

Answers

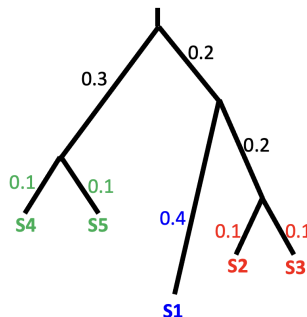
Practical 1

Let's plot and manipulate this tree:

1) represent this tree in Newick format.

2) Save the Newick string in a text file, then open it (File->Open) in FigTree

Does the tree look the same? Try to make it look similar (select branch, use "Rotate")!



Newick: "((S1:0.4,(S2:0.1,S3:0.1):0.2):0.2,(S4:0.1,S5:0.1):0.3);"

Practical 2

1) Run MAFFT of the primate-mtDNA-unaligned.fasta file

```
Making a distance matrix ..
  1 / 12
done.

Constructing a UPGMA tree (efffree=0) ...
 10 / 12
done.

Progressive alignment 1/2...
STEP   11 / 11   f
done.

Making a distance matrix from msa..
  0 / 12
done.

Constructing a UPGMA tree (efffree=1) ...
 10 / 12
done.

Progressive alignment 2/2...
STEP   11 / 11   f
done.

disttbfast (nuc) Version 7.520
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 thread(s)

Strategy:
FFT-NS-2 (Fast but rough)
Progressive method (guide trees were built 2 times.)
```

Practical 2

2) Compare aligned and unaligned files : what's the difference ?

Unaligned

```
>Gorilla
AAGCTTCACGGCGCAGTTGTTCTTATAATTGCCACGGACTTACATCATCATTATTATT
CTGCCTAGCAAACCTCAAACCTACGAACGAACCCACAGCCGCATCATAATTCTCTCTCAAGG
ACTCCAACCCCTACTCCCCTAATAGCCCTTTGATGACTTCTGGCAAGCCTCGCCAACCT
CGCCTTACCCCCACCATTAACTACTAGGAGAGCTCTCCGTACTAGTAACCACTTCTC
CTGATCAAAACCCACCCCTTTTACTTACAGGATCTAACAATACTAATTACAGCCCTGTACTC
CCTTTATATATTTACCAACAACAAATGAGGCCCACTCACACACCACATCACCAACATAAA
ACCCCTCATTACACGAGAAAAATCCTCATATTATGCACCTATCCCCATCCTCCTCCT
ATCCCTCAACCCCGATATTATCACCGGTTACCTCCTGTAATATAGTTTAAACAAAAC
ATCAGATTGTGAATCTGATAACAGAGGCTCACAACCCCTTATTTACCGAGAAAGCTCGTA
AGAGCTGCTAACTCATACCCCGTGTGACAACATGGCTTTCTCAACTTTTAAAGGATA
ACAGCTATCCATTGGTCTTAGGACCCAAAATTTTGGTGCAACTCCAATAAAAGTAATA
ACTATGTACGCTACCATAACCCCTTAGCCCTAACTTCCTTAAATCCCCCTATCCTTACC
ACCTTCATCAATCCTTAACAAAAAAGCTCATACCCCATACGTAATAATCTATCGTCGCA
TCCACCTTTATCATCAGCTCTTCCCCACAACAATATTTCTATGCCTAGACCAAGAAGCT
ATTATCTCAAGCTGACACTGAGCAACAACCCCAACAATTCAACTCTCCCTAAGCTT
```

Aligned

```
>Gorilla
aagccttcacggcgagttgttcttataattgccacggacttacatcatcattattatt
ctgccttagcaaactcaaactacgaacgaacccacagccgcacataattctctctcaagg
actccaacccctactcccactaatagccctttgatgactcttggaagcctcgccaacct
cgcttaccocccaccattaaactactaggagagctctccgtactagtaaccacattctc
ctgatcaaacaccacccctttacttacaggatctaacatactaattacagccctgtactc
cctttatataatttaccacaacacaatgaggccactcacacaccacatcaccacataaa
accctcatttacagagaaaaacatcctcataattcatgcacctatcccccacctcctcct
atccctcaaccccgatattatcacgggttcacctcctgtaaatatagtttaacaaaaac
atcagattgtgaatctgataacagaggctcaacaaccccttattaccgagaaagctcg
taagagctgctaactcatacccccgtgcttgacaacatggctttctcaacttttaaagga
taacagctatccattggctcttaggacccaaaaatttgggtgcaactcacaataaaagtaa
taactatgtacgctaccataacccacttagccctaactccttaattccccctatcctta
ccaccttcatacctaatacaaaaaaagctcatacccccattacgtaaaatctatgctg
catccaccttatacatcagcctcttcccccacaacaatatcttatgcttagaccaagaag
ctattatctcagctgacactgagcaacaacccaacaattcaactctccctaagctt
```

