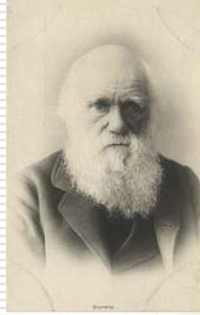
Adaptive Evolution on Genes and Genomes

Hernán J. Dopazo

e-mail: hdopazo@cipf.es *http://bioinfo.cipf.es/hdopazo*

Bioinformatics Department
Comparative Genomics Unit
Centro de Investigación Príncipe Felipe (CIPF)
Valencia, Spain

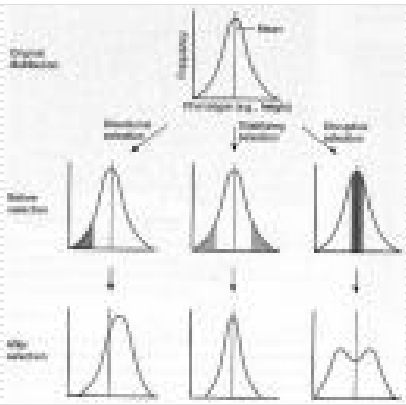


Natural Selection

- The mechanisms in which relative frequencies of genotypes change according to their relative fitnesses in the population

- Positive Selection

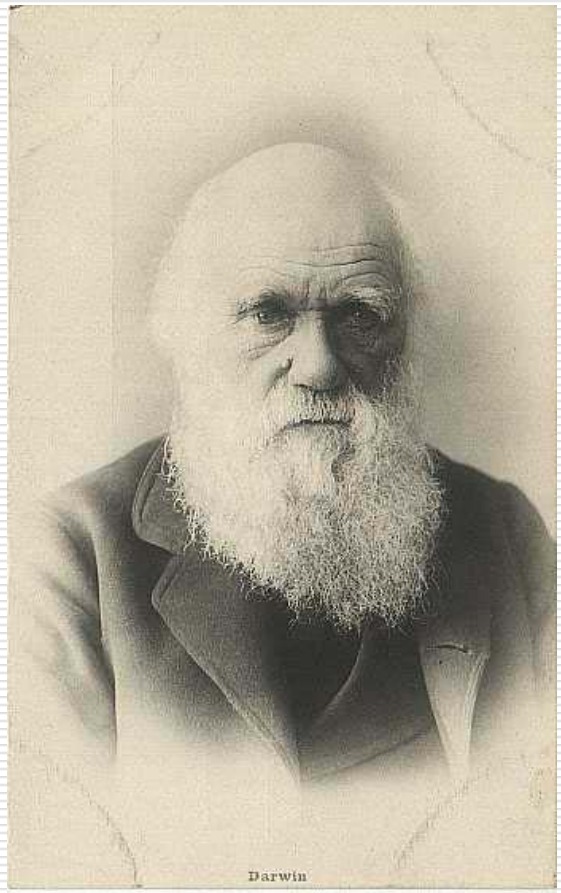
- Purifying Selection



Natural Selection

- **Natural selection is a process of pervasive importance in the biological word, which includes our own species, and on which that species is utterly depenent. Progress in evolutionary biology and its applications is perhaps most obviously relevant to medical and environmental issues, but there is no aspect of human life for which an understanding of evolution is not a vital neccestity.**

George C. Williams, Plan and Purpose in Nature, 1996



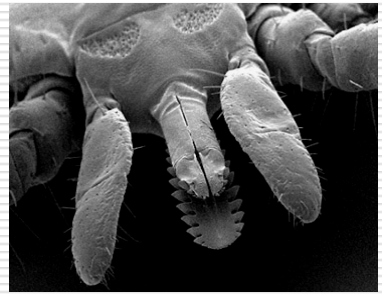


Positive Selection

- The mechanism whereby newly produced mutants have **higher fitnesses** than the average in the population and the frequencies of the mutants **increase** in the following generations

- The biological consequence of such a mechanism is: **adaptation**





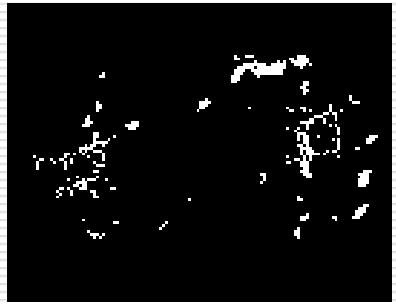




Purifying Selection

- The mechanism whereby newly produced mutants have **lower fitnesses** than the average in the population, and the frequencies of the mutants **decrease** in the following generations

- The biological consequence of such a mechanism is the **mantainement of adaptation** in the population





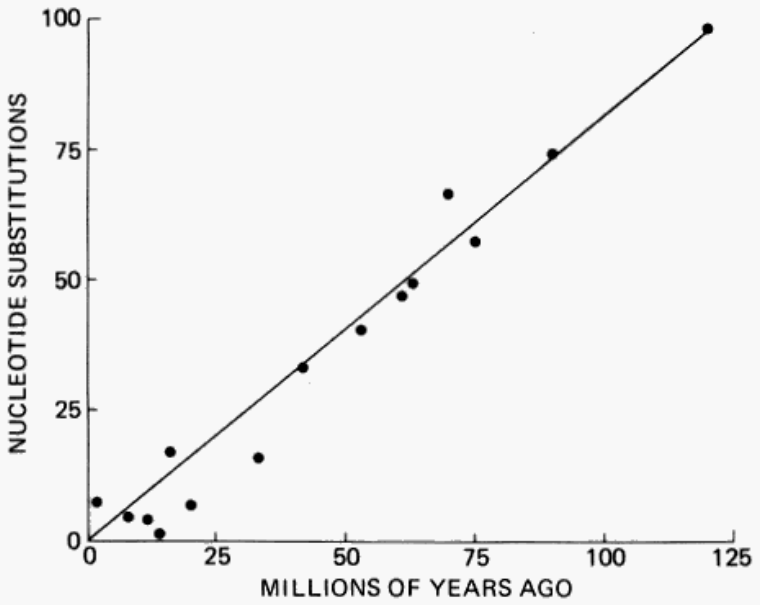




Neutral Evolution

- The great majority of evolutionary changes at the molecular level are caused by random drift of selectively neutral or nearly neutral mutations

- The relevant consequence of such a mechanism is the **fixation of mutations at a constant rate with** in the population



Measuring natural selection at molecular level

Selective pressure at the protein level can be measured as: $\omega = dN/dS$

where, dN is the number of nonsynonymous substitutions per nonsynonymous site and dS is the number of synonymous substitutions per synonymous site between two protein-coding sequences

