Technical University of Denmark

Written examination, date June 2, 2015 Page 1 of 7 pages

Course name: Molecular Evolution

Course number: 27615

Aids allowed: All aids allowed

Exam duration: 4 hours

Weighting: Indicated for each problem on following pages

Molecular Evolution, course # 27615

Final exam, 2015

This exam includes a total of 5 problems. The maximum number of points obtainable is 100, and at least 50 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. Prepare a file (word, pdf, …) with your responses and upload on CampusNet when finished. MAKE SURE TO SAVE BEFORE UPLOADING FINAL ANSWER.

You should do all work in the virtual machine you used for the course.

If a problem requires the answer from a previous problem, and you for some reason have not managed to finish this, then feel free to come up with a guess at what the requirement would look like, indicate that you are using this, and continue. All following questions will then be graded as if you had the correct input.

The duration of the exam is 4 hours. All non-living resources may be used (books, internet including exercise web-pages, notes, etc., but no email, skype, etc).

Have fun (!),
Anders

1: Preparation of data (5 points)

1.1: Download the 4 data files to a suitable directory on your virtual machine using the following commands:

wget http://wiki.bio.dtu.dk/~agpe/course_material/27615/all.fasta
wget http://wiki.bio.dtu.dk/~agpe/course_material/27615/all_subseq500.fasta
wget http://wiki.bio.dtu.dk/~agpe/course_material/27615/elephants.fasta
wget http://wiki.bio.dtu.dk/~agpe/course_material/27615/nd5_all.fasta

1.2: Create an alignment from each of the 4 files. Describe what software you used and why you picked it.

Note: the file "all.fasta" contains full mitochondrial genomes, and will take a while to align.

1.3: Convert alignments for the first 3 files to NEXUS format. Describe the software used.

1.4: Convert the alignment for the 4th file (nd5_all) to interleaved Phylip format (Phylip4).

NOTE 1: most conversion software cuts off names after 10 characters for Phylip format - make sure to open the resulting file and edit names so they are returned to being identical to the names in the original file.

NOTE 2: We will be using this file for PAML package software, which expects interleaved Phylip files to have an uppercase "i" ("I") at the end of the header line (separated by a space from the preceding information). Make sure the file also conforms to this format.

2: Characterization of full data set (20 points)

The file "all.fasta" (and the alignment you constructed from it) contains full mitochondrial genomes from a set of 11 mammals and a single bird. (The bird should be used as outgroup in phylogenetic analyses below). The data set includes sequences for three extinct species (Woolly Mammoth, North American Mastodon, and Neanderthal).

2.1: How many possible unrooted trees are there for this data set? How many rooted?

2.2: Load the alignment in PAUP, and find the uncorrected distance ("p-distance") between "Chicken" and "Woolly_Mammoth".

This observed distance ignores the possibility of multiple substitutions: Perform a manual calculation of what the estimated real distance is, assuming the sequences evolve according to the Jukes and Cantor model ("JC").

2.3: Use PAUP to also find the following distances between "Chicken" and "Woolly_Mammoth":

(a) Uncorrected distance, counting only transitions
(b) Uncorrected distance, counting only transversions

- List and explain the commands used to find these numbers.
- What is the observed ratio between transitions and transversions?
- What transition/transversion ratio would be expected if all possible nucleotide substitutions occurred at exactly the same rate? Why?
- Provide at least one explanation for why the observed ratio may differ from the expected for this data set.

2.4: Use model selection to determine the best substitution model for this data set.

NOTE: the full data set is too large for a thorough analysis so instead of analysing all data, use the data from the file "all_subseq500", which contains the first 500 nucleotides from all sequences. (We will assume that this is a representative sample as far as substitutions are concerned).

- List and explain the commands you used
- Indicate which model was selected, and what parameters are included in this model.

3: Phylogenetic reconstruction based on of full data set: Comparison of methods (20 points)

For all subproblems, please use the Chicken as outgroup, and include clearly labeled (or named) images of the tree plots with your response.

3.1: Reconstruct phylogeny using parsimony. Save tree in file "all_pars_tree.ph"

- List and explain the commands used
- What is the score for the best parsimony tree?
- What does this score quantify (i.e., what is the criterion used to choose the best tree in parsimony)?

3.2: Reconstruct phylogeny by least squares distance method using the substitution model found above to correct for multiple substitutions. Save tree in file "all_dist_tree.ph"

- List and explain the commands used
- What is the score for the best distance based tree?
- What does this score quantify (i.e., what is the criterion used to select the best distance tree?)

3.3: Reconstruct tree by maximum likelihood, using the substitution model found above. Save tree in file "all_lik_tree.ph"

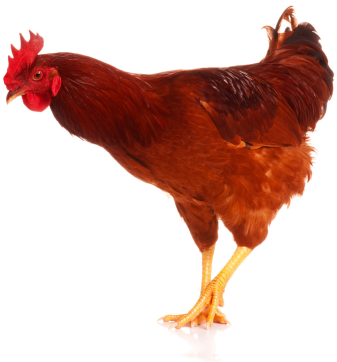
- List and explain the commands used
- What is the likelihood for the best tree?
- What does this score quantify (i.e., what is "likelihood"?)

