Lucía Martín Fernández

EJERCICIO DE EVOLUCIÓN MOLECULAR Y FILOGENIA

The alignment attached (Rabies_ALL_SPECIES_N.FULL.DATED) includes 70 sequences of the gene encoding the N protein of the Rabies virus. The name of each sequence includes information about the host of origin (bats, dogs, foxes, raccoon, skunks, wolves and humans), GenBank the accession number and the date (year) of sampling.

To answer the following questions, we will use the BEAST software and all its complementary tools (Bouckaert *et al.*, 2019).

1) ¿From which host are the virus sequences genetically closer to those from skunk?

To determine which Rabies viruses are genetically closest to those affecting skunks, we need to build a phylogenetic tree from the given sequences and compare the branching patterns of the different host groups.

One popular method for building a phylogenetic tree is Bayesian inference, which can be modelled using the BEAST software. To run BEAST, we first generated an XML file from the fasta alignment using BEAUTi where some parameters need to be defined.

Considering we don't have much information about the evolution model followed by our sequences, we selected the most flexible parameters. In this context, we selected the GTR (Generalized time reversible) evolution model with estimated base frequencies and both gamma and invariant sites heterogeneity models with partitions into the 3 codons positions. The Markov's Chain will have 100 million states and start with a UPGMA tree. Log data and trees will be save every 10 thousand steps.

The analysis with Tracer of the resulting log file from BEAST shows the robustness of the results since the ESS (Effective size samples), the distribution values and the trace drawings are adequate.

Once we have confirmed the reliability of our data, we can generate a consensus tree using TreeAnnotator. The result is a maximum credibility tree defined by median heights nodes with a burn-in of 10% of the total Markov's Chain length.

In the rooted (Figure 1) and un-rooted (Figure 2) consensus trees we can observe how skunks Rabies' viruses are present in 3 different branches. It seems that the eldest skunk's viruses (1992-1998) are closest to fox Rabies viruses and have a common ancestor with most of the Rabies viruses affecting to the Canidae family species in our phylogenetic tree. However, the most recent skunk's Rabies viruses

from 2001 are more genetically similar to the human Rabies viruses from 1994. This human branch shares an internal node with other Rabies viruses affecting bats. In addition, there is another skunk virus from 2001 that seems closest to the raccoons Rabies viruses than to the rest of species. Nevertheless, in Figure 2 it is observed that this genetic distance is pretty high, especially when comparing it to the similitude between the rest of skunk's viruses and its respective closer sequences. In this context, the closest genetic distance between a skunk virus and another host is found in the eldest skunk's viruses relation with the early 2000s fox and human viruses.

