MCB 5472

Bayesian approaches and types of selection

Peter Gogarten
Office: *BSP 404*
phone: *860 486-4061,*
Email: *gogarten@uconn.edu*

Old Assignment

•Given a multiple fasta sequence file*, write a script that for each sequence extract the gi number and the species name. and 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 clustalw will recognize the number and name as handle for the sequence.)

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]
```

Example multiple sequence file is here.

- Work on your student project

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/\[(.+)\]//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.".giSpec";
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;
    }
}
```

More brute force

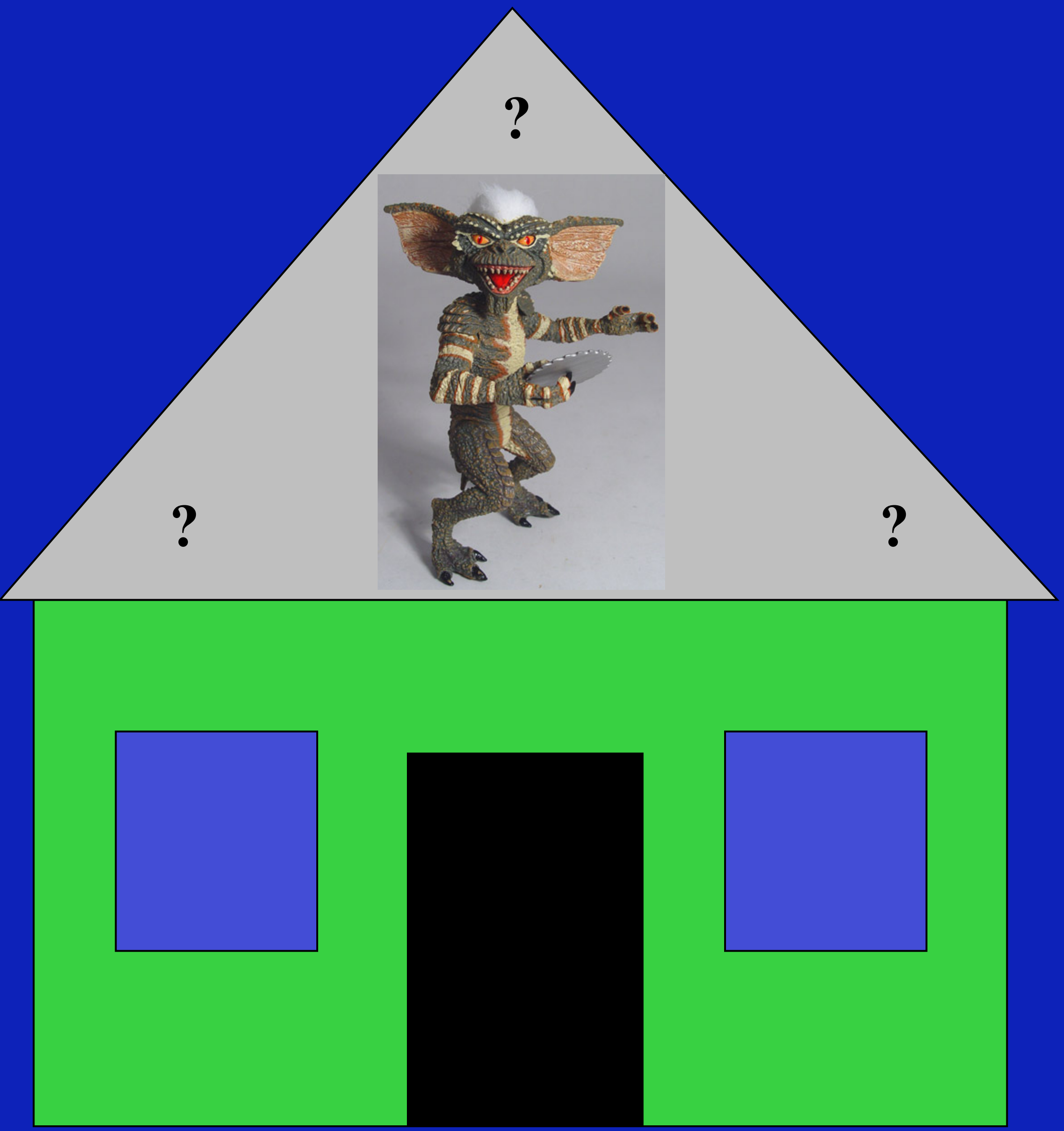
```
open (FILE,"glcosylTransferases.fasta");

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;
}}
```

Write script from exam

Work of Student project

Elliot Sober’s Gremlins



Observation: Loud noise in the attic

Hypothesis: *gremlins in the attic playing bowling*

Likelihood = *P(noise|gremlins in the attic)*

P(gremlins in the attic|noise)

Bayes' Theorem

Reverend Thomas Bayes
(1702-1761)

$$P(model|data, I) = P(model, I)\,\frac{P(data|model, I)}{P(data, I)}$$

Likelihood

describes how well the model predicts the data

Posterior Probability

represents the degree to which we believe a given **model** accurately describes the situation given the available **data** and all of our prior information **I**

Prior Probability

describes the degree to which we believe the model accurately describes reality based on all of our prior information.

Normalizing constant

Alternative Approaches to Estimate Posterior Probabilities

Bayesian Posterior Probability Mapping with MrBayes

(Huelsenbeck and Ronquist, 2001)

Problem:

Strimmer's formula $$p_i = \frac{L_i}{L_1 + L_2 + L_3}$$

only considers 3 trees
(those that maximize the likelihood for the three topologies)

Solution:

Exploration of the tree space by sampling trees using a biased random walk
(Implemented in MrBayes program)

Trees with higher likelihoods will be sampled more often

$$p_i \approx \frac{N_i}{N_{total}}$$

,where N_i - number of sampled trees of topology *i*, *i*=1,2,3

N_{total} – total number of sampled trees (has to be large)

Illustration of a biased random walk

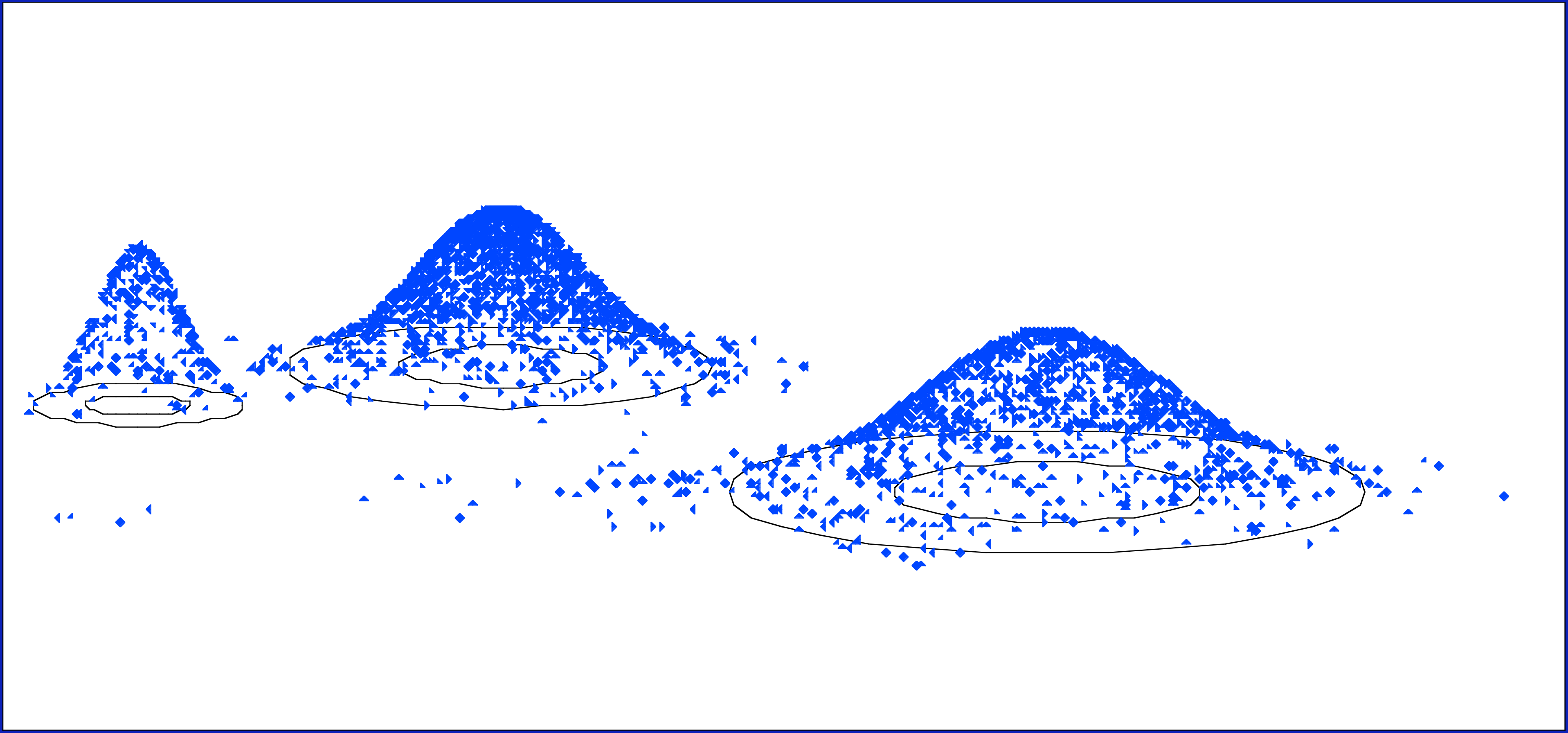
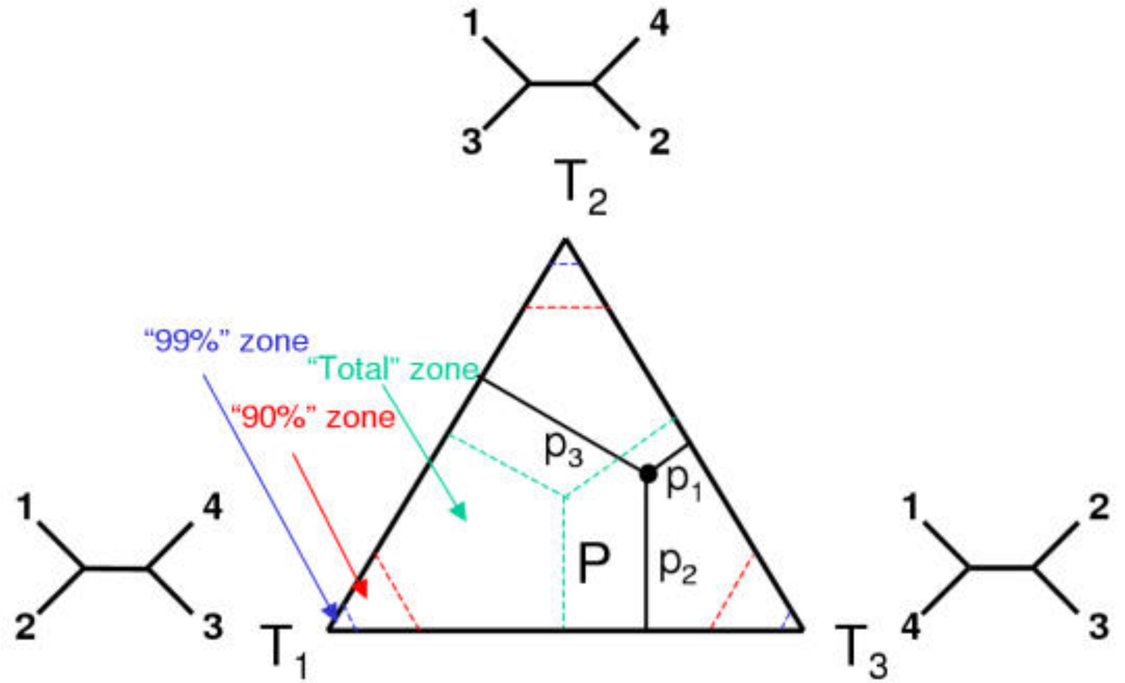


Figure generated using MCRobot program (Paul Lewis, 2001)

ml mapping

- 

From: Olga Zhaxybayeva and J Peter Gogarten *BMC Genomics* 2002, 3:4

ml mapping

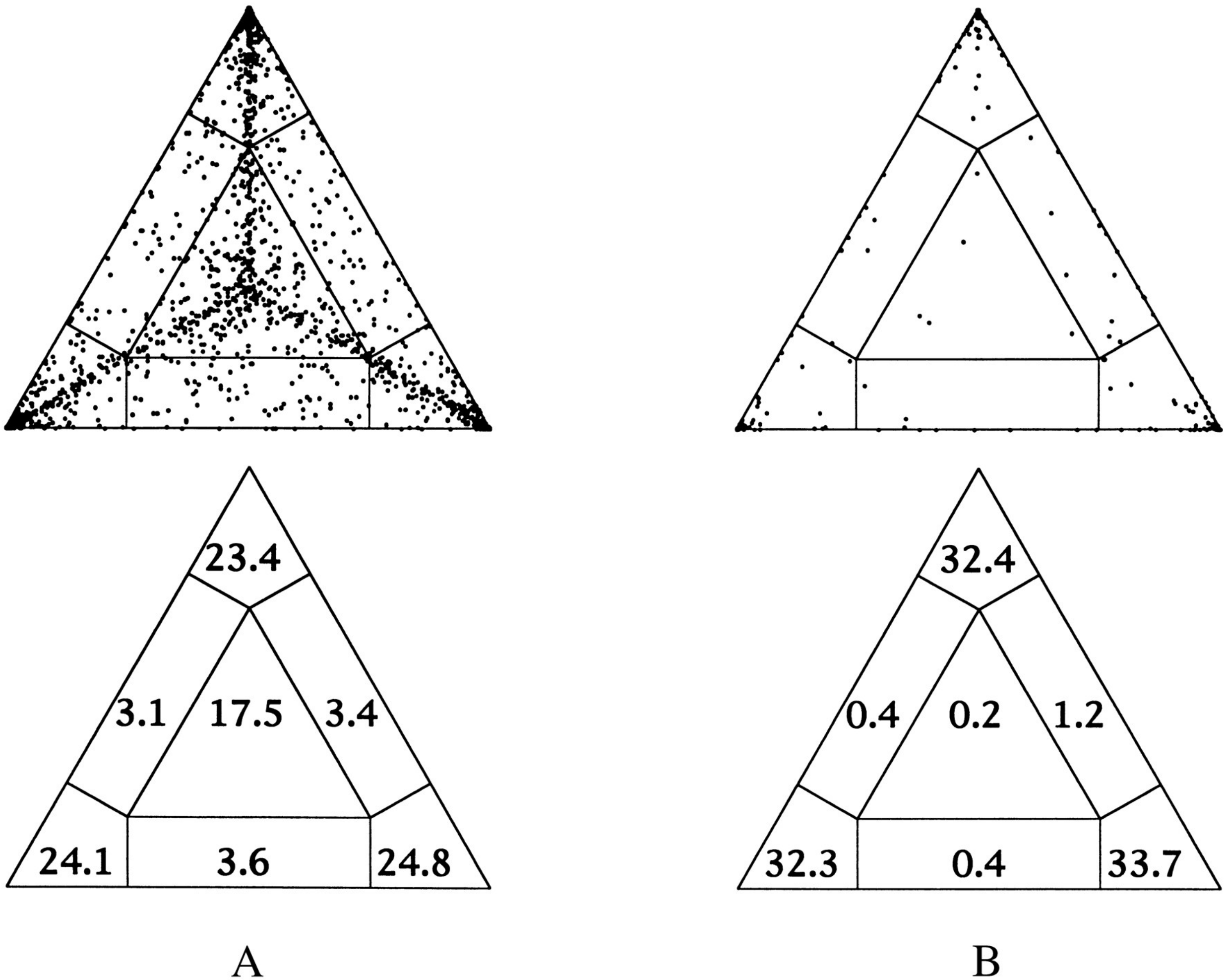
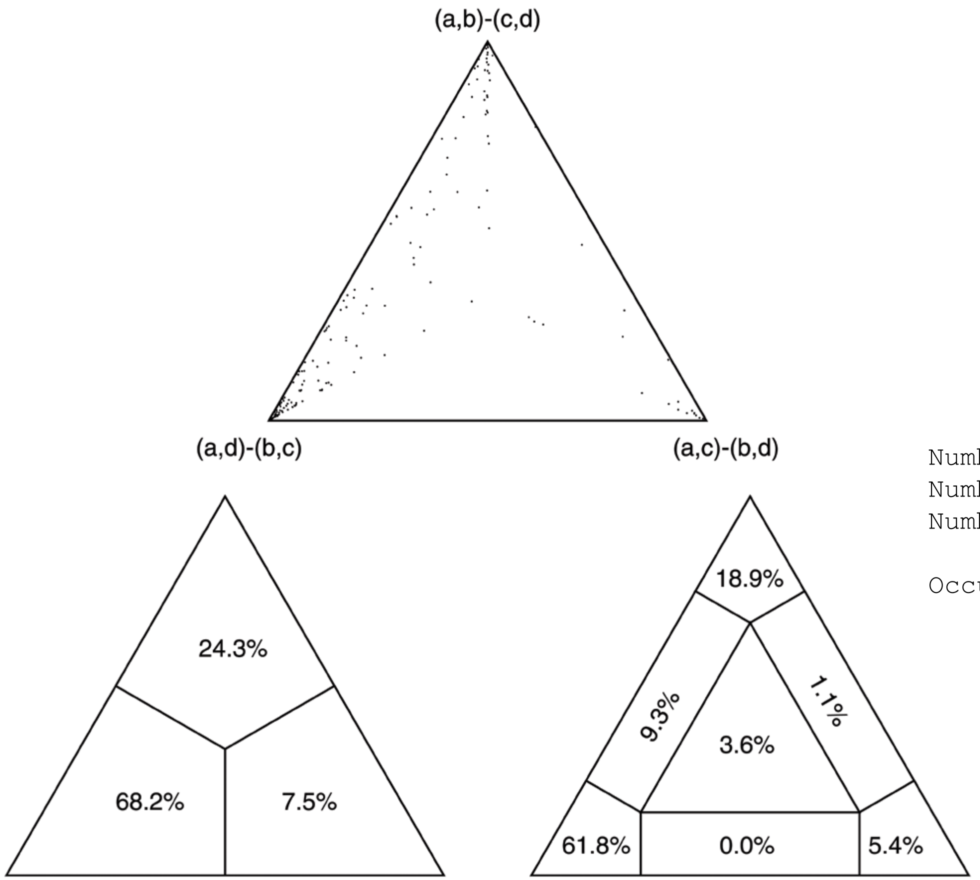


Figure 5. Likelihood-mapping analysis for two biological data sets. (*Upper*) The distribution patterns. (*Lower*) The occupancies (in percent) for the seven areas of attraction.

(*A*) Cytochrome-*b* data from ref. 14. (*B*) Ribosomal DNA of major arthropod groups (15).

