Instructions for the BIO 332 final assignment.

Files are available on the student portal. The PC-lab (where we did the practicals) is available for your work. Unfortunately, there is a shortage of extra keys, but you may usually find people who are willing to let you in. One key may be borrowed (overnight?) from former Department of Botany. Contact:
Oddfrid T.K. Førland
Studiekonsulent / Study counsellor
Institutt for biologi / Department of biology
P.O. Box 7800, N-5020 Bergen
Besøksadresse / Visiting address: Allégate 41
Tlf. +47 55 58 22 24. Fax. +47 55 58 96 67

In case of technical problems, contact Arild Breistøl (zoology 1st floor) (tlf. 82233) or E.Willassen (tlf. 82901).

When to submit:
Final deadline on December 1 (2004).

How to submit:
Submit by Email to Oddfrid.Forland@bio.uib.no

She will collect your files in electronic folders marked with your candidate number so that material is anonymous when evaluated and marked.

What to submit:
1) Your name and candidate number
2) PAUP* script files and files produced by the scripts.
3) A report with text, tables, and figures (trees) presenting the methods, results, your interpretations, conclusions, and possible references to literature. The final part of the report must contain a table with the following format:

Supplementary files

File number	contents	File name
1	Paup script	Myfile_1.nex
2	Results model testing	modeltest.txt

Please refer to the file numbers in the text if you want to document details of your work, for instance, ‘ – the GTR+G model was selected based on results (2) from the hierarchical testing with Modeltest – ‘

UoB Nov.16. 2004. E.Willassen

BIO332 Final assignment.
November 2004 (E.Willassen)

A key issue in public interest about evolutionary questions is the relationship between humans and the great apes. (Some background information may be found in the file *human_evol.pdf*, Willassen 2004: lecture notes from BIO210). For this assignment, you will examine some of the evidence for the relationships of the hominids by analyzing two data sets of mitochondrial sequences.
Data set 1 (Hayasaka et al. 1988) includes two mitochondrial protein coding genes and three tRNAs from twelve primates. The range of each gene is defined in the available nexus data file (***Data1.nex***).
Data set 2 (***Data2.nex***) is an alignment of additional mitochondrial tRNAs from some of the same taxa.
Use the text book and additional material with your previous knowledge from the course. Check the PAUP* and MrBayes manuals to find additional information on commands.

Analyze Data set 1

Write a *PAUP script** that does the following when executed:

1) Logs the run
2) computes uncorrected distance (p-distance) for each gene so that you may subsequently compute average p-distance in each gene by importing the results to Excel.
3) computes total pair wise (uncorrected) differences between the sequences and also computes the numbers of transitions and transversions in 1st, 2nd, and 3rd positions of the protein coding genes.
4) computes "empirical" base frequencies for the alignment.
5) computes MP tree(s) and saves the MP tree(s) to a file.
6) reconstructs character state changes on the branch representing the most recent ancestor of Homo_sapiens and Pan. [hint: describetrees /apolist=yes].
7) Computes and saves a strict consensus tree to indicate unresolved nodes.
8) computes bootstrap support and saves the bootstrap consensus tree to a file.

Presentation:
Present the results from 2) and 4) in tables. 3) Plot transitions and transversions (y-axis) in each codon position versus absolute pair wise differences (x-axis). Present the numbers of substitution types (C<->T, etc) in the most recent common ancestor (MRCA) of Homo and Pan in a table. Present your strict consensus tree in your report [hint: export graphic from Treeview or Mesquite?]

Interpretation:
Is higher sequence divergence indicated in particular parts of the alignment?
Is saturation indicated to a larger degree in any of the codon positions, and if so, how would you explain that?
Is there a bias in nucleotide frequencies?
Do the results from 6) indicate equal rates of nucleotide substitution types?
Does this exploration of the data suggest what sort of properties we would require of a model to be used with this data set for phylogeny reconstruction with distance or ML methods?

Prepare, execute, and log scripts for ML analysis of Data set 1

9) use the script *ModelblockPAUPb10.nex* and hierarchical log-likelihood ratio testing with *Modeltest* to find an evolutionary model that describes all data best under the ML criterion.
10) Use the model without the parameter estimates from the model testing [hint: see how your model is phrased in Modelblock], and compute the

likelihood scores for your MP trees. [Hint: by adding the option khtest=normal to your commands, the Kishino-Hasegawa test will tell you whether one tree is significantly better than alternative trees].

