MCB5472
Computer methods in molecular evolution

Lecture 4/14/2014

MCrobot demo?

http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php

OldAssignment

•Given a multiple fasta sequence file*, write a script that for each sequence extract the gi number and the species name, and then rewrites the file so that the annotation line starts with the gi number, followed by the species/strain name, followed by a space. (The gi number and the species name should not be separated by or contain any spaces – replace them by _. This is useful, because many programs will recognize the number and name as handle for the sequence (e.g., clustalw2 and phyml)

Assume that the annotation line follows the NCBI convention and begins with the
> followed by the gi number, and ends with the species and strain designation given in []
Example:

```
>gi|229240723|ref|ZP_04365119.1| primary replicative DNA helicase; intein [Cellulomonas flavigena DSM 20109]
```

*An example multiple sequence file in the unofficial NCBI formatted annotation line is here.

More elegant

```
#!/usr/bin/perl
# namerewrite.pl
use strict; use warnings;

die "usage: namerewrite.pl <limit>\n" unless @ARGV == 1;
my $filename=$ARGV[0];
open(IN, "< $filename") or die "cannot open $filename:$!";
open(OUT, "> namerewrite.out");
my $line='';
my $species='';
my $rest='';
while(defined(my $line=<IN>)){
        chomp($line);
        if ($line=~/^>/) {
                $line =~ s/\s/_/g;
                #$line =~ s/\./_/g;
                #$line =~ s/\>/_/g;
                $line =~ s/\>gi\|(\w+)\|//g;
                my $gi=$1;
                $line =~ s/\|[C.+]\|//g;
               my $rearrange = '>'.$gi.'_'.$1.' '.$line;
                print "$rearrange\n";
                }
        else {
                $line =~ tr/atgc/ATGC/;
                $line =~ s/\s//g;
                }
        print "$line\n";
        }
close(IN);
close(OUT);
```

less elegant

```
#!/usr/bin/perl
unless(@ARGV==1) {die "please provide file name in command line \n
file should contain multiple sequences in fasta format \n\n";}
$filename=$ARGV[0];
open(IN, "< $filename") or die "cannot open $filename:$!";
$outfile=$filename."_giSeq";
open(OUT, "> $outfile") or die "cannot open $outfile:$!";

while(<IN>){
    $line = $_;
    if($line =~ m/^>/){ #find annotation line
        if ($line =~ m/gi\|\d*/) # find gi number
            {$gi=$& ;  # assign match to $gi
            $gi =~ s/gi\|//g; #sub gi| with nothing
            # $gi =~ s/\|//g;} #sub | with nothing no longer needed reg ex does not
            else {$nogi++;
            $gi="noGInumber".$nogi}; #in case no match to gi\|\d* found
        if ($line =~ m/\[.*\]/){  #look for species/strain name
            $name = $&;  # assign match to $name
            $name =~ s/\[//g; #sub [ with nothing - note \ before [ in Reg Ex
            $name =~ s/\]//;g}  #sub ] with nothing
            else {$name="NoNameFound"};  #report that no name was included
    $id="$gi_"."$name";
    $id =~ tr/ /_/;
    chomp($id);
    print OUT ">$id\n";
    }else{
    print OUT $line;
    }
}
```

See P17 for info on $&

More brute force

```
open (FILE,"glcosylTransferases.fasta");

while ($in=<FILE>) {
if ($in =~ '\>') {
$in=~ s/\>//g;
@split = split ('\[',$in);
@splat = split ('\]',$split[1]);
@split = split ('\/',$in);
$splat[0] =~ s/ /_/g;
print ">$split[1]$splat[0] $in";
} else {
print $in;
}}
```

HGT as a force a creative force

New biochemical pathways

Oxygen producing PS
Acetoclasitc Methanogenesis (here) (cause of the Permian extinctions? here)

New substrates, new weapons, new resistance genes, breaks up linkage in case of selective sweeps.

Discussion:
Selfish genes *versus* altruism. (Evolutionary stable strategies).
- Group selection? (plasmid sharing in Agrobacteria after plant transformation)
- Under which conditions is it useful for an organism to sacrifice itself (e.g. GTAs), so that other members of the population reap a benefit? -> social parasites

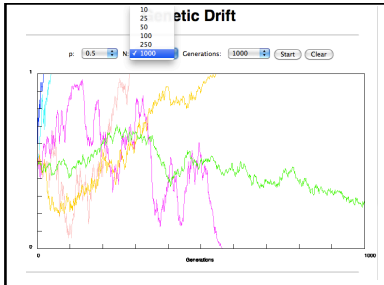
Evolution of the holobiont? (sushi wrapper digesting intestinal symbionts)
- Is selection really acting on the holobiont?

selection versus drift

The larger the population the longer it takes for an allele to become fixed.
Note: Even though an allele conveys a strong selective advantage of 10%, the allele has a rather large chance to go extinct.
Note#2: Fixation is faster under selection than under drift.

Genetic Drift

If one waits long enough, one of two alleles with equal fitness will be fixed



Time till fixation depends on population size

s=0

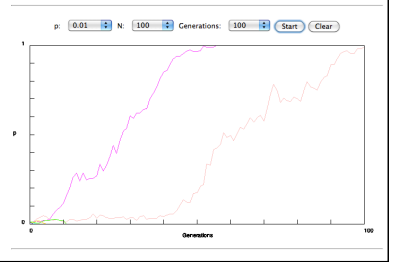
Probability of fixation, P, is equal to frequency of allele in population.
Mutation rate (per gene/per unit of time) = u ;
freq. with which allele is generated in diploid population size N =u*2N
Probability of fixation for each allele = 1/(2N)

Substitution rate =
frequency with which new alleles are generated * Probability of fixation=
u*2N *1/(2N) = u = Mutation rate
Therefore:
If f s=0, the substitution rate is independent of population size, and equal to the mutation rate !!!! (NOTE: Mutation unequal Substitution!)
This is the reason that there is hope that the molecular clock might sometimes work.

