Notes for codon-based Clade models

by Joseph Bielawski and Ziheng Yang, 2004
Last modified by Ziheng Yang: August 2010, June 2014

Noted added by Ziheng in August 2010: The clade models now allow more than two branch types (clades). The example here uses two clades (branch types). Also note that Weadick and Chang (2012) suggest using a modified version of M2a as the null hypothesis for testing against Clade Model C. Please see the PAML manual.

1. Contents of folder

The folder contains a sequence data file, a tree file and two control files, for the ECP-EDN gene family dataset (15 sequences) analyzed by Bielawski and Yang (2004) under the clade models. The tree is rooted, but please note that in some cases an unrooted tree may be more appropriate.

```
ECP_EDN_15.nuc
tree.txt
codeml.CladeC.ctl  (Clade Model C: model = 3  NSsites = 2)
codeml.CladeD.ctl  (Clade Model D: model = 3  NSsites = 3)
```

To run the program, do something like the following (assuming that the executable codeml is in the bin/ folder).

```
cd paml4.8\examples\CladeModelCD\Dataset1.CladeC
..\..\..\bin\codeml
```

2. Clade models C and D

```
Model C:  model = 3  NSsites = 2  (ncatG = 3 is fixed)
Model D:  model = 3  NSsites = 3 (ncatG = 3 or 2)
```

The clade models are specified in the control file codeml.ctl as above. Under model D the number of site classes is set by the user using the variable ncatG (= 2 or 3). Under model C, the number of site classes is fixed at 3 by the program so that ncatG is ignored. The current version of codeml implements a version of the Model C that is different from that described in Bielawski and Yang (2004). The new Model C is shown below.

Site class	Proportion	Clade 1	Clade 2	Clade 3 …
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$
2	$p_2 = 1 - p_0 - p_1$	ω_2	ω_3	$\omega_4 …$

The new model C is compared with the new model M1a (Yang, Wong and Nielsen, 2005) with df = 2. Note that the modified version of Model C is described in Yang, Wong and Nielsen (2005). The Bayes Empirical Bayes (BEB) procedure (Yang, Wong and Nielsen, 2005), which is recommended over the Naive Empirical Bayes procedure, is implemented for Model C but not for Model D.

3. Labelling branches in the tree file

Clade models require branches in the tree to be labelled. The tree is specified using the parenthesis notation in the tree file, and the labels are assigned to nodes in the tree. The default label (0) does not have to be specified. If there are two clades or branch types, the labels will be 0 and 1. Labels are preceeded by the symbol # or $.

For example, in the ECP-EDN tree (see figure on the right), we would like to label the ECP and EDN paralogs as two different
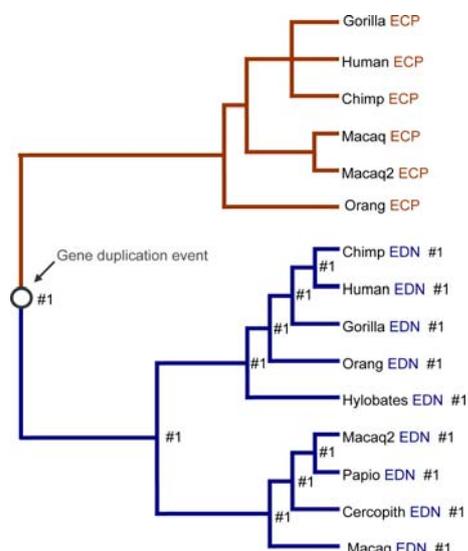
clades, so that the tree file is as follows:

```
(((Hylobates_EDN #1,(Orang_EDN #1,(Gorilla_EDN #1,(Chimp_EDN #1,Human_EDN
#1)#1)#1)#1)#1,(Macaq_EDN #1,(Cercopith_EDN #1,(Macaq2_EDN #1,Papio_EDN
#1)#1)#1)#1)#1,(Orang_ECP,((Macaq_ECP,Macaq2_ECP),
(Goril_ECP,Chimp_ECP,Human_ECP))));
```

You can also use the symbol "$" to label an entire clade. Again an integer value should follow the "$" symbol, and the number 0 is the default and does not have to be specified in the tree file. The tree below is equivalent to the tree shown above:

```
(((Hylobates_EDN,(Orang_EDN,(Gorilla_EDN,(Chimp_EDN,Human_EDN)))),(Macaq_EDN,(Cerco
pith_EDN,(Macaq2_EDN,Papio_EDN
))))$1,(Orang_ECP,((Macaq_ECP,Macaq2_ECP),(Goril_ECP,Chimp_ECP,Human_ECP))));
```

You can open the tree file tree.txt in Rod Page's (1996) TreeView program, which will display the labels.

