Rucsanda Juncu
r0872274
Vijgen et al., 2005
Assignment 1

Introduction:

This paper is an evaluation of the evolution rate of Human Respiratory Coronavirus OC43 reported by St-Jean *et al.* (2004). In that paper, sequences from a sample in 1967 and 2001 were shown to be unusually similar. Another paper by Vijgen *et al.* (2005) evaluated this claim, showing the statistical improbability of such a low evolution rate. Statistical analysis of sequence phylogeny is used in this paper to evaluate the hypothesis that the two samples had a substitution rate much lower than those reported in other coronaviruses.

1)

a) Sequences were aligned in TEMPest using a muscle algorithm. Alignments were used to create a BioNJ tree (Figure 1), with evolution being modeled using an HKY model. Both the bovine and human samples formed monophyletic clusters. The bovine samples all shared the same common ancestor without sharing it with a member from another taxa. Similarly, all human samples are connected to the same ancestor and did not share it with any non-human sample. The root was calculated by finding the midpoint between the two most diverged samples, and represents a hypothetical common ancestor of both taxa.

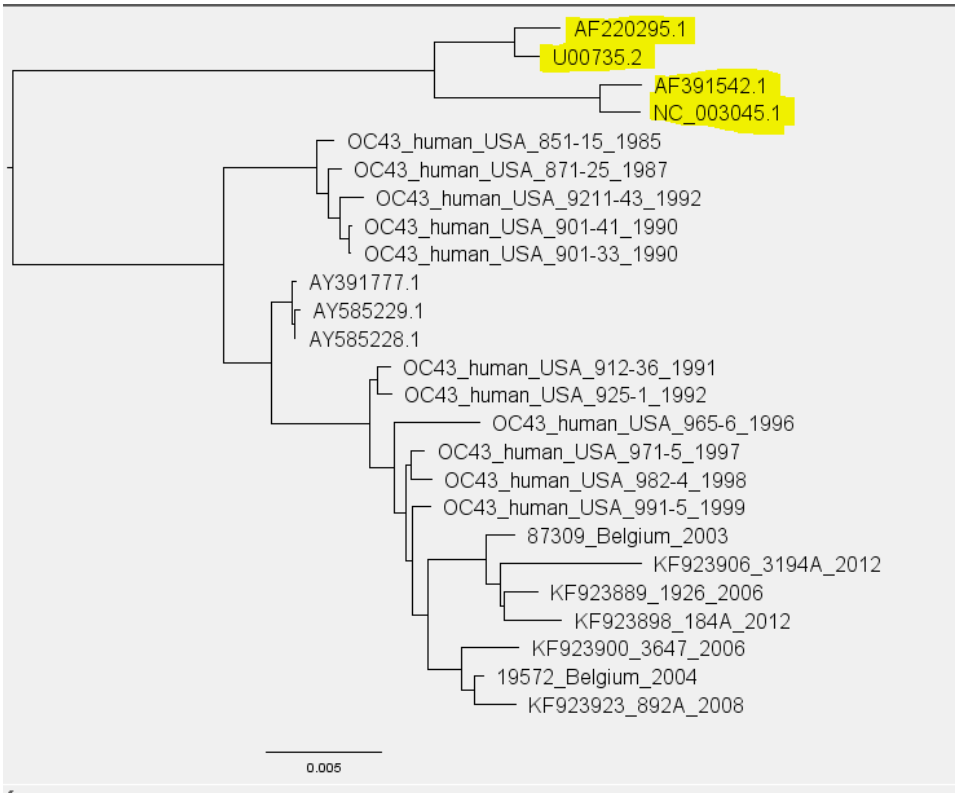

Figure 1: Phylogenetic tree made of the original 18 samples, 2 test samples, 4 bovine samples, and 1 sample imported from Vijver et al. data. Bovine samples are highlighted in yellow. The tree was created in TEMPest and viewed in FigTree.

Rucsanda Juncu
r0872274
Vijgen et al., 2005
Assignment 1

b) There is general evidence to support the clusters generated by the tree. Samples with similar IDs clustered together, indicating that there was similarity between replicates within experiments. This makes sense as those viral replicates should have similar genetic backgrounds. Additionally, clusters generally formed by year, with older 1990's samples clustering further away from 2000's samples. There is also evidence to support the similarity between the 1967 AY58228 virus and the 2001 AY58229 since they clustered together and exclusive of other samples.

c) In order to find the sequence most similar to OC43_human_USA_965-6_1996, the tree was labeled with branch times up to four significant digits. The units of the times were irrelevant since only the distance with the smallest value was needed to know the closest sequence.

