INFERENCIA FILOGENÉTICA USANDO MODELOS EVOLUTIVOS

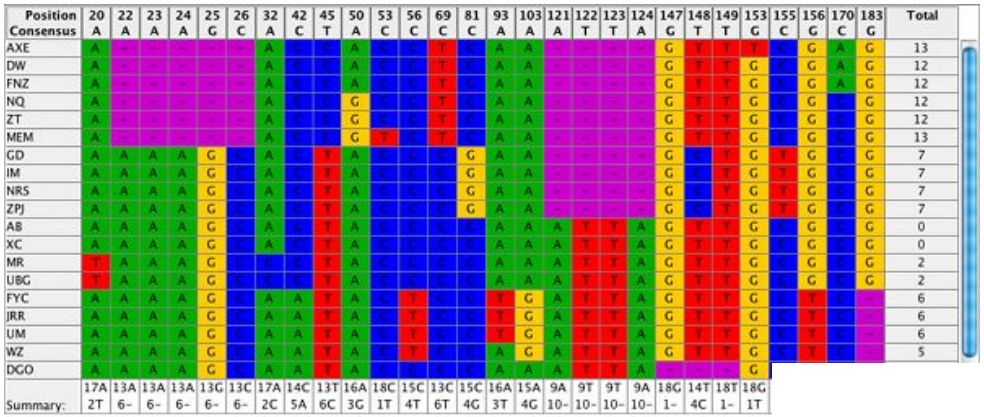
¿QUÉ ES UN MODELO?

Descripción de un sistema o proceso usando un lenguaje matemático para poder hacer predicciones

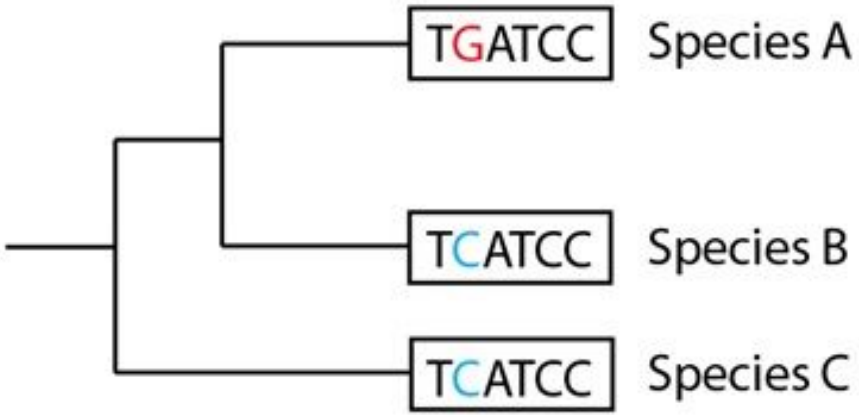
- Parámetros (variables)
- Función matemática

¿QUÉ ES UN MODELO DE SUSTITUCIÓN DE CARACTERES?

Modelos matemáticos que predicen como los caracteres evolucionan entre sus estados y tasas relativas de cambio en las ramas de un árbol



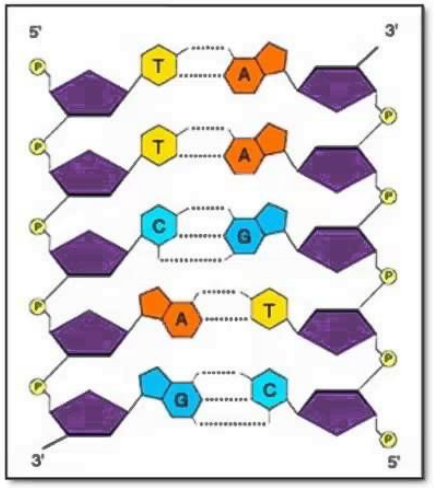
Position	20	22	23	24	25	26	32	42	45	50	53	56	69	81	93	103	121	122	123	124	147	148	149	153	155	156	170	183	Total
Consensus	A	A	A	A	G	C	A	C	T	A	C	C	C	C	A	A	A	T	T	A	G	T	T	G	C	G	C	G	
AXE	A	-	-	-	-	-	A	-	-	A	-	-	T	-	A	A	-	-	-	-	G	T	T	T	-	G	A	G	13
DW	A	-	-	-	-	-	A	C	-	A	-	-	T	-	A	A	-	-	-	-	G	T	T	G	-	G	A	G	12
FNZ	A	-	-	-	-	-	A	-	-	A	-	-	T	-	A	A	-	-	-	-	G	T	T	G	-	G	A	G	12
NQ	A	-	-	-	-	-	A	C	-	G	-	-	T	-	A	A	-	-	-	-	G	T	T	G	-	G	-	G	12
ZT	A	-	-	-	-	-	A	C	-	G	-	-	T	-	A	A	-	-	-	-	G	T	T	G	-	G	C	G	12
MEM	A	-	-	-	-	-	A	C	-	G	T	-	T	-	A	A	-	-	-	-	G	T	T	G	-	G	C	G	13
GD	A	A	A	A	G	-	A	-	T	A	-	-	-	G	A	A	-	-	-	-	G	-	T	G	T	G	-	G	7
IM	A	A	A	A	G	-	A	C	T	A	-	-	-	G	A	A	-	-	-	-	G	-	T	G	T	G	C	G	7
NRS	A	A	A	A	G	-	A	C	T	A	-	-	-	G	A	A	-	-	-	-	G	-	T	G	T	G	C	G	7
ZPJ	A	A	A	A	G	-	A	-	T	A	-	-	-	G	A	A	-	-	-	-	G	-	T	G	T	G	-	G	7
AB	A	A	A	A	G	-	A	-	T	A	-	-	-	-	A	A	A	T	T	A	G	T	T	G	-	G	-	G	0
XC	A	A	A	A	G	-	A	C	T	A	-	-	-	-	A	A	A	T	T	A	G	T	T	G	C	G	C	G	0
MR	T	A	A	A	G	-	C	C	T	A	-	-	-	-	A	A	A	T	T	A	G	T	T	G	-	G	-	G	2
UBG	T	A	A	A	G	-	T	-	T	A	-	-	-	-	A	A	A	T	T	A	G	T	T	G	C	G	C	G	2
FYC	A	A	A	A	G	-	A	A	T	A	-	T	-	-	T	G	A	T	T	A	G	T	T	G	-	T	-	-	6
JRR	A	A	A	A	G	-	A	A	T	A	-	T	-	-	T	G	A	T	T	A	G	T	T	G	C	T	C	-	6
UM	A	A	A	A	G	-	A	A	T	A	-	T	-	-	T	G	A	T	T	A	G	T	T	G	C	T	-	-	6
WZ	A	A	A	A	G	-	A	A	T	A	-	T	-	-	A	G	A	T	T	A	G	T	T	G	-	T	-	-	5
DGO	A	A	A	A	G	-	A	A	T	A	-	-	-	-	A	A	A	T	T	A	-	-	-	G	-	-	-	-	
Summary:	17A 2T	13A 6-	13A 6-	13A 6-	13G 6-	13C 6-	17A 2C	14C 5A	13T 6C	16A 3G	18C 1T	15C 4T	13C 6T	15C 4G	16A 3T	15A 4G	9A 10-	9T 10-	9T 10-	9A 10-	18G 1-	14T 4C	18T 1-	18G 1T					

