Molecular Evolution, course # 27615

Final exam, May 6, 2009

This exam includes a total of 8 problems on 8 pages (including this cover page). The maximum number of points obtainable is 240, and at least 130 points are required to pass. Note that different problems are worth different amounts of points. If a problem has more than one sub-problem, then each sub-problem has the same weight. If a solution is only partially right, it will give a fraction of the total amount of points. Problems should be answered in the space provided on these pages.

The duration of the exam is 4 hours (13:00-17:00). All non-living resources may be used (books, internet including exercise web-pages, notes, etc., but no email, skype, etc). Coffee and cookies will be served.

Have fun (!),
Anders

Table of chi-square critical values

	Significance level						
DF	0.9950	0.9750	0.9000	0.5000	0.1000	0.0500	0.0100	0.0010
1	0.0000	0.0010	0.0158	0.4549	2.7055	3.8415	6.6349	10.8276
2	0.0100	0.0506	0.2107	1.3863	4.6052	5.9915	9.2103	13.8155
3	0.0717	0.2158	0.5844	2.3660	6.2514	7.8147	11.3449	16.2662
4	0.2070	0.4844	1.0636	3.3567	7.7794	9.4877	13.2767	18.4668
5	0.4117	0.8312	1.6103	4.3515	9.2364	11.0705	15.0863	20.5150
6	0.6757	1.2373	2.2041	5.3481	10.6446	12.5916	16.8119	22.4577
7	0.9893	1.6899	2.8331	6.3458	12.0170	14.0671	18.4753	24.3219
8	1.3444	2.1797	3.4895	7.3441	13.3616	15.5073	20.0902	26.1245
9	1.7349	2.7004	4.1682	8.3428	14.6837	16.9190	21.6660	27.8772
10	2.1559	3.2470	4.8652	9.3418	15.9872	18.3070	23.2093	29.5883
11	2.6032	3.8157	5.5778	10.3410	17.2750	19.6751	24.7250	31.2641
12	3.0738	4.4038	6.3038	11.3403	18.5493	21.0261	26.2170	32.9095
13	3.5650	5.0088	7.0415	12.3398	19.8119	22.3620	27.6882	34.5282
14	4.0747	5.6287	7.7895	13.3393	21.0641	23.6848	29.1412	36.1233
15	4.6009	6.2621	8.5468	14.3389	22.3071	24.9958	30.5779	37.6973

1: (20 points)

For each of the following observed branch lengths compute the corresponding distances corrected for multiple substitutions assuming that the sequences in question evolve according to the Jukes and Cantor model. Also compute the ratio between the corrected and the observed distance. Above what branch length is the correction larger than 10%?

Branch length (v)	JC corrected (JC)	Ratio (JC/v)
0.01	0.01004	1.004
0.05	0.0517	1.034
0.10	0.107	1.070
0.25	0.304	1.216
0.50	0.824	1.648
1.00	∞	∞
2.00	∞	∞

$$D_{JC} = -\frac{3}{4}\ln\left(1-\frac{4}{3}D_{OBS}\right)$$

2: (15 points)

You have a set of 12 aligned sequences to which you have fitted two alternative models: (1) K2P, and (2) HKY with gamma-distributed rates. The maximized log likelihoods were found to be:

lnL(K2P) = -4378.98 lnL(HKY+gamma) = -4369.23

Use a likelihood ratio test and a significance level of 0.01 to decide whether the HKY+gamma model is significantly better than the K2P model. Explain how you decided the degrees of freedom to use.

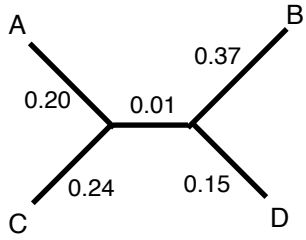
$$\Delta = 2\left(\ln L(HKY+gamma) - \ln L(K2P)\right) = 2(-4369.23 - -4378.98) = 19.5$$

```
K2P has 1 free parameter:
      1 transition/transversion rate ratio

HKY + gamma has 5 free parameters:
      1 transition/transversion rate ratio
      3 nucleotide frequencies (the fourth is not free since they have to sum to one)
      1 gamma distribution shape parameter

df = 5-1 = 4

Compare test statistic (19.5) to chi-square table with 4 degrees of freedom:
      19.5>13.2767   =>  HKY+gamma is significantly better than K2P
```

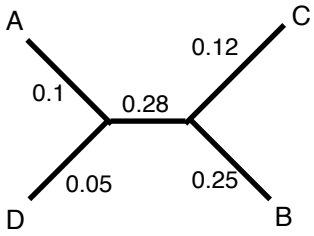
3: (40 points)

Below we have shown a distance matrix giving the pairwise distances between a set of 4 sequences. Also shown are two alternative trees with branch lengths indicated.