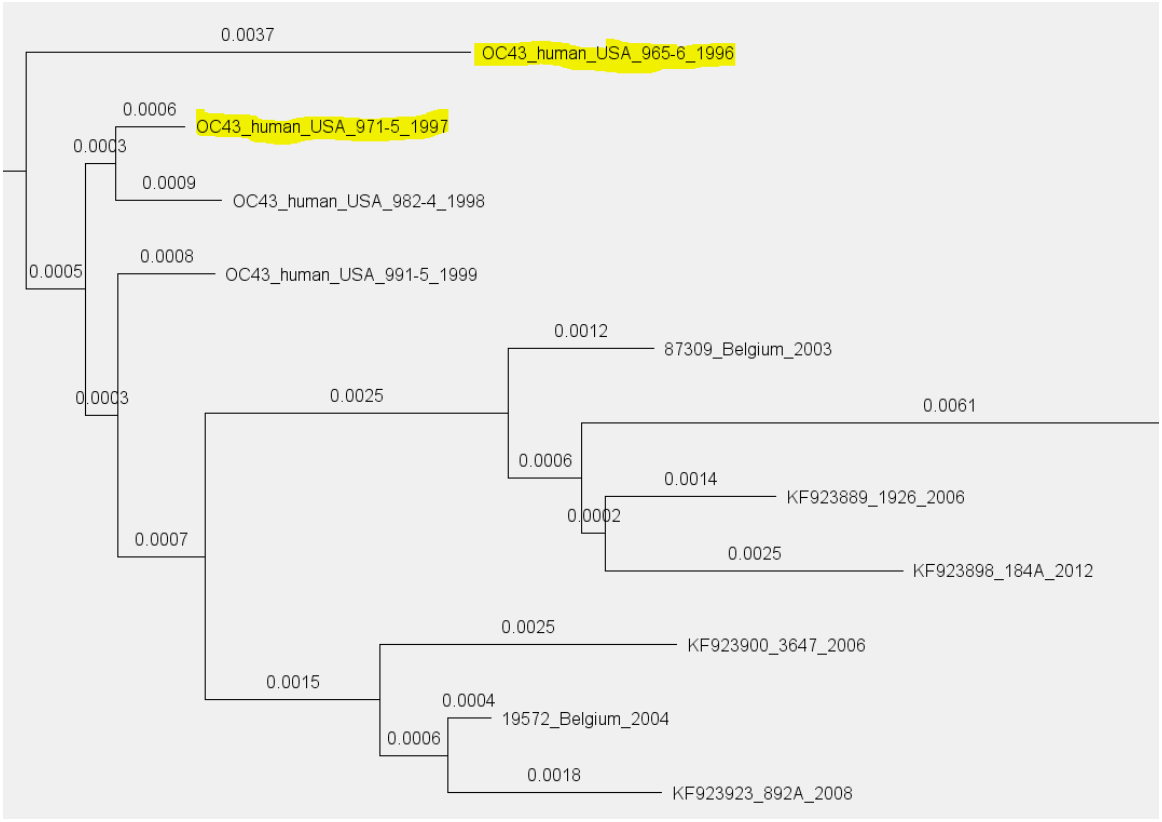

Figure 2: Segment of Figure 1 with distances displayed.

Patristic distances:

i) OC43_human_USA_971-5_1997: 0.0006+0.0003+0.0005+0.0037 = 0.0051

ii) OC43_human_USA_982-4_1998: 0.0009+0.0003+0.0005+0.0037 = 0.0054

iii) OC43_human_USA_991-5_1999: 0.0008+0.0003+0.0005+0.0037 = 0.0053

The sample with the smallest patristic distance is OC43_human_USA_971-5_1997, making it the most closely related.

d) Genetic diversity was calculated by summing the distances for the sequences being analyzed, and dividing by the total number of sequences. Using the same distances from the previous question, the genetic diversity calculations are shown below.
   i) Cluster 1: OC43_human_USA_982-4_1998, OC43_human_USA_971-5_1997, and C43_human_USA_991-5_1999 (0.0006+0.0009+0.0003+0.0003+0.0008)/3=0.000967
   ii) Cluster 2: KF923898_184A_2012, 87309_Belgium_2003, and KF923889_1926_2006 (0.0025+0.0002+0.0014+0.0006+0.0012)/3=0.001967
   iii) Cluster 2 was the most genetically diverse.

e) The sample OC43_ATCC_VR759_1967 was relabelled to AY391777.1. The sample most closely related to it (the sample with the smallest patristic distance) was AY585228.1

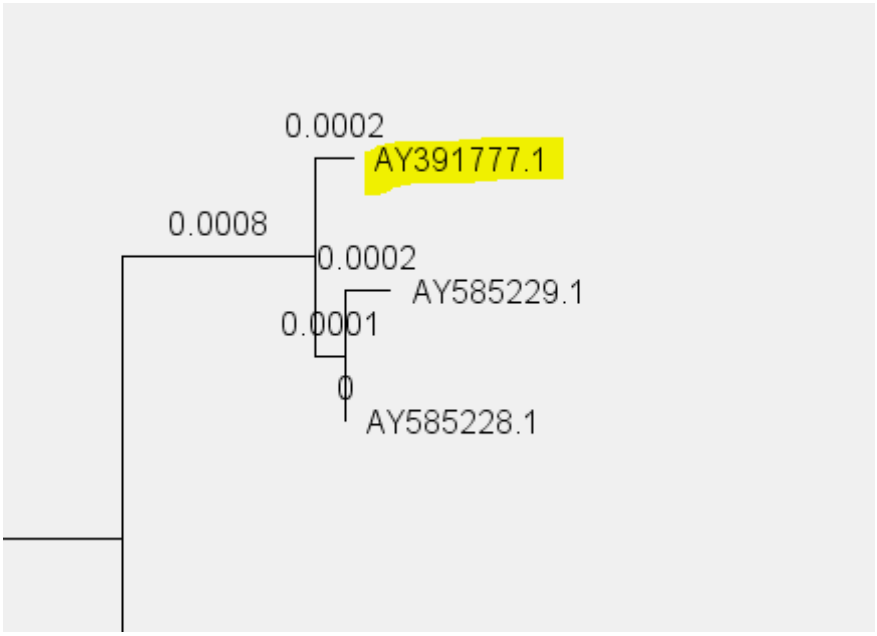

Figure 3: Segment of Figure 1 with distances displayed.

