MCB 5472

The Queue,
Phylogenetic Reconstruction and Selection

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

Old exercises:

Write a program that it uses hashes to calculates mono-, di-, tri-, and quartet-nucleotide frequencies in a genome.

Go over tetraA.pl

New exercises:

modify tetraA.pl so that the user (or another program) can assign the size of the nmer as a variable!

Old Assignments:

- Re-read chapter P16-P18 in the primer

- 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.)

- Work on your student project

- 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.

an error prone solution is at convertannotationline.pl

New Assignment:

Rewrite this script so that is uses the $& variable to extract the gi number and the species name.

Symbol	Meaning
.	any character
\w	alphanumeric and _
\W	any non-word character
\s	any whitespace
\S	any non-whitespace
\d	any digit character
\D	any non-digit character
\t	tab
\n	newline
*	match 0 or more times
+	match 1 or more times
?	match 1 or 0 times
{n}	match exactly n times
{n,m}	match n to m times
^	match from start
$	match to end

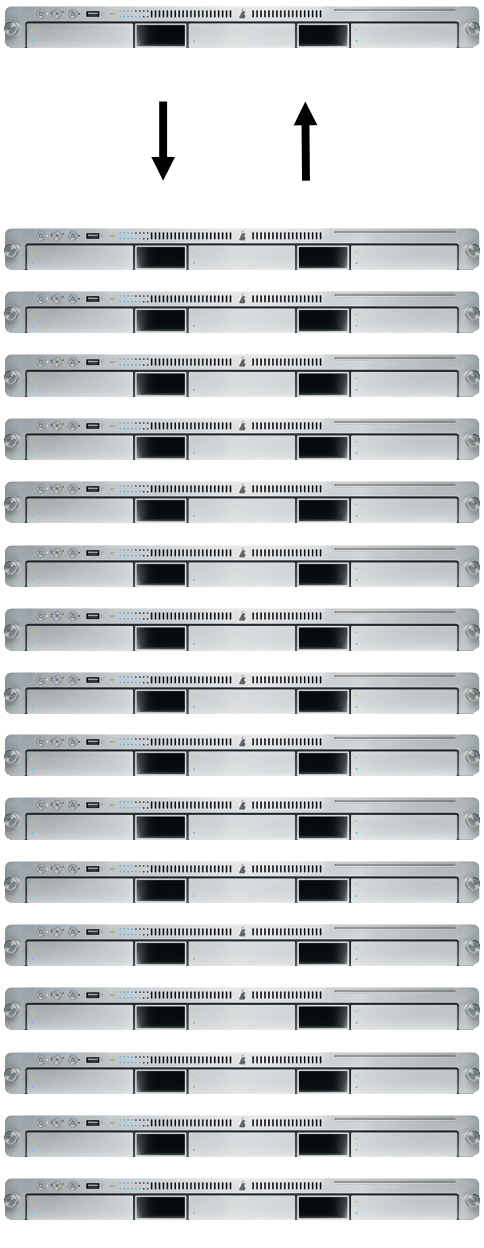
Bioinformatics Facility of the Biotechnology

The Do and Don't of the Xserve Cluster

Pascal Lapierre, Facility Scientist
Biotech Center, G05
486-8742

The next ten slides were provided Pascal Lapierre

Head Node



Node001 To Node017

Xserve Cluster Physical Organization

- 2 x 2.3GHz G5 processors

- 2 GB of memories per node (8 GB on node 17)

-2.3 TeraBytes of Storage on the head node.

- 2 other mini clusters assigned for special projects

Basic Rules

- For research purpose only. Not a place to put your favorite MP3 or backup your HD.

- Do not overload the systems. It is ok to use ~6 nodes in period of low activities but when it gets busy, limit yourself to only 2-3 nodes if absolutely necessary.