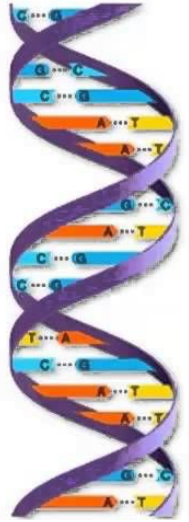


Homología en secuencias de ADN

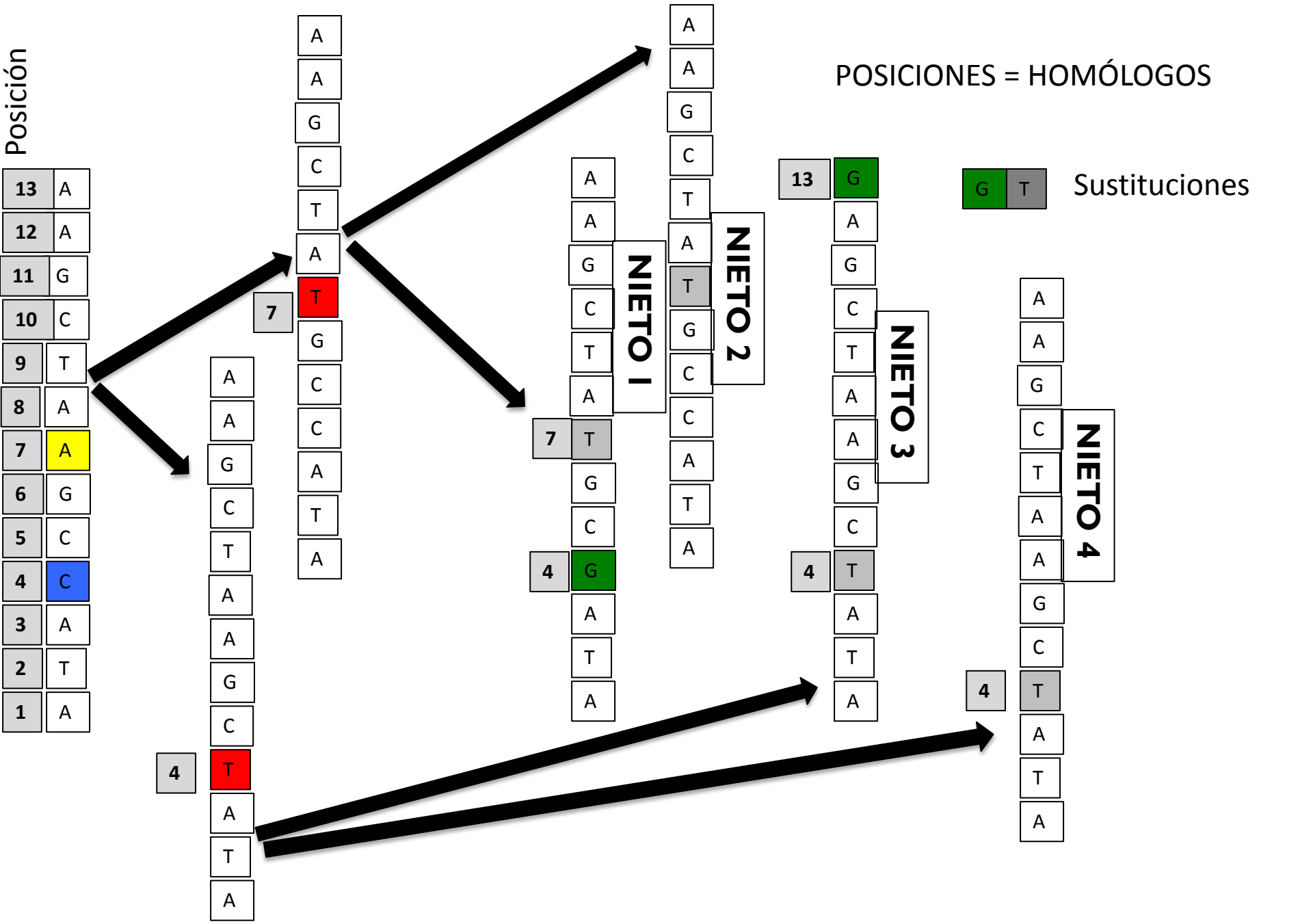
Alineamiento

Scarites	C	T	T	A	G	A	T	C	G	T	A	C	C	A	A	-	-	-	A	A	T	A	T	T	A	C
Carenum	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	A	-	T	A	C	-	T	T	T	A	C
Pasimachus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	T	A	T	A	A	G	T	T	T	A	C
Pheropsophus	C	T	T	A	G	A	T	C	G	T	T	C	C	A	C	-	-	-	A	C	A	T	A	T	A	C
Brachinus armiger	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	T	C
Brachinus hirsutus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	A	C
Aptinus	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	C	A	A	T	T	A	C
