Exam I
Name: _____

1. a) Define the different schools in phylogenetic systematics and order them based on how they progressed/replaced each other. Use the terms ***evolutionary taxonomy***, ***phenetics, cladistics*****,** and ***statistical phylogenetics*****.** (1 paragraph or list, 10 pts).

   b) Parsimony can be inconsistent. Explain what this means, when it occurs, why it occurs, and why it is used as a justification for statistical phylogenetics (1 paragraph, 10 pts)

2. What is a Likelihood and how do you find a Maximum Likelihood? Use a linear regression as an example. You should define the parameters of the model and **verbally** describe the steps you would use to calculate and identify the Maximum Likelihood estimates in software such as R (be sure to specify how you calculate for more than 1 independent data point). (10 pts)

3. List the parameters the Jukes-Cantor model vs. the Generalized Time-Reversible Model. (10 pts)

4. A 2014 paper by Alexander Pyron and Frank Burbrink entitled *Early origin of viviparity and multiple reversions to oviparity in squamate reptiles* performed ancestral state reconstruction for a single trait, Oviparity (egg-laying) vs. viviparity (life-birth) using a Continuous Time Markov Chain model. They came to the surprising conclusion that viviparity evolved early in the history of squamate evolution, and that there have been a high number of reversals to oviparity. This was surprising to many scientists who assumed that viviparity would exhibit Dollo's law (i.e. once egg laying was lost in a lineage, it would be unlikely to re-evolve). Pyron and Burbrink's paper received spirited criticism and many, many response papers. How appropriate do you feel a CTMC model is for this application? If you feel it may be inappropriate, discuss any assumptions you feel are violated in this case. (1 paragraph, 10 pts)

5. We will be discussing the application of CTMC models to the study of biogeography. Here is the Q-matrix for one of these models:

$$Q = \left[\begin{array}{c|cccccccc} & \emptyset & 1 & 2 & 3 & 12 & 13 & 23 & 123 \\ \hline \emptyset & — & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & E_1 & — & 0 & 0 & D_{12} & D_{13} & 0 & 0 \\ 2 & E_2 & 0 & — & 0 & D_{21} & 0 & D_{23} & 0 \\ 3 & E_3 & 0 & 0 & — & 0 & D_{31} & D_{32} & 0 \\ 12 & 0 & E_2 & E_1 & 0 & — & 0 & 0 & D_{13}+D_{23} \\ 13 & 0 & E_3 & 0 & E_1 & 0 & — & 0 & D_{12}+D_{32} \\ 23 & 0 & 0 & E_3 & E_2 & 0 & 0 & — & D_{21}+D_{31} \\ 123 & 0 & 0 & 0 & 0 & E_3 & E_2 & E_1 & — \end{array}\right].$$

The basic model considers 3 biogeographic areas (e.g. think Asia, Europe and Africa). Species can occupy 0 of these areas (i.e. complete extinction), 1 of these areas (1, 2 and 3), 2 of these areas (12, 13, 23) or all 3 simultaneously (123).

a) How do you interpret biologically the first row of this Q-matrix? (5 pts)

b) How many free parameters does this Q-matrix have? (3 pts)

c) Suppose you are applying this model to frogs, and you consider it impossible for such frogs (which can't be exposed to salt water) to cross directly over an ocean that has always separated areas 2 and 3. How would you modify the model to reflect this hypothesis? (4 pts)

d) What does this Q-matrix assume about the possibility of simultaneous colonization of two areas, and how can you tell? (3 pts)

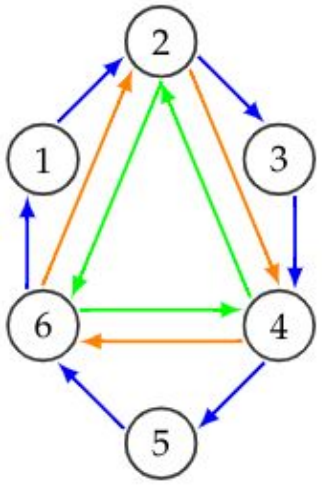
6.