11) Compute the maximum likelihood tree with the parameter estimates suggested by model test. Save the tree.

Analyze Data set 1 with MrBayes

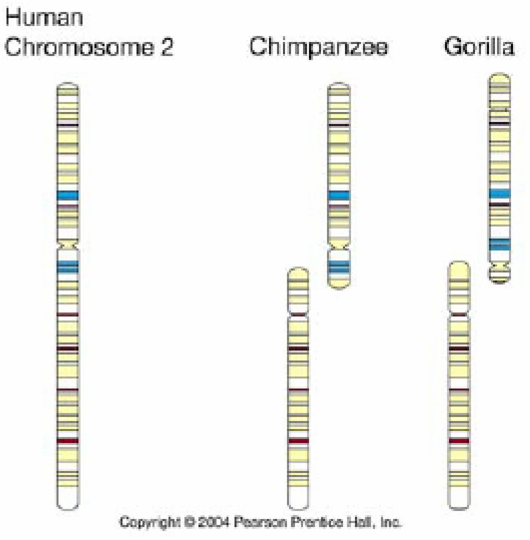
12) Find a suitable model for the data by either adapting the most similar model to the one used with ML above [hint: see file 'ML_models…' by J.Nylander], or by running *MrModeltest* .
13) Apply the model for Bayesian estimation of phylogeny.
14) Use *Tracer* to decide when likelihood estimates (and preferably other parameters) are in equilibrium, and effective sample size (ESS) is sufficient. (Make sure that you have a large number of trees for the computation of posterior probabilities on branches.)
15) Present your tree with branch support in your report.

Analyze Data set 2

16) Use you knowledge and skills to decide whether a phylogeny reconstructed from Data set 2 is congruent with results obtained with Data set 1.

Sum up your analyzes with respect to the question of monophyly human-chimps.

How would you explain this striking similarity: that the homologous sites of the human chromosome 2 are found on two separate chromosomes in chimps and gorillas?
What sort of information would you need in order to decide whether these characteristics are actually in conflict with the hypothesis of humans and chimps as sister groups?



Dating the diversification of hominids

17) Use Dataset 1 and load your previously achieved unrooted ML tree to memory
18) Estimate likelihood score, parameters, and branch lengths for the tree under the chosen model.
19) print the tree with branch lengths to the log. [hint: describetree / `plot=ph brlens=yes].` Do the terminal branches indicate an ultrametric tree? [hint: see textbook]
20) Root the tree with *Tarsius* as outgroup.
21) Estimate likelihood score, parameters, and branch lengths for the tree under a molecular clock constraint [hint: clock=yes].

22) print the tree with branch lengths to the log.
23) Use the log likelihood ratio test to decide whether evolution in the unconstrained ML tree significantly deviates from a molecular clock. **Note: The degrees of freedom for the test are N-2 (Not N-1 !)** (N=#taxa). (The ultrametric tree is the null hypothesis.)

A 'standard molecular clock' for animal mitochondrial DNA is 2% nucleotide divergence per million years, i.e. a substitution rate of 0.01 nucleotides per nucleotide per million years. Why is the divergence rate two times the evolutionary rate?

24) If the ultrametric tree can be used to model the evolution of the primates, use the evolutionary rate and branch lengths to date the nodes in the tree.

In 2002, about 6-7 million year fossil remains of a hominid species called *Sahelanthropus tchadensis* were discovered in Tchad (Nature 418, 145–151). It has been suggested that *S. tchadensis* represents the MRCA of chimps and humans. Other researchers claim that these fossils have characteristics that are more gorilla-like. How does your molecular dating contribute to this discussion?

