BS32010: Applied Bioinformatics: Assignment 4: Phylogenetic

Fiona Macfarlane (110010712)

Phylogenetic analysis can be used to assess the evolutionary relationships between organisms. The RAG1 gene, a recombination-activating gene, is important for the development of mature T and B lymphocytes, which are crucial in an adaptive immune system. By investigating the similarities and differences between the RAG1 genes in mammals, a phylogenetic analysis can be undertaken to investigate evolutionary relationships. (1)

103 sequences of Rag1 genes from different species were compared in the multiple sequence alignment. However some of the sequences had the same name, to overcome this the sequence identifiers were altered to numbers, 1 to 103.

The multiple sequence alignment was read into R studio and the row names (sequence identifiers) altered, using the commands;

```
x<-read.dna("marsup_rag1.fasta",format="fasta")

rownames(x)<-(1:103)
```

This changes the species names to numbers and allows the data from the fasta file to be used in phylogenetic analysis.

To compare the number of substitutions separating any pair of gene sequences, a distance matrix can be calculated and inserted into a table. The default, K80 model is used here .This table can then be saved as a csv file, distances1.csv.

```
d<-dist.dna(x)

write.table(as.matrix(d),"distances1.csv")
```

Using the distances that have been calculated, phylogenetic trees can be constructed. There are several methods of tree construction, including BIONJ (Bio Neighbour Join), NJ (Neighbour Join) and UPGMA (Unweighted pair group method in arithmetic mean). The three methods were used to construct trees and their residuals calculated.

Tree construction method 1) BIONJ

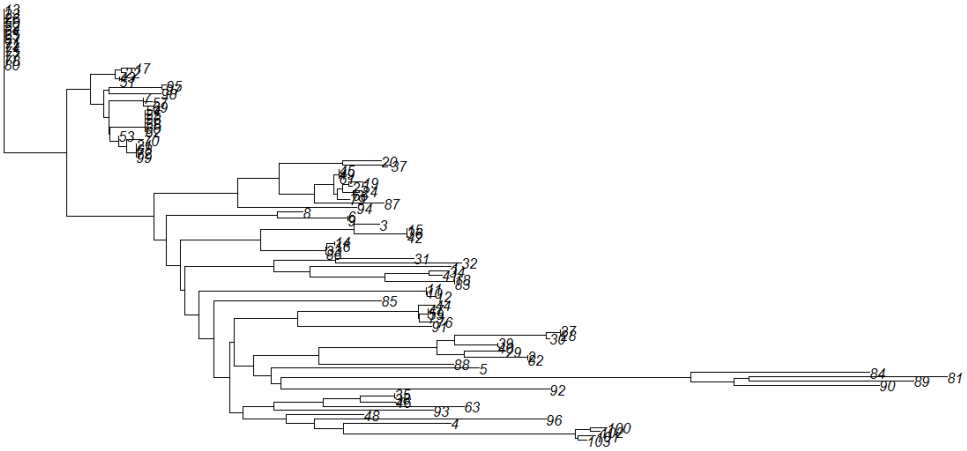
The tree can then be created using the BIONJ method, using the distances calculated;

```
tr.Bionj<-Bionj(d)
```

The tree can then be visualised by plotting it;

```
plot(tr.Bionj)
```

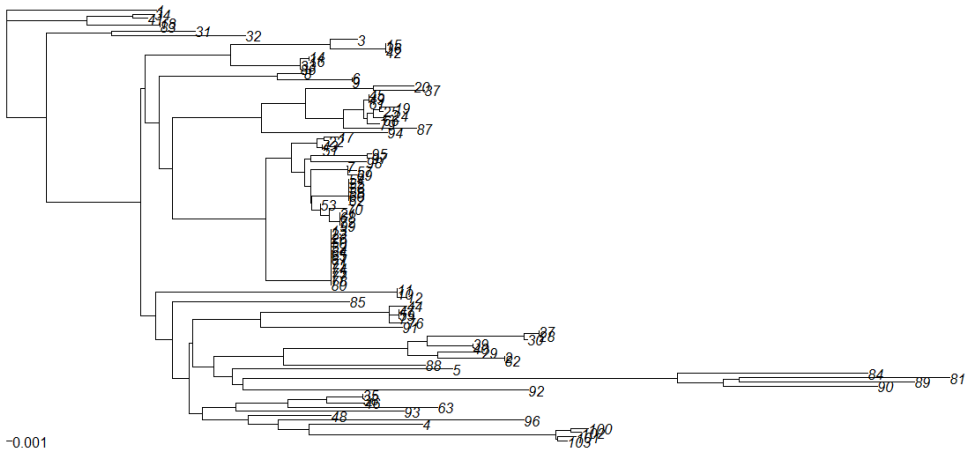
Plot 1: Unrooted tree using Bionj method:



The phylogenetic tree can also be rooted, one node is selected then the tree is plotted using this as a base. The other nodes are plotted in accordance to their relationships with the base node.

```
tr.Bionjr<-root(tr.Bionj, outgroup="1", resolve.root=TRUE)

plot(tr.Bionjr);add.scale.bar(length=0.001)
```

Plot 2: Rooted tree using Bionj method



Tree construction can distort the distances between taxa. The distortion can be calculated by using cophenetic analysis. This measures the new distance from the tree from the original distance in the distance matrix.

The following command calculates the cophenetic analysis;

```
dt.Bionj<-cophenetic(tr.Bionj)
```

We can pull out the taxa data as a matrix;

dmat<-as.matrix(d)

We can then pull out the rownames of the taxa data matrix;

nms<-rownames(dmat)

This allows us to update the cophenetic analysis with the rownames (nms);

dt.Bionj<-dt.Bionj[nms, nms]

This is then used to calculate the new distances;

dt.Bionj<-as.dist(dt.Bionj)

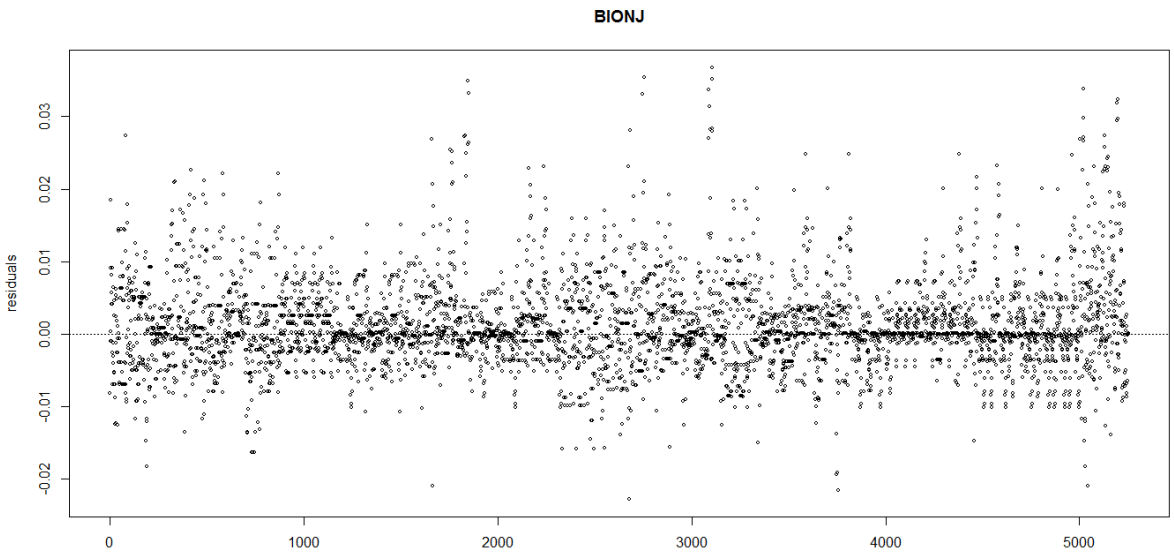
Then we can plot the difference between the new and original distances;

plot(dt.Bionj-d,ylab="residuals", cex=0.5,main="BIONJ")

abline(h=0,lty=3)

This plot of the difference displays the residuals of the tree construction method.

