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:

Simply use the provided wget commands

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

Many options here - I would simply use MAFFT for all alignments, since it has good performance and an easy to use interface at EBI

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

The EBI ReadSeq server with output option set to PAUP/NEXUS

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

Again, EBI ReadSeq server, with output option Phylip4. (NOTE: This file was never used in the exam - this was a mistake - the file was left from an earlier version of the exam I had prepared).

2: Characterization of full data set (20 points)

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

The file contains 12 species. According to the table (and formula) from week 2 this corresponds to 654,729,075 possible unrooted trees.

Each of these trees have 21 branches (2*N-3). Since the root can be placed on each branch on each tree, the number of possible rooted trees is: 654,729,075 * 21 = 13,749,310,575

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

```
paup> exe all_align.nexus

paup> dset dist=p

paup> showdist
```

Chicken to Woolly Mammoth: 0.35185

Using the formula from the quizzes in week 6:

D_JC = -0.75 * ln(1 - 1.33 * D_OBS) = -0.75 * ln (1 - 1.33*0.35185) = 0.473

NOTE: The slides in the lecture on this have a typo in the same formula: The minus sign is missing!

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.

```
paup> dset dist=p subst=ti
```

```
paup> showdist
```

Transition distance = 0.16241

```
paup> dset dist=p subst=tv
```

```
paup> showdist
```

Transversion distance = 0.18944

What is the observed ratio between transitions and transversions?

0.16241 / 0.18944 = 0.86

What transition/transversion ratio would be expected if all possible nucleotide substitutions occurred at exactly the same rate? Why?

Exepected transition/transversion ratio = 0.5

Because there are twice as many possible transversions than transitions

Provide at least one explanation for why the observed ratio may differ from the expected for this data set.

It’s a mitochondrial data set. These are known for higher transition/transversion ratios

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

```
paup> exe all_subseq500_align.nexus
```

```
paup> exe modelblock3.gorm
```

(NOTE: the modelblock3.gorm file was used in earlier exercises, and can be copied from those directories).

```
paup> quit
```

```
student> modeltest < model.scores > model.results
```

(NOTE: modeltest should already be installed in your virtualbox after the model selection exercise)

Inspection of the model.results file (in nedit) shows that the selected model (under the AIC criterion) is: GTR+G. The full model (with paraneter estimates) can be specified in PAUP using this command:

```
Lset  Base=(0.3629 0.2338 0.1565)  Nst=6  Rmat=(4.5101 11.4371 5.9199 1.1211 25.0844)
Rates=gamma  Shape=0.4663  Pinvar=0;
```

This model has separate parameters for all 6 possible nucleotide substitutions, for the 4 nucleotide frequencies, and for the shape of the gamma distribution

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

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

List and explain the commands used

```
paup> exe all_align.nexus

paup> set crit=parsimony

paup> outgroup Chicken

paup> hsearch

paup> savetrees file=all_pars_tree.ph brlens=yes
```

What is the score for the best parsimony tree?

Score of best tree(s) found = 24074

What does this score quantify (i.e., what is the criterion used to choose the best tree in parsimony)?

This is the minimum number of substitutions required to explain how the aligned sequences evolved from a common ancestor (parsimony criterion = best tree is the one requiring the fewest substitutions)

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

```
paup> set crit=dist

paup> dset dist=GTR rates=gamma shape=0.4663 objective=lsfit

paup> outgroup Chicken

paup> hsearch

paup> savetrees file=all_dist_tree.ph brlens=yes
```

What is the score for the best distance based tree?

```
paup> dscore
```

```
SS          0.02600 (%SD         3.58032,  g%SD        4.28098)
```

What does this score quantify (i.e., what is the criterion used to select the best distance tree?)

Here SS is the score we usually use - sum of squared errors between pairwise distances in alignment, and measured along 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

```
paup> set crit=lik
```

```
Lset  Base=(0.3629 0.2338 0.1565)  Nst=6  Rmat=(4.5101 11.4371 5.9199 1.1211 25.0844)
Rates=gamma  Shape=0.4663  Pinvar=0;
```

```
paup> outgroup Chicken
```

```
paup> hsearch
```

```
paup> savetrees file=all_lik_tree.ph brlens=yes
```

What is the likelihood for the best tree?

```
Score of best tree(s) found = 116148.6
```

Note: this is the negative log likelihood, meaning log likelihood is: -116148.6

What does this score quantify (i.e., what is “likelihood”?)

