Molecular Evolution, course #27615

Final Exam, May 14, 2008

This exam includes a total of 12 problems on 11 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

Name:______________________________
Student ID number:__________________
Signature:__________________________
Table of chi-square critical values

<table>
<thead>
<tr>
<th>DF</th>
<th>0.9950</th>
<th>0.9750</th>
<th>0.9000</th>
<th>0.5000</th>
<th>0.1000</th>
<th>0.0500</th>
<th>0.0100</th>
<th>0.0010</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0000</td>
<td>0.0010</td>
<td>0.0158</td>
<td>0.4549</td>
<td>2.7055</td>
<td>3.8415</td>
<td>6.6349</td>
<td>10.8276</td>
</tr>
<tr>
<td>2</td>
<td>0.0100</td>
<td>0.0506</td>
<td>0.2107</td>
<td>1.3863</td>
<td>4.6052</td>
<td>5.9915</td>
<td>9.2103</td>
<td>13.8155</td>
</tr>
<tr>
<td>3</td>
<td>0.0717</td>
<td>0.2158</td>
<td>0.5844</td>
<td>2.3660</td>
<td>6.2514</td>
<td>7.8147</td>
<td>11.3449</td>
<td>16.2662</td>
</tr>
<tr>
<td>4</td>
<td>0.2070</td>
<td>0.4844</td>
<td>1.0636</td>
<td>3.3567</td>
<td>7.7974</td>
<td>9.4877</td>
<td>13.2767</td>
<td>18.4668</td>
</tr>
<tr>
<td>5</td>
<td>0.4117</td>
<td>0.8312</td>
<td>1.6103</td>
<td>4.3515</td>
<td>9.2364</td>
<td>11.0705</td>
<td>15.0863</td>
<td>20.5150</td>
</tr>
<tr>
<td>6</td>
<td>0.6757</td>
<td>1.2373</td>
<td>2.2041</td>
<td>5.3481</td>
<td>10.6446</td>
<td>12.5916</td>
<td>16.8119</td>
<td>22.4577</td>
</tr>
<tr>
<td>7</td>
<td>0.9893</td>
<td>1.6899</td>
<td>2.8331</td>
<td>6.3458</td>
<td>12.0170</td>
<td>14.0671</td>
<td>18.4753</td>
<td>24.3219</td>
</tr>
<tr>
<td>8</td>
<td>1.3444</td>
<td>2.1797</td>
<td>3.4895</td>
<td>7.3441</td>
<td>13.3616</td>
<td>15.5073</td>
<td>20.0902</td>
<td>26.1245</td>
</tr>
<tr>
<td>13</td>
<td>3.5650</td>
<td>5.0088</td>
<td>7.0415</td>
<td>12.3398</td>
<td>19.8119</td>
<td>22.3620</td>
<td>27.6882</td>
<td>34.5282</td>
</tr>
</tbody>
</table>

The Genetic code:

![Genetic code diagram]
1: (20 points):

Above we have shown an unrooted tree with 5 leaves. Make a sketch of all the “neighboring” trees that can be constructed using NNI (nearest neighbor interchange) starting from this tree.
2: (10 points)

Imagine that you want to compare the fit of the following two models to a DNA data set using a likelihood ratio test:

- Jukes and Cantor (JC)
- General Time Reversible model (GTR)

How many degrees of freedom should you use for the chi-squared test when determining the p-value?

3: (15 points)

Below, we have shown a distance matrix giving the observed (uncorrected), pairwise distances between four sequences. These distances do not take multiple substitutions into account. Assuming the Jukes and Cantor model of evolution, compute the estimated, corrected distance between each pair of sequences (enter the results in the empty table).

<table>
<thead>
<tr>
<th></th>
<th>Man</th>
<th>Chimp</th>
<th>Gorilla</th>
<th>Gibbon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man</td>
<td>-</td>
<td>0.24</td>
<td>0.35</td>
<td>0.64</td>
</tr>
<tr>
<td>Chimp</td>
<td>-</td>
<td>-</td>
<td>0.37</td>
<td>0.59</td>
</tr>
<tr>
<td>Gorilla</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.68</td>
</tr>
<tr>
<td>Gibbon</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Man</th>
<th>Chimp</th>
<th>Gorilla</th>
<th>Gibbon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chimp</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gorilla</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gibbon</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
4: (10 points)

Shown below is a short protein-encoding DNA sequence (dots indicate codon-codon boundaries):

\[ 5'-\text{GGT.CAT.AGG.TCT.CGT.AAC.GAC}-3' \]

What fraction of the possible point mutations will result in an amino acid change?

5: (15 points)

Shown on the right is an unrooted phylogenetic tree relating 6 vertebrate species.