Measuring natural selection at molecular level

Selective pressure at the protein level can be measured as: $\omega = dN/dS$

where, dN is the number of nonsynonymous substitutions per nonsynonymous site and dS is the number of synonymous substitutions per synonymous site between two protein-coding sequences

- If non-synonymous mutations are favoured by positive selection, non-synonymous mutations will be fixed at faster rate than synonymous mutations, then $\omega > 1$

Measuring natural selection at molecular level

Selective pressure at the protein level can be measured as: $\omega = dN/dS$

where, dN is the number of nonsynonymous substitutions per nonsynonymous site and dS is the number of synonymous substitutions per synonymous site between two protein-coding sequences

- If non-synonymous mutations are favoured by positive selection, non-synonymous mutations will be fixed at faster rate than synonymous mutations, then $\omega > 1$
- If non-synonymous mutations are deleterious (purifying selection), synonymous mutations will be fixed at faster rate, then $\omega < 1$

Measuring natural selection at molecular level

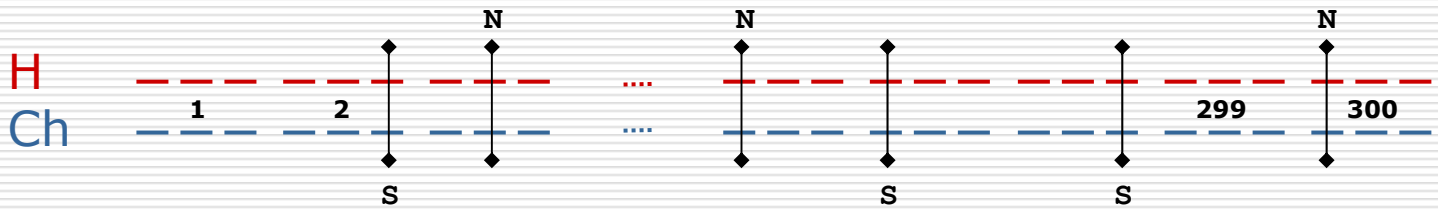
Selective pressure at the protein level can be measured as: $\omega = dN/dS$

where, dN is the number of nonsynonymous substitutions per nonsynonymous site and dS is the number of synonymous substitutions per synonymous site between two protein-coding sequences

- If non-synonymous mutations are favoured by positive selection, non-synonymous mutations will be fixed at faster rate than synonymous mutations, then $\omega > 1$
- If non-synonymous mutations are deleterious (purifying selection), synonymous mutations will be fixed at faster rate, then $\omega < 1$
- If selection has no effect on fitness (neutral evolution), synonymous and non synonymous mutations will be fixed at equal rate, then $\omega = 1$

Sequence pairwise comparison:
Counting methods

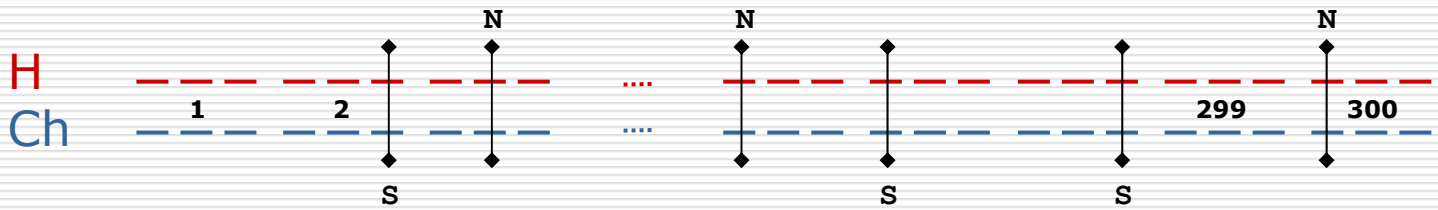
- Suppose a gene has 300 codons and we observe 3 synonymous and 3 nonsynonymous diferences between two sequences.



Is $\omega = 1$?

Sequence pairwise comparison:
Counting methods

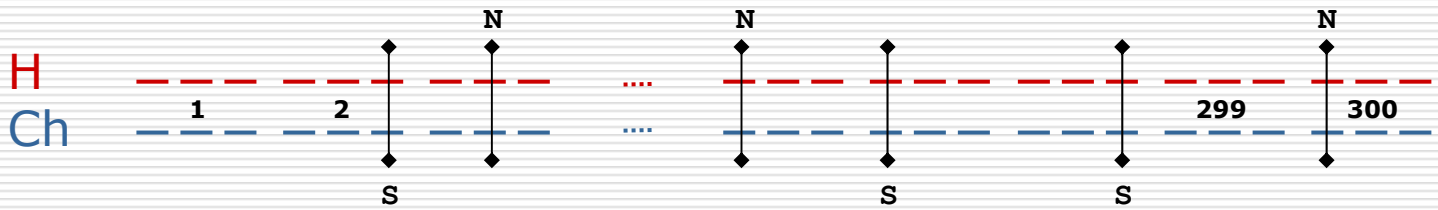
- Suppose a gene has 300 codons and we observe 3 synonymous and 3 nonsynonymous diferences between two sequences.



Is $\omega = 1$?

Sequence pairwise comparison:
Counting methods

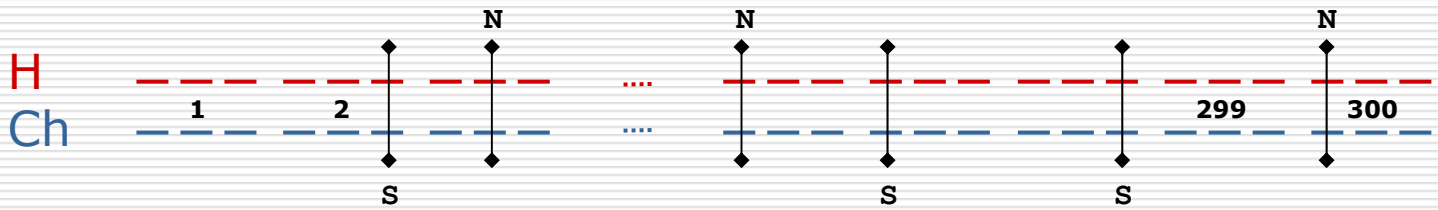
- Suppose a gene has 300 codons and we observe 3 synonymous and 3 nonsynonymous diferences between two sequences.



- If mutations from any one nucleotide to any other occur at the same rate, we expect 25.5% of mutations to be synonymous and 74.5% to be nonsynonymous (a default consequence of the genetic code table).