Likelihood = probability of data (alignment) given model (GTR+G+tree)

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

Report all pairwise distances between the three trees

```
paup> gettrees file=all_pars_tree.ph mode=3
paup> gettrees file=all_dist_tree.ph mode=7

paup> gettrees file=all_lik_tree.ph mode=7

paup> treedist
```

```
Symmetric difference (RF x 2) distances between trees

        1   2   3

1       -

2      18   -

3      18   8   -
```

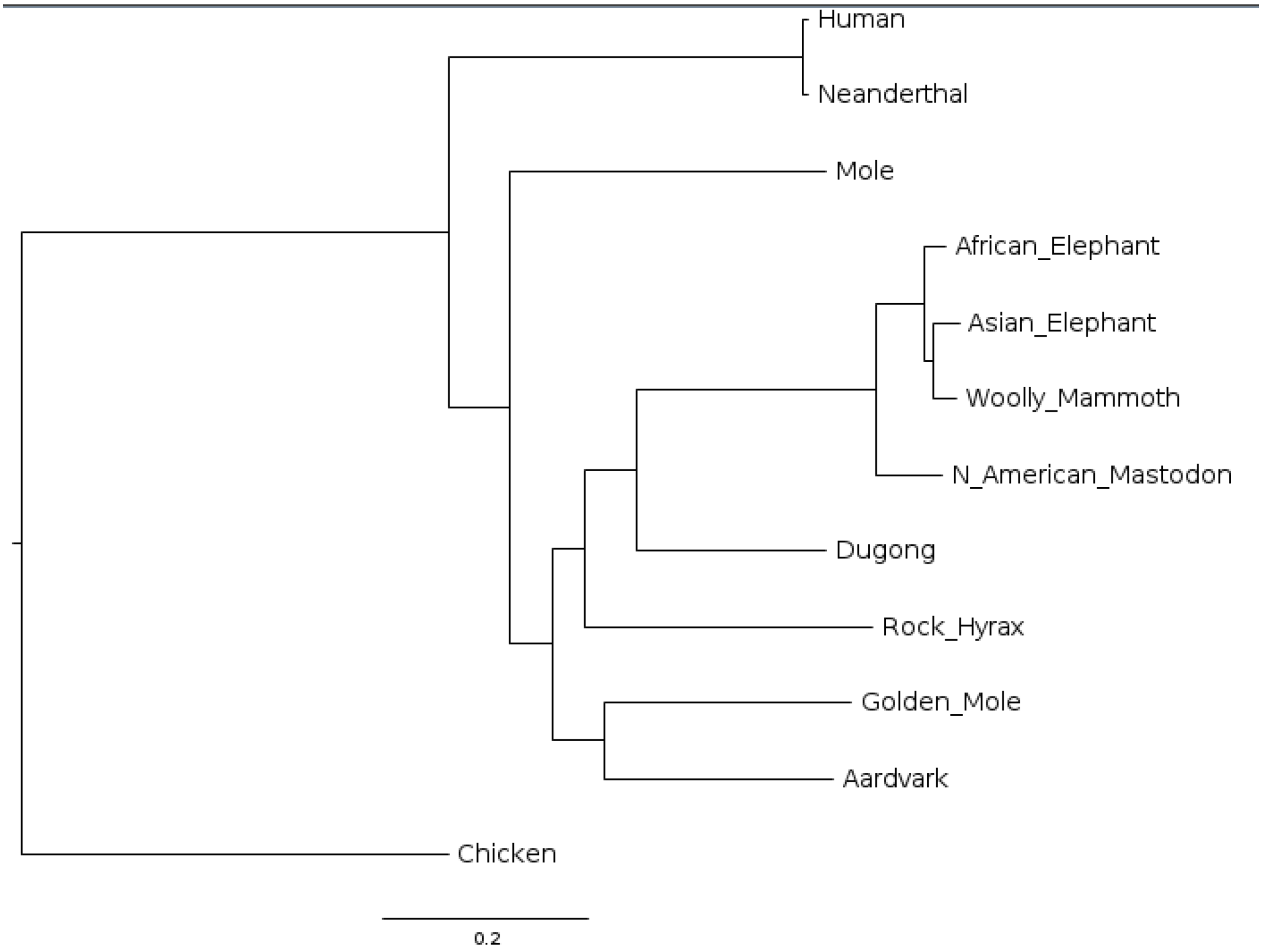
Based on distances and tree plots: Do the trees agree? If not - where to they disagree

Several differences…. Max likelihood tree is correct (if done as above)

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

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?

Based on tree below, golden mole is more closely related to dugong than to mole, because the dugong and mole share a more recent common ancestor: Starting at the golden mole, you have to go back two nodes before you find an ancestor shared with the dugong, but three nodes before you find one shared with the mole



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

```
paup> set crit=like
```

```
paup> Lset  Base=(0.3629 0.2338 0.1565)  Nst=6  Rmat=(4.5101 11.4371 5.9199 1.1211 25.0844)  Rates=gamma  Shape=0.4663  Pinvar=0;
```

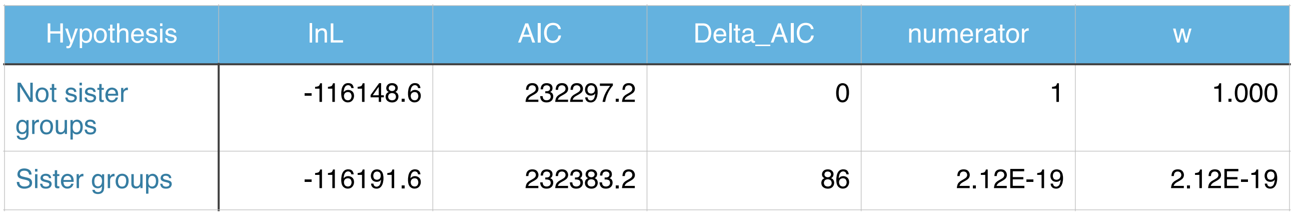
```
paup> constraints mole (monophyly)=((Mole,Golden_Mole));
```

```
paup> hsearch constraints=mole enforce=yes
```

```
Score of best tree(s) found = 116191.6
```

4.3: Compare hypotheses using AIC and Akaike weights

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.