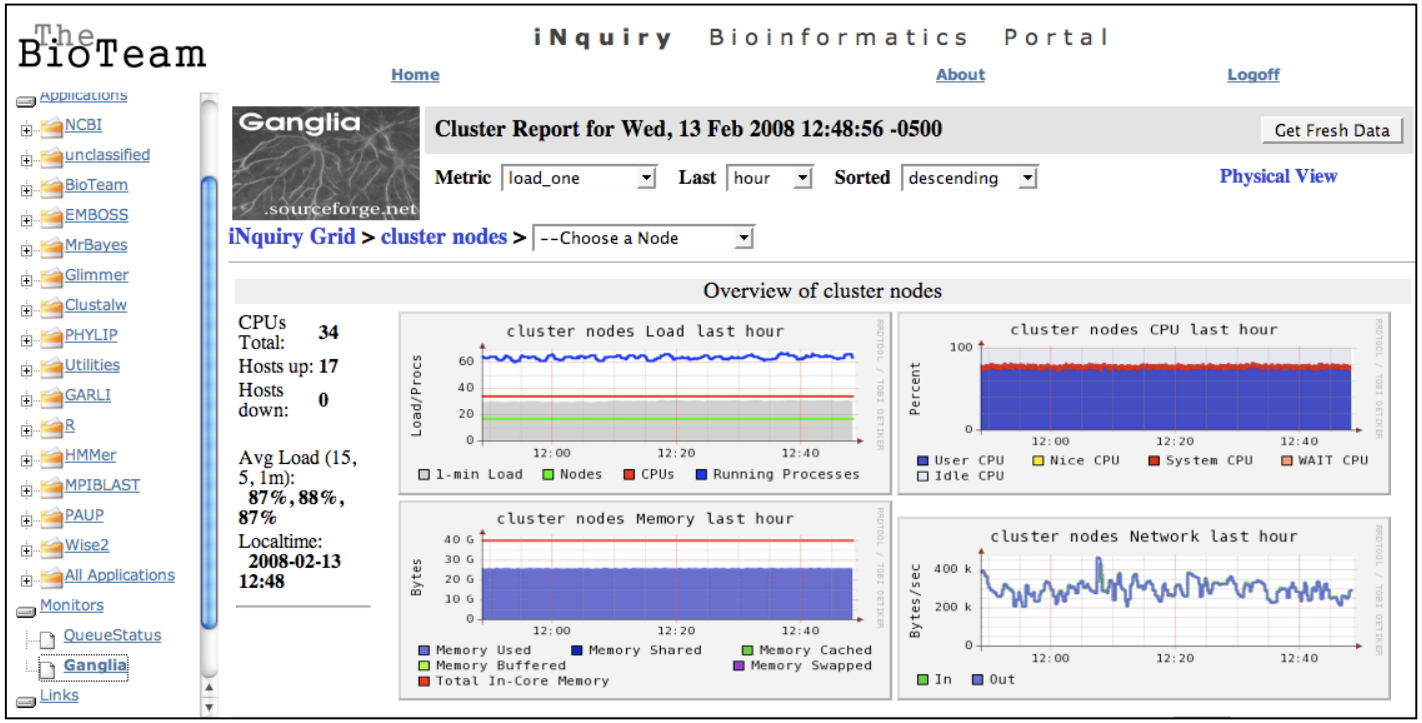
- Always keep track of your jobs. Don’t let things running unattended for months.

- Use the queue system whenever you can.

- Do not run jobs on the Head node.

Remote Access

- Via SSH or Web Interface
  - ssh your_name@bbcxsrv1.biotech.uconn.edu
  - http://www.biotech.uconn.edu/bf/



Useful Commands

(Help page available at : http://137.99.46.188/wiki/index.php/Main_Page)

`qstat` : Shows the current status of the available Grid Engine queues and the jobs associated with the queues.

`ls` : List directory contents

`ps` : Display the process status. Allow to get process ID.

-ps ux : Displays your process only

-ps aux : displays all the process running on the node

`du` : display disk usage statistics. Use `du -h` for a readable output (`df` for disk space)

Useful Commands (cont)

`mkdir` and `rmdir` : create and remove directories

`cp` : copy files

`mv` : moved files (can be used to rename files)

`rm` : remove files. `rm -r` to remove files and sub-directories

`kill` : to kill a running process. `Kill -9 'proc_id'`

The queue system

"Managing Workload by Managing Resources and Policies"

`qstat` : Display the queue status.

`qrsh` : Queue remote shell. Automatically select an available node to log on.

`qsub` : Queue submit. Automatically submit a job to an available node. Used in conjuncture with a shell script (see next slide).

`qdel` : Delete a job running in the queue.

qdel - process_ID

How to submit a job using qsub?

A shell script is just a small text file pointing to what you want to run in the queue.

For example, if I want to submit a perl script (phyml_trees1.pl), I will create a text file name phyml.sh :

```
#!/bin/sh
cd /Users/nucleus/evolver
perl phyml_trees1.pl
#end of script
```

To submit the shell :

```
qsub phyml.sh
```

Things to be cautious :

-While highly reliable, the cluster might sometimes run into problems and needed to be rebooted. This will cause to loose all the processes that were running at the time. Try to think of ways to break up or save at different stage of your analyses.

-The NFS (Network File System) have temporary amnesia. When overwhelm, the system will forget to write part of the output files. A workaround is to save to the scratch drive of the individual nodes (cd /scratch).

- blastall -p blastp -d nr -o /scratch/pascal/blast.out -a 2 -F F -m 9

Tricks that I have learned

In Perl, Array of Arrays are useful for grid-like manipulations of data :

```
##!/usr/bin/perl -w

$infile = ”seq.txt";
open (FILE,$infile);
while ($in = <FILE>) { #go infile line by line

chomp $in;
@data = split ('',$in); #split using ‘’
push @matrix,[@data]; # put the array @data
                        into @matrix
}
```

Seq.txt =

```
MRRAIATNQQ
MRLAIISRQD
MRRLSISRQQ
MRLAIIIRQQ
```

↓

```
  0123456789
0 MRRAIATNQQ
1 MRLAIISRQD
2 MRRLSISRQQ
3 MRLAIIIRQQ
```

Print $matrix[2][4]; ⟶ S

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.

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

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.

