Stat 547 — Assignment 3

Release Date:	Saturday April 16, 2011
Due Date:	Wednesday, April 27, 2011 at 4:30 PST

Note that the deadline for this assignment is one day before the final project deadline, and both are hard deadline; I will not accept assignments past Wednesday, April 27, 2011 at 4:30 PST. You should submit a written report under my office door (LSK 330) as well as a zipped file by email containing the "answer" (in which you should copy one exec folder for each experimental question) and "src" folders (do *not* include the other directory to avoid email quotas problems). The report should contain your work for the written questions as well as a summary of what worked/did not work in your experiments.

1 Getting the code and data

First, make sure as early as possible that you can access the course materials.

`http://www.stat.ubc.ca/~bouchard/pri/stat547-assignment3.zip`

The authentication restrictions are due to licensing terms. The username and password have been announced in class (same as assignment 1 and 2), but if for any reason you did not get it, please let me know by email.
Unzip the downloaded file to your local working directory. It contains both the data that you will need, some evaluation code, and some harness code that will help you do the assignment.

2 Technical stuff

Use the same procedure as assignment 1 and 2 to get the code harness setup. Create a new, fresh project and close the previous projects (right-click in eclipse and pick 'close project'). If you need code from the previous assignments or solutions in this assignment, copy contents from the previous assignments manually later on (if you copy files into a project, you might need to refresh it by right-clicking and selecting refresh on the project).
Important: Make sure the libraries were imported properly. This should be automatic, but double-check by right clicking on the project (left hand side part of the window), select project properties, then java build path, then libraries,

then click on add external jars. Use this button as many times as needed to get all the jar files in the lib folder of the project directory imported.

3 Phylogenetic inference

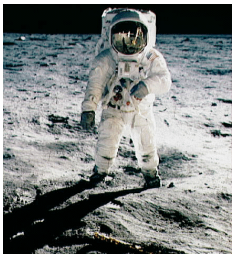
In this question, we are going to use molecular data from twelve modern primate species to infer the phylogenetic relationships between them. Open the file 'data/primates.msf'. It contains mitochondrial DNA that has been aligned for you. For background information, refer to:

`http://en.wikipedia.org/wiki/Mitochondrial_DNA`
`http://en.wikipedia.org/wiki/Multiple_sequence_alignment`

The names used in the first column correspond to abbreviations of the scientific names of the species under consideration. We show some examples here, for information on the other species, refer to

`http://tolweb.org/tree/`

Pan troglodytes
Chimpanzee



Homo Sapiens
Human

Pongo borneo
Orangutan

3.1 Phylogenetic inference: theory

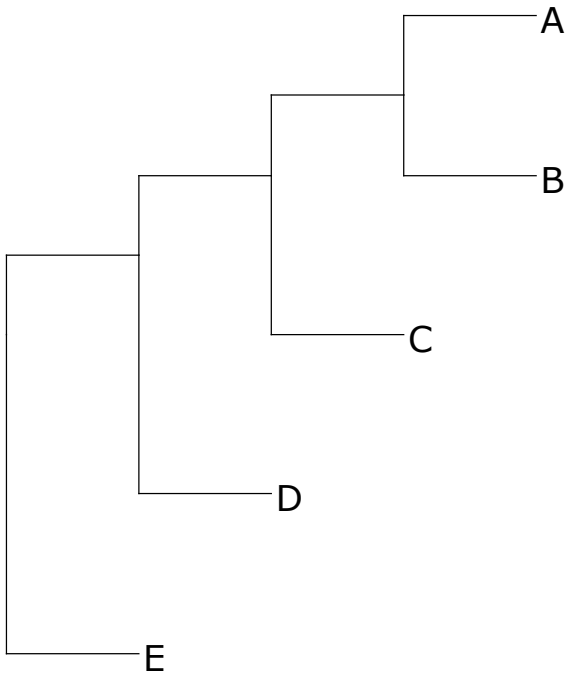
In the next few questions, we will develop the theory behind the standard model in Bayesian phylogenetic: non-clock unrooted phylogenetic trees.
A (non-clock, unrooted) phylogenetic tree t is specified by a connected acyclic graph (V, E) (an undirected tree), called the *topology*, and a positive number associated with each edge, called a *branch length*: $b : E \to (0, \infty)$. Nodes at the periphery of the graph (leaves) represent modern species, points inside the tree represent ancestral species, bifurcation represent speciation events, and branch lengths specify the amount of evolution between species (as measured by the amount of molecular change on the sequences). We also assume that we have a bifurcating tree, in which all nodes in V have either one or three neighbors (corresponding to the case of leaves and internal nodes respectively).
We will consider a simple prior over such trees: a uniform distribution over topologies, and an independent exponential distribution for each branch length (note that there are always the same number of internal branches for a fixed number of leaves in a bifurcating tree).