Sequence pairwise comparison:
Counting methods

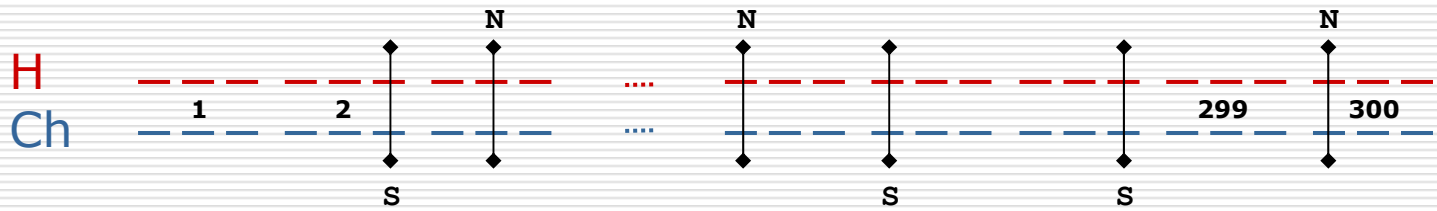
- Suppose a gene has 300 codons and we observe 3 synonymous and 3 nonsynonymous diferences between two sequences.



- If mutations from any one nucleotide to any other occur at the same rate, we expect 25.5% of mutations to be synonymous and 74.5% to be nonsynonymous (a default consequence of the genetic code table).
- We do not expect synonymous and nonsynonymous mutations at equal proportions even if there is no selection at the protein level.

Sequence pairwise comparison: *Counting methods*

- Suppose a gene has 300 codons and we observe 3 synonymous and 3 nonsynonymous diferences between two sequences.



Is $\omega = 1$?

- The total number of synonymous (*S*) and nonsynonymous (*N*) sites probably are close to:

$$N = 3\times 300\times 0.745 = 670.5 \;\; ; \;\; S = 3\times 300\times 0.255 = 229.5$$

- Therefore, $dN = \frac{3}{670.5} = 0.0044 \; ; \;\; dS = \frac{3}{229.5} = 0.0131 \Rightarrow \omega = \frac{0.0044}{0.0131} = 0.336$

Counting methods (3 steps)

Complicated by factors such as ts/tv bias and base/codon frequency bias!!!

1. Count the total number of N and S nucleotide sites

H	CCA	CCG	AAA	ACC	…	TCA	GTA	CGG	AAT
Ch	GCA	CCG	AAA	AGC	…	ACA	GTA	GGG	AAC
	1	2	3	4	…	697	698	699	700

$$N = 3 \times 700 \times 0.634 = 1331.4 \;\; ; \;\; S = 3 \times 700 \times 0.366 = 768.6$$

Counting methods (3 steps)

Complicated by factors such as ts/tv bias and base/codon frequency bias!!!

1. Count the total number of N and S nucleotide sites

H	CCA	CCG	AAA	ACC	…	TCA	GTA	CGG	AAT
Ch	GCA	CCG	AAA	AGC	…	ACA	GTA	GGG	AAC
	1	2	3	4	…	697	698	699	700

$$N = 3 \times 700 \times 0.634 = 1331.4 \;\; ; \;\; S = 3 \times 700 \times 0.366 = 768.6$$

2. Count the synonymous (S) and nonsynonymous (N) differences

H	CCA	CCG	AAA	ACC	…	TCA	GTA	CGG	AAT
	S			N		S		N	S	S
Ch	GCA	CCG	AAA	AGC	…	ACA	GTA	GGG	CAC
	1	2	3	4	…	697	698	699	700

This is straigthtforward if the two compared codons differ at onecodon position only. When they differ at 2 or 3 codon positions, there exists 4 or 6 pathways from one codon to the other. The multiple pathways may involve different number of synonymous and nonsynonymous and should ideally be weighted appropriately according to their likelihood of occurrence. Most counting methods use equal weighting.

Counting methods (3 steps)

Complicated by factors such as ts/tv bias and base/codon frequency bias!!!

1. Count the total number of N and S nucleotide sites

H	CCA	CCG	AAA	ACC	…	TCA	GTA	CGG	AAT
Ch	GCA	CCG	AAA	AGC	…	ACA	GTA	GGG	AAC
	1	2	3	4	…	697	698	699	700

$$N = 3 \times 700 \times 0.634 = 1331.4 \;\; ; \;\; S = 3 \times 700 \times 0.366 = 768.6$$

2. Count the synonymous (S) and nonsynonymous (N) differences

H	CCA	CCG	AAA	ACC	…	TCA	GTA	CGG	AAT
	S			N		S		N	S S
Ch	GCA	CCG	AAA	AGC	…	ACA	GTA	GGG	CAC
	1	2	3	4	…	697	698	699	700

This is straigthtforward if the two compared codons differ at onecodon position only. When they differ at 2 or 3 codon positions, there exists 4 or 6 pathways from one codon to the other. The multiple pathways may involve different number of synonymous and nonsynonymous and should ideally be weighted appropriately according to their likelihood of occurrence. Most counting methods use equal weighting.

3. Apply a correction for multiple substitution at the same site

Counting methods:

- The method of **Miyana-Yasunaga** (1980, J Mol Evol,16(1):23), and its simplified version (**Nei-Gojobori**, 1986, Mol Biol Evol,3(5):418) are based on nucleotide substitution model of Jukes and Cantor (1969) and ignore the ts/tv bias or base codon frequency.

- Since ts are more likely to be synonymous than tv at 3rd. position, ignoring the ts/tv rate bias understimate the number of S and overestimate N. This effect is well known, and different methods account for this ratio
  - (**Li, et al. 1985**, Mol Biol Evol, 2(2):150, **Li, 1993**, J Mol Evol, 36(1):96, **Pamilo and Bianchi** 1993, Mol Biol Evol, 10(2):271, **Ina**, 1995, J Mol Evol, 40(2):190)

- The effect of biased base/codon frequencies can have devastating effects on the estimation of dN and dS. Qualitatively different conclusions were reached depending on whether codon usage bias is accomodated for nucler genes from mammals and Drosophila.

- A counting method incorporating both the ts/tv bias and the base/codon frequency bias was implemented by **Yang and Nielsen, 2000**, Mol Biol Evol,17(1):32.

- Many, if not all of them, are incorporated in codeml (**PAML**) program.