Pseudomorpha	C	T	T	A	G	A	T	C	G	T	A	C	C	-	-	-	-	-	A	C	A	A	A	T	A	C





Evolución de secuencias de ADN



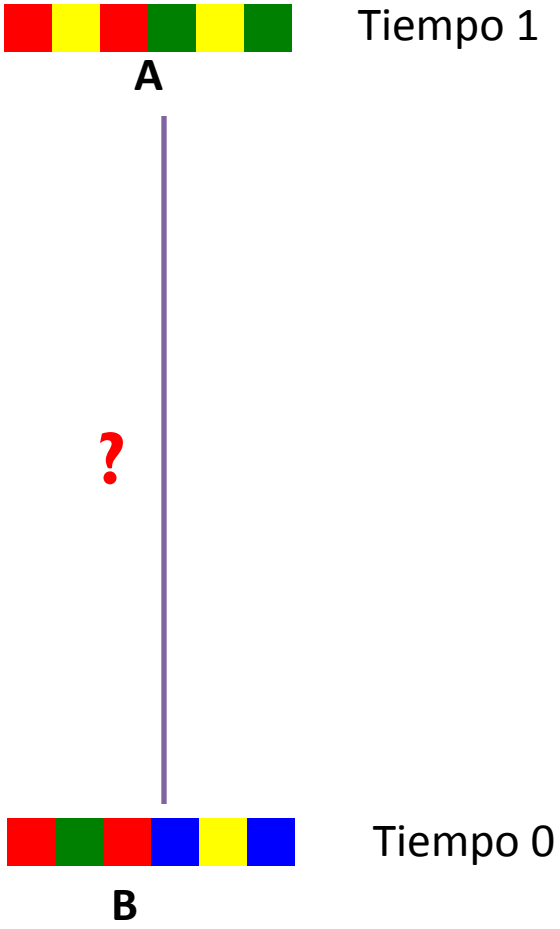
SUPUESTOS DE LOS PARÁMETROS DE LOS MODELOS

- **Propiedad de Markov:** La probabilidad de sustitución de un caracter depende únicamente del estado presente (independiente de estados anteriores)
- **Homegeneidad:** Las tasas de sustitución en un sitio no cambian con el tiempo
- **Estacionalidad:** Las frecuencias relativas de los estados de caracter están en equilibrio
- Las tasas de cambio son reversibles en el tiempo

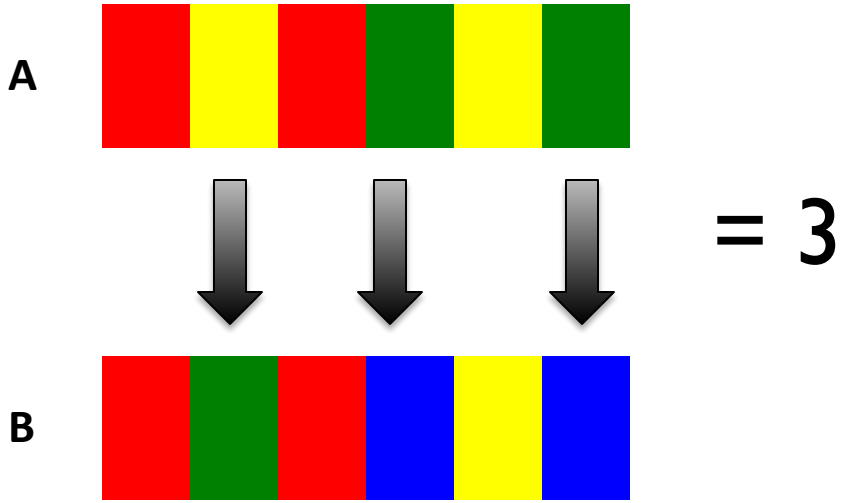
SUPUESTOS DE EVOLUCIÓN DE LOS CARACTERES