Reference

Hayasaka, K., T. Gojobori, and S. Horai. 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. Mol. Biol. Evol., 5:626-644)

Extra files

Data1.nex
Data2.nex
ModelblockPAUPb10.nex
(There is a bugged version of Modelblock out there. Make sure to use this file)
Human_evol.pdf
Paup.pdf
Mrbayes3.pdf

E.Willassen

Comments to BIO332 2004 Final Assignment

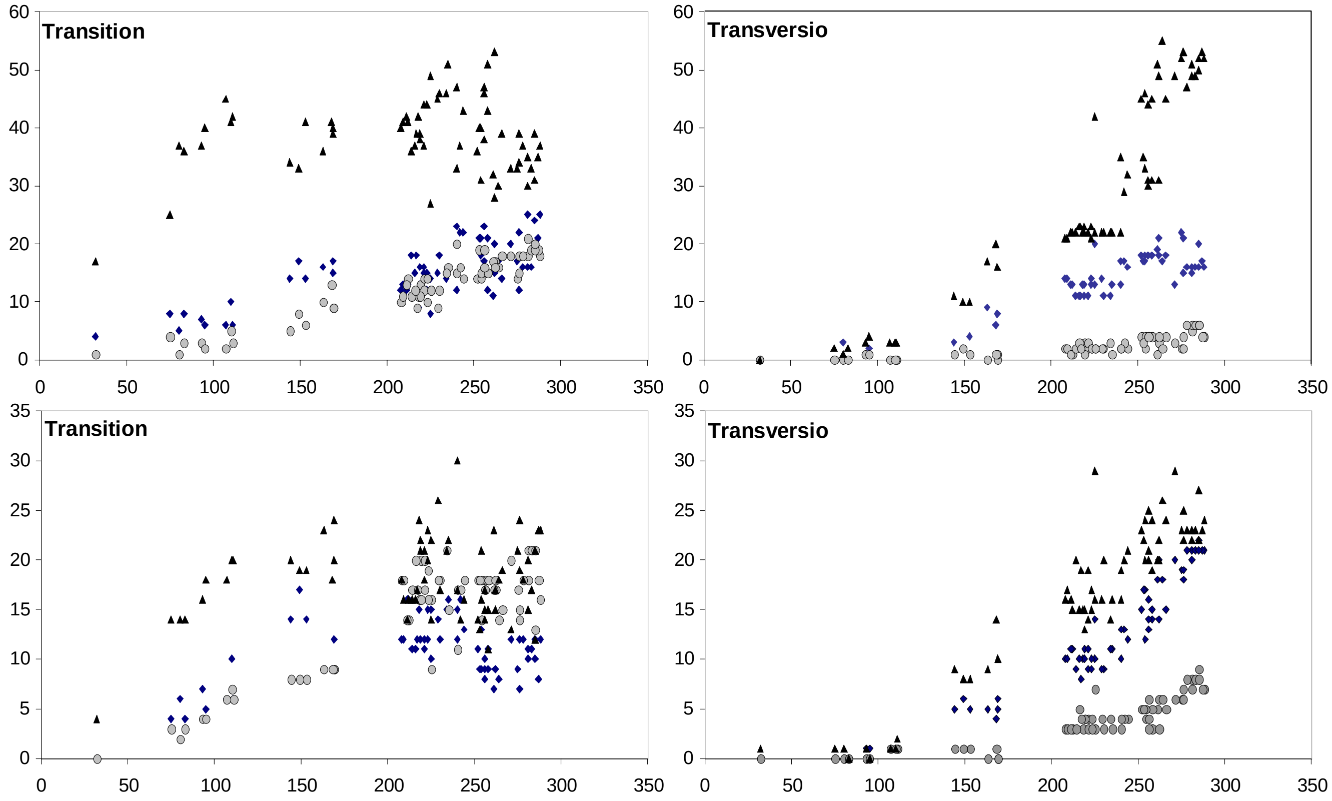
Data set 1

Exploration

Mean p-distance computed for each gene separately indicates slightly higher rates in the protein coding genes, so the tRNAs appear more conserved.

	ND4	tRNA1	tRNA2	tRNA3	ND5
mean p dist	0.0950	0.0400	0.0704	0.0304	0.1326

Pairwise number of substitutions in ND4 (upper) and ND5 (lower) plotted against pairwise total substitutions in the alignment. (♦=1^{st}, ●=2^{nd}, filled Δ=3^{rd} codon position)



Saturation is indicated in 3^{rd} codon transitions of ND4 and in both 1^{st} and 3^{rd} positions of ND5. Codon degeneracy (synonymous codons) explains why constraints on substitution increases in the order 3^{rd}<1^{st}<2^{nd}, so that 3^{rd} codons are more likely to become saturated. Saturated sites may contribute to homoplasy and so raw p-distances between a pair of taxa are likely to become smaller that the distance between the taxa over branches in the tree (see textbook p.180-).