From: Korbinian Strimmer and Arndt von Haeseler Proc. Natl. Acad. Sci. USA Vol. 94, pp. 6815-6819, June 1997



Cluster a: 14 sequences
outgroup (prokaryotes)

Cluster b: 20 sequences
other Eukaryotes

Cluster c: 1 sequences
Plasmodium

Cluster d: 1 sequences
Giardia

```
                 (a,b)-(c,d)
                      /\
                     /  \
                    /    \
                   /   1  \
                  / \    / \
                 /   \  /   \
                /     \/     \
               /  3    :   2  \
              /        :       \
             /__________________\
     (a,d)-(b,c)                  (a,c)-(b,d)
```

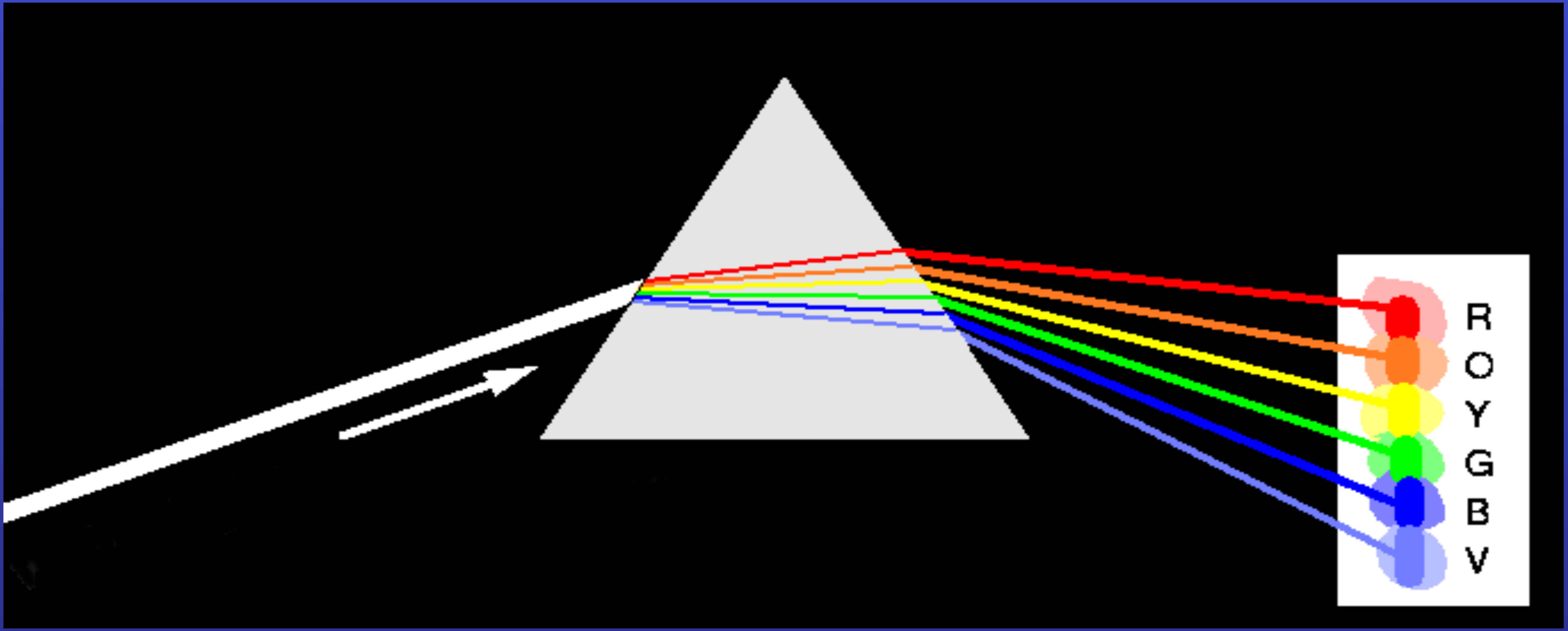
Number of quartets in region 1: 68 (= 24.3%)
Number of quartets in region 2: 21 (= 7.5%)
Number of quartets in region 3: 191 (= 68.2%)

Occupancies of the seven areas 1, 2, 3, 4, 5, 6, 7:

```
                 (a,b)-(c,d)
                      /\
                     /  \
                    /  1 \
                   / \  / \
                  /   /\   \
                 / 6 /  \ 4 \
                /   /  7 \   \
               / \ /______\ / \
              / 3  :  5  :  2  \
             /__________________\
     (a,d)-(b,c)                  (a,c)-(b,d)
```

Number of quartets in region 1: 53 (= 18.9%)
Number of quartets in region 2: 15 (= 5.4%)
Number of quartets in region 3: 173 (= 61.8%)
Number of quartets in region 4: 3 (= 1.1%)
Number of quartets in region 5: 0 (= 0.0%)
Number of quartets in region 6: 26 (= 9.3%)
Number of quartets in region 7: 10 (= 3.6%)

Decomposition of Phylogenetic Data



Phylogenetic information present in genomes

Break information into small quanta of information (bipartitions or embedded quartets)

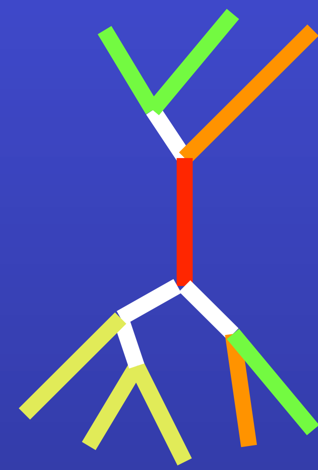
Analyze spectra to detect transferred genes and plurality consensus.

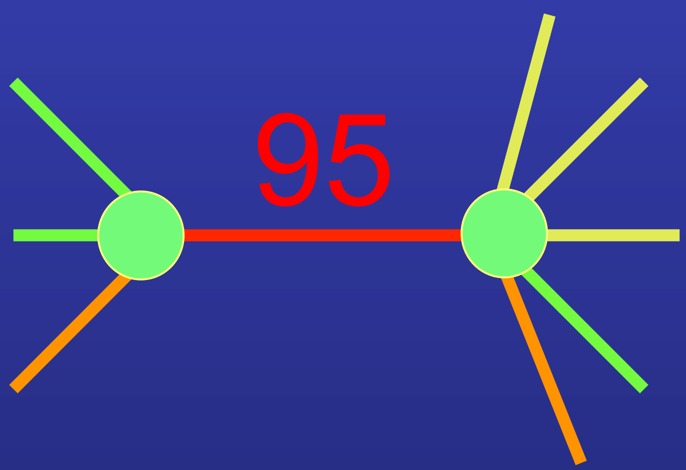
TOOLS TO ANALYZE PHYLOGENETIC INFORMATION FROM MULTIPLE GENES IN GENOMES:

Bipartition Spectra (Lento Plots)

BIPARTITION OF A PHYLOGENETIC TREE

Bipartition (or split) – a division of a phylogenetic tree into two parts that are connected by a single branch.
It divides a dataset into two groups, but it does not consider the relationships within each of the two groups.





* * * . . . . .

Yellow *vs* Rest

* * * . . . * *

compatible to illustrated bipartition

Orange vs Rest

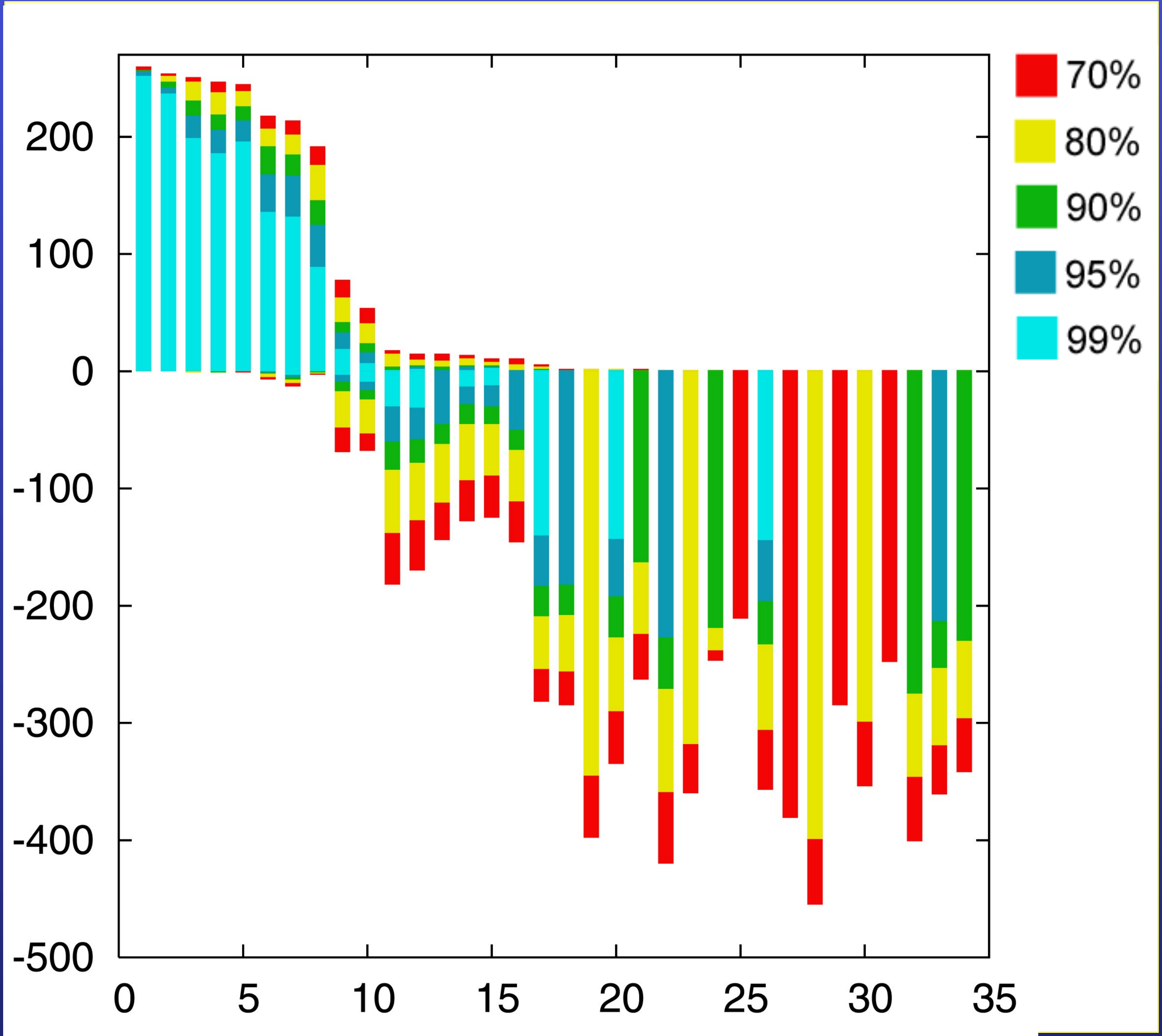
. . * . . . . *

incompatible to illustrated bipartition

“Lento”-plot of 34 supported bipartitions (out of 4082 possible)

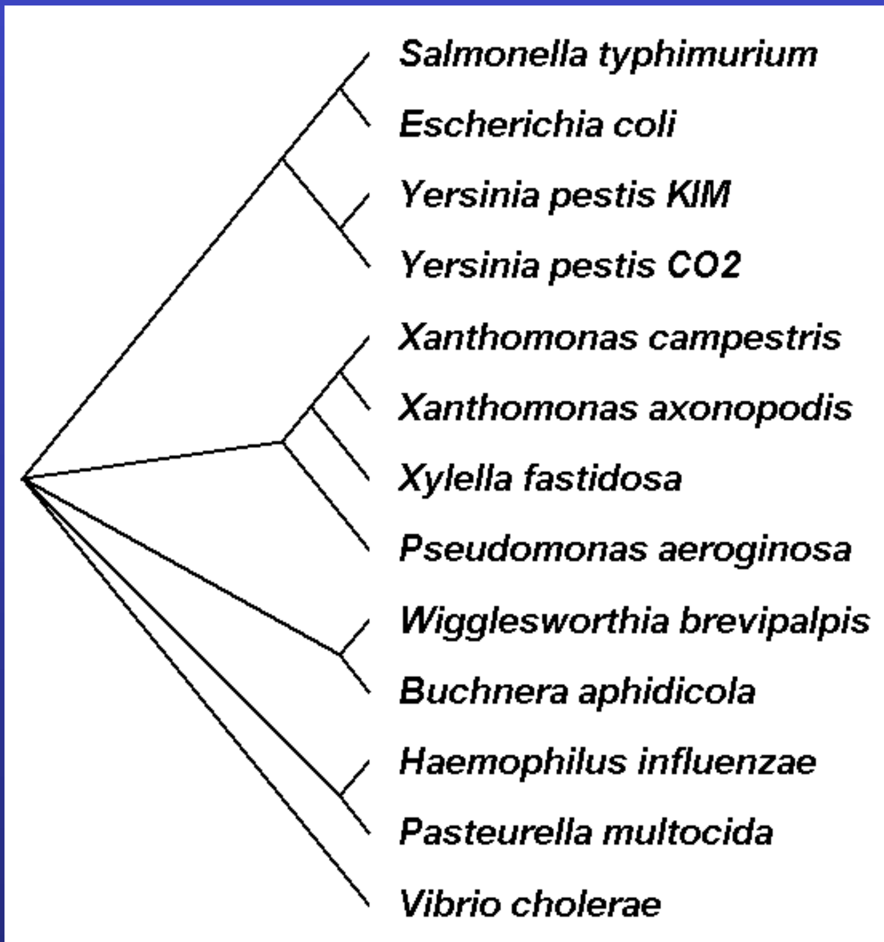
13 gamma-proteobacterial genomes (258 putative orthologs):

- E.coli
- Buchnera
- Haemophilus
- Pasteurella
- Salmonella
- Yersinia pestis (2 strains)
- Vibrio
- Xanthomonas (2 sp.)
- Pseudomonas
- Wigglesworthia



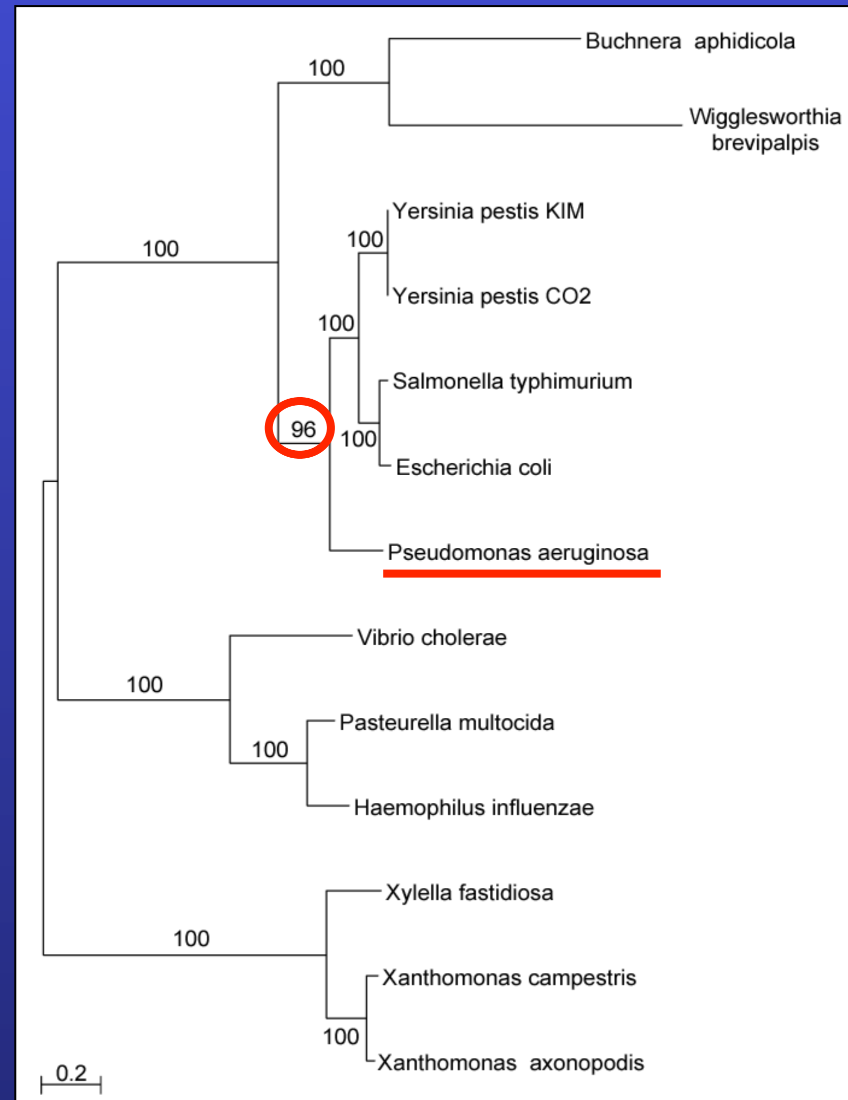
There are 13,749,310,575 possible unrooted tree topologies for 13 genomes

Consensus clusters of eight significantly supported bipartitions



only 258 genes analyzed

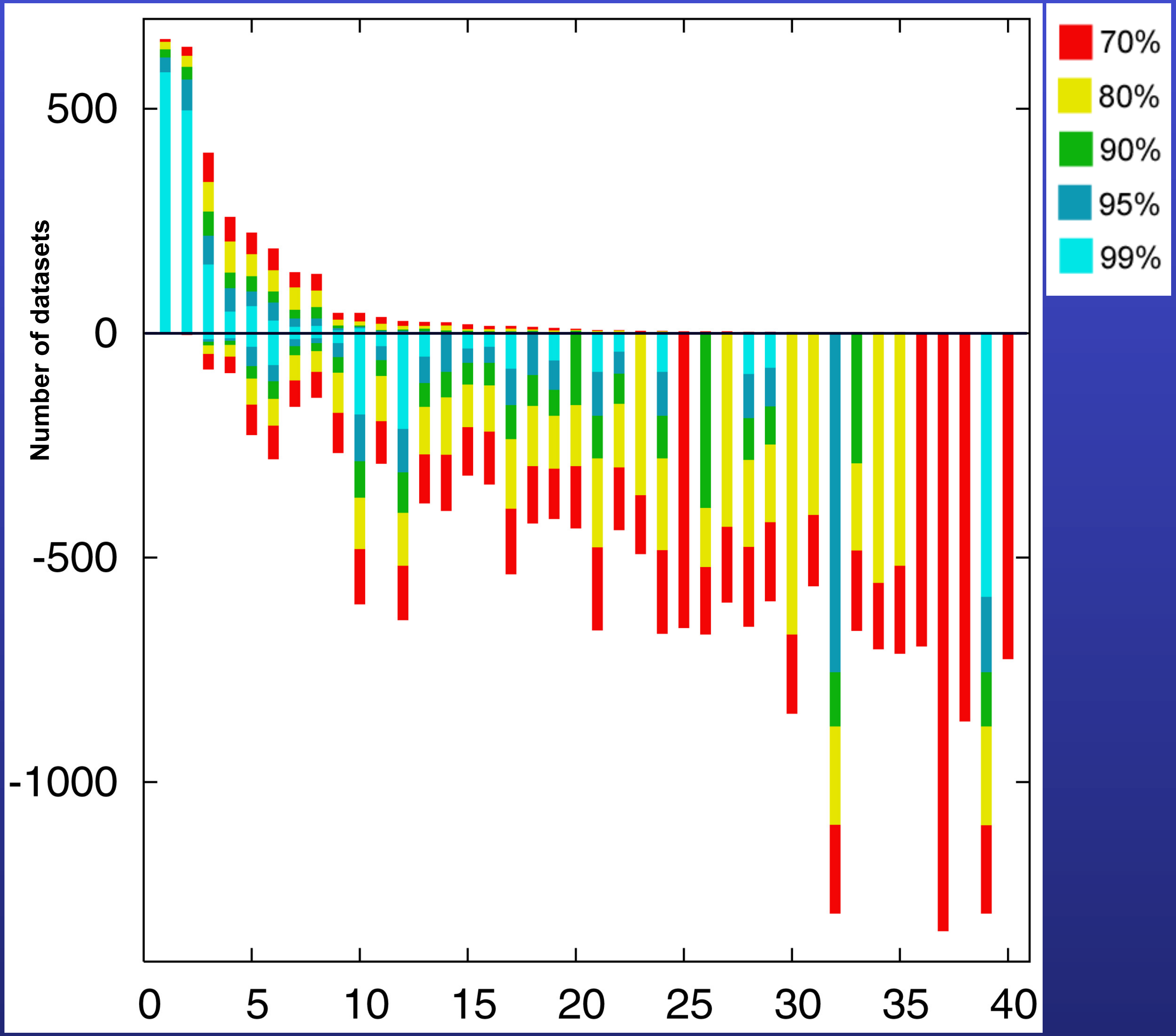
Phylogeny of putatively transferred gene (virulence factor homologs (mviN))



"Lento"-plot of supported bipartitions (out of 501 possible)

10 cyanobacteria:

- *Anabaena*
- *Trichodesmium*
- *Synechocystis* sp.
- *Prochlorococcus marinus* (3 strains)
- Marine *Synechococcus*
- *Thermo-synechococcus elongatus*
- *Gloeobacter*
- *Nostoc punctioforme*



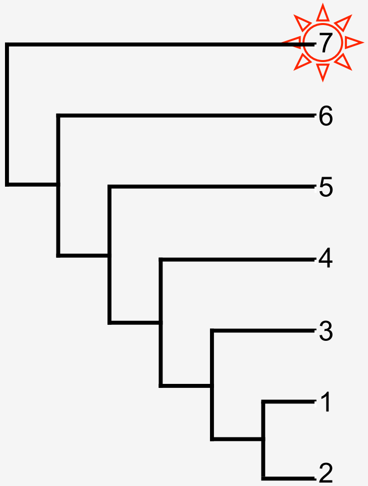
Based on 678 sets of orthologous genes

Zhaxybayeva, Lapierre and Gogarten, *Trends in Genetics*, 2004, **20(5)**: 254-260.