The IMPORTANT difference is that in the aligned file all sequences have the same length, and gap characters “-” have been added

Practical 2

3) Run MAFFT also on the SARS-CoV-2 sequences. Why does it take longer ?

```
nthread = 0
nthreadpair = 0
nthreads = 0
ppenalty_sv = 0
stacksize: 8192 kb
generating a scoring matrix for nucleotide (dist=200) ... done
Gap Penalty = -1.53, +0.00, +0.00

Making a distance matrix ..
101 / 136
done.

Constructing a UPQMA tree (efftree=0) ...
130 / 136
done.

Progressive alignment 1/2...
STEP 135 / 135 f
done.

Making a distance matrix from msa..
130 / 136
done.

Constructing a UPQMA tree (efftree=1) ...
130 / 136
done.

Progressive alignment 2/2...
STEP 135 / 135 f
done.

disttofast (nuc) Version 7.520
aln=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 threads(s)

Strategy:
FFT-NS-2 (Fast but rough)
Progressive method (guide trees were built 2 times.)

If unsure which option to use, try 'mafft --auto input > output'.
For more information, see 'mafft --help', 'mafft --man' and the mafft page.

The default gap scoring scheme has been changed in version 7.110 (2013 Oct).
It tends to insert more gaps into gap-rich regions than previous versions.
To disable this change, add the --leavegappyregion option.
```

We now have more (136 vs 11) and longer (30,000 vs 100s of bp) genomes, hence the longer runtime. Higher divergence also leads to longer runtime

Practical 3

1) Run IQ-TREE 2 with GTR substitution model on the alignments we previously generated

```
Computing ML distances based on estimated model parameters...
Computing ML distance took 0.000790 sec (of wall-clock time) 0.000672 sec (of CPU time)
Computing RapidNJ tree took 0.000043 sec (of wall-clock time) 0.000051 sec (of CPU time)
Log-likelihood of RapidNJ tree: -5945.376
```

Distance matrix calculation and NJ

```
|-----|
|           INITIALIZING CANDIDATE TREE SET           |
|-----|
Generating 98 parsimony trees... 0.066 second
Computing log-likelihood of 97 initial trees ... 0.045 seconds
Current best score: -5945.484
```

Estimate initial parsimony trees

```
Do NNI search on 20 best initial trees
Estimate model parameters (epsilon = 0.100)
BETTER TREE FOUND at iteration 1: -5944.643
Iteration 10 / LogL: -5944.665 / Time: 0h:0m:0s
Iteration 20 / LogL: -5944.646 / Time: 0h:0m:0s
Finish initializing candidate tree set (1)
Current best tree score: -5944.643 / CPU time: 0.167
Number of iterations: 20
```

Optimize parsimony trees with ML

```
|-----|
|           OPTIMIZING CANDIDATE TREE SET           |
|-----|
Iteration 30 / LogL: -5949.319 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 40 / LogL: -5944.647 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 50 / LogL: -5944.651 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 60 / LogL: -5944.656 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 70 / LogL: -5944.645 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 80 / LogL: -5944.728 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 90 / LogL: -5944.654 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 100 / LogL: -5944.648 / Time: 0h:0m:0s (0h:0m:0s left)
TREE SEARCH COMPLETED AFTER 102 ITERATIONS / Time: 0h:0m:0s
```

```
|-----|
|           FINALIZING TREE SEARCH                 |
|-----|
```

```
Performs final model parameters optimization
Estimate model parameters (epsilon = 0.010)
1. Initial log-likelihood: -5944.643
Optimal log-likelihood: -5944.643
Rate parameters: A-C: 10.05648 A-G: 28.54357 A-T: 5.60546 C-G: 2.88589 C-T: 35.20844 G-T: 1.00000
Base frequencies: A: 0.324 C: 0.304 G: 0.186 T: 0.266
Parameters optimization took 1 rounds (0.001 sec)
BEST SCORE FOUND : -5944.643
Total tree length: 1.535
```

