Curso fundamenteal de Inferencia Filogenética Molecular

Pablo Vinuesa (vinuesa@ccg.unam.mx)

Progama de Ingeniería Genómica, CCG, UNAM

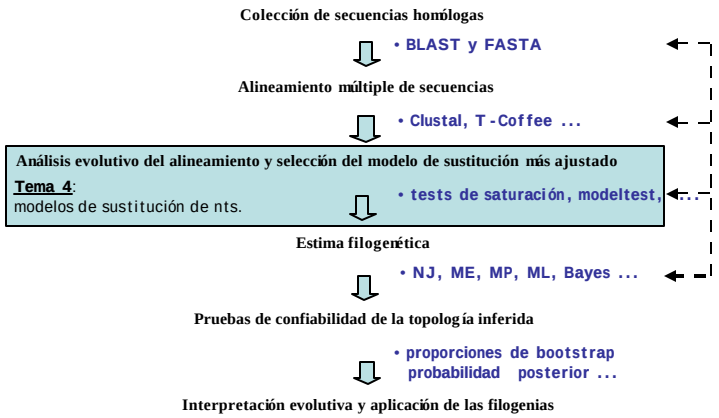
http://www.ccg.unam.mx/~vinuesa/

Tutor: PDCBM, Ciencias Biológicas, PDCBioq. y
Profesor de la Lic. Ciencias Genómicas y posgrado

- **Tema 4: Modelos de sustitución de secuencias nucleotídicas**

1. Modelos en ciencia y filogenética
2. Procesos y modelos Markovianos
3. Aproximaciones empíricas vs. paramétricas al modelado del proceso de sustitución
4. Parámetros relevantes para describir el proceso de sustitución en secuencias de DNA
5. La familia GTR de modelos de sustitución de nucleótidos
6. Condiciones de aplicabilidad de los modelos y sus violaciones
7. Acomodo de la heterogeneidad de tasas de sustitución entre sitios en modelos de sustitución: proporción de sitios invariantes y distribución gamma.

Protocolo básico para un análisis filogenético de secuencias moleculares



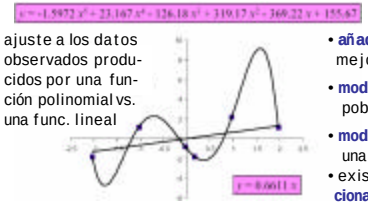
Modelos de evolución de secuencias -introducción

- Para el análisis filogenético de secuencias alineadas virtualmente todos los métodos describen la evolución de las secuencias usando un modelo que consta de dos componentes:

1. un árbol filogenético
2. una descripción de la manera (frecuencias de sustitución) en que las sustituciones de aas o nts de las secuencias individuales evolucionan a lo largo de las ramas de dicho árbol

- **¿Porqué necesitamos modelos y para qué sirven?**

- Los modelos nos sirven para interpolar adecuadamente entre nuestras observaciones con el fin de poder hacer predicciones inteligentes sobre observaciones futuras



ajuste a los datos observados producidos por una función polinomial vs. una func. lineal

- **añadir parámetros** a un modelo generalmente mejora su ajuste a los datos observados
- **modelos infra-parametrizados** conducen a un pobre ajuste a los datos observados
- **modelos supra-parametrizados** conducen a una pobre predicción de eventos futuros
- existen métodos estadísticos para **seleccionar modelos ajustados** a cada set de datos

Modelos de evolución de secuencias -introducción

- **Modelos de evolución del proceso de sustitución y métodos de reconstrucción filogenética: consideraciones generales**

1.- La reconstrucción o estima filogenética es un **problema de inferencia estadística**, y como tal **requiere un modelo** de sustitución de resíduos (aa o nt), es decir, un modelo de evolución molecular de las secuencias. Todos los modelos, por no ser más que aproximaciones de los procesos naturales, hacen una serie de **suposiciones** (simplificaciones)

2.- Los **modelos de evolución de secs.** son usados en filogenética **para describir las probabilidades con las que se dan los distintos eventos de sustitución** entre aa o nt, con el fin de **corregir o compensar las sustituciones no observadas a lo largo de la filogenia**