Empirical base frequencies indicate that particularly Gs are less frequent than other bases, but although the nucleotide base content is slightly variable among species, this heterogeneity is not significant.

	A	C	G	T	# sites
Mean	0.32412	0.30402	0.10553	0.26633	895.50

Parsimony

Including all characters, with 367 parsimony informative sites, returns two trees of 1153 steps. The strict consensus tree is unresolved with respect to the relationship of *Homo*, *Pan*, and *Gorilla*. Bootstrapping with 250 replicates may yield about 50% support for either *Homo* and *Pan*, or *Pan* and *Gorilla* as sister groups.

ACCTRAN reconstruction of character change in the MRCA of *Pan* and *Homo* indicates a majority of C<>T transitions, and that all data partitions contribute evidence for a sister group relationship between *Pan* and *Homo*.

TREE1: Apomorphies in MCRA of (Pan,Homo)

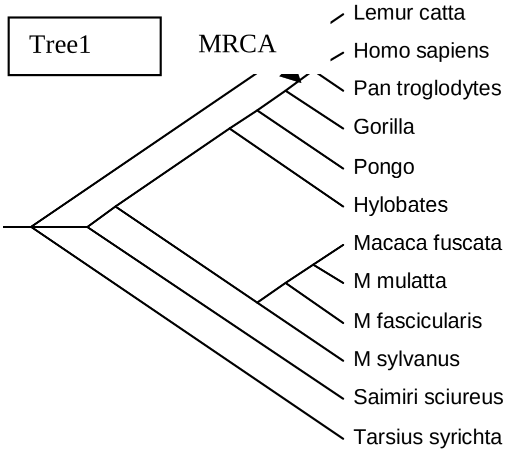
	Pos	Steps	CI	Subst
ND4	31	1	0.200	T --> C
	88	1	0.500	A ==> C
	97	1	0.500	C ==> T
	149	1	0.500	C --> T
	223	1	1.000	A ==> G
	250	1	0.500	C ==> T
	256	1	0.500	C --> T
	307	1	0.333	T --> C
	332	1	0.600	C --> T
	340	1	1.000	A ==> C
	349	1	0.750	C ==> T
tRNA	499	1	0.500	T --> C
	514	1	0.500	A ==> G
	569	1	0.500	T --> C
	625	1	1.000	A ==> C
ND5	776	1	0.500	C ==> T
	821	1	0.333	T ==> C
	880	1	1.000	T ==> C
	884	1	1.000	A ==> G

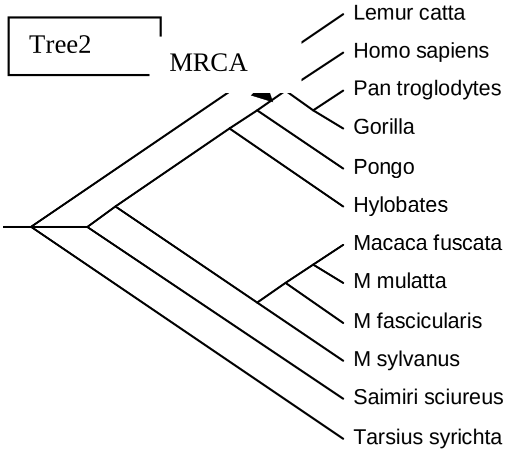
Unambiguous changes (=>) are largely transitions.

However, there are two MP trees. In **Tree2** the relationship is ((Pan,Gorilla),Homo), so the MRCA of *Homo* and *Pan* in this tree is on a deeper node with a lot more substitutions.

TREE1	A	C	G	T	
A	-	3	3	0	A
C	8	-	0	13	C
G	13	1	-	0	G
T	3	24	0	-	T
	A	C	G	T	**TREE2**

In both hypothetical MRCAs, we see that transitions (black) are more abundant than transversions (pink).





It appears that nucleotide frequencies are not equal, and that transitions are more frequent than transversions. Inspection of character changes in two hypothetical MRCAs additionally suggests that CT transitions may be more frequent than AG. However, this may not necessarily apply to all nodes.

It seems fair to conclude from the exploration of the data that a ML approach should account for unequal nucleotide frequencies and unequal substitution rates. The heterogeneous variability over sites may additionally suggest a gamma model. With Modeltest3, we have a tool to examine further details in order to select an appropriate model specification.

