Morphological sampled ancestors

This is going to walk through a couple simulated examples to test the morphological likelihood calculation on sampled ancestor and bifurcating trees.

Example using equal branch lengths

We are going to simulate some data using **INDELible** (http://abacus.gene.ucl.ac.uk/software/indelible/tutorial/settings.shtml) and test the abilty of **mandos** to correctly identify ancestral and branching relationships. First let's create a small tree with equal branch lengths:

```
(E:0.1,((A:0.1,B:0.1):0.1,(C:0.1,D:0.1):0.1):0.1);
```

Now, we will simulate some simple data in **INDELible** under Jukes-Cantor by saving this text in `control.txt`.

```
[TYPE] NUCLEOTIDE 1
[MODEL] jukes
 [submodel] JC
[TREE] tree (E:0.1,((A:0.1,B:0.1):0.1,(C:0.1,D:0.1):0.1):0.1);

[PARTITIONS] JC69
 [tree jukes 100]

[SETTINGS]
 [ancestralprint] SAME

[EVOLVE]
 JC69 1 pJC
```

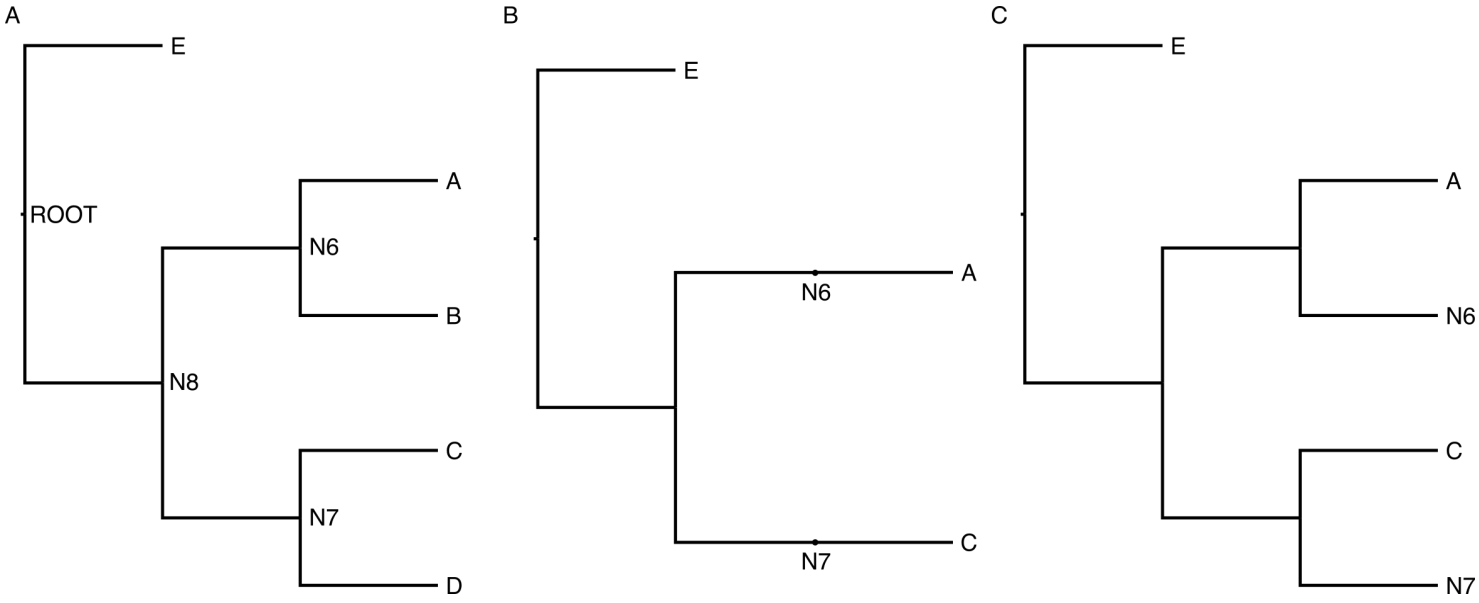
The `[ancestralprint]` option in the control file indicates that we want to save the ancestral states. Run the simulation by just running **INDELible** from the command line in the directory where the control file is saved. Since our primary focus here is morphological data, let's recode the resulting DNA alignment to binary characters by combining purines and pyrimadines:

(running from the top directory of the repo)

```
cd ../../
python recode_binary.py sims/equal_brlens/pJC_TRUE.phy 2
```

By deleting taxa *B* and *D* from the tree above, we can treat the sequences generated at *N6* and *N7* as lineages directly ancestral to tips *A* and *B* (Fig. S1b). To test the identifiability of directly ancestral relationships, let's assume that only taxa *E*, *A*, *C*, *N6*, and *N7* have been sampled in the 'fossil record'.

Figure S1:



Copy the recoded simulated characters to a new file:

```
cp BIN_pJC_TRUE.phy anc.phy
```

Open `anc.phy` and delete all sequences but those for taxa *E*, *A*, *C*, *N6*, and *N7*. Make sure the phylip-style header is correct. We then want to compare the likelihood of these simulated data given an anagenetic representation (Fig. S1b) to a purely cladogenetic representation (Fig. S1c). If our method works, the tree shown in Fig. S1b should have a better log-likelihood than Fig. S1c. To test, we will need to make a quick partitons file. Open a new file and paste this line:

```
MULTI, 2 = 1-100
```

Save this as something sensible (I am going to call it `pJC.phy.models`).

Save newick representations of the bifurcating and anagenetic topologies in the files `h1.tre` and `h2.tre`, respectively (these are what we are going to use to calculate likelhoods).

```
(E,((A,N6),(C,N7)));
(E,((A)N6,(C)N7));
```

We will now calculate some likelhoods using the script `morphliketest.py`. Let's try the cladogenetic representation first:

```
python ../../scripts/morphliketest.py h1.tre anc.phy pJC.phy.models
```

This yields the output (Copy the recoded simulated characters to a new file: (the actual numbers may vary due to imperfections in optimization):

```
optimizing branch lengths on: (E,((A,N6),(C,N7)))
optimizing 8 parameters under Mk
Morphological log-likelihood: -188.395205688
AIC: 392.790411376
```

Next, we'll do the same for the anagenetic tree:

```
python ../../scripts/morphliketest.py h2.tre anc.phy pJC.phy.models
```