Part A

Given such a tree, we then define a process that generate a DNA sequence $s \in \Sigma^* = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^*$ for each species (both internal and at the periphery). The process is parameterized by a rate matrix Q, which we assume is symmetric in this problem set. We will denote the marginal transitions of Q by P_t.
The process proceeds as follows: first, a sequence length is picked according to a geometric distribution. Next, the first symbol of each sequence will be generated, then the second symbol of each sequence, etc. The n-th symbol of each sequence are called the n-th site. Our process generates each site independently.
The symbols in one site are generated as follows. First, pick a node arbitrarily in V (this is an instance of what is called rooting a tree). Second, generated a character from Σ for that node using the stationary distribution of P_1. Third, a CTMC is used to generate in preorder all the other symbols for this site.
In this question, you have to show that this process is well-defined: i.e. given a tree, the root placement does not change the distribution over the random sequences generated by this process.

Part B

We will denote the internal nodes by z, the nodes at the leaves by y, and the process described in the previous question by $\mathbb{P}(\,\mathrm{d}y,\,\mathrm{d}z|t)$. Show how your result in question 3.1 of the first assignment can be used to compute $\mathbb{P}(\,\mathrm{d}y|t)$ when the tree has the following imbalanced form:

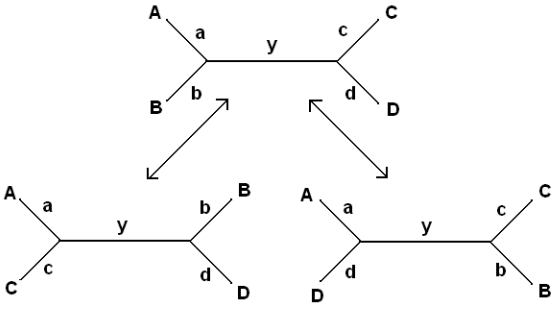


Part C (Optional)

Generalize the equations of question 3.1.B to compute efficiently $\mathbb{P}(\,\mathrm{d}y|t)$ for any phylogenetic tree t.

Part D

While $\mathbb{P}(\,\mathrm{d}y|t)$ is tractable, $\mathbb{P}(\,\mathrm{d}y)$ is not, so we will resort to MCMC sampling to compute the posterior over trees. The simplest irreducible chain has two moves: one move that proposes a local change in the topology, called a nearest neighbor interchange (NNI), and one that proposes a change to a single branch length. NNI works as follows:



where the start state has the topology on the top, and one of the three topologies shown in the figure is proposed uniformly at random.
For the branch lengths, one of the edges is picked at random, and its current value is multiplied by a random number distributed uniformly in $[\frac{1}{a}, a]$ for some fixed parameter $a > 1$ (controlling how bold the move is).
Derive the MCMC acceptance ratio for such a proposal.

3.2 Phylogenetic inference: implementation

Launch 'PhyloMain.java'. This simulates a Markov chain over phylogenetic trees. You will see a summary of the current loglikelihood ($\mathbb{P}(\,\mathrm{d}y|t)$), and of the acceptance ratio of each move (sNNI stands for stochastic NNI, and MB, for multiplicative branch length proposals). You can ignore the lines that start by 'Estimate of posterior' for now, we will come back to it in Part C. Currently, the chain is not ergodic and gets stuck at the same state. We will make the chain stationary in this question.
You can still look in 'state/execs', where the output of each execution is recorded in numbered folders. The MAP (maximum a posteriori) tree is periodically recorded. Use the following tool, archaeopteryx, to visualize them:

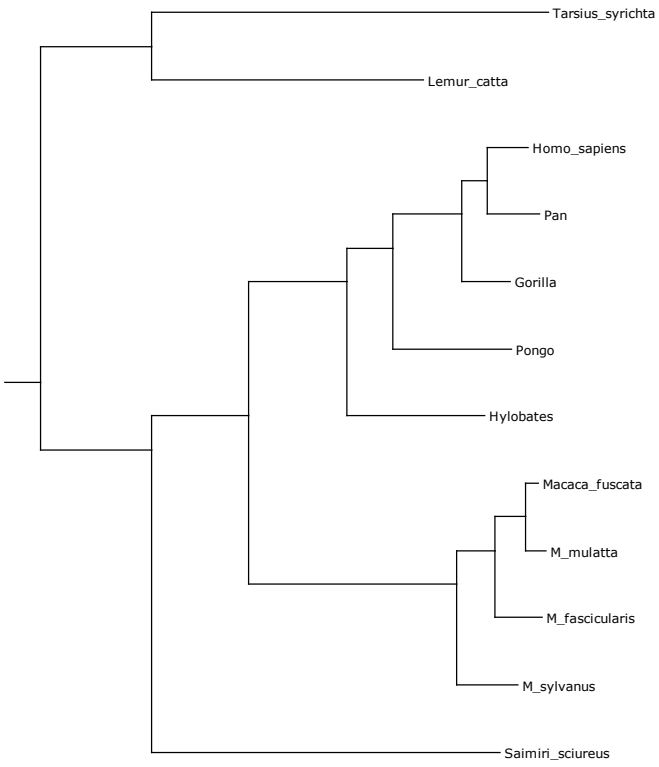
`http://www.phylosoft.org/archaeopteryx/`

Part A

The first step is to implement the Metropolis-Hastings correction. Open the file 'PhyloMCMC.java'. Look at the hints in the function 'sample()', and implement it.
Once this is done, use archaeopteryx to create a pdf of the estimated tree and include it in your report after hiding the internal nodes and rerooting it such that it is more or less balanced.
The tree should be more reasonable now, for example Chimpanzees and Humans should form a clade (a clade is a subset of the leaves of a tree that can be obtained as follows: note that removing a single edge in a tree creates a bipartition of the leaves, the set of all the blocks of these bipartition, ranging over all edges is the set of all clades induced by a tree).

Part B

The tree is still imperfect though, because we are not yet resampling branch lengths. Open the file 'MultiplicativeBranchProposal.java'. The function 'propose()' should do two things: sample a proposed new value for the random edge, and compute the log of the proposal ratio. Use your result of 3.1.D to complete this function.
After filling in this gap, you should get a tree similar to the following tree (up to rerooting):



Part C (Optional)

Suppose now that we only care about the question of whether Chimpanzees and Humans form a clade. What is the Bayes estimator for this decision assuming zero-one loss? Implement in 'CladePosterior.posteriorPanHSapiensClosestCousins()' an estimator for the posterior probability of this event.

4 Advanced topics in non-parametric Bayesian statistics

In this question, we explore some of the more advanced topics covered in the non-parametric Bayesian section of this course.

Part A

Show, using Campbell's theorem, that the Lévy process Γ_t with compensator

$$\mathbb{P}_0(A \times B) = G_0(B) \int_A \frac{1}{z} e^{-z} \, \mathrm{d}z$$

has Gamma finite-dimensional marginals.

Part B

Prove or disprove; or test empirically, the following claims:

1. A random transition matrix (equivalently, Markov chain model) sampled from the Infinite HMM is positive recurrent.

2. A random transition matrix (equivalently, Markov chain model) sampled from the Dependent Dirichlet Process is positive recurrent.

Part C (Optional)

Open-type question: construct a non-trivial distribution over (countably) infinite transition matrices such that some non-trivial expectations with respect to the random stationary distribution can be computed analytically.

Part D (Optional)

Open-type question: revisit question 3.2 of the first assignment using the Infinite HMM or a state-split Infinite HMM.