Woods Hole Molecular Evolution Course 2012
# Modifying a Tree File for PAML
Nicholas R. Meyerson

This tutorial will teach you how to properly modify a Newick tree file so that you can assign a unique ω (also known as dN/dS) value to a branch or branches of interest using PAML. For details about how to set up a control file for such an analysis, please refer to exercise #3 from Joe Bielawski's PAML lab. Here, I will mainly address the tree file and how to properly interpret the output file. This tutorial will not address calculating LRT values from log likelihoods because that is covered in Joe's lab.

Example 1

We will begin by modifying the most simple tree, one that contains only 3 taxa. I will be showing the results for Gene X (whose identity is not relevant for this tutorial). The Newick code for the tree shown in Figure 1 is:

((Human,Chimpanzee),Rhesus);

First I have run the analysis to assign a single ω value to each branch in the tree. By comparing with later analyses you will see how ω changes as we isolate particular branches. Because only a single value was calculated, it represents ω for each of the three branches shown.


$$\omega_0 = 0.1889$$
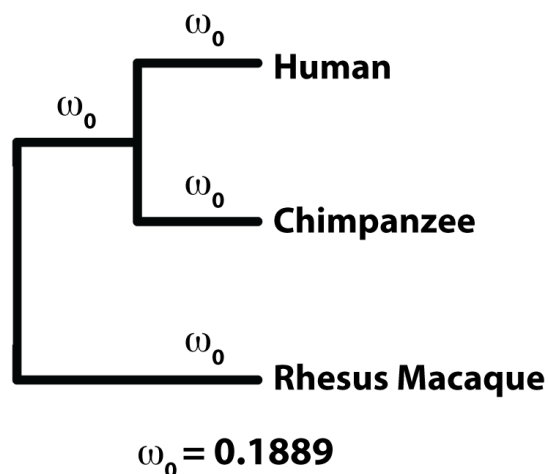
Figure 1: Phylogenetic relationship of three primate species and the result of a PAML analysis using model M0 (NSsites = 0) where a single ω value is assigned to each branch (model = 0).

Now let's tell PAML that we want it to assign a unique ω value to just the branch leading to the Human taxa. I might do this if I have a particular reason to believe that Gene X has experienced a differential evolutionary pressure that is specific to Humans. To do this we must modify the tree file so that it now reads:

((Human #1,Chimpanzee),Rhesus);

Notice the addition of '#1' immediately after 'Human' has been specified. Anytime you are interested in a terminal branch, i.e. a single species, adding this modification immediately after the taxa name will work. I now re-run the analysis (it is critical two change the control file from model = 0 to model = 2, which tells PAML that you have specified particular branches that should be isolated).

$$\omega_0 = 0.1495 \qquad \omega_1 = 0.3910$$
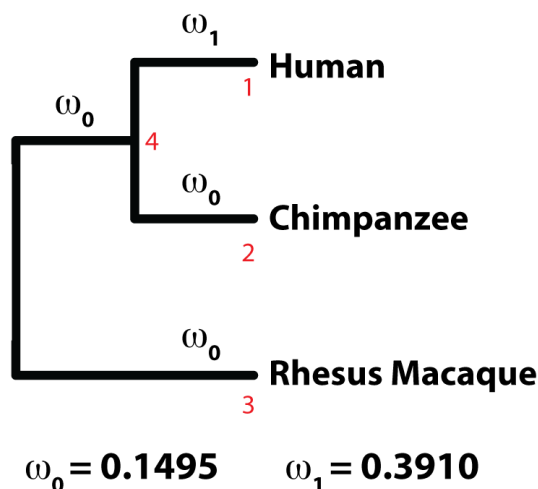
```
(1: 0.011975, 2: 0.006144, 3: 0.036960);

(Human: 0.011975, Chimp: 0.006144, Rhesus: 0.036960);

Detailed output identifying parameters

kappa (ts/tv) =  4.52636

w (dN/dS) for branches:  0.14953 0.39099

dN & dS for each branch
```

branch	t	N	S	dN/dS	dN	dS	N*dN	S*dS
4..1	0.012	362.0	139.0	0.3910	0.0028	0.0071	1.0	1.0
4..2	0.006	362.0	139.0	0.1495	0.0008	0.0053	0.3	0.7
4..3	0.037	362.0	139.0	0.1495	0.0048	0.0320	1.7	4.4

Figure 2: Results from a M0 PAML run using a modified tree file that isolates the Human branch. Shown in red text are the node labels which are found by comparing the data highlighted in red from the output file. Green highlight in the output file shows the ω values. These data are found near the end of the output file.