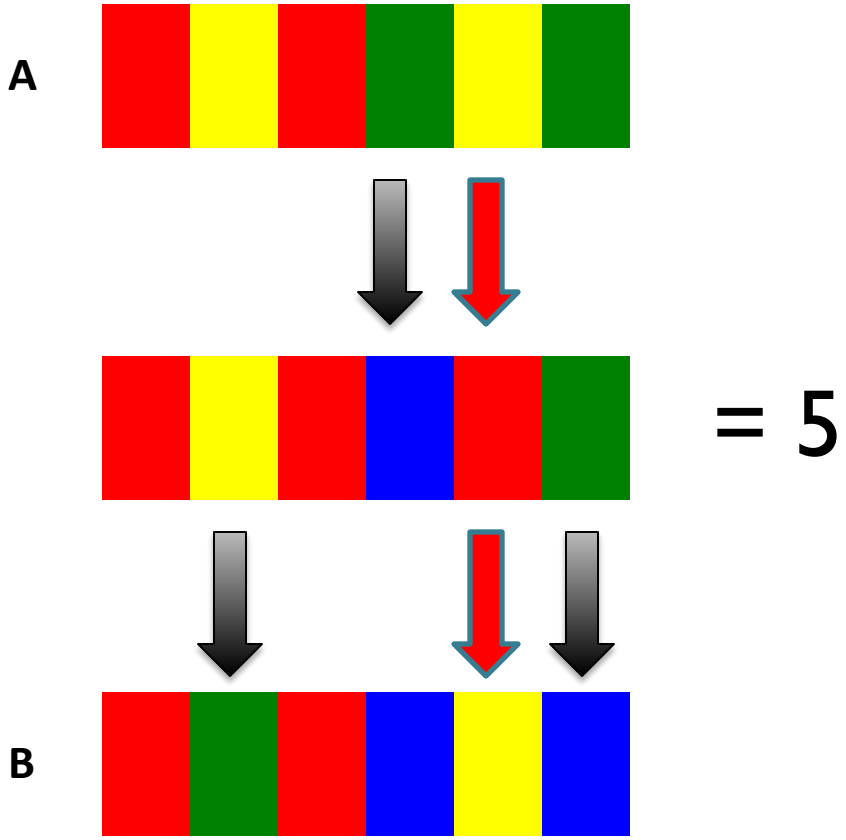
- Caracteres son neutrales
- Caracteres evolucionan independientemente
- Caracteres cambian a un número finito de estados
- Caracteres evolucionan a lo largo de las ramas
- Las ramas tienen una duración específica

Distancia evolutiva



Promedio de sustituciones que han ocurrido en cada caracter


A
B
= 3


A
= 5
B

¿Cómo estimar el promedio de cambios que hubo entre las dos secuencias observadas a través del tiempo?

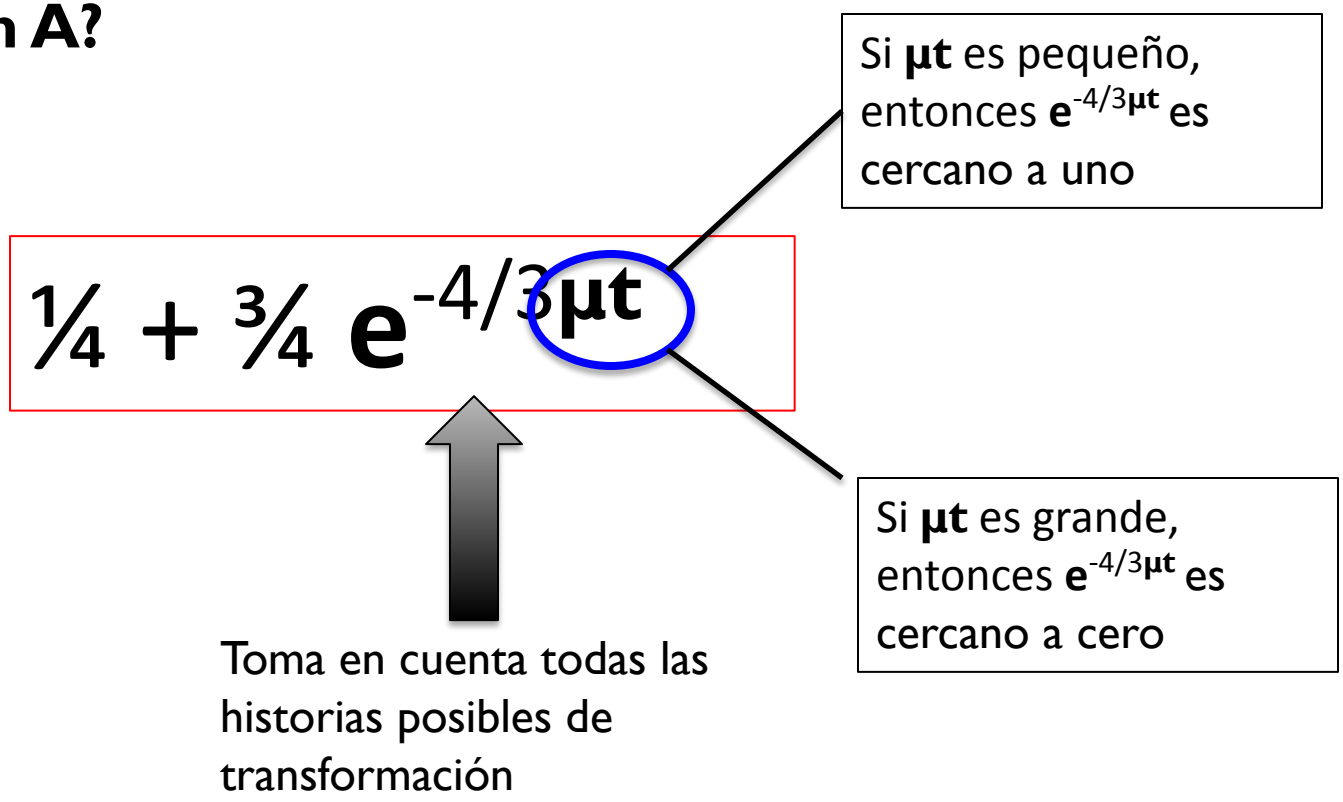
Calculando la probabilidad de observar cada cambio en función de:

- Tasa de sustitución: μ
- Tiempo: t

MATRIZ DE PROBABILIDAD DE SUSTITUCIÓN

		To:			
		A	C	G	T
From:	A	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	C	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	G	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	T	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

¿Cuál es la probabilidad de comenzar con A y terminar en A?



MATRIZ DE PROBABILIDAD DE SUSTITUCIÓN

Jukes-Cantor (1969) = JC69

		To:			
		A	C	G	T
From:	A	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	C	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	G	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	T	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

- Cuatro bases son igualmente frecuentes
- Los tipos de sustituciones ocurren a la misma tasa
- La tasa de sustitución en igual a través de todos los sitios de la secuencia

Pero sabemos que:

- Las cuatro bases generalmente no se encuentran con la misma frecuencia
- Algunos tipos de sustitución ocurren a diferentes tasa que otros (Ts vs. Tv)
- Algunas posiciones de la secuencia evolucionan a tasa más rápidas que otras

MODELOS MÁS REALISTAS DE EVOLUCIÓN MOLECULAR

Felsenstein (1981) = F81

		To: A	To: C	To: G	To: T
From:	A	$\pi_A + (1 - \pi_A)e^{-mt}$	$\pi_C(1 - e^{-mt})$	$\pi_G(1 - e^{-mt})$	$\pi_T(1 - e^{-mt})$
	C	$\pi_A(1 - e^{-mt})$	$\pi_C + (1 - \pi_C)e^{-mt}$	$\pi_G(1 - e^{-mt})$	$\pi_T(1 - e^{-mt})$
	G	$\pi_A(1 - e^{-mt})$	$\pi_C(1 - e^{-mt})$	$\pi_G + (1 - \pi_G)e^{-mt}$	$\pi_T(1 - e^{-mt})$
	T	$\pi_A(1 - e^{-mt})$	$\pi_C(1 - e^{-mt})$	$\pi_G(1 - e^{-mt})$	$\pi_T + (1 - \pi_T)e^{-mt}$

Permite que las cuatro bases estén a differentes frecuencias

Parámetro Π

($\Pi_A \neq \Pi_T \neq \Pi_G \neq \Pi_C$)

MODELOS MÁS REALISTAS DE EVOLUCIÓN MOLECULAR

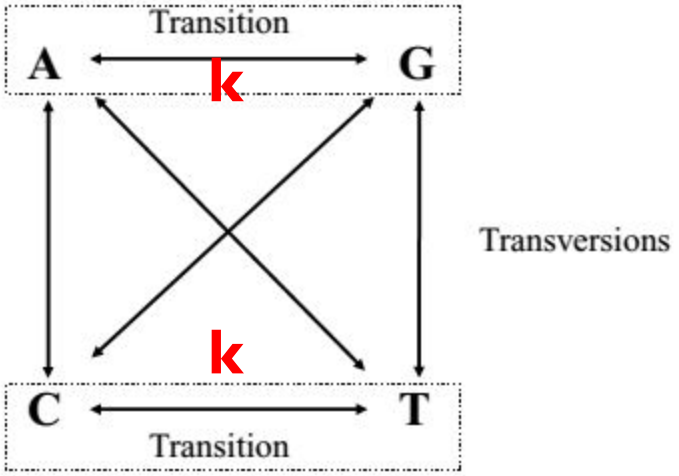
HKY85

		To:			
		A (freq = π_A)	C (freq = π_C)	G (freq = π_G)	T (freq = π_T)
From:	A (freq = π_A)	$-m(\pi_C + \kappa\pi_G + \pi_T)$	$\pi_C m$	$\pi_G \kappa m$	$\pi_T m$
	C (freq = π_C)	$\pi_A m$	$-m(\pi_A + \pi_G + \kappa\pi_T)$	$\pi_G m$	$\pi_T \kappa m$
	G (freq = π_G)	$\pi_A \kappa m$	$\pi_C m$	$-m(\kappa\pi_A + \pi_C + \pi_T)$	$\pi_T m$
	T (freq = π_T)	$\pi_A m$	$\pi_C \kappa m$	$\pi_G m$	$-m(\pi_A + \kappa\pi_C + \pi_G)$