3.- Mientras que los métodos de **MP** asumen un **modelo implícito** de evolución (número mínimo de sustituciones a lo largo de la filogenia), los métodos de **distancia** (UPGMA, NJ), los de **ML y Bayesianos** requieren de un **modelo explícito** de evolución

4.- Los **métodos de distancia** estiman finalmente **un sólo parámetro** (**no.esperado sust./sitio**) dado el modelo y el valor de los parámetros del mismo; en cambio, los **métodos de ML y Bayesianos** pueden **estimar el valor de cada uno de los parámetros del modelo** explicitado, dada una topología y la matriz de datos (alineamiento)

Modelos de evolución de secuencias -introducción

- **Modelos de evolución del proceso de sustitución y métodos de reconstrucción filogenética: consideraciones generales**

Corolario:

1. El grado de confianza que tengamos en una filogenia particular realmente depende de la que tengamos en el modelo subyacente

2. Por lo tanto, siempre que usemos un método de reconstrucción basado en un modelo explícito de evolución (NJ, ML, By) es necesario usar rigurosas pruebas estadísticas para seleccionar el modelo y el valor de sus parámetros que mejor se ajusten a la matriz de datos a analizar

Modelos de evolución de secuencias -introducción

- **Modelos de evolución del proceso de sustitución y métodos de reconstrucción filogenética: consideraciones generales**

• Existen dos aproximaciones para construir modelos de evolución de secuencias.

1. construcción de **modelos empíricos** basados en propiedades del proceso de sustitución calculadas a partir de comparaciones de un gran número de alineamientos. Los modelos empíricos **resultan en valores fijos de los parámetros**, los cuales son estimados sólo una vez, suponiéndose que son adecuados para el análisis de otros sets de datos. Esto los hace fácil de usar e implementar en términos computacionales, pero su utilidad real para cada caso particular ha de ser evaluada críticamente. Se usan principalmente en el análisis evolutivo de **secs. de AAs** (modelos BLOSUM, PAM, JTT, WAG...)

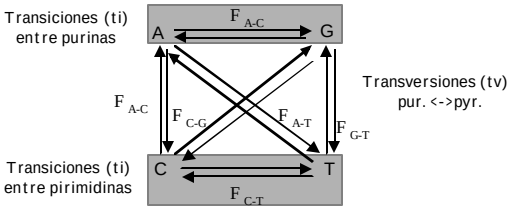
2. construcción de **modelos paramétricos** basado en el modelado de propiedades químicas o genéticas del aas y nts. Los modelos paramétricos tienen la ventaja de que los **valores de los parámetros pueden ser inferidos de cada set de datos** al hacer un análisis de los mismos usando métodos de ML o By, por tanto ajustándolos a cada matriz de datos particular. Usados principalmente en el análisis de **secs. de nts.**

Modelos de evolución de secuencias de DNA -introducción

- **Modelos de evolución del proceso de sustitución nucleotídica y métodos de reconstrucción filogenética: consideraciones generales**

• Ambos métodos resultan en modelos de **procesos de Markov**, definidos por **matrices** que contienen las **tasas relativas** (nos. relativos, en promedio y por u. de t.) de ocurrencia de todos los tipos posibles de sustituciones

• La mayoría de los modelos asume que esta **matriz es reversible**, es decir, que no se puede definir de antemano la direccionalidad temporal del proceso evulutivo, resultando por lo tanto en **árboles no enraizados**. Para definir la polaridad o dirección evolutiva se requiere información biológica externa adiconal (p. ej registro fósil o grupo externo)



Modelos de evolución de secuencias de DNA -introducción

- **Modelos de evolución del proceso de sustitución: supuestos básicos**

• Las **sustituciones** (reemplazos evolutivos de los estados de caracter) se describen como el el resultado de mutaciones al azar. Su aparición en cada posición de una secuencia a lo largo del tiempo es modelado por un **proceso de Markov.** La probabilidad de intercambio de un estado de caracter por otro viene modelada en esencia por una **distribución de Poisson (de eventos raros)**