Figuring out where the appropriate ω values are found seems trivial for this example, however, you will see in the next example that it can be quite tricky. So let's just walk through this one. The top two boxes I have highlighted in red are kind of like a decoder. You can quickly see that Human has been assigned '1', Chimp has been assigned '2', and Rhesus has been assigned '3'. These are their tip labels. In the third red box at the bottom, it says that the ancestral node for all of the tips (1-3) has been labeled '4'.

Notice that the top two red boxes are tree files. I usually print them out and then use them to label the nodes by hand. This really helps to interpret results and create a visualization of your analysis.

Example 2a

Now we will do some analyses for a bit more complicated of a case: 5 taxa! Just like last time, we will start by assigning a single ω value to each branch on the tree. In this case there are 7 branches.

((((Human,Chimpanzee),Orangutan),Rhesus),Marmoset);
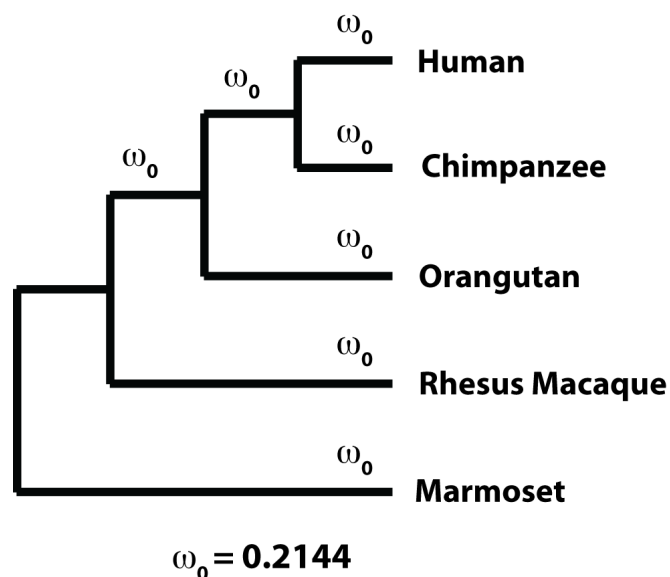
Figure 3: PAML results from M0 for 7 species where a single ω value was assigned to every branch.

Next we will isolate a single branch on the tree to answer the question, "How different is ω on the branch leading to hominoids?" In this example there are three hominoids: Human, Chimpanzee, and Orangutan. Our hypothesis might be something like, "Gene X underwent rapid evolution in the ancestral hominoid but has since been relaxed in extant species." And so to isolate the hominoid branch our tree file would be modified as such:

((((Human,Chimpanzee),Orangutan)#1,Rhesus),Marmoset);

In the case of isolating a branch leading to a clade, you can think of it like this, "I will add the '#1' just after the parenthesis that describes my clade of interest." Try your best to make sense of that, but if it is not coming clear it is OK, you will be able to double check yourself when we analyze the output.

$$\omega_0 = 0.2073 \quad \omega_1 = 0.3281$$
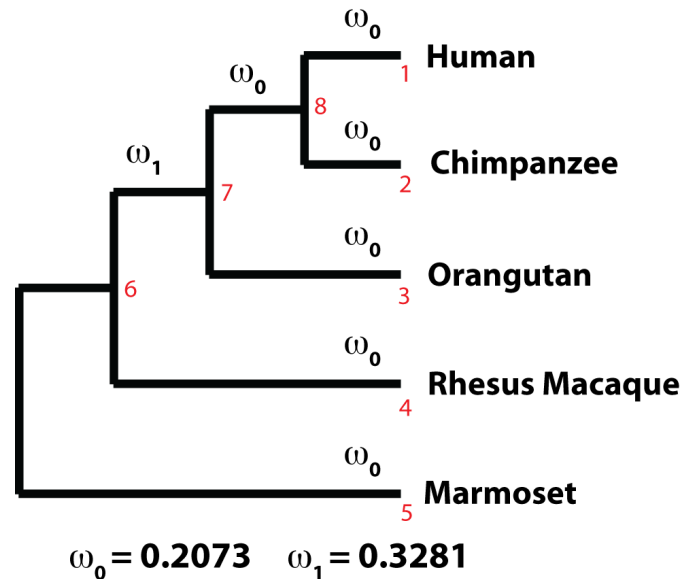
```
(((1: 0.006158, 2: 0.012046): 0.005620, 3: 0.031229): 0.012693, 4: 0.024691, 5: 0.088423);

(((Human: 0.006158, Chimp: 0.012046): 0.005620, Orangutan: 0.031229): 0.012693, Rhesus: 0.024691, Marmoset: 0.088423);

Detailed output identifying parameters

kappa (ts/tv) =  3.55962

w (dN/dS) for branches:  0.20729 0.32808

dN & dS for each branch
```