ML analysis

The best-fit model selected by hLRT in Modeltest is TVM+G, which is a constrained GTR+G model with five substitution types. As opposed to the tendency indicated by inspection of the apomorphy list above, the results from Modeltest suggest that the two types of transitions (AG and CT) have about equal rates. Accordingly, we may treat transitions as one single substitution class. We can specify the model in PAUP* with parameter estimates based on the NJ-tree initially calculated by PAUP* while running the Modelblock script. However, the parameters may also be re-estimated given one or several other trees computed from the same data. To estimate all the parameters of the TVM+G model we use the following model specification

```
lset nst=6 base=est rmat=est rclass=( a b c d b e)
rate=gamma shape=est pinvar=0
```

The rclass option specifies that A-G and C-T (transitions) have equal rates. It is not a good idea to search for a ML tree while simultaneously estimating all these parameters. However, with the lscore command we may optimise the parameter estimates on previously computed trees. (The same procedure is used in Modelblock). To load the MP trees into memory we use the command

```
gettrees file=my_mp.tre mode=3
```

We then execute the command

```
lscore all/ khtest=normal
```

to obtain parameter estimates for the trees while simultaneously executing the the Kishino-Hasegawa test.

```
Tree                 1           2
------------------------------
-ln L    5709.63215  5716.79123
Base frequencies:
  A        0.358075    0.359484
  C        0.318592    0.316888
  G        0.084591    0.084624
  T        0.238743    0.239004
Rate matrix R:
  AC        3.99887     4.63405
  AG       40.57875    45.30845
  AT        3.41193     3.68819
  CG        2.39085     2.89510
  CT       40.57875    45.30845
  GT        1.00000     1.00000
Shape      0.375152    0.378177
```

The KH test using normal approximation indicates that the tree with *Pan* and *Homo* as sisters is significantly better than the alternative tree in likelihood score.

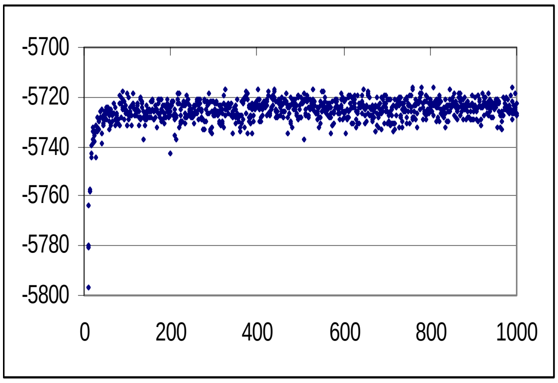
```
                KH-test
Tree     -ln L       Diff -ln L   P
---------------------------------------
-
 1     5709.63215     (best)
 2     5716.79123    7.15909    0.000*
  * P < 0.05
```

The parameter values for Tree1 are very close to estimates obtained when searching for the best model using Modeltest (with the TVM+G model).

To search for the best tree under the ML criterion, we may use the model settings from Modeltest (including parameter values). By doing so, we obtain a ML tree that is congruent with MP Tree1 (above) and has approximately the same likelihood score.

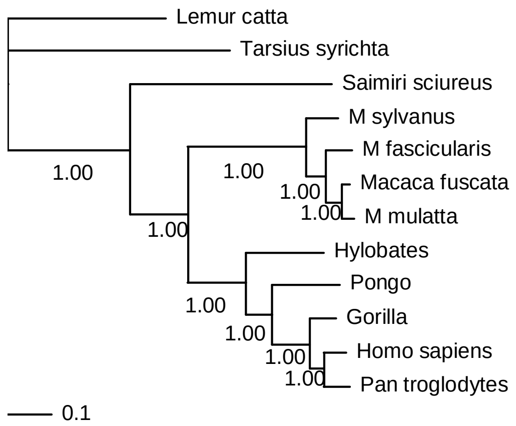
Bayesian inference

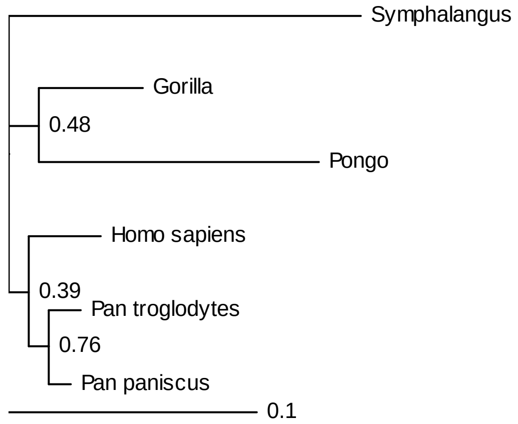
I applied a GTR+G model with four gamma categories for the whole data set and ran 100000 MCMC generations with six chains, sampling every 100 generation.