Model optimization

```
Total number of iterations: 102
CPU time used for tree search: 0.661 sec (0h:0m:0s)
Wall-clock time used for tree search: 0.481 sec (0h:0m:0s)
Total CPU time used: 0.692 sec (0h:0m:0s)
Total wall-clock time used: 0.520 sec (0h:0m:0s)
```

Summary

Analysis results written to:

```
IQ-TREE report: /Users/demaio/Desktop/presentations/2023/phylogenetics/materials/01-foundations/primate-mtDNA_mafft-aligned_iqtreeGTR.iqtree
Maximum-likelihood tree: /Users/demaio/Desktop/presentations/2023/phylogenetics/materials/01-foundations/primate-mtDNA_mafft-aligned_iqtreeGTR.treefile
Likelihood distances: /Users/demaio/Desktop/presentations/2023/phylogenetics/materials/01-foundations/primate-mtDNA_mafft-aligned_iqtreeGTR.nldist
Screen log file: /Users/demaio/Desktop/presentations/2023/phylogenetics/materials/01-foundations/primate-mtDNA_mafft-aligned_iqtreeGTR.log
```

Output

Practical 3

2) Now let IQ-TREE 2 investigate which model best fits the data (ModelFinder). Does IQ-TREE 2 select a model with rate variation ?

```
NOTE: ModelFinder requires 1 MB RAM!
ModelFinder will test up to 286 DNA models (sample size: 898) ...
No. Model      -LnL      #F AIC      AICc      BIC
1  GTR+F      9644.739   29 11947.478  11949.483  12086.083
2  GTR+F+I    9748.157   30 11588.314  11582.459  11724.219
3  GTR+F+G4    9721.695   38 11080.190  11085.335  11647.195
4  GTR+F+I+G4  9722.005   31 11586.011  11588.302  11654.816
5  GTR+F+R2    9726.477   31 11513.233  11517.244  11663.759
6  GTR+F+R3    9721.563   33 11509.127  11511.724  11667.532
16  TVM+G4     9822.223   27 11698.446  11700.184  11828.051
17  TVM+I+G4   9821.499   28 11698.198  11700.457  11832.403
29  TVM+F+G4   9721.748   29 11581.488  11583.485  11648.685
38  TVM+F+I+G4  9722.135   38 11584.270  11586.415  11648.275
42  TVM+G4     9871.328   26 11794.056  11796.158  11919.461
43  TVM+I+G4   9870.562   27 11795.124  11796.862  11924.729
55  TIM3+F+G4   9728.695   28 11513.390  11515.258  11647.794
56  TIM3+F+I+G4  9729.225   29 11516.451  11518.455  11655.056
68  TIM3+G4     9896.586   25 11843.168  11844.658  11963.172
69  TIM3+I+G4   9895.803   26 11843.685  11845.217  11968.418
81  TIM2+F+G4   9723.844   28 11503.688  11505.587  11638.893
82  TIM2+F+I+G4  9724.338   29 11506.676  11508.681  11645.881
94  TIM2+G4     9828.287   25 11786.414  11787.984  11826.418