Permite que las tasas relativas de los tipos de sustitución sean diferentes

Transiciones más probables que transversiones

Parámetro **"k"** de tasa relativa de cambio



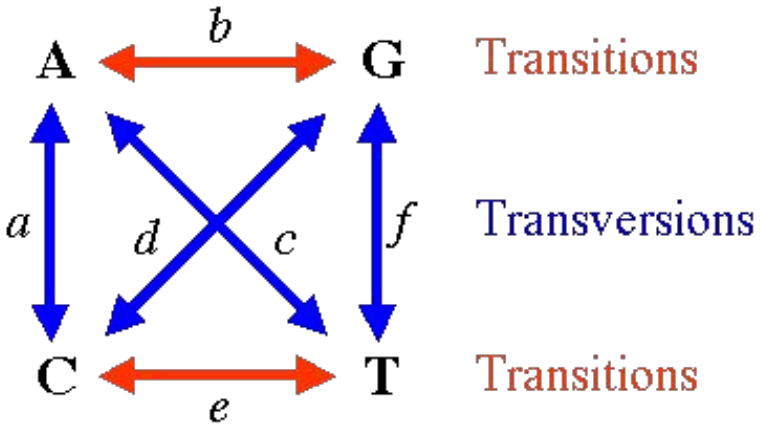
MODELOS MÁS REALISTAS DE EVOLUCIÓN MOLECULAR

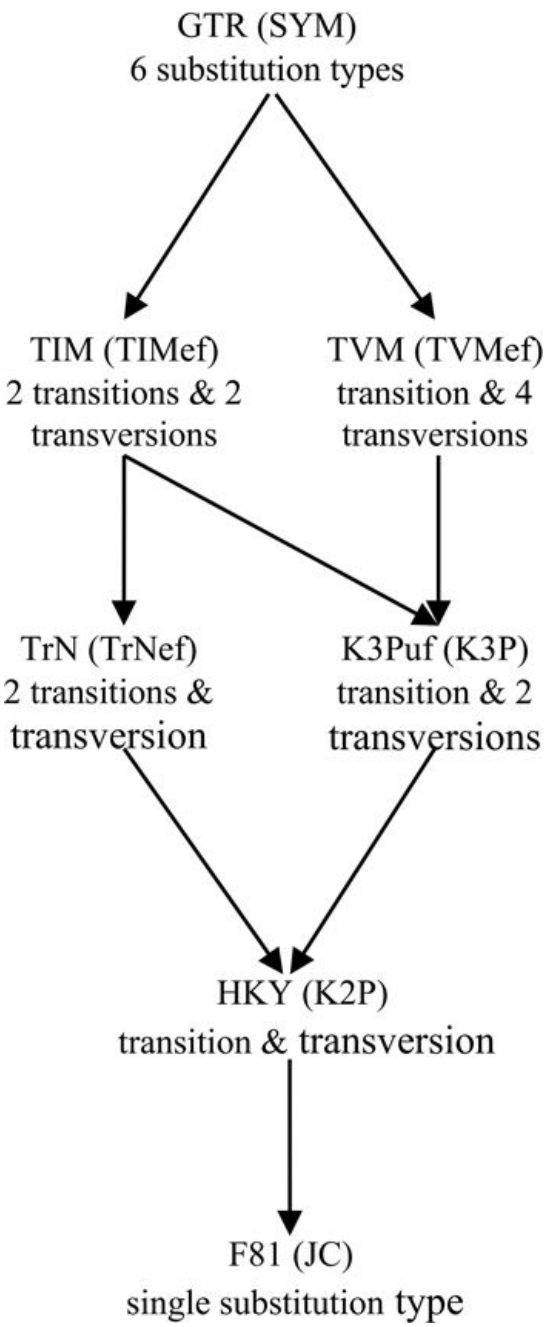
General time-reversible model (GTR)

$$Q = \begin{array}{c} \begin{matrix} \quad\mathbf{A} & \quad\mathbf{C} & \quad\mathbf{G} & \quad\mathbf{T} \end{matrix} \\ \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ a\mu\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & d\mu\pi_G & e\mu\pi_T \\ b\mu\pi_A & d\mu\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & f\mu\pi_T \\ c\mu\pi_A & e\mu\pi_C & f\mu\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix} \end{array}$$

