Lab exercise
Phylogenetic analysis of carnivore rabies
Part One: Estimating and interpreting a phylogeny

Roman Biek, University of Glasgow

Background

Rabies, caused by an RNA virus in the family Lyssaviridae, continues to be one of the most significant zoonoses worldwide. Although rabies can infect most mammal species, reservoir hosts are generally either carnivore or bat species. For this exercise, you will be analyzing rabies nucleoprotein (N) sequences from a rabies variant circulating among carnivore species worldwide. In addition to the genetic data, you also have an excel spreadsheet (*RabiesN_samples.xls*) that contains information about the time, the place and the species these sequences originated from.

Using both sources, you will be investigating questions of host range, geographic distribution and demographic history of carnivore rabies. Make sure to take notes on results and your thoughts along the way as material for later questions and discussions.

Estimating and interpreting a phylogeny

Download the data file *RabiesNWorld.phy* and view it in a text editor. You see the aligned genetic sequences for 70 taxa. Note that the last 7 sequences are included as an outgroup (this includes sequences from different bat species as well as a raccoon and skunk variant, both from North America).

Before loading the sequences into R, load the package `ape`

```
rabies.seq <-read.dna("RabiesNWorld.phy", format="sequential")
```

We will now calculate all pairwise genetic distances between the sequences using two different methods. The first (“raw”) will simply calculate the number of different sites between sequences, the second (“TN93”) will adjust distances based on Timura & Nei’s model, which assumes different base frequencies and different rates of transitions and transversions. We will also allow nucleotide sites to vary in their propensity to change according to a gamma distribution. We will then compare the two types of distances graphically.

```
RawDist<-dist.dna(rabies.seq,model="raw")
TN93Dist<-dist.dna(rabies.seq,model="TN93", gamma=TRUE)
plot(TN93Dist, RawDist)
abline(a=0, b=1, lty=2)
```

Make sure you understand what this graph tells you. What does the straight line represent and what does the deviation from this line mean?

We will use the distances calculated under the TN93 model to build a tree using the neighbor joining algorithm.

```
nj.rabies.TN93<-nj(TN93Dist)
plot(nj.rabies.TN93, cex=0.5)
add.scale.bar(y=0.5, length=0.01)
```

Look at the tree - where does it have its root?

Use sequences 64-70 as an outgroup to root the part of the tree we are actually interested in (the carnivore rabies sequences represented by sequences 1-63).

```
nj.rabies.TN93<-root(nj.rabies.TN93, 64:70)
```

Plot the tree again and notice what has changed.

Next you will perform a bootstrap analysis to assess the reliability of your phylogenetic estimate.

```
nj.boot.rabies<-boot.phylo(phy=nj.rabies.TN93, x=rabies.seq,
FUN=function(xx) nj(dist.dna(xx,pairwise.deletion=TRUE)), B=200)
nj.rabies.est<-nj.rabies.TN93
nj.rabies.est<-root(nj.rabies.est, 64:70)
nj.rabies.est$node.label <- nj.boot.rabies / 2
write.tree(nj.rabies.est, "nj.rabies.boot.tre")
```

You can plot the saved NJ tree in R if you want but it is easier to edit it in a standalone (and free!) program called FigTree. You can use different layout options to make the tree easier to read (for example, I recommend Trees->OrderNodes->decreasing). It may help to color code branches or clades in order to distinguish ingroup and outgroup, host groups, geographic regions, etc

Use the first NJ tree and the consensus tree (along with the bootstrap proportions) to consider the following questions

- Do rabies variants tend to cluster by species or rather by geographic region?
- Which major groups/geographic regions can be distinguished?
- How well are these groups supported by bootstrap values?