Log likelihoods started to converge towards equilibrium after about 10000 generations. Hence, 100 trees were excluded in the 'burnin' (i.e. 10000/100).

Consensus of all compatible trees returned posterior probabilities of 1 on all branches so the phylogeny is very well supported with a Bayesian approach. In conclusion, likelihood and Bayesian inference support a Homo-Pan relationship that is unresolved with parsimony.

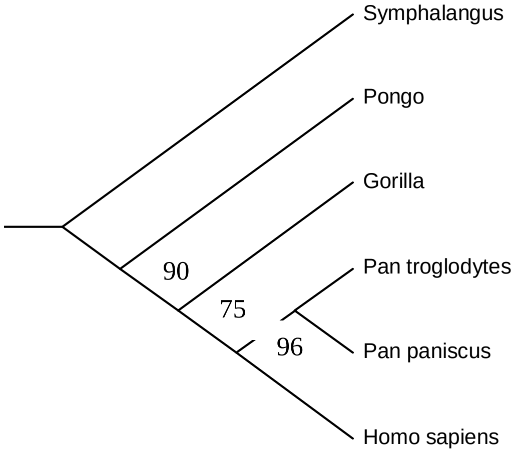




Data set 2

Parsimony

Only 43 of 94 variable characters are parsimony informative. MP analysis gives one tree with 168 steps and CI 0.85. Bootstrap support on branches was obtained with 250 replicates.



To compare the tree with those inferred from data set 1, we root the tree on *Symphalangus*. There is evidence in both data sets for a sister group relationship between *Pan* and *Homo.* The MP tree also is congruent the trees obtained with Data set 1 concerning the branching order of *Pongo* and *Gorilla*.

While a MP approach reconstructs relationships that correspond with the results from Data set 1, ML and Bayesian analyses with the HKY+G model return short internal branches that are generally poorly supported. The high proportion of parsimony uninformative substitutions contributes to long terminal branches in some taxa and the internal branches are short. A complication with tRNA is that the data may be affected by compensating mutations. Modelling evolution of base pairs in the secondary structure may be more appropriate for this type of sequences (but this was not expected for this assignment).

Chromosome characters

The observation that chimps and gorillas have two chromosomes whereas humans have just one homologous chromosome is simply a *similarity* between chimps and gorillas. However, the phylogenetic significance of this observation is ambiguous unless we know the state in other taxa. One chromosome may be an autapomorphy for humans (fusion). If *Pongo* has two chromosomes, this state is more likely plesiomorphic, and so it cannot be taken to indicate a closest relationship between chimp and gorilla. Conversely, if *Pongo* shares the human state (one chomosome), it would conflict with the (Pan,Homo) phylogeny. In other words, we need information about the homologous chromosomes in *Pongo* (and preferably other primates in the data set).

Dating nodes with a molecular clock

The relationships between these primates seem well supported from Data set 1 when we are using a TVM+G model. We see from a phylogram representation of the tree (see for example the Bayesian tree above) that the tips of the terminal branches are not on a straight line. Therefore, the tree is not exactly ultrametric, indicating that evolutionary rates may have been slightly different among clades. (Remember that cladograms may look like ultrametric trees, but branch lengths are irrelevant in cladograms / MP trees.)