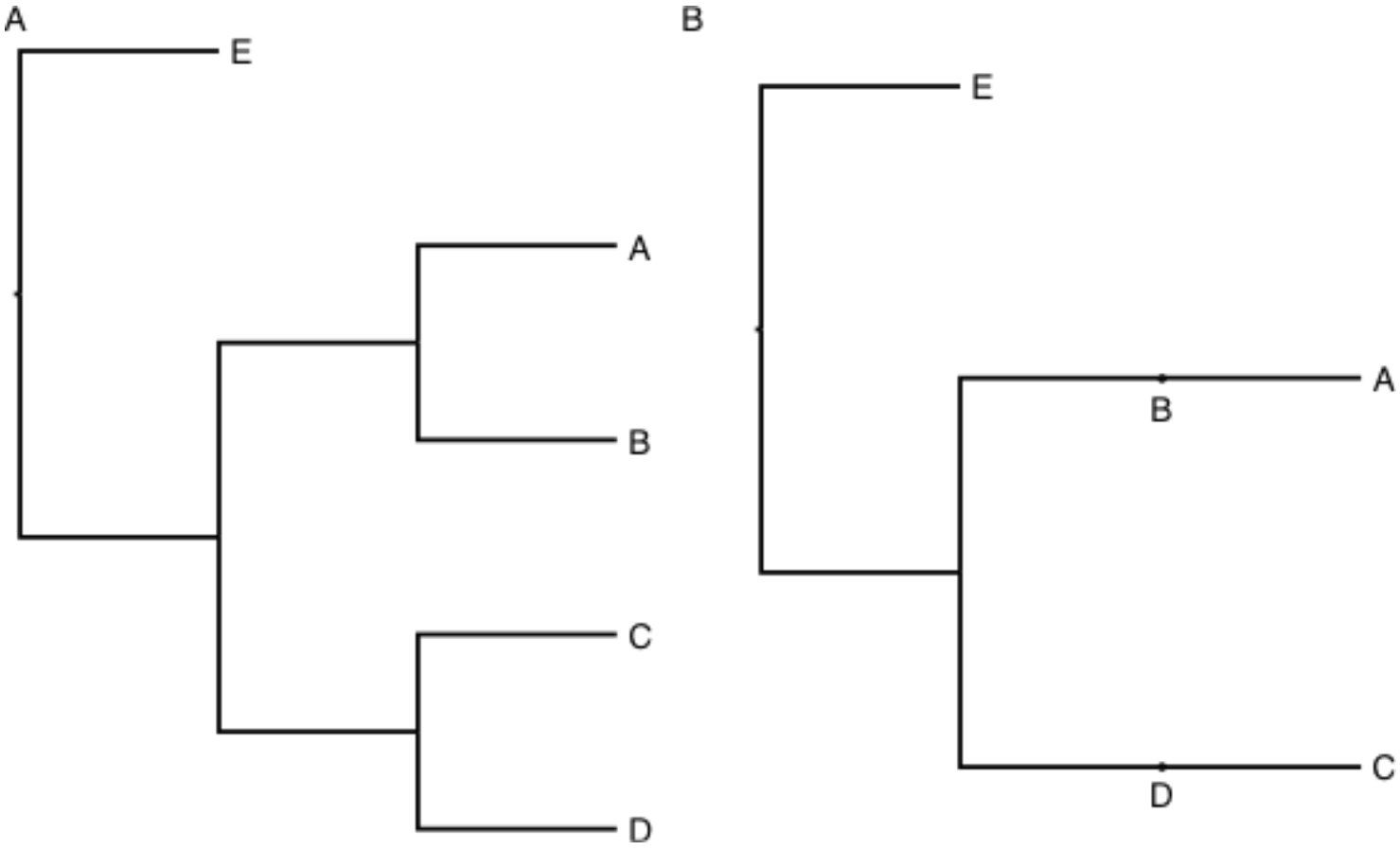
with the result:

```
optimizing branch lengths on: (E,((A)N6,(C)N7))
optimizing 6 parameters under Mk
Morphological log-likelihood: -183.483717854
AIC: 378.967435708
```

Cool! We (correctly) find that the anagenetic tree has a higher likelihood than one where the ancestral *N6* and *N7* are represented as collateral to *A* and *C*. So looks like we are safe from false negatives when simulated branch lengths are equal.

Now let's try for false positives. Copy the alignment with all of the simulated sequences into a new file and delete all of the ancestral sequences. I will call mine `bif.phy`. This will leave us with something that looks like:

```
5  100
E   0101111111010100011011000110101000000110011010011010110011011111010110011111000011
0110100111110001010
A   0111111111000110011011000110111100000110011110011010111011000110010100011111010011
0110101111110011101
B   1101011011000100011011000010111101000110001110011010101011001110010110001110101111
0110111111111001101
C   0010111011000110011011000010101101000010011110010100011011011010010110011010100111
0110101111110010100
D   0010111110000100011010000101111101001010011110001000011000011000010110010010100101
0110101111110010101
```



We are now going to test whether we can correctly identify that all of these taxa should be treated as collateral (Fig. S2a) rather than ancestral (Fig. S2b). In a new file, save the newick string representing the tree that we used to simulate the data:

```
(E:0.1,((A:0.1,B:0.1):0.1,(C:0.1,D:0.1):0.1):0.1);
```

I will call mine `bif_h1.phy`. Let's calculate the likelihood of this tree as we did above:

```
python ../../scripts/morphliketest.py bif_h1.tre bif.phy pJC.phy.models
```

Giving the output:

```
optimizing branch lengths on: (E,((A,B),(C,D)))
optimizing 8 parameters under Mk
Morphological log-likelihood: -254.899555693
AIC: 525.799111385
```

Let's compare this to a tree where taxa *B* and *D* are ancestral to *A* and *C*. Instead of making a new file, we can just collapse them using the `make_ancestor()` function. Uncomment the following line in our script (should be line 15):

```
mandos.tree_utils2.make_ancestor(tree,"B,D")
```

and run as above:

```
python ../../scripts/morphliketest.py bif_h1.tre bif.phy pJC.phy.models
```