2)

a) Calendar dates were parsed in TempEST only by order and with variable precision. Figure 4 shows a root to tip diagram of the midpoint rooted tree from the previous analysis, the X axis representing time in years and the Y axis representing root-to-tip divergence. The four functions provided by TempEST: heuristic residual mean squared (HRMS), residual mean squared (RMS), correlation, and R squared, provided multiple models for comparison. The functions RMS and HRMS had the same output, as did correlation and R squared. R squared and correlation had slightly higher values for all statistics compared to HRMS and RMS, which are listed below. Reasons for this are unclear, but both results are within the same order of magnitude and are similar in value. The dates ranged over 45 years, and a R squared was chosen to calculate a strict clock model for divergence accumulation. The slope was 2.9473E-4 mutations/genomic

sites/year, with an X intercept (tMRCA being the most common ancestor) occurring in 1947. The correlation coefficient showed strong positive correlation, and R^2 indicated that the model explained about half of the variation in the data. The strong correlation indicates a positive relationship between the genetic divergence of the samples from the root, providing evidence to support the presence of a significant temporal signal.

Statistic	RMS/HRMS	R squared/Correlation
Slope (rate)	2.5999E-4	2.9473E-4
X-Intercept (tMRCA)	1943.7783	1946.5699
Correlation Coefficient	0.75	0.7701
R squared	0.5624	0.593
Residual Mean Squared	7.886E-6	8.8293E-6

Table 1: Comparison of statistics given by different by the four TempEST algorithms for strict clock models.

Rucsanda Juncu
r0872274
Vijgen et al., 2005
Assignment 1

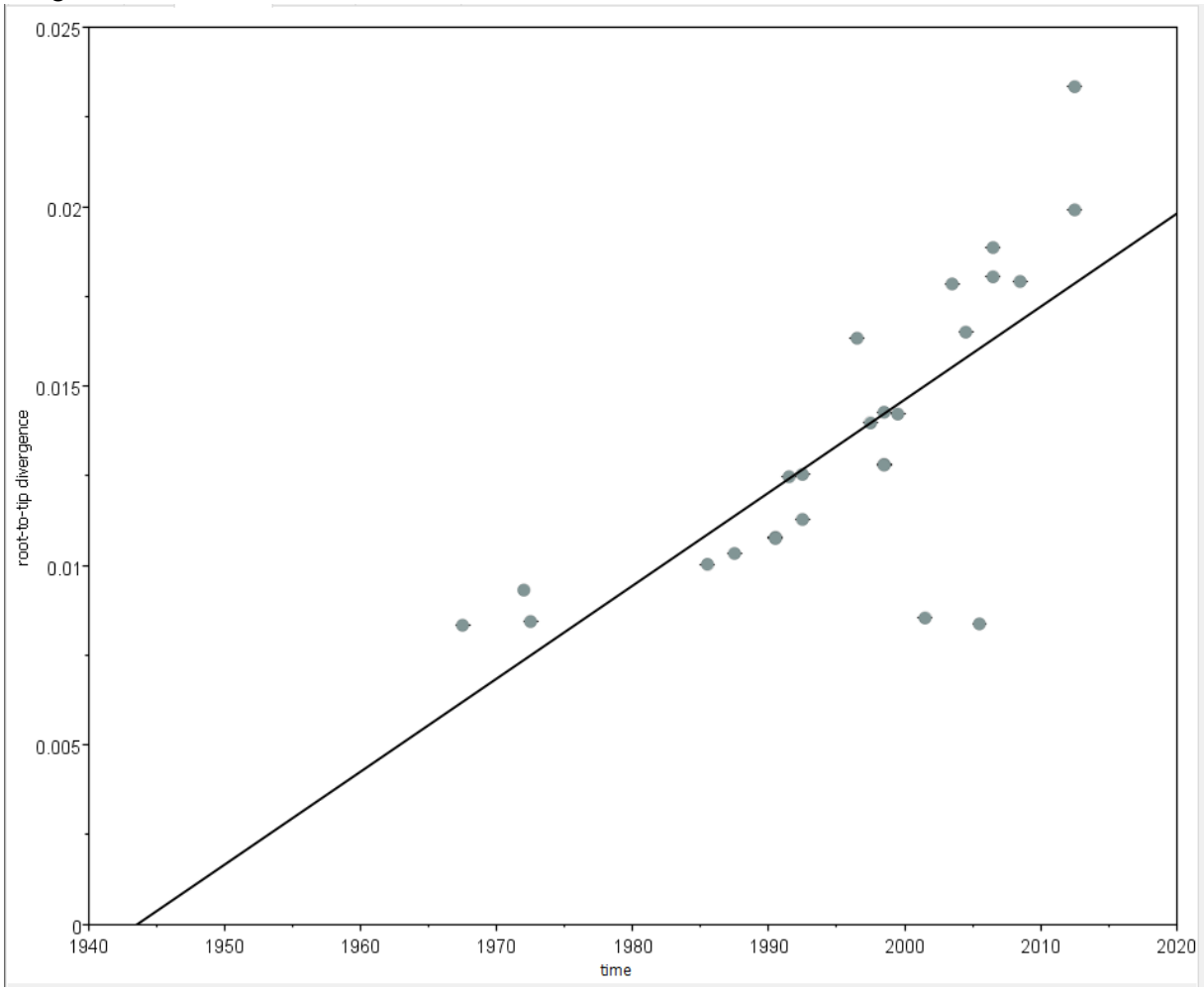

Figure 4: Root to tip diagram of samples present in Figure 1.

b) Residuals shown in Figure 5 do not appear to show any obvious non random clustering or trends, implying that the data is well explained by the strict clock model in Figure 4. Some data points in the early 2000s show a similar amount of divergence as those from the 1960s in Figure 4, implying a possible relationship between the points or divergence from the clock model.

