

Molecular Evolution & Phylogenetics (BIOS 659: Advanced Studies in Genetics)

Assignment I – 21 Sept 2004

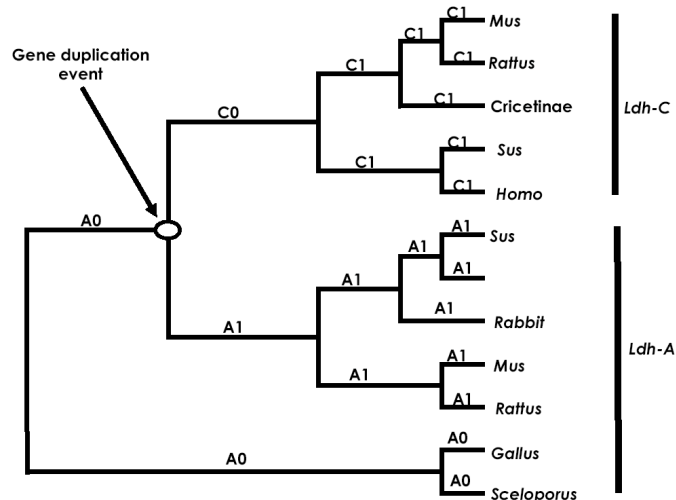
You'll need to use PAML-CodeML & MEGA3 (optional) for the following problems. MEGA is installed on the PC's in room 205; PAML is running on iNquiry (<http://inquiry.egg.isu.edu>). Data required for the following problems is obtained from the course website. Each problem requires the generation of some data and a brief interpretation (1-2 paragraphs). **The assignment is due on Thursday, September 30.**

1. Given the GstD1 genes of *Drosophila melanogaster* & *D. simulans*, consisting of 600 codons (1800bp):
 - a) Evaluate the likelihood for the following ten fixed values of ω : 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 2.0. Plot the results (likelihood score vs. ω). Use runmode= -2 (pairwise); CodonFreq=3 (F61); Model=0; Nssites=0.
 - b) Allow PAML to calculate the value of ω (i.e., no fixed rate).
 - c) Caculate Nei-Gojobori dN/dS (use PAML or MEGA3).
 - d) Determine the effect of codon frequency estimates and κ (tv/ts) fixed values & estimates on calculates of substitution rates. **Complete the following table and provide a brief interpretation**, using runmode= -2 (pairwise); Model=0; Nssites=0; **CodonFreq= 0, 2 or 3 (change according to assumptions); fix_kappa= fixed or 1.0 (change according to assumptions).**

Assumptions	κ	S	N	dS	dN	ω	lnL
Fequal + $\kappa=1$	1.0						
Fequal + $\kappa=est.$							
F3x4 + $\kappa=1$	1.0						
F3x4 + $\kappa=est.$							
F61 + $\kappa=1$	1.0						
F61 + $\kappa=est.$							

2. Conduct a likelihood ratio test (LRT) for variation in the selection pressure among branches of the *Ldh* gene. The *Ldh* gene family is an important model system for molecular evolution of isozyme multigene families. The rate of evolution is known to have increased in in *Ldh-C* following the gene duplication event (see figure). The sequence file and four tree files are available – each tree designates the branches to be tested (check to ensure they reflect the appropriate tests. Test four hypotheses:

- a) $H_0: \omega_{A0}=\omega_{A1}=\omega_{C1}=\omega_{C0}$ – “Was selection equal in all parts of the tree?” To test this, use “**tree2.h0**”, runmode= 0 (user defined tree); CodonFreq=2 (F3x4); **Model=0**; Nssites=0.
- b) $H_1: \omega_{A0}=\omega_{A1}=\omega_{C1}\neq\omega_{C0}$ – “Was selection different in the *Ldh-C* ancestor?” To test this, use “**tree2.h1**”, runmode= 0 (user defined tree); CodonFreq=2 (F3x4); **Model=2**;



- Nssites=0.
- c) H2: $\omega_{A0}=\omega_{A1}\neq\omega_{C1}=\omega_{C0}$ – “Was selection different in the entire *Ldh*-C lineage?” To test this, use “**tree2.h2**”, runmode= 0 (user defined tree); CodonFreq=2 (F3x4); **Model=2**; Nssites=0.
- d) H3: $\omega_{A0}\neq\omega_{A1}\neq\omega_{C1}=\omega_{C0}$ – “Was selection different in the entire *Ldh*-C lineage *and* in the entire *Ldh*-A lineage?” To test this, use “**tree2.h3**”, runmode= 0 (user defined tree); CodonFreq=2 (F3x4); **Model=2**; Nssites=0.

Complete the following table and provide a brief interpretation (1-2 paragraphs):

Model	ω_{A0}	ω_{A1}	ω_{C1}	ω_{C0}	lnL	LRT ^a
H0: $\omega_{A0}=\omega_{A1}=\omega_{C1}=\omega_{C0}$		$=\omega_{A0}$	$=\omega_{A0}$	$=\omega_{A0}$		N/A
H1: $\omega_{A0}=\omega_{A1}=\omega_{C1}\neq\omega_{C0}$		$=\omega_{A0}$	$=\omega_{A0}$			
H2: $\omega_{A0}=\omega_{A1}\neq\omega_{C1}=\omega_{C0}$		$=\omega_{A0}$		$=\omega_{C1}$		
H3: $\omega_{A0}\neq\omega_{A1}\neq\omega_{C1}=\omega_{C0}$				$=\omega_{C1}$		

a: the LRT's correspond to H0 vs. H1, H0 vs. H2, and H2 vs. H3, respectively (df=3); provide a p-value for the LRT's and indicate which (if any) are significant.

3. Test for adaptive evolution among sites in the *nef* gene of human HIV-2. The data file contains 44 *nef* alleles from a population of 37 humans infected with HIV-2. You will test 3 pairs of nested hypotheses. Use “**tree3**”, runmode= 0 (user defined tree); CodonFreq=2 (F3x4); Model=0; Nssites=**0, 1, 2, 3, 7 or 8 (change according to model used)**; ncatG=**2 (for M0, M1, M2), 3 (for M3) or 10 (for M7, M8)**.

A.

Model	ω_0	ω_1	PSS	lnL	LRT
H0: one ratio (M0, 1)		$=\omega$	N/A		
H1: variable (M3, 5)					

B.

Model	ω_0	ω_1	PSS	lnL	LRT
H0: neutral (M1, 1)			N/A		
H1: selection (M2, 3)					

C.

Model	ω_0	ω_1	PSS	lnL	LRT
H0: beta (M7, 2)			N/A		
H1: beta & ω (M8, 4)					

For each table: the numbers in parentheses after the hypothesis is the model number (from Bielawski & Yang, 2003) and the number of free parameters, ω_0 is the ratio averaged over all sites, ω_1 is the ratio for sites allowed to vary, and PSS is the number of positively selected sites (by definition: none under M0 and not allowed under M1 & M7).

Highlight the hypothesis that is implicated by the LRT and provide a brief interpretation (1-2 paragraphs).