Fixation time due to drift alone:
t_{av}=4*N_e generations
(N_e=effective population size; For n discrete generations
N_e= n/(1/N_1+1/N_2+…..1/N_n)

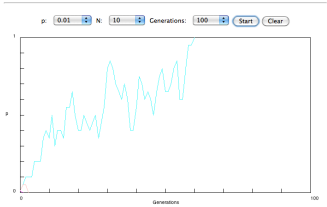
w_{11} = 1.0, w_{12} = 0.9, and w_{22} = 0.8.

Selection and Genetic Drift



w_{11} = 1.0, w_{12} = 0.9, and w_{22} = 0.8.

Selection and Genetic Drift

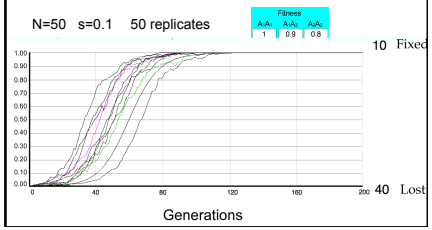


For advantageous mutations:
Probability of fixation, P, is approximately equal to 2s;
e.g., if selective advantage s = 1% then P = 2%

How does the survival of multiple alleles in a population depend on the population size.

Drift and Selection (interesting setting: P=0.01, N=50)

Note: Even though the allele conveys a strong selective advantage of 10%, the allele has a rather large chance to go extinct quickly.



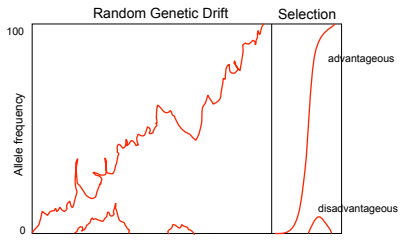
s>0

Time till fixation on average:
t_{av}= (2/s) ln (2N) generations
(also true for mutations with negative "s")

E.g.: N=10^6,
s=0: average time to fixation: $4*10^6$ generations
s=0.01: average time to fixation: 2900 generations

N=10^4,
s=0: average time to fixation: 40.000 generations
s=0.01: average time to fixation: 1.900 generations

⇒ **substitution rate of mutation under positive selection is larger than the rate with which neutral mutations are fixed.**
⇒ **This is easily detected in case of diversifying selection.**



Modified from from www.tcd.ie/Genetics/staff/Aoife/GE3026/GE3026_1+2.ppt

Positive selection (s>0)

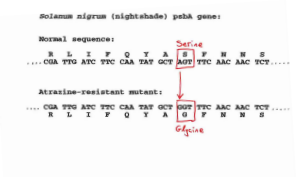
- A new allele (mutant) confers some increase in the **fitness** of the organism
- Selection acts to favour this allele
- Also called adaptive selection or Darwinian selection.

NOTE: **Fitness** = ability to survive and reproduce

Modified from from www.tcd.ie/Genetics/staff/Aoife/GE3026/GE3026_1+2.ppt

Advantageous allele

Herbicide resistance gene in nightshade plant



Modified from from www.tcd.ie/Genetics/staff/Aoife/GE3026/GE3026_1+2.ppt

Negative selection (s<0)

- A new allele (mutant) confers some decrease in the fitness of the organism
- Selection acts to remove this allele
- Also called purifying selection

Modified from from www.tcd.ie/Genetics/staff/Aoife/GE3026/GE3026_1+2.ppt

Neutral mutations

- Neither advantageous nor disadvantageous
- Invisible to selection (no selection)
- Frequency subject to 'drift' in the population
- **Random drift** – random changes in small populations

Types of Mutation-Substitution

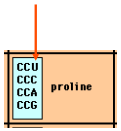
- Replacement of one nucleotide by another
- Synonymous (Doesn't change amino acid)
  - Rate sometimes indicated by Ks
  - Rate sometimes indicated by d_s
- Non-Synonymous (Changes Amino Acid)
  - Rate sometimes indicated by Ka
  - Rate sometimes indicated by d_n

(this and the following 4 slides are from mentor.lscf.ucsb.edu/course/ spring/eemb102/lecture/Lecture7.ppt)

Genetic Code – Note degeneracy of 1st vs 2nd vs 3rd position sites

UUU UUC	phenylalanine	UCU UCC UCA UCG	serine	UAU UAC	tyrosine	UGU UGC	cysteine
UUA UUG	leucine			UAA UAG	stop	UGA	stop
						UGG	tryptophan
CUU CUC CUA CUG	leucine	CCU CCC CCA CCG	proline	CAU CAC	histidine	CGU CGC CGA CGG	arginine
				CAA CAG	glutamine		
AUU AUC AUA	isoleucine	ACU ACC ACA ACG	threonine	AAU AAC	asparagine	AGU AGC	serine
AUG	methionine			AAA AAG	lysine	AGA AGG	arginine
GUU GUC GUA GUG	valine	GCU GCC GCA GCG	alanine	GAU GAC	aspartic acid	GGU GGC GGA GGG	glycine
				GAA GAG	glutamic acid		

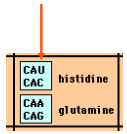
Genetic Code



Four-fold degenerate site – Any substitution is synonymous

From: mentor.lscf.ucsb.edu/course/spring/eemb102/lecture/Lecture7.ppt

Genetic Code



Two-fold degenerate site – Some substitutions synonymous, some non-synonymous

From: mentor.lscf.ucsb.edu/course/spring/eemb102/lecture/Lecture7.ppt

Genetic Code

Degeneracy of 1st vs 2nd vs 3rd position sites results in 25.5% synonymous changes and 74.5% non synonymous changes (Yang&Nielsen,1998).

