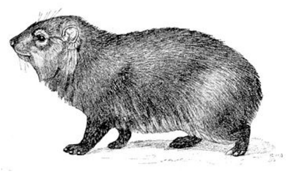
Madingley field course computer practical:

Evolutionary Relationships of the Caniform Carnivora

1. Molecular phylogenetics

Molecular phylogenetics (inferring evolutionary relationships from DNA sequence data) has filled many gaps in our understanding of life. For example, until quite recently, zoologists were puzzled by hyraxes.



These mammals look superficially like rodents, but have retained a number of characteristics from the very earliest mammals (e.g., poor temperature regulation). In recent years, molecular phylogenetics has shown clearly that the four living species of hyrax are most closely related to elephants (Proboscidea) and sea cows (Sirenia)...

In this practical, we revisit another recently solved puzzle in mammalian evolution. We will concentrate on members of the order Carnivora, and particularly the suborder Caniformia (i.e., species more closely related to dogs than to cats). This group contains diverse animals such as seals, skunks, badgers, bears and otters, as well as dogs (the feliform Carnivora include meerkats and hyaenas as well as cats).

2. Checking the sequence alignment

We will construct our phylogeny from an alignment of the cytochrome c oxidase 1 gene (often labelled COI or cox1). This gene is encoded in the maternally inherited mitochondrial genome, and is involved in oxidative phosphorylation.

Begin by finding the file “Caniform.COI.fasta”. Open this file in a text editor to visualise the data, which is in “FASTA” format. Then try opening the file with manual alignment software. Some publicly available alignment programs are *Se-Al* (http://tree.bio.ed.ac.uk/software/seal/) for Macs and *Bioedit* (http://www.mbio.ncsu.edu/bioedit/bioedit.html) for PCs.

Checking your alignment carefully is very important. In the subsequent analyses, each column in the alignment will be assumed to contain only homologous bases (i.e., bases descended directly from a single ancestral base). If this assumption is incorrect (i.e., if we have misaligned our sequences) then our results will be worthless.

COI is a protein-coding gene, so to check the alignment, it is helpful to translate the sequence and view the amino acids for which the DNA codes. It is noticable that these vary less than the underlying DNA. Remember that vertebrate mitochondria have an unusual genetic code, so you'll need to specify this correctly in the alignment software. What happens if you translate the sequence using the wrong genetic code?

While you are looking at the alignment, also take note of the mammal species that we have sampled. The names are in the format: "Genus_species__Family" (this is non-standard, but will be helpful for visualising the relationships later on). Feel free to look up any species names that you don't recognise.

3. Estimating the tree

We are now ready to estimate the phylogeny of the species from which the COI genes were sequenced. We will do this in the *R* software package. This freely available software is extremely useful for many kinds of analyses.

First, allow R to access the directory where your alignment file can be found. To do this, at the R terminal, type something like:

```
> setwd("~/Desktop/MadingleyFiles")
```

changing the directory name, as appropriate. To learn more about the command "setwd" or any of the other commands we use below, you can use the help commands:

```
> ?setwd
> help(setwd)
> ??directory
```

(`??` is useful when you've forgotten the exact name of a function).

We now need to install an R “library” which contains the functions necessary to do phylogenetics. We are going to use a package called “ape”, so type:

```
> install.packages(“ape”, dep = TRUE)

> library(ape)
```

Now we need to load in our alignment data to R. The following command will read our fasta file, and place its contents into a variable called “f”, that we can use for further analysis.

```
> f <- read.FASTA(“Caniform.COI.fasta”)
```

If this had loaded successfully, typing f should show you the following:

```
> f

53 DNA sequences in binary format stored in a list.

All sequences of same length: 1533

Labels: Canis_latrans__Canidae Canis_lupus__Canidae
Cuon_alpinus__Canidae Nyctereutes_procyonoides__Canidae
Vulpes_vulpes__Canidae Spilogale_putorius__Mephitidae ...

Base composition:

a     c     g     t

0.274 0.247 0.178 0.302
```

Note that R tells us how long the COI gene sequence is (1533 bases), how many species are included in our alignment (53) and the average base composition of the sequences. We can also look at the alignment directly, by typing something like:

```
> as.character(f)
```

We will estimate the phylogeny of our 53 species using a simple “distance method”, which first estimates of genetic distance between each pair of sequences, and then arranges these into a tree using a hierarchical clustering algorithm. Details of the algorithm can be found here: http://en.wikipedia.org/wiki/Neighbor_joining.

To estimate the matrix of genetic distances, type:

```
> d <- dist.dna(x, model = “F84”)
```

Note that this command is not simply counting the numbers of differences between each sequence pair. Such a simple method can give misleading results, because it ignores the possibility that some bases will have changed multiple times along a single evolutionary lineage. To account for this, we have used a Markov model of DNA sequence evolution introduced by Joe Felsenstein in 1984 (hence “F84”). This model also accounts for some important features of molecular evolution, e.g., the fact that transitions (changes between two pyrimidines or two purines) occur more frequently than transversions (changes between a purine and a pyrimidine), and that some bases occur more frequently than others (mammalian mitochondrial genome often contain more Ts than Gs for example). The mathematical details of these nucleotide substitution models can be found here: http://en.wikipedia.org/wiki/Models_of_DNA_evolution.

