Molecular Evolution, course #27615

Final Exam, May 2, 2007

This exam includes a total of 14 problems on 8 pages (including this cover page). The maximum number of points obtainable is 150, and at least 85 points are required to pass. Note that different problems are worth different amounts 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. You should, however, not contact your uncle who is a professor of evolutionary biology at Harvard). You can work on your own, or in the groups you worked in during the computer exercises. Please do not discuss the test outside your group.

Members of the group must sign below:

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>2.7055</td>
<td>3.8415</td>
<td>6.6349</td>
<td>10.8276</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0100</td>
<td>0.0506</td>
<td>0.2107</td>
<td>4.6052</td>
<td>5.9915</td>
<td>9.2103</td>
<td>13.8155</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.0717</td>
<td>0.2158</td>
<td>0.5844</td>
<td>2.3660</td>
<td>7.8147</td>
<td>11.3449</td>
<td>16.2662</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.2070</td>
<td>0.4844</td>
<td>1.0636</td>
<td>3.5657</td>
<td>7.7994</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](image-url)
1: (10 points)

Below we have listed four widely used nucleotide substitution models and a list of four PAUP model specifications. Use lines to connect the model names with the corresponding PAUP specifications:

(A) Felsenstein 81 (F81)  (1) lset nst=1 basefreq=empirical
(B) General Time Reversible model (GTR)  (2) lset nst=2 basefreq=equal tratio=estimate
(C) Jukes and Cantor model (JC)  (3) lset nst=1 basefreq=equal
(D) Kimura 2 parameter model (K2P)  (4) lset nst=6 basefreq=empirical rmatrix=estimate

2: (5 points)

You have a rooted, bifurcating tree with 89 leaves. How many branches are there in this tree?

3: (15 points)

You are using parsimony to reconstruct the phylogeny of a set of 10 sequences. Each of your sequences is 1 nucleotide long. What is the length of the two trees below? Which tree is the best one according to the parsimony criterion?

4: (15 points)

Below is an alignment of two 47 bp long DNA sequences:

Seq_1: AGCTGCGATCTTGTGGATCATCTTGATCATATGGCAGCGTTGCAGTA
Seq_2: AGCAGAGTCTCCATGTTTCCCATCTTGCCTTACAGTAGCATTGGAGAA

What is the (uncorrected) distance between these two sequences?

The above result does not take multiple substitutions into account. Assuming the Jukes and Cantor model of evolution, compute the estimated, corrected distance between the sequences:

Now, assume the Kimura 2-parameter model of evolution, and again compute the estimated, corrected distance between the sequences:
5: (5 points)

Two sequences are evolving according to the Kimura 2 parameter model of evolution. Which of the following statements are true (select one):

1) \( P_{A\rightarrow C} (1 \text{ mio}) > P_{C\rightarrow A} (1 \text{ mio}) \)
2) \( P_{A\rightarrow C} (1 \text{ mio}) = P_{C\rightarrow A} (1 \text{ mio}) \)
3) \( P_{A\rightarrow C} (1 \text{ mio}) < P_{C\rightarrow A} (1 \text{ mio}) \)

where \( P_{A\rightarrow C} (1 \text{ mio}) \) is the probability that an A will be substituted with a C in 1 million years, while \( P_{C\rightarrow A} (1 \text{ mio}) \) is the probability of the reverse event (i.e., that a C will be substituted with an A in 1 million years).

6: (5 points)

The posterior probability of a model is (select one):

1) The probability of the data given the model – \( P(\text{Data} | \text{Model}) \)
2) The probability of the model given the data – \( P(\text{Model} | \text{Data}) \)
3) The probability of the model before any evidence has been collected – \( P(\text{Model}) \)
4) The total probability of the data – \( P(\text{Data}) \)

7: (5 points)

Heuristic searching of tree-space is guaranteed to always find the best possible tree:

1) True
2) False

8: (10 points)

Below we have listed the names of five different methods for reconstructing phylogenies as well as five very brief descriptions of the methods. Use lines to connect the method names with the corresponding descriptions:

(A) Parsimony
(B) Least squares
(C) Maximum likelihood
(D) Bayesian
(E) Minimum evolution

(A) Parsimony: (1) The ultimate goal of this method is to find the full posterior probability distribution over trees and other parameters

(B) Least squares: (2) The best tree is the one implying fewest mutations

(C) Maximum likelihood: (3) For a given tree topology, the best branch lengths are those having the smallest deviation between observed and patristic distances; the best tree is the shortest one.

(D) Bayesian: (4) The best tree, and the best branch lengths for a given tree, are those having the smallest deviation between observed and patristic distances.

(E) Minimum evolution: (5) The best tree is the one giving the highest probability of the aligned sequences having evolved from some common ancestor.
9: (25 points)

Below we have shown an unrooted tree relating four present-day sequences. Each sequence is 1 nucleotide long. As the result of archeological work, and using PCR, we have managed to also find the ancestral sequences, which are given at the two internal nodes.