First Position	Second Position: U		C		A		G		Third Position
	code	Amino Acid	code	Amino Acid	code	Amino Acid	code	Amino Acid	
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

Measuring Selection on Genes

- Null hypothesis = neutral evolution
- Under neutral evolution, synonymous changes should accumulate at a rate equal to mutation rate
- Under neutral evolution, amino acid substitutions should also accumulate at a rate equal to the mutation rate

From: mentor.lscf.ucsb.edu/course/spring/eemb102/lecture/Lecture7.ppt

Counting #s/#a

	Ser	Ser	Ser	Ser	Ser
Species1	TGA	TGC	TGT	TGT	TGT
	Ser	Ser	Ser	Ser	Ala
Species2	TGT	TGT	TGT	TGT	GGT

#s = 2 sites
#a = 1 site
#a/#s=0.5

To assess selection pressures one needs to calculate the rates (Ka, Ks), i.e. the occurring substitutions as a fraction of the possible syn. and nonsyn. substitutions.

Things get more complicated, if one wants to take transition transversion ratios and codon bias into account. See chapter 4 in Nei and Kumar, Molecular Evolution and Phylogenetics.

Modified from: mentor.lscf.ucsb.edu/course/spring/eemb102/lecture/Lecture7.ppt

Testing for selection using dN/dS ratio

dN/dS ratio (aka **Ka/Ks** or **ω** (omega) ratio) where

dN = number of non-synonymous substitutions / number of possible non-synonymous substitutions

dS =number of synonymous substitutions / number of possible non-synonymous substitutions

dN/dS >1 positive, Darwinian selection

dN/dS =1 neutral evolution

dN/dS <1 negative, purifying selection

PAML (codeml) the basic model

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

The equilibrium frequency of codon j (π_j) can be considered a free parameter, but can also be calculated from the nucleotide frequencies at the three codon positions (control variable `CodonFreq`). Under this model, the relationship holds that $\omega = d_N/d_S$, the ratio of nonsynonymous/synonymous substitution rates. This basic model is fitted by specifying model = 0 NSsites = 0, in the control file codeml.ctl. It forms the basis for more sophisticated models implemented in codeml.

dambe

Three programs worked well for me to align nucleotide sequences based on the amino acid alignment,

One is DAMBE (works well for windows). This is a handy program for a lot of things, including reading a lot of different formats, calculating phylogenies, it even runs codeml (from PAML) for you.

The procedure is not straight forward, but is well described on the help pages. After installing DAMBE go to HELP -> general HELP -> sequences -> align nucleotide sequences based on ...->

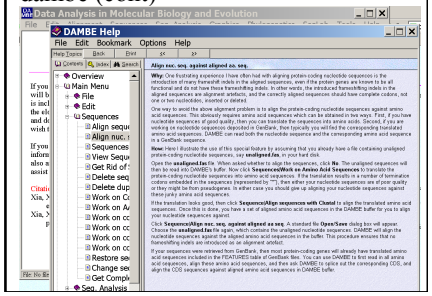
If you follow the instructions to the letter, it works fine.

DAMBE also calculates Ka and Ks distances from codon based aligned sequences.

Alternatives are

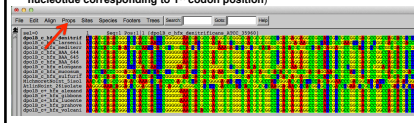
- tranalign from the EMBOSS package, and
- Seaview (see below)

dambe (cont)

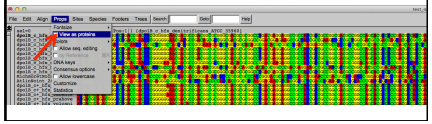


Codon based alignments in Seaview

Load nucleotide sequences (**no gaps in sequences, sequence starts with nucleotide corresponding to 1st codon position**)

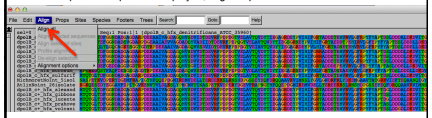


Select view as proteins



Codon based alignments in Seaview

With the protein sequences displayed, align sequences



Select view as nucleotides

PAML (codeml) the basic model

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

The equilibrium frequency of codon j (π_j) can be considered a free parameter, but can also be calculated from the nucleotide frequencies at the three codon positions (control variable `CodonFreq`). Under this model, the relationship holds that $\omega = d_N/d_S$, the ratio of nonsynonymous/synonymous substitution rates. This basic model is fitted by specifying model = 0 NSsites = 0, in the control file codeml.ctl. It forms the basis for more sophisticated models implemented in codeml.

sites versus branches

You can determine omega for the whole dataset; however, usually not all sites in a sequence are under selection all the time.

PAML (and other programs) allow to either determine omega for each site over the whole tree, *Branch Models* **, or determine omega for each branch for the whole sequence,** *Site Models* **.**

It would be great to do both, i.e., conclude codon 176 in the vacuolar ATPases was under positive selection during the evolution of modern humans – alas, a single site does not provide much statistics ….

Sites model(s)

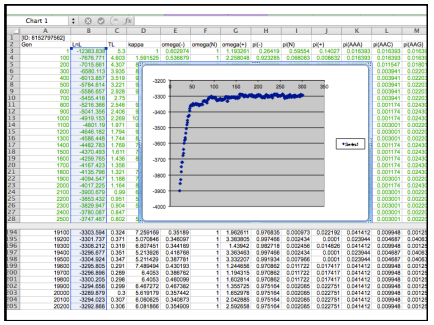
work great have been shown to work great in few instances. The most celebrated case is the influenza virus HA gene.

