

# Inferring Cases of Lineage-Specific & Site-Specific Adaptive Evolution

## An Introduction to Ziheng Yang's PAML.

David Lynn M.Sc., Ph.D., Department of Molecular Biology & Biochemistry,  
Simon Fraser University, Vancouver, B.C., Canada.

## Conservation Genetics Data Analysis Course

11-16 September 2007, Flathead Lake Biological Station, Montana, USA

### ***Why Study Adaptive Evolution?***

- The detection of adaptive evolution at the molecular level is of interest not only as an insight into the process of evolution but also because of its functional implications for genes of interest.
- From an evolutionary biology perspective – much interest in detecting whether most mutations are deleterious, advantageous or neutral – The Neutral Theory.
- Identification of selected loci provide insight into the events that have shaped a species' evolution & can indicate which genes have been particularly important in the evolution of a species e.g. positive selection in human CCR5 suggested to have been driven by selective agents e.g. bubonic plague or smallpox.
- The capacity of a species or population to respond to and survive novel infectious disease challenge is one of the most significant selective forces shaping genetic diversity.
- By screening for selective signatures associated with immunity or disease susceptibility, we may be able to identify those genes that have been of critical importance to the development of disease resistance.
- Can indicate which amino acid sites/domains are functionally important in a molecule. E.g. MHC – antigen recognition domain subject to positive selection (Hughes & Nei, 1988); TLR4 – sites subject to PS suggest location of ligand binding domain (White *et al.*, 2003); Alpha-defensins – PS in mature antimicrobial peptide (Lynn *et al.*, 2004).

### ***Detecting Positive Selection using Molecular Sequence Data:***

- Widely used method to detect adaptive evolution → accelerated rate of  $d_N/d_S$ 
  - $d_N$  = # nonsynonymous (protein changing) substitutions per nonsynonymous site
  - $d_S$  = # synonymous substitutions per synonymous site
- $d_S$  (silent subs) are assumed to be neutral.
  - But see recent papers on selection acting on synonymous sites
- If amino acid changes selectively neutral → fixed at the same rate as synonymous mutations and  $\omega = 1$ .
- If nonsynonymous mutations are slightly deleterious, then  $\omega < 1$ .
- If amino acid changes selectively advantageous → fixed at a higher rate and  $\omega > 1$ .

### ***Ziheng Yang's PAML Program:***

- The most widely used and accepted set of methods to detect positive selection.
- <http://abacus.gene.ucl.ac.uk/software/paml.html>
- Available for 

### ***PAML Models:***

- ***Models to detect positive selection acting on***
  - Particular branches/lineages of a phylogeny / certain genes in particular species (lineage/branch-specific models).
  - Particular codon (amino acid) sites (site-specific models).
  - On both simultaneously (branch-site model).

### ***Lineage-specific / Branch Models:***

- *Branch models* – allow  $\omega$  vary among branches of phylogeny & are used to detect PS acting on particular lineages.
- Model = 0 → *one-ratio model* → assumes an equal  $\omega$ -ratio for all branches in the phylogeny (The null model).
- Model = 1 → *free-ratios model* → assumes an independent  $\omega$ -ratio for each branch.
- These models can be compared by Likelihood Ratio Test (LRT) → compares *lnL* values for each model and tests if they are significantly different.
- P value determined → Twice the log-likelihood difference between the two models compared to a  $\chi^2$  distribution with d.f. = difference in # parameters between one-ratio and free-ratios model (Yang 1998; Yang and Nielsen 1998).
- Where *free-ratios* model significantly favored → can conclude that there is variable selective pressure in the phylogeny.
- If some branches have  $\omega > 1$  → weak evidence of PS.
- Problem: free-ratios model is very parameter-rich → use of a more specific model is preferred.
- Model = 2 → “two” ratios model → allows you to have two or more  $\omega$  ratios. User must specify how many ratios and which branches should have which rates in the tree file by using branch labels.
- Allows testing of specific hypotheses e.g. PS acting on genes after duplication, PS acting on particular species etc.
- LRT again used to compared two-ratios and one-ratio model.
- Where *two-ratios* model significantly favored & branches tested have  $\omega > 1$  → can conclude that there is evidence of PS on those lineages.
- To confirm  $\omega$  significantly  $> 1$  → compare *two-ratios* model to same model but with  $\omega$  fixed =1.