Plot3: Residuals of the Bionj method:



Tree construction method 2) NJ

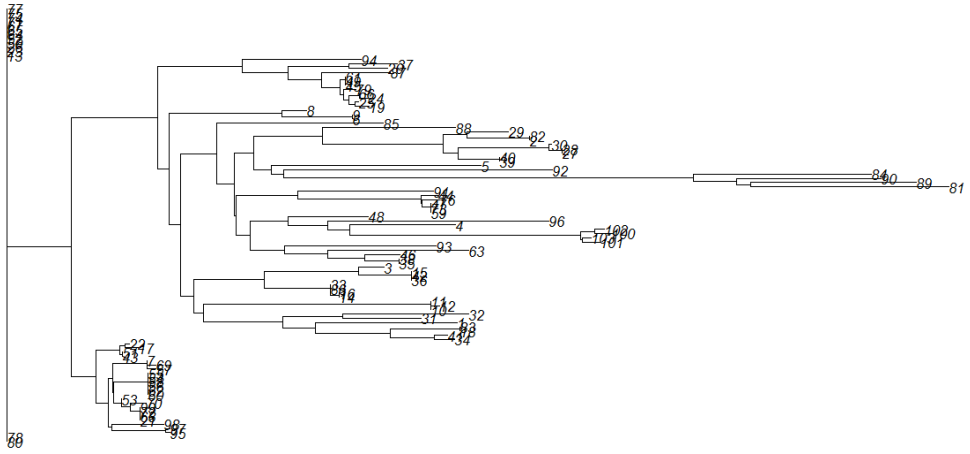
The above process was repeated for the Neighbour join method (NJ).

The NJ tree was constructed and plotted;

tr.nj<-nj(d)

plot(tr.nj)

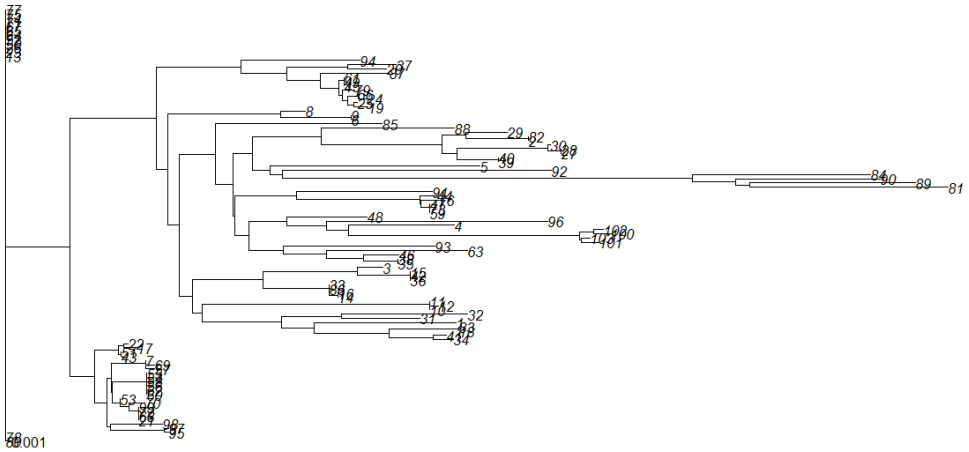
Plot4: NJ unrooted tree



A rooted tree was also constructed and plotted:

```
tr.njr<-root(tr.nj,outgroup="1", resolve.root=TRUE )
plot(tr.nj);add.scale.bar(length=0.001)
```

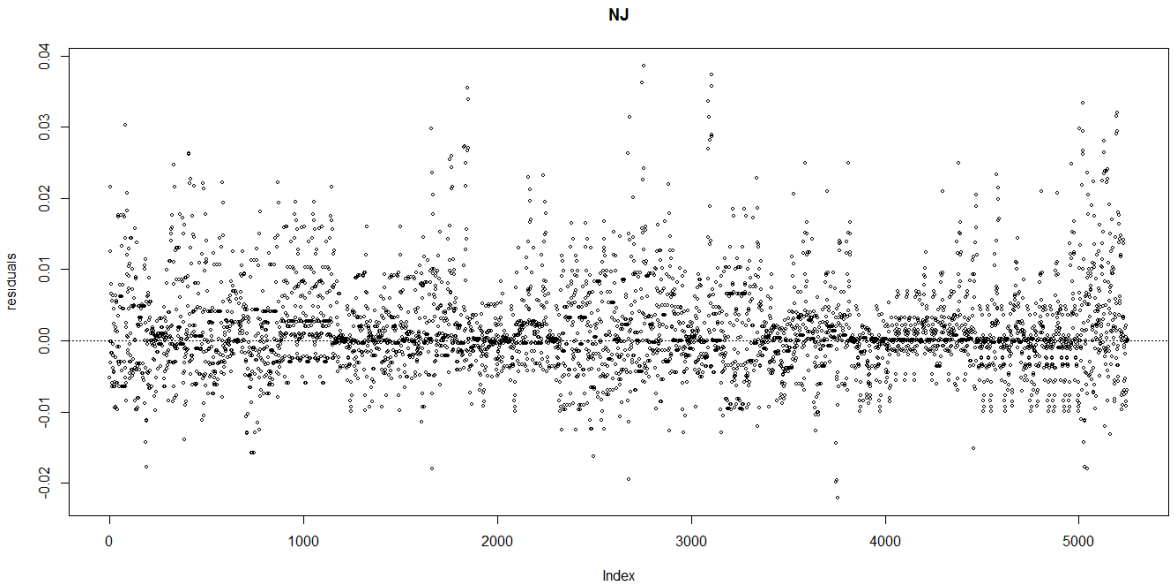
Plot5: Rooted NJ tree:



The distortion was calculates and the residual values plotted:

```
dt.nj<-cophenetic(tr.nj)
dmat<-as.matrix(d)
nms<-rownames(dmat)
dt.nj<-dt.nj[nms, nms]
dt.nj<-as.dist(dt.nj)
plot(dt.nj-d,ylab="residuals", cex=0.5,main="NJ")
abline(h=0,lty=3)
```

Plot 6: Residuals of the NJ method



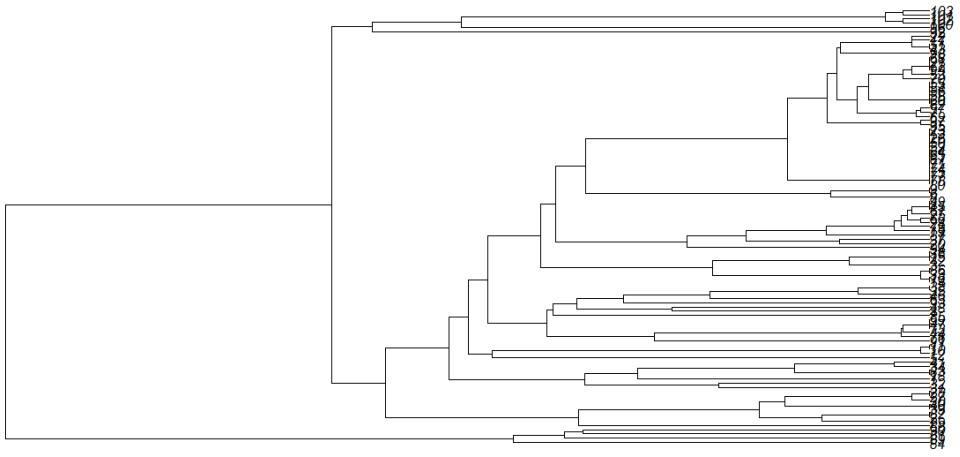
Tree construction method 3: UPGMA

The same process was applied using the UPGMA method , as that that had been used for NJ and BIONJ.