4. Results
The results for Clade Models C and D applied to these data are shown in the tables below.

Model C (model = 3 NSsites = 2 ncatG = 3), lnL = −1702.90

	Proportion	Clade 1	Clade 2
site class 0	$p_0 = 0.36$	$\omega_0 = 0.00$	$\omega_0 = 0.00$
site class 1	$p_1 = 0.33$	$\omega_1 = 1$	$\omega_1 = 1$
site class 2	$p_2 = 0.31$	$\omega_2 = 2.23$	$\omega_3 = 0.06$

Model D (model = 3 NSsites = 3 ncatG = 2), lnL = − 1696.09

	Proportion	Clade 1	Clade 2
site class 0	$p_0 = 0.16$	$\omega_0 = 3.49$	$\omega_0 = 3.49$
site class 1	$p_1 = 0.84$	$\omega_1 = 1.26$	$\omega_1 = 0.15$

Model D (model = 3 NSsites = 3 ncatG = 3), lnL = − 1691.30

	Proportion	Clade 1	Clade 2
site class 0	$p_0 = 0.42$	$\omega_0 = 0.07$	$\omega_0 = 0.07$
site class 1	$p_1 = 0.13$	$\omega_1 = 3.76$	$\omega_1 = 3.76$
site class 2	$p_2 = 0.45$	$\omega_2 = 3.21$	$\omega_3 = 0.28$

5. Warnings and recommendations:
We found that Clade Model C often has multiple local peaks. The ECP-EDN dataset is a good example. You are advised to run the program multiple times with different initial values (change the values of kappa, omega, Small_Diff etc. in the control file) and use the set of estimates that have the highest log likelihood value. Also try to use Model C instead of Model D. Use BEB and instead of NEB.

6. References
Bielawski, J. P. and Z. Yang. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. *Journal of Structural and Functional Genomics*, 3:201-212.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* 22:1107-1118.

Test runs

Clade Model C (model = 3 NSsites = 2 ncatG = 3)

-1702.903599
tree length =   1.38484
kappa (ts/tv) = 1.93880
dN/dS for site classes (K=3)
site class	0	1	2
proportion	**0.36048**	**0.33321**	**0.30632**
background w	**0.00000**	**1.00000**	**2.22918**
foreground w	**0.00000**	**1.00000**	**0.05875**

```
 -1707.415478
tree length =   1.34867
kappa (ts/tv) =  1.89871
dN/dS for site classes (K=3)
site class          0          1          2
proportion       0.20319   0.16770   0.62910
background w      1.00000   1.00000   0.71627
foreground w      1.00000   1.00000   0.00000
```

```
-1702.955613
tree length =   1.39416
kappa (ts/tv) =  2.16114
dN/dS for site classes (K=3)
site class          0          1          2
proportion       0.43347   0.00000   0.56653
background w      0.00451   1.00000   3.44402
foreground w      0.00451   1.00000   0.97336
```

```
-1703.278810
tree length =   1.45479
kappa (ts/tv) =  2.22531
dN/dS for site classes (K=3)
site class          0          1          2
proportion       0.33508   0.59710   0.06782
background w      0.00000   1.00000   1.91106
foreground w      0.00000   1.00000   7.95363
```

Clade Model D (model = 3 NSsites = 3 ncatG = 3)

```
-1691.295786
tree length =   1.45147
kappa (ts/tv) =  2.24728
dN/dS for site classes (K=3)
site class          0          1          2
proportion       0.41833   0.13212   0.44955
background w      0.07131   3.76238   3.21545
foreground w      0.07131   3.76238   0.27716
```

```
-1702.956997
tree length =   1.45739
kappa (ts/tv) =  2.26756
dN/dS for site classes (K=3)
site class          0          1          2
proportion       0.36673   0.57440   0.05886
background w      0.00000   1.21062   1.91787
foreground w      0.00000   1.21062   9.14846
```