| branch	t	N	S	dN/dS	dN	dS	N*dN	S*dS
6..7	0.013	366.7	134.3	0.3281	0.0027	0.0083	1.0	1.1
7..8	0.006	366.7	134.3	0.2073	0.0009	0.0045	0.3	0.6
8..1	0.006	366.7	134.3	0.2073	0.0010	0.0049	0.4	0.7
8..2	0.012	366.7	134.3	0.2073	0.0020	0.0096	0.7	1.3
7..3	0.031	366.7	134.3	0.2073	0.0051	0.0248	1.9	3.3
6..4	0.025	366.7	134.3	0.2073	0.0041	0.0196	1.5	2.6
6..5	0.088	366.7	134.3	0.2073	0.0146	0.0702	5.3	9.4

Figure 4: Results from a M0 PAML run. Tree file was modified to isolate the branch leading to hominoids.

To figure out what the node labeling is, we will use the same strategy as before. First, use the top two red boxes in the output file to decode the taxa assignments. Next, use the third red box to decode the node assignments. In this example, our branch of interest lies between node 6 and 7, represented by '6..7'. You will also notice that the ω value for this branch is different than all the others. And so, we set up our tree file correctly! It is at this point that you can conduct quality control on your tree file modification.

Example 2b

OK, we will do one more example and then I am letting you go off into the world on your own. Let's suppose we want to set up a tree file to address the following hypothesis, "Gene X has experienced a differential evolutionary pressure on the branch leading to hominoids, and continued to experience this pressure throughout the divergence of the hominoids into their extant taxa." So now we need to add the '#1' character not only to the branch leading to hominoids, but also to the branches leading to Human-Chimp clade, Human alone, Chimpanzee alone, and Orangutan alone. A good rule to follow is that if you are trying to isolate 5 branches, which we are here, then you need to add 5 '#1' characters into your tree file. Because you are now a professional and resident expert, I will let you analyze this final example on your own.

((((Human #1,Chimpanzee #1)#1,Orangutan #1)#1,Rhesus),Marmoset);
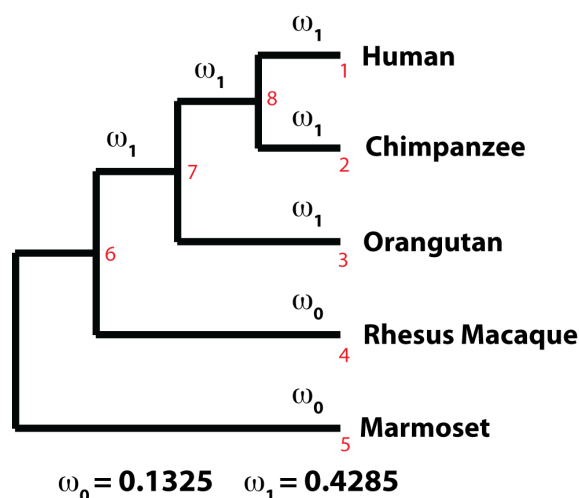
branch	t	N	S	dN/dS	dN	dS	N*dN	S*dS
6..7	0.013	366.6	134.4	0.4285	0.0031	0.0072	1.1	1.0
7..8	0.006	366.6	134.4	0.4285	0.0014	0.0032	0.5	0.4
8..1	0.006	366.6	134.4	0.4285	0.0015	0.0035	0.5	0.5
8..2	0.012	366.6	134.4	0.4285	0.0030	0.0069	1.1	0.9
7..3	0.031	366.6	134.4	0.4285	0.0076	0.0178	2.8	2.4
6..4	0.025	366.6	134.4	0.1325	0.0030	0.0228	1.1	3.1
6..5	0.089	366.6	134.4	0.1325	0.0108	0.0816	4.0	11.0

Figure 5: Results from a M0 PAML run. Tree file was modified to isolate the branch leading to hominoids and all resulting branches from then on.