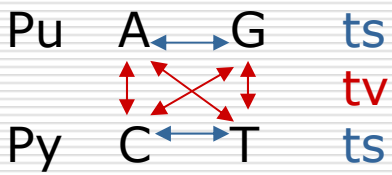
Codon Model

Markov model of codon substitution (61 states or sense codons)

(Goldman & Yang, 1994 Mol Biol Evol, 11, Pp 725; Muse & Gaut, 1994, Mol Biol Evol, 11, Pp 715

Codon instantaneous rate matrix

$$Q = \{q_{ij}\} = \left\{ \begin{array}{ccccc|c} q_{1,1} & q_{1,2} & \cdots & q_{1,60} & q_{1,61} & \sum_j q_{1,j} = 0 \\ q_{2,1} & q_{2,2} & \cdots & q_{2,60} & q_{2,61} & \sum_j q_{2,j} = 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{60,1} & q_{60,2} & \cdots & q_{60,60} & q_{60,61} & \sum_j q_{60,j} = 0 \\ q_{61,1} & q_{61,2} & \cdots & q_{61,60} & q_{61,61} & \sum_j q_{61,j} = 0 \end{array} \right\}$$



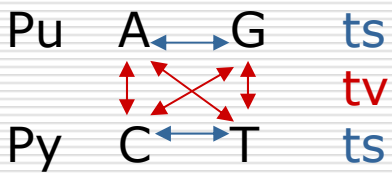
Codon Model

Markov model of codon substitution (61 states or sense codons)

(Goldman & Yang, 1994 Mol Biol Evol, 11, Pp 725; Muse & Gaut, 1994, Mol Biol Evol, 11, Pp 715

Codon instantaneous rate matrix

$$Q = \{q_{ij}\} = \begin{array}{|ccccc|c} q_{1,1} & q_{1,2} & \cdots & q_{1,60} & q_{1,61} & \sum_j q_{1,j} = 0 \\ q_{2,1} & q_{2,2} & \cdots & q_{2,60} & q_{2,61} & \sum_j q_{2,j} = 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{60,1} & q_{60,2} & \cdots & q_{60,60} & q_{60,61} & \sum_j q_{60,j} = 0 \\ q_{61,1} & q_{61,2} & \cdots & q_{61,60} & q_{61,61} & \sum_j q_{61,j} = 0 \end{array}$$



The instantaneous substitution rate from codon i to codon j

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at 2 or 3 codon positions,} \\ \mu\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \mu\kappa\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition.} \end{cases}$$

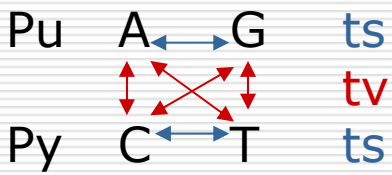
Codon Model

Markov model of codon substitution (61 states or sense codons)

(Goldman & Yang, 1994 Mol Biol Evol, 11, Pp 725; Muse & Gaut, 1994, Mol Biol Evol, 11, Pp 715

Codon instantaneous rate matrix

$$Q = \{q_{ij}\} = \begin{array}{|cccccc|} q_{1,1} & q_{1,2} & \cdots & q_{1,60} & q_{1,61} & \sum_j q_{1,j} = 0 \\ q_{2,1} & q_{2,2} & \cdots & q_{2,60} & q_{2,61} & \sum_j q_{2,j} = 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{60,1} & q_{60,2} & \cdots & q_{60,60} & q_{60,61} & \sum_j q_{60,j} = 0 \\ q_{61,1} & q_{61,2} & \cdots & q_{61,60} & q_{61,61} & \sum_j q_{61,j} = 0 \end{array}$$



The instantaneous substitution rate from codon i to codon j

Frequency of codon j

ts/tv rate ratio

dN/dS rate ratio

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at 2 or 3 codon positions,} \\ \mu\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \mu\kappa\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition.} \end{cases}$$

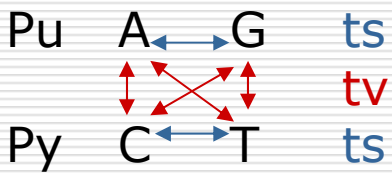
Codon Model

Markov model of codon substitution (61 states or sense codons)

(Goldman & Yang, 1994 Mol Biol Evol, 11, Pp 725; Muse & Gaut, 1994, Mol Biol Evol, 11, Pp 715

Codon instantaneous rate matrix

$$Q = \{q_{ij}\} = \begin{array}{|cccccc|} q_{1,1} & q_{1,2} & \cdots & q_{1,60} & q_{1,61} & \sum_j q_{1,j} = 0 \\ q_{2,1} & q_{2,2} & \cdots & q_{2,60} & q_{2,61} & \sum_j q_{2,j} = 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{60,1} & q_{60,2} & \cdots & q_{60,60} & q_{60,61} & \sum_j q_{60,j} = 0 \\ q_{61,1} & q_{61,2} & \cdots & q_{61,60} & q_{61,61} & \sum_j q_{61,j} = 0 \end{array}$$



The instantaneous substitution rate from codon i to codon j

Frequency of codon j

ts/tv rate ratio

dN/dS rate ratio

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at 2 or 3 codon positions,} \\ \mu\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \mu\kappa\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition.} \end{cases}$$

- **The model accounts for ts/tv bias, unequal synonymous and nonsynonymous substitution, and biased base/codon frequencies.**

Codon Model

For example, consider the substitution rates to codon CTG (Leu). We have:

$$q_{CTC,\,CTG} = \mu\,\pi_{CTG} \quad since\ the\ CTC(Leu) \mapsto CTG(Leu)\ change\ is\ a\ synonymou\ s\ transversi\ on,$$

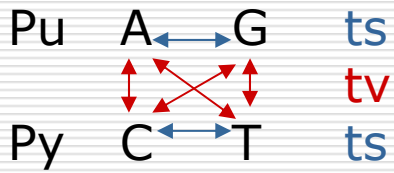
$$q_{TTC,\,CTG} = \mu\,\kappa\pi_{CTG} \quad since\ the\ TTC(Leu) \mapsto CTG(Leu)\ change\ is\ a\ synonymou\ s\ transition\ ,$$

$$q_{GTG,\,CTG} = \mu\,\omega\pi_{CTG} \quad since\ the\ GTG(Val) \mapsto CTG(Leu)\ change\ is\ a\ nonsynonym\ ous\ transversi\ on,$$

$$q_{CCG,\,CTG} = \mu\,\kappa\omega\pi_{CTG} \quad since\ the\ GTG(Pro) \mapsto CTG(Leu)\ change\ is\ a\ nonsynonym\ ous\ transition\ ,$$

$$q_{TTT,\,CTG} = 0 \quad since\ the\ TTT \mapsto CTG\ differ\ at\ more\ than\ one\ position$$

$$q_{ij} = \begin{cases} 0, & if\ i\ and\ j\ differ\ at\ 2\ or\ 3\ codon\ positions, \\ \mu\pi_j, & if\ i\ and\ j\ differ\ by\ a\ synonymou\ s\ transversi\ on, \\ \mu\kappa\pi_j, & if\ i\ and\ j\ differ\ by\ a\ synonymou\ s\ transition\ , \\ \mu\omega\pi_j, & if\ i\ and\ j\ differ\ by\ a\ nonsynonym\ ous\ transversi\ on, \\ \mu\kappa\omega\pi_j, & if\ i\ and\ j\ differ\ by\ a\ nonsynonym\ ous\ transition\ . \end{cases}$$