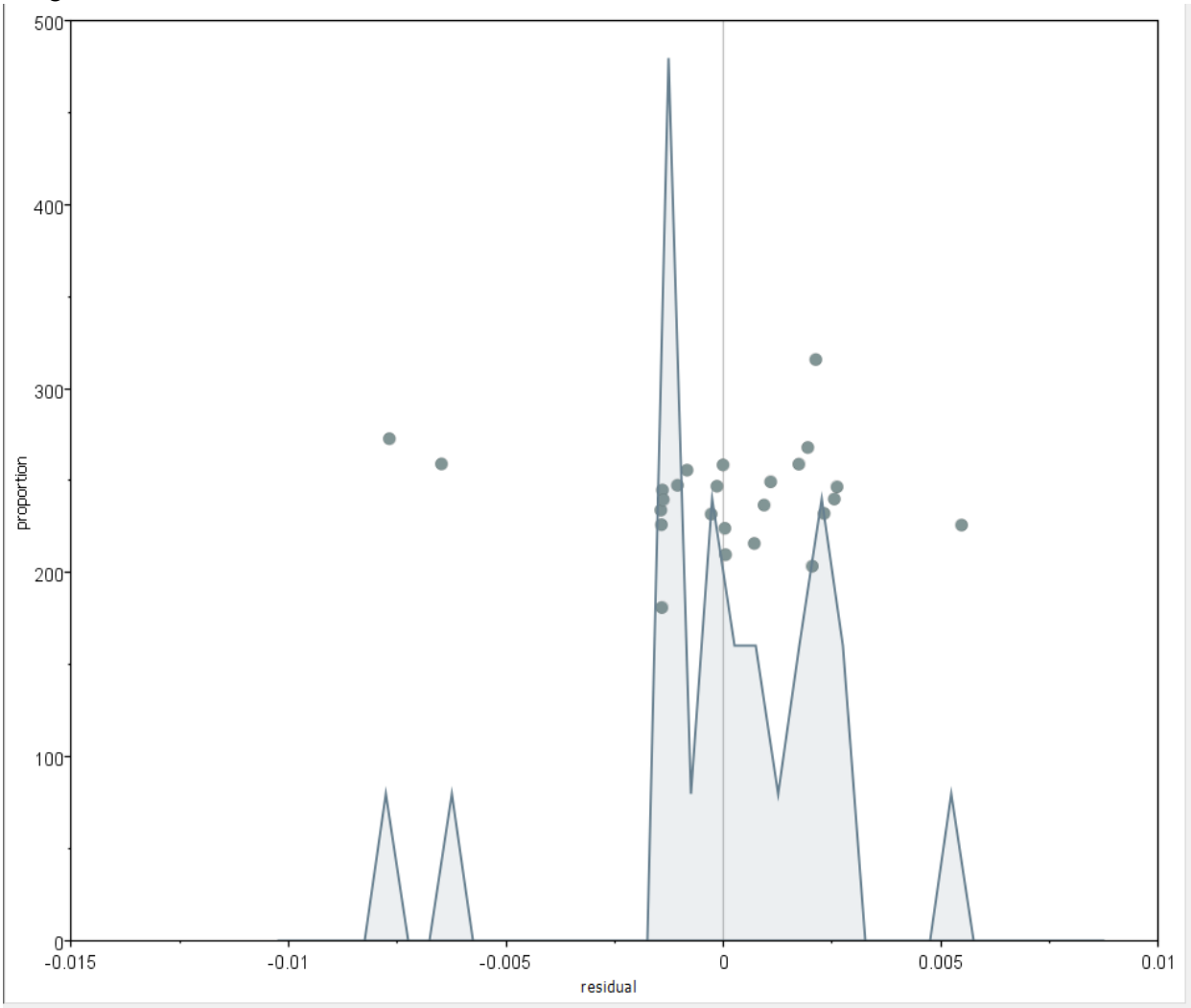

Figure 5: Residuals associated with the model in Figure 4.

c) The evolutionary rate is 2.9473E-4 mutations/genomic sites/year.

3)

a) BEAUti and BEAST were used to generate an HKY model with 10,000,000 MCMC (Markov chain Monte Carlo) chain steps. Only one taxa was used, the Paris 2001 isolate, in order to estimate an age without explicitly assigning one to the sequence. Trace plot quality was evaluated using ESS (effective sample size), which gave the number of independent samples in the trace. ESS was calculated as the proportion of the chain length to the ACT (auto-correlation time). ACT measured the number of states for which two points on the MCMC remained correlated. ESS values in Table 2 were well above 100, indicating stationarity in the trace plot. The population mean for the Paris taxa was 1966.4908, meaning the average age estimated for that taxa was before 2001. A histogram plot of the Paris samples is shown in Figure 6. An estimate for the probability that the Paris sample is younger than 1970 is less than 5%.