Node	Connected to node	Branch length	Distance from root	Distance to tips	Time lenght MYrs
22	(root)		0.0000	0.6503	
Lemur catta (1)	22	0.6503	0.6503	0.0000	65.03
21	22	0.1744	0.1744	0.4759	17.44
20	21	0.1230	0.2974	0.3529	12.30
16	20	0.1480	0.4455	0.2048	14.80
15	16	0.0508	0.4962	0.1541	5.08
14	15	0.0824	0.5786	0.0717	8.24
13	14	0.0191	0.5977	0.0526	1.91
Homo sapiens (2)	13	**0.0526**	0.6503	0.0000	5.26
Pan troglodytes (3)	13	**0.0526**	0.6503	0.0000	5.26
Gorilla (4)	14	0.0717	0.6503	0.0000	7.17
Pongo (5)	15	0.1541	0.6503	0.0000	15.41
Hylobates (6)	16	0.2048	0.6503	0.0000	20.48
19	20	0.2663	0.5637	0.0866	26.63
18	19	0.0326	0.5963	0.0540	3.26
17	18	0.0350	0.6313	0.0190	3.50
Macaca fuscata (7)	17	0.0190	0.6503	0.0000	1.90
M mulatta (8)	17	0.0190	0.6503	0.0000	1.90
M fascicularis (9)	18	0.0540	0.6503	0.0000	5.40
M sylvanus (10)	19	0.0866	0.6503	0.0000	8.66
Saimiri sciureus (11)	21	0.4759	0.6503	0.0000	47.59
Tarsius syrichta (12)*	22	0.6503	0.6503	0.0000	65.03

We use the log-likelihood-ratio test to decide whether these observed deviancies from ultrametricity are significant. If this is not the case, we may tentatively date the nodes in the tree by using an empirical rate to convert branch lengths to time.

The first step in this procedure is to obtain the log-likelihood score for the unconstrained tree.

```
gettrees file=dat1_ml_trees.tre mode=3;
set crit=likelihood;
Lset  clock=no Base=est  Nst=6  Rmat=est rclass=(a b c
d b e)  Rates=gamma  Shape=est  Pinvar=0;
describetree/plot=ph brlens=yes;
```

Second, we root the tree with an outgroup, enforce the tree to become ultrametric, and compute the log-likelihood score for the constrained tree (criterion is still likelihood).

```
outgroup 12;
root;
Lset  clock=yes Base=est  Nst=6  Rmat=est rclass=(a b c
d b e)  Rates=gamma  Shape=est  Pinvar=0;
describetree/plot=ph brlens=yes;
```

PAUP* will respond to these commands by printing phylograms (plot=ph) and tabulated branch lengths. To obtain branch lengths for the ultrametric tree it is vital that these commands are not performed under the parsimony criterion, so: set crit=likelihood;. The tips of the terminal branches will then be shown on a straight line. The table shows results concerning the ultrametric tree. We see that the branch lengths (in blue) of the sisters *Homo* and *Pan* have been constrained to be equal. (The 'noclock tree' had lengths 0.050041 for Homo and 0.060643 for Pan.) While the uncorrected distance (p-distance) between these taxa is 8.93% in Data set 1, we see that the divergence has been adjusted to 10.52% (5.26+5.26) by our modelling.

We may use the likelihood ratio calculator in Modeltest or Mrmodeltest to compute the test statistics 2(ln clock - ln noclock). The score for the null model was 5717.686 and for the alternative model (noclock) 5709.632. We have 12 taxa (N), so the degrees of freedom is 10 (N-2). The probability of observing the resulting ratio of 16.108398 under a correct null model is 0.096571. This is not significant with an alpha level of 0.01 so we cannot reject the hypothesis that the sequences have evolved under constant rates.

Hence we adopt the ultrametric tree and use the 'empirical divergence rate' of 2% per million years to compute (with Excel) time intervals from the branch lengths (see table, yellow cells). Under the assumption of an ultrametric tree, the distance between two sister species is the result of equal evolution rates in both species. Accordingly, the rate of evolution in a lineage must be half of the divergence rate, that is 0.01 per million years. Thus in the case of *Homo*, which has a branch length of 0.0526 we compute 0.0526/0.01= 5.26 million years from the tip to the most recent common ancestor with *Pan*.

By saving the ultrametric tree with branch lengths

```
savetree file=dat1_ultrametric.tre brlens=yes
```

we may open the tree in a text editor and manually replace branch lengths

```
tree PAUP_1 = [&R]
(1:0.650300,((((((2:0.052598,3:0.052598):0.019092
```

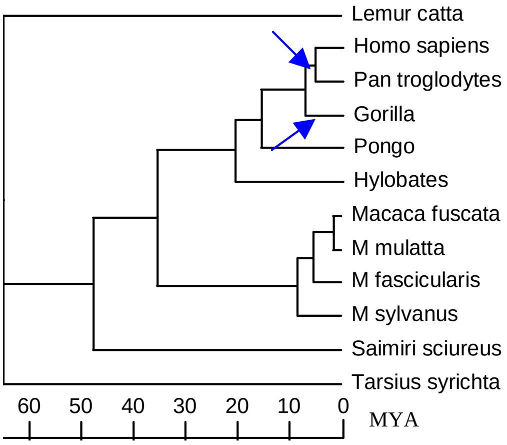
with lineage duration **time** (see table)

```
tree PAUP_1 = [&R]
(1:65.0300,((((((2:5.2598,3:5.2598):1.9092
```

to produce a calibrated tree. When the file is opened in Rod Page's program TreeViewX and displayed as a phylogram, a scale is generated for the branch lengths.