• **Proceso Markoviano**: es un modelo matemático de eventos raros de cambios en estados (discretos o de caracter) a lo largo del tiempo, en el que los eventos futuros suceden por azar y dependen únicamente del estado actual, y no de la historia que llevó a dicho estado. En filogenética, los estados del proceso son los aas o nts presentes en una posición particular de una secuencia (estados de caracter) en un tiempo dado; los cambios de estado representan las mutaciones en dichas secuencias

Modelos de evolución de secuencias de DNA -introducción

- **Modelos de sustitución de nucleótidos**

- El modelado del proceso de sustitución nucleotídica se ha concentrado en la aproximación paramétrica. Se manejan **tres tipos principales de parámetros** en estos modelos:

  1. parámetros de **frecuencia**
  2. parámetros de **tasas de intercambio**
  3. parámetros de **heterogeneidad de tasas de sustitución** entre sitios

Modelos de evolución de sustitución de nucleótidos -modelos paramétricos

- Los **parámetros de frecuencia describen las frecuencias de las bases** (A, C, G, T) promediadas sobre todas las posiciones y a lo largo del árbol. Estos parámetros representan constricciones a las sustituciones posibles debido a causas tales como el contenido de GC del genoma. Funcionan como factores de ponderación en el modelo al hacer unas sustituciones más probables que otras. Se trata de frecuencias relativas (sumatoria = 1).

- Los **parámetros de tasa (de intercambiabilidad)** describen las tendencias relativas de las bases de ser sustituídas unas por otras (hasta 6 parámetros que representan las tasas de sustitución relativas entre A<->C, A<->G, A<->T, C<->G, C<->T y G<->T (esta última generalmente es definida como = 1). Generalmente a las *ti* se les asigna una tasa, κ, relativa a una tasa de 1 para las *tv*, siendo generalmente $\kappa >> 1$

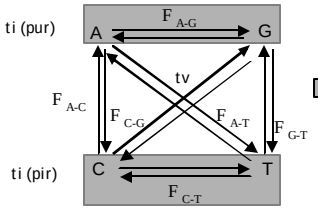
- La aproximación más utilizada para **modelar la heterogeneidad de tasas entre sitios** es la de describir la tasa de sustitución de cada posición como una muestra aleatoria de una **distribución GAMMA.** El uso de esta distribución para modelar la heterogeneidad de tasas entre sitios representa un factor muy significativo para incrementar el ajuste entre los modelos y los datos. Esta distribución gamma se emplea en conjunción con los parámetros de frecuencia y tasa, describiéndose el modelo resultante como **sufijo+G** ó **sufijo+T** (por ejemplo, TrN+G ó K2P+T)

Modelos de evolución de sustitución de nucleótidos -modelos paramétricos

- **los diversos modelos evolutivos se distinguen por su grado de parametrización**

I. Frecuencias de nt : pA = pC = pG = pT = 0.25 ó pA ? pC ? pG ? pT
- modelos de = frecuencia: JC69; K2P, K3P ...
- modelos de ? frecuencia: F81, F84, HKY85, TrN93, GTR ...

II. Tasas de sustitución transicionales/transversionales



- Existen 4 tipos de sustituciones ti y 8 tv; cuando **ti/tv ? 0.5 existe un sesgo en sustituciones ti (o tv) en el set de datos. ti generalmente >> 1**

- los modelos evolutivos se diferencian también en la cantidad de parámetros que utilizan para acomodar diversas tasas de sustitución:

tasas	modelo
1	**JC69** (ti=tv)
2	**K2P, F84** (ti ?tv)
3	**TrN** ó **K3P** (2 ti, 1 tv)
6	**GTR** (cada sust. su tasa)