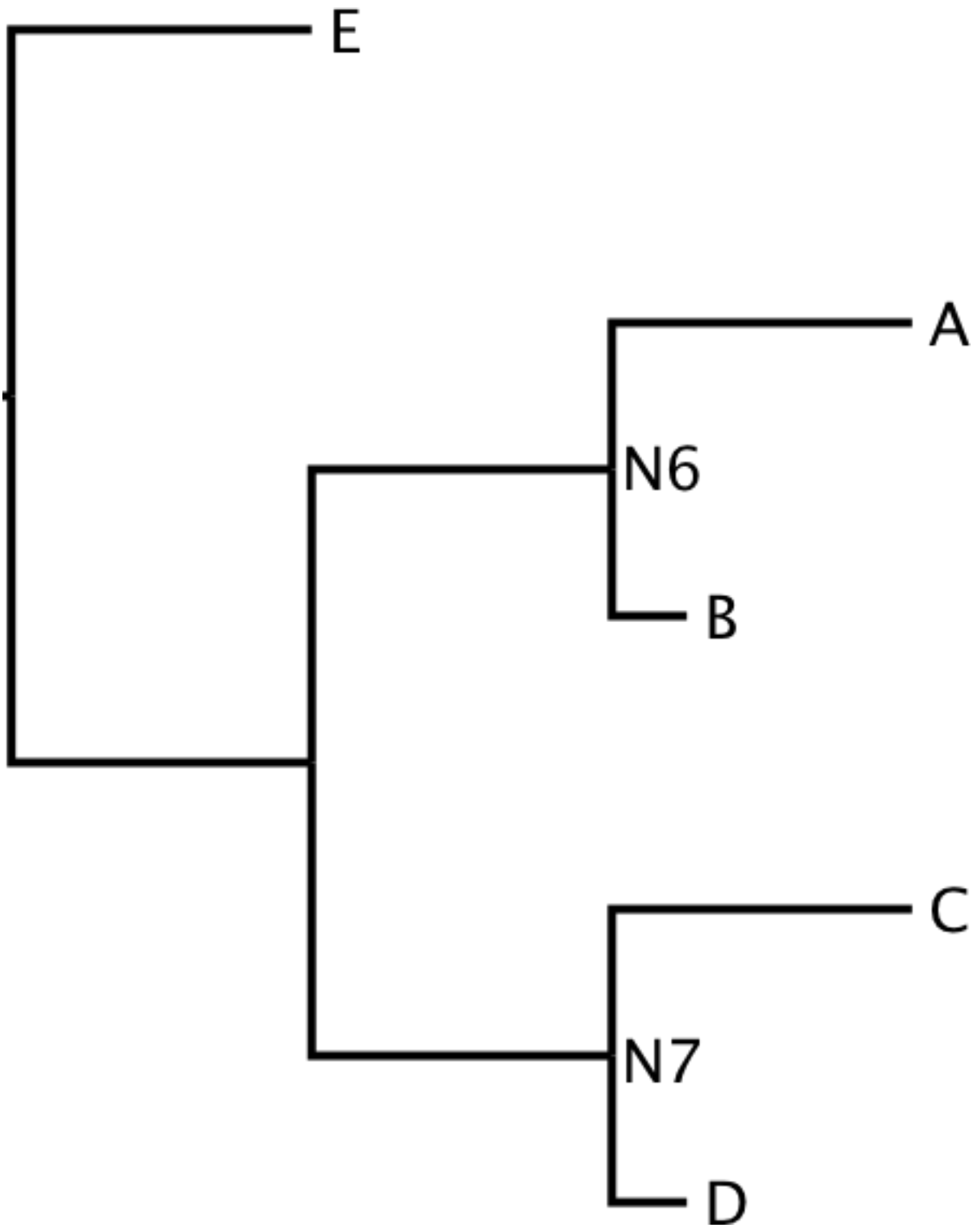
Output:

```
optimizing branch lengths on: (E,((A)B,(C)D))
optimizing 6 parameters under Mk
Morphological log-likelihood: -275.387679604
AIC: 562.775359209
```

Ok cool, our AIC is better for the bifurcating tree.

