (Continuación)

```
Model selected: TVM+G
-lnL = 5709.6323
K = 8
Base frequencies:
freqA = 0.3581
freqC = 0.3186
freqG = 0.0846
freqT = 0.2387
Substitution model:
Rate matrix
R(a) [A-C] = 3.9989
R(b) [A-G] = 40.5788
R(c) [A-T] = 3.4119
R(d) [C-G] = 2.3909
R(e) [C-T] = 40.5788
R(f) [G-T] = 1.0000
Among-site rate variation
Proportion of invariable sites = 0
Variable sites (G)
Gamma distribution shape parameter = 0.3752
```

-interpretación de la salida de modeltest: **2. AIC** = $-2 \ln L + 2 K$; Akaike 1974
(cantidad de información perdida cuando la realidad es aproximada por un modelo)

```
-----
*          AKAIKE INFORMATION CRITERION (AIC)
*
-----
Model selected: TrN+G
-lnL = 5710.5513
K = 6
AIC = 11433.1025

Base frequencies:
freqA = 0.3581
freqC = 0.3252
freqG = 0.0765
freqT = 0.2402
Substitution model:
Rate matrix
R(a) [A-C] = 1.0000
R(b) [A-G] = 16.0043
R(c) [A-T] = 1.0000
R(d) [C-G] = 1.0000
R(e) [C-T] = 11.6796
R(f) [G-T] = 1.0000
Among-site rate variation
Proportion of invariable sites = 0
Variable sites (G)
Gamma distribution shape parameter = 0.3566
```

- interpretación de la salida de modeltest: **2. AIC** (continuación)

PAUP* Commands Block: If you want to implement the previous estimates as likelihod settings in PAUP*, attach the next block of commands after the data in your PAUP file:

```
[!
Likelihood settings from best-fit model (TrN+G) selected by
AIC in Modeltest 3.7 on Sat May 20 17:12:56 2006
]

BEGIN PAUP;
Lset Base=(0.3581 0.3252 0.0765) Nst=6 Rmat=(1.0000
16.0043 1.0000 1.0000 11.6796) Rates=gamma Shape=0.3566
Pinvar=0;
END;
```

- interpretación de la salida de modeltest: **2. AIC** (continuación)

* MODEL SELECTION UNCERTAINTY : Akaike Weights

Model	-lnL	K	AIC	delta	weight	cumWeight
TrN+G	5710.5513	6	11433.1025	0.0000	0.2463	0.2463
HKY+G	5711.9385	5	11433.8770	0.7744	0.1672	0.4135
TIM+G	5710.4355	7	11434.8711	1.7686	0.1017	0.5152
TrN+I+G	5710.5513	7	11435.1025	2.0000	0.0906	0.6058
TVM+G	5709.6323	8	11435.2646	2.1621	0.0835	0.6894
K81uf+G	5711.8125	6	11435.6250	2.5225	0.0698	0.7591
GTR+G	5708.9224	9	11435.8447	2.7422	0.0625	0.8217
HKY+I+G	5711.9385	6	11435.8770	2.7744	0.0615	0.8832
TIM+I+G	5710.4355	8	11436.8711	3.7686	0.0374	0.9206
TVM+I+G	5709.6323	9	11437.2646	4.1621	0.0307	0.9513
K81uf+I+G	5711.8125	7	11437.6250	4.5225	0.0257	0.9770
GTR+I+G	5708.9224	10	11437.8447	4.7422	0.0230	1.0000

intervalo de credibilidad del 95 %

- interpretación de la salida de modeltest: **2. AIC** (continuación)

* MODEL AVERAGING AND PARAMETER IMPORTANCE (using Akaike Weights)
Including all 56 models (índices normalizados y relativos de Akaike)

Parameter	Importance	Model-averaged estimates
fA	1.0000	0.3596
fC	1.0000	0.3223
fG	1.0000	0.0794
fT	1.0000	0.2387
TiTv	0.2287	5.4113
rAC	0.1998	3.7999
rAG	0.5615	19.9668
rAT	0.1998	3.2371
rCG	0.1998	2.3657
rCT	0.5615	14.9960
pinv(I)	0.0000	0.3717
alpha(G)	0.7311	0.3621
pinv(IG)	0.2689	0.0000
alpha(IG)	0.2689	0.3621

Values have been rounded.
(I): averaged using only +I models.
(G): averaged using only +G models.
(IG): averaged using only +I+G models.

- Interpretación de la importancia de parámetros
1.

los params. de frec. son un componenete esencial del modelo
2.

Ti/Tv también es significativa
3.

El pto. 2 se ratifica en la import. de rAG y rCT respecto a tasas de Tv
4.

El parámetro alpha (uso de distrib. gamma) es mucho más imp. que asumir sólo pinv.

Modelos de base evaluados por Modeltest

Table 1. Model names. Some models have no reference (TNef, K81uf, TIMef, TIM, TVMef, TVM), they are just some variations of some existing models, and they were no developed, only named, by D. Posada.

Model	Name
JC	Jukes and Cantor (Jukes and Cantor, 1969)
F81	Felsenstein 81 (Felsenstein, 1981)
K80	Kimura 80 (=K2P) (Kimura, 1980)
HKY	Hasegawa, Kishino, Yano 85 (Hasegawa, Kishino and Yano, 1985)
TNef	Tamura-Nei equal frequencies
TN	Tamura-Nei (Tamura and Nei, 1993)
K81	Two transversion-parameters model 1 (=K81=K3P) (Kimura, 1981)
K81uf	Two transversion-parameters model 1 unequal frecuencies
TIMef	Transitional model equal frequencies
TIM	Transitional model
TVMef	Transversional model equal frequencies
TVM	Transversional model
SYM	Symmetrical model (Zharkikh, 1994)
GTR	General time reversible (=REV) (Tavaré, 1986)

Modelos de base evaluados por Modeltest

Table 2. Model parameters. The substitution codes are just two ways of indicating the substitution scheme. Any of these models can ignore rate variation or include invariable sites (+I), rate variation among sites (+G), or both (+I+G).

Model	Free parameters	Base frequencies	Substitution rates	Substitution code 1	Substitution code 2
JC	0	equal	a=b=c=d=e=f	000000	aaaaaa
F81	3	unequal	a=b=c=d=e=f	000000	aaaaaa
K80	1	equal	a=c=d=f, b=e	010010	abaaba
HKY	4	unequal	a=c=d=f, b=e	010010	abaaba
TNef	2	equal	a=c=d=f, b, e	010020	abaaca
TN	5	unequal	a=c=d=f, b, e	010020	abaaca
K81	2	equal	a=f, c=d, b=e	012210	abccba
K81uf	5	unequal	a=f, c=d, b=e	012210	abccba
TIMef	3	equal	a=f, c=d, b, e	012230	abccda
TIM	6	unequal	a=f, c=d, b, e	012230	abccda
TVMef	4	equal	a, c, d, f, b=e	012314	abcdbe
TVM	7	unequal	a, c, d, f, b=e	012314	abcdbe
SYM	5	equal	a, c, d, f, b, e	012345	abcdef
GTR	8	unequal	a, c, d, f, b, e	012345	abcdef