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

- Matriz de tasas de sustitución instantáneas del modelo GTR

	A	C	G	T
A	–	$\pi_C a\mu$	$\pi_G b\mu$	$\pi_T c\mu$
C	$\pi_A a\mu$	–	$\pi_G d\mu$	$\pi_T e\mu$
G	$\pi_A b\mu$	$\pi_C d\mu$	–	$\pi_T f\mu$
T	$\pi_A c\mu$	$\pi_C e\mu$	$\pi_G f\mu$	–

$-\mu\,(\pi_A c + \pi_C e + \pi_G f)$

9 parámetros
π_A
π_C
π_G
a
b
c
d
e
m

El modelo GTR es idéntico al de JC69 si $a = b = c = d = e = f = 1$ y todas las bases se asumen que tienen igual frecuencia (¼)

m = tasa del proceso generador de todos los tipos de sustituciones
a, ... *e* = modificadores de tasa relativa de cada tipo particular de sustitución
p = frecuencia de cada nt

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

Jukes-Cantor 1969 (JC69)
igual frecuencia de bases: $p_A = p_C = p_G = p_T$
todas las sustituciones tienen igual tasa $\alpha=\beta$

acomodan sesgo ti/tv — acomodan distintas frecuencias de bases

Kimura 2 parameters (K2P)
igual frec. de bases: $p_A = p_C = p_G = p_T$
distintas tasas de sustitución ti y tv; $\alpha?\beta$

Felsenstein 1981 (F81)
distinta frec. de bases: $p_A? p_C? p_G? p_T$
igual tasa de sustitución ti y tv; $\alpha=\beta$

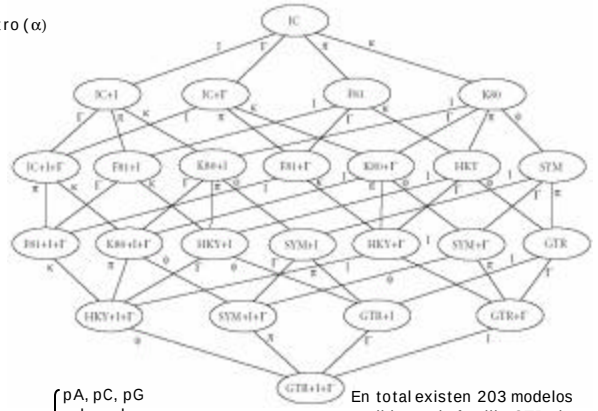
acomodan ? frec. bases — acomodan sesgo tasas sust. ti/tv

distintas frecs. bases: pA ? pC ? pG ? pT
distintas tasas de sust. ti and tv; $\alpha?\beta$
Hasegawa-Kishino-Yano (HKY85), 2 tasas y Felsenstein 1984 (F84)
o
Tamura-Nei 1993 (TN93), 3 tasas
o
Modelo general tiempo reversible (GTR), 6 tasas

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

1 parámetro (α)

Incremento en el número de parámetros
modelos más generales



11 parámetros libres a estimar
pA, pC, pG
a, b, c, d, e
μ,
I, T

En total existen 203 modelos posibles en la familia GTR al combinar params. de frec., tasa, G e I
La mayoría de ellos carecen de nombre

Modelos de evolución de sustitución de nucleótidos -modelos paramétricos

Condiciones de aplicabilidad de los modelos (supuestos)

1.- Supuesto de independencia: las mutaciones en un sito no afectan a otros en la secuencia. Violado por ej. en el caso de rRNAs suelen seleccionarse mutaciones compensatorias (evolución covariada entre sitios)

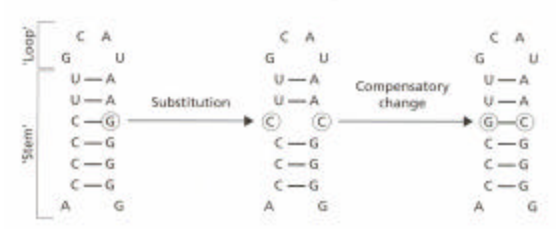
2.- Supuesto de homogeneidad de tasas de sustitución a lo largo del tiempo y entre linajes: en este supuesto se basa el reloj molecular y de su cumplimiento depende la posibilidad de poder utilizar un "**reloj molecular**" para datar clados