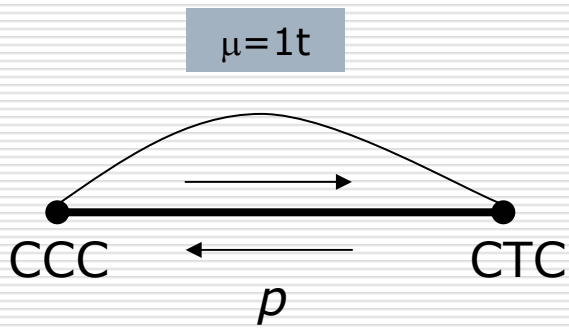
Codon Model

- Molecular sequence data do not allow separate estimation of rate (μ) and time (t), and only their product (μt) can be identified.
- We thus fix the rate μ such that the expected number of nucleotide substitutions per codon is one.
- This scaling means that time (t) is measured by genetic distance -the expected number of (nucleotide) substitutions per codon.
- The transition probability matrix over time t is:

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$

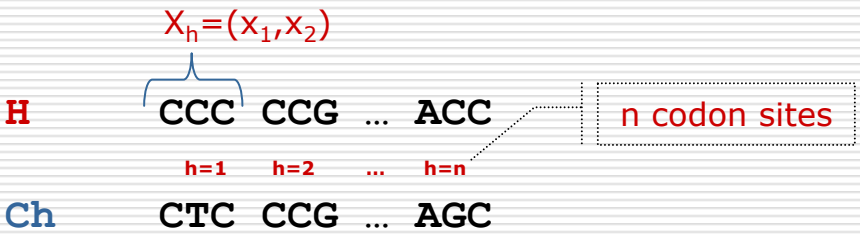
Lastly, the model is time-reversible

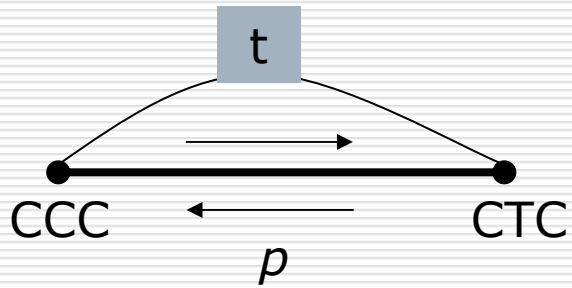
This means, $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$



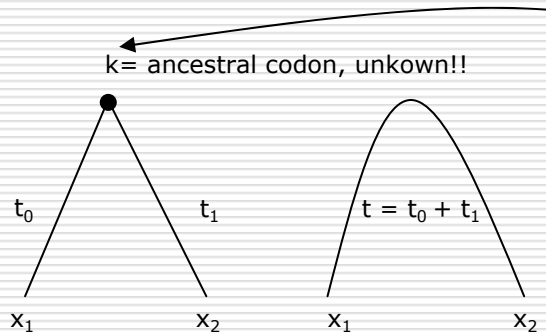
Pairwise Comparison:

Maximum Likelihood (ML) estimation of ω





The probability of site h,



$$f(x_h) = \sum_{k=1}^{61} \pi_k \, p_{k,CCC}(t_0) \, p_{k,CTC}(t_1) = \pi_{CCC} \, p_{CCC,CTC}(t)$$

Pairwise Comparison: Maximum Likelihood (ML) estimation of ω

$$f(x_h) = \pi_{x_1} p_{x_1,x_2}(t)$$

Parameters in the model:

- ✓ the sequence divergence (t),
- ✓ transition/transversion rate ratio (κ),
- ✓ the nonsynonymous/synonymous rate ratio (ω) ,
- ✓ the codon frequency, (π_j)

...estimated from the observed data (base –F3x4- or codon frequencies –F61)

The log-likelihood function for these sequence is then given by:

$$l(t,\kappa,\omega) = \sum_{h=1}^{n} \log f(x_h)$$

Maximum Likelihood (ML) estimation of ω

- Since there is no analytic solution, a numerical optimization algorithm is used to maximize "*l*"

$$l(t,\kappa,\omega) = \sum_{h=1}^{n} \log f(x_h)$$

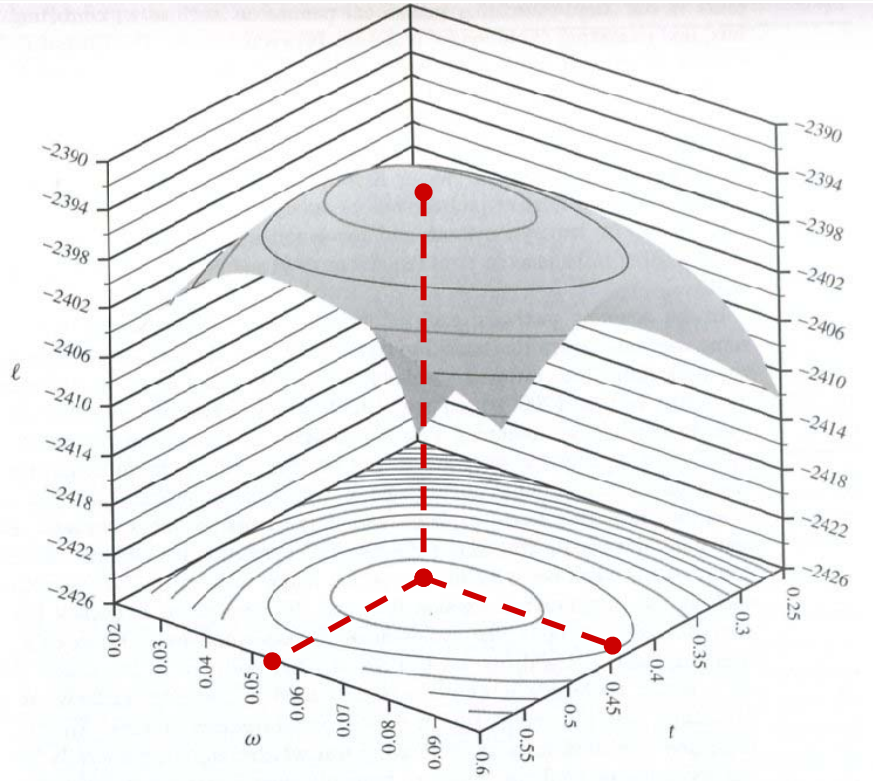


Figure 9.2 The log-likelihood surface contour as a function of parameters t and ω for the comparison of the human and mouse acetylcholine receptor α genes. The maximum likelihood method estimates parameters by maximizing the likelihood function. For these data, the estimates are $\hat{t} = 0.444$, $\hat{\omega} = 0.059$, with optimum log-likelihood $\ell = -2392.83$.