Rucsanda Juncu
r0872274
Vijgen et al., 2005
Assignment 1

Summary Statistics	age(Paris)	age(Root)	clock.rate
mean	1966.4908	1825.4115	1.3433E-4
stderr of mean	0.0374	0.5483	4.381E-7
stdev	0.9457	11.4882	8.9101E-6
variance	0.8943	131.9778	7.939E-11
median	1966.5359	1826.0894	1.3426E-4
value range	[1963.1307, 1971.2523]	[1764.1356, 1857.8741]	[9.8606E-5, 1.6707E-4]
geometric mean	1966.4905	1825.3753	1.3403E-4
95% HPD interval	[1964.5189, 1968.215]	[1803.117, 1847.8834]	[1.1675E-4, 1.5132E-4]
auto-correlation time (ACT)	14104.4966	20506.7819	21762.6415
effective sample size (ESS)	638.2	438.9	413.6
number of samples	9001	9001	9001

Table 2: Summary Statistics for Paris and Root Taxa. age(Paris) represents the model to estimate the Paris taxa age, age(Root) represents the remaining 24 samples, and clock.rate shows statistics for the estimated substitution rate for all 25 sequences.

Rucsanda Juncu
r0872274
Vijgen et al., 2005
Assignment 1

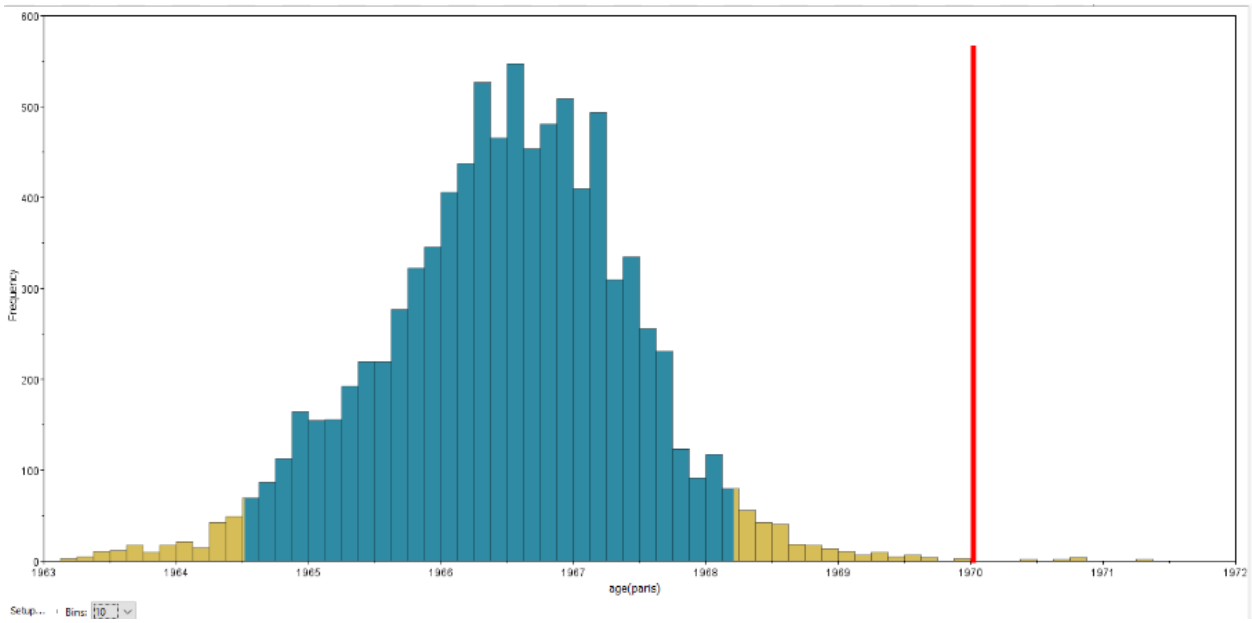

Figure 6 age(Paris) Histogram with 10 bins: A histogram of the sample ages is shown above. The blue section represents the HPD (highest posterior density) interval, which contains 95% of the data. The yellow tails on either side represent statistically significant regions of the remaining 5%. The red line is marked at 1970, showing that less than 2.5% of the data occurs more recently than 1970.

b) Table 2 contains the statistical information for the estimated rate of substitution for all aligned sequences. The mean (1.3433E-4 substitutions/site/year) is also visible in Figure 7, along with the credible intervals. Table 3 contains a list of molecular clock estimates from this research, Vijgen *et al*. (2005), and St. Jean *et al.* (2001). The rates from this analysis are within the same order of magnitude as the ones in Vijgen's paper, while St. Jean's rates show a difference of 10E-2. Despite the rates generated in this paper being smaller than most of the rates in Vijgen's paper, there is no evidence to suggest they would be anywhere near St. Jean's reported rate.

Calculation Source	Evolution Estimate (substitutions/site/year)
BEAST analysis (clock.rate)	1.3433E-4
R squared/Correlation	2.9473E-4
RMS/HRMS	2.5999E-4

SARS (Salemi, M., *et al.*, 2004)	4.0E-4
CoV (Sanchez, C., *et al*., 1992)	7.5E-4
ML estimate with Rhino (Rambaut, A., *et al*., 2000)	1.54E-4
St. Jean Lab Estimate (2001)	5.7E-6

Table 3: Molecular Clock Models. The first three estimates come from this paper, created either using BEAST/BEAUti or TempEST. The next three are rates mentioned in Vijgen et al. (2005) as comparisons to St. Jean's reported rate (final row).