The unrooted tree was constructed and plotted:

```
tr.Upgma<-Upgma(d)

plot(tr.Upgma)
```

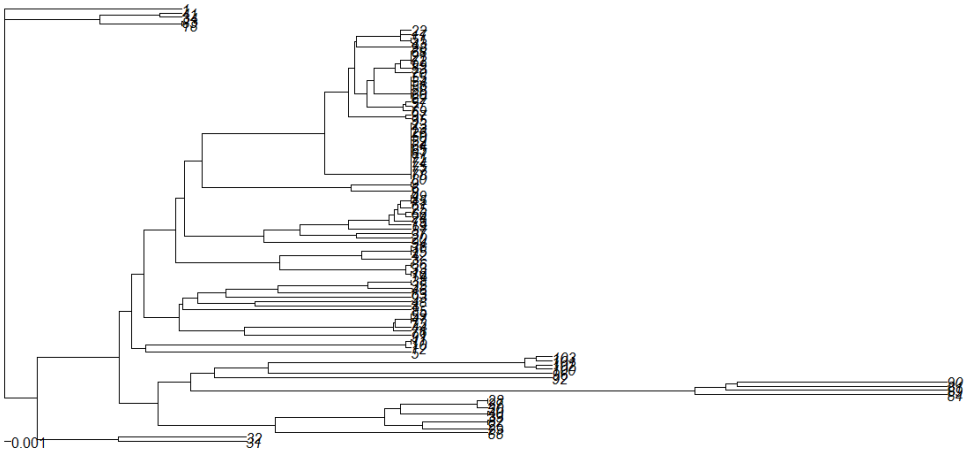
Plot7: UPGMA unrooted tree:



A rooted tree was then constructed and plotted using the UGMA method:

```
tr.Upgmar<-root(tr.Upgma,outgroup="1", resolve.root=TRUE )

plot(tr.Upgmar);add.scale.bar(length=0.001)
```

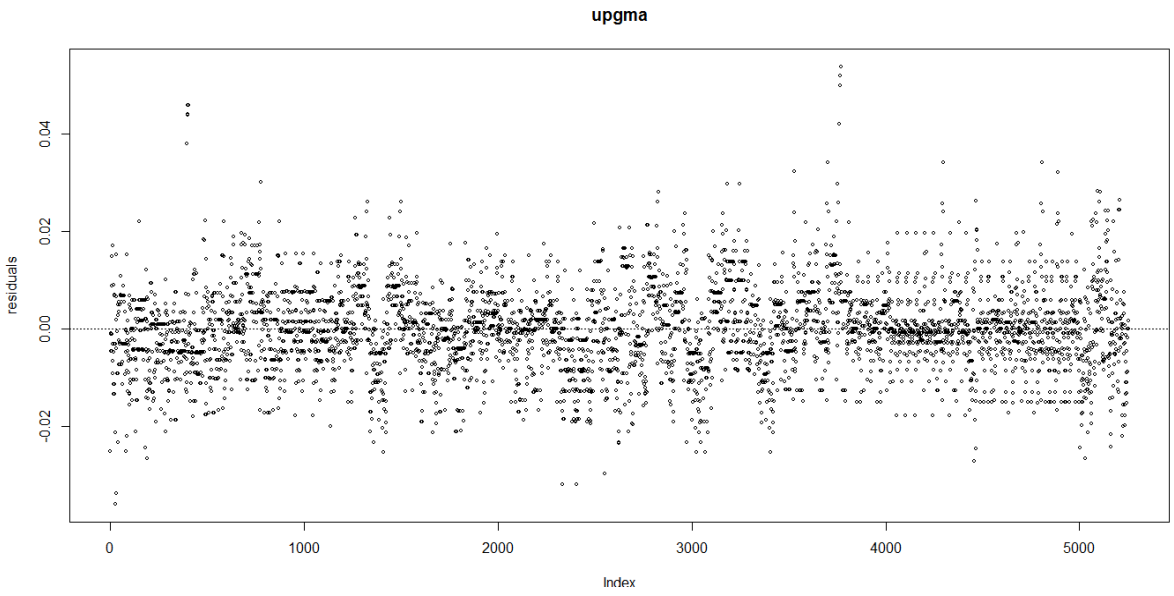
Plot8:UPGMA rooted tree



The distortion was calculated and the residuals,plotted :

```
dt.Upgma<-cophenetic(tr.Upgma)
dmat<-as.matrix(d)
nms<-rownames(dmat)
dt.Upgma<-dt.Upgma[nms, nms]
dt.Upgma<-as.dist(dt.Upgma)
plot(dt.Upgma-d,ylab="residuals", cex=0.5,main=Upgma")
abline(h=0,lty=3)
```

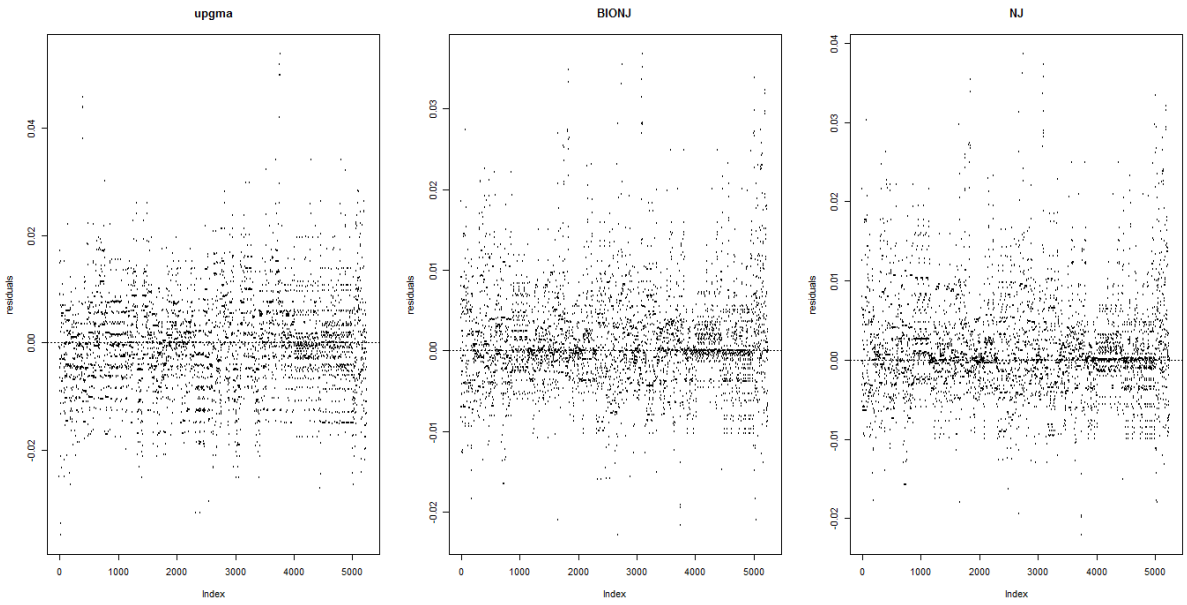
Plot9: Residuals of the UPGMA method



The 3 methods can then be compared. The trees plotted for each method do show some differences. However the best way to test which method is best is to compare the distortion. The residual plots for each of the methods can be plotted together to allow comparrisons to be made.

```
par(mfrow=c(1,3))# allows 3 plots to be displayed side by side

plot(dt.Upgma-d,ylab="residuals", cex=0.5,main="Upgma")

abline(h=0,lty=3)

plot(dt.nj-d,ylab="residuals", cex=0.5,main="nj")

abline(h=0,lty=3)

plot(dt.Bionj-d,ylab="residuals", cex=0.5,main="Bionj")

abline(h=0,lty=3)
```

Plot10: Comparisons of the 3 methods:



In this case, the BIONJ method gave the lowest residual values. This indicates that the BIONJ method gives the least amount of distortion, and therefore is the most favourable method to use for tree construction. This is to be expected since the BIONJ method is an improved version of the NJ method and is suitable for creating trees using the distances. BIONJ also has a better topological accuracy than other construction methods. (2)