Heterogeneous branch lengths

We are going to try an example where branch lengths are unequal. We will simulate a dataset as above, retaining the ancestral states, on this tree:



This can be seen in the control file `control.txt`. As before, run **INDELible** and recode the characters using our recoding script (see above). This time, let's leave the taxa *A* and *C* and see if we can identify signal that these taxa are ancestral. To calculate these likelihoods, we will need to name the internal nodes in a new newick string:

```
(E:0.1,((A:0.1,B:0.025)N6:0.1,(C:0.1,D:0.025)N7:0.1):0.1);
```

I have saved this in the file `anc_h1.tre`. We must also copy our recoded characters into a new file and delete the sequences we are not using (*N8* and *ROOT*). I have done this in the file `anc.phy`. Create a partition file as above and calculate the AIC:

```
python ..//../scripts/morphliketest.py anc_h1.tre anc.phy pJC.phy.models
```

(open the script and comment the line `#mandos.tree_utils2.make_ancestor(tree,"B,D")`)

This yields the output:

```
optimizing branch lengths on: (E,((A,B)N6,(C,D)N7))
optimizing 8 parameters under Mk
Morphological log-likelihood: -215.165520315
AIC: 446.33104063
(E:0.101228330063,((A:0.133298549229,B:2.1389079548e-05)N6:0.0735060635546,(C:0.19296
7495819,D:0.00566237583368)N7:0.0446557527041):0.0832053128468):0.0
```

Ok, so our branch lengths are not perfect. Let's compare this to a tree where *N6* and *N7* have their own branches:

```
(E,(((A,B),N6),((C,D),N7)));
```

I have saved this newick in the file `anc_h2.tre`. Running the test script yields the output:

```
optimizing branch lengths on: (E,(((A,B),N6),((C,D),N7)))
optimizing 12 parameters under Mk
Morphological log-likelihood: -257.708296634
AIC: 539.416593269
(E:0.137511119066,(((A:0.179544478792,B:6.06731438873e-05):0.0787686134428,N6:0.13058
5867758):0.162754179761,((C:0.233890733548,D:0.122266960171):0.049943633744,N7:0.1058
23234607):0.0103339364608):0.163320993373):0.0
```

Ok, so our method correctly prefers the tree where *N6* and *N7* are direct ancestors. Lets remove the ancestral taxa and see if we get a false positive when making taxa *B* and *D* are made ancestral. Save the newick string contained in the control file to `bif.tre`. Also copy the binary matrix to the file `bif.phy` and delete all ancestral sequences.

Calculate the AIC of this tree:

```
python ..//../scripts/morphliketest.py bif.tre bif.phy pJC.phy.models
```

output:

```
optimizing branch lengths on: (E,((A,B),(C,D)))
Morphological log-likelihood: -202.555096089
AIC: 405.110192178
(E:0.1,((A:0.1,B:0.025):0.1,(C:0.1,D:0.025):0.1):0.1):0.0
```

Compare this to a tree where tips *B* and *D* are made ancestral by uncommenting line 15 in the script as above.

```
optimizing branch lengths on: (E,((A)B,(C)D))
Morphological log-likelihood: -201.46350255
AIC: 402.927005099
(E:0.1,((A:0.075)B:0.125,(C:0.075)D:0.125):0.1):0.0
```

Here, the method breaks. So now we know that when testing between anagenetic and cladogenetic arrangements between collateral sister taxa where one has a much shorter morphological branch length, we might falsely prefer an arrangement where the short branch length taxon is collapsed as a direct ancestor. In palaeontological applications, this might be in part a matter of interpretation, but should encourage caution nonetheless.

This result highlights the importance of temporal data in helping to constrain the set of possible hypotheses. It is not likely that one would actually test the above scenario in a real application with stratigraphic data, since the temporal co-occurrence of the terminal taxa would preclude directly ancestral hypotheses from the outset. However, the relationship between morphological and temporal branch lengths can of course vary wildly, and so one should of course be thoughtful.