A talk by Walter Fitch (slides and sound) on the evolution of this molecule is here .
This article by Yang et al, 2000 gives more background on ml aproaches to measure omega. The dataset used by Yang et al is here: flu_data.paup .

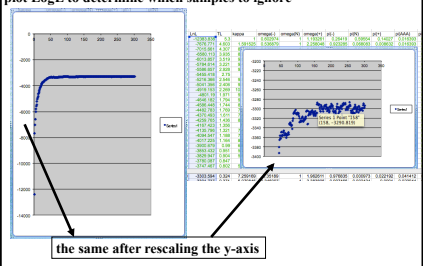
sites model in MrBayes

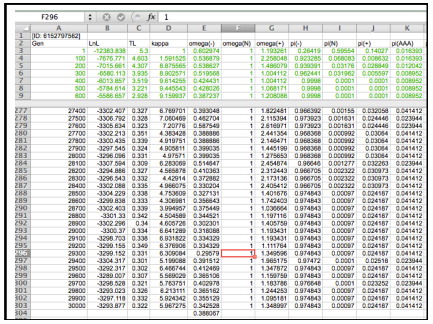
The MrBayes block in a nexus file might look something like this:

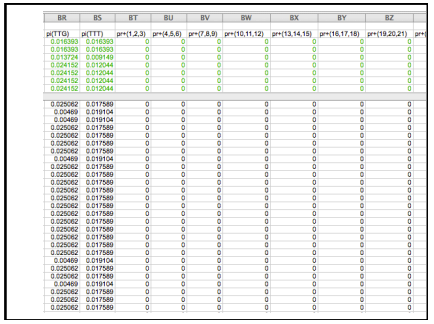
```
begin mrbayes;
set autoclose=yes;
lset nst=2 rates=gamma nucmodel=codon omegavar=Ny98;
mcmcp samplefreq=500 printfreq=500;
mcmc ngen=500000;
sump burnin=50;
sumt burnin=50;
end;
```



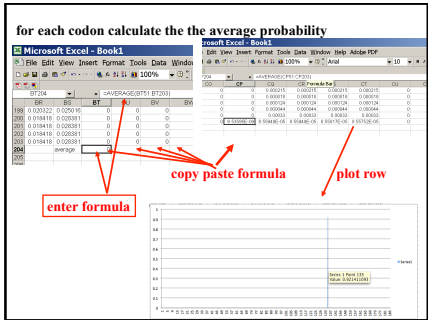
plot LogL to determine which samples to ignore



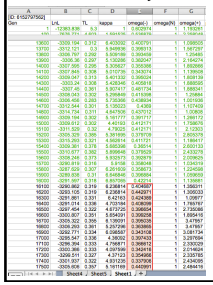




for each codon calculate the the average probability

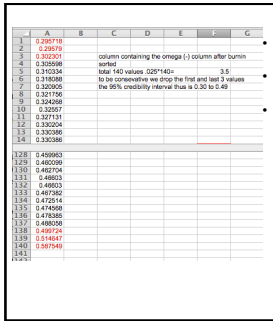


To determine credibility interval for a parameter (here omega<1):



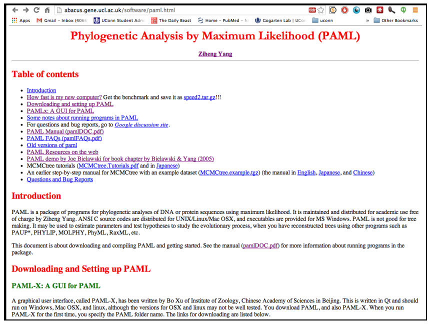
Select values for the parameter, sampled after the burning.

Copy paste to a new spreadsheet,



- **Sort values according to size,**
- **Discard top and bottom 2.5%**
- **Remainder gives 95% credibility interval.**

Slides on codeml are at
http://abacus.gene.ucl.ac.uk/ziheng/data/pamlDEMO.pdf
http://m.docente.unife.it/silvia.fuselli/dispense-corsi/EM_Lezione%208.pdf



abacus.gene.ucl.ac.uk/software/paml.html

Phylogenetic Analysis by Maximum Likelihood (PAML)

Ziheng Yang

Table of contents

Introduction

PAML is a package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood. It is maintained and distributed for academic use free of charge by Ziheng Yang. ANSI C source codes are distributed for UNIX/Linux/Mac OSX, and executables are provided for MS Windows. PAML is not good for tree making. It may be used to estimate parameters and test hypotheses to study the evolutionary process, when you have reconstructed trees using other programs such as PAUP*, PHYLIP, MOLPHY, PhyML, RaxML, etc.

This document is about downloading and compiling PAML and getting started. See the manual (pamlDOC.pdf) for more information about running programs in the package.

Downloading and Setting up PAML

PAML-X: A GUI for PAML

A graphical user interface, called PAML-X, has been written by Bo Xu of Institute of Zoology, Chinese Academy of Sciences in Beijing. This is written in Qt and should run on Windows, Mac OSX, and linux, although the versions for OSX and linux may not be well tested. You download PAML, and also PAML-X. When you run PAML-X for the first time, you specify the PAML folder name. The links for downloading are listed below.

Major weaknesses:

- Poor tree search
- Poor user interface

Major strength:

- Sophisticated likelihood models

From:
http://abacus.gene.ucl.ac.uk/ziheng/data/pamlDEMO.pdf

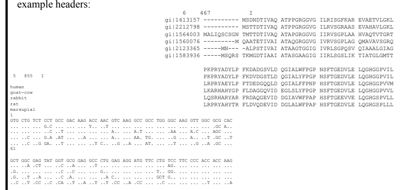
PAML – codeml – sites model

the paml package contains several distinct programs for nucleotides (baseml) protein coding sequences and amino acid sequences (codeml) and to simulate sequences evolution.