a. Write down Bayes Rule. Label the two components of the numerator and the one component of the denominator. (5 pts)

b. How does the question asked by a Bayesian analysis differ from the answer in a Maximum Likelihood analysis? (1 short paragraph, 5 pts)

7.

a. Fill out the two missing squares. In Model 1, translate the Q-matrix depicted into a graphical cartoon as shown for Model 2. Leave out all transitions that are absent (=0) and label or indicate which transitions are what parameter. For model 2, do the reverse, and conver the depicted CTMC as a Q-matrix. Different colors indicate different transition rates, with transitions of the same color occurring at the same rate. (10 pts)

	Q-matrix	Graphical model
Model 1	$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} - & G & 0 & 0 & 0 & 0 \\ L & - & G & D & 0 & 0 \\ 0 & L & - & G & 0 & D \\ 0 & 0 & L & - & G & 0 \\ 0 & 0 & 0 & L & - & G \\ 0 & 0 & 0 & 0 & L & - \end{bmatrix} \end{array}$	
Model 2		

b. Which of the two models would be a good model for chromosome evolution in plants? Here, the trait would be the haploid number of chromosomes. The processes affecting chromosome evolution would be polyploidization (chromosome doubling) and gain/loss of individual chromosomes. (2 pts)

8. The following methods section is taken from Zhang et al. 2018, “Phylogeny, evolution, and mitochondrial gene order rearrangement in scale worms (Aphroditiformia, Annelida)”. Molecular Phylogenetics & Evolution:

The model of nucleotide evolution for each gene was selected based on Akaike information criterion (AIC) in jModelTest (Darriba et al., 2012). The concatenated matrix was partitioned by genes and the GTR + I + G substitution model was applied for all partitions of the four-gene dataset. The data from different genes were concatenated using SequenceMatrix (Vaidya et al., 2011). Phylogenetic analyses were performed using Maximum Likelihood (ML), Bayesian Inference (BI) and Parsimony analysis (PA). The ML analysis was conducted using RAxML GUI1.3 (Silvestro & Michalak, 2012) with 1000 bootstrap replicates. The Bayesian analysis was performed using MrBayes v3.2.6 (Ronquist & Huelsenbeck, 2003) using parameters identical to those in Norlinder et al. (2012). The PA analysis was conducted using the heuristic search algorithm based on 1000 replicates and 100 random addition sequences in PAUP* V4.0 (Swofford, 2003), with Tree Bisection and Reconnection (TBR) applied to alter trees and evaluate different tree topologies. Gaps were treated as missing data. Phylogenetic analyses were also conducted using few taxa (16 newly sequenced species) but more genes for each species: two nuclear genes (18S and 28S rRNA), as 13 mitochondrial PCGs [Protein Coding Genes] and two mitochondrial rRNA genes (12S and 16S rRNA). The nucleotide sequences of 11 out of 13 mitochondrial PCGs were detected to show signs of saturation using DAMBE, but they were still included in phylogenetic analyses because they contained the majority of phylogenetic information. Each mitochondrial PCG was partitioned based on codon position. The best model for each partition is listed in Supplementary Table S3. GTR+I+G substitution model was applied for each gene in the ML analysis. The concatenated dataset including DNA sequences of four rRNA and protein sequences of 13 mitochondrial PCGs was also used for phylogeny. The best model of amino acid evolution of each gene was determined using prottest3.4 (Darriba et al., 2011). Methods for sequence alignment, data concatenation, and phylogenetic analyses were identical to those described above.

   a. How did the authors account for heterogeneity in substitution rates across sites in their Likelihood analysis? List all the ways (Bonus point if you can point out another common way they didn’t address). (3 pts)

   b. In their parsimony analysis, they conduct an heuristic search based on 1000 replicates. Why do they do it so many times? (3 pts)

   c. They use a program called DAMBE and state that their protein-coding genes “were detected to show signs of saturation”, but that “they were still included in the phylogenetic analyses because they contained the majority of phylogenetic information”. Why are they concerned about saturation in the first place, and how might their decision impact their results? (4 pts)

9. Why was Felsenstein's Pruning Algorithm a big deal? (1 sentence, 3 pts)