BUT

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
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:

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

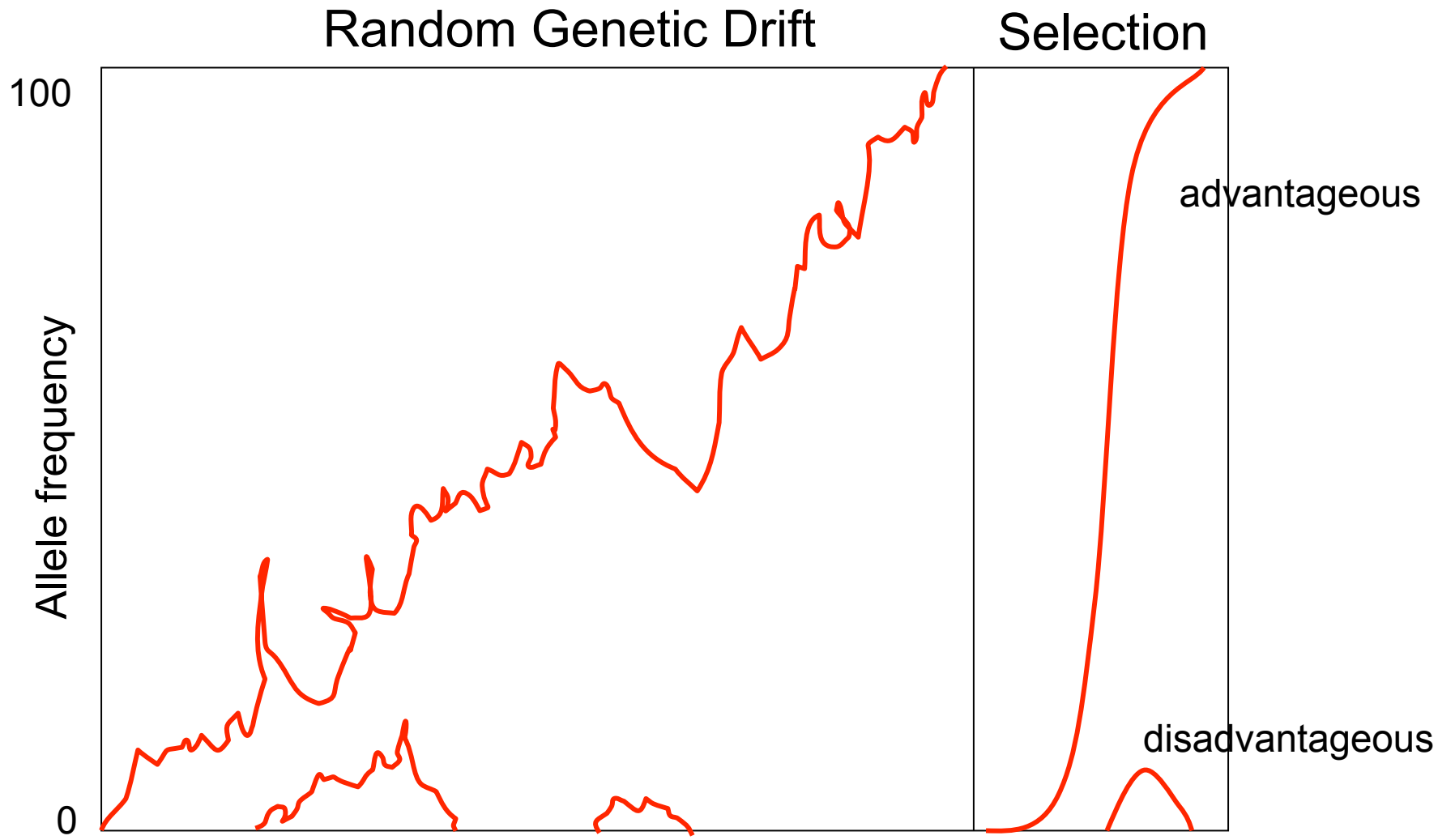
s>0

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

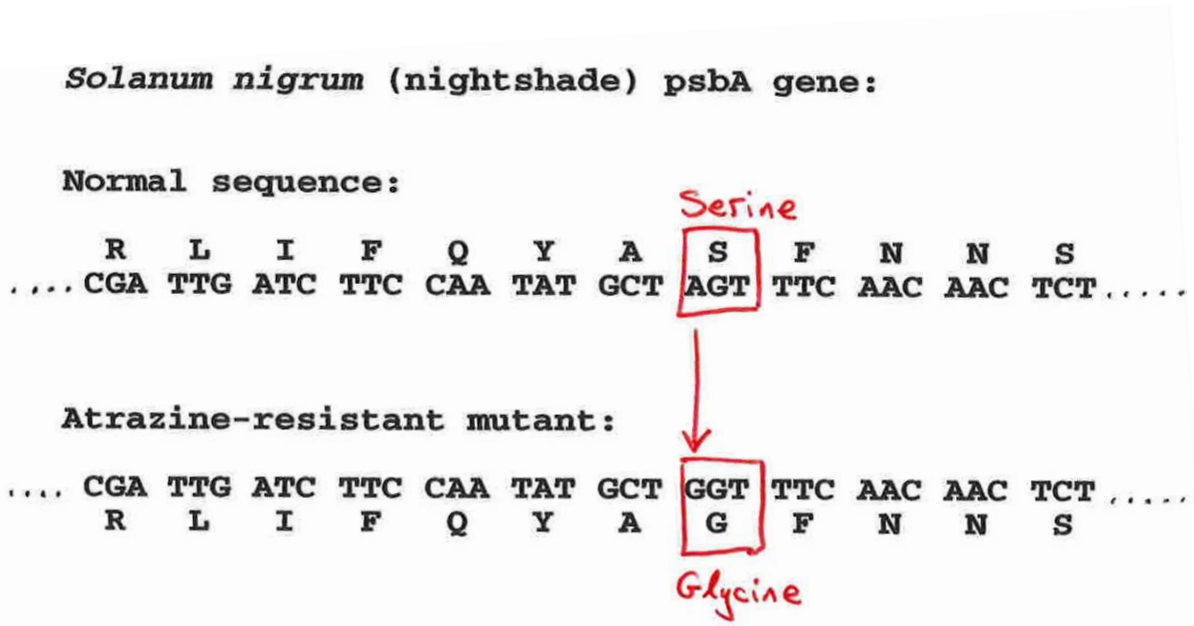
- A new allele (mutant) confers some increase in the **fitness** of the organism
- Selection acts to favour this allele
- Also called adaptive selection or Darwinian selection.