Hypothesis	lnL	AIC	Delta_AIC	numerator	w
Not sister groups	-116148.6	232297.2	0	1	1.000
Sister groups	-116191.6	232383.2	86	2.12E-19	2.12E-19

a: not sister groups

b: sister groups

AIC

a: -2*-116148.6 = 232297.2

b:-2*-116191.6 = 232383.2

deltaAIC

a: 232297.2 -232297.2 = 0

b: 232383.2 - 232297.2 = 86

numerator

a: exp(-0.5*0) = 1

b: exp(-0.5*86)= 2.12E-19

sum = 1 + 2.12E-19 ~ 1

w

a:1/sum ~ 1/1 = 1.00

b:2.12E-19/1 = 2.12E-19

ratio

1/2.12E-19 =4.7E18

Based on the Akaike weights: Which hypothesis has most support from the data? What is the ratio between the Akaike weights.

The hypothesis that they are not sister groups has much more support (in fact within rounding error it has the probability 1)

The ratio is 4.7E18 meaning the support for the non-sister hypothesis is immensely stronger

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

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

According to (my) parsimony tree: B

According to (my) distance tree: A (but they are very close to being equally distant)

According to (my) likelihood tree: C (Asian closer)

Do the trees agree? If not - which tree do you believe the most and why?

No agreement! Likelihood tree is probably most correct (explicit model-based approach that accounts for all aspects of substitution, finds a global substitutions process (same over entire tree) and includes gappy columns in principled manner.)

5.2: Compare hypotheses using likelihood and constraints

Rerun the likelihood analysis but with constraints corresponding to the other two hypotheses about mammoth placement

```
paup> constraints A (monophyly)=((African_Elephant,Woolly_Mammoth));

paup> constraints B (monophyly)=((Asian_Elephant,African_Elephant));

paup> hsearch constraints=A enforce=yes

paup> hsearch constraints=B enforce=yes
```

(Constraint C is the same as the tree found above)

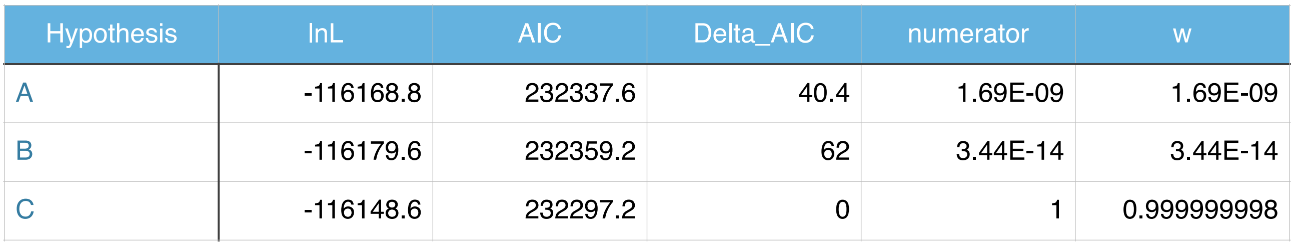
Report likelihoods

A: -116168.8

B: -116179.6

C: -116148.6

Compute AIC and Akaike weights for the three hypotheses - which has more support?



Hypothesis	lnL	AIC	Delta_AIC	numerator	w
A	-116168.8	232337.6	40.4	1.69E-09	1.69E-09
B	-116179.6	232359.2	62	3.44E-14	3.44E-14
C	-116148.6	232297.2	0	1	0.999999998

Option C has by far the most support (being at least 10^9 times more probable than the second best hypothesis)

5.3: Compare hypotheses using Bayesian phylogeny

In Bayesian phylogeny (or Bayesian statistics more generally) probability is used as a way of quantifying degree of belief in some aspect of reality. The result of a Bayesian phylogenetic analysis is a joint probability distribution (the posterior) over all possible values of all parameters in the model. Typically, model parameters include tree topology, branch lengths, substitution rates, nucleotide frequencies, gamma distribution shape parameter.

```
student> mb

MrBayes > exe elephants_align.nexus

MrBayes > lset nst=6 rates=gamma

MrBayes > mcmc nchains=3 ngen=200000 samplefreq=200
```

Hypothesis C (Asian Elephant with Woolly Mammoth) has far the most support (clade credibility = 98%)