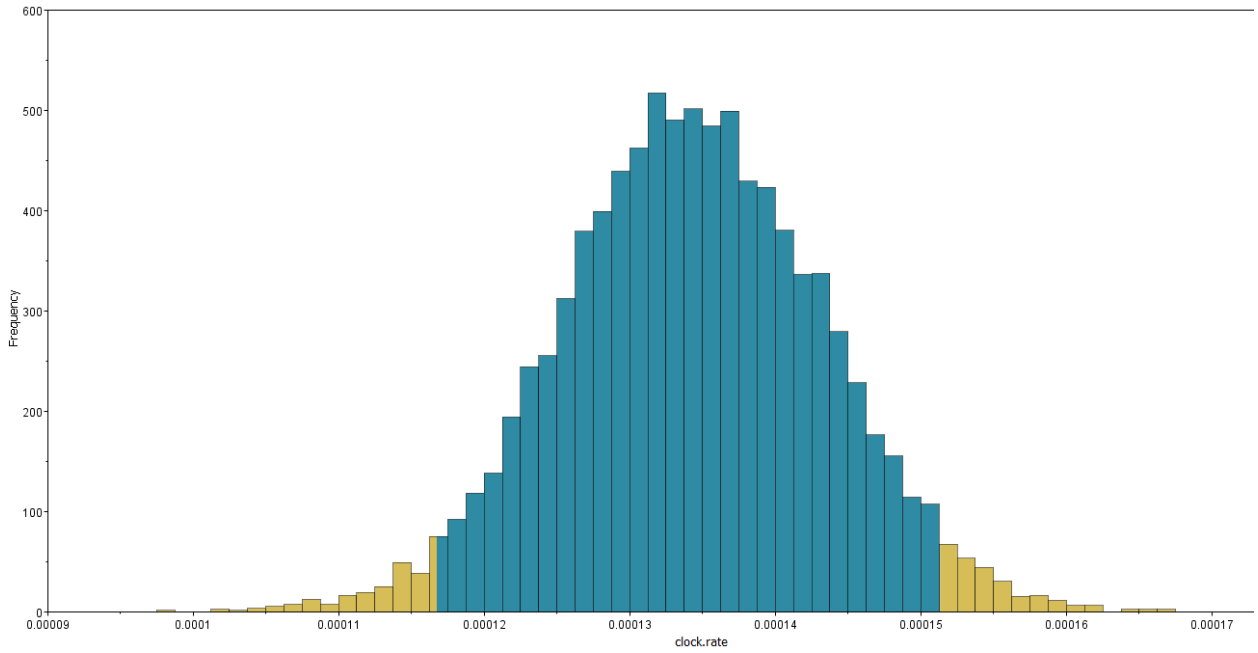

Figure 7 clock.rate Histogram with 10 bins. Plot of distribution of sampled rate of substitution/site/year for all sequences.

c) The .trees file generated by BEAST was viewed in TreeAnnotator and compared to the BioNJ tree generated by TEMPest in earlier steps (Figure 8). Both trees maintained the midpoint position between the samples from 1972-1998 and the rest of the samples. The clusters from 1985 to 1992 seem mostly the same except for OC43_human_USA_912-36_1991 and OC43_human_USA_925-1_1992 having five branching events in the original tree and four in the TreeAnotator tree. AY391777.1_2005 has a longer distance between the two other samples in its cluster, and is also more closely related to AY585229.1_2001 in the new tree. KF923889_184A_2012 also branches off earlier than 87309_Belgium_2003 in the new tree. The trees have different scales, with the TreeAnotator tree

illustrating a longer passage of time from the tMRCA than BioNJ. This may be due to BEAST estimating tMRCA to be in the 1800s, while the BioNJ tree only had a midpoint calculation based on the largest distance between sequence clusters.