95  TIM2+I+G4   9827.271   26 11786.562  11788.154  11831.346
187  TIM+F+G4    9729.896   28 11514.193  11516.462  11648.598
188  TIM+F+I+G4  9729.512   29 11517.025  11519.029  11656.238
128  TIM+G4      9891.368   26 11832.719  11834.218  11952.723
129  TIM+I+G4   9890.185   26 11832.289  11833.821  11957.814
133  TPM3u+F+G4  9728.784   27 11511.488  11513.146  11641.812
134  TPM3u+F+I+G4  9729.223   28 11514.446  11516.315  11648.858
146  TPM3+F+G4   9728.784   27 11511.488  11513.146  11641.812
147  TPM3+F+I+G4  9729.223   28 11514.446  11516.315  11648.858
159  TPM2u+F+G4  9723.978   27 11501.948  11503.677  11631.544
168  TPM2u+F+I+G4  9724.449   28 11504.898  11506.767  11639.383
172  TPM2+F+G4   9723.964   27 11501.928  11503.666  11631.533
173  TPM2+F+I+G4  9724.449   28 11504.898  11506.767  11639.383
185  K3P+F+G4    9729.888   27 11512.161  11513.899  11641.765
186  K3P+F+I+G4  9729.584   28 11515.088  11516.877  11649.413
198  K3P+G4     9644.214   24 11936.428  11937.803  12051.632
199  K3P+I+G4   9643.358   25 11936.699  11938.198  12056.784
211  TN+F+G4     9738.635   27 11515.269  11517.007  11644.874
212  TN+F+I+G4   9731.187   28 11518.374  11520.243  11652.778
224  TNe+G4     9888.735   24 11845.471  11848.845  11964.675
225  TNe+I+G4   9889.938   25 11849.861  11851.352  11969.865
237  HKY+F+G4    9738.635   26 11513.271  11514.883  11638.075
238  HKY+F+I+G4  9731.191   27 11516.382  11518.128  11645.986
250  K2P+G4      9952.992   23 11951.985  11953.248  12062.389
251  K2P+I+G4    9952.671   24 11953.363  11954.717  12068.547
263  F81+F+G4    6127.854   25 12384.189  12385.690  12424.113
264  F81+F+I+G4  6124.888   26 12381.777  12383.389  12426.581
276  JC+G4       6273.172   22 12598.363  12591.588  12695.947
277  JC+I+G4    6278.751   23 12597.582  12598.765  12697.986

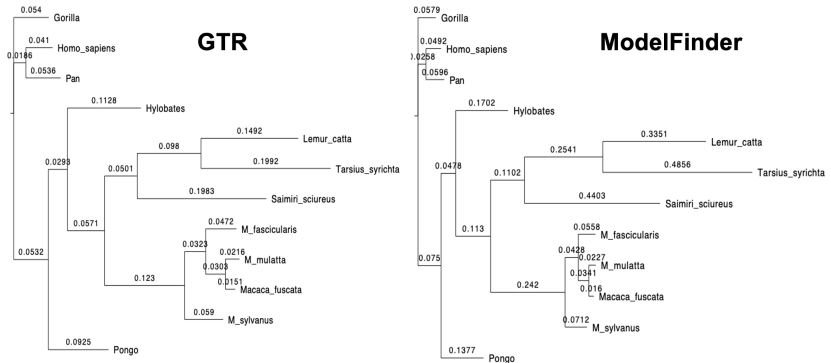
Akaike Information Criterion:      TVM+F+G4
Corrected Akaike Information Criterion: TVM+F+G4
Bayesian Information Criterion:      TPM2+F+G4
Best-fit model: TPM2+F+G4 chosen according to BIC

All model information printed to /Users/demais/Desktop/presentations/2023/phylogenetics/materials/BI-foundations/primate-mtDNA_nafft-aligned_iqtreeModelSelection.model.gz
CPU time for ModelFinder: 0.996 seconds (0h:0m:0s)
Wall-clock time for ModelFinder: 1.805 seconds (0h:0m:1s)
```