Comparing pairwise methods

Table 5.1. Estimation of d_S and d_N between *Drosophila melanogaster* and *D. simulans GstD1* genes.

Method	κ	S	N	d_S	d_N	ω	ℓ
ML methods							
Fequal, $\kappa = 1$	1	152.9	447.1	0.0776	0.0213	0.274	-927.18
Fequal, κ estimated	1.88	165.8	434.2	0.0221	0.0691	0.320	-926.28
F3×4, $\kappa = 1$	1	70.6	529.4	0.1605	0.0189	0.118	-844.51
F3×4, κ estimated	2.71	73.4	526.6	0.1526	0.0193	0.127	-842.21
F61, $\kappa = 1$	1	40.5	559.5	0.3198	0.0201	0.063	-758.55
F61, κ estimated	2.53	45.2	554.8	0.3041	0.0204	0.067	-756.57
Counting methods							
Nei and Gojobori	1	141.6	458.4	0.0750	0.0220	0.288	
Yang and Nielsen (F3×4)	3.28	76.6	523.5	0.1499	0.0190	0.127	

PAML

Codeml

YN00

Codon bias

Transition bias

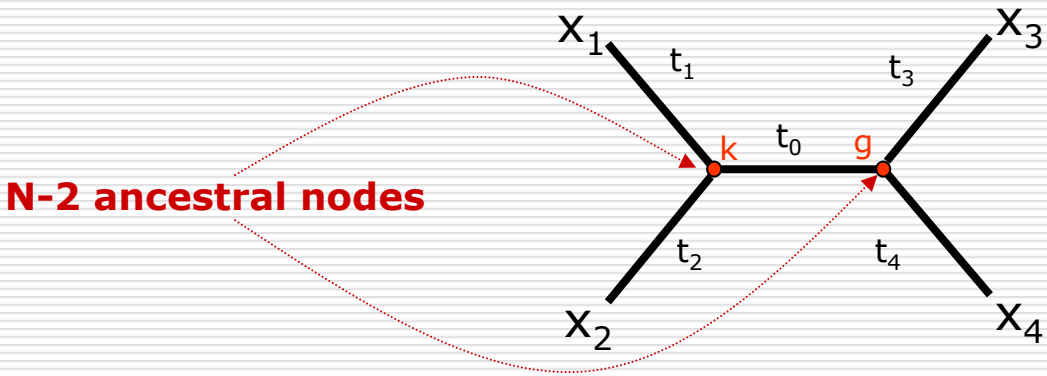
Phylogenetic (ML) estimation of ω

- Pairwise methods has low power to detect positive selection
  - **It averages over all sites along all the evolutionary distance separating sequences.**
    - **Only 17 out of 3600 genes** (*Endo et al. 1996, MolBiolEvol, 13. Pp. 685)*
- Power is improved **if selective pressure is allowed to vary over sites or branches.**
- Increasing the complexity of the codon model in this way requires that likelihood be calculated for **multiple sequences on a phylogeny**.

Likelihood calculation for multiple sequences on a phylogeny

- Given an unrooted tree with N=4 species

$X_h=(x_1,x_2,x_3,x_4)$



The data at each site will be the sum over $61^{(N-2)}$ possible combinations of ancestral nodes

- **The probability of observing the data at codon h, is:**

$$f(x_h)=\sum_k\sum_g\left\{\pi_k p_{k,x_1}(t_1)\pi_k p_{k,x_2}(t_2)\pi_k p_{k,g}(t_0)\pi_g p_{g,x_3}(t_3)\pi_g p_{g,x4}(t_4)\right\}$$

- **The log-likelihood function is the sum over all (*h*=*n*) codon sites:**

$$l(t_{(2N-3)},\kappa,\omega)=\sum_{h=1}^{n}\log\{f(x_h)\}$$

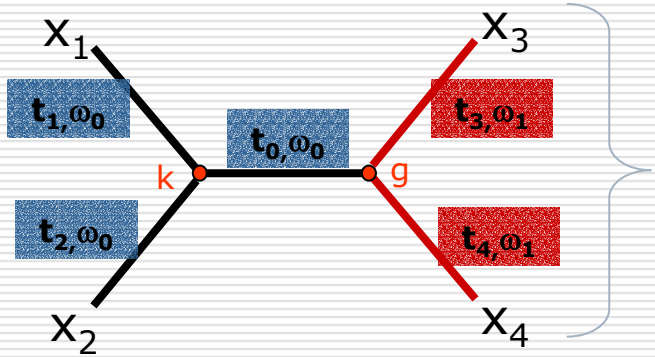
We assumed a single ω value for all the sequences in the tree

Modelling variable selective pressure among lineages

Yang, Z. 1998. Mol. Biol. Evol. 15:568.

- Adaptive evolution is most likely to occur in an episodic fashion.

The most general model called the **"free-ratio model"** specified an independent ω for each branch of the tree



The transition probabilities for the 2 set of branches are calculated from different rate matrices (Q) generated by using different ω ratios

- **The probability of observing the data at codon h, $X_h=(x_1,x_2,x_3,x_4)$**

$$f(x_h)=\sum_k\sum_g\left\{\pi_k p_{k,x_1}(t_1,\omega_0)\pi_k p_{k,x_2}(t_2,\omega_0)\pi_k p_{k,g}(t_0,\omega_0)\pi_g p_{g,x_3}(t_3,\omega_1)\pi_g p_{g,x4}(t_4,\omega_1)\right\}$$

- **The log-likelihood function is the sum over all (*h*=*n*) codon sites:**

$$l(t_{(2N-3)},\kappa,\omega_0,\omega_1)=\sum_{h=1}^{n}\log\{f(x_h)\}$$

2 ω values estimated for different lineages in the tree averaging over all sites!!

Modelling variable selective pressure among lineages

Yang, Z. 1998. Mol. Biol. Evol. 15:568.

- Adaptive evolution is most likely to occur in an episodic fashion.

Table 9.2 Log-likelihood values and parameter estimates under different models for the lysozyme *c* genes.