The position of *Sahelanthropus*

According to the computations above, *Gorilla* diverged from the most recent common ancestor (MRCA) of *Homo* and *Pan* about 7.17 million years ago. MRCA of *Homo* and *Pan* sustained about 1.9 myrs before it speciated. If *Sahelanthropus* is 6-7 myrs, the fossil may indeed represent MRCA of *Homo* and *Pan* or an extinct deviation from this very same lineage. However, the fossil may also be associated with the *Gorilla* branch (arrows).



We did not compute some sort of confidence intervals on the dates of nodes, and the gap in the time line from ((Homo,Pan) Gorilla) to the maximum estimate of *Sahelanthropus* age is just 0.17 myrs. If we had used the model parameter estimates found during model testing to optimise the branch lengths of the ultrametric tree, this time gap would shrink to 0.15 myrs. Although this is not a big difference, it may be a reminder of the inherent approximations of molecular clock estimates. Our initial calculations of proportions differing between sequences indicate different rates in different genes. It is obvious that the total substitution rate depends on the relative composition of fast and slow evolving genes in the dataset, so application of a 'universal empirical rate' is potentially misleading and may affect the dating substantially. The importance of the model is underscored by the fact that application of a HKY model would reduce the ultrametric branch length of *Homo* by one million years! The imprecise dating of *Sahelanthropus* to 6-7 myrs also is a problem. Alternative placement of *Sahelanthropus* on the branch representing MRCA of *Homo*, *Pan* and *Gorilla* thus does not seem as a very far-fetched proposal. After all, the phylogenetic relationship of *Sahelanthropus* cannot be resolved unless the fossils show shared synapomorphic character states that are unique to any of the candidate clades.

Some advice based on review of submitted material

- It seems that some of you may have unintentionally forgotten an **include all** command in some instances and reconstructed phylogenies on very small data sets. MtDNA genes are linked on the same chromosome. They should reflect the 'same phylogenetic history '. Include all the data in a phylogeny reconstruction unless the alignment is ambiguous in some partitions.
- **Lset** is used to define likelihood models, while **Lscore** with options is used to compute log-likelihood for trees residing in memory. You don't need **Lscore** in order to compute trees.
- **Begin paup;** (or **Begin MrBayes;**) is required just once in a batch script.
- Remember that default or changed settings involving search-options, model-specifications, character-exclusion, criterion etc. remain in PAUP* memory until they are actually changed. Reflect on implications that this might have for the composition and order of commands (with options) in your batch file. Are the results going to be identical the next time you run the batch script?
- Remember that a consensus tree is a summary of two or more trees. There is no point in computing and presenting a consensus tree of one tree.
- The **bootstrap** command with option **treefile** stores the bootstrap replicate trees in a file. You may use the **contree** command to compute and save a consensus bootstrap tree in a separate file.
- Make sure that the root of your tree(s) reflects your intended direction of the time dimension. (Confer the comparison of trees from the two data sets.) (The default in PAUP* is to automatically root with the first taxon in the matrix as outgroup.)
- Make sure that you understand the fundamental concepts of monophyly and paraphyly, and remember that alternative roots on a tree make a difference in this respect. (Apomorphies are certainly also affected by rooting.)
- Remember that computation of posterior probabilities in Bayesian analysis requires a large sample of trees. However, extremely large result files may overrun the capacity of programs that you might want to use for post run analysis (Excel and others). It is better to run many MCMC generations with lengthy sampling intervals than few generations with frequent sampling.
- Remember that a ML model that fits a large data set (comprised of protein coding and tRNA genes) may not necessarily be the best model for separate analyses of its partitions. MrBayes allows for mixed models. Thus, a possible procedure would be to first 'modeltest' each partition in separate analyses, and to subsequently apply the best model for each of the partitions in a Bayesian analysis of the combined data set.