Lab exercise
Phylogenetic analysis of carnivore rabies
Part Two: Phylogenetic applications – inferring movement and testing hypotheses

Roman Biek, University of Glasgow

Estimating ancestral spatial states

In the introduction, 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 (in program PAUP) to examine the geographic origin and migration of carnivore rabies.

First, you will be re-estimating a NJ tree, this time using a slightly more sophisticated (and better fitting) substitution model. Open up PAUP and execute the file "*RabiesNWorld.nex".* Construct the neighbor joining tree and save it by typing

```
NJ
```

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

For the spatial analysis, we will distinguish four geographic regions (Europe/North Africa, sub-Saharan Africa, North/South-America, Asia/Middle East. The input file "*RabiesNWorld_Continent.nex*" contains the same carnivore rabies taxa you analyzed previously but instead of sequences in contains variables 1-4, representing the geographic region of origin. Note that for outgroup taxa, region is shown as '?', so their locations are not included in the estimation.

Execute the file "*RabiesNWorld_Continent.nex*" (click o.k. when program asks about replacing data file) and load up your first NJ tree.

```
gettrees file=RabiesNWorld_NJ.tre;
```

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;
```

Use your results to address 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 can color the branches in FigTree to visualize these events)

Normally we would want to quantify the uncertainty of our reconstruction but for the sake of time we won't do this here. 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. Think about which source of uncertainty in the estimate this approach addresses. Can you think of other sources of uncertainty that you should also consider?

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.

First, load the sequence data again ("*RabiesNWorld.nex*"). The next 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). Find the four American taxa mentioned above to define the constraint. Next, we will find a NJ tree that is compatible with this constraint:

```
NJ enforce=yes
```

```
showtrees
```

Save this tree as you did before and add the original NJ tree (without the constraint) into memory (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.