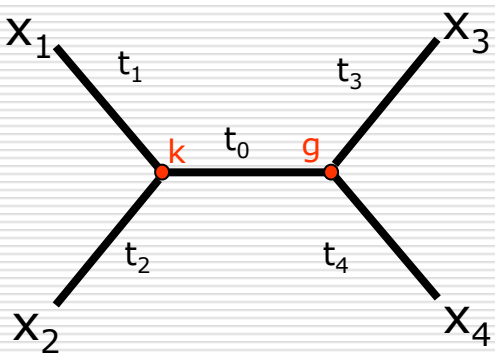
Model	p	ℓ	$\hat{\kappa}$	$\hat{\omega}_0$	$\hat{\omega}_h$	$\hat{\omega}_c$
A. 1 ratio: $\omega_0 = \omega_h = \omega_c$	22	−906.02	4.5	0.81	$= \hat{\omega}_0$	$= \hat{\omega}_0$
B. 2 ratios: $\omega_0 = \omega_h, \omega_c$	23	−904.64	4.6	0.69	$= \hat{\omega}_0$	3.51
C. 2 ratios: $\omega_0 = \omega_c, \omega_h$	23	−903.08	4.6	0.68	∞	$= \hat{\omega}_0$
D. 2 ratios: $\omega_0, \omega_h = \omega_c$	23	−901.63	4.6	0.54	7.26	$= \hat{\omega}_H$
E. 3 ratios: $\omega_0, \omega_h, \omega_c$	24	−901.10	4.6	0.54	∞	3.65
F. 2 ratios: $\omega_0 = \omega_h, \omega_c = 1$	22	−905.48	4.4	0.69	$= \hat{\omega}_0$	1
G. 2 ratios: $\omega_0 = \omega_c, \omega_h = 1$	22	−905.38	4.4	0.68	1	$= \hat{\omega}_0$
H. 2 ratios: $\omega_0, \omega_h = \omega_c = 1$	22	−904.36	4.3	0.54	1	1
I. 3 ratios: $\omega_0, \omega_h, \omega_c = 1$	23	−902.02	4.5	0.54	∞	1
J. 3 ratios: $\omega_0, \omega_h = 1, \omega_c$	23	−903.48	4.4	0.54	1	3.56

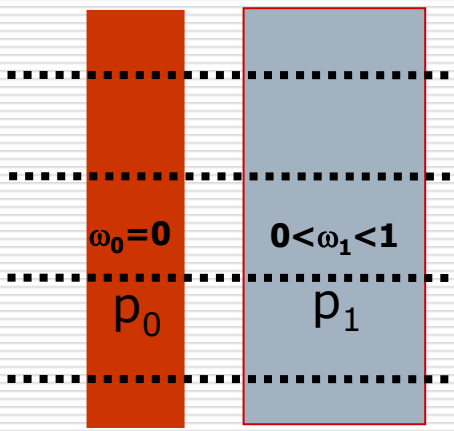
Note: p is the number of parameters. All models include the following 21 common parameters: 11 branch lengths in the tree (Figure 9.5), 9 parameters for base frequencies at codon positions used to calculate codon frequencies, and the transition/transversion rate ratio κ. *Source*: Yang (1998).

Modelling variable selective pressure among sites

Yang, Z. and R. Nielsen, 2002. Mol. Biol. Evol. 19: 908-917.

- The most used approach is to use a statistical distribution to model the random variation in ω over sites (ie.: beta, gamma distribution, etc.)
- Codeml has 13 alternative models available.
- Continuous distributions are approximated by discrete categories: *K*
  - ***K*$_i$ classes, with ω$_i$ different ratios and *p*$_i$ proportions**





If K=2, i=(0,1)

Since we do do not known to wich class each site h belongs, we sum over both classes

$$f(x_h|\omega_i) = \sum_k \sum_g \left\{\pi_k p_{k,x_1}(t_1)\pi_k p_{k,x_2}(t_2)\pi_k p_{k,g}(t_0)\pi_g p_{g,x_3}(t_3)\pi_g p_{g,x4}(t_4)\right\}$$

$$f(x_h) = \sum_{i=0}^{i=1} p_i f(x_h|\omega_i)$$

$$l = \sum_{h=1}^{n} \log\{f(x_h)\}$$

Models of variable ω ratio among sites

Table 5.3. Parameter estimates and likelihood scores under models of variable ω ratios among sites for HIV-2 *nef* genes. (Note: The number after the model code, in parentheses, is the number of free parameters in the ω distribution. The d_N/d_S ratio is an average over all sites in the HIV-2 *nef* gene alignment. Parameters in parentheses are not free parameters and are presented for clarity. PSS is the number of positive selected sites, inferred at the 50% (95%) posterior probability cutoff.)

Model	d_N/d_S	Parameter estimates	PSS	ℓ
M0: one ratio (1)	0.51	$\omega = 0.505$	none	−9775.77
M3: discrete (5)	0.63	$p_0 = 0.48, p_1 = 0.39, (p_2 = 0.13)$ $\omega_0 = 0.03, \omega_1 = 0.74, \omega_2 = 2.50$	31 (24)	−9232.18
M1: neutral (1)	0.63	$p_0 = 0.37, (p_1 = 0.63)$ $(\omega_0 = 0), (\omega_1 = 1)$	not allowed	−9428.75
M2: selection (3)	0.93	$p_0 = 0.37, p_1 = 0.51, (p_2 = 0.12)$ $(\omega_0 = 0), (\omega_1 = 1), \omega_2 = 3.48$	30 (22)	−9392.96
M1a: nearly neutral (2)	0.48	$p_0 = 0.55, (p_1 = 0.45)$ $(\omega_0 = 0.06), (\omega_1 = 1)$	not allowed	−9315.53
M2a: positive selection (4)	0.73	$p_0 = 0.51, p_1 = 0.38, (p_2 = 0.11)$ $(\omega_0 = 0.05), (\omega_1 = 1), \omega_2 = 3.00$	26 (15)	−9241.33
M7: beta (2)	0.42	$p = 0.18, q = 0.25$	not allowed	−9292.53
M8: beta & ω (4)	0.62	$p_0 = 0.89, (p_1 = 0.11)$ $p = 0.20, q = 0.33, \omega = 2.62$	27 (15)	−9224.31

Bayes Empirical Bayes (BEB)

- After ML estimation of model's parameters, an **empirical Bayes approach** is used to infer which ω class each site is most likely to belong.