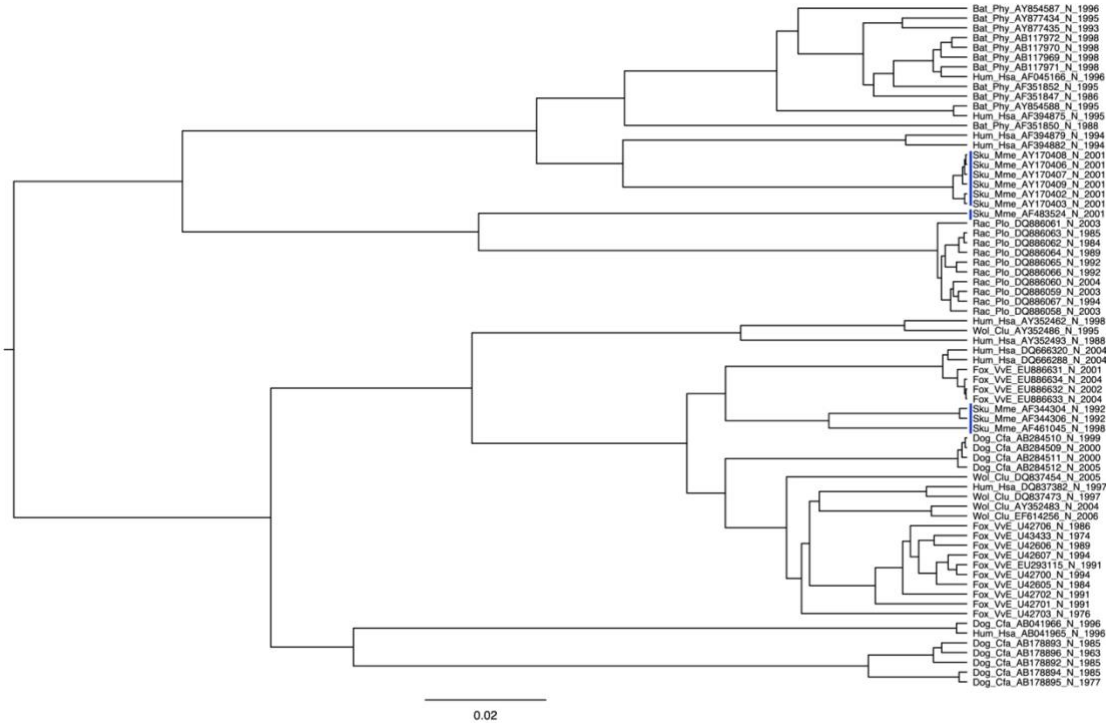

Figure 1. Rooted maximum credibility tree generated with Bayesian inference and a Markov's chain of 100 million states. Skunks' Rabies viruses are marked in blue.

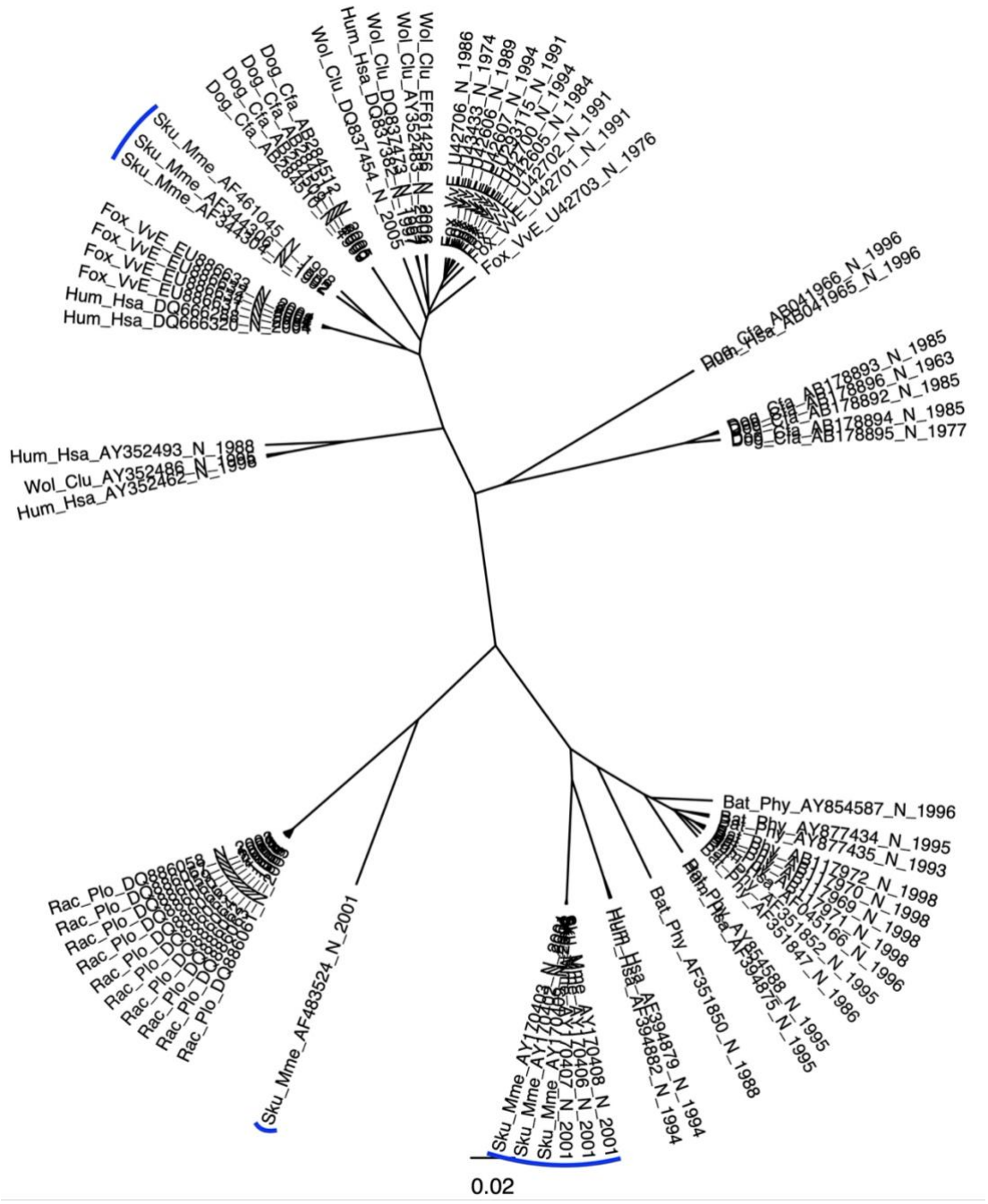

Figure 2. Un-rooted maximum credibility tree generated with Bayesian inference and a Markov's chain of 100 million states. Skunks' Rabies viruses are marked in blue.