**Detecting Positive Selection on CD2 – A step by step example - Lynn et al., Genetics 2005:**

**Requirements for PAML Analysis:**

- A coding DNA sequence alignment in PAML format.
- A treefile in newick format.
- codeml.ctl parameter file

**Coding DNA sequence alignment in PAML format:**

- Coding DNA sequences should never be aligned at the DNA level as alignment programs may insert gaps in codons and can end up with out-of-frame alignment.
- First align CD2 protein sequences using T-coffee program (or similar) <http://www.tcoffee.org/>
- Copygaps.pl → Uses protein alignment as template to generate DNA alignment. Every gap in protein aln → 3 gaps in DNA aln.

# of sequences  
length of aln

10 1086  
CD2\_baboon

```

ATGAGCTTCCATGTAAATTTGTAGCCAGCTTCCTTCTAATTTTCCACGTTTCTTCCAAA
GGTGCAGTCTCCAAAGAGATTAGGAATGCTTTGGAAACCTGGGGAGCGCTGGGTCAGGAC
ATCGACTTGGACATTTCCTAGTTTTTCAAATGAGTGATGATATTGATGATATAAAATGGGAG
AAAACTTCAGACAAGAAAAAGATTGCAAAATTCAGAAAAGAGAAAGAGACTTTCGAGGAA
AAAGATGCATATAAGCTATTTAAAAACGGAACTCTGAAAATTAAGCAT---CTGAAGATC
CATGATCAGGATAGCTACAAAGGTATCAATATACGATACAAAAGGAAAAAATGTTGGAA
AAAAATTGATTTGAAGATTCAAGAGAGGGTCTCAGAACCAAAAGATCTCTGGACTTGT
ATCAACACAACCCTGACCTGTGAAGTAATGAAATGGAAC TGACCCCGAATTAACCTGTAT
CAAGATGGGAAACATCTAAAACCTTCT---CAGAGGGTCATCACACAAAGTGGACCAACC
    
```

- PAML requires that a phylogeny of the sequences to be analyzed is constructed using a 3rd party program.

# of sequences  
# of trees

10 1

```

((CD2_rat,CD2_mouse),((CD2_pig,CD2_cow),CD2_cat,CD2_horse),((CD2_human,CD2_chimp),(CD2_baboon,CD2_rhesusmonkey)));
    
```

- For model =2 tests:

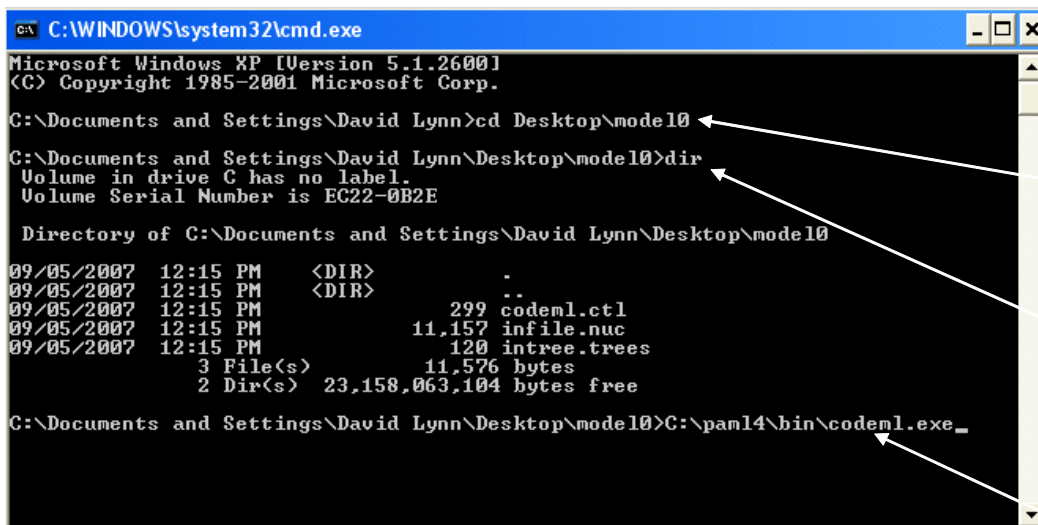
```

10 1
((CD2_rat,CD2_mouse),((CD2_pig #1),CD2_cow #1),CD2_cat,CD2_horse),((CD2_human,CD2_chimp),(CD2_baboon,CD2_rhesusmonkey)));
    
```