- The **posterior probability** that site *h* with data x_h is from site *k* is:

$$prob(\omega_k | x_h) = \frac{p_k f(x_h | \omega_k)}{f(x_h)}$$

Bayes empirical bayes results

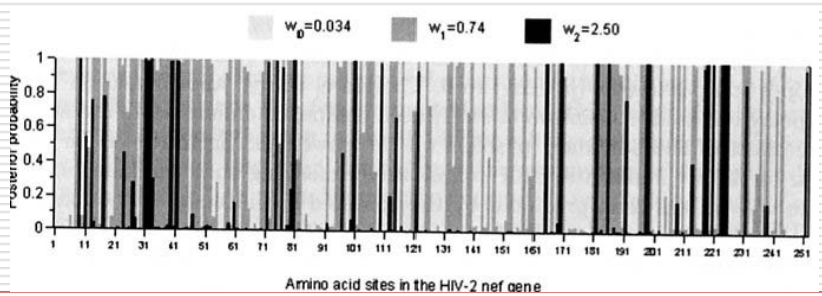
Bayes Empirical Bayes (BEB) analysis
Positively selected sites (*: P>95%; **: P>99%)

		Pr(w>1)	post mean +- SE for w
1	M	0.007	0.156 +- 0.298
2	E	0.009	0.169 +- 0.353
3	E	0.009	0.169 +- 0.353
4	P	0.125	0.893 +- 1.263
5	Q	0.010	0.182 +- 0.370
6	S	0.015	0.212 +- 0.436
7	D	0.010	0.180 +- 0.375
8	P	0.368	2.207 +- 2.473
9	S	0.007	0.160 +- 0.310
10	V	0.139	0.969 +- 1.480
11	E	0.009	0.169 +- 0.353
12	P	0.091	0.722 +- 1.043
13	P	0.014	0.208 +- 0.450
14	L	0.013	0.200 +- 0.411
15	S	0.009	0.178 +- 0.371
16	Q	0.010	0.182 +- 0.370
17	E	0.011	0.186 +- 0.405

Purifying

Positive

Neutral



Yang, Z. and R. Nielsen, 2002. Mol Biol Evol 19. Pp. 908

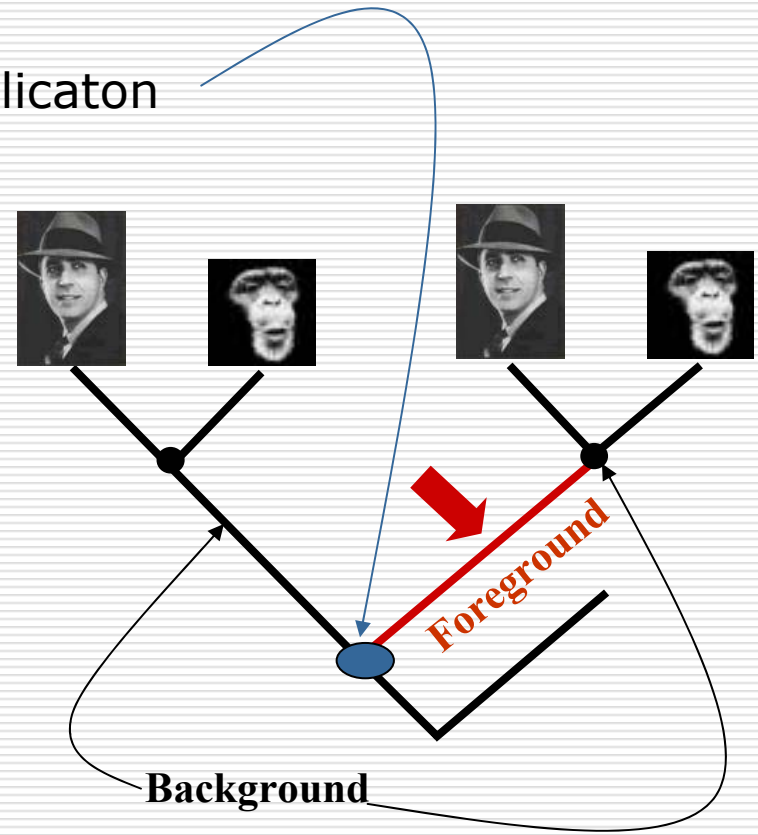
Branch-sites models

- Usefull to test adaptive process ocurred at a subset of sites for a limited period of time.
- Usefull to test adaptation after gene duplicaton

Test I: M_{1a} (neutral) vs **A_1**(Positive selection) models

k= 3, site classes (0, 1, 2)

	$\omega_2>1$ (p2a , p2b)	Positive selection
$\omega_1=1$,	$\omega_1=1$ (p_1)	Neutrality
$0<\omega_0<1$,	$0<\omega_0<1$ (p_0)	Purifyng selection



Likelihood ratio test …nested models comparison

$$l_{M_{1a}} - l_{A_1} = 2\Delta l > \chi^2_{0.05,df=2}$$

Accept the alternative hypothesis of positive selection!!!

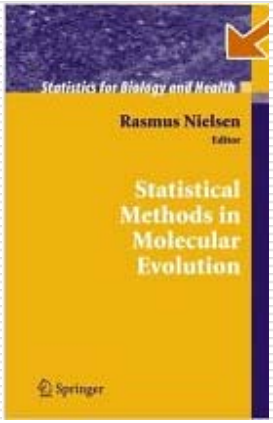
Table 9.3 Likelihood ratio statistics ($2\Delta\ell$) for testing hypotheses concerning lysozyme evolution.

Hypothesis tested	Assumption made	Models compared	$2\Delta\ell$
A. $(\omega_h = \omega_c) = \omega_0$	$\omega_h = \omega_c$	A & D	8.78**
B. $\omega_c = \omega_0$	$\omega_h = \omega_0$	A & B	2.76
C. $\omega_c = \omega_0$	ω_h free	C & E	3.96*
D. $\omega_h = \omega_0$	$\omega_c = \omega_0$	A & C	5.88*
E. $\omega_h = \omega_0$	ω_c free	B & E	7.08**
A′. $(\omega_h = \omega_c) \leq 1$	$\omega_h = \omega_c$	D & H	5.46*
B′. $\omega_c \leq 1$	$\omega_h = \omega_0$	B & F	1.68
C′. $\omega_c \leq 1$	ω_h free	E & I	1.84
D′. $\omega_h \leq 1$	$\omega_c = \omega_0$	C & G	4.60*
E′. $\omega_h \leq 1$	ω_c free	E & J	4.76*

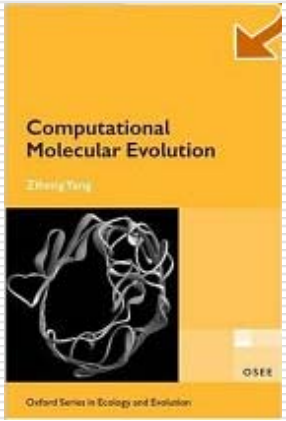
*Significant at the 5% level ($\chi^2(1) = 3.84$).
**Significant at the 1% level ($\chi^2(1) = 6.63$).
Source: Yang (1998).

Thanks!

☐ Acknowledgements



Chapter 5.
Bielawsky J. & Z. Yang, 2005



Computational Molecular Evolution
by Ziheng Yang



Micostrium Vulgaris
Phylum: Chordata
Subphylum: Vertebrata
Class: Mammalia