	A	B	C	D
A	-	0.67	0.54	0.15
B		-	0.41	0.62
C			-	0.49
D				-



Tree 1



Tree 2

3.1: Compute the tree length for each of the two trees:

Tree 1: 0.20+0.24+0.01+0.37+0.15 = 0.97
Tree 2: 0.1+0.05+0.28+0.12+0.25 = 0.80

3.2: Compute the sum of squared errors for each of the two trees:

Taxon pair	Observed dist	T1 dist	T1 squared error	T2 dist	T2 squared error
AB	0.67	0.58	0.0081	0.63	0.0016
AC	0.54	0.44	0.01	0.50	0.0016
AD	0.15	0.36	0.0441	0.15	0
BC	0.41	0.62	0.0441	0.37	0.0016
BD	0.62	0.52	0.01	0.58	0.0016
CD	0.49	0.40	0.0081	0.45	0.0016
SUM			0.1244		0.008

3.3: Which tree is better according to the minimum evolution criterion? Which tree is better according to the "least squares" criterion?

Least squares: Tree 2 is better because it has the smallest sum of squared errors
Minimum evolution: Tree 2 is better because it has the smallest tree length

4: (30 points)

You are investigating a data set with homologous sequences from 21 different species of insects.

4.1: How many possible bifurcating, unrooted trees are there for this data set?

$$N = \prod_{i=2}^{20} (2i - 3) = 8.2 \times 10^{21}$$

4.2: Imagine you have randomly selected one of these trees. How many "neighboring" trees can you reach from this tree using the NNI type of tree rearrangement? Explain the formula used.

An unrooted, bifurcating tree with n tips has 2n-3 branches. Of these n are external (they lead to a tip - recall there are n tips). Therefore n-3 of the branches are internal. The present tree therefore has 21-3 = 18 internal branches.

NNI can construct two neighbors for each internal branch, giving:

$N_{neighbors}$ = 2 x 18 = 36

4.3: How many rounds of NNI re-arrangements will the heuristic search algorithm at least need to go through before it has investigated more than 1,000,000 different trees for this data set? (One round of re-arrangements consists of looking at all the NNI-neighbors of the currently best tree, as in problem 4.2).

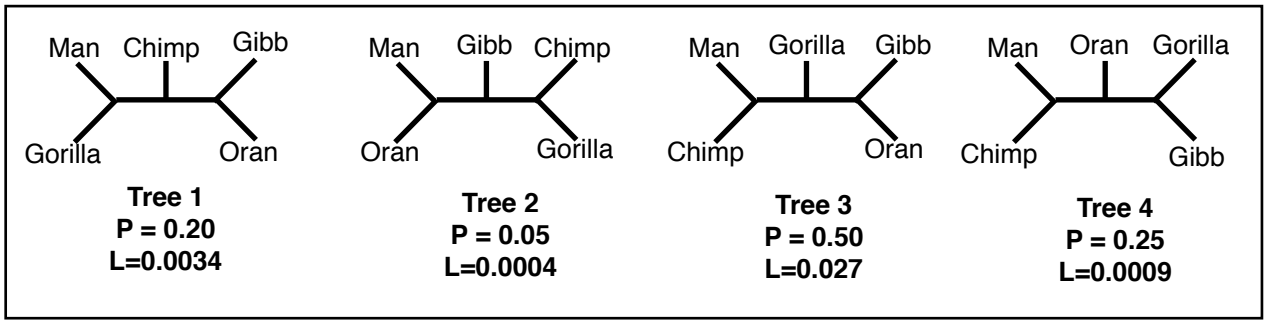
In each round of NNI there are 36 neighbors to investigate. Except for the first round one of these has already been visited (that's where the present tree was reached from). Therefore 35 new trees are investigated each round, giving:

N = 1,000,000 / 35 = 28572 rounds

5: (50 points)

You are investigating a data set with 5 aligned primate sequences using Bayesian statistics. Below is a list of the 4 trees that you believe represent plausible hypotheses about the phylogeny. Also listed is your prior probability (P) and the likelihood (L) for each tree.

5.1: Compute the posterior probability for each tree, and indicate the maximum posterior tree.



Tree	Prior	Likelihood	Prior x likelihood	Posterior
1	0.20	0.0034	0.00068	0.05
2	0.05	0.0004	0.00002	0.00
3	0.50	0.0270	0.0135	0.94
4	0.25	0.0009	0.000225	0.02
SUM			0.014425	1.00

5.2: Based on the above analysis: what is the posterior probability that the closest relative of man is Chimp (i.e., the posterior probability of the bipartition Man+Chimp vs. Gorilla+Oran+Gibb)?

Man + Chimp are present as a monophyletic group in tree 3 and tree 4, so the posterior probability of this group is found by adding the posteriors for these two trees:

P = P(3) + P(4) = 0.94 + 0.02 = 0.96

6: (20 points)

You are investigating the growth of a virus that infects human liver cells. As a result of infecting one cell the virus produces 1000 new copies of itself. 20% of these are functional and go on to infect new cells. Imagine that in the first round of replication (immediately after infecting the liver), a single new mutated version of the virus occurs among the 1000 offspring. This mutant virus is also able to produce 1000 copies of itself, but it has a higher survival rate: 25% of its offspring eventually infect new cells. One viral life cycle (generation time) lasts 48 hours. How much time will elapse before the new virus makes up 90% of the viral population?

$p = f_A = 1/1000, \quad R_A = 25\% \times 1000 = 250$
$q = f_a = 999/1000, \quad R_a = 20\% \times 1000 = 200$

$$p_t = \frac{p}{p + q\left(\frac{R_a}{R_A}\right)^t} > 0.9 \quad \Leftrightarrow \quad p > 0.9p + 0.9q\left(\frac{R_a}{R_A}\right)^t$$

$$\Leftrightarrow \left(\frac{R_a}{R_A}\right)^t < \frac{0.1p}{0.9q} \quad \Leftrightarrow \quad t > \frac{\ln\left(\frac{0.1p}{0.9q}\right)}{\ln\left(\frac{R_a}{R_A}\right)}$$

$$\Leftrightarrow t > 40.8 \text{generations}$$

$$\Leftrightarrow t > 81.6 \text{days}$$

7: (15 points)

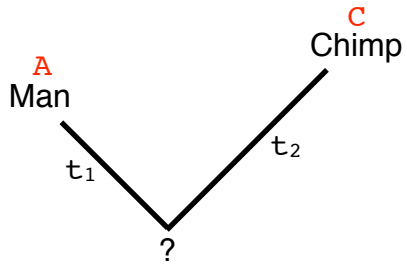
Write the PAUP commands that you would use to (1) load the sequence file "swineflu.nexus", (2) specify that you want to conduct a maximum likelihood analysis, (3) using the K2P model, and (4) perform a branch-and-bound search for the best tree.

```
(1) paup swineflu.nexus    (or: paup, execute swineflu.nexus)

(2) set criterion=likelihood

(3) lset nst=2 basefreq=equal  (optional: tratio=estimate)

(4) bandb
```

8: (50 points)

Shown below is a rooted phylogenetic tree for Man and Chimp. The two branches in the tree have the lengths t_1 and t_2 as indicated. Shown below the tree is the corresponding alignment of sequences (both sequences are only 1 nucleotide long). At the right are the two substitution probability matrices corresponding to the branch lengths t_1 and t_2. The equilibrium frequencies of the nucleotides are: π_A=0.6, π_C=0.1, π_G=0.2, π_T=0.1

What is the maximum likelihood estimate of the nucleotide present in the ancestor of chimp and man (i.e., at the root node indicated by the question mark)? (The full computation must be shown).



Man: A
Chimp: C

To

From	A	C	G	T
A	0.896	0.026	0.052	0.026
C	0.156	0.767	0.052	0.026
G	0.156	0.026	0.793	0.026
T	0.156	0.026	0.052	0.767

t_1

To

From	A	C	G	T
A	0.780	0.055	0.110	0.055
C	0.330	0.504	0.110	0.055
G	0.330	0.055	0.559	0.055
T	0.330	0.055	0.110	0.504

t_2

The ancestral nucleotide (present at the node labeled "?" in the tree) can be either A, C, G, or T. For each of these possibilities I compute the likelihood (i.e., the probability of the data given the model). Starting at the "Man" node the general likelihood expression becomes: $P(data|model) = \pi_A P_{A?}(t_1)P_{?C}(t_2)$

A: $\pi_A P_{AA}(t_1)P_{AC}(t_2) = 0.6 \times 0.896 \times 0.055 = 0.02957$ ←

C: $\pi_A P_{AC}(t_1)P_{CC}(t_2) = 0.6 \times 0.026 \times 0.504 = 0.00786$

G: $\pi_A P_{AG}(t_1)P_{GC}(t_2) = 0.6 \times 0.052 \times 0.055 = 0.00172$

T: $\pi_A P_{AT}(t_1)P_{TC}(t_2) = 0.6 \times 0.026 \times 0.055 = 0.00086$

(the same result would have been found regardless of where the computation was started, since the substitution model used here is time-reversible by design).

The maximum likelihood is found for "A" meaning this is the best estimate of the ancestral nucleotide.