3.- Las frecuencias de nucleótido son homogéneas entre linajes: este supuesto es frecuentemente violado cuando usamos secuencias de linajes muy distantes, particularmente en procariontes, ya que los contenidos de G+C de distintos grupos microbianos varía mucho, del 22 % (*Wigglesworthia*, *gamma-Proteobacteria*) al 75 % (*Anaeromyxobacter*, *delta-Proteobacteria)*

4.- Las probabilidades de sustitución son las mismas para cada sitio: este supuesto es violado casi sin excepción. Así por ejemplo, las 3as. pos. de los codones acumulan mutaciones mucho más rápidadmenteque la 2a y 1a. Los distintos dominios de una proteína o rRNA también evolucionan con tasas distintas. **Distribución Gamma (G)**

Condiciones de aplicabilidad de los modelos (supuestos)

1.- Supuesto de independencia y modelos de covarión: las mutaciones en un sito no afectan a otros en la secuencia. En el caso de rRNAs suelen seleccionarse mutaciones compensatorias (**covariación de sitios**). Existen modelos que acomodan este hecho.

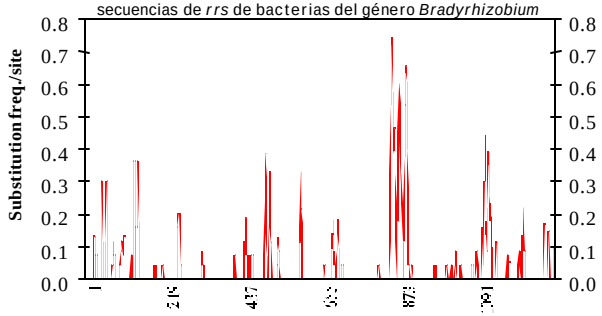


Condiciones de aplicabilidad de los modelos (supuestos)

Acomodo de la heterogeneidad de tasas de sustitución entre sitios

(**I**) acomoda las posiciones invariables (**proporción de sitios invariantes**)
(**G**) acomoda la **heterogeneidad de tasas de sust**. entre las posiciones variables

Heterogeneidad de tasas de sust. a lo largo de un aln. múltiple de 29 secuencias de *rrs* de bacterias del género *Bradyrhizobium*

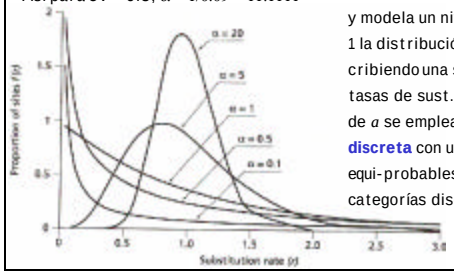


Condiciones de aplicabilidad de los modelos (supuestos)

2.- Distribución gamma y heterogeneidad de tasas de sust. entre sitios: Para modelar con cierto realismo el proceso de sustitución es esencial acomodar adecuadamente la heterogeneidad de tasas de sustitución entre sitios de un alineamiento

Pdf(*r*) = $\alpha^{\alpha} r^{\alpha-1} / \exp(\alpha r) \Gamma(\alpha)$

Diversas formas de la distribución gamma (Γ) para un rango de valores del parámetro $\alpha = 1/CV^2$, donde CV = coef. de var. de las tasas.
Así para CV = 0.3, $\alpha = 1/0.09 = 11.1111$



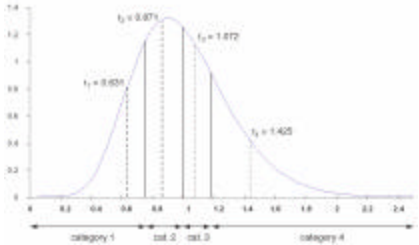
Para ello se asume generalmente una **distribución gamma (G)** de las tasas y que cada sitio tiene una tasa tomada aleatoriamente de dicha distribución, e independientemente de los demás sitios. El parámetro α controla la forma de la distribución. Para $\alpha > 1$ la distribución tiene forma de campana y modela un nivel bajo de heterogeneidad. Para $\alpha < 1$ la distribución toma forma de L invertida, describiendo una situación de fuerte heterog. de tasas de sust. entre sitios. Para calcular el valor de α se emplea generalmente una **distribución G discreta** con un número c finito de tasas des sust. equi-probables (q_1, q_2 ..., q_c). El uso de 4 a 8 categorías discretas permite obtener una buena aprox. de la distrib. contínua.