IQ-TREE 2 selects a model with rate variation (“G4”)

Practical 3

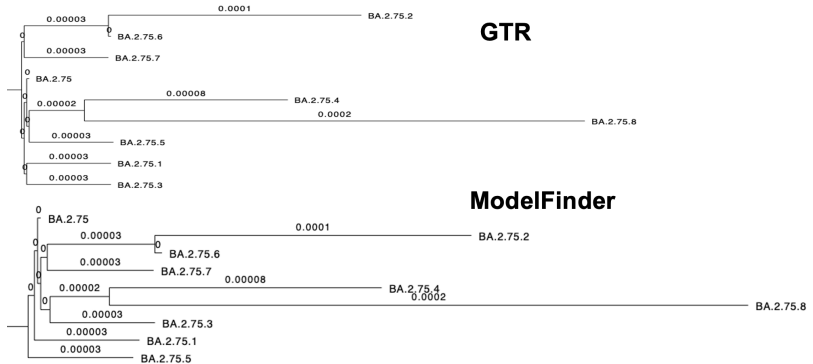
2) Now let IQ-TREE 2 investigate which model best fits the data (ModelFinder). Does the tree differ from before? How and why?



IQ-TREE 2 selects a model with rate variation ("G4")

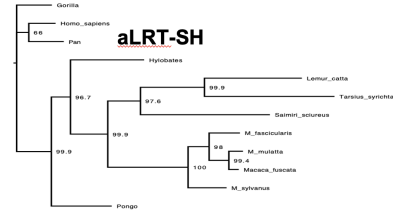
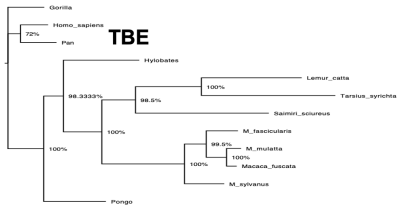
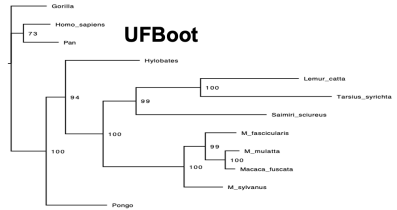
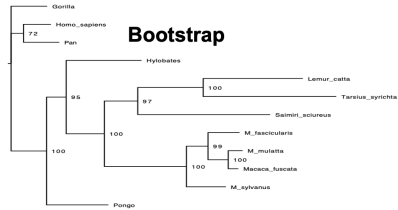
Practical 3

2) Now let IQ-TREE 2 investigate which model best fits the data (ModelFinder). Does the tree differ from before? How and why?

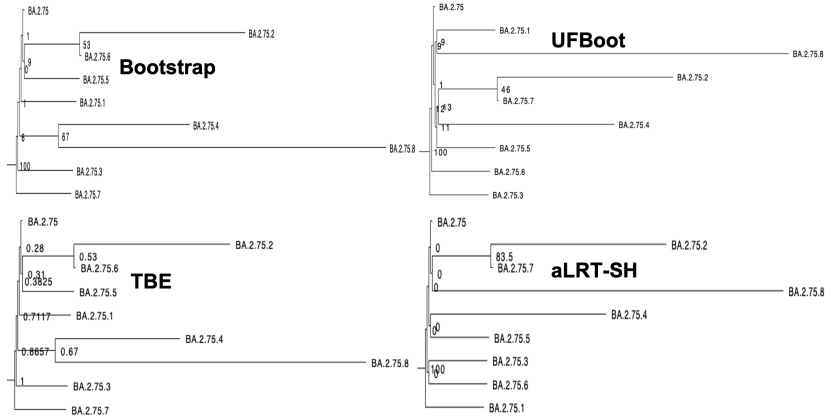


IQ-TREE 2 selects a model with rate variation ("G4")

Practical 4



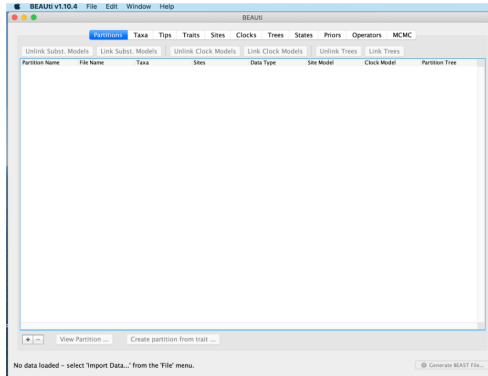
Practical 4



Practical 5

1) Create an input xml file using BEAUti

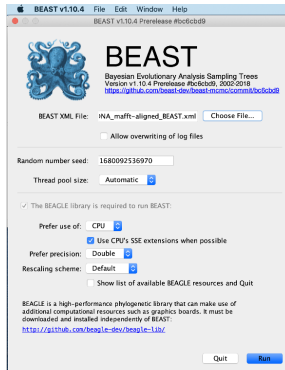
- Open BEAUti
- File->"Import data"; or, drag alignment file onto BEAUti window
- Sites-> select GTR substitution model
- Trees-> Yule process
- Click "Generate BEAST File"



Practical 5

2) Run the xml file in BEAST

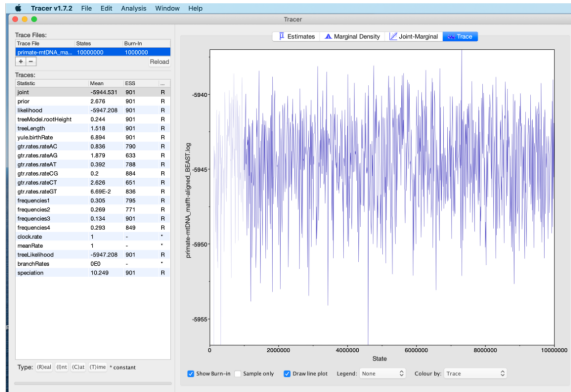
- Give BEAST the .xml file created by BEAUti ; or from the command line : `beast primate-mtDNA-mafft-aligned-BEAST.xml`
- Click “Generate BEAST File”