El más complejo: todas las tasas relativas de cambio son diferentes

Adiciona un parámetro diferente para cada tipo de cambio



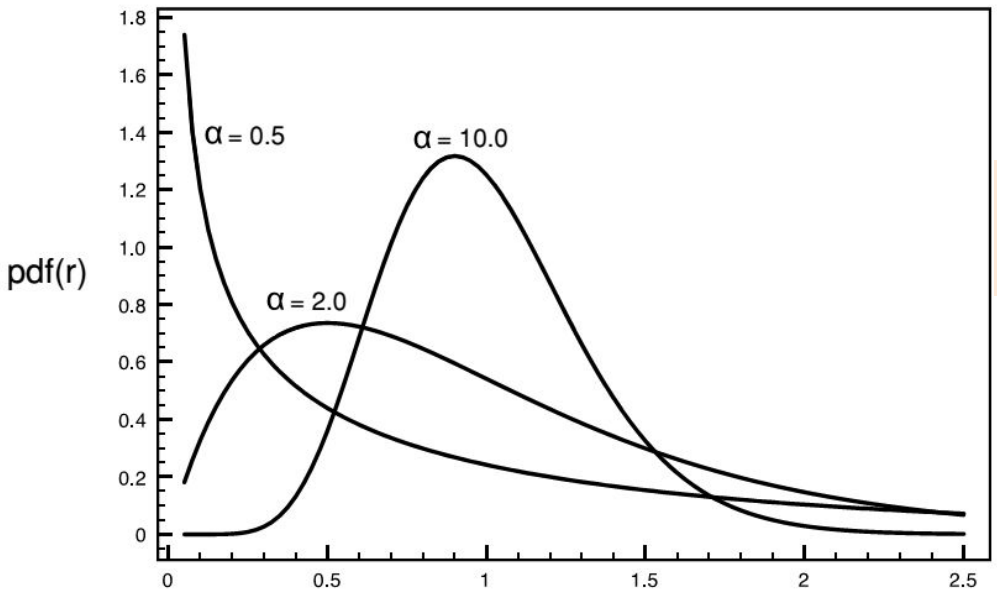

GTR (SYM)
6 substitution types
TIM (TIMef)
2 transitions & 2 transversions
TVM (TVMef)
transition & 4 transversions
TrN (TrNef)
2 transitions & transversion
K3Puf (K3P)
transition & 2 transversions
HKY (K2P)
transition & transversion
F81 (JC)
single substitution type

¿Que pasa si las posiciones evolucionan a tasas diferentes dentro de la secuencia de ADN?

P.e. Codones (lento en bases que modifican aminoácidos)

ESTRATÉGIAS:

- Particionar los datos y agruparlos de acuerdo a sus tasas de evolución.
- Asumir que la tasa de sustitución es diferente a través de las posiciones y muestrear valores con respecto a una distribución de tasas.

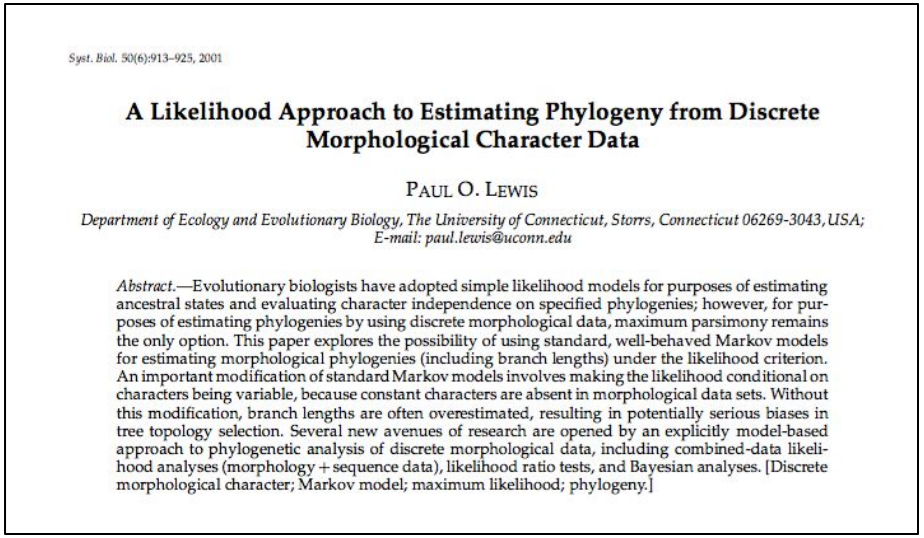


Distribución gamma de tasas de sustitución

¿EXISTEN MODELOS DE EVOLUCIÓN MORFOLÓGICA?

Modelo "Mk" (Markov y número de estados observados)

- Generalización de JC69 (k = 4 estados)
- Usado también para AFLPs, INDELS, proteínas



Syst. Biol. 50(6):913–925, 2001

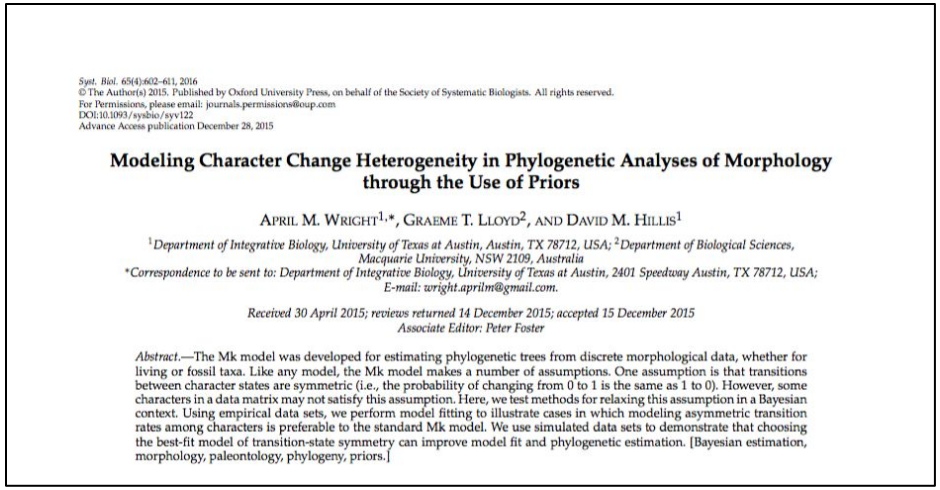
A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data

PAUL O. LEWIS

Department of Ecology and Evolutionary Biology, The University of Connecticut, Storrs, Connecticut 06269-3043, USA; E-mail: paul.lewis@uconn.edu