The internal branch has the length $t_1$ while the four external branches all have the length $t_2$. Indicated below the tree are two time-reversible substitution-probability matrices corresponding to these two branch lengths (specifically, these are of the F81 type). The equilibrium frequencies of the nucleotides are: $\pi_A = 0.2$, $\pi_C = 0.3$, $\pi_G = 0.4$, $\pi_T = 0.1$. Compute the probability of the data (the likelihood) given the information provided here.

\[
\begin{array}{cccc}
& A & C & G & T \\
A & 0.685 & 0.118 & 0.157 & 0.039 \\
C & 0.079 & 0.725 & 0.157 & 0.039 \\
G & 0.079 & 0.118 & 0.764 & 0.039 \\
T & 0.079 & 0.118 & 0.157 & 0.646 \\
\end{array}
\]

\[
\begin{array}{cccc}
& A & C & G & T \\
A & 0.775 & 0.084 & 0.112 & 0.028 \\
C & 0.056 & 0.803 & 0.112 & 0.028 \\
G & 0.056 & 0.084 & 0.831 & 0.028 \\
T & 0.056 & 0.084 & 0.112 & 0.747 \\
\end{array}
\]
10: (20 points)

On a cold and rainy Sunday afternoon you decide to cheer yourself up by performing a Bayesian analysis of a dataset you were investigating in a previous, extremely interesting exercise on a Molecular Evolution course (taught by the charismatic Professor Sredna Mrog Nesredep). As part of this analysis you are using a discretized version of the JC+gamma model, where you are considering just 8 possible values of the gamma distribution shape parameter (denoted $\alpha$). Based on a previous analysis of a similar data set you have specified the following prior probabilities for the 8 possible parameter values:

Prior probabilities:
- $P(\alpha=0.1) = 0.05$
- $P(\alpha=0.3) = 0.15$
- $P(\alpha=0.5) = 0.35$
- $P(\alpha=0.7) = 0.15$
- $P(\alpha=0.9) = 0.10$
- $P(\alpha=1.1) = 0.10$
- $P(\alpha=1.3) = 0.05$
- $P(\alpha=1.5) = 0.05$

For each of these possible values you have furthermore computed the probability of the data given that particular parameter value (the likelihood):

Likelihoods:
- $P(D | \alpha=0.1) = 0.01$
- $P(D | \alpha=0.3) = 0.05$
- $P(D | \alpha=0.5) = 0.15$
- $P(D | \alpha=0.7) = 0.23$
- $P(D | \alpha=0.9) = 0.89$
- $P(D | \alpha=1.1) = 0.55$
- $P(D | \alpha=1.3) = 0.21$
- $P(D | \alpha=1.5) = 0.05$

Find the posterior probability of the eight possible parameter values:
- $P(\alpha=0.1 | D) =$
- $P(\alpha=0.3 | D) =$
- $P(\alpha=0.5 | D) =$
- $P(\alpha=0.7 | D) =$
- $P(\alpha=0.9 | D) =$
- $P(\alpha=1.1 | D) =$
- $P(\alpha=1.3 | D) =$
- $P(\alpha=1.5 | D) =$
11: (10 points)

You are comparing the fit of two alternative substitution models (M0 and M3) to a data set derived from several species of Puerto Rican bats. Below we have listed the number of parameters in the different models (K, this includes the branch lengths) along with the maximal log likelihood (lnL) of the fitted models. Use likelihood ratio testing to decide if M3 is significantly better than M0 (use a significance level of 0.01):

<table>
<thead>
<tr>
<th>Model</th>
<th>K</th>
<th>lnL</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0</td>
<td>68</td>
<td>-1245.98</td>
</tr>
<tr>
<td>M3</td>
<td>72</td>
<td>-1237.24</td>
</tr>
</tbody>
</table>

12: (15 points)

The following is a short piece of protein-encoding DNA. Codon-codon boundaries are indicated by dots. What is the probability that a single point mutation in this sequence will result in an amino-acid change? (Use the genetic code table included on page 2).

5’-GTC.GGA.CCC.CGA-3’
13: (5 points)

The two trees shown below represent two phylogenies estimated independently by two research groups. Both groups used maximum likelihood to estimate the trees. Tree 1 was built using a 2000 nucleotide segment of the sloD gene, while tree 2 was built using a 150 nucleotide segment from the same gene. As you can see, the groups obtained different topologies. How do you assess the trustworthiness of the two trees (select one):

1) Tree 1 is more trustworthy
2) Tree 2 is more trustworthy
3) I trust both trees equally

14: (5 points)

You are performing an experiment on a population of haploid, asexual organisms. Assume that the organisms exhibit exponential growth with discrete, non-overlapping generations, and that genetic drift can be ignored. At a certain locus you observe two alleles: “Hunter” and “Gonzo”. The absolute fitness of both alleles is R = 2.75 offspring/generation. At the start of the experiment the Gonzo allele is present in 7% of the population. What is the frequency of the Gonzo allele after 17 generations?