The input file needs to be in phylip format.

By default it assumes a sequential format (e.g. here).

If the sequences are interleaved, you need to add an "I" to the first line, as in these example headers:



PAML – codeml – sites model (cont.)

the program is invoked by typing codeml followed by the name of a control file that tells the program what to do (or it uses codeml.ctl by default).

paml can be used to find the maximum likelihood tree, however, the program is rather slow. Phyml is a better choice to find the tree, which then can be used as a user tree.

An example for a codeml.ctl file is codeml.hv1.sites.ctl
This file directs codeml to run three different models:
one with an omega fixed at 1, a second where each site can be either have an omega between 0 and 1, or an omega of 1, and third a model that uses three omegas as described before for MrBayes.
The output is written into a file called Hv1.sites.codeml_out (as directed by the control file).

Point out log likelihoods and estimated parameter line (kappa and omegas)

Additional useful information is in the rst file generated by the codeml

Discuss overall result.

PAML – codeml – branch model

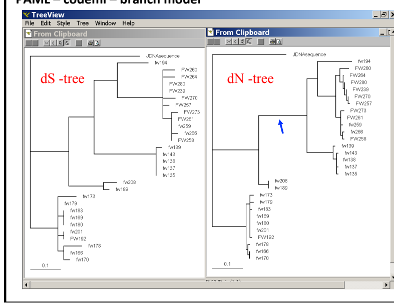
For the same dataset to estimate the dN/dS ratios for individual branches, you could use this file codeml.hv1.branches.ctl as control file.

The output is written, as directed by the control file, into a file called Hv1.branch.codeml_out

A good way to check for episodes with plenty of non-synonymous substitutions is to compare the dn and ds trees.

Bottom line: one needs plenty of sequences to detect positive selection.

PAML – codeml – branch model



Power and accuracy of LRT to detect positive selection

- χ^2 distribution does not apply when sample sizes are small
- χ^2 distribution (or mixture distributions) do not apply due to boundary problems
- χ^2 makes LRT conservative (type I error rate < alpha)
- LRT based on χ^2 can be powerful !!!
- Power is affected by (i) sequence divergence, (ii) number of lineages, and (iii) strength of positive selection
- The most efficient way to increase power is to add lineages !

Data from: Anisimova, Bielawski, and Yang, 2001, *Mol. Bio. Evol.* 18:1585-1592.

From:
http://abacus.gene.ucl.ac.uk/ziheng/data/pamlDEMO.pdf

CODEML.ctl parameter file

```
      seqfile = VHL.fas   * sequence data file name
      outfile = mlc       * main result file name
     treefile = VHL.nwk   * tree structure file name

        noisy = 3         * 0,1,2,3,9: how much rubbish on the screen
      verbose = 1         * 1: detailed output, 0: concise output

      runmode = 0         * 0: user tree; 1: semi-automatic; 2: automatic;
                          * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

      seqtype = 1         * 1:codons; 2:AAs; 3:codons-->AAs
    CodonFreq = 2         * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table

        model = 0         * codonml:  0:one, 1:b, 2:2 or more dN/dS ratios for branches
      NSsites = 0 1 2 3   * 0:one w;1:neutral;2:positive; 3:discrete;4:freqs;
                          * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
                          * 10:beta&1+gamma; 11:beta&1>normal; 12:0&2normal; 13:3normal

        icode = 0         * 0:standard genetic code; 1:mammalian mt; 2-10:see below
        Mgene = 0         * codonml: 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff

    fix_kappa = 0         * 1: kappa fixed, 0: kappa to be estimated
        kappa = 2

    fix_alpha = 1         * shape parameter of gamma distribuion 1:fixed, 0: estimated
        alpha = 0         * alpha = 0 means infinity

    fix_omega = 0         * 1: omega or omega_1 fixed, 0: estimate
        omega = 0.5

        ncatG = 3

    cleandata = 1         * remove sites with ambiguity data (1:yes, 0:no)?
```

$k \longrightarrow$ transition/transversion rate ratio

$\pi \longrightarrow$ codon frequency

- F_{EQUAL}: each codon has the same frequency
- F1X 4: codon frequency are expected from the frequencies of four nucleotides.
- **F3 X 4**: codon frequencies are expected from 3 sets of nucleotide frequencies for the three codon positions.
- F61: all codon frequnecies as parameters

- Ziheng Yang: Computational Molecular Evolution
- PAML User Guide
- Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 2000; **17**: 32-43
- Yang Z. PAML 4: phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 2007; **24**: 1586-1591

H_0: variable selective pressure but NO positive selection (M1a)

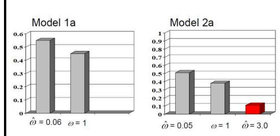
H_1: variable selective pressure with positive selection (M2a)

Likelihood ratio test (LRT)

$2\Delta l = 2\,(l_1 - l_0)$ χ^2 distribution

Parameters in Site Models

Model	NSsites	p	Parameters
M1a (neutral)	1	2	p_0 ($p_1 = 1 - p_0$), $\omega_0 < 1$, $\omega_1 = 1$
M2a (selection)	2	4	p_0, p_1 ($p_2 = 1 - p_0 - p_1$), $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$



SE:

- IL MODELLO M2a SI ADATTA MEGLIO AI DATI
- IL VALORE DI ω STIMATO E' >1 PER LA CLASSE DI SITI p_2

UNA PROPORZIONE DI SITI PARI A p2 E' SOTTOPOSTA A SELEZIONE POSITIVA

From: http://m.docente.unife.it/silvia.fuselli/dispense-corsi/EM_Lezione%208.pdf

H_0: Beta distributed variable selective pressure (M7)

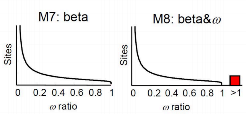
H_1: Beta plus positive selection(M8)

Likelihood ratio test (LRT)

$2\Delta l = 2\,(l_1 - l_0)$ χ^2 distribution

Parameters in Site Models

Model	NSsites	p	Parameters
M7 (beta)	7	2	p, q
M8 (beta&ω)	8	4	p_0 $(p_1 = 1 - p_0)$, $p, q, \omega_s > 1$



SE:

- IL MODELLO M8 SI ADATTA MEGLIO AI DATI
- IL VALORE DI ω_S STIMATO E' >1 PER LA CLASSE DI SITI p1

UNA PROPORZIONE DI SITI PARI A p1 E' SOTTOPOSTA A SELEZIONE POSITIVA

Quando i test suggeriscono la presenza di selezione positiva si utilizzano dei metodi Bayesiani (BEB) per calcolare la probabilità a posteriori che **ciascun codone** provenga dalla classe di siti sotto selezione positiva.

From: http://m.docente.unife.it/silvia.fuselli/dispense-corsi/EM_Lezione%208.pdf

---

Table 2. Parameters in the site models

Model	NSsites	#p	Parameters
M0 (one ratio)	0	1	ω
M1a (neutral)	1	2	p_0 $(p_1 = 1 - p_0)$, $\omega_0 < 1$, $\omega_1 = 1$
M2a (selection)	2	4	p_0, p_1 $(p_2 = 1 - p_0 - p_1)$, $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$
M3 (discrete)	3	5	p_0, p_1 $(p_2 = 1 - p_0 - p_1)$ $\omega_0, \omega_1, \omega_2$
M7 (beta)	7	2	p, q
M8 (beta&ω)	8	4	p_0 $(p_1 = 1 - p_0)$, $p, q, \omega_s > 1$

NOTE.— #p is the number of free parameters in the ω distribution. Parameters in parentheses are not free and should not be counted: for example, in M1a, p_1 is not a free parameter as $p_1 = 1 - p_0$. In both likelihood ratio tests comparing M1a against M2a and M7 against M8, df = 2. The site models are specified using `NSsites`.

The *site models* allow the ω ratio to vary among sites (among codons or amino acids in the protein) (Nielsen and Yang 1998; Yang et al. 2000b). A number of such models are implemented in `codeml` using the variable `Nssites` (and `model` = 0). You can run several Nssites models in one go, by specifying several values for NSsites. For example, NSsites = 0 1 2 7 8 will fit 5 models to the same data in one go. The site models have been used in real data analyses and evaluated in computer simulation studies (Anisimova et al. 2001; Anisimova et al. 2002; Anisimova et al. 2003; Wong et al. 2004). Two pairs of models appear to be particularly useful, forming two likelihood ratio tests of positive selection. The first compares M1a (NearlyNeutral) and M2a (PositiveSelection), while the second compares M7 (beta) and M8 (beta&ω). M1a (NearlyNeutral)

From: PAML manual

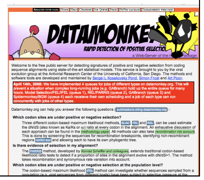
Hy-Phy –

Hypothesis Testing using Phylogenies.

Using Batchfiles or GUI

Information at http://www.hyphy.org/

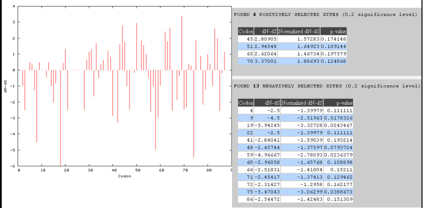
Selected analyses also can be performed online at http://www.datamonkey.org/



---

hy-phy

Results of an anaylsis using the SLAC approach



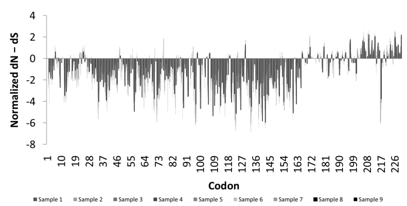
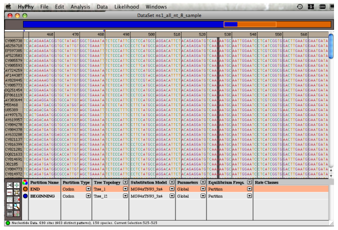


Fig 1. Patterns of substitutions: Bars represent dN > dS (positive) or dN < dS (negative) in random samples of 148 – 150 sequences (A) and the whole dataset of 1312 viruses (B). Included in B are regions of mapped activity and 3D structures of the RNA-binding domain (RBD, panel I) [21] and Effector domain (ED, rotated to expose the 7 β-sheets (panel II) and 2 α-helices (panel II)) [7] with residues under negative (yellow/brown), neutral (gray) or positive (red) selection highlighted. Residues 208-230 not included in the 3D structure of the ED are disordered (compare with figure 5). Note sites with dN > dS map on the helix motifs of the ED or the linkers flanking them or the disordered region.

