Molecular evolution analyses using PAML

Kawther Abdilleh
MOCB graduate student
Machado Lab

Phylogenetic Analysis by Maximum Likelihood (PAML)

- One of the most widely used programs to analyze molecular evolution data

- Suite of programs that uses maximum likelihood to infer evol. relationships of genes or proteins

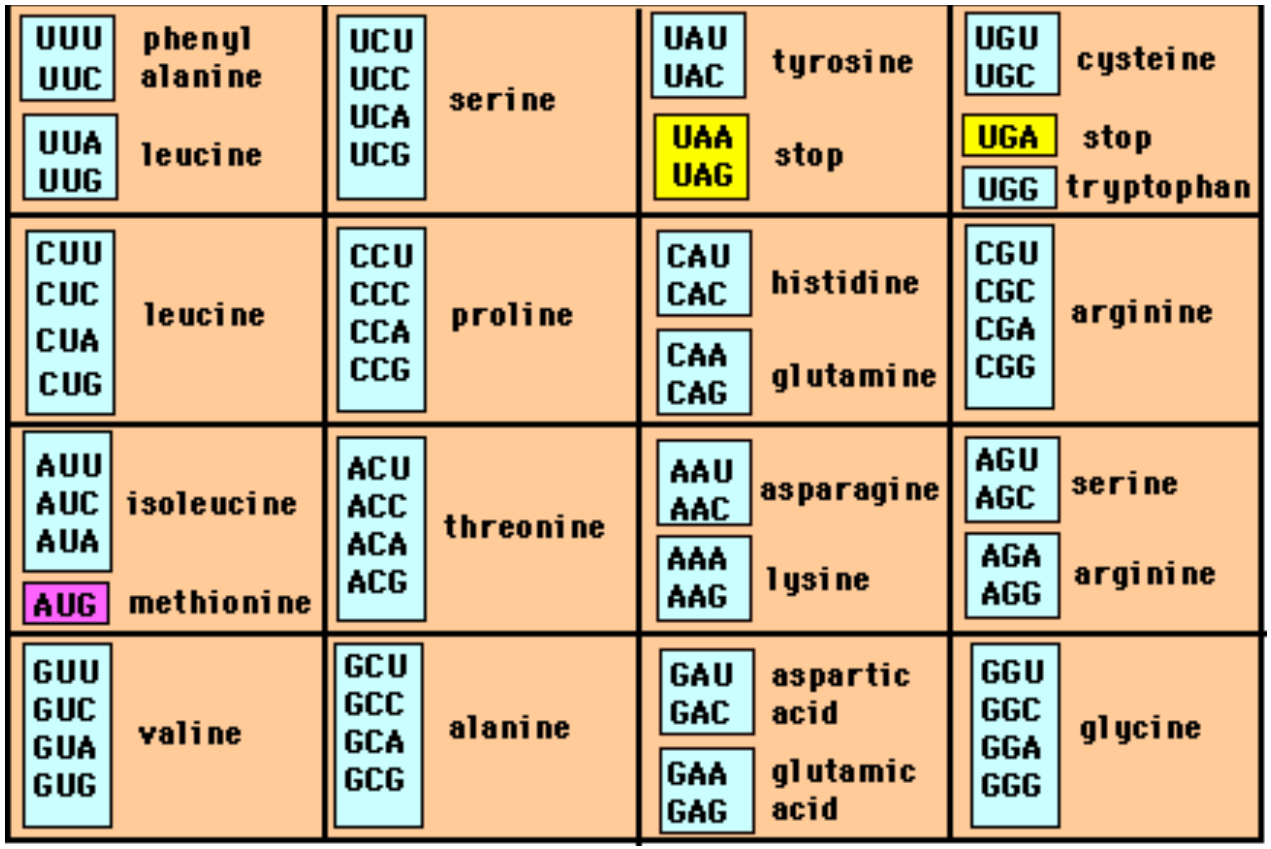
baseml
basemlg
codeml
evolver
yn00
chi2
pamp
mcmctree

Types of analyses you can do with PAML…

- Determining rates of sequence (nuc &aa) evolution
- Reconstruction of ancestral nuc or AA sequences
- Detecting positive selection

Molecular evolution review…
Synonymous vs non-Synonymous mutations

- Synonymous (d_s) – no change to amino acid
- Non-Synonymous (d_n) – changes amino acid



Molecular evolution review…

Using the ratio of non-synonymous to synonymous substitutions can measure selection @ the codon level

- Purifying selection – d_N/d_S (ω) < 1
  (genes highly conserved between species)
- Neutral evolution – d_N/d_S (ω) ~ 1
  (pseudogenes)
- Positive selection – d_N/d_S (ω) > 1
  (genes involved in immunity/immune response)

- PAML models can detect positive selection
  - Specific branches/lineages (branch sites models)
  - Specific codon sites (site-specific models)
  - Or both (branch-site model).

Branch Models

- Allow ω vary among branches of phylogeny & are used to detect PS acting on specific lineages.

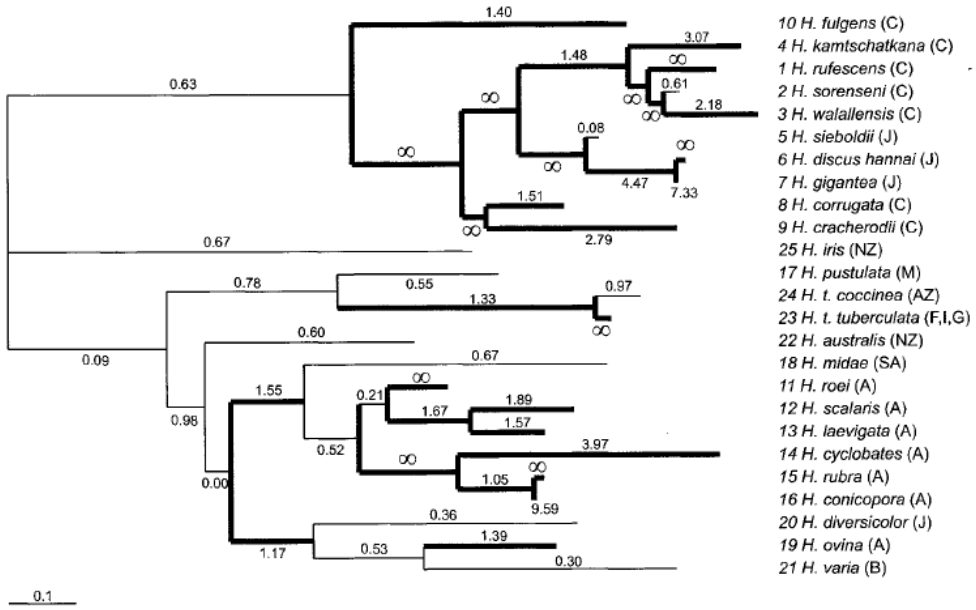
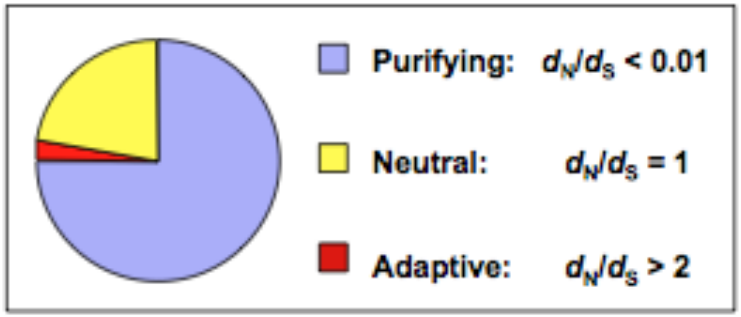


FIG. 1.—Phylogeny for the sperm lysin genes from 25 abalone (genus *Haliotis*) species. Letters in parentheses indicate the collecting sites: Australia (A), Azores (AZ), Borneo (B), California (C), France (F), Greece (G), Italy (I), Japan (J), Madagascar (M), and New Zealand (NZ). Analysis in this paper used the unrooted topology only. Branches are drawn in proportion to their lengths, defined as the expected number of nucleotide substitutions per codon. Maximum-likelihood estimates of branch lengths were obtained under the "free-ratios" model, which assumes an independent ω ratio (d_N/d_S) for each branch in the tree. Estimates of the ω ratios under that model are shown along branches, and branches for which the estimated ω ratios are >1 are drawn in thick lines.

Yang *et al.*, MBE 2000

Selection pressure varies **among sites**

ATG CTT GTG CTA ... CGC TAA



If we average over sites, we do NOT detect positive selection;

$\omega = 0.31$

J. Bielawski

Example using Codeml (codon based model of seq. evol.)

- Detect positive selection on certain sites/branches or both
- Sophisticated model that corrects for unequal transition/ transversion ratios, codon usage bias and GC content.

A model of codon substitution

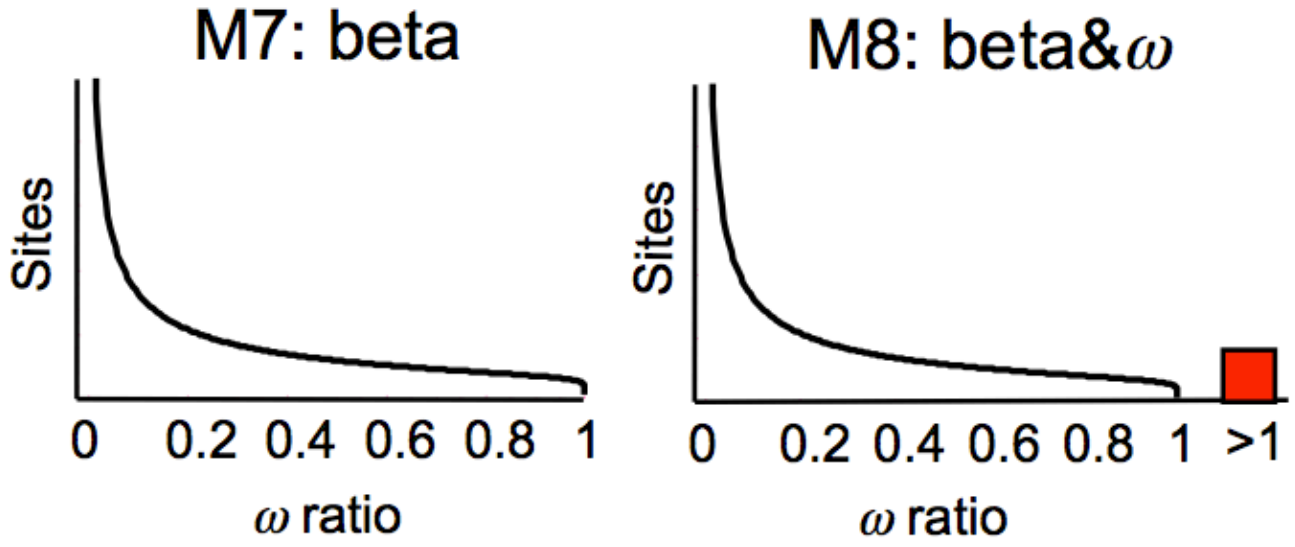
The codon is considered the unit of evolution. The substitution rate from codons i to j ($i \neq j$) is given as:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition.} \end{cases}$$

Parameter κ is the transition/transversion rate ratio, π_j is the equilibrium frequency of codon j and ω ($= d_N/d_S$) measures the selective pressure on the protein.

H_0: Beta distributed variable selective pressure (M7)
H_1: Beta plus positive selection (M8)

Compare $2\Delta l = 2(l_1 - l_0)$ with a χ^2 distribution



What you need to run PAML

- PAML (obviously..)
- Sequence files in PHYLIP format
- Tree file in parenthetical notation
  - ((((Human:0.1, Chimpanzee:0.2):0.8, Gorilla:0.3):0.7, Orangutan:0.4, Gibbon:0.5);
  - (6,(5,(4,(1,(2,3)))));
- Control file (.ctl)

Interweaved phylip format

```
 3 384
CYS1_DICDI    -----MKVIL LFVLAVFTVF VSS------- --------RG IPPEEQ---- --------SQ
ALEU_HORVU    MAHARVLLLA LAVLATAAVA VASSSSFADS NPIRPVTDRA ASTLESAVLG ALGRTRHALR
CATH_HUMAN    ------MWAT LPLLCAGAWL LGV------- -PVCGAAELS VNSLEK---- --------FH

              FLEFQDKFNK KY-SHEEYLE RFEIFKSNLG KIEELNLIAI NHKADTKFGV NKFADLSSDE
              FARFAVRYGK SYESAAEVRR RFRIFSESLE EVRSTN---- RKGLPYRLGI NRFSDMSWEE
              FKSWMSKHRK TY-STEEYHH RLQTFASNWR KINAHN---- NGNHTFKMAL NQFSDMSFAE

              FKNYYLNNKE AIFTDDLPVA DYLDDEFINS IPTAFDWRTR G-AVTPVKNQ GQCGSCWSFS
              FQATRL-GAA QTCSATLAGN HLMRDA--AA LPETKDWRED G-IVSPVKNQ AHCGSCWTFS
              IKHKYLWSEP QNCSAT--KS NYLRGT--GP YPPSVDWRKK GNFVSPVKNQ GACGSCWTFS

              TTGNVEGQHF ISQNKLVSLS EQNLVDCDHE CMEYEGEEAC DEGCNGGLQP NAYNYIIKNG
              TTGALEAAYT QATGKNISLS EQQLVDCAGG FNNF------ --GCNGGLPS QAFEYIKYNG
              TTGALESAIA IATGKMLSLA EQQLVDCAQD FNNY------ --GCQGGLPS QAFEYILYNK

              GIQTESSYPY TAETGTQCNF NSANIGAKIS NFTMIP-KNE TVMAGYIVST GPLAIAADAV
              GIDTEESYPY KGVNGV-CHY KAENAAVQVL DSVNITLNAE DELKNAVGLV RPVSVAFQVI
              GIMGEDTYPY QGKDGY-CKF QPGKAIGFVK DVANITIYDE EAMVEAVALY NPVSFAFEVT

              E-WQFYIGGV F-DIPCN--P NSLDHGILIV GYSAKNTIFR KNMPYWIVKN SWGADWGEQG
              DGFRQYKSGV YTSDHCGTTP DDVNHAVLAV GYGVENGV-- ---PYWLIKN SWGADWGDNG
              QDFMMYRTGI YSSTSCHKTP DKVNHAVLAV GYGEKNGI-- ---PYWIVKN SWGPQWGMNG

              YIYLRRGKNT CGVSNFVSTS II--
              YFKMEMGKNM CAIATCASYP VVAA
              YFLIERGKNM CGLAACASYP IPLV
```

Sequential Phylip format

```
 3 384
CYS1_DICDI-----MKVILLFVLAVFTVFVSS---------------RGIPPEEQ-----------SQFLEFQDKFNKKY-SHEEY
LERFEIFKSNLGKIEELNLIAINHKADTKFGVNKFADLSSDEFKNYYLNNKEAIFTDDLPVADYLDDEFINSIPTAFDWRTRG-AVTP
VKNQGQCGSCWSFSTTGNVEGQHFISQNKLVSLSEQNLVDCDHECMEYEGEEACDEGCNGGLQPNAYNYIIKNGGIQTESSYPYTAET
GTQCNFNSANIGAKISNFTMIP-KNETVMAGYIVSTGPLAIAADAVE-WQFYIGGVF-DIPCN--PNSLDHGILIVGYSAKNTIFRKN
MPYWIVKNSWGADWGEQGYIYLRRGKNTCGVSNFVSTSII--
ALEU_HORVUMAHARVLLLALAVLATAAVAVASSSSFADSNPIRPVTDRAASTLESAVLGALGRTRHALRFARFAVRYGKSYESAAEVR
RRFRIFSESLEEVRSTN----RKGLPYRLGINRFSDMSWEEFQATRL-GAAQTCSATLAGNHLMRDA--AALPETKDWREDG-IVSPVK
NQAHCGSCWTFSTTGALEAAYTQATGKNISLSEQQLVDCAGGFNNF--------GCNGGLPSQAFEYIKYNGGIDTEESYPYKGVNGV-
CHYKAENAAVQVLDSVNITLNAEDELKNAVGLVRPVSVAFQVIDGFRQYKSGVYTSDHCGTTPDDVNHAVLAVGYGVENGV-----PYW
LIKNSWGADWGDNGYFKMEMGKNMCAIATCASYPVVAA
CATH_HUMAN------MWATLPLLCAGAWLLGV--------PVCGAAELSVNSLEK-----------FHFKSWMSKHRKTY-STEEYH
HRLQTFASNWRKINAHN----NGNHTFKMALNQFSDMSFAEIKHKYLWSEPQNCSAT--KSNYLRGT--GPYPPSVDWRKKGNFVSPVK
NQGACGSCWTFSTTGALESAIAIATGKMLSLAEQQLVDCAQDFNNY--------GCQGGLPSQAFEYILYNKGIMGEDTYPYQGKDGY-
CKFQPGKAIGFVKDVANITIYDEEAMVEAVALYNPVSFAFEVTQDFMMYRTGIYSSTSCHKTPDKVNHAVLAVGYGEKNGI-----PYW
IVKNSWGPQWGMNGYFLIERGKNMCGLAACASYPIPLV
```