Model Test

There are various models that can be used to calculate the distances between taxa. A model test can be run on the data from the fasta file to evaluate which model is the most suitable.

```
mt<-modelTest(as.phyDat(x),G=F,I=F)
```

View(mt)

Table 1: Results of the Model test:

	Model	df	logLik	AIC	BIC
1	JC	203	-6643.162	13692.32	14585.58
2	F81	206	-6629.199	13670.40	14576.85
3	K80	204	-6343.507	13095.01	13992.67
4	HKY	207	-6324.797	13063.59	13974.45
5	SYM	208	-6308.113	13032.23	13947.48
6	GTR	211	-6298.045	13018.09	13946.54

The Akaike information criterion (AIC) values are a measure of the relative quality of the model for the specific data. (3) The preferred model is the model with the lowest AIC values, since the AIC is

$$AIC = 2k - 2\ln(L)$$

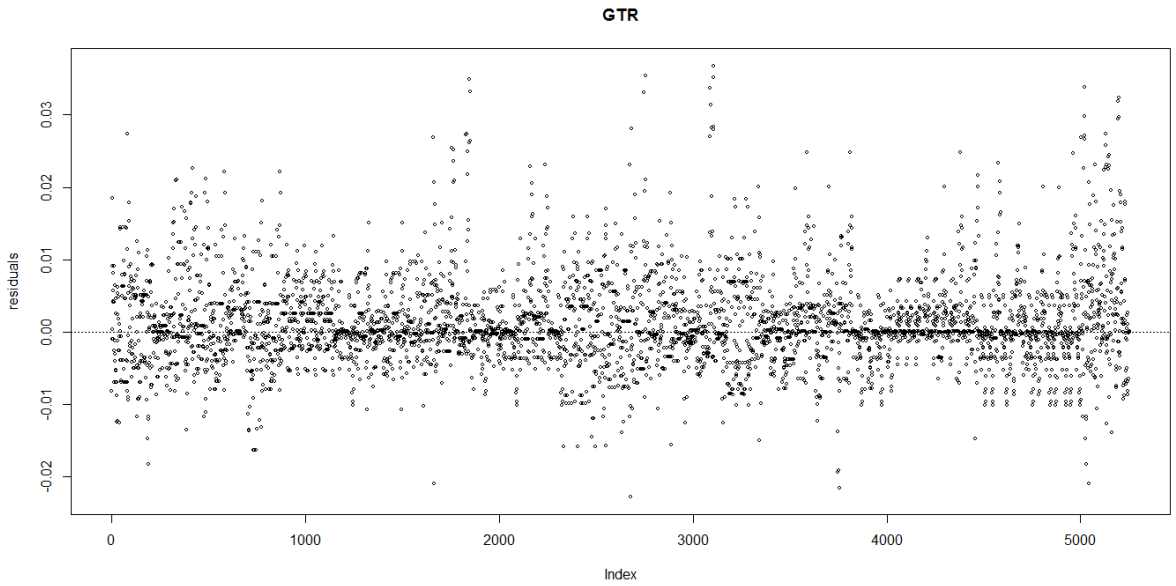
Where *k* is the number of parameters in the statistical model, and *L* is the maximized value of the likelihood function for the estimated model. From the results of the model test on the marsupial data, here the most suitable model to use is the GTR model, as it has the lowest AIC value.

The GTR (Generalised Time Reversible) model was first described by Simon Tavare in 1986. It is a one of the more generally used substitution models. The model is independent, preventing the changes in one site affecting the probability of changes in another site. The model is also a finite site model, allowing a single site to be changed multiple times. (4)

The K80 (default) model that was used in previous calculations can be replace by the GTR method. The distortion levels of this model can be calculated using cophenetic analysis, and the residuals plotted:

```
fittedtreeGTR<-pml(tr.Bionj,as.phyDat(x),k=4,inv=0.2)
dt.fittedtreeGTR<-cophenetic(fittedtreeGTR$tree)
dt.fittedtreeGTR<-dt.fittedtreeGTR[nms, nms]
dt.fittedtreeGTR<-as.dist(dt.fittedtreeGTR)
plot(dt.fittedtreeGTR-d,ylab="residuals", cex=0.5,main="GTR")
```

Plot11: Residuals of the GTR method



The GTR model produces very little distortion, this allows it to be a valid model to be used in phylogenetic analysis.

Bootstrapping

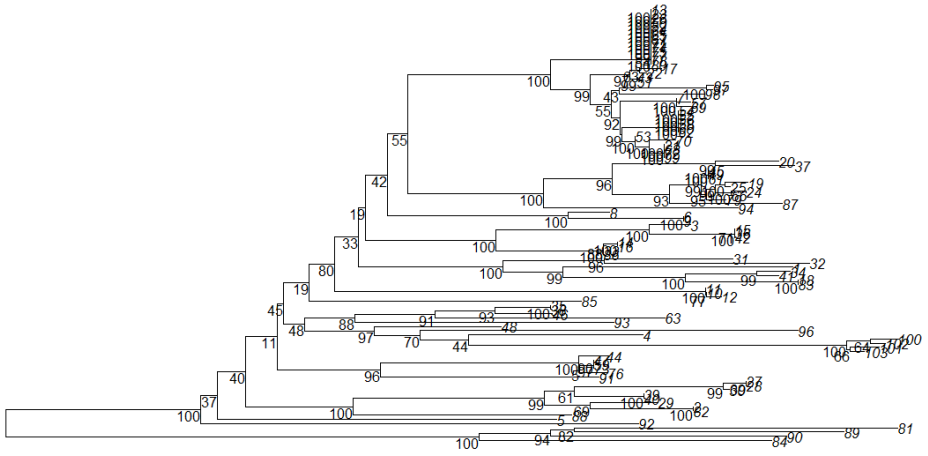
Bootstrapping can be used to evaluate the strength of the phylogeny produced from the data. Bootstrapping gives a value for the confidence in the phylogeny at each node.

We can use the GTR method to boostrap the data;

```
bs<-bootstrap.pml(fittedtreeGTR,bs=100,optNni=T)
```

```
plotBS(fittedtreeGTR$tree,bs)
```

Plot12: Bootstrapped tree using GTR model



The trees created using the BIONJ method can also be bootstrapped:

The data must be fitted to the tree:

```
fit<-pml(tr.Bionj,as.phyDat(x))
```

The fit is then optimised and a random seed set:

```
fit=optim.pml(fit,T)

set.seed(8)
```

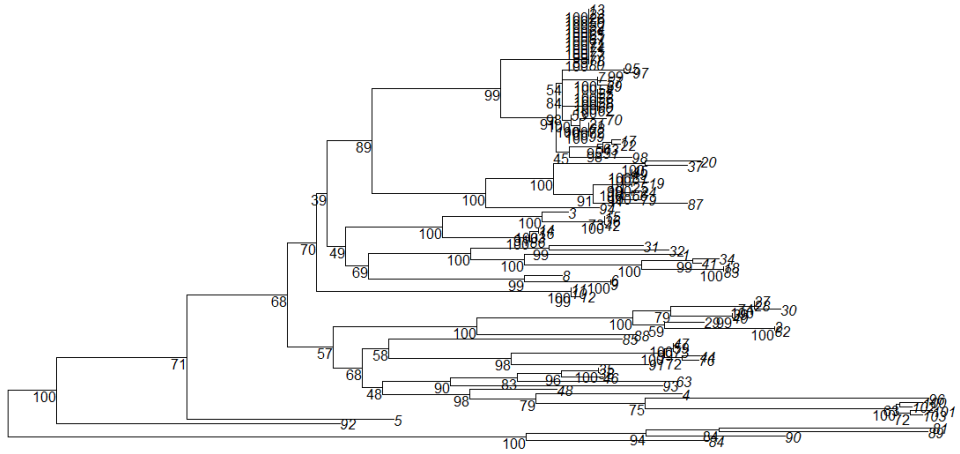
The data can then be bootstrapped ,100 times:

```
bs<-bootstrap.pml(fit,bs=100,optNni=T)
```

The boostrapped tree can then be plotted:

```
treeBS<-plotBS(fit$tree, type="p", bs)
```

Plot13: Bootstraped tree using BIONJ and K80 model



There are some differences in confidence levels at certain nodes , when comparing the two bootstrapped trees (plot 12 and 13).