3.4: Quantify the differences between the three trees using symmetric tree distance in PAUP

- Report all pairwise distances between the three trees
- Based on distances and tree plots: Do the trees agree? If not - where to they disagree

Aardvark

Rock Hyrax



Chicken

Golden mole

European mole

Dugong

4: Phylogenetic reconstruction on full data set: Testing golden mole relationships (20 points)

In the data set used here, all animals except Chicken and Mole are of African origin. In recent years, molecular evidence has indicated that many African mammals are more closely related to each other, than they are to superficially more similar species, and that they form a monophyletic group now named Afrotheria. This evolutionary episode, where a single African ancestor gave rise to species filling many different ecological niches, appears to have happened while Africa was drifting away from South America and had not yet reached Europe and Asia.

4.1: Based on your maximum likelihood tree from 3.3: Is the golden mole more closely related to the mole or to the dugong?

Explain how you assess relatedness from the tree

4.2: Rerun the maximum likelihood analysis, but with a constraint forcing the Golden mole and the Mole to form a monophyletic group

- List the commands used
- Report the likelihood

4.3: Compare hypotheses using AIC and Akaike weights

We now want to formally compare two competing hypotheses about the phylogenetic placement of the Golden mole: Hypothesis 1: The golden mole is placed as a sister group to the Mole. Hypothesis 2: The golden mole is not sister to the mole (but is placed as in your max likelihood tree).

- Compute AIC and Akaike weights for the two competing hypotheses based on the two likelihoods. You can arbitrarily set K=0 parameters for both hypotheses (the two hypotheses use the same number of parameters, so K will not matter). Show how you computed the values.
- Based on the Akaike weights: Which hypothesis has most support from the data? What is the ratio between the Akaike weights.

5: How are Asiatic and African elephants related to Woolly mammoths? (35 points)

Woolly mammoths (Mammuthus primigenius) were a very successful species that are thought to have existed in huge numbers. They ranged from Spain to North America. The oldest fossils of woolly mammoths are 150,000 years old. Most woolly mammoths died out at the end of the Pleistocene (10–12,000 years ago), while the most recent remains date from just 3,700 years ago.

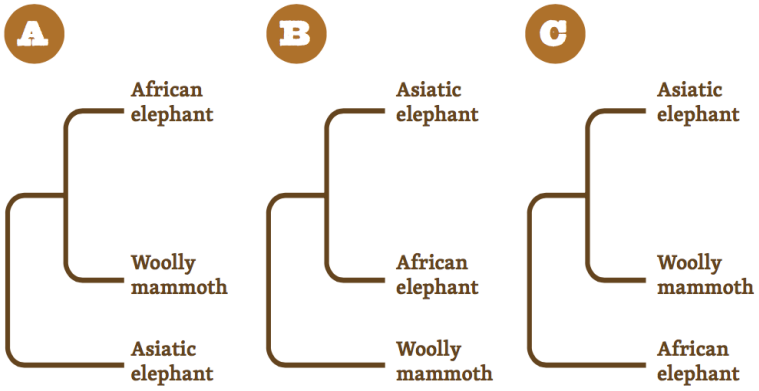
Woolly mammoth depicted in Rouffignac cave

Woolly mammoth (left) compared to North American Mastodon (right)

African elephant

Asian elephant

Mammoths are closely related to present-day elephants, but until very recently the exact relationship between these species was unknown. Did mammoths share a common ancestor of today's elephants (B)? Or were they more closely related to one of the modern elephant species (A or C)?



5.1: Based on previously constructed trees: which hypothesis is correct?

- Inspect the parsimony, distance, and likelihood trees, and comment on which of the three hypotheses they each support
- Do the trees agree? If not - which tree do you believe the most and why?

5.2: Compare hypotheses using likelihood and constraints

- Rerun the likelihood analysis but with constraints corresponding to the other two hypotheses about mammoth placement
- Report likelihoods
- Compute AIC and Akaike weights for the three hypotheses - which has more support?

5.3: Compare hypotheses using Bayesian phylogeny

Perform a Bayesian phylogenetic reconstruction based on the alignment made from the file "elephants". This contains the two present-day elephants, the two extinct elephantids, and the Chicken (outgroup). Use the model determined using model selection above. Preferably continue the run until convergence (standard deviation of split frequencies < 0.01). If this takes too long: Stop earlier, but indicate the number of generations and the standard deviation of split frequencies.

Tip: Run-time is directly proportional to the number of chains used. You need two independent runs to assess convergence, but you can use as little as one chain per run.

- Briefly explain the interpretation and use of probability in Bayesian phylogeny
- List the commands used
- Which of the three hypotheses has the most support according to this method? (Report the clade support)