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

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

- Also called purifying selection

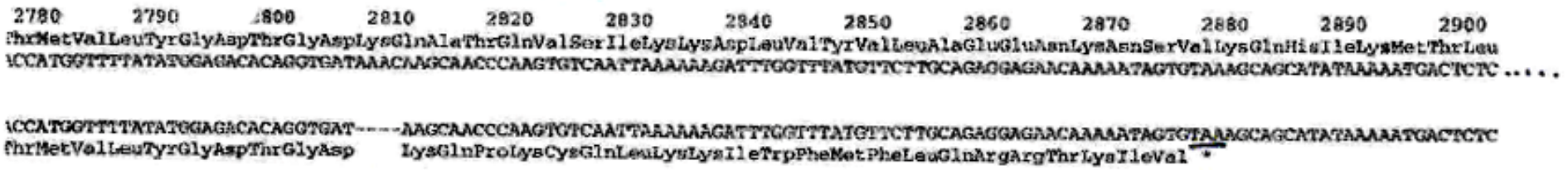
Modified from from www.tcd.ie/Genetics/staff/Aoife/GE3026/GE3026_1+2.ppt

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)

4 base pair deletion
Causes frameshift

Stop codon

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

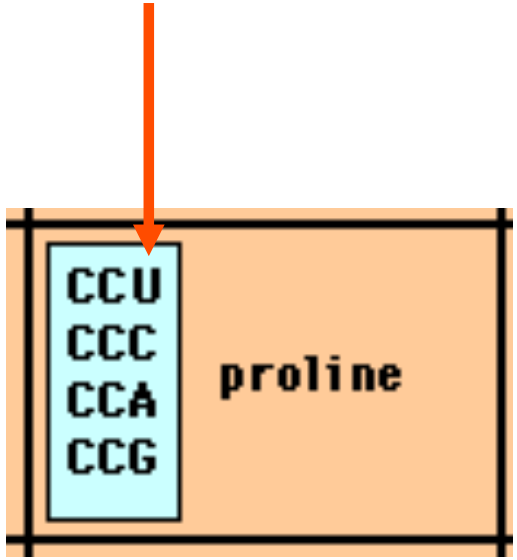
- Replacement of one nucleotide by another
- Synonymous (Doesn't change amino acid)
  - Rate sometimes indicated by Ks
  - Rate sometimes indicated by d_s
- Non-Synonymous (Changes Amino Acid)
  - Rate sometimes indicated by Ka
  - Rate sometimes indicated by d_n