Bayesian Trees

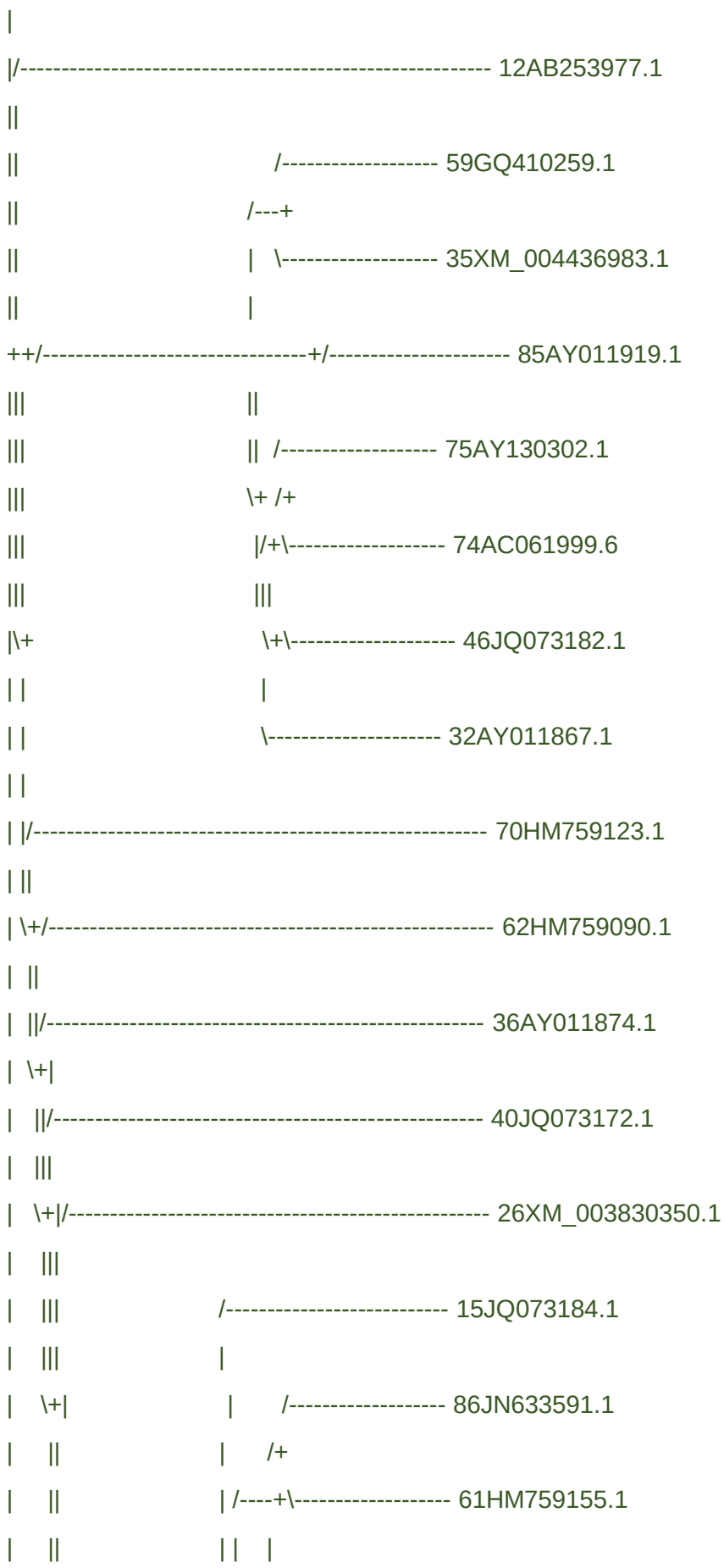
The Bayesian inference creates a posterior distribution for a parameter, based of the likelihood of data from the multiple alignment.(5).To produce a Bayesian inferred tree on the data the package MrBayes was used on the Xming machine (6). The fasta file had to be converted to a nexus file to run on MrBayes, this was done using the file converting website bugaco (7).

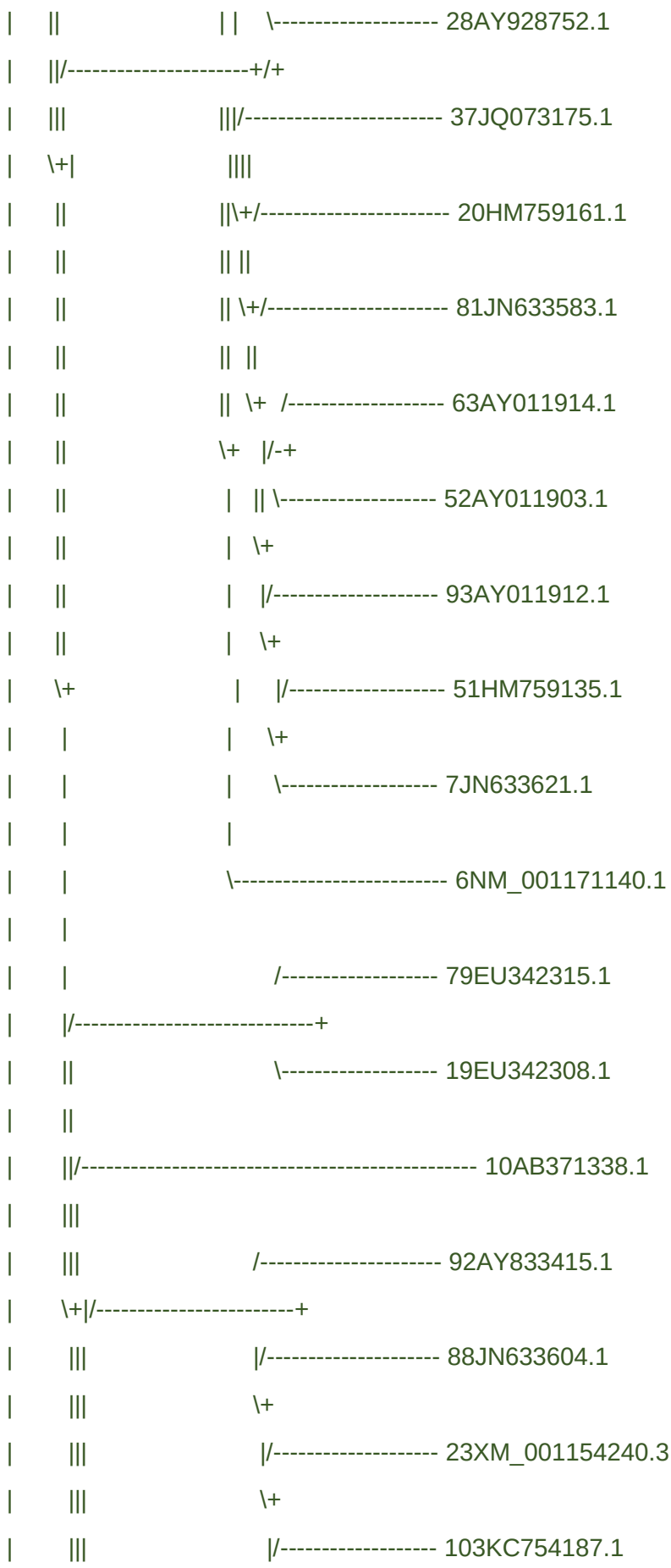
The following commands were then used in MrBayes to produces the trees:

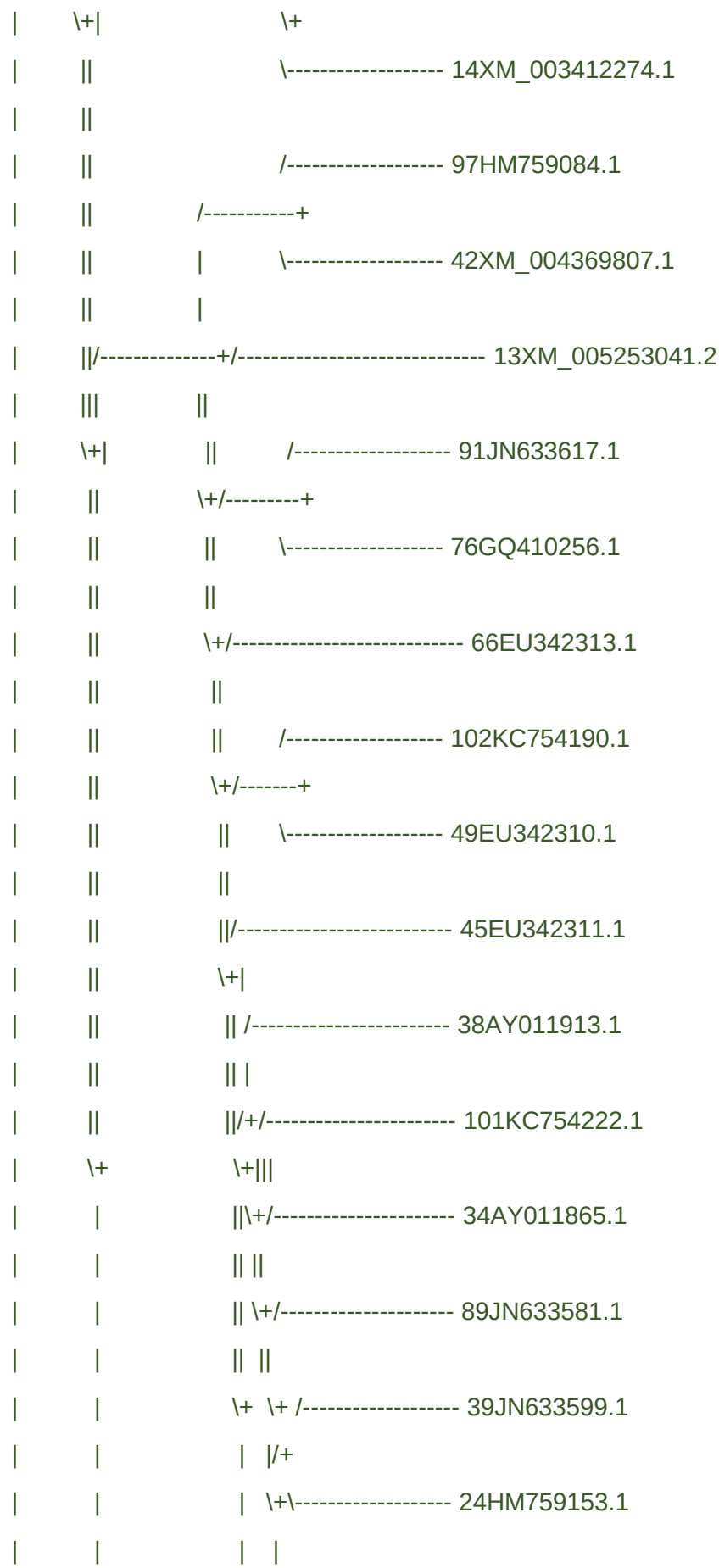
```
>execute marsup_rag1.2.nexus

>lset nst=6 rates=Invgamma

>mcmc ngen=20000 samplefreq=100 printfreq=100 diagnfreq=100 nchains=6 nruns=2

>sump

>sumpt
```

This was ran until the standard deviation as less than 0.05. The following tree was produced:

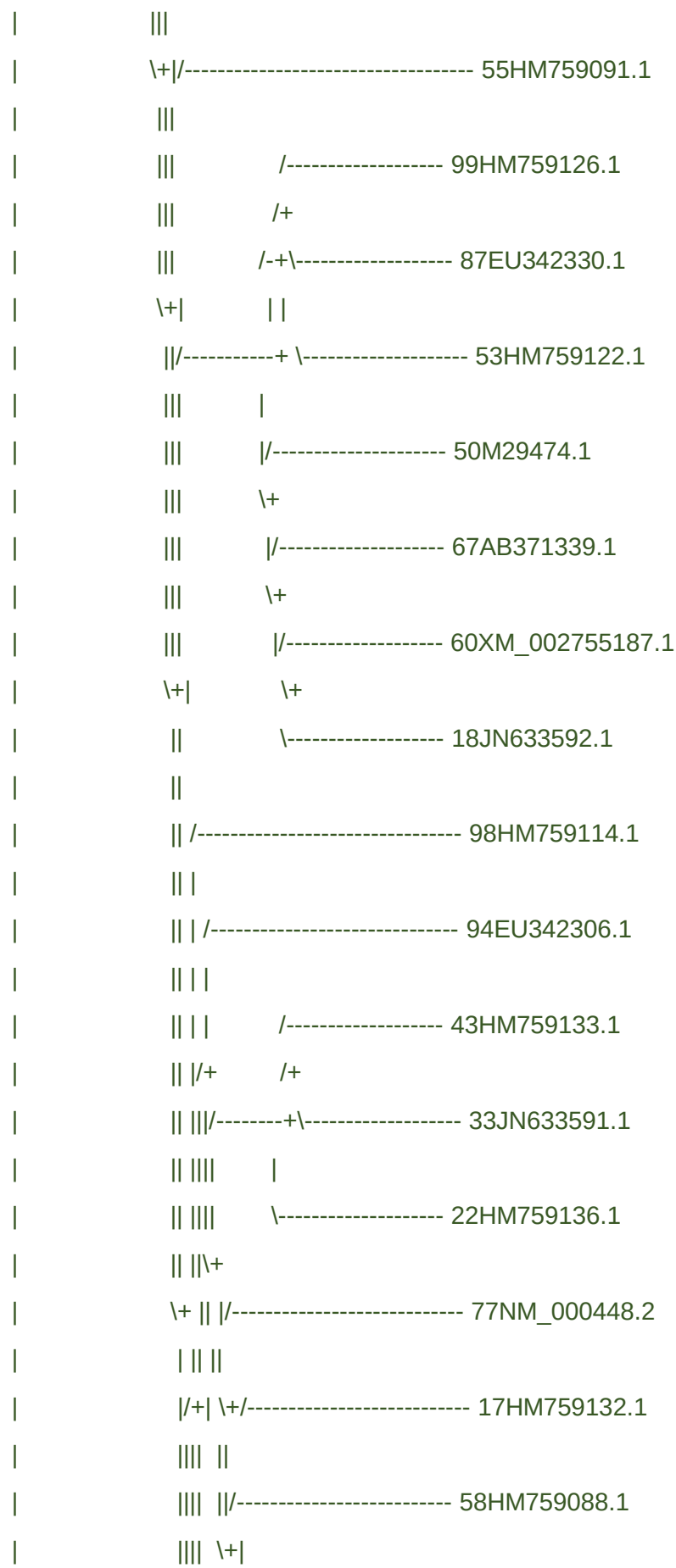
/---------- 72HM759128.1

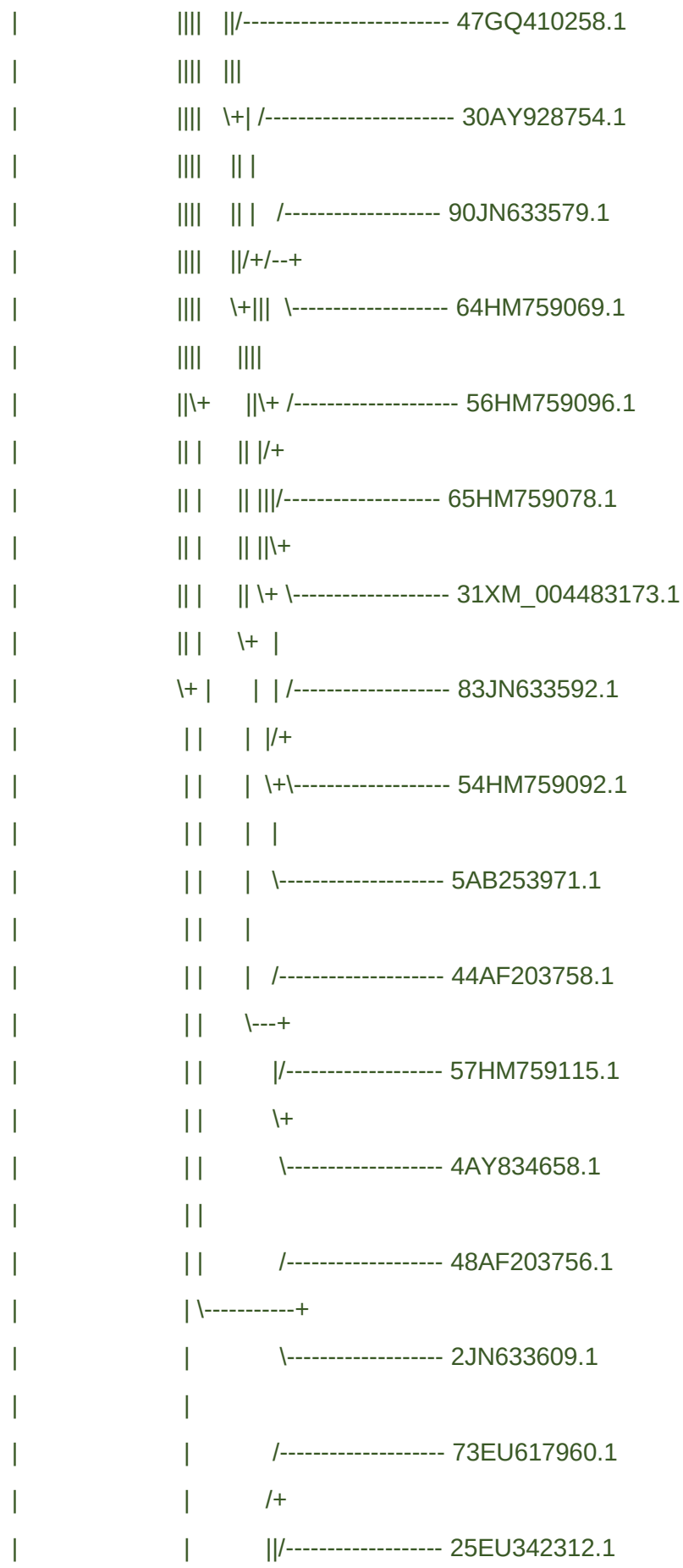

12AB253977.1
59GQ410259.1
35XM_004436983.1
85AY011919.1
75AY130302.1
74AC061999.6
46JQ073182.1
32AY011867.1
70HM759123.1
62HM759090.1
36AY011874.1
40JQ073172.1
26XM_003830350.1
15JQ073184.1
86JN633591.1
61HM759155.1


28AY928752.1
37JQ073175.1
20HM759161.1
81JN633583.1
63AY011914.1
52AY011903.1
93AY011912.1
51HM759135.1
7JN633621.1
6NM_001171140.1
79EU342315.1
19EU342308.1
10AB371338.1
92AY833415.1
88JN633604.1
23XM_001154240.3
103KC754187.1


14XM_003412274.1
97HM759084.1
42XM_004369807.1
13XM_005253041.2
91JN633617.1
76GQ410256.1
66EU342313.1
102KC754190.1
49EU342310.1
45EU342311.1
38AY011913.1
101KC754222.1
34AY011865.1
89JN633581.1
39JN633599.1
24HM759153.1


9M77666.1
69HM759117.1
8AY011895.1
68HM759126.1
96JX276328.1
78NG_007528.1
80BC037344.2
29AB109367.1
41AY011866.1
16AY011875.1
71HM759079.1
84AY011864.1
27JN633601.1
100KC754152.1
11AB253978.1
82JN633609.1
95HM759083.1


55HM759091.1
99HM759126.1
87EU342330.1
53HM759122.1
50M29474.1
67AB371339.1
60XM_002755187.1
18JN633592.1
98HM759114.1
94EU342306.1
43HM759133.1
33JN633591.1
22HM759136.1
77NM_000448.2
17HM759132.1
58HM759088.1


47GQ410258.1
30AY928754.1
90JN633579.1
64HM759069.1
56HM759096.1
65HM759078.1
31XM_004483173.1
83JN633592.1
54HM759092.1
5AB253971.1
44AF203758.1
57HM759115.1
4AY834658.1
48AF203756.1
2JN633609.1
73EU617960.1
25EU342312.1

```
|                    \----------+\+
|                               | \------------------ 21HM759124.1
|                               |
|                               \-------------------- 3JN633590.1
|
\------------------------------------------------------------ 1JN633593.1
```

The Bayes inferred tree that was created, shows different grouping of phylogeny than those seen in the trees produced using other methods.

ED Scores

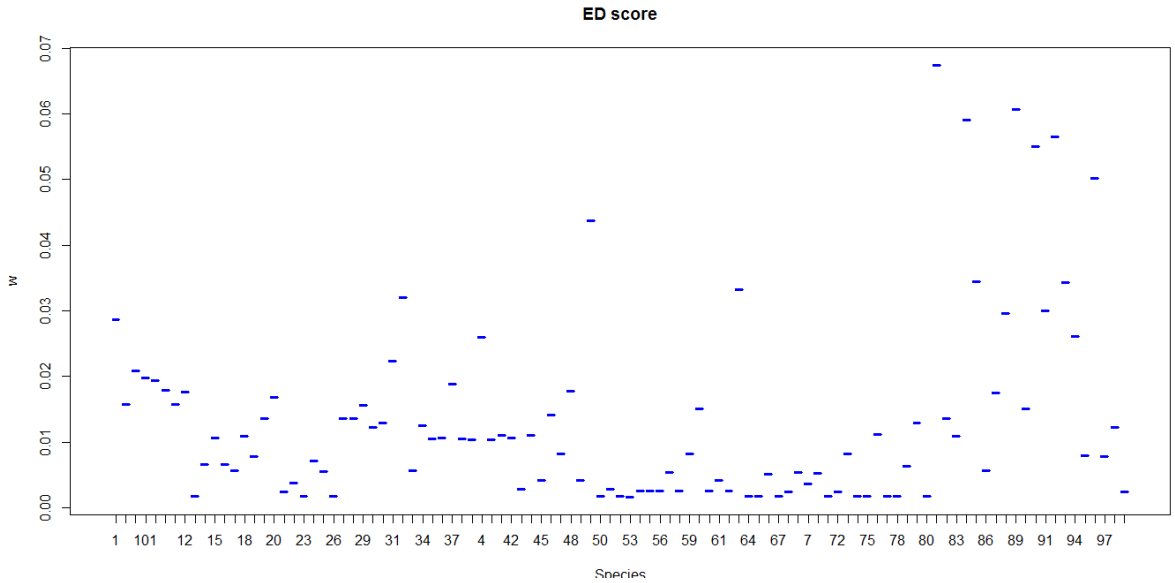
Evolutionary distinctiveness scores are useful in phylogenetic analysis. ED scores can be calculated using the evol.distinct command in R studio:

```
orig<-evol.distinct(tr.Bionjr,type="fair.proportion")
orig
```

This produces a table ,of the taxa and their respective ED score, which can then be plotted :

```
plot(orig, col=" blue", border=" blue" ,main="ED score", density=100)
```

Plot 14: Plot of ED Scores:

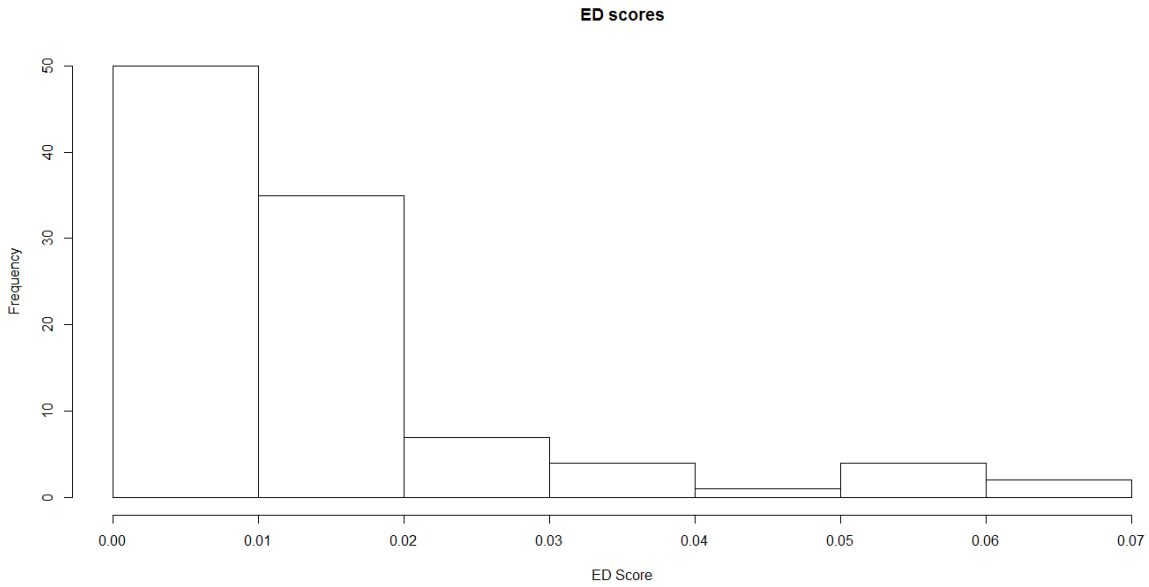


The ED score indicates how distinct a species is from the other species in the phylogeny. Some species are more distinct tan others, these distinct species generally represent a large amount of evolution that is unique to them. (8)

A histogram of ED scores can also be plotted to see the general trend of the data:

```
hist(orig$w, main='ED scores', xlab='ED Score')
```

Graph 1: Histogram of ED scores:



The above histogram shows that the majority of the taxa have very low ED scores.

The lower the ED score the less unique the species is in the phylogenetic tree. In this case all of the ED scores are very low, this indicates that most of the taxa in this phylogeny are not distinct, but are related and evolved similarly. Comparisons can be made between the ED scores and the bootstrapped data trees. If we look at the GTR bootstrapped tree, plot 12 , species '89' (JN633617.1/1-602), which is from a tubed nose fruit bat, is more distinct than other species, this is then supported by the fruit bat having a larger ED score than other species in the phylogeny. (9) Whereas species '13' (XM_005253041.2/1-602), which is a human gene, has a very low ED score, and is positioned very close to other species in the phylogenetic tree. (10) This suggests that the human is more closely related to other species in the phylogeny than the fruit bat.

The Bayes inferred tree suggest a close link between the Taphozous, a sac winged bat (species 48: AF203756) and the Small spotted Genet (species 2: JN633609.1). (11, 12) However the bootstrapped trees (plots 11 and 12) do not indicate the same level of similarity between the genes. When looking at the ED scores for the species, the Taphozous, has a high ED score compared to the rest of the phylogeny, while the genet has a very low ED score. This suggest that the Taphozous is more distinct, in terms of evolutionary background, than the Genet, however there rag1 gene sequences are similar. There could be a possible common ancestor in this case.

The table of species and there ED scores can also be visualised, however it is more difficult to identify trends using the table.

Table 2: ED scores

	Species	w
1	1	0.028702816
2	2	0.013687635
3	3	0.012316966
4	4	0.025950131
5	5	0.043752320
6	6	0.015032060
7	7	0.003623183
8	8	0.012969045
9	9	0.015032060
10	10	0.015720613
11	11	0.015768358
12	12	0.017655967
13	13	0.001780132
14	14	0.006551513
15	15	0.010647708
16	16	0.006551513
17	17	0.005626027
18	18	0.010844931
19	19	0.007752403
20	20	0.016860873
21	21	0.002443749
22	22	0.003814725
23	23	0.001780132
24	24	0.007105937
25	25	0.005589995
26	26	0.001780132
27	27	0.013646222
28	28	0.013646222
29	29	0.015644292
30	30	0.012953752
31	31	0.022388164
32	32	0.031969212
33	33	0.005659539
34	34	0.012457274
35	35	0.010481212
36	36	0.010647708
37	37	0.018871295
38	38	0.010481212
39	39	0.010485156
40	40	0.010485156
41	41	0.011045088

	Species	w
42	42	0.010647708
43	43	0.002782746
44	44	0.011047076
45	45	0.004178885
46	46	0.014153093
47	47	0.008257697
48	48	0.017832429
49	49	0.004178885
50	50	0.001780132
51	51	0.002782746
52	52	0.001780132
53	53	0.001624220
54	54	0.002519640
55	55	0.002519640
56	56	0.002519640
57	57	0.005395798
58	58	0.002519640
59	59	0.008257697
60	60	0.002519640
61	61	0.004178885
62	62	0.002519640
63	63	0.033200582
64	64	0.001780132
65	65	0.001780132
66	66	0.005188416
67	67	0.001780132
68	68	0.002443749
69	69	0.005404035
70	70	0.005308449
71	71	0.001780132
72	72	0.002443749
73	73	0.008257697
74	74	0.001780132
75	75	0.001780132
76	76	0.011207828
77	77	0.001780132
78	78	0.001780132
79	79	0.006396117
80	80	0.001780132
81	81	0.067435844
82	82	0.013687635

	Species	w
83	83	0.010844931
84	84	0.059068785
85	85	0.034397779
86	86	0.005668158
87	87	0.017446206
88	88	0.029628217
89	89	0.060730813
90	90	0.055079870
91	91	0.029957398
92	92	0.056473127
93	93	0.034345909
94	94	0.026043382
95	95	0.007951756
96	96	0.050136062
97	97	0.007815531
98	98	0.012208263
99	99	0.002443749
100	100	0.020818519
101	101	0.019724764
102	102	0.019324871
103	103	0.017903541

Note:

The complete script and *.tre files, created using MrBayes, used to complete this assignment can be found at the following Github address:

https://github.com/fmacfarlane/Assignment_4.git

References

1. Jones, Jessica M.; Gellert, Martin (2004). "The taming of a transposon: V(D)J recombination and the immune system". *Immunological Reviews* **200**: 233–48.
2. http://www.atgc-montpellier.fr/Bionj/
3. Akaike, Hirotugu (1974), "A new look at the statistical model identification", *IEEE Transactions on Automatic Control* **19** (6): 716–723
4. Tavaré S. "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences". *Lectures on Mathematics in the Life Sciences* (American Mathematical Society)**17**: 57–86.)
5. Geyer, C.J. (1991). "Markov chain Monte Carlo maximum likelihood". In Keramidas, E.M. *Computing Science and Statistics: Proceedings of the 23rd Symposium of the Interface*. Fairfax Station VA: Interface Foundation. pp. 156–163.
6. http://mrbayes.sourceforge.net/
7. http://sequenceconversion.bugaco.com/converter/biology/sequences/
8. http://www.edgeofexistence.org/about/edge_science.php
9. http://www.uniprot.org/uniprot/G3M7M7
10. http://www.ncbi.nlm.nih.gov/nuccore/578820674
11. http://www.ncbi.nlm.nih.gov/nuccore/6694640
12. http://www.uniprot.org/uniprot/G3M7L9