Discretización de la distribución Gamma

- Acomodo de la heterogeneidad en tasas de sustitución entre sitios (HTSES): **distribución gamma (G)**

- Y expandiendo para *m* tasas, tenemos: $L_k = \sum_{i=1}^{m} (p_i)\, L_k^{(ri)}$

- Bajo el modelo gamma, las *m* tasas relativas r_i se escogen de tal manera que cada una es la media de una sección de igual área de la distribución gamma. Dado que la distribución es partida en pedazos con igual área, las probabilidades p_i son todas = 1/*m*.



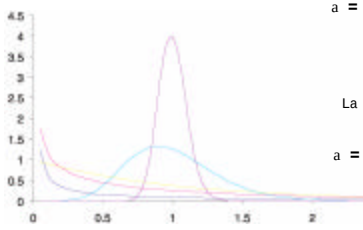
Distribución gamma para $\alpha = 10$ y 4 categorías discretas de tasas de sust.

El área bajo la curva de densidad gamma completa es = 1, por lo que la media de las r_i = 1. Esto es adecuado, al tratarse de tasas relativas. Estos valores de r_i son utilizados para modificar los parámetros de tasa en las fórmulas de probabilidades de transición. Así para JC, la tasa de sustitución αt para la primera categoría es substituída por $r_1 \alpha$

Distribución gamma y acomodo de heterogeneidad de tasas entre sitios

- Acomodo de la heterogeneidad en tasas de sustitución entre sitios (HTSES): **distribución gamma (G)**

La distribución gamma puede adoptar muchas formas dependiendo únicamente del valor del parámetro α, conocido también como parámetro de forma ("shape parameter"): para valores < 1 la distribución adopta forma de L invertida y representa una gran heterogeneidad de tasas entre sitios. Valores grandes de α representan poca heterogeneidad. La distribución puede ir desde casi 0 hasta el 8 , 8 representando homogeneidad perfecta



a = **0.1**

i	lower	upper	r_i
1	0.00000000	0.00000579	0.00000063
2	0.00000579	0.00593391	0.00107809
3	0.00593391	0.35306358	0.09375338
4	0.35306358	∞	3.90516801

La tasa más alta y baja se diferencian en > 7X10⁶

a = **300**

i	lower	upper	r_i
1	0.00000000	0.96047646	0.92759503
2	0.96047646	0.99888911	0.98031495
3	0.99888911	1.03831262	1.01778024
4	1.03831262	∞	1.07430888

La tasa más alta y baja se diferencian en sólo ~ 1.5

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

- El método de momentos es de utilidad limitada en estadística (y filogenética) ya que no permite obtener una fórmula explícita para calcular la distancia entre secuencias usando modelos más complejos como el HKY85 o GTR

- La fórmula explícita de distancia para el modelo K2P es:

$$d = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

-este modelo tiene 2 parámetros, ***P*** y ***Q*** (proporción de *ti* y *tv* en que difieren 2 secuencias, donde $p = P + Q$

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

- Comparación de los modelos de JC69 y K2P en su capacidad de corregir distancias observadas (p) entre pares de secuencias según su grado de divergencia

$$d_{JC69} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \quad \text{vs.} \quad d_{K2P} = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

- Escenario I:

- sean 2 secs. de long. = 200 nt, que difieren en 20 *ti* y 4 *tv*

por lo tanto L = 200, $P = 20/200 = 0.1$ y $Q = 4/200 = 0.02$

$p = 24/200 = 0.12$

d_{JC69} ~ 0.13 (sust./sitio)

no. de sust. esperadas = 0.13 X 200 ~ 26

d_{K2P} ~ 0.13 (sust./sitio)

no. de sust. esperadas = 0.13 X 200 ~ 26

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

- Comparación de los modelos de JC69 y K2P en su capacidad de corregir distancias observadas (p) entre pares de secuencias según su grado de divergencia

$$d_{JC69} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \quad \text{vs.} \quad d_{K2P} = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

- Escenario II:

- sean 2 secs. de long. = 200 nt, que difieren en 50 *ti* y 16 *tv*

por lo tanto L = 200, $P = 50/200 = 0.25$ y $Q = 16/200 = 0.08$

$p = 66/200 = 0.33$

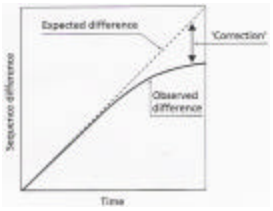
d_{JC69} ~ 0.43 (sust./sitio)

no. de sust. esperadas = 0.43 X 200 ~ 86

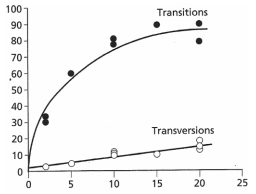
d_{K2P} ~ 0.48 (sust./sitio)

no. de sust. esperadas = 0.48 X 200 ~ 96

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV



- El objetivo de los modelos de sustitución es el de **compensar para los eventos homoplásicos de múltiples sustituciones**, y así obtener estimas de distancias evolutivas corregidas



- El número de *ti* es generalmente > que el de *tv*, fenómeno que se acentúa cuanto mayor es la divergencia entre las secuencias a comparar. De ahí que en nuestro ejemplo las diferencias entre los escenarios I y II sólo se hicieron notar en el caso en el que la divergencia entre las secuencias era mayor (escenario II)

Selección de modelos de sustitución para nucleótidos y Proteínas usando ModelGenerator Web Server

http://distributed.cs.nuim.ie/multiphyl.php

Formato FASTA o Phylip

Selección de modelos de sustitución para nucleótidos y Proteínas usando ModelGenerator Web Server

http://distributed.cs.nuim.ie/multiphyl.php

Selección de modelos de sustitución para nucleótidos y Proteínas usando ModelGenerator Web Server

http://distributed.cs.nuim.ie/multiphyl.php

```
**********Report file for MultiPhyl v1.0.6**********

File Created on Mon Nov 19 19:07:18 GMT 2007

To cite MultiPhyl:
Thomas M Keane, Thomas J Naughton, James O McInerney (2007) MultiPhyl:
      A high-throughput phylogenomics webserver using distributed computing, Nucleic Acids Research, 35:W33-W37

ALIGNMENT INFORMATION

Alignment File                          my_46U_recA510.phy
Number of taxa                          46
Datatype is                     NUCLEOTIDE
Alignment format is             PHYLIP INTERLEAVED
Number of sites                         510
Number of constant sites is 337 (66.08% of sequence)


...


MODEL CHOOSEN FOR TREESEARCH: GTR+I+G
```

Selección de modelos de sustitución para nucleótidos FindModel Web Server

http://hcv.lanl.gov/content/hcv-db/findmodel/findmodel.html

Selección de modelos de sustitución para nucleótidos
FindModel Web Server

http://hcv.lanl.gov/content/hcv-db/findmodel/findmodel.html

****
AIC-SELECTED MODEL: GTR+G :
General Time Reversible plus Gamma
(model 55) lnL = -2388.407125 AIC = 4794.81425
****

Selección de modelos de sustitución para nucleótidos
ModelTest

- Los servidores anteriores se basan en el software ModelTest escrito por David Posada (Posada & Crandall, 1998).
- Modeltest evalúa los valores de máxima verosimilitud calculados por PAUP* sobre una filogenia NJ-JC para 56 modelos de sustitución anidados de la familia GTR.
- Pueden consultar mi detallado tutorial sobre uso de ModelTest. Veremos cómo usar PAUP* y ModelTest en el tema de máxima verosimilitud.