Lab exercise
Spatial phylogenetics of carnivore rabies
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 carnivora 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 phylogeography of carnivore rabies. Make sure to take notes on results and your thoughts along the way as material for later questions and discussions.

Before you start, download program Figtree from the following website
http://tree.bio.ed.ac.uk/software/figtree/

1) Estimating and interpreting a phylogeny

Start program Paup* and open the file *RabiesNWorld.nex*. Choose 'edit' rather than 'execute' to take a look at the data. You see the aligned genetic sequences for 70 taxa. Following the data you see a series of commands. For example, one command specifies the last 7 sequences as outgroup sequences (this includes sequences from different bat species as well as a raccoon and skunk variant, both from North America). Further below, parameters for the nucleotide substitution model are given.

Execute the file. Then, in the Paup command line, type

```
Nj
```

for 'neighbor joining'. This will create a 'quick and dirty' tree based on pairwise distances between sequences (searching for the best ML tree for example would take too much time for this many sequences). Save the tree by typing

```
SaveTrees File=RabiesNWorld_NJ.tre BrLens=yes
```

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

```
Set WarnTSave=no MaxTrees=1000;
```

```
Bootstrap NReps=1000 Search=NJ Treefile= RabiesNWorld.out
BrLens=no GrpFreq=no;
```

This will take a few minutes to finish. In the meantime, open the saved NJ tree in program 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

Check if your bootstrap analysis has finished. You can save the consensus tree (but not the bootstrap frequencies) with the following commands

```
GetTrees File= RabiesNWorld.out;
```

```
ConTree all/strict=no MajRule=yes ShowTree=no GrpFreq=no;
```

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?

2) Estimating ancestral spatial states

In the lecture, I mentioned a recent paper by Wallace et al (2007, PNAS) reconstructing origin and global migration patterns of avian influenza H5N1 using parsimony reconstruction. You will be using the same methodology to examine the geographic origin and migration of carnivore rabies.

Decide on which geographic regions you want to distinguish (I would recommend 4-8 groups). Then, build a new input file by keeping the taxa names but replacing the sequences with variables between 0 and 7, representing the geographic region of origin. Use '?' for the origin of outgroup taxa, so their locations are not included in the estimation.

You will also need to modify the input file header so something like

```
BEGIN DATA;
      DIMENSIONS NTAX=70 NCHAR=1;
      FORMAT DATATYPE=Standard symbols="01234" MISSING=?;
      MATRIX
```

Note that you need to define all symbols here that you are planning on using for the different geographic regions. Also, change the commands below the data to the following

```
Begin Paup;
Outgroup 64-70;
set criterion=pars;
pset opt=Deltran
End;
```

Execute the file and load up your first NJ tree. Remember that this tree summarizes our knowledge about the genetic relationships among rabies sequences. Based on this tree and the geographic regions assigned to taxa, the program will try to find the minimum number of migration events, working its way backwards from the tree trips. We can visualize these reconstructions by typing

```
reconstruct 1;
```

We can also get a table of all state changes using

```
DescribeTrees 1/ChgList=y;
```

you can copy and paste this results from this table to text editor or spread sheet. Summarize your results with regard to the following questions:

- what is the inferred region of origin for carnivore rabies?

- How many migration events have there been from and into each region? (if you have time you could try and visualize those with arrows on a map)

So far, we have made no attempt to quantify the uncertainty of our reconstruction. In their paper, Wallace et al use a Monte Carlo test with 10,000 trials to randomize the localities on the tree tips. They then ask whether the probability of the migration events inferred from the original tree is higher than the frequency of the same migration event if localities are randomly distributed.

Because we don't have their code and because of time constraints we will have to limit ourselves to a few randomizations. Copy the taxa names and localities from your geographic input file and paste them into a spreadsheet. Separate entries into two columns ('text to columns'), enter a third column with random numbers and use these to randomize locations only. Paste taxa names and localities back into the input file, save under a different name (like 'replicate1.nex') and repeat the analysis above. Determine the frequency of migration events from however many randomizations you manage to do and compare to your original results.

Obviously, this test only addresses one source of uncertainty in your estimate (can you name which one?). What is the null hypothesis this test is based on and do you find this null biologically reasonable? What other sources of uncertainty would you consider important and how could you go about assessing those?

3) Testing spatial hypotheses using topological constraints

According to our inference for the NJ tree, rabies virus at one point was introduced from Africa into the New World (giving rise to samples from Brasil, Wisonsin, Canada and Mexico) and subsequently moved back to Africa from there. An alternative hypothesis (requiring one less migration event) would be that this lineage was maintained in Africa the whole time but at one point was introduced into the Americas. The second hypothesis is equivalent to the four American samples forming a monophyletic group (a group that can be traced back to a single ancestor and contains all descendents of this ancestor).

Asking that the tree should contain a certain monophyletic group is called enforcing a constraint. Once we have obtained an estimate of the phylogeny with this constraint we can compare its likelihood to that of the original estimate and determine whether it is significantly worse.

The first step is to define the constraint

```
constraints NewWorld = ((taxon1, taxon2, taxon3, etc));
```

where taxa can either be represented by their name or by their entry number in the sequence file). Next, we will find a NJ tree that is compatible with this constraint and save this tree.

```
NJ enforce=yes
```

Load up the original NJ tree into memory as well (make sure the program tells you there are two trees in the memory)

```
Gettrees file= RabiesNWorld_NJ.tre mode=7
```

We will compare these two trees by obtaining a likelihood score for each one and asking whether the difference in scores is large enough to favor one tree over the other (using a test named after Shimodaira and Hasegawa):

```
lscores all/displayOut=y SHTest=RELL
```

What does the result tell you about the two hypotheses regarding the introduction of rabies in the New World?

Note that there are two other New World sequences. Define a constraint that is compatible with a single introduction of rabies virus in the Americas and repeat the test above.