Practical 5

3) Analyse the output in Tracer

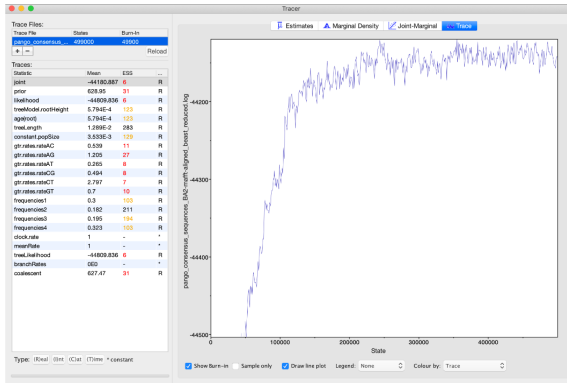
- File->”Import trace file”-> pick .log file created by BEAST ; or just drag it on Tracer the window



Practical 5

3) Analyse the output in Tracer Small ESS values (<100 , in red) mean that the MCMC needs to run longer

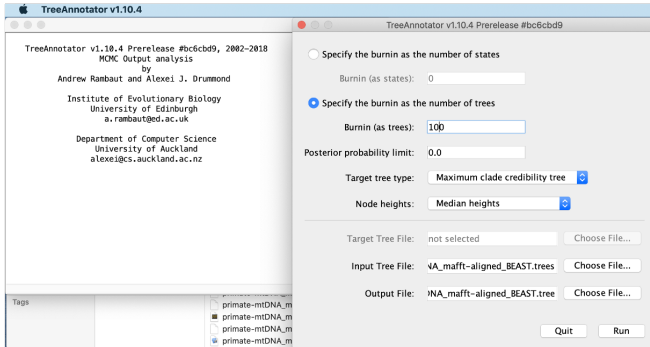
- BEAUti->MCMC-> Length of chain
- File->"Import trace file"-> pick .log file created by BEAST ; or just drag it on Tracer the window



Practical 5

3) Analyse the output in Tracer

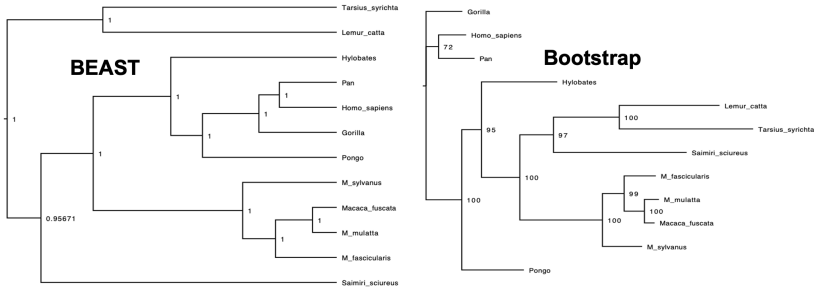
- Pick as input the .trees file created by BEAST. Choose output name



Practical 5

3) Analyse the output in Tracer

- Pick as input the tree file created by TreeAnnotator

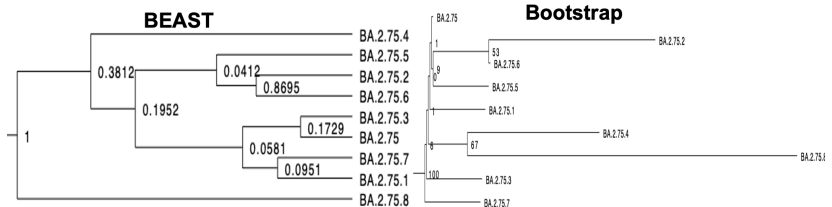


4) Compare to maximum likelihood branch support

Practical 5

3) Analyse the output in Tracer

- Pick as input the tree file created by TreeAnnotator

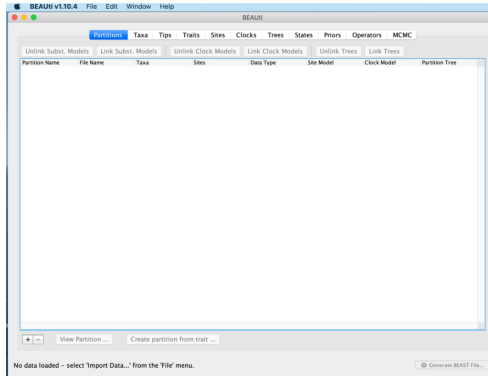


4) Compare to maximum likelihood branch support

Practical 6

1) Create an input xml file using BEAUti

- Open BEAUti
- File->"Import data"; or, drag alignment file onto BEAUti window
- Sites-> select GTR substitution model
- Trees-> Yule process
- Click "Generate BEAST File"