Specifies that cow and pig have the same  $\omega$  rate and other branches have different independent rate → allows test of PS specifically on these lineages.

### Running the Null Model – The one-ratio model:

- Edit the codeml.ctl file so that model = 0
- Create a directory (“model0” for example) to run analysis. Copy the codeml.ctl file, the “infile.nuc” file (coding sequence alignment in PAML format) & the treefile “intree.trees” into this directory.
- These files can be downloaded from <http://www.pathogenomics.ca/~dlynn/congen/>
- PAML only runs from command line. To open command prompt in windows XP → start → Run → type “cmd”
- Running *Codeml – PAML Program that Implements Models*



```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\David Lynn>cd Desktop\model0
C:\Documents and Settings\David Lynn\Desktop\model0>dir
Volume in drive C has no label.
Volume Serial Number is EC22-0B2E

Directory of C:\Documents and Settings\David Lynn\Desktop\model0

09/05/2007  12:15 PM    <DIR>          .
09/05/2007  12:15 PM    <DIR>          ..
09/05/2007  12:15 PM                299  codeml.ctl
09/05/2007  12:15 PM            11,157  infile.nuc
09/05/2007  12:15 PM                120  intree.trees
               3 File(s)              11,576 bytes
               2 Dir(s)  23,158,063,104 bytes free

C:\Documents and Settings\David Lynn\Desktop\model0>C:\paml4\bin\codeml.exe _
```

“cd Desktop\model0” → changes directory to model0

“dir” → lists files in model0 directory

“C:\paml4\bin\codeml.exe” → reads in codeml.ctl parameters and runs *codeml*.

Note: The exact paths shown here may differ on your own PC.

## The Output Files:

- Several different output files produced e.g. rst, rst1, rub, Inf, 2NG.ds, 2NG.dn, 2NG.t, mlc
- Main output file = mlc

Codon position x base (3x4) table, overall

```
position 1:  T:0.13578  C:0.25199  A:0.34281  G:0.26942
position 2:  T:0.24495  C:0.22722  A:0.34098  G:0.18685
position 3:  T:0.24893  C:0.25566  A:0.25138  G:0.24404
Average     T:0.20989  C:0.24495  A:0.31172  G:0.23344
```

np = # of paramters

TREE # 1: ((9, 7), (18, 4), 2, 5), ((6, 3), (1, 10))) MP score: 879

```
lnL(ntime: 16 np: 18): -5712.563755 +0.000000
11..12 12..9 12..7 11..13 13..14 14..8 14..4 13..2 13..5 11..15 15..16 16..6 16..3 15..17 17..
0.814628 0.165759 0.273220 0.162777 0.260052 0.386361 0.537402 0.671473 0.443011 0.282663 0.039155 0.012715 0.002755 0.063086 0.0
```

lnL value for one-ratio model →  
record this!