(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

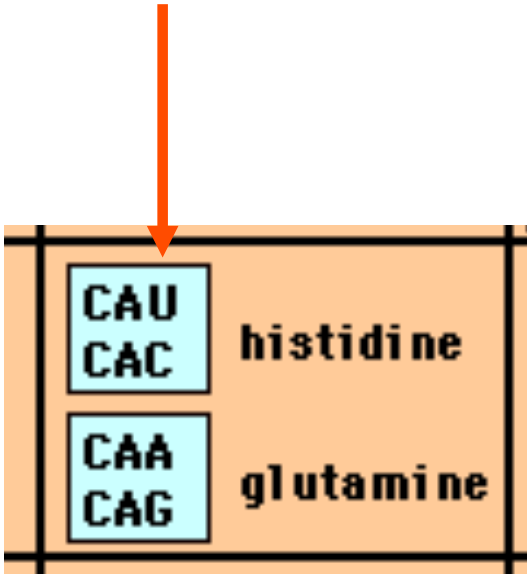
Genetic Code



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,

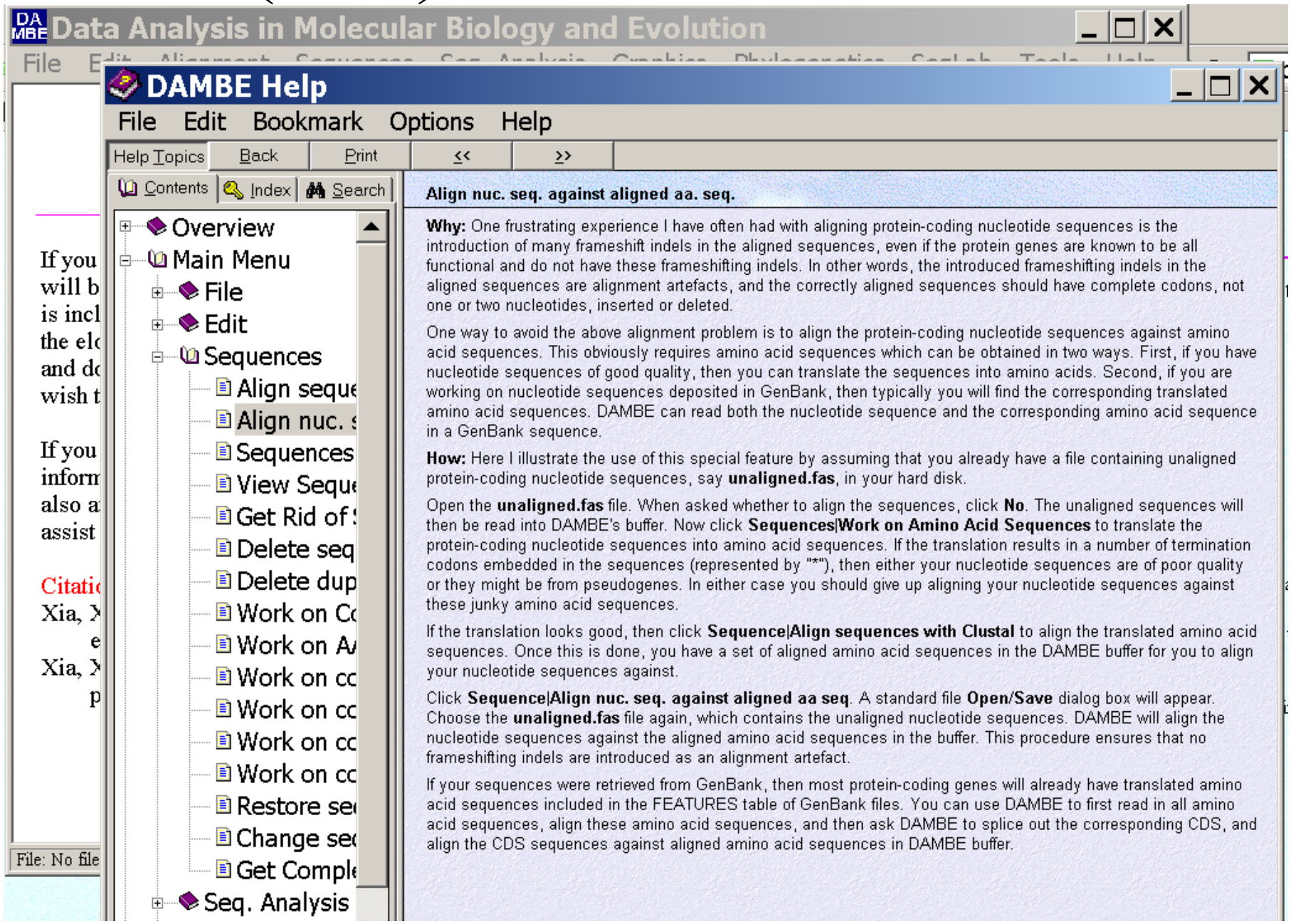
One is DAMBE (only for windows). This is a handy program for a lot of things, including reading a lot of different formats, calculating phylogenies, it even runs codeml (from PAML) for you.

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
Align seque
Align nuc. s
Sequences
View Seque
Get Rid of S
Delete seq
Delete dup
Work on Co
Work on A/
Work on cc
Work on cc
Work on cc
Work on cc
Restore se
Change se
Get Comple
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.
File: No file

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.

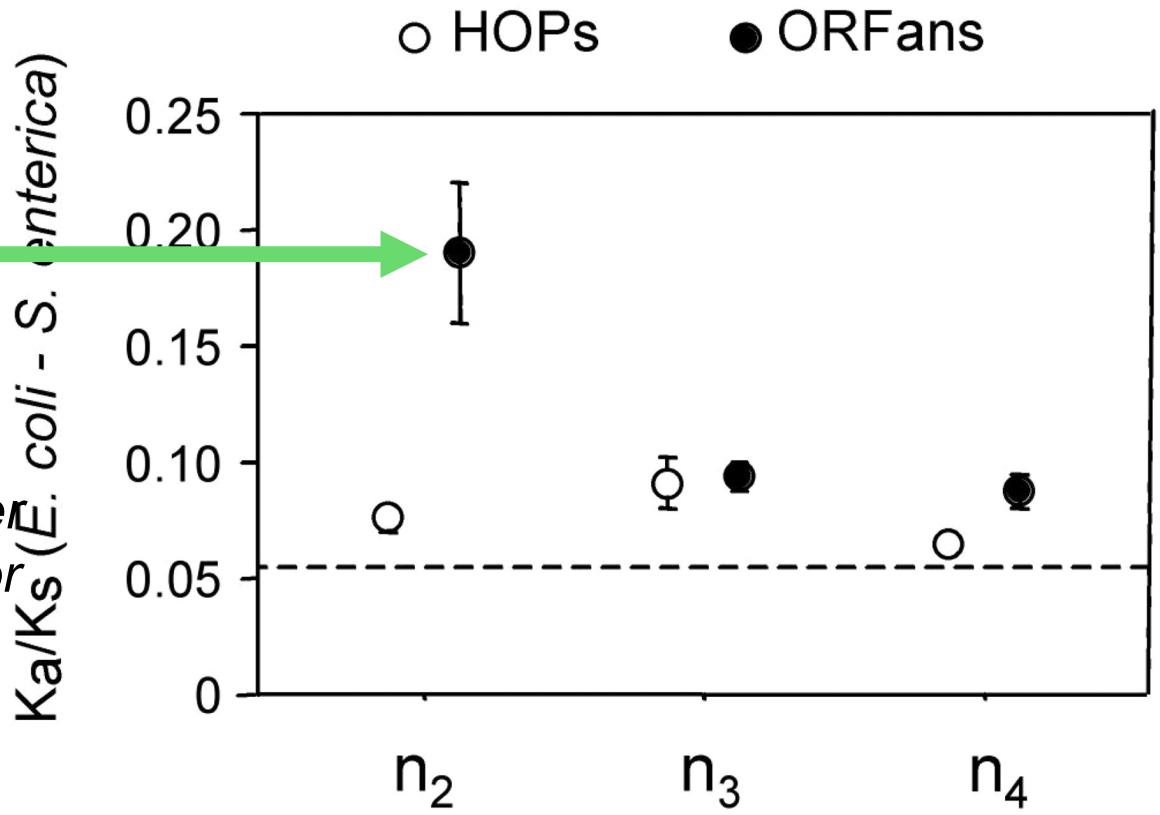


Fig. 3 from Vincent Daubin and Howard Ochman, *Genome Research 14:1036-1042, 2004*

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)?

where to get help

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

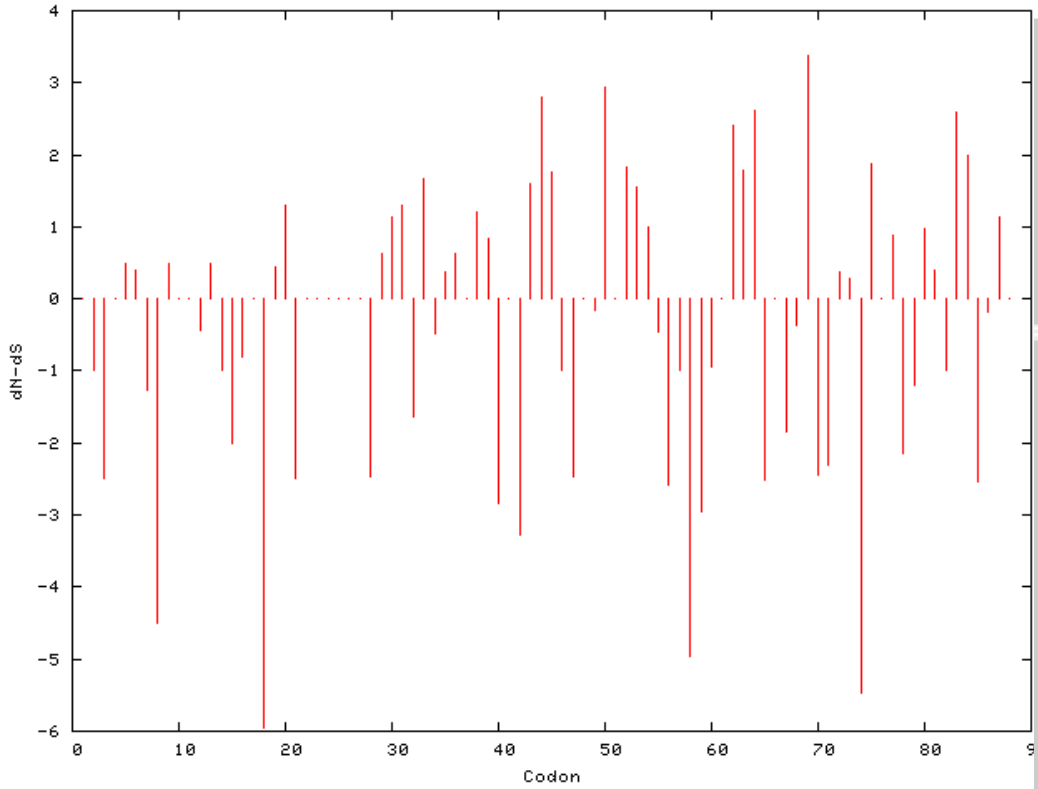
else

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.

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

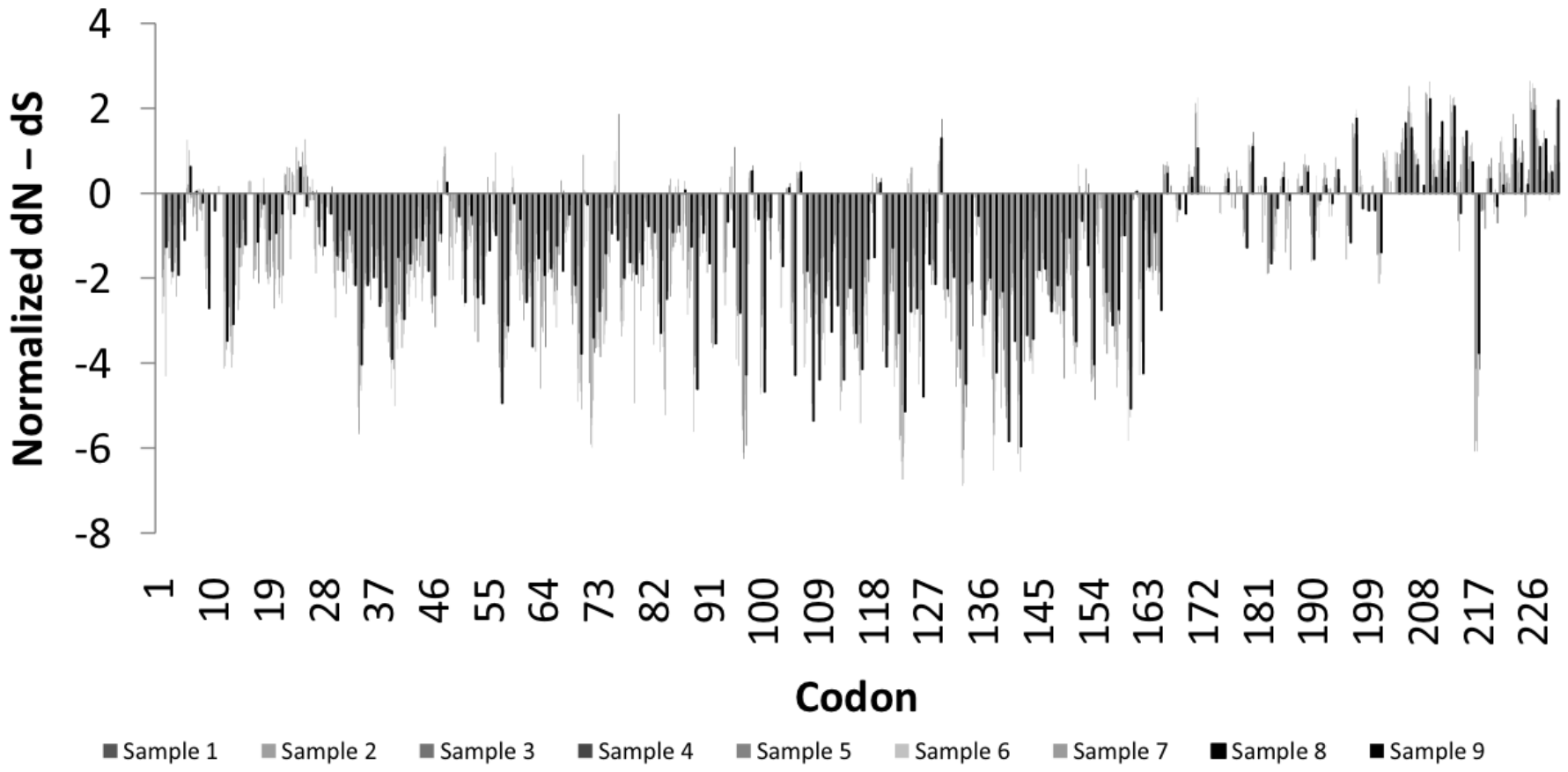


Fig 1. Patterns of substitutions: Bars represent dN > dS (positive) or dN < dS (negative) in random samples of 148 – 150 sequences (A) and the whole dataset of 1312 viruses (B). Included in B are regions of mapped activity and 3D structures of the RNA-binding domain (RBD, panel I) [21] and Effector domain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

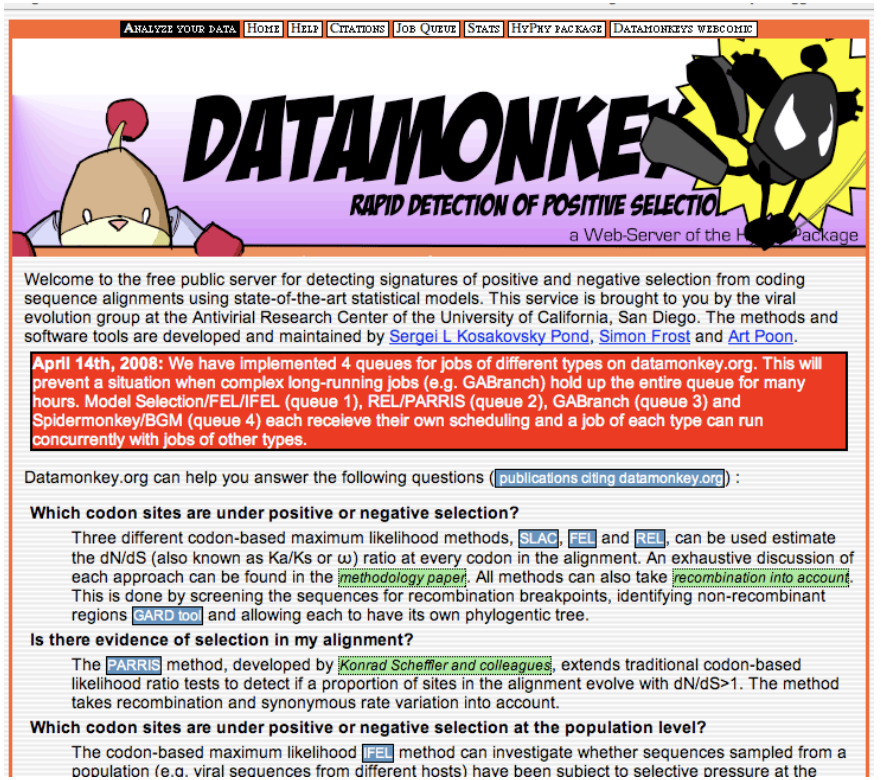
HYPHY PACKAGE
Hypothesis Testing Using Phylogenies

Hypothesis Testing using Phylogenies.

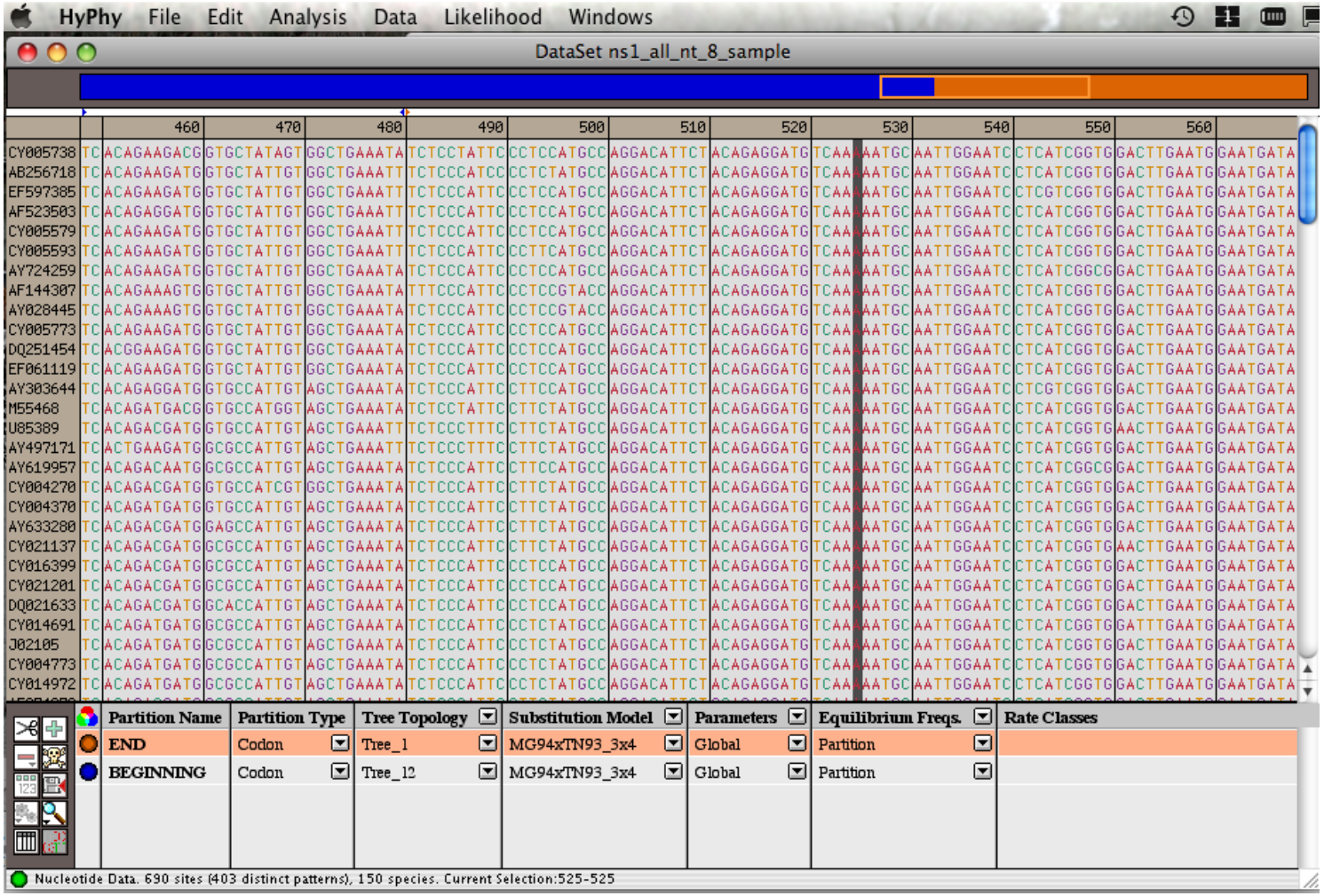
Using Batchfiles or GUI

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

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