Practical 6

2) Include time and location data

- Tips->"Import dates"; select given "dates.txt" file
- Traits->"Import traits"; select given "locations.txt" file
- Traits->"Create partition from trait"

Partitions Taxa **Tips** Traits Sites Clocks Trees States Priors Operators MCMC

☒ Use tip dates

Parse Dates Import Dates Set Dates Clear Dates Set Uncertainty

Dates as: Years ☒ Since some time in the past ☒ Specify origin date: unable to parse date

Name	Date	Uncertainty	Height
BA.2	2022.7232876712328	0.0	0.2575342465754602
BA.2.1	2022.2931506849316	0.0	0.6876712328767098
BA.2.10	2022.2219178082191	0.0	0.7589041095891589
BA.2.10.1	2022.4191780821918	0.0	0.5616438356164508
BA.2.10.2	2022.7260273972602	0.0	0.2547945205481028
BA.2.10.3	2022.4164383561645	0.0	0.5643835616438082
BA.2.10.4	2022.2465753424658	0.0	0.7342465753424676
BA.2.11	2022.9287671232876	0.0	0.05205479452069994
BA.2.12	2022.3671232876711	0.0	0.6136986102171508
BA.2.12.1	2022.6109589041096	0.0	0.3698630136987049
BA.2.12.2	2022.0986301369862	0.0	0.8821917808220405
BA.2.13	2022.890410958904	0.0	0.09041095890415818
BA.2.13.1	2022.6383561643836	0.0	0.3424657534246762
BA.2.14	2022.1917808219177	0.0	0.789041095890525
BA.2.15	2022.6849315068494	0.0	0.29589041095891844
BA.2.16	2022.9	0.0	0.9808219178082709
BA.2.17	2022.7780821917809	0.0	0.20273972602740287
BA.2.18	2022.4191780821918	0.0	0.5616438356164508
BA.2.19	2022.5150684931507	0.0	0.4657534246575785
BA.2.2	2022.9369863013699	0.0	0.04383561643840039
BA.2.2.1	2022.5479452054794	0.0	0.4328767123288344
BA.2.20	2022.545205479452	0.0	0.4356164383561618
BA.2.21	2022.7589041095891	0.0	0.22191780821913198
BA.2.22	2022.2986301369863	0.0	0.682191780821995
BA.2.23	2022.9342465753425	0.0	0.046575342465757785
BA.2.23.1	2022.4657534246576	0.0	0.515068493150689
BA.2.24	2022.4438356164383	0.0	0.5369863013700069
BA.2.25	2022.7342465753425	0.0	0.24657534246580126
BA.2.25.1	2022.9780821917801	0.0	0.002739726027397392
BA.2.26	2022.4164383561645	0.0	0.5643835616438082

Tip date sampling: Off ☒ Apply to taxon set: All taxa ☒

Data: 136 taxa, 2 partitions [Generate BEAST File...](#)

Practical 6

Reduce the number of MCMC steps if you don't want to wait 20 minutes

- MCMC->"Length of chain"->1,000,000
- Click "Generate BEAST File"

The screenshot shows the 'MCMC' tab in the BEAST2 configuration interface. The 'Length of chain' is set to 10000000. 'Echo state to screen every' and 'Log parameters every' are both set to 10000. The 'File name stem' is 'pango_consensus_sequences_BA2-mafft-aligned', with an option to 'Add .txt suffix'. The 'Log file name' is 'pango_consensus_sequences_BA2-mafft-aligned.log' and the 'Trees file name' is 'pango_consensus_sequences_BA2-mafft-aligned.trees'. There is an option to 'Create tree log file with branch length in substitutions'. The 'Substitutions trees file name' is empty. The 'Create operator analysis file' checkbox is checked, and the 'Operator analysis file name' is 'pango_consensus_sequences_BA2-mafft-aligned.ops'. There is an option to 'Sample from prior only - create empty alignment'. A note states: 'Select the option below to perform marginal likelihood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS) or generalized stepping-stone sampling (GSS) which performs an additional analysis after the standard MCMC chain has finished.' The 'Marginal likelihood estimation (MLE)' dropdown is set to 'None'. A 'Settings' button is next to it. At the bottom left, it says 'Data: 136 taxa, 2 partitions'. At the bottom right, there is a 'Generate BEAST File...' button.

Partitions Taxa Tips Traits Sites Clocks Trees States Priors Operators **MCMC**

Length of chain: 10000000

Echo state to screen every: 10000

Log parameters every: 10000

File name stem: pango_consensus_sequences_BA2-mafft-aligned

☐ Add .txt suffix

Log file name: pango_consensus_sequences_BA2-mafft-aligned.log

Trees file name: pango_consensus_sequences_BA2-mafft-aligned.trees

☐ Create tree log file with branch length in substitutions:

Substitutions trees file name:

☒ Create operator analysis file:

Operator analysis file name: pango_consensus_sequences_BA2-mafft-aligned.ops

☐ Sample from prior only - create empty alignment

Select the option below to perform marginal likelihood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS) or generalized stepping-stone sampling (GSS) which performs an additional analysis after the standard MCMC chain has finished.

Marginal likelihood estimation (MLE): None

Settings

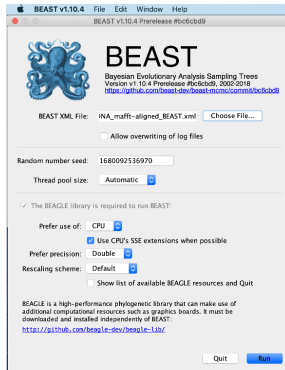
Data: 136 taxa, 2 partitions

Generate BEAST File...

Practical 6

Now you can run BEAST with the new xml file

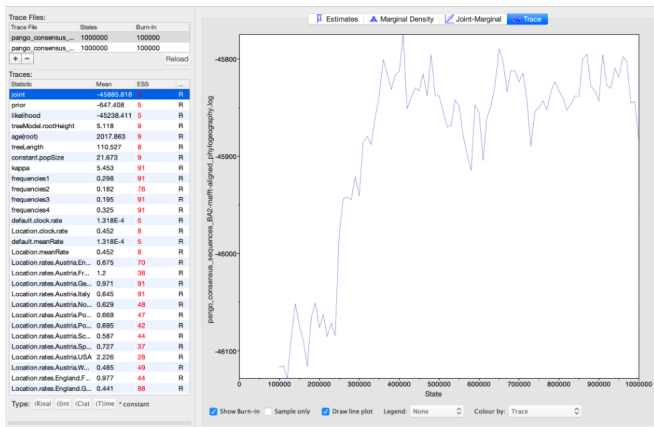
- Give BEAST the .xml file created by BEAUti ; or from the command line : `beast primate-mtDNA-mafft-aligned-BEAST.xml`
- Click “Generate BEAST File”



Practical 6

Check results in Tracer

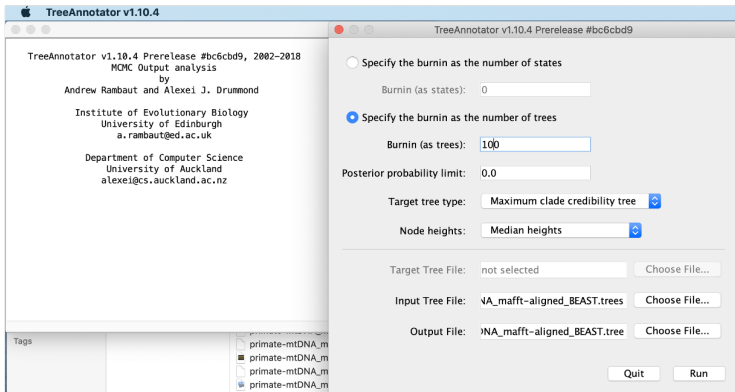
- File->"Import trace file"-> pick .log file created by BEAST ; or just drag it on Tracer the window
- A proper analysis will need to run longer than 1,000,000 MCMC steps



Practical 6

Process the output in TreeAnnotator

- Pick as input the .trees file created by BEAST. Choose output name



Practical 6

Visualize tree in Figtree

- Pick as input the tree file created by TreeAnnotator