Example testing for dN/dS in two partitions of the data -- John's dataset

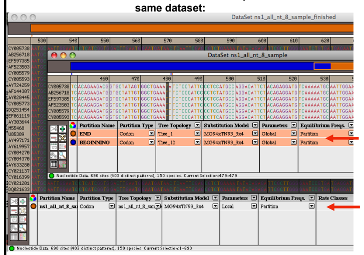


Set up two partitions, define model for each, optimize likelihood

---

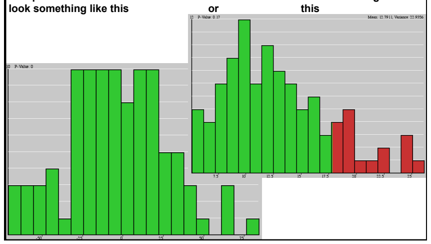
Example testing for dN/dS in two partitions of the data -- John's dataset

Alternatively, especially if the the two models are not nested, one can set up two different windows with the same dataset:



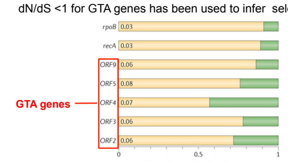
Example testing for dN/dS in two partitions of the data -- John's dataset

Simulation under model 2, evaluation under model 1, calculate LR Compare real LR to distribution from simulated LR values. The result might look something like this or this

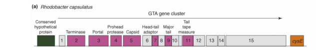


Purifying selection in GTA genes

dN/dS <1 for GTA genes has been used to infer selection for function



Lang AS, Zhaxybayeva O, Beatty JT. Nat Rev Microbiol. 2012 Jun 11;10(7):472-82



Lang, A.S. & Beatty, J.T. Trends in Microbiology , Vol.15, No.2 , 2006

Purifying selection in *E.coli* ORFans

dN-dS < 0 for some ORFan *E. coli* clusters seems to suggest they are functional genes.

Gene groups	Number	dN-dS>0	dN-dS<0	dN-dS=0
E. coli ORFan clusters	3773	944 (25%)	1953 (52%)	876 (23%)
Clusters of *E.coli* sequences found in *Salmonella sp., Citrobacter sp.*	610	104 (17%)	423(69%)	83 (14%)
Clusters of *E.coli* sequences found in some *Enterobacteriaceae* only	373	8 (2%)	365 (98%)	0 (0%)

Adapted after Yu, G. and Stoltzfus, A. Genome Biol Evol (2012) Vol. 4 1176-1187

Vincent Daubin and Howard Ochman: Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in *E. coli*. *Genome Research 14:1036-1042, 2004*

The ratio of non-synonymous to synonymous substitutions for genes found only in the E.coli - Salmonella clade is lower than 1, but larger than for more widely distributed genes.

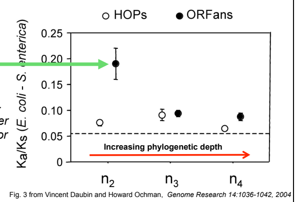


Fig. 3 from Vincent Daubin and Howard Ochman, *Genome Research 14:1036-1042, 2004*

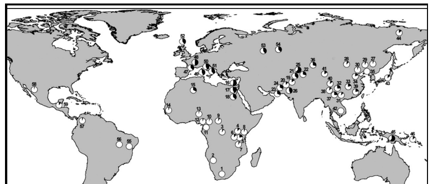
Trunk-of-my-car analogy: Hardly anything in there is the is the result of providing a selective advantage. Some items are removed quickly (purifying selection), some are useful under some conditions, but most things do not alter the fitness.

Could some of the inferred purifying selection be due to the acquisition of novel detrimental characteristics (e.g., protein toxicity, HOPELESS MONSTERS)?

Other ways to detect positive selection

Selective sweeps -> fewer alleles present in population (see contributions from archaic Humans for example)

Repeated episodes of positive selection -> high dN

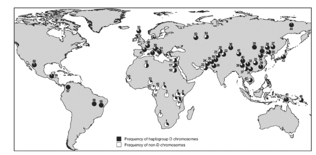


Variant arose about 5800 years ago

Each population's geographic origin, number of individuals, and frequency of haplogroup D chromosomes are given in parentheses as follows:

1, Southeastern and Southwestern Bantu (South Africa, 8, 0%); 2, San (Namibia, 7, 0%); 3, Mbuti Pygmy (Democratic Republic of Congo, 14, 7.1%); 4, Masaai (Tanzania, 26, 11.5%); 5, Sandawe (Tanzania, 33, 22.7%); 6, Burunge (Tanzania, 28, 12.5%); 7, Turu (Tanzania, 27, 7.4%); 8, Northeastern Bantu (Kenya, 12, 4.2%); 9, Biaka Pygmy (Central African Republic, 32, 6.3%); 10, Zime (Cameroon, 25, 2%); 11, Bakola Pygmy (Cameroon, 27, 0%); 12, Bamoun (Cameroon, 29, 5.2%); 13, Yoruba (Nigeria, 24, 2.1%); 14, Mandenka (Senegal, 24, 4.2%); 15, Mozabite [Algeria (Mzab region), 29, 31%]; 16, Druze [Israel (Carmel region), 45, 52.2%]; 17, Palestinian [Israel (Central), 43, 46.5%]; 18, Bedouin [Israel (Negev region), 46, 37%]; 19, Hazara (Pakistan, 19, 15.8%); 20, Balochi (Pakistan, 23, 13%); 21, Pathan (Pakistan, 21, 40.5%); 22, Burusho (Pakistan, 23, 28.3%); 23, Makrani (Pakistan, 25, 32%); 24, Brahui (Pakistan, 24, 33.3%); 25, Kalash (Pakistan, 25, 60%); 26, Sindhi (Pakistan, 25, 44%); 27, Hezhen (China, 9, 5.6%); 28, Mongola (China, 9, 11.1%); 29, Daur (China, 8, 6.3%); 30, Orogen (China, 10, 5%); 31, Miaozu (China, 10, 10%); 32, Yizu (China, 10, 25 %); 33, Tujia (China, 10, 20%); 34, Han (China, 42, 17.9%); 35, Xibo (China, 9, 0%); 36, Uygur (China, 10, 30%); 37, Dai (China, 10, 25%); 38, Lahu (China, 10, 10%); 39, She (China, 7, 21.4%); 40, Naxi (China, 9, 11.1%); 41, Tu (China, 9, 16.7%); 42, Cambodian (Cambodia, 10, 0%); 43, Japanese (Japan, 28, 10.7%); 44, Yakut [Russia (Siberia region), 24, 12.5%]; 45, Papuan (New Guinea, 16, 59.4%); 46, NAN Melanesian (Bougainville, 18, 11.1%); 47, French Basque (France, 15, 40%); 48, French (France, 29, 50%); 49, Sardinian (Italy, 27, 46.3%); 50, North Italian [Italy (Bergamo region), 12, 45.8%]; 51, Tuscan (Italy, 8, 37.5%); 52, Orcadian (Orkney Islands, 16, 40.6%); 53, Russian (Russia, 25, 38%); 54, Adygei [Russia (Caucasus region), 15, 40%]; 55, Karitiana (Brazil, 24, 0%); 56, Surui (Brazil, 21, 0%); 57, Colombian (Colombia, 13, 3.8%); 58, Pima (Mexico, 25, 2%); 59, Maya (Mexico, 24, 12.5%).