Abstract.—Evolutionary biologists have adopted simple likelihood models for purposes of estimating ancestral states and evaluating character independence on specified phylogenies; however, for purposes of estimating phylogenies by using discrete morphological data, maximum parsimony remains the only option. This paper explores the possibility of using standard, well-behaved Markov models for estimating morphological phylogenies (including branch lengths) under the likelihood criterion. An important modification of standard Markov models involves making the likelihood conditional on characters being variable, because constant characters are absent in morphological data sets. Without this modification, branch lengths are often overestimated, resulting in potentially serious biases in tree topology selection. Several new avenues of research are opened by an explicitly model-based approach to phylogenetic analysis of discrete morphological data, including combined-data likelihood analyses (morphology + sequence data), likelihood ratio tests, and Bayesian analyses. [Discrete morphological character; Markov model; maximum likelihood; phylogeny.]

Lewis (2001)



Syst. Biol. 65(4):602–611, 2016

DOI:10.1093/sysbio/syv122
Advance Access publication December 28, 2015

Modeling Character Change Heterogeneity in Phylogenetic Analyses of Morphology through the Use of Priors

APRIL M. WRIGHT[1,*], GRAEME T. LLOYD[2], AND DAVID M. HILLIS[1]

[1]*Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA;* [2]*Department of Biological Sciences, Macquarie University, NSW 2109, Australia*

**Correspondence to be sent to: Department of Integrative Biology, University of Texas at Austin, 2401 Speedway Austin, TX 78712, USA; E-mail: wright.aprilm@gmail.com.*

Received 30 April 2015; reviews returned 14 December 2015; accepted 15 December 2015
Associate Editor: Peter Foster

Abstract.—The Mk model was developed for estimating phylogenetic trees from discrete morphological data, whether for living or fossil taxa. Like any model, the Mk model makes a number of assumptions. One assumption is that transitions between character states are symmetric (i.e., the probability of changing from 0 to 1 is the same as 1 to 0). However, some characters in a data matrix may not satisfy this assumption. Here, we test methods for relaxing this assumption in a Bayesian context. Using empirical data sets, we perform model fitting to illustrate cases in which modeling asymmetric transition rates among characters is preferable to the standard Mk model. We use simulated data sets to demonstrate that choosing the best-fit model of transition-state symmetry can improve model fit and phylogenetic estimation. [Bayesian estimation, morphology, paleontology, phylogeny, priors.]

Wright et al. (2016)

Pruebas estadística para comparar ln(L) entre modelos

Ln(ML) modelo 1= – 5.92 (1 parámetro)
Ln(ML) modelo 2= – 2.50 (2 parámetros)

$$LRT = 2(\ell_1 - \ell_0)$$

Likelihood Ratio Test: para comparar pares de modelos anidados

$$AIC = -2\ell + 2K$$

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}$$

Akaike Information Criterion: penaliza modelos con más parámetros libres
K= No. de parámetros libres
n = No. de taxones

$$BIC = -2\ell + K \log n$$

Para comparar pares de modelos anidados usando Bayes factors

LTR: Likelihood Ratio Test

Compara modelos anidados

Ln(ML) modelo 1= – 5.92 (1 parámetro)
Ln(ML) modelo 2= – 2.50 (2 parámetros)

$$\Delta = 2 \cdot (\ln L_2 - \ln L_1)$$
$$\Delta = 2 \cdot (-2.50 - -5.92)$$
$$\Delta = 6.84$$

Comparamos con una distribución chi cuadrado con 1 grado de libertad (parámetros de diferencia entre los dos modelos)

P=0.009

LTR: Likelihood Ratio Test

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582

AIC y AICc: Akaike Information Criterion

Ln(ML) modelo 1= – 5.92
Ln(ML) modelo 2= – 2.50

$$AIC = -2\ell + 2K$$

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}$$

$$\begin{aligned} AIC_1 &= 2k_1 - 2lnL_1 = 2 \cdot 0 - 2 \cdot -5.92 \\ AIC_1 &= 11.8 \\ AIC_2 &= 2k_2 - 2lnL_2 = 2 \cdot 1 - 2 \cdot -2.50 \\ AIC_2 &= 7.0 \end{aligned}$$

$$\begin{aligned} \Delta AIC_{c_1} &= AIC_{c_1} - AICc_{min} \\ &= 11.8 - 7.0 \\ &= 4.8 \\ \Delta AIC_{c_2} &= AIC_{c_2} - AICc_{min} \\ &= 7.0 - 7.0 \\ &= 0 \end{aligned}$$

$$\begin{aligned} AIC_{c_1} &= AIC_1 + \frac{2k_1(k_1+1)}{n-k_1-1} \\ AIC_{c_1} &= 11.8 + \frac{2 \cdot 0(0+1)}{100-0-1} \\ AIC_{c_1} &= 11.8 \\ AIC_{c_2} &= AIC_2 + \frac{2k_2(k_2+1)}{n-k_2-1} \\ AIC_{c_2} &= 7.0 + \frac{2 \cdot 1(1+1)}{100-1-1} \\ AIC_{c_2} &= 7.0 \end{aligned}$$

ΔAICc < 4: el modelo i es casi equivalente al modelo de valor más bajo

ΔAICc 4–8: el modelo i soporta débilmente a los datos

ΔAICc > 8: el modelo i puede ser descartado con seguridad