We now turn our distance matrix into a phylogenetic tree using the “neighbour joining algorithm”.

```
> p <- nj(d)
```

Typing p should show you something like this:

```
Phylogenetic tree with 53 tips and 51 internal nodes.

Tip labels:

Canis_latrans__Canidae, Canis_lupus__Canidae,
Cuon_alpinus__Canidae, Nyctereutes_procyonoides__Canidae,
Vulpes_vulpes__Canidae, Spilogale_putorius__Mephitidae,
...
```

4. Estimating the uncertainty in our inferences

Before viewing our phylogeny, we should also estimate how much uncertainty is associated with our inference. One way to do this is to ask whether different sites in the COI gene (i.e., different columns in our alignment) suggest different evolutionary relationships. We can do this by using “bootstrapping”. Bootstrapping means resampling sites from our alignment at random with replacement, until we have a new alignment of the same length, we then reestimate the phylogeny from this random subsample of sites. After repeating this process a large number of times, we check how frequently the phylogenetic groupings from our initial tree also appear in our bootstrapped trees (which were inferred from a random subsample of the sites). Groups which appear in most of these trees are said to have high support (in that most sites in the alignment support the same grouping). (The phrase bootstrapping refers to this reuse of the same data, which is analogous to “pulling ourselves up by our own bootstraps”).

To generate bootstrap confidence intervals, we use the following command:

```
> p$node.label <- boot.phylo(p,f,function(x)
nj(dist.dna(x, model='F84')))
```

The syntax of this command is complicated, because to boostrap our data, the function needs to know our alignment, `f`, our phylogeny, `p`, and the method that we used to generate this phylogeny, `nj(d, model = “F84”)`.

Each node (branching point) in our tree is now associated with a bootstrap confidence value, telling us the proportion of the randomised resamplings of the data that contained the same node.

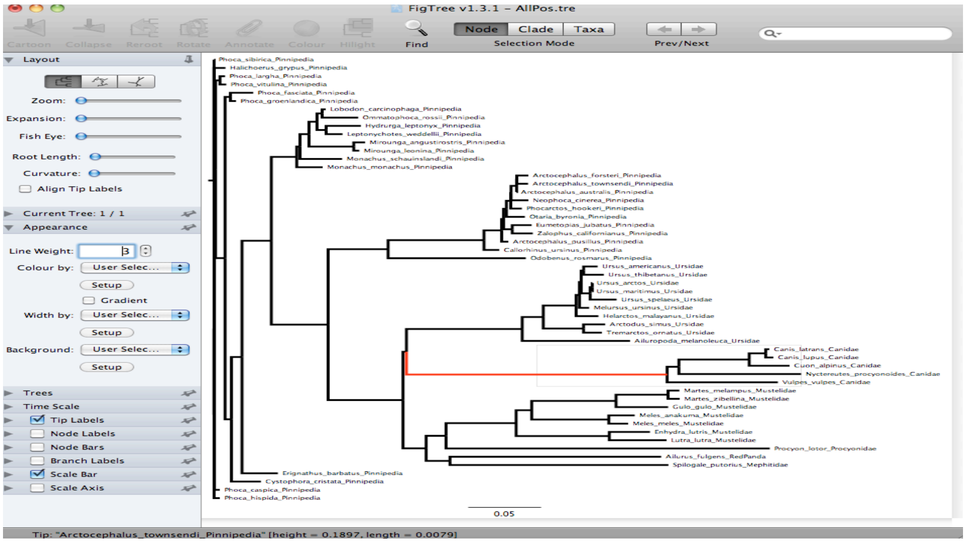
Phylogenetic trees can be visualised in R using `plot.phylo(p, show.node.label=TRUE)`. The help file for this function describes many ways of changing the appearence of these plots. However, in many cases, it is easier to use alternative software to visualise phylogenetic trees. To do this, we first need to save our phylogeny to disk:

```
> write.tree(p,file=”Caniform.tre”)
```

5. Visualing the tree

To view and manipulate our tree, will use the software *Figtree*, which can be downloaded here (http://tree.bio.ed.ac.uk/software/figtree/).

Open *Figtree*, and then load in your Caniform phylogeny (*Figtree* will also ask you to give a name to our bootstrap confidence intervals). Once the tree is open, it should look something like this:


FigTree v1.3.1 – AllPos.tre
Cartoon
Collapse
Reroot
Rotate
Annotate
Colour
Hilight
Find
Node
Clade
Taxa
Selection Mode
Prev/Next
Layout
Zoom:
Expansion:
Fish Eye:
Root Length:
Curvature:
☐ Align Tip Labels
Current Tree: 1 / 1
Appearance
Setup
☐ Gradient
Setup
Setup
Trees
Time Scale
☑ Tip Labels
☐ Node Labels
☐ Node Bars
☐ Branch Labels
☑ Scale Bar
☐ Scale Axis
Phoca_sibirica_Pinnipedia
Halichoerus_grypus_Pinnipedia
Phoca_largha_Pinnipedia
Phoca_vitulina_Pinnipedia
Phoca_fasciata_Pinnipedia
Phoca_groenlandica_Pinnipedia
Lobodon_carcinophaga_Pinnipedia
Ommatophoca_rossii_Pinnipedia
Hydrurga_leptonyx_Pinnipedia
Leptonychotes_weddellii_Pinnipedia
Mirounga_angustirostris_Pinnipedia
Mirounga_leonina_Pinnipedia
Monachus_schauinslandi_Pinnipedia
Monachus_monachus_Pinnipedia
Arctocephalus_forsteri_Pinnipedia
Arctocephalus_townsendi_Pinnipedia
Arctocephalus_australis_Pinnipedia
Neophoca_cinerea_Pinnipedia
Phocarctos_hookeri_Pinnipedia
Otaria_byronia_Pinnipedia
Eumetopias_jubatus_Pinnipedia
Zalophus_californianus_Pinnipedia
Arctocephalus_pusillus_Pinnipedia
Callorhinus_ursinus_Pinnipedia
Odobenus_rosmarus_Pinnipedia
Ursus_americanus_Ursidae
Ursus_thibetanus_Ursidae
Ursus_arctos_Ursidae
Ursus_maritimus_Ursidae
Ursus_spelaeus_Ursidae
Melursus_ursinus_Ursidae
Helarctos_malayanus_Ursidae
Arctodus_simus_Ursidae
Tremarctos_ornatus_Ursidae
Ailuropoda_melanoleuca_Ursidae
Canis_latrans_Canidae
Canis_lupus_Canidae
Cuon_alpinus_Canidae
Nyctereutes_procyonoides_Canidae
Vulpes_vulpes_Canidae
Martes_melampus_Mustelidae
Martes_zibellina_Mustelidae
Gulo_gulo_Mustelidae
Meles_anakuma_Mustelidae
Meles_meles_Mustelidae
Enhydra_lutris_Mustelidae
Lutra_lutra_Mustelidae
Procyon_lotor_Procyonidae
Ailurus_fulgens_RedPanda
Spilogale_putorius_Mephitidae
Erignathus_barbatus_Pinnipedia
Cystophora_cristata_Pinnipedia
Phoca_caspica_Pinnipedia
Phoca_hispida_Pinnipedia
0.05

Line Weight: 3

Colour by: User Selec...

Width by: User Selec...

Background: User Selec...

Tip: "Arctocephalus_townsendi_Pinnipedia" [height = 0.1897, length = 0.0079]

This is a bit of a mess, mainly because our tree is *unrooted*. That means that we have not inferred a common ancestor for the group of species as a whole. Figtree plot has guessed the root for us, and has not done a very good job. In this case, we know from external evidence that the dog family (Canidae) is most distantly related to the other Caniformia. To use this knowledge to root out tree, highlight the branch on your tree that leads to the 5 members of the Canidae family (the branch shown coloured in red above). Then click on the yellow "Reroot" icon on the *Figtree* toolbar.

Now you have a correctly rooted tree of the Carnivora. Looking at this tree we can infer quite a lot about the relationships within the group. For example, three families of caniform species have adopted a semi-aquatic lifestyle: the seals (Phocidae), walrus (Odobenidae) and sea lions (Otariidae). Looking at your phylogeny, how many evolutionary transitions do you think occured from a fully terrestrial lifestyle? Looking more closely at the true seals, many taxonomists have argued that the genus *Phoca* be replaced with three new genera; can you tell why?

Another finding of our analysis is that the bootstrap support values are quite low for almost all of the groupings. This means that we cannot be very certain about any of our inferences. How might we improve power in future analyses?

6. What is a red panda?

The analysis above included only some of the extant species of Caniformia. One absent species was the red panda (*Ailurus fulgens*), which lives in temperate forests in the Himalayas.

Like hyraxes (see above) red pandas were a zoological mystery until quite recently. Because they specialise in eating bamboo (unusually for a member of the Carnivora), many zoologists argued that this species was a type of panda - hence the name - and thus belonged in the family Ursidae alongside the giant panda (*Ailuropoda*). Others disagreed, arguing that the red panda was more closely related to racoons (family Procyonidae). To see how molecular phylogenetics solved this puzzle, we'll need to add a COI sequence from the red panda to our alignment.

The best place to look for publicly available DNA sequences is the Genbank database, where geneticists the world over deposit their data. Go to Genbank, at http://www.ncbi.nlm.nih.gov/genbank/, and see if you can find the red panda COI sequence. When you've found it, record its "Accession number" (this is a unique identifier for every sequence deposited in Genbank).

You can add the red panda sequence to your existing COI alignment in two ways. First, the sequence can be read directly into R, using the command `read.GenBank`, and then saved to disk, using `write.fasta`. Alternatively, you can copy and paste directly into a text file. Remember to check your alignment thoroughly before re-estimating the phylogeny.

What are the closest relatives of red pandas? Why do you think that their phylogenetic position proved so hard to determine?