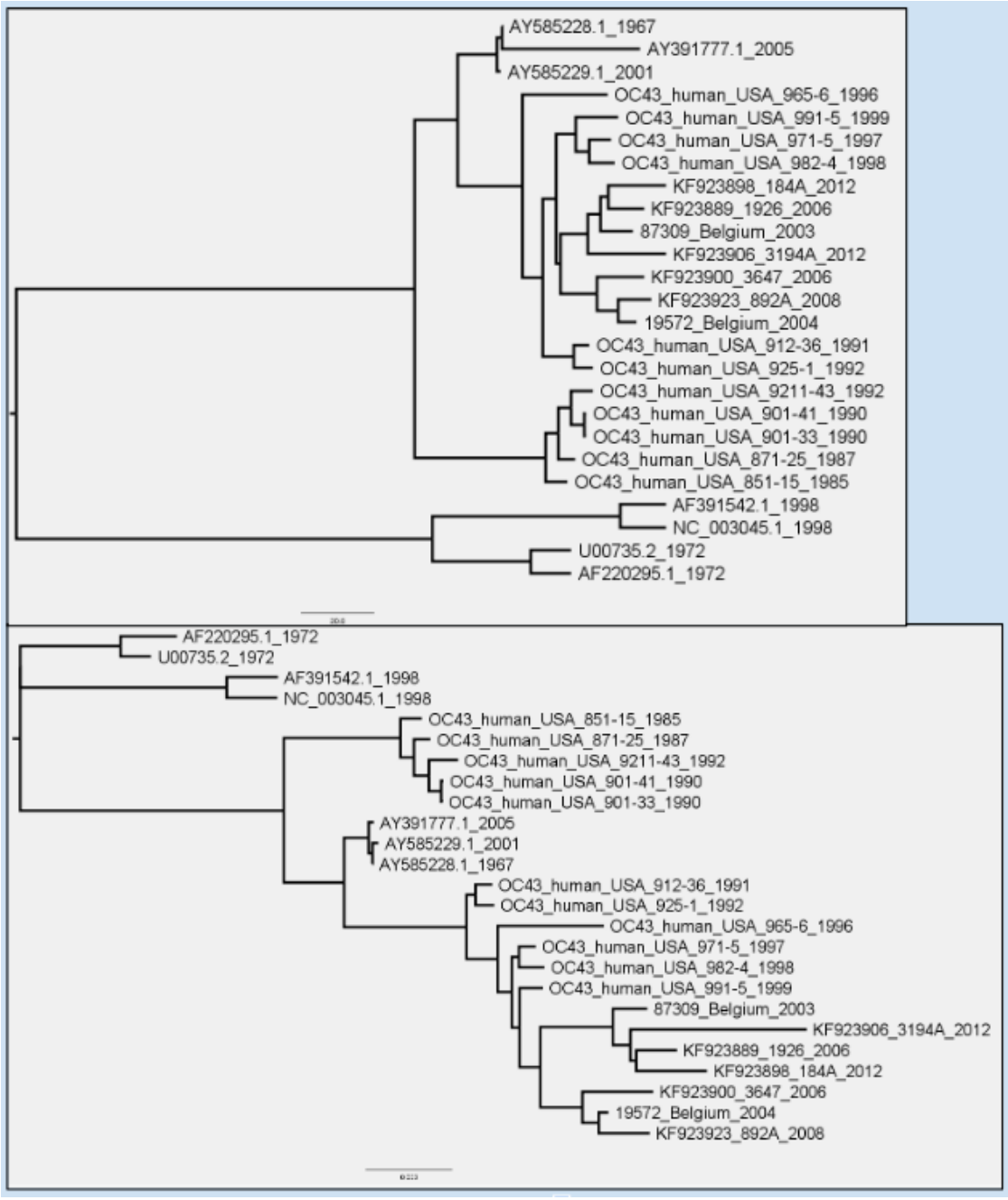

Figure 8 Tree comparisons. Top: Tree developed by running TreeAnnotator on the .trees file generated by BEAST. The scale at the bottom of the image is 20.0. Bottom: Original BioNJ tree made by TEMPest. The scale is 0.003.

Rucsanda Juncu
r0872274
Vijgen et al., 2005
Assignment 1

4)

a) The BEAST data was generated using a discrete gamma distribution with four rate categories (one for every base). Doing so accommodated for rate heterogeneity between base substitutions, meaning substitutions were not assumed to be independent and identically distributed. This was paired with an HKY model to generate distributions for α and Pinv values, α being the main shaping value for the gamma distribution. α values less than one produce L shaped distributions with high rate heterogeneity, while α values greater than one create bell shaped distributions have lower rate heterogeneity. The α and Pinv distributions are shown in Figure 9 after being smoothing with a kernel density estimator. Figure 9 also shows the HPD for the Pinv distribution, spanning from 0.6420 to 0.7985. The distribution skewed right, with a mean of 0.7255. This insinuates that the majority of sequences (72.55%) should be invariant, with a rate heterogeneity of zero. The remaining rate variation on other sites indicating sites of potential mutation increase.