Note: Branch length is defined as number of nucleotide substitutions per codon (not per nucleotide site).

tree length = 4.13055

((9: 0.165759, 7: 0.273220): 0.814628, ((8: 0.386361, 4: 0.537402): 0.260052, 2: 0.671473, 5: 0.443011): 0.162777, ((6: 0.012715, (

((CD2\_rat: 0.165759, CD2\_mouse: 0.273220): 0.814628, ((CD2\_pig: 0.386361, CD2\_cow: 0.537402): 0.260052, CD2\_cat: 0.671473, CD2\_hor:

Detailed output identifying parameters

```
kappa (ts/tv) = 2.58330
omega (dN/dS) = 0.78529
```

Estimated transition/transversion ratio

dN & dS for each branch

branch	t	N	S	dN/dS	dN	dS	N*dN	S*dS
11..12	0.815	714.9	266.1	0.7853	0.2528	0.3219	180.7	85.7
12..9	0.166	714.9	266.1	0.7853	0.0514	0.0655	36.8	17.4
12..7	0.273	714.9	266.1	0.7853	0.0848	0.1080	60.6	28.7
11..13	0.163	714.9	266.1	0.7853	0.0505	0.0643	36.1	17.1
13..14	0.260	714.9	266.1	0.7853	0.0807	0.1028	57.7	27.3
14..8	0.386	714.9	266.1	0.7853	0.1199	0.1527	85.7	40.6
14..4	0.537	714.9	266.1	0.7853	0.1668	0.2124	119.2	56.5
13..2	0.671	714.9	266.1	0.7853	0.2084	0.2653	149.0	70.6
13..5	0.443	714.9	266.1	0.7853	0.1375	0.1751	98.3	46.6
11..15	0.283	714.9	266.1	0.7853	0.0877	0.1117	62.7	29.7
15..16	0.039	714.9	266.1	0.7853	0.0122	0.0155	8.7	4.1
16..6	0.013	714.9	266.1	0.7853	0.0039	0.0050	2.8	1.3
16..3	0.003	714.9	266.1	0.7853	0.0009	0.0011	0.6	0.3
15..17	0.063	714.9	266.1	0.7853	0.0196	0.0249	14.0	6.6
17..1	0.009	714.9	266.1	0.7853	0.0029	0.0037	2.1	1.0
17..10	0.006	714.9	266.1	0.7853	0.0019	0.0024	1.4	0.6

one-ratio model assumes same  $\omega$  ratio on all lineages

```
tree length for dN: 1.28178
tree length for dS: 1.63224
```

## Running the free-ratios model:

- Edit the codeml.ctl file so that model = 1
- Create a directory (“model1” for example) to run analysis. Copy the codeml.ctl file, the “infile.nuc” file (coding sequence alignment in PAML format) & the treefile “intree.trees” into this directory.
- Run codeml as before, but in the model1 directory.
- Record free-ratios model lnL from mlc output file.

TREE # 1: ((9, 7), ((8, 4), 2, 5), ((6, 3), (1, 10))); MP score: 875  
 check convergence .  
 lnL (ntime: 16 np: 33): -5698.412673 +0.000000  
 11..12 12..9 12..7 11..13 13..14 14..8 14..4 13..2 13..5 11..15 15..16  
 0.834576 0.167658 0.272873 0.165458 0.259631 0.383114 0.532855 0.668739 0.441724 0.285256 0.038977

Note: Branch length is defined as number of nucleotide substitutions per codon (not per nucleotide)  
 tree length = 4.14616

((9: 0.167658, 7: 0.272873): 0.834576, ((8: 0.383114, 4: 0.532855): 0.259631, 2: 0.668739, 5: 0.441  
 ((CD2\_rat: 0.167658, CD2\_mouse: 0.272873): 0.834576, ((CD2\_pig: 0.383114, CD2\_cow: 0.532855): 0.259

Detailed output identifying parameters  
 kappa (ts/tv) = 2.59070  
 dN & dS for each branch

branch	t	N	S	dN/dS	dN	dS	N*dN	S*dS
11..12	0.835	714.8	266.2	0.5359	0.2252	0.4203	161.0	111.9
12..9	0.168	714.8	266.2	0.6296	0.0482	0.0765	34.4	20.4
12..7	0.273	714.8	266.2	0.6753	0.0805	0.1191	57.5	31.7
11..13	0.165	714.8	266.2	0.5836	0.0462	0.0792	33.0	21.1
13..14	0.260	714.8	266.2	1.2062	0.0908	0.0752	64.9	20.0
14..8	0.383	714.8	266.2	1.0849	0.1305	0.1203	93.3	32.0
14..4	0.533	714.8	266.2	1.4149	0.1930	0.1364	137.9	36.3
13..2	0.669	714.8	266.2	0.8972	0.2162	0.2410	154.5	64.2
13..5	0.442	714.8	266.2	0.7793	0.1367	0.1755	97.7	46.7
11..15	0.285	714.8	266.2	0.7569	0.0875	0.1156	62.5	30.8
15..16	0.039	714.8	266.2	0.4120	0.0094	0.0227	6.7	6.1
16..6	0.012	714.8	266.2	0.3737	0.0029	0.0077	2.0	2.0
16..3	0.003	714.8	266.2	999.0000	0.0014	0.0000	1.0	0.0
15..17	0.064	714.8	266.2	0.2283	0.0112	0.0490	8.0	13.0
17..1	0.009	714.8	266.2	0.1838	0.0014	0.0077	1.0	2.0
17..10	0.006	714.8	266.2	999.0000	0.0028	0.0000	2.0	0.0

tree length for dN: 1.28369  
 tree length for dS: 1.64612

*lnL value for free-ratio model → record this!*

*np = # of paramters*

*free-ratios model assumes independent  $\omega$  ratio on all lineages. Note: lineages where  $\omega > 1$ .*

### ***Likelihood Ratio Test (LRT) Comparing One-ratio Model to Free-Ratios Model:***

- P value determined → Twice the log-likelihood difference between the two models compared to a  $\chi^2$  distribution with d.f. = difference in # of parameters between one-ratio and free-ratios models.
- $2\delta l = \text{abs}(2 \times (-5712.56 - -5698.41)) = 28.3$ ;  $df = (33 - 18) = 15$ .
- Type C:\paml4\bin\chi2.exe → display  $\chi^2$  table.
- $P < 0.05$  → free-ratios model favored.

### ***Running the “two”-ratios model:***

- Note: It is statistically incorrect to use free-ratios model to develop hypotheses and then use two-ratios model to test them.
- Edit the codeml.ctl file so that model = 2.
- Edit “intree.trees” to specify which branches to test.
- Create a directory (“model2” for example) to run analysis. Copy the codeml.ctl file, the “infile.nuc” file (coding sequence alignment in PAML format) & the treefile “intree.trees” into this directory.
- Run codeml as before, but in the model2 directory.

### ***Likelihood Ratio Test (LRT) Comparing One-ratio Model to Two-Ratios Model:***

- P value determined → Twice the log-likelihood difference between the two models compared to a  $\chi^2$  distribution with d.f. = difference in number of parameters between models.
- $2\delta l = \text{abs}(2 \times (-5712.56 - -5706.99)) = 11.14$ ;  $df = 1$ 
  - $P < 0.001$  → two-ratios model is significantly favored.
  - $d_N/d_S > 1$  on lineages of interest (cow and pig) is supportive of PS on those lineages.
- Compare previous two-ratios model to same model but where  $\omega$  is fixed = 1.

- Considered most stringent test for positive selection → but very conservative and lacks power.

### ***Testing for Evidence of Site-Specific Positive Selection:***

- Branch models require  $\omega > 1$  over whole sequence → conservative → PS tends to act only on specific amino acids or domains.
- Site models allow the  $\omega$  ratio to vary among sites (among codons or amino acids in the protein) (Nielsen and Yang 1998; Yang et al. 2000).
- Main recommended models
  - LRT M1a Vs M2a.
  - LRT M7 (beta) and M8 (beta& $\omega$ ).

### ***Identifying Which Sites are Subject to Positive Selection:***

- Posterior Bayesian probabilities of site classes calculated for each amino acid site (Nielsen & Yang 1998).
- Posterior Bayesian probabilities used to be calculated by naïve empirical Bayes (NEB) method.
  - In small datasets or when sequences very similar → estimates unreliable.
  - PAML now implements improved Bayes Empirical Bayes (BEB) (Yang 2005) → use this one!
  - More sequences → better accuracy & power.
- If the  $\omega$ -ratios for some site classes are  $>1$  (from M2a or M8).
  - Sites with high posterior probabilities ( $>0.95$ ) for those classes are likely to be under positive selection.

### ***Testing for Evidence of Site-Specific Positive Selection on the Mammalian CD2 Gene:***

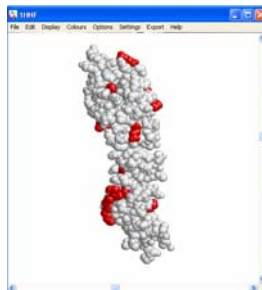
#### ***Running the site-specific models M1a, M2a, M7 & M8:***

- Edit the codeml.ctl file so that model = 0 & NSsites = 1 2 7 8
- Create a directory (“sites” for example) to run analysis. Copy the codeml.ctl file, the “infile.nuc” file (coding sequence alignment in PAML format) & the treefile “intree.trees” into this directory.
- Run codeml as before, but in the “sites” directory.

- mlc file contains the lnL values from M1a, M2a, M7 and M8.
- Construct LRTs of
  - M1a vs M2a (test of positive selection).
  - $LRT = 2\delta l = \text{abs}(2 \times (-5631.80 - -5615.92)) = 31.76; df = 2.$
  - M7 vs M8 (test of positive selection).
  - $LRT = 2\delta l = \text{abs}(2 \times (-5641.42 - -5616.05)) = 50.74; df = 2.$
- Both models (M2a & M8) which have site classes with  $\omega > 1$  are significantly favored → particular codon (amino acid) sites are subject to adaptive evolution.
- Sites with high posterior probabilities (BEB > 0.95) of belonging to site classes with  $\omega > 1$  → codon sites subject to PS.

***Structure for Human CD2 Extracellular Domain is Available – Plot Sites Subject to PS on Structure:***

- If a 3D structure is available for your protein of interest you can download a pdb structure file from the *Protein Data Bank*  
<http://www.rcsb.org/pdb/home/home.do>
- E.g. 1HNF.pdb = pdb id for Human CD2 extracellular domain.
- Download & install rasmol from <http://www.openrasmol.org/> - Open 1HNF.pdb file in rasmol.
- Open 1HNF.pdb file as text → match up position numbering in mlc file to numbering of residues in .pdb file.
- Use rasmol command line to select residues → “select LEU7, LYS51, LYS55” etc → “color red”.



### ***Power & Accuracy of LRT:***

- $\chi^2$  distribution does not apply when sample sizes are small.
- $\chi^2$  makes LRT conservative (type I error rate < alpha).
- Power is affected by (i) sequence length (ii) sequence divergence (low power for highly similar or highly divergent seqs) (iii) number of lineages, and (iv) strength of positive selection (Anisimova 2001).
- The most efficient way to increase power is to add lineages
  - Power low in datasets of 5/6 seqs but ~100% for 17 seqs.
- High recombination rates can cause high false positive rates in LRT
  - Empirical Bayes less affected (Anisimova 2003).

### ***Branch-site models:***

- Allow  $\omega$  ratio vary both among sites and among lineages.
  - attempt to detect PS that affects only a few sites along a few lineages.
  - Original branch-site model (Yang & Nielsen 2002) unrealistic & found to high false positive rate (20%-70%) (Zhang 2004).
  - New model now implemented that is more accurate (Zhang 2005).
- Model A (Model =2; NSsites = 1) → assumes variable selective pressure on sites on particular specified branches.
  - Compared to site model 1A (NSsites = 1) variable selective pressure on sites on all branches.
  - Comparison of Model A to Model A  $\omega$  fixed = 1 → most stringent branch-site test of PS.
- Bayes Empirical Bayes (BEB) output for sites should be used.