Fig. 3. Global frequencies of Microcephalin haplogroup D chromosomes (defined as having the derived C allele at the G37995C diagnostic SNP) in a panel of 1184 individuals



P. D. Evans et al., Science 309, 1717 -1720 (2005)

The age of haplogroup D was found to be ~37,000 years

Published by AAAS

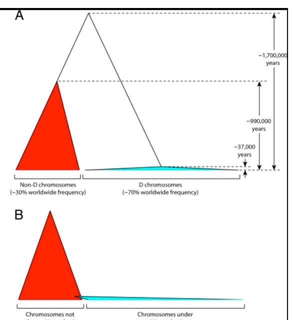
Science AAAS

Proc Natl Acad Sci U S A. 2006 November 28; 103(48): 18178–18183. PMCID: PMC1635020
Published online 2006 November 7. doi: 10.1073/pnas.0606966103.

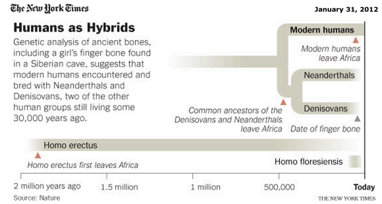
Evolution

Evidence that the adaptive allele of the brain size gene *microcephalin* introgressed into *Homo sapiens* from an archaic *Homo* lineage

Patrick D. Evans,*†‡ Nitzan Mekel-Bobrov,*†‡ Eric J. Vallender,*†‡ Richard R. Hudson,§ and Bruce T. Lahn*†¶



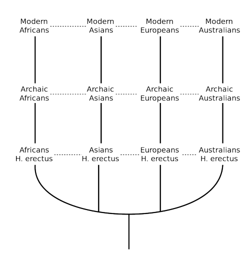
The New York Times — January 31, 2012



Close Window

From:
http://www.nytimes.com/2012/01/31/science/gains-in-dna-are-speeding-research-into-human-origins.html?_r=1

The multiregional hypothesis



From http://en.wikipedia.org/wiki/Multiregional_Evolution

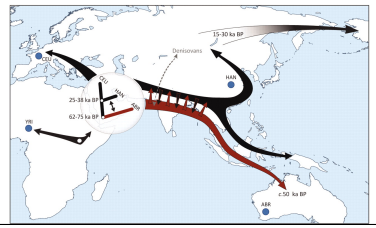
Published Online September 22 2011
Science 7 October 2011:
Vol. 334 no. 6052 pp. 94-98
DOI: 10.1126/science.1211177

‹ Prev | Table of Contents

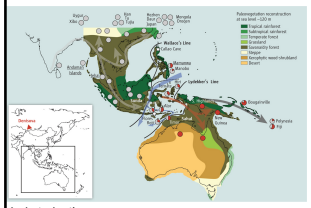
REPORT

An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia

Morten Rasmussen[1,2,*], Xiaosen Guo[2,3,*], Yong Wang[2,*], Kirk E. Lohmueller[2,*], Simon Rasmussen[1],

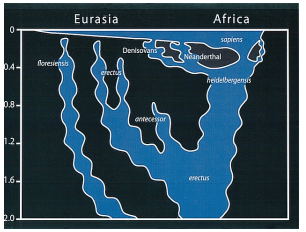


Did the Denisovans Cross Wallace's Line?
Science 18 October 2013:
vol. 342 no. 6156 321-323



Ancient migrations.
The proportions of Denisovan DNA in modern human populations are shown as red in pie charts, **relative to New Guinea and Australian Aborigines** (3). Wallace's Line (8) is formed by the powerful Indonesian flow-through current (blue arrows) and marks the limit of the Sunda shelf and Eurasian placental mammals.

Archaic human admixture with modern Homo sapiens



From: http://en.wikipedia.org/wiki/Archaic_human_admixture_with_modern_Homo_sapiens

For more discussion on archaic and early humans see:
http://en.wikipedia.org/wiki/Denisova_hominin

http://www.nytimes.com/2012/01/31/science/gains-in-dna-are-speeding-research-into-human-origins.html

http://www.sciencedirect.com/science/article/pii/S0002929711003958
http://www.abc.net.au/science/articles/2012/08/31/3580500.htm

http://www.sciencemag.org/content/334/6052/94.full
http://www.sciencemag.org/content/334/6052/94/F2.expansion.html

http://haplogroup-a.com/Ancient-Root-AJHG2013.pdf