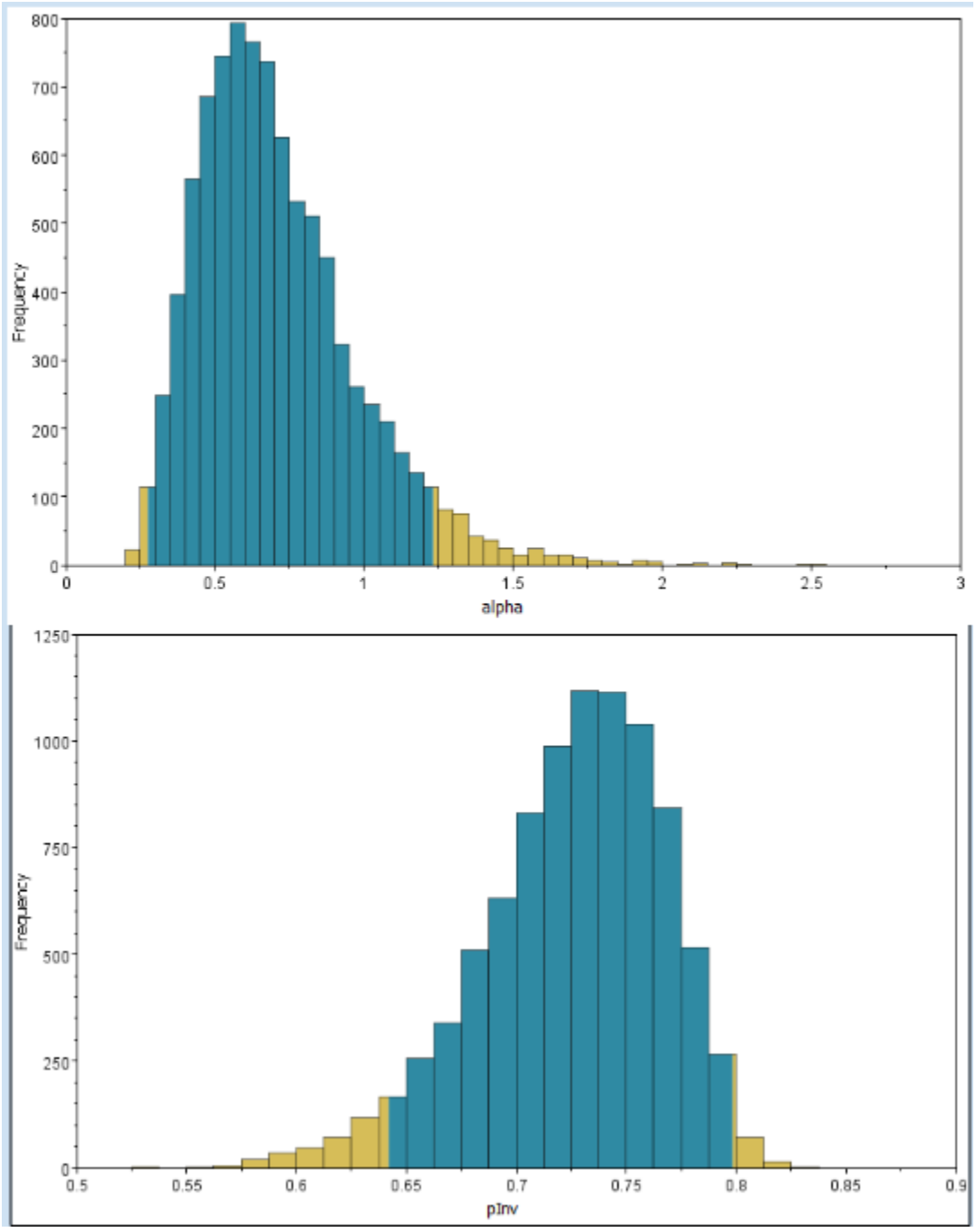

Figure 9: Top: Histogram of α values with HPD. Note that α values have a mean less than one, indicating stronger heterogeneity, but with a considerable tail to the right. Bottom: Pinv histogram with HPD. Pinv values show around 70% of sites are invariant.

5)

a) BEAST was run with no specified taxa in BEAUti in order to compare the previous model with one having all dates explicitly specified. The mean estimated substitution rate of the new model was 1.334E-4 substitutions/site/year, and the

mean tMRCA was in 1824.273. Comparing the substitution rate with the previous model (1.3433E-4 substitutions/site/year) shows similarity up to 1 significant digit. The mean tMRCA (1825.412) are also similar, having a difference of one year. It seems that specifying the 2001 date had little difference from allowing the model to estimate the model statistics using evolutionary distance between samples. Additionally, despite different ordering of taxa, the topology of both the unspecified and specified trees appear to be identical.

Conclusion:

Several pieces of evidence gathered in this analysis support the hypothesis that the evolutionary rate from St. Jean's paper is due to some technical error or is otherwise unreputable. The estimated mean substitution rate being much larger than the St. Jean's estimate, the strong indication of the 2001 sample having an age before 1970, and the predicted ~25% site heterogeneity with very little mutations actually occurring, all indicate that there is little scientific evidence to support the claim that the viruses are 30 years apart. There should be further investigation and attempts to replicate the research of the St. Jean's paper in order for their research to be considered scientifically sound.

Rucsanda Juncu
r0872274
Vijgen et al., 2005
Assignment 1

Resources:

- Bryant JE, Holmes EC and Barrett ADT (2007) Out of Africa: A Molecular Perspective on the Introduction of Yellow Fever Virus into the Americas. *PLoS Pathogens*, **3**: e75. doi: 10.1371/journal.ppat.0030075
- Jia, F., Lo, N., & Ho, S. Y. (2014). The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. *PloS one*, *9*(5), e95722. https://doi.org/10.1371/journal.pone.0095722
- Rachel B. Bevan, David Bryant, B. Franz Lang, Accounting for Gene Rate Heterogeneity in Phylogenetic Inference, *Systematic Biology*, Volume 56, Issue 2, April 2007, Pages 194–205, https://doi.org/10.1080/10635150701291804
- Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution*, *2*(1), vew007. https://doi.org/10.1093/ve/vew007
- . Salemi, M., W. M. Fitch, M. Ciccozzi, M. J. Ruiz-Alvarez, G. Rezza, and M. J. Lewis. 2004. Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. J. Virol. 78:1602–1603.
- Sanchez, C. M., F. Gebauer, C. Sune, A. Mendez, J. Dopazo, and L. Enjuanes. 3223 1992. Genetic evolution and tropism of transmissible gastroenteritis coronaviruses. Virology 190:92–105.
- St-Jean, J. R., H. Jacomy, M. Desforges, A. Vabret, F. Freymuth, and P. J. Talbot. 2004. Human respiratory coronavirus OC43: genetic stability and neuroinvasion. J. Virol. 78:8824–8834.
- Vijgen, L., Keyaerts, E., Moës, E., Thoelen, I., Wollants, E., Lemey, P., Vandamme, A. M., & Van Ranst, M. (2005). Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *Journal of virology*, *79*(3), 1595–1604. https://doi.org/10.1128/JVI.79.3.1595-1604.2005
- Vivek Jayaswal, John Robinson, Lars Jermiin, Estimation of Phylogeny and Invariant Sites under the General Markov Model of Nucleotide Sequence Evolution, *Systematic Biology*, Volume 56, Issue 2, April 2007, Pages 155–162, https://doi.org/10.1080/10635150701247921