2) ¿Could you estimate the time to the most recent common ancestor and the evolutionary rate for the rabies virus?

Estimating the time to the most recent common ancestor (TMRCA) and evolutionary rate for a virus such as rabies requires a detailed phylogenetic analysis, taking into account multiple factors such as the number of substitutions per site and the coalescent theory.

We reused the consensus tree generated with BEAST in the previous step to see if the given sequences follow the molecular clock model in TempEst (Rambaut *et al.*, 2016). Usually, a temporal structure is expected with an R squared between 0.25 and 0.30 or a correlation coefficient higher than 0.5. When parsing the tree's dates and finding the best fitting root by optimizing correlation values we observed that a priori the initial assumption is not held (Table 1). However, TempEst assumes a strict linear clock where the evolutionary rate is constant over time. This assumption may not always be true, especially in cases of rapid evolution such as viruses. Thus, since the observed values are not too far from the expected considering the initial molecular clock assumption, we will continue with the analysis implementing an uncorrelated relaxed clock.

Table 1. Initial temporary structure estimation in TempEst

Date range	43
Slope (rate)	4,45E-3
X-Intercept (TMRCA)	1937,6518
Correlation Coefficient	0,4658
R squared	0,217
Residual Mean Squared	4,9163E-3

In this context, BEAST is rerun, this time, including time-related information. To prepare the XML file, the same parameters from section 1 are introduced in the BEAUTi software. In addition, we parsed the sequences' sampling dates and assumed an uncorrelated relaxed clock with lognormal distribution. Since the relaxed clock model assumes that different branches may have different evolution rates and that these values are independent, we may expect a temporal structure in spite of the initial analysis.

Results' robustness is once again confirmed with Tracer and temporal structural results are obtained (Table 2). The estimation shows that the last common ancestor (LCA) could have been sampled around the year 1040 (95% HPDI [337.9811, 1544.5818]) and the evolutionary rate of the Rabies virus is 1.93E-4 (95% HPDI [8.0825E-5, 3.0022E-4]). The latest result is consistent with the

reviewed bibliography, since other viruses have also shown evolution rates of the 10^{-4} order (Hanada, Suzuki and Gojobori, 2004).

Table 2. Final temporary structural estimation with Tracer

	TMRCA	**Substitution rate**
mean	1039.9658	1.9287E-4
stderr of mean	21.3927	2.3501E-6
median	1140.2814	1.9118E-4
95% HPD interval	[337.9811, 1544.5818]	[8.0825E-5, 3.0022E-4]
effective sample size (ESS)	742.7	569.6

In addition, we generated the consensus tree with TreeAnnotator and introduced it into TempEst to compare both estimated TMRCA results (Table 3). The TempEst estimation shows the X-Intercept to fall in the year 1152 which is closes to the median TMRCA obtain with Tracer than to the mean value. Since all of this values fall into the high posterior density interval (95% HPDI [337.9811, 1544.5818]), we decided to take the median value obtained with Tracer as the most accurate estimate of the TMRCA value for the Rabies virus.

Table 3. Final temporary structural estimation with TempEst.

Date range	43
Slope (rate)	1
X-Intercept (TMRCA)	1152,0191
Correlation Coefficient	1
R squared	1
Residual Mean Squared	6,9218E-14

References

Bouckaert, R. *et al.* (2019) 'BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis', *PLOS Computational Biology*. Public Library of Science, 15(4), p. e1006650.

Hanada, K., Suzuki, Y. and Gojobori, T. (2004) 'A Large Variation in the Rates of Synonymous Substitution for RNA Viruses and Its Relationship to a Diversity of Viral Infection and Transmission Modes', *Molecular Biology and Evolution*, 21(6), pp. 1074–1080. doi: 10.1093/molbev/msh109.

Rambaut, A. *et al.* (2016) 'Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)', *Virus Evolution*, 2(1), p. vew007. doi: 10.1093/ve/vew007.