Notes for codon-based Clade models

by Joseph Bielawski and Ziheng Yang
Last modified: September 2005

1. Contents of folder:

The folder contains a control file, data file and tree file for two example datasets for demonstrating the clade models (Bielawski and Yang 2004). The first dataset is the ECP-EDN gene family dataset (15 sequences) used in that paper; in this case the tree is rooted. The second dataset is an expanded sample of the ECP-EDN family (17 sequences) which illustrates the use of clade models on an unrooted tree.

Dataset1:
codeml.ctl
ECP_EDN_15.nuc
tree.txt
rooted.PDF

Dataset2:
codeml2.ctl
ECP_EDN_17.nuc
tree2.txt
unrooted.PDF

2. Clade models:

The clade models are specified in the control file by setting the "model" and Nssites" variables. The number of site classes under model D is variable, and is set by the user by setting the variable ncatG (2 or 3). Under model C, the number of site classes is fixed at 3. 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
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2	$p_2 = 1 - p_0 + p_1$	ω_2	ω_3

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 also 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 only implemented for Model C . The control file setting for Models C and D are:

Model C: model = 3 NSsites = 2
Model D: model = 3 NSsites = 3

3. Labelling a tree file:

Clade models require the nodes of the tree (denoted in a text file by using parenthetical notation) are labelled to indicate the clades that will be assigned independent omega parameters. Such clades are labelled by indicating all the branches of the tree that belong to that particular clade with a symbol "#"followed by an integer number. Note the default integer number is 0 and does not have to be included in the tree.

For example, the labelled ECP-EDN gene tree for dataset 1 is as follows:

```
(((Hylobates_EDN #1,(Orang_EDN #1,(Gorilla_EDN #1,(Chimp_EDN #1,Human_EDN
#1)#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))));
```

This is a rooted tree and the labels indicate independent omega parameters for the ECP and EDN clades. You can open the tree files (tree.txt and tree2.txt) in Rod Page's (1996) TreeView program, which will display the labels.

The files rooted.PDF and unrooted.PDF illustrate the relationship between the labeled parenthetical tree file and the clade labels.

4. Example1: ECP-EDN with no outgroup

The purpose of this example is to allow you to use the control files, tree file, and data file to reproduce the results reported in Bielawski and Yang (2004) for Model D, k=3. Note that for this dataset there is one suboptimal peak in the likelihood surface (logL = -1702.96), so you should run the program several times, by using several different initial values of omega, to find the globally optimum likelihood score (log L = - 1691.30).

5. Example2: ECP-EDN with an unrooted tree

This dataset illustrates the use of Models C and D with an unrooted tree. When biologically relevant, we favour the use of an unrooted tree; model C appears to perform better with an unrooted tree. The results for dataset 2 under Model C and D are shown below.

Model C, k = 3:

LogL = -2049.506			
	Proportion	Clade 1	Clade 2
site class 0	p_0 = 0.62	ω_0 = 0.20	ω_0 = 0.20
site class 1	p_1 = 0.00	ω_1 = 1	ω_1 = 1
site class 2	p_2 = 0.38	ω_2 = 3.67	ω_3 = 1.94

Model D, k = 3:

LogL = -2046.471			
	Proportion	Clade 1	Clade 2
site class 0	p_0 = 0.49	ω_0 = 0.11	ω_0 = 0.11
site class 1	p_1 = 0.11	ω_1 = 4.17	ω_1 = 4.17
site class 2	p_2 = 0.41	ω_2 = 3.05	ω_3 = 0.95

Note: these parameter estimates are similar to those obtained for dataset 1

6. Recommendations and warnings:

- Try to use Model C instead of Model D. Use BEB and try to avoid NEB. Note that BEB is implemented for Model C only.
- Model C might not perform well with rooted trees (e.g., example 1 ECP-EDN with no outgroup, as shown in the file "rooted.PDF"); several sub-optimal peaks in likelihood appear characteristic of such cases.
- Local peaks are common. Run the program multiple times, using different initial values.

7. References:

Page, R.D.M. 1996 TreeView: An application to display phylogenetic trees on personal computers. Computer Applications in the Biological Sciences, 12, 357-358.

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. Mol Biol Evol. 22:1107-1118.

Running notes for dataset 1

Dataset 1: 15 sequences

Model D

K=3

Initial Omega	-ln L
0.001	-1691.295786
0.01	-1691.295786
0.1	-1691.295786
0.25	-1691.295786
0.5	-1691.295786
0.75	-1691.295786
1	-1691.295786
2	-1702.956997
3	-1702.956997
4	-1702.956997
5	-1702.956997
10	-1702.956997

-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
```

Model C

Initial Omega	-ln L
0.001	-1702.903599
0.01	-1702.903599
0.1	-1702.903599
0.25	-1702.955613
0.5	-1707.415478
0.75	-1702.955613
1	-1703.278810
2	-1703.278810
3	-1703.278810
4	-1703.278810
5	-1703.278810
10	-1703.278810

-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
```

Running notes for dataset 2

Dataset 2: 17 sequences

Model D

K=3

Initial Omega	-ln L
0.001	-2046.471142
0.01	-2046.471142
0.1	-2046.471142
0.25	-2046.471142
0.5	-2046.471142
0.75	-2046.471142
1	-2046.471142
2	-2046.471142
3	-2049.506364
4	-2049.506364
5	-2050.832612
10	-2050.827965

-2046.471142

```
tree length =   1.98422
kappa (ts/tv) =  2.16857
dN/dS for site classes (K=3)
site class         0         1         2
proportion       0.48590  0.10718  0.40692
background w     0.10557  4.17186  3.05340
foreground w     0.10557  4.17186  0.95081
```

-2049.506364

```
tree length =   1.93446
kappa (ts/tv) =  2.12727
dN/dS for site classes (K=3)
site class         0         1         2
proportion       0.61760  0.00087  0.38153
background w     0.20165  0.20165  3.66737
foreground w     0.20165  0.20165  1.94075
```

-2050.827965

```
tree length =   1.96104
kappa (ts/tv) =  2.15037
dN/dS for site classes (K=3)
site class         0         1         2
proportion       0.60568  0.34204  0.05228
background w     0.19913  2.05976  3.53599
foreground w     0.19913  2.05976  5.35787
```

Model C

Initial Omega	-ln L
0.001	-2062.055212
0.01	-2062.055212
0.1	-2062.055212
0.25	-2062.055212
0.5	-2062.055212
0.75	-2049.506364
1	-2049.506364
2	-2049.506364
3	-2049.506364
4	-2049.506364
5	-2049.506364
10	-2049.506364

-2049.506364

```
tree length =   1.93446
kappa (ts/tv) =  2.12726
dN/dS for site classes (K=3)
site class         0         1         2
proportion       0.61847  0.00000  0.38153
background w     0.20165  1.00000  3.66737
foreground w     0.20165  1.00000  1.94075
```

-2062.055212

```
tree length =   1.87336
kappa (ts/tv) =  1.85303
dN/dS for site classes (K=3)
site class         0         1         2
proportion       0.00000  0.54316  0.45684
background w     0.12624  1.00000  0.18562
foreground w     0.12624  1.00000  0.00000
```