Q 5.1: Given that “Frog” can be used as an outgroup, make a sketch of the *rooted* phylogenetic tree below.

Q 5.2: Are hippopotamuses more closely related to whales or to sheep?
Q 6.1: Assuming that all ancestral nucleotides were “A”, compute the probability of the data (the likelihood) given the information provided here.

Q 6.2: It is also possible that all ancestral nucleotides were C (there are many other possibilities of course). Which of the two possibilities (“all ancestral nucleotides = A” or “all ancestral nucleotides = C”) is more likely based on the information provided here?
7: (20 points)

In a population of 100 haploid, asexual organisms a mutation at a certain locus has given rise to a single individual with a novel allele (named “New”). The remaining 99 individuals have the allele “Old” at this locus. The Old allele has an absolute fitness of 2.15 offspring/generation. Assume the population exhibits exponential growth with discrete, non-overlapping generations, and that genetic drift and mutation can be ignored. After 20 generations of growth the New allele is found to constitute 50% of the population. What is the absolute fitness of the New allele?

8: (20 points)

Construct a majority rule consensus tree based on the set of trees below. Label each branch in the consensus tree with the corresponding frequency.
Imagine that you are investigating a data set consisting of DNA sequences from 5 species. You are considering two different hypotheses about the phylogenetic tree relating them (see below): Tree number 1, which is strictly bifurcating, and tree number 2, where the three bird species (Great Tit, Pygmy Parrot, and Golden Eagle) form a polytomy. You have determined the maximum likelihood for the K2P model fitted to the data on each of these trees (log likelihood values are shown below the respective trees). Use a likelihood ratio test to investigate whether tree number 1 is significantly better than tree number 2. Use a significance level of 5%.

<table>
<thead>
<tr>
<th></th>
<th>Tree 1</th>
<th>Tree 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>lnL</td>
<td>-28904.25</td>
<td>-28906.37</td>
</tr>
</tbody>
</table>

Tree 1

Tree 2
10: (10 points)

Indicate which of the following four sets of species that do NOT form a monophyletic group (pick as many as relevant).

(A) Birds + Crocodiles
(B) Crocodiles + Snakes + Turtles
(C) Mammals + Turtles
(D) Snakes + Turtles

11: (15 points)

Below is an alignment of two 50 bp long DNA sequences:

\[
\text{Seq}_1: \text{AGCTGGCATTGCTTGCATGCATCTTGTCATATGGCAGCGCGCTTGCAGTA} \\
\text{Seq}_2: \text{AGCAGAGTCTCCATGTTTCCATCTTGCCTTACAGTAGTGTTGGAGAA}
\]

Assume the sequences are evolving according to the Kimura 2-parameter model of evolution

Q 11.1: What is the estimated pairwise distance corrected for multiple substitutions?

Q 11.2: What is the transition transversion rate ratio?
Imagine that you are investigating a single branch in a phylogenetic tree where there has been a change from A to G:

\[ \text{A} \rightarrow \text{G} \]

Further imagine that you have somehow determined that the Kimura 2-parameter model is a suitable description in this case, and that the transition/transversion rate ratio has a value of R=15.

Your first goal is to find the maximum likelihood estimate of the branch length “t”. Due to technical considerations you limit yourself to a discretized version of the problem where you only consider the following five possible branch lengths:

\[
\begin{align*}
    t &= 1.0 \\
    t &= 1.5 \\
    t &= 2.0 \\
    t &= 2.5 \\
    t &= 3.0
\end{align*}
\]

**Q 12.1:** Find the likelihood of the 5 possible branch lengths, and determine which one is the maximum likelihood estimate of “t”. Use the fact that under the Kimura 2-parameter model of evolution, the probability that a given nucleotide will undergo a transition-type mutation, depends on the branch length (t) and the transition/transversion rate ratio (R) according to the equation shown below. Also recall that under K2P all four nucleotides are assumed to have the same frequency.

\[
P(\text{transition}) = 0.25 - 0.5 \times \exp\left( t \times \frac{-2R - 1}{R + 1} \right) + 0.25 \times \exp\left( t \times \frac{-2}{R + 1} \right)
\]
Q 12.2: You also want to perform a Bayesian analysis of the same problem. Based on a previous analysis of the data set, you have the following prior probabilities for the five possible values of the branch length:

\[
\begin{align*}
P(t = 1.0) &= 0.25 \\
P(t = 1.5) &= 0.55 \\
P(t = 2.0) &= 0.10 \\
P(t = 2.5) &= 0.05 \\
P(t = 3.0) &= 0.05 
\end{align*}
\]

Given these priors, compute the posterior probability of all 5 possible values, and determine the maximum posterior estimate of the branch length “t”: 