PROBLEMS WITH BIPARTITIONS (A)

Q_1={
4 5 6 7
1 5 6 7
2 5 6 7
3 5 6 7
3 4 6 7
1 4 6 7
2 4 6 7
2 3 6 7
1 3 6 7
1 2 6 7
1 2 3 7
1 2 4 7
1 3 4 7
2 3 4 7
2 3 5 7
1 3 5 7
1 2 5 7
1 4 5 7
2 4 5 7
3 4 5 7
3 4 5 6
1 4 5 6
2 4 5 6
2 3 5 6
1 3 5 6
1 2 5 6
1 2 3 6
1 2 4 6
1 3 4 6
2 3 4 6
2 3 4 5
1 3 4 5
1 2 4 5
1 2 3 5
1 2 3 4}

embedded quartets



B_1={
**.....,
***....,
****...,
*****..
}

bipartitions

A single rogue sequence that moves from one end of a Hennigian comb to the other changes all bipartition

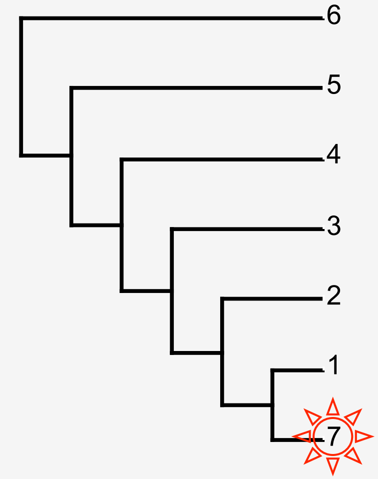
supported quartets

$Q1 \cap Q2 =$

{3 4 5 6, 1 4 5 6, 2 4 5 6, 2 3 5 6,
1 3 5 6, 1 2 5 6, 1 2 3 6, 1 2 4 6,
1 3 4 6, 2 3 4 6, 2 3 4 5, 1 3 4 5,
1 2 4 5, 1 2 3 5, 1 2 3 4}

supported bipartitions:

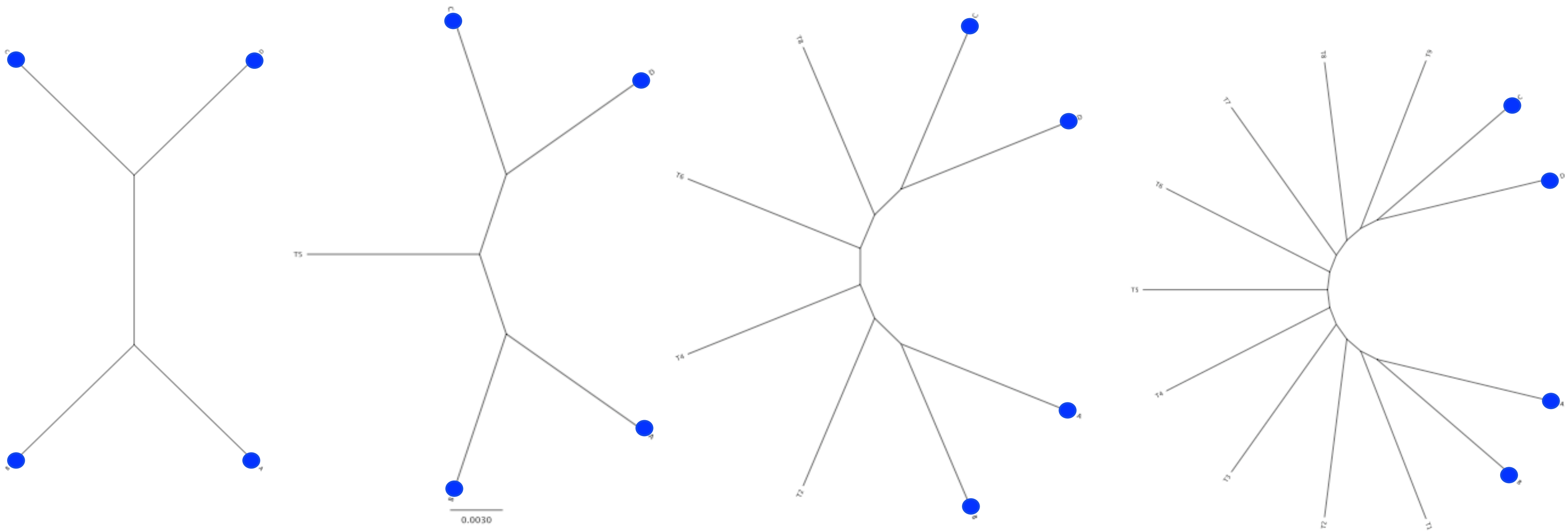
$B1 \cap B2 = \emptyset$



B_2={
.....,
**....*,
***...*,
****..*
}

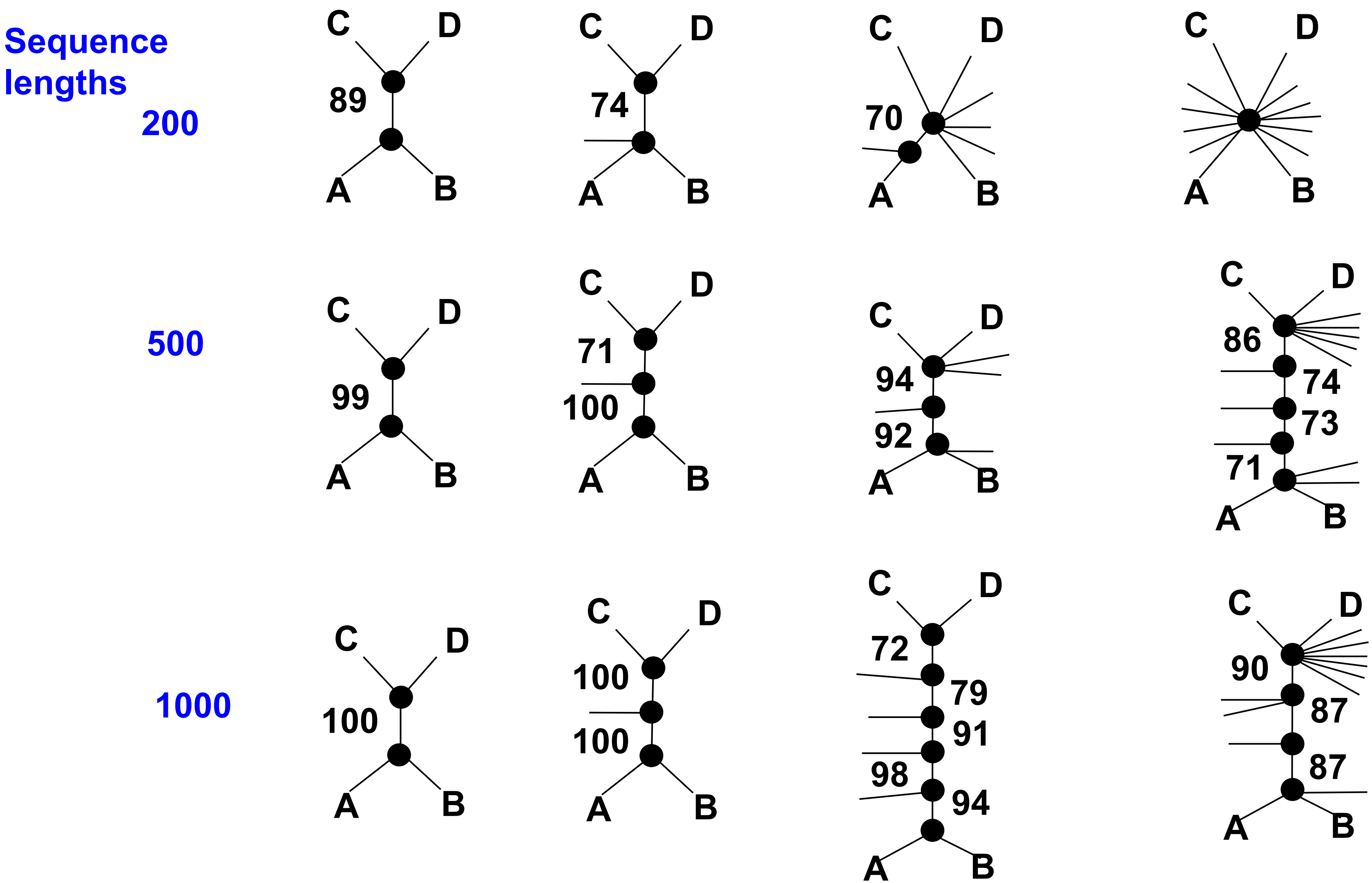
Q_2={
3 4 5 6
1 4 5 6
7 4 5 6
2 4 5 6
2 3 5 6
1 3 5 6
7 3 5 6
7 2 5 6
1 2 5 6
1 7 5 6
1 7 2 6
1 7 3 6
1 2 3 6
7 2 3 6
7 2 4 6
1 2 4 6
1 7 4 6
1 3 4 6
7 3 4 6
2 3 4 6
2 3 4 5
1 3 4 5
7 3 4 5
7 2 4 5
1 2 4 5
1 7 4 5
1 7 2 5
1 7 3 5
1 2 3 5
7 2 3 5
7 2 3 4
1 2 3 4
1 7 3 4
1 7 2 4
1 7 2 3}

Decay of bipartition support with number of OTUs



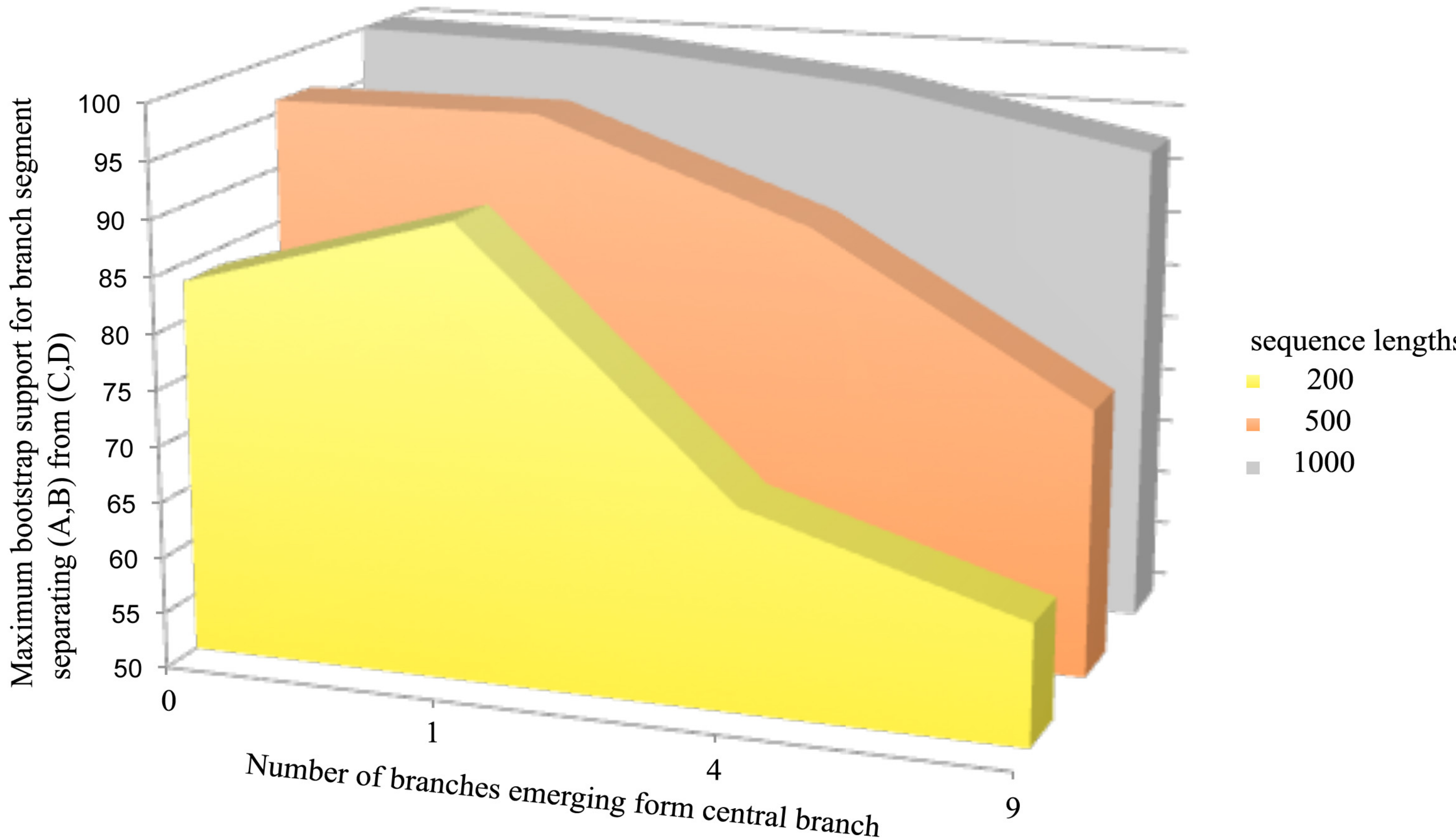
Phylogenies used for simulation

Example for decay of bipartition support with number of OTUs



Only branches with better than 70% bootstrap support are shown

Decay of bipartition support with number of OTUs

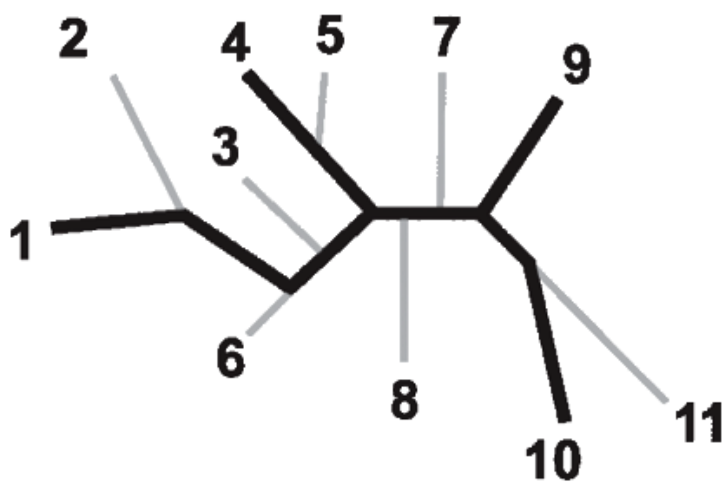


Each value is the average of 10 simulations using seq-gen.
Simulated sequences were evaluated using PHYML.
Model for simulation and evaluation WAG + Γ(α=1, 4 rate categories)

Bipartition Paradox:

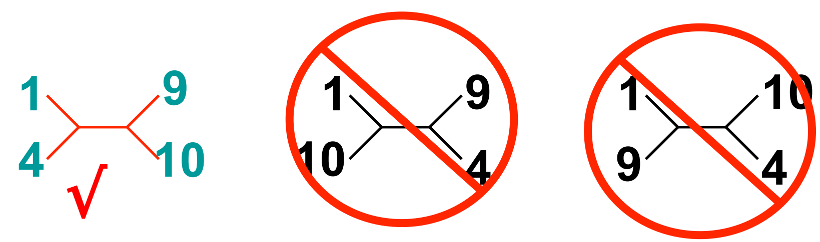
- The more sequences are added, the lower the support for bipartitions that include all sequences. The more data one uses, the lower the bootstrap support values become.
- This paradox disappears when only embedded splits for 4 sequences are considered.

Bootstrap support values for embedded quartets



— + — : tree calculated from one pseudo-sample generated by bootstraping from an alignment of one gene family present in 11 genomes

— : embedded quartet for genomes 1, 4, 9, and 10 .
This bootstrap sample supports the topology ((1,4),9,10).



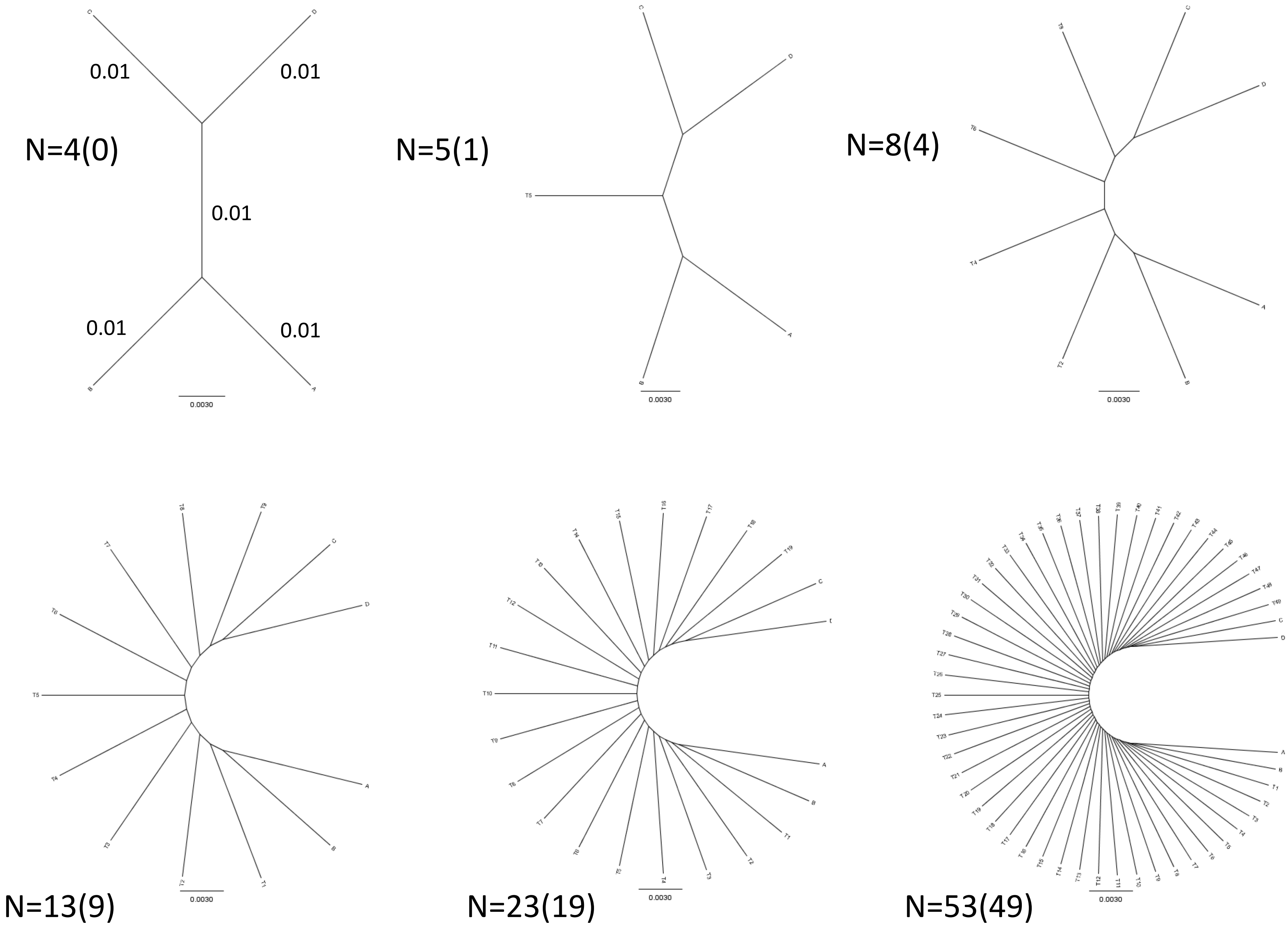
Quartet spectral analyses of genomes iterates over three loops:

- ➢Repeat for all bootstrap samples.
- ➢Repeat for all possible embedded quartets.
- ➢Repeat for all gene families.

Zhaxybayeva et al. 2006, Genome Research, 16(9):1099-108

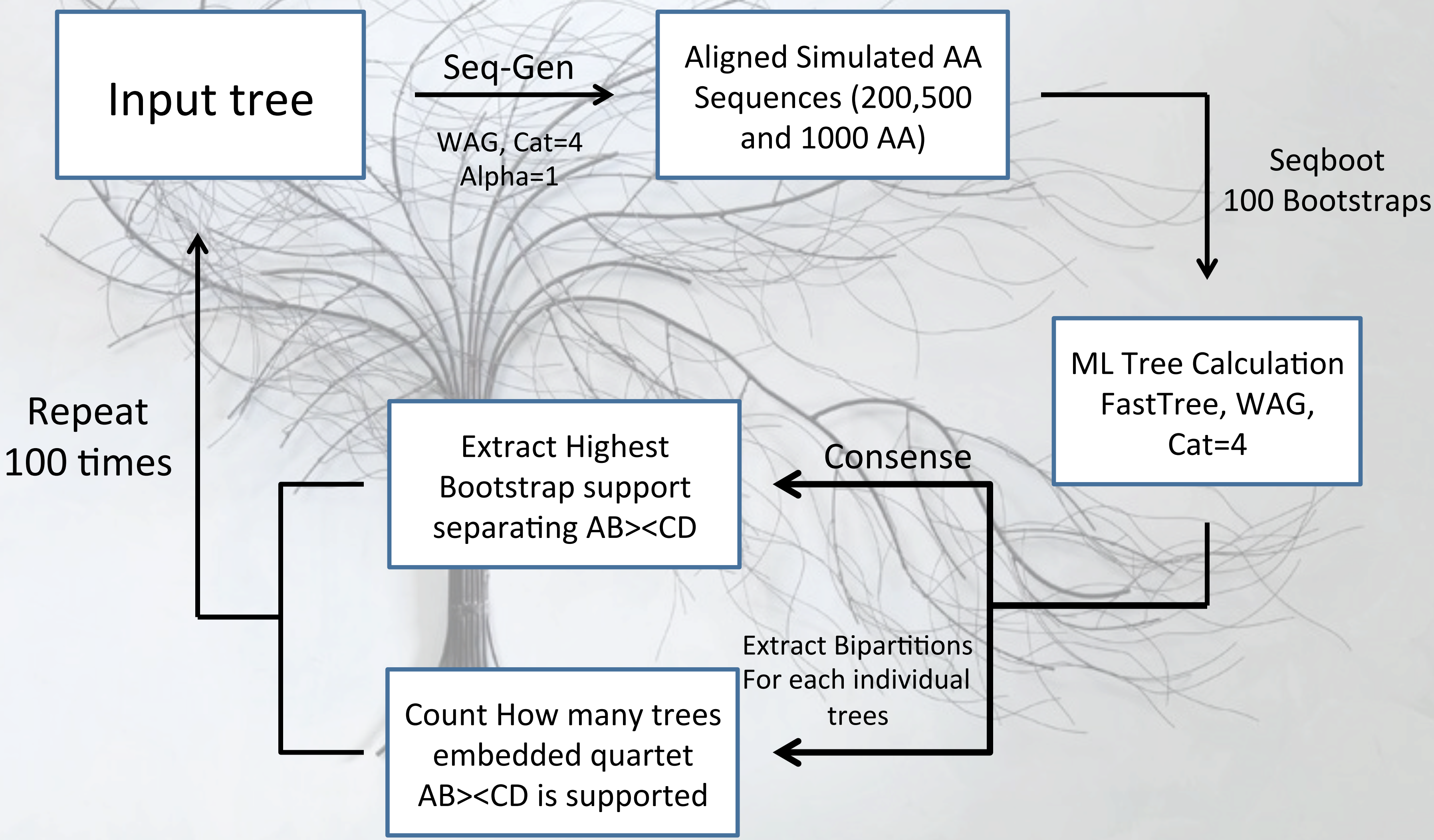
Boostrap Support Values for Embedded Quartets vs. Bipartitions:

Performance evaluation using sequence simulations and phylogenetic reconstructions



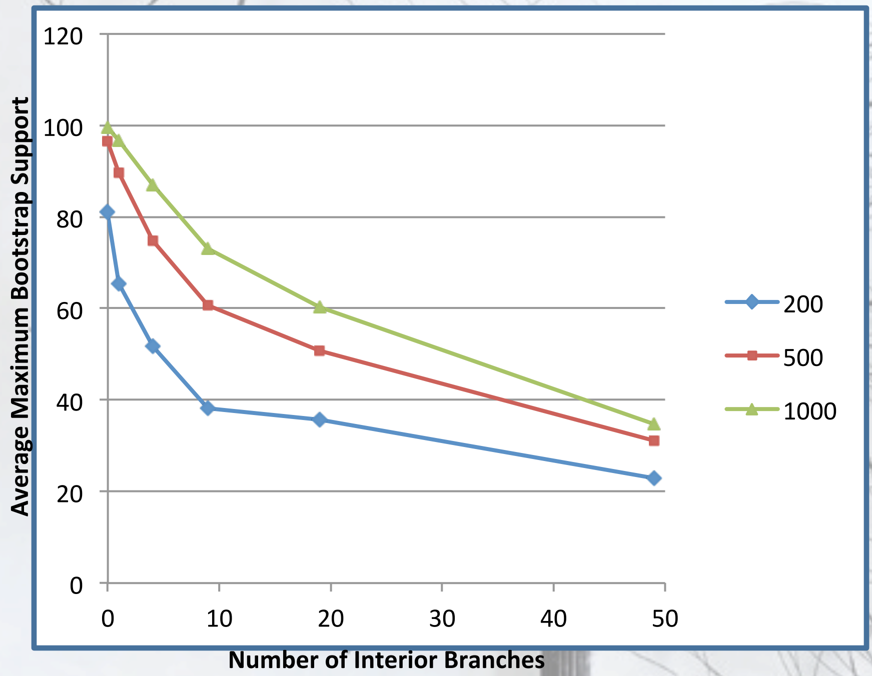
N=4(0)
0.01
0.01
0.01
0.01
0.01
0.0030
N=5(1)
0.0030
N=8(4)
0.0030
N=13(9)
0.0030
N=23(19)
0.0030
N=53(49)
0.0030

Methodology :

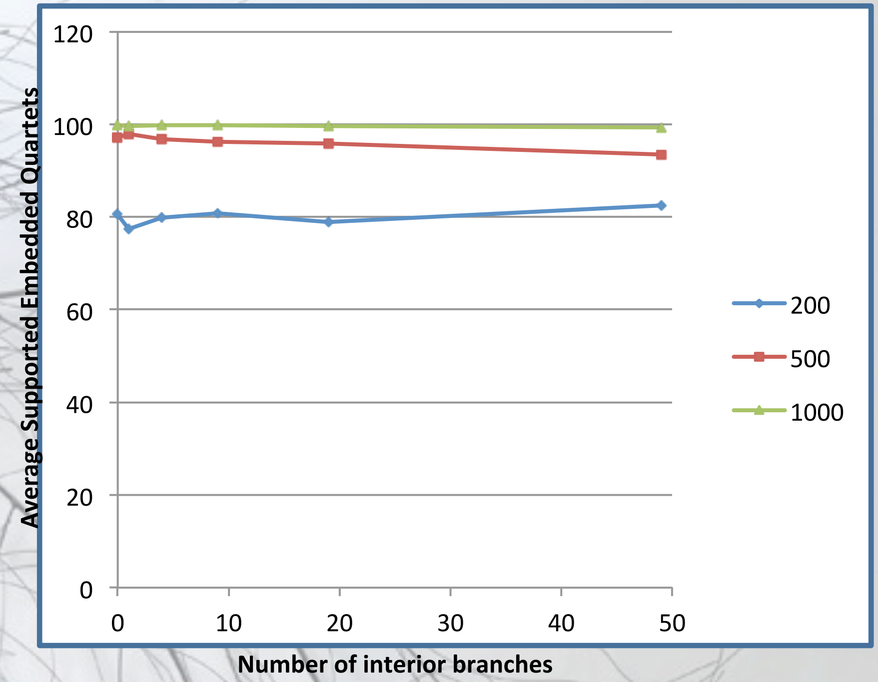


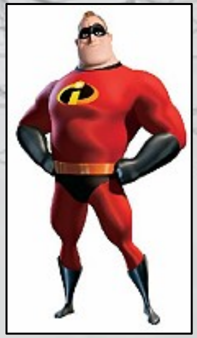
Results :

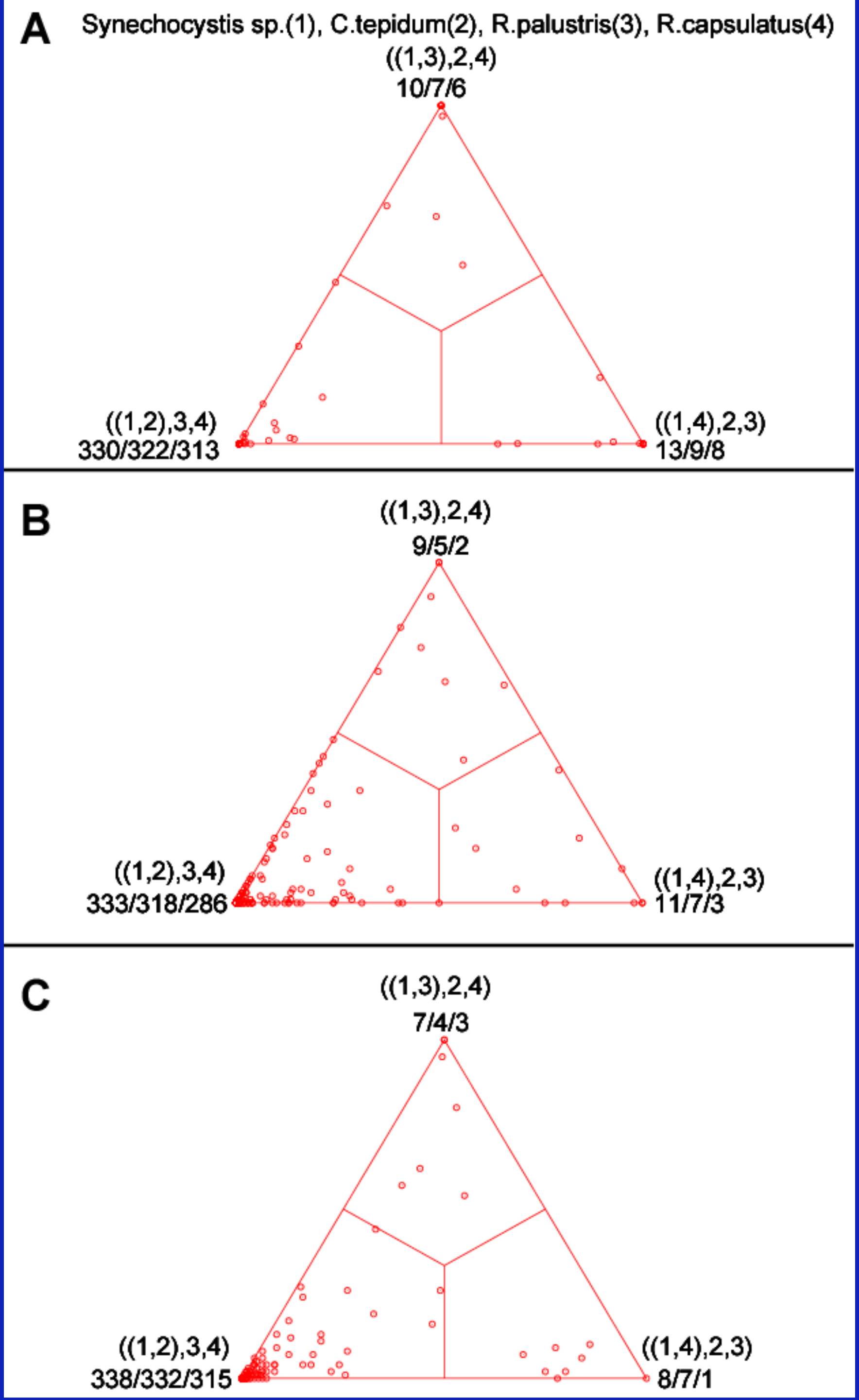
Maximum Bootstrap Support value for Bipartition separating (AB) and (CD)



Maximum Bootstrap Support value for embedded Quartet (AB),(CD)







Zhaxybayeva and Gogarten, *BMC Genomics* 2003 4: 37

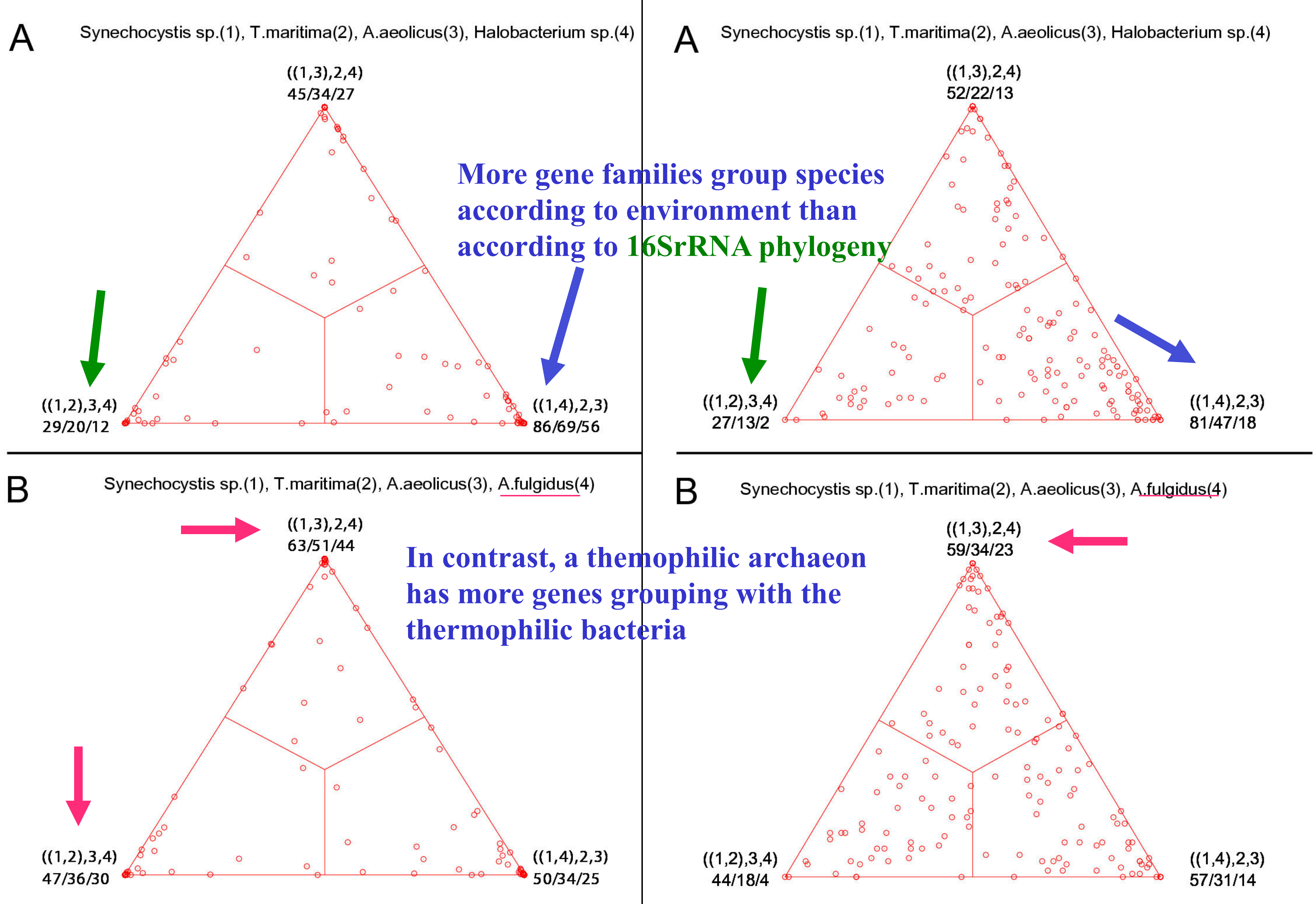
COMPARISON OF DIFFERENT SUPPORT MEASURES

A: mapping of posterior probabilities according to Strimmer and von Haeseler

B: mapping of bootstrap support values

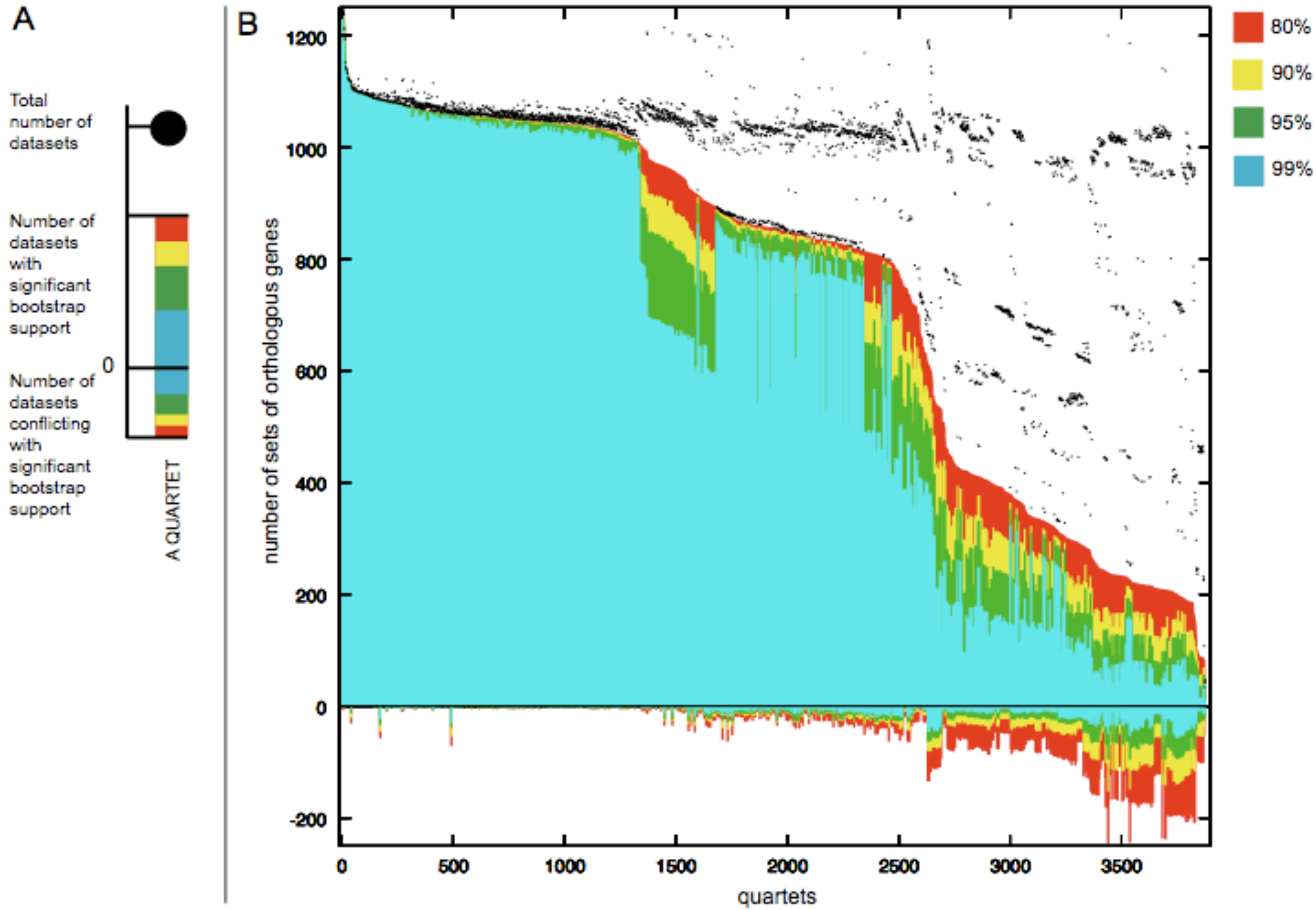
C: mapping of bootstrap support values from extended datasets

ml-mapping *versus* bootstrap values from extended datasets



More gene families group species according to environment than according to 16SrRNA phylogeny

In contrast, a themophilic archaeon has more genes grouping with the thermophilic bacteria



Quartet decomposition analysis of 19 *Prochlorococcus* and marine *Synechococcus* genomes. Quartets with a very short internal branch or very long external branches as well those resolved by less than 30% of gene families were excluded from the analyses to minimize artifacts of phylogenetic

the gradualist point of view

Evolution occurs within populations where the fittest organisms have a selective advantage. Over time the advantages genes become fixed in a population and the population gradually changes.
This reasoning (with many more details) is known as the modern synthesis.
Note: this is not in contradiction to the the theory of neutral evolution. (which says what ?)

Processes that MIGHT go beyond inheritance with variation and selection?

- Horizontal gene transfer and recombination
- Polyploidization (botany, vertebrate evolution) see here
- Fusion and cooperation of organisms (Kefir, lichen, also the eukaryotic cell)
- Targeted mutations (?), genetic memory (?) (see Foster's and Hall's reviews on directed/adaptive mutations; see here for a counterpoint)
- Random genetic drift
- Gratuitous complexity
- Selfish genes (who/what is the subject of evolution??)
- Parasitism, altruism, Morons
- Mutationism, hopeful monsters (see here for a critical discussion by Arlin Stolzfus)

selection versus drift

see Kent Holsinger's java simulations at
http://darwin.eeb.uconn.edu/simulations/simulations.html
The law of the gutter.
compare drift versus select + 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.

$s=0$

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 (the rate with which mutations are fixed in a lineage) =
frequency with which new alleles are generated * Probability of fixation=
u*2N *1/(2N) = u
Therefore:
If f s=0, the substitution rate is independent of population size, and equal to the mutation rate !!!!

This is the reason that there is hope that the molecular clock might sometimes work.

Fixation time due to drift alone:

$\mathbf{t}_{av}=4*N_e$ generations
(N_e=effective population size; For n discrete generations
$N_e= n/(1/N_1+1/N_2+\ldots..1/N_n)$

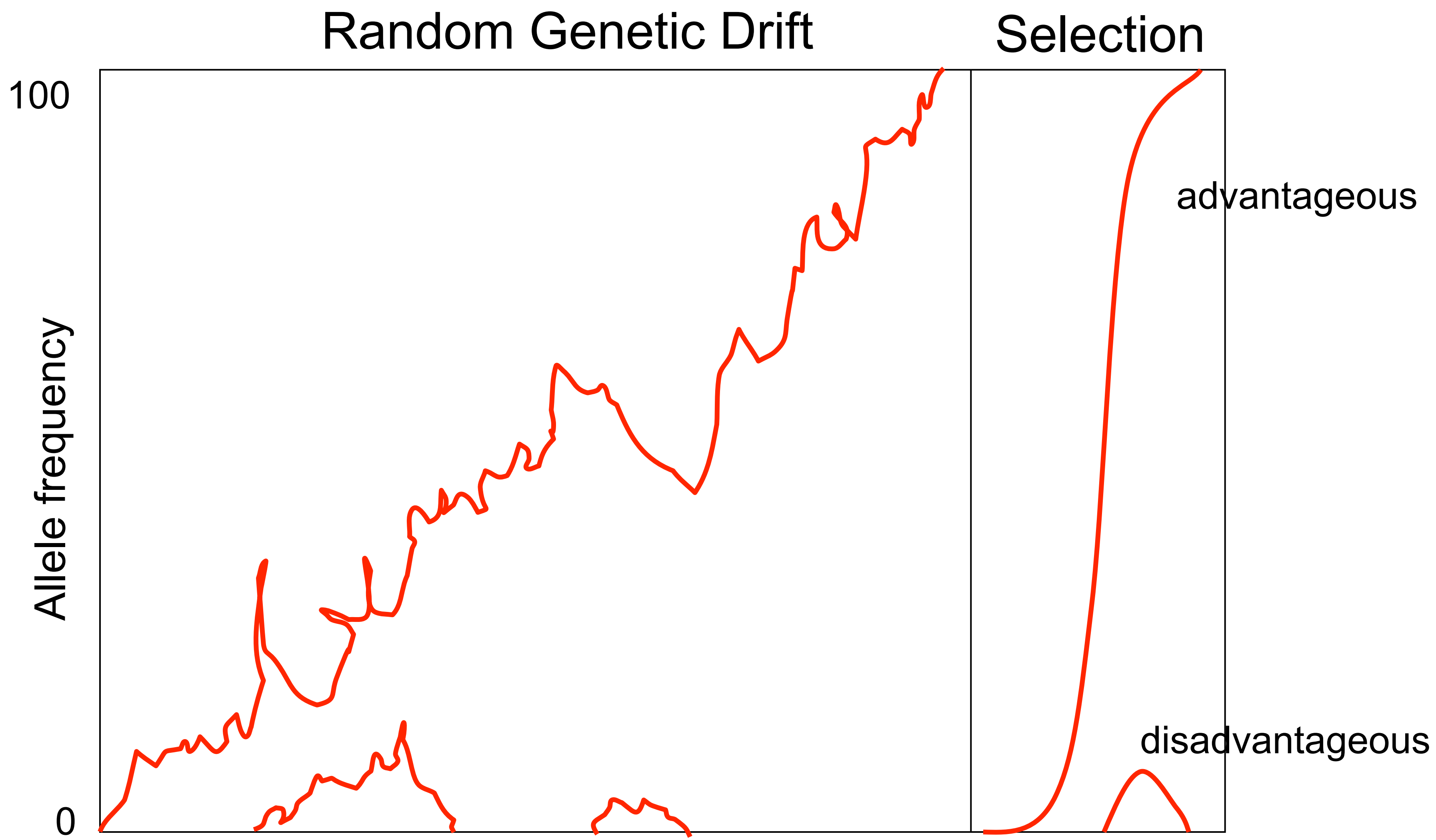
s>0

Time till fixation on average:
t_{av}= (2/s) ln (2N) generations
(also true for mutations with negative “s” ! discuss among yourselves)

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 wite which neutral mutations are fixed.



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

Positive selection

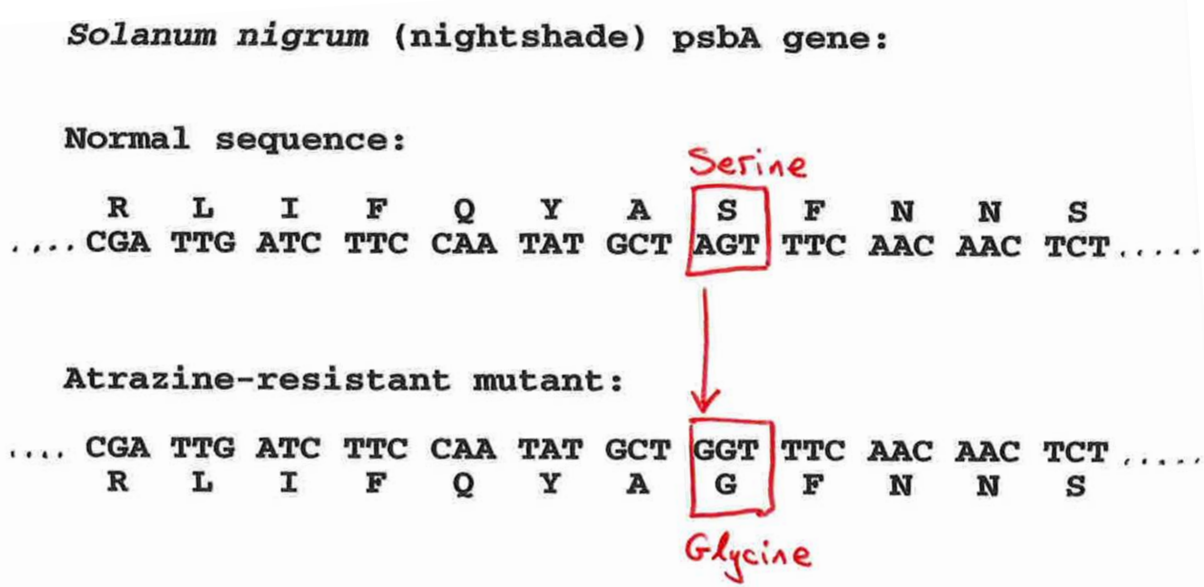
- A new allele (mutant) confers some <u>increase</u> in the **fitness** of the organism
- Selection acts to favour this allele
- Also called adaptive selection or Darwinian selection.

NOTE: **Fitness** = ability to survive and <u>reproduce</u>

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

- A new allele (mutant) confers some <u>decrease</u> in the fitness of the organism
- Selection acts to remove this allele

- Also called purifying selection

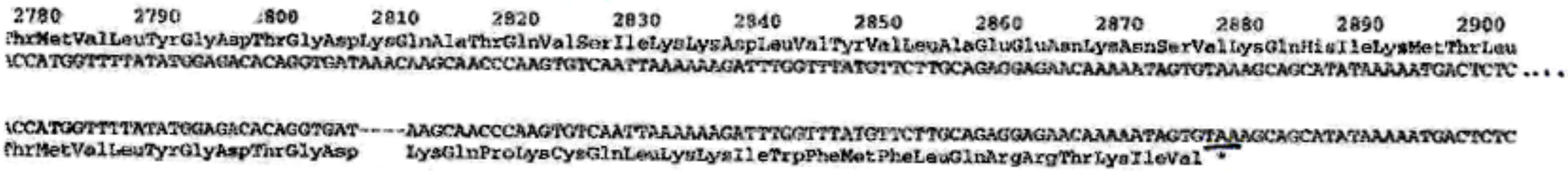
Modified from from www.tcd.ie/Genetics/staff/Aoife/GE3026/GE3026_1+2.ppt

Deleterious allele

Human breast cancer gene, BRCA2

5% of breast cancer cases are familial
Mutations in BRCA2 account for 20% of familial cases

Normal (wild type) allele



Mutant allele (Montreal 440 Family)

Stop codon

4 base pair deletion
Causes frameshift

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

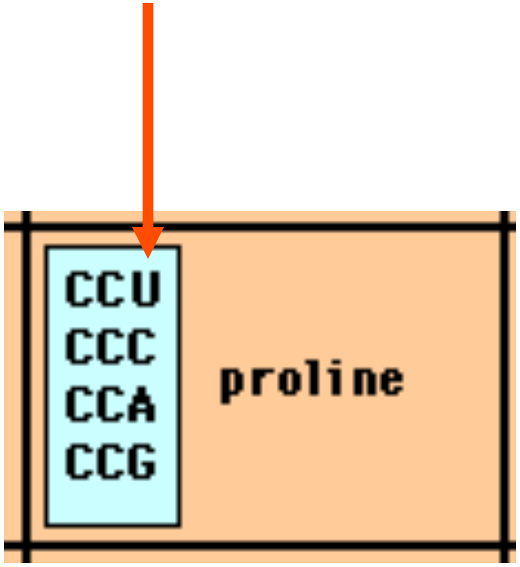
- Replacement of one nucleotide by another
- Synonymous (Doesn't change amino acid)
  - Rate sometimes indicated by Ks
  - Rate sometimes indicated by d_s
- Non-Synonymous (Changes Amino Acid)
  - Rate sometimes indicated by Ka
  - Rate sometimes indicated by d_n

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

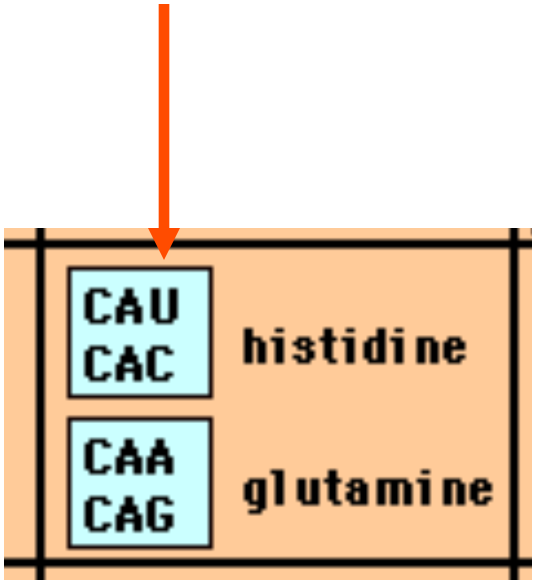
Genetic Code



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

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 <u>possible</u> 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

dambe

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

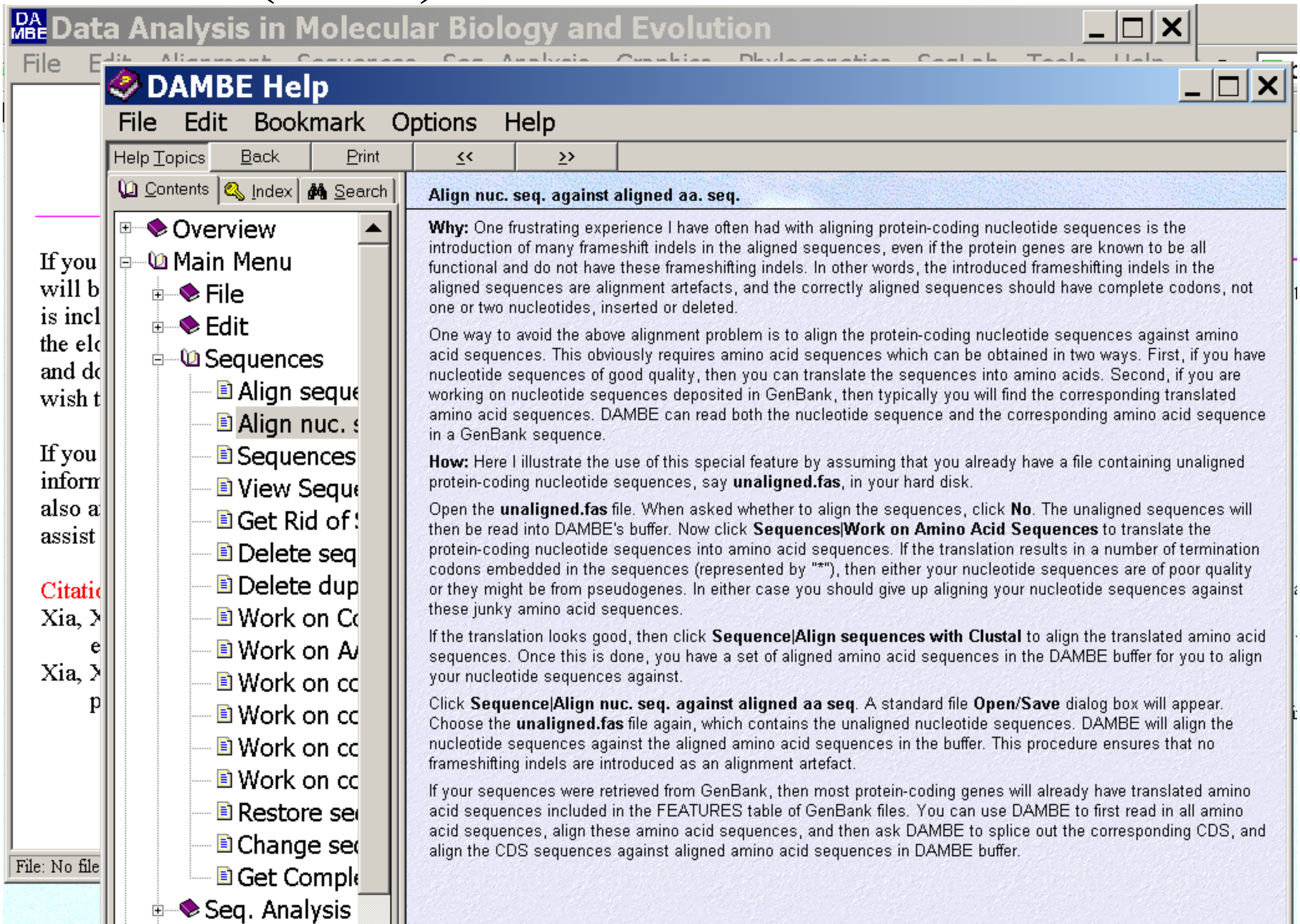
One is DAMBE (only for windows). This is a handy program for a lot of things, including reading a lot of different formats, calculating phylogenies, it even runs codeml (from PAML) for you.

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.

dambe (cont)


Data Analysis in Molecular Biology and Evolution
DAMBE Help
File Edit Bookmark Options Help
Help Topics
Back
Print
<<
>>
Contents
Index
Search
Overview
Main Menu
File
Edit
Sequences
Seq. Analysis
Align nuc. seq. against aligned aa. seq.
Why: One frustrating experience I have often had with aligning protein-coding nucleotide sequences is the introduction of many frameshift indels in the aligned sequences, even if the protein genes are known to be all functional and do not have these frameshifting indels. In other words, the introduced frameshifting indels in the aligned sequences are alignment artefacts, and the correctly aligned sequences should have complete codons, not one or two nucleotides, inserted or deleted.
One way to avoid the above alignment problem is to align the protein-coding nucleotide sequences against amino acid sequences. This obviously requires amino acid sequences which can be obtained in two ways. First, if you have nucleotide sequences of good quality, then you can translate the sequences into amino acids. Second, if you are working on nucleotide sequences deposited in GenBank, then typically you will find the corresponding translated amino acid sequences. DAMBE can read both the nucleotide sequence and the corresponding amino acid sequence in a GenBank sequence.
How: Here I illustrate the use of this special feature by assuming that you already have a file containing unaligned protein-coding nucleotide sequences, say unaligned.fas, in your hard disk.
Open the unaligned.fas file. When asked whether to align the sequences, click No. The unaligned sequences will then be read into DAMBE's buffer. Now click Sequences|Work on Amino Acid Sequences to translate the protein-coding nucleotide sequences into amino acid sequences. If the translation results in a number of termination codons embedded in the sequences (represented by "*"), then either your nucleotide sequences are of poor quality or they might be from pseudogenes. In either case you should give up aligning your nucleotide sequences against these junky amino acid sequences.
If the translation looks good, then click Sequence|Align sequences with Clustal to align the translated amino acid sequences. Once this is done, you have a set of aligned amino acid sequences in the DAMBE buffer for you to align your nucleotide sequences against.
Click Sequence|Align nuc. seq. against aligned aa seq. A standard file Open/Save dialog box will appear. Choose the unaligned.fas file again, which contains the unaligned nucleotide sequences. DAMBE will align the nucleotide sequences against the aligned amino acid sequences in the buffer. This procedure ensures that no frameshifting indels are introduced as an alignment artefact.
If your sequences were retrieved from GenBank, then most protein-coding genes will already have translated amino acid sequences included in the FEATURES table of GenBank files. You can use DAMBE to first read in all amino acid sequences, align these amino acid sequences, and then ask DAMBE to splice out the corresponding CDS, and align the CDS sequences against aligned amino acid sequences in DAMBE buffer.

aa based nucleotide alignments (cont)

An alternative is the tranalign program that is part of the emboss package. On bbcxsrv1 you can invoke the program by typing tranalign.

Instructions and program description are here .

If you want to use your own dataset in the lab on Monday, generate a codon based alignment with either *dambe* or *tranalign* and save it as a nexus file **and** as a phylip formated multiple sequence file (using either clustalw, PAUP (export or tonexus), dambe, or **readseq** on the web)

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 any 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;
```

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.

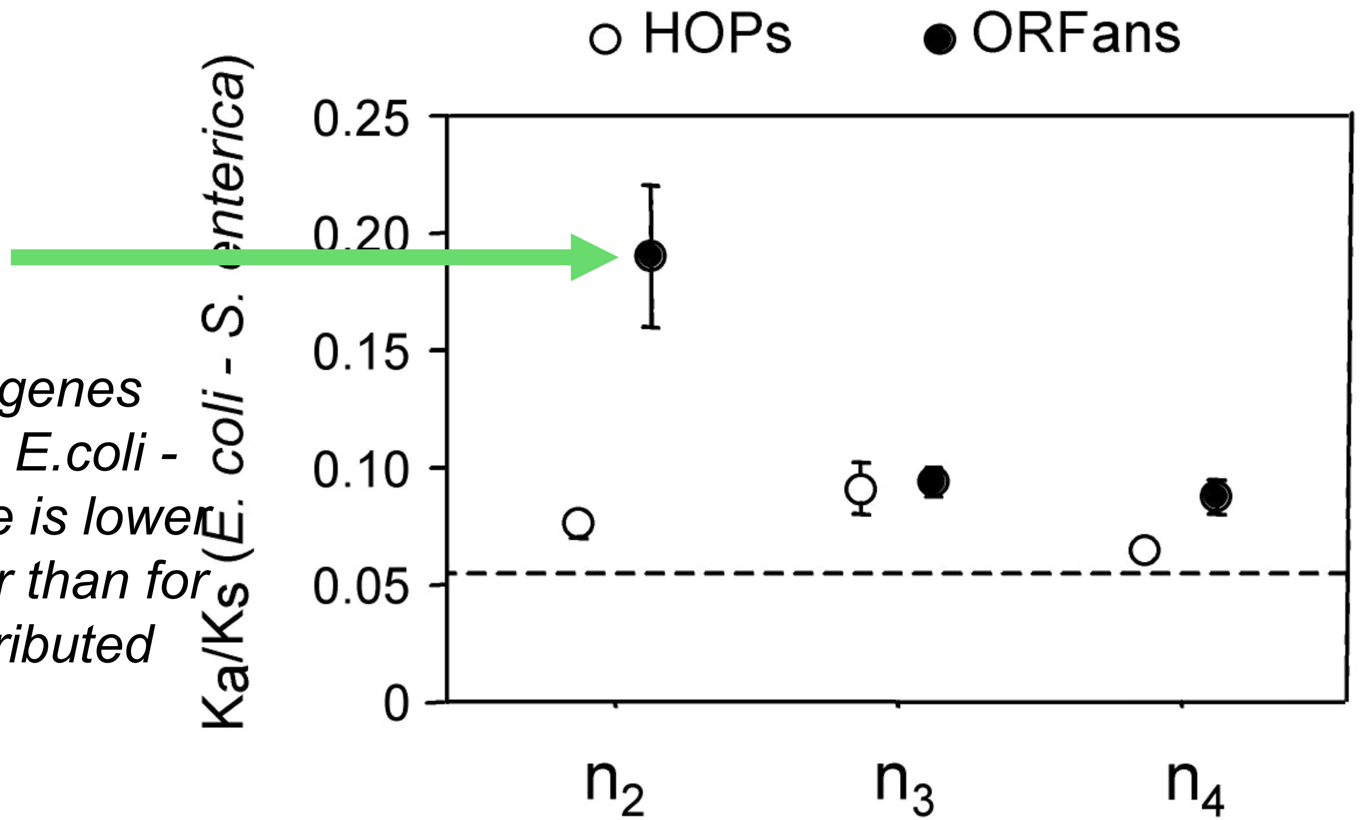


Fig. 3 from Vincent Daubin and Howard Ochman, *Genome Research 14:1036-1042, 2004*

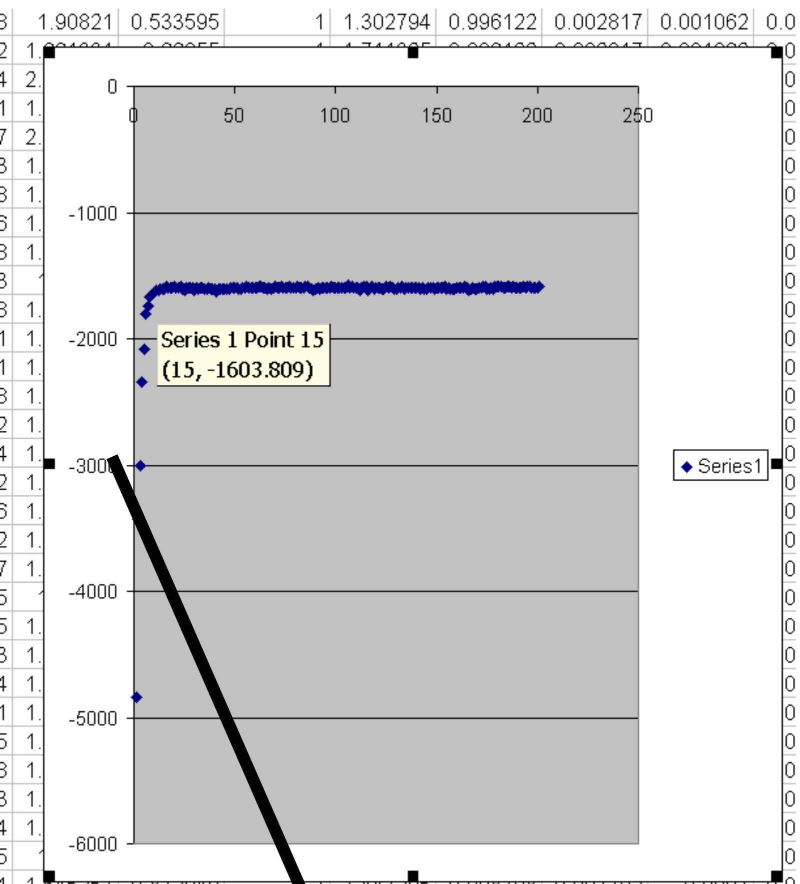
Trunk-of-my-car analogy: Hardly anything in there is the is the result of providing a selective advantage. Some items are removed quickly (purifying selection), some are useful under some conditions, but most things do not alter the fitness.

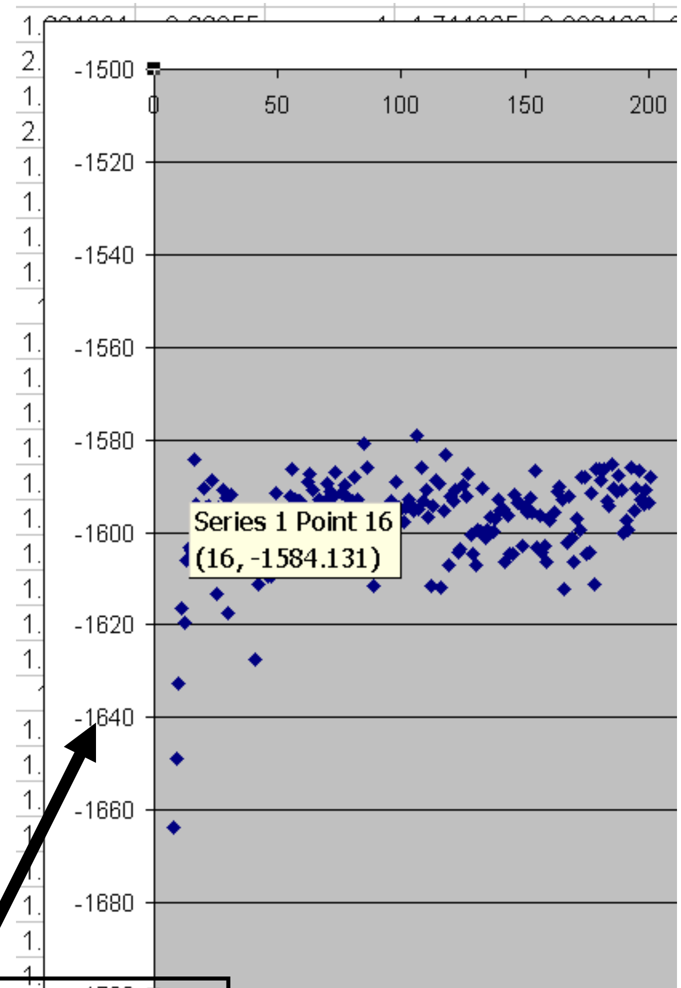
Could some of the inferred purifying selection be due to the acquisition of novel detrimental characteristics (e.g., protein toxicity)?

MrBayes analyzing the *.nex.p file

1. **The easiest is to load the file into excel (if your alignment is too long, you need to load the data into separate spreadsheets – see here execise 2 item 2 for more info)**
2. **plot LogL to determine which samples to ignore**
3. **for each codon calculate the the average probability (from the samples you do not ignore) that the codon belongs to the group of codons with omega>1.**
4. **plot this quantity using a bar graph.**

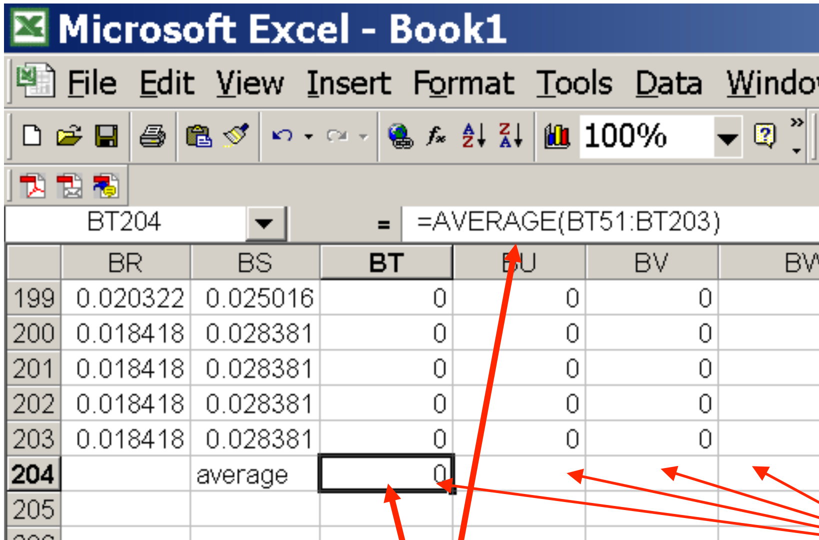
plot LogL to determine which samples to ignore

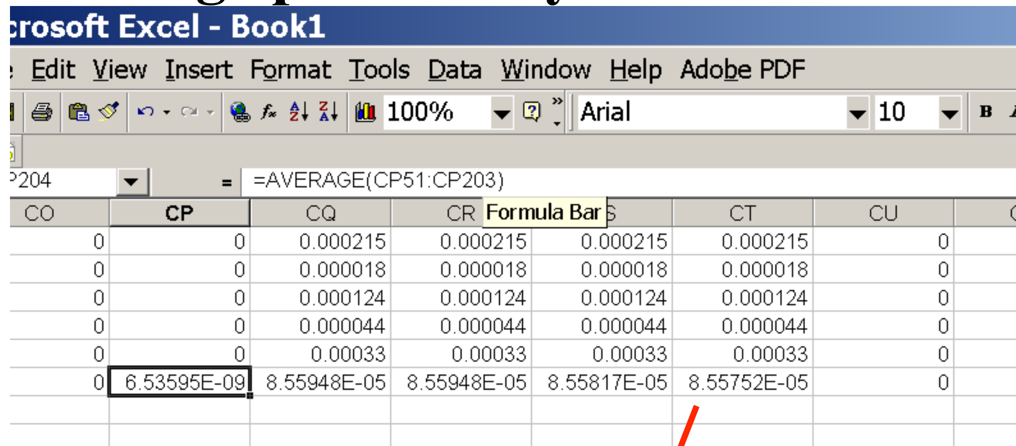




the same after rescaling the y-axis

for each codon calculate the the average probability

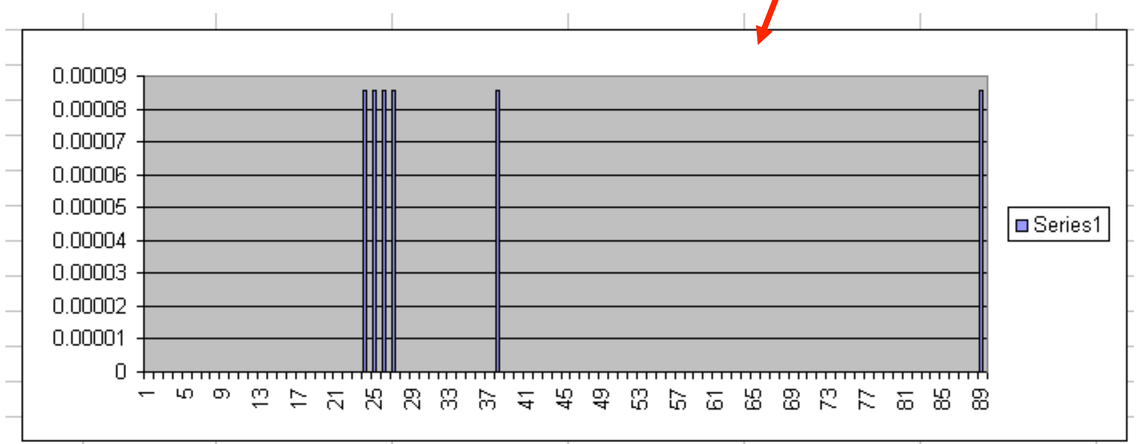




copy paste formula

plot row

enter formula



MrBayes on bbcxrv1

If you do this for your own data,
- **run the procedure first for only 50000 generations (takes about 30 minutes) to check that everthing works as expected,**
- **then run the program overnight for at least 500 000 generations.**
- **Especially, if you have a large dataset, do the latter twice and compare the results for consistency. (I prefer two runs over 500000 generations each over one run over a million generations.)**

The preferred wa to run mrbayes is to use the command line:

`>mb`

Do example on threonlyRS

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:

```
      6    467           I
gi|1613157 ---------- MSDNDTIVAQ ATPPGRGGVG ILRISGFKAR EVAETVLGKL
gi|2212798 ---------- MSTTDTIVAQ ATPPGRGGVG ILRVSGRAAS EVAHAVLGKL
gi|1564003 MALIQSCSGN TMTTDTIVAQ ATAPGRGGVG IIRVSGPLAA HVAQTVTGRT
gi|1560076 ---------M QAATETIVAI ATAQGRGGVG IVRVSGPLAG QMAVAVSGRQ
gi|2123365 -----MN--- -ALPSTIVAI ATAAGTGGIG IVRLSGPQSV QIAAALGIAG
gi|1583936 -----MSQRS TKMGDTIAAI ATASGAAGIG IIRLSGSLIK TIATGLGMTT

           PKPRYADYLP FKDADGSVLD QGIALWFPGP NSFTGEDVLE LQGHGGPVIL
           PKPRYADYLP FKDVDGSTLD QGIALYFPGP NSFTGEDVLE LQGHGGPVIL
           LRPRYAEYLP FTDEDGQQLD QGIALFFPNP HSFTGEDVLE LQGHGGPVVM
           LKARHAHYGP FLDAGGQVID EGLSLYFPGP NSFTGEDVLE LQGHGGPVVL
           LQSRHARYAR FRDAQGEVID DGIAVWFPAP HSFTGEEVVE LQGHGSPVLL
           LRPRYAHYTR FLDVQDEVID DGLALWFPAP HSFTGEDVLE LQGHGSPLLL
```

```
 5   855   I

human
goat-cow
rabbit
rat
marsupial
1
GTG CTG TCT CCT GCC GAC AAG ACC AAC GTC AAG GCC GCC TGG GGC AAG GTT GGC GCG CAC
... ... ... G.C ... ... ... T.. ..T ... ... ... ... ... ... ... ... ... .GC A..
... ... ... ..C ..T ... ... ... ... A.. ... A.T ... ... .AA ... A.C ... AGC ...
... ..C ... G.A .AT ... ..A ... ... A.. ... AA. TG. ... ..G ... A.. ..T .GC ..T
... ..C ..G GA. ..T ... ... ..T C.. ..G ..A ... AT. ... ..T ... ..G ..A .GC ...
61

GCT GGC GAG TAT GGT GCG GAG GCC CTG GAG AGG ATG TTC CTG TCC TTC CCC ACC ACC AAG
... ..A .CT ... ..C ..A ... ..T ... ... ... ... ... ... AG. ... ... ... ... ...
.G. ... ... ... ..C ..C ... ... G.. ... ... ... ... T.. GG. ... ... ... ... ...
.G. ..T ..A ... ..C .A. ... ... ..A C.. ... ... ... GCT G.. ... ... ... ... ...
..C ..T .CC ..C .CA ..T ..A ..T ..T .CC ..A .CC ... ..C ... ... ... ..T ... ..A
```

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.

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.

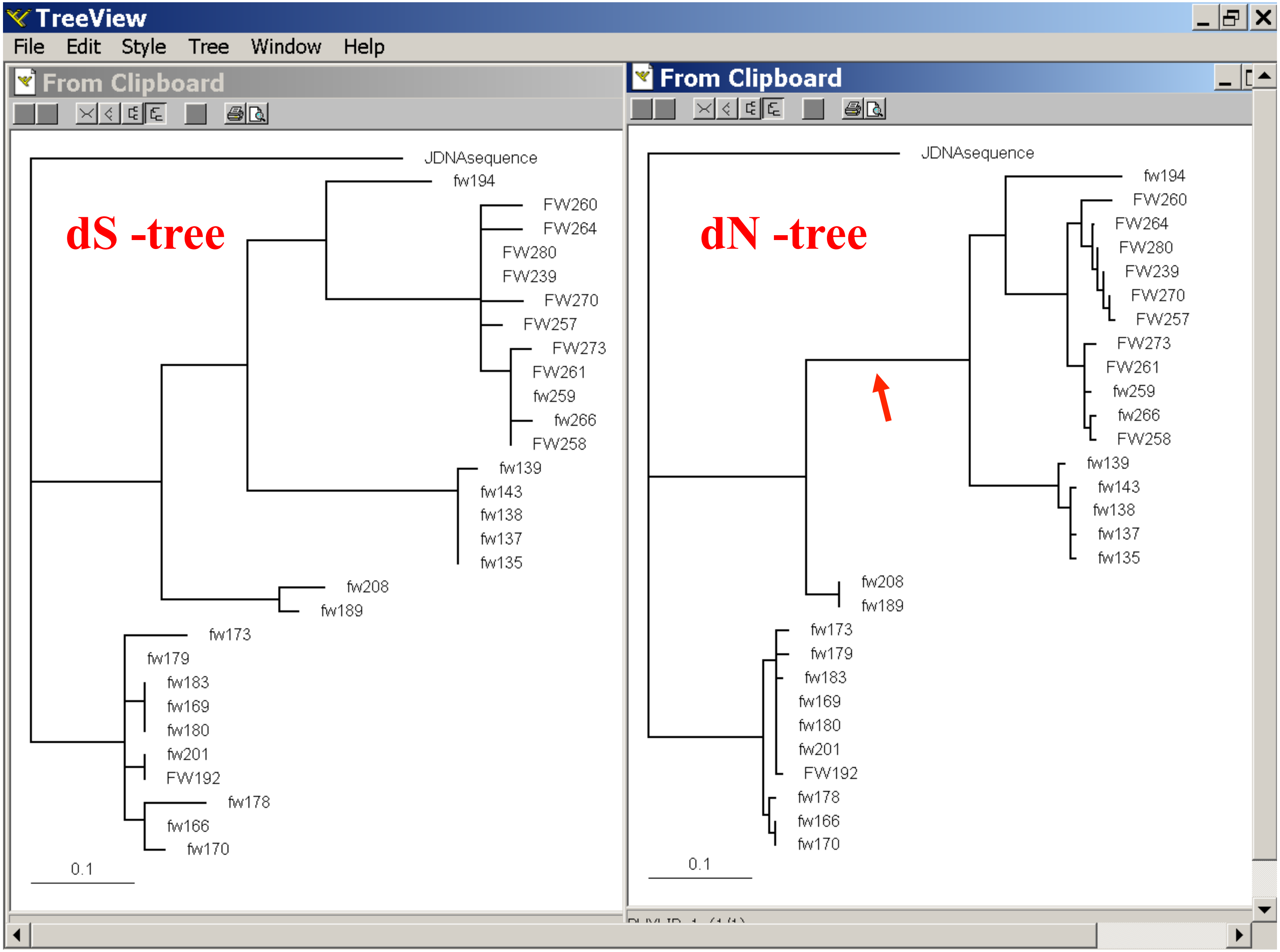
The output is written, as directed by the control file, into a file called Hv1.branch.codeml_out

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

Also, it might be a good idea to repeat the analyses on parts of the sequence (using the same tree). In this case the sequences encode a family of spider toxins that include the mature toxin, a propeptide and a signal sequence (see here for more information).

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

PAML – codeml – branch model



where to get help

read the manuals and help files
check out the discussion boards at http://www.rannala.org/phpBB2/

else

there is a new program on the block called hy-phy
(=hypothesis testing using phylogenetics).

The easiest is probably to run the analyses on the authors datamonkey.

Discussion: Other ways to detect positive selection?

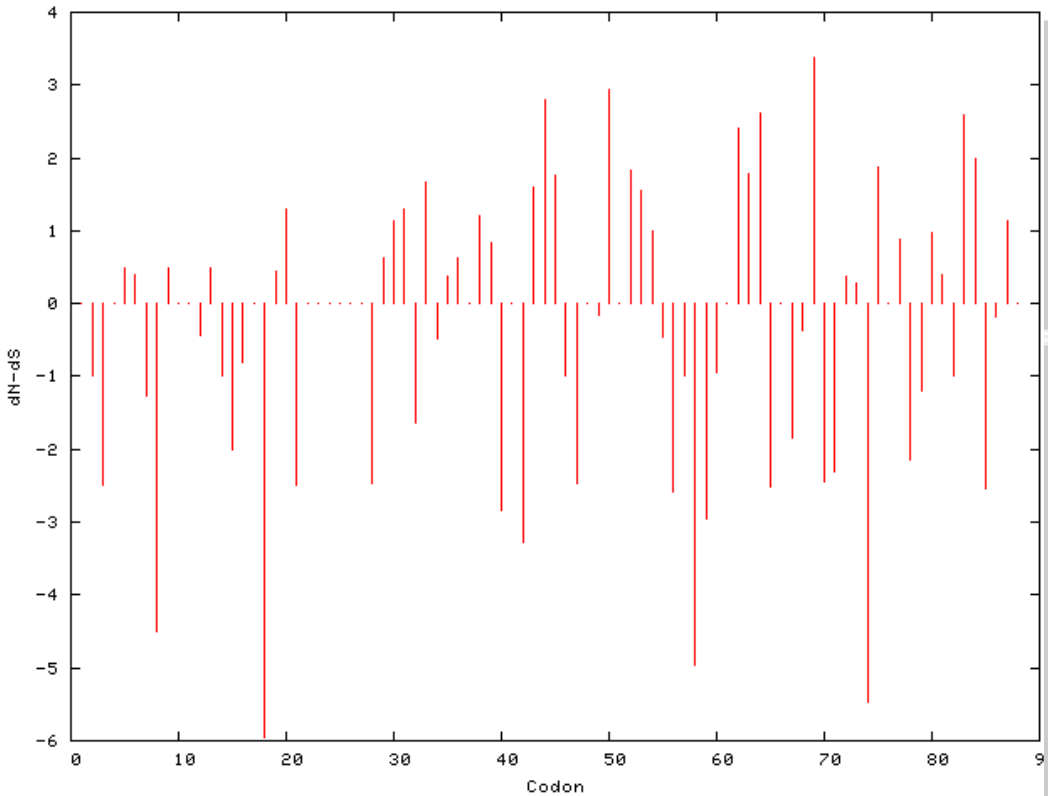
Selective sweep -> fewer alleles present in population

Repeated episodes of positive selection -> high dN

If time discuss http://online.itp.ucsb.edu/online/infobio01/fitch1/

hy-phy

Results of an anaylsis using the SLAC approach



FOUND **4** POSITIVELY SELECTED SITES (0.2 significance level)

Codon	dN-dS	Normalized dN-dS	p-value
45	2.80905	1.57283	0.174148
51	2.94548	1.64923	0.109144
65	2.62064	1.46734	0.197579
70	3.37001	1.88693	0.124868

FOUND **13** NEGATIVELY SELECTED SITES (0.2 significance level)

Codon	dN-dS	Normalized dN-dS	p-value
4	-2.5	-1.39979	0.111111
9	-4.5	-2.51963	0.0178326
19	-5.94245	-3.32728	0.0243467
22	-2.5	-1.39979	0.111111
41	-2.84041	-1.59039	0.193214
48	-2.45744	-1.37597	0.0793724
59	-4.96667	-2.78093	0.0236379
60	-2.96058	-1.65768	0.108898
66	-2.51831	-1.41004	0.15211
71	-2.45417	-1.37413	0.129462
72	-2.31427	-1.2958	0.162177
75	-5.47043	-3.06299	0.0388673
86	-2.54472	-1.42483	0.151309

more output might still be here

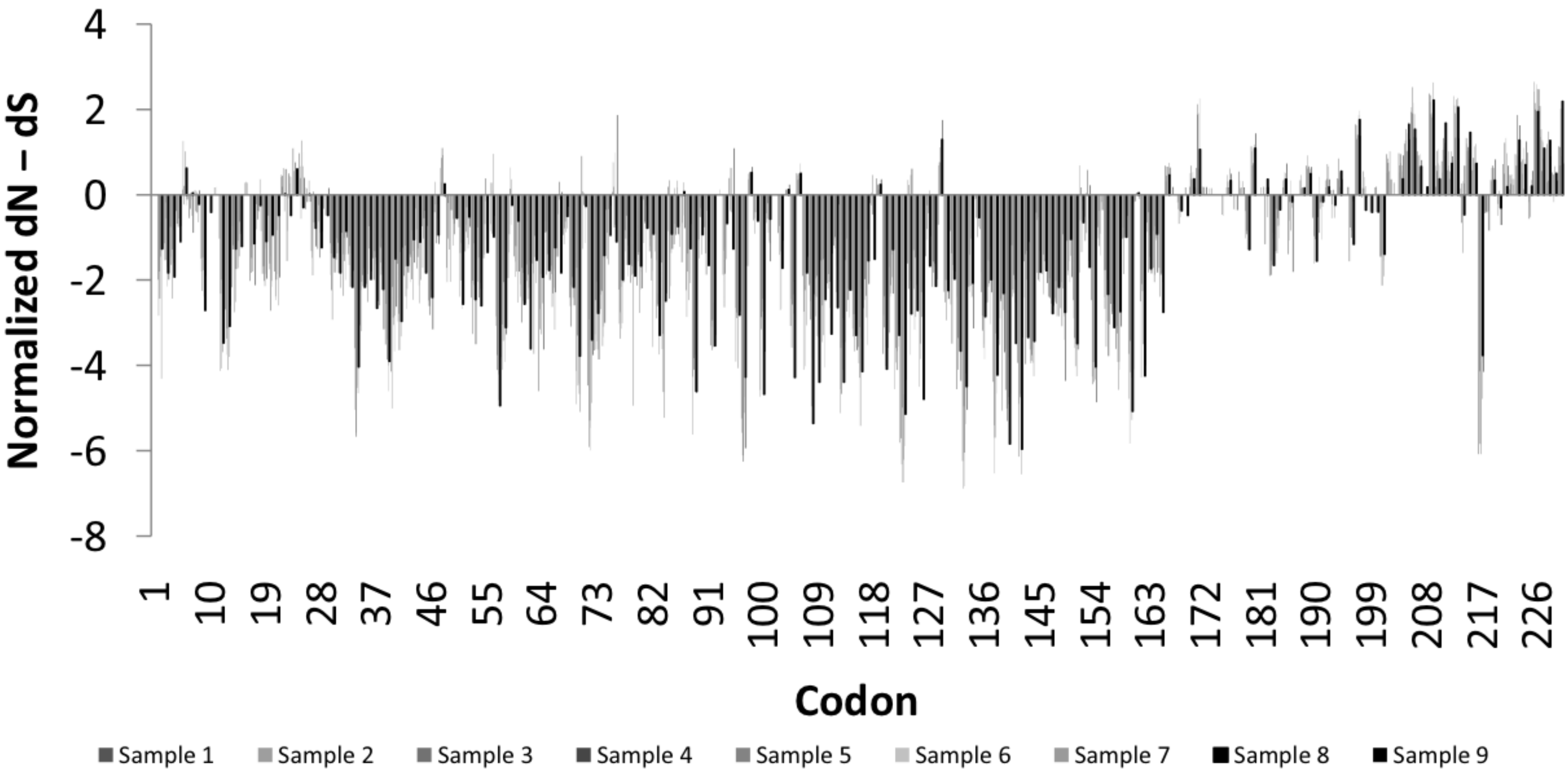


Fig 1. Patterns of substitutions: Bars represent dN > dS (positive) or dN < dS (negative) in random samples of 148 – 150 sequences (A) and the whole dataset of 1312 viruses (B). Included in B are regions of mapped activity and 3D structures of the RNA-binding domain (RBD, panel I) [21] and Effector domain (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.

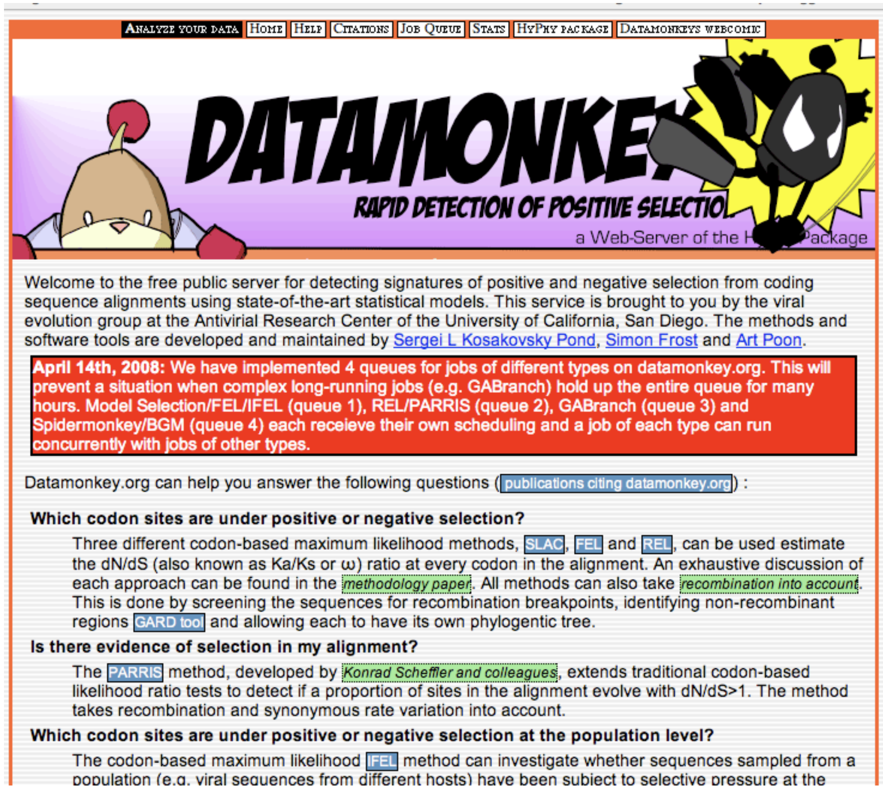
Hy-Phy -

Hypothesis Testing using Phylogenies.

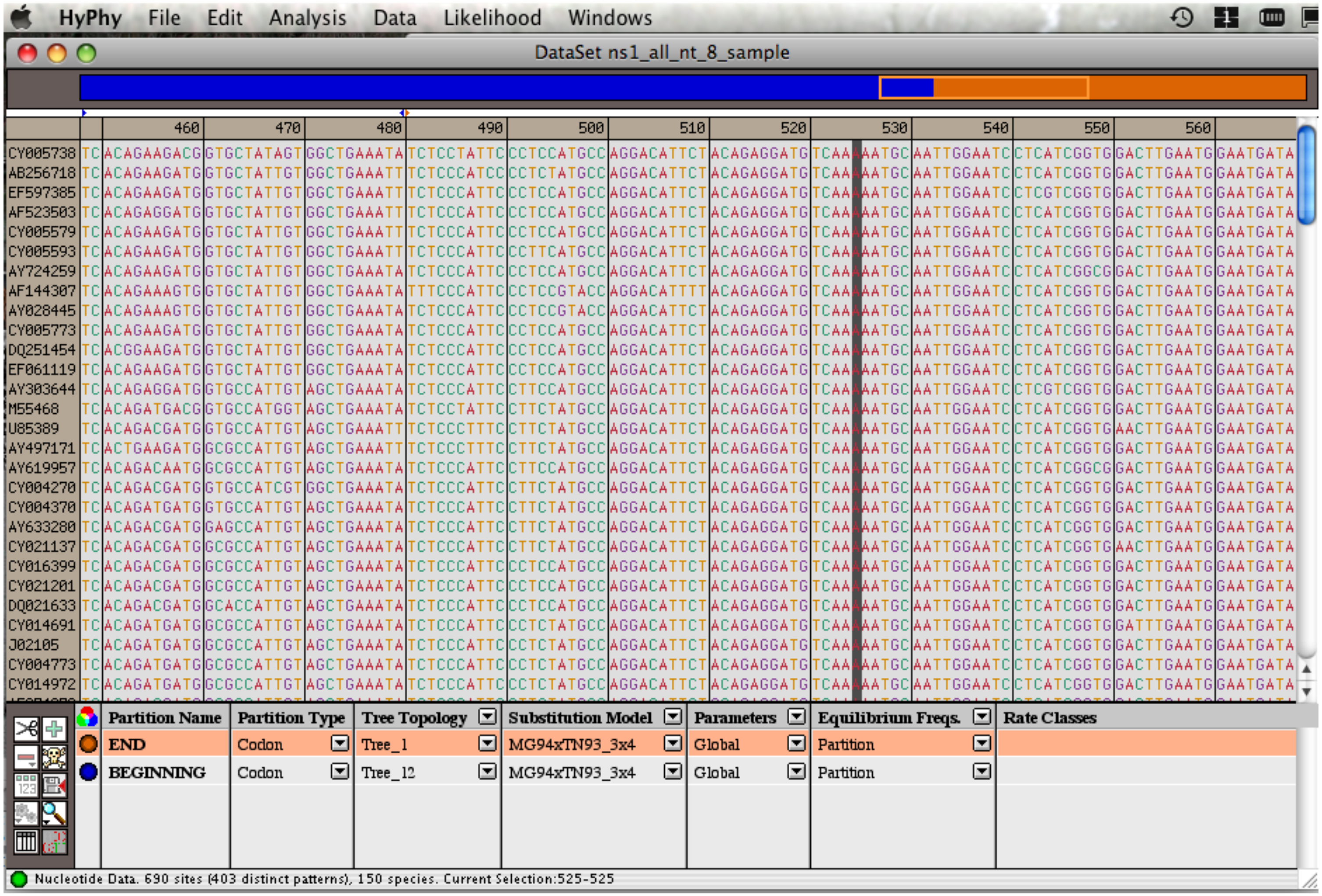
Using Batchfiles or GUI

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

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

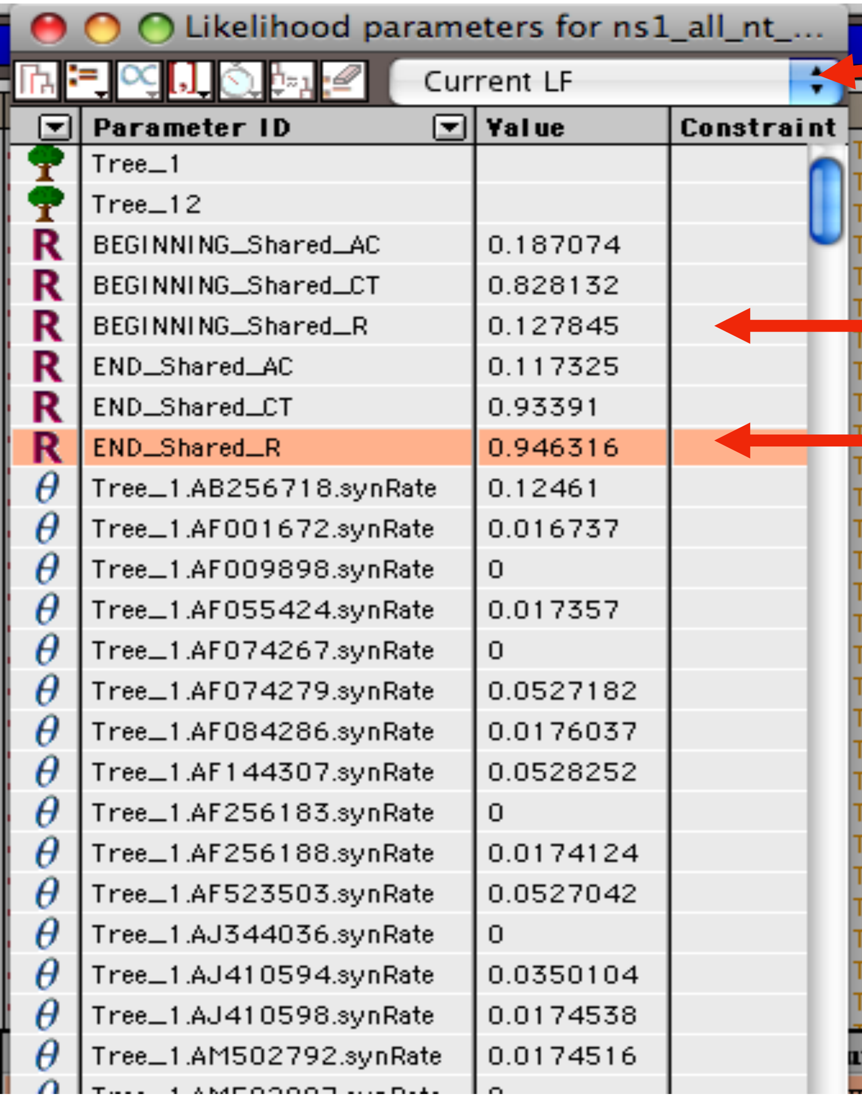


Example testing for dN/dS in two partitions of the data -- John's dataset



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

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



Likelihood parameters for ns1_all_nt_...

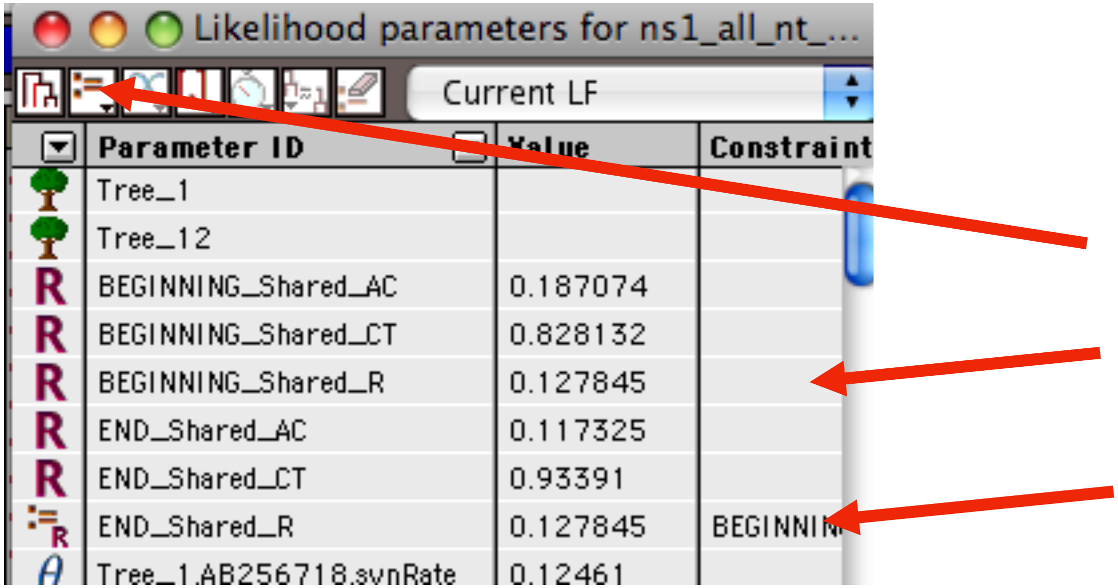
Current LF

Parameter ID	Value	Constraint
Tree_1		
Tree_12		
BEGINNING_Shared_AC	0.187074	
BEGINNING_Shared_CT	0.828132	
BEGINNING_Shared_R	0.127845	
END_Shared_AC	0.117325	
END_Shared_CT	0.93391	
END_Shared_R	0.946316	
Tree_1.AB256718.synRate	0.12461	
Tree_1.AF001672.synRate	0.016737	
Tree_1.AF009898.synRate	0	
Tree_1.AF055424.synRate	0.017357	
Tree_1.AF074267.synRate	0	
Tree_1.AF074279.synRate	0.0527182	
Tree_1.AF084286.synRate	0.0176037	
Tree_1.AF144307.synRate	0.0528252	
Tree_1.AF256183.synRate	0	
Tree_1.AF256188.synRate	0.0174124	
Tree_1.AF523503.synRate	0.0527042	
Tree_1.AJ344036.synRate	0	
Tree_1.AJ410594.synRate	0.0350104	
Tree_1.AJ410598.synRate	0.0174538	
Tree_1.AM502792.synRate	0.0174516	

Save Likelihood Function then
select as alternative

The dN/dS ratios for the two partitions are different.

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

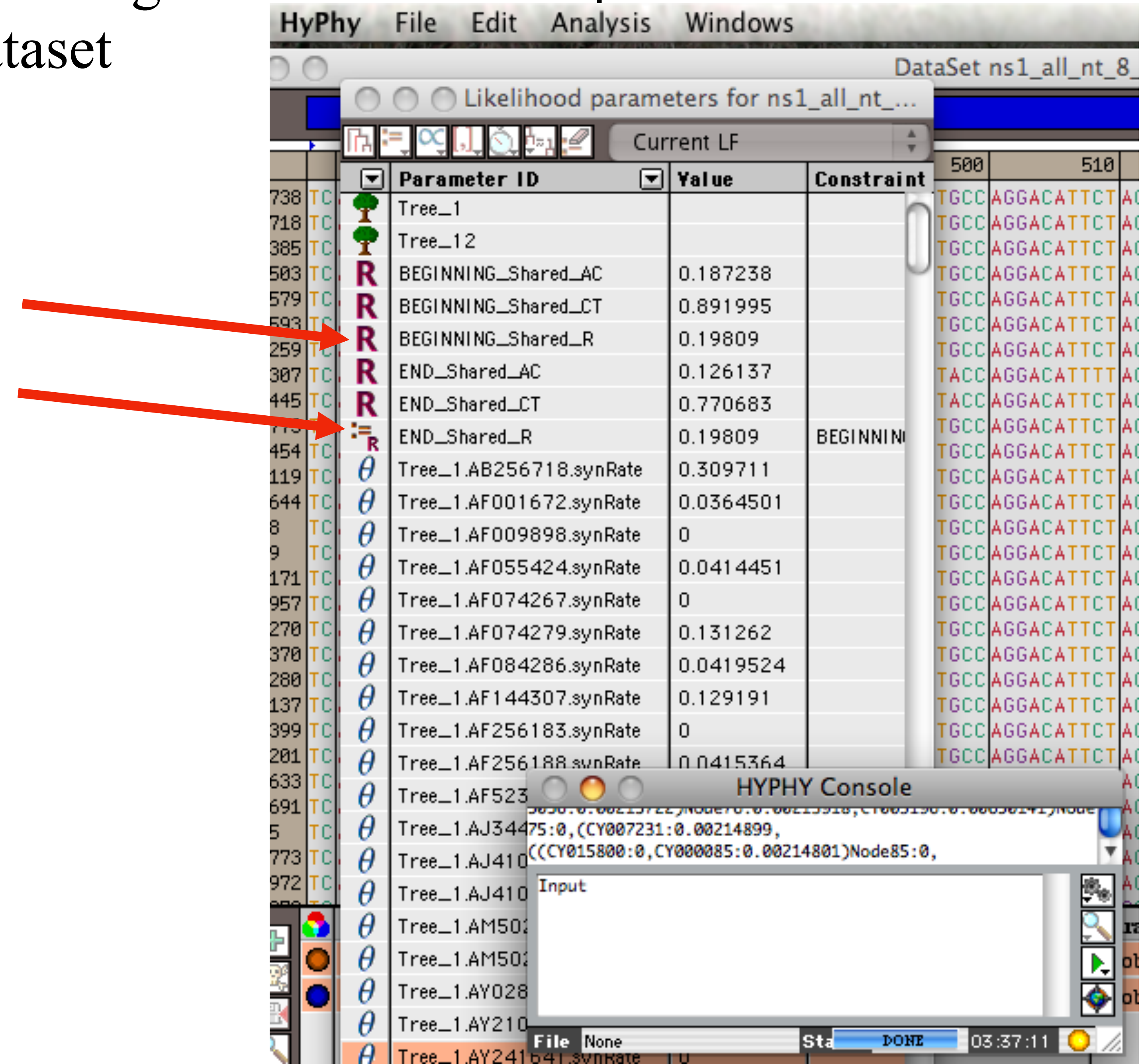


Set up null hypothesis, i.e.:

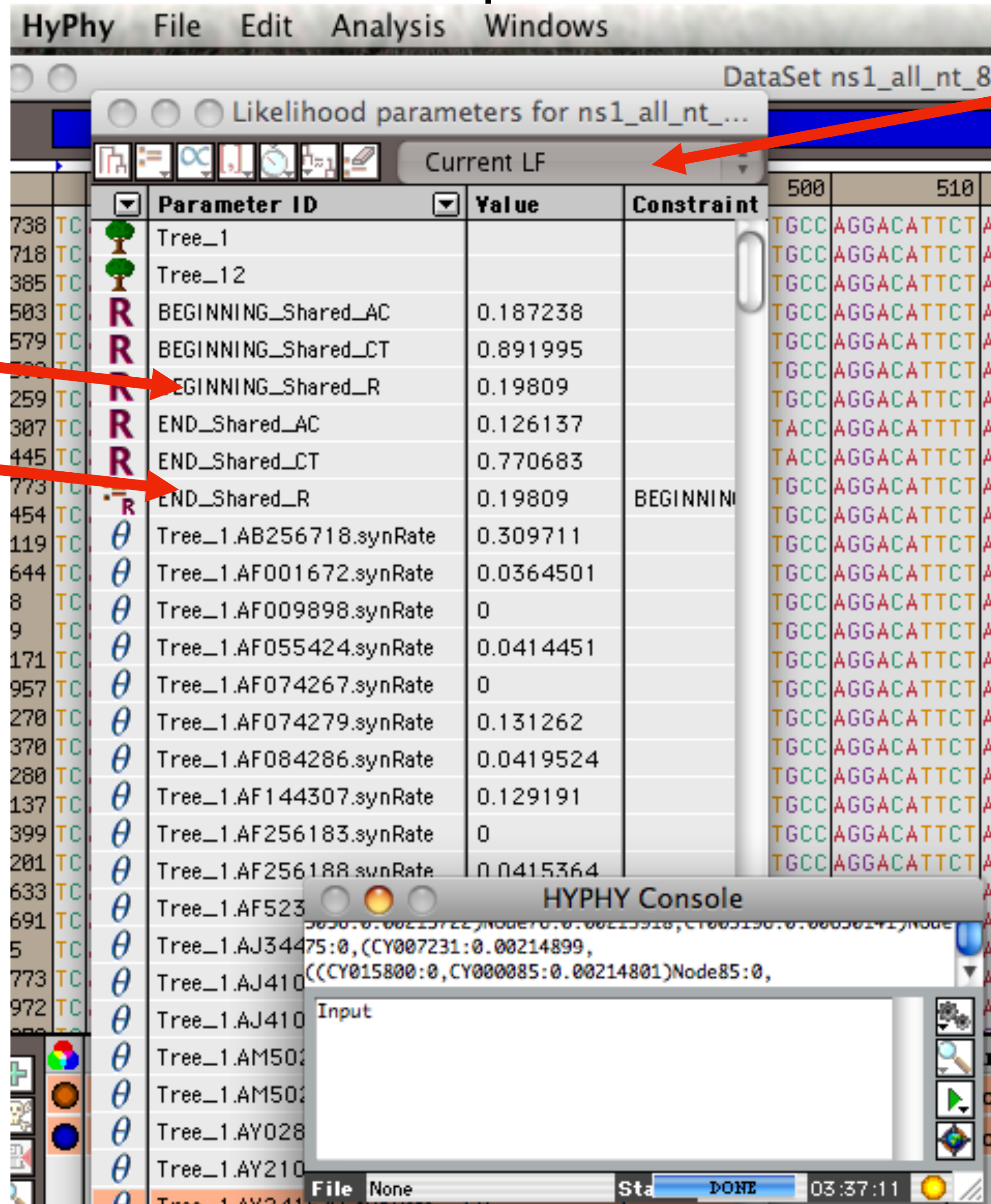
The two dN/dS are equal

(to do, select both rows and then click the define as equal button on top)

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

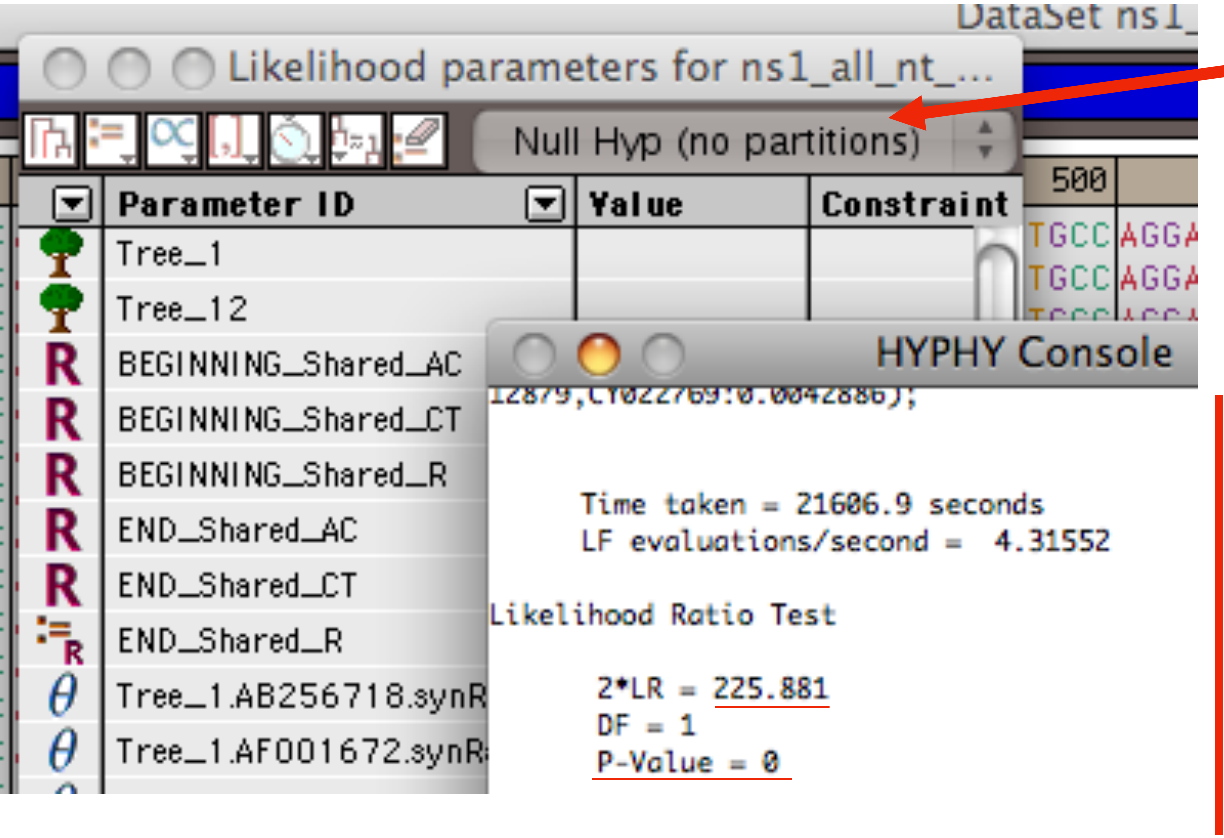


Example testing for dN/dS in two partitions of the data -- John's dataset



Name and save as Null-hyp.

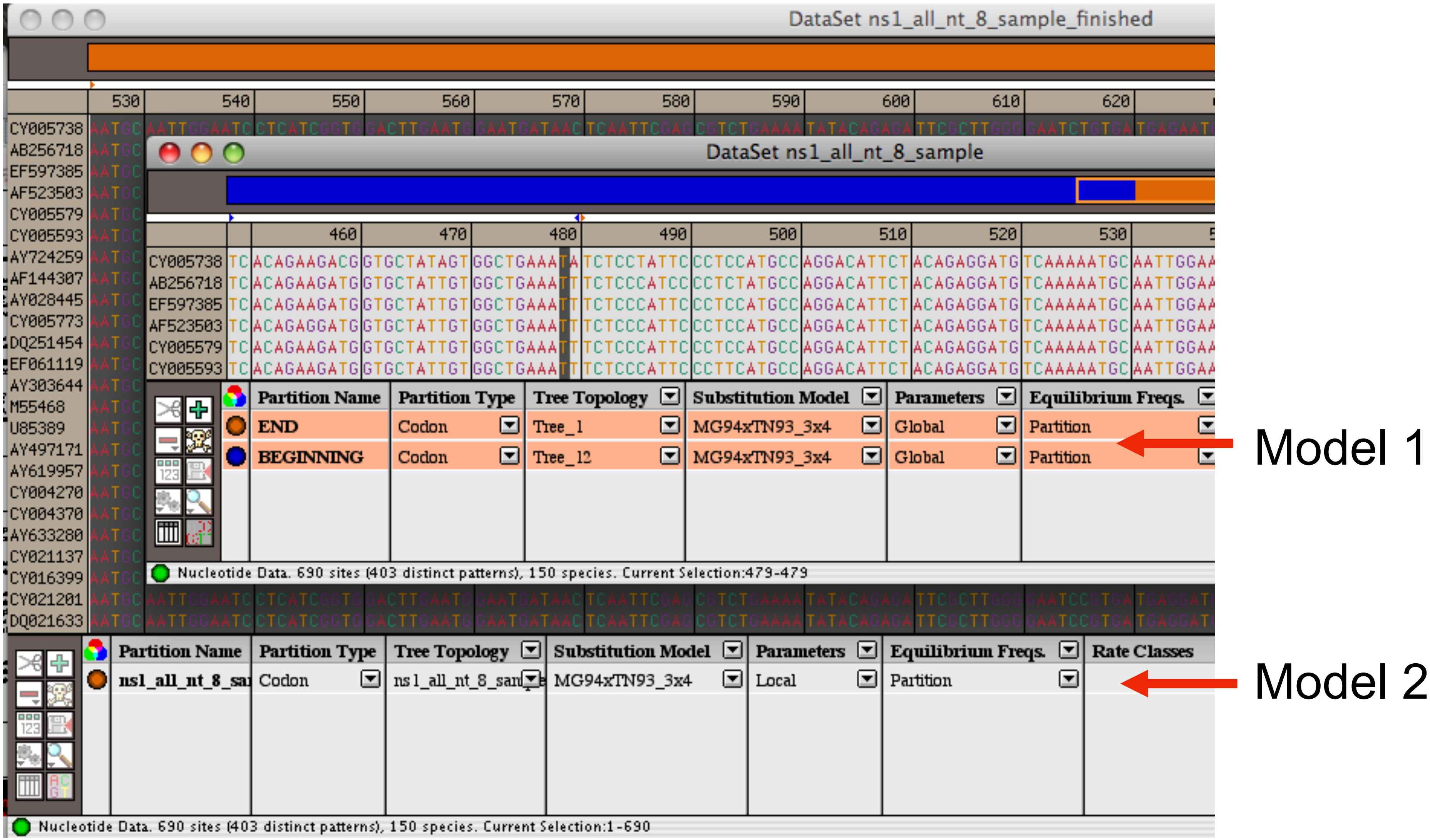
Example testing for dN/dS in two partitions of the data -- John's dataset



After selecting LRT (= Likelihood Ratio test), the console displays the result, i.e., **the beginning and end of the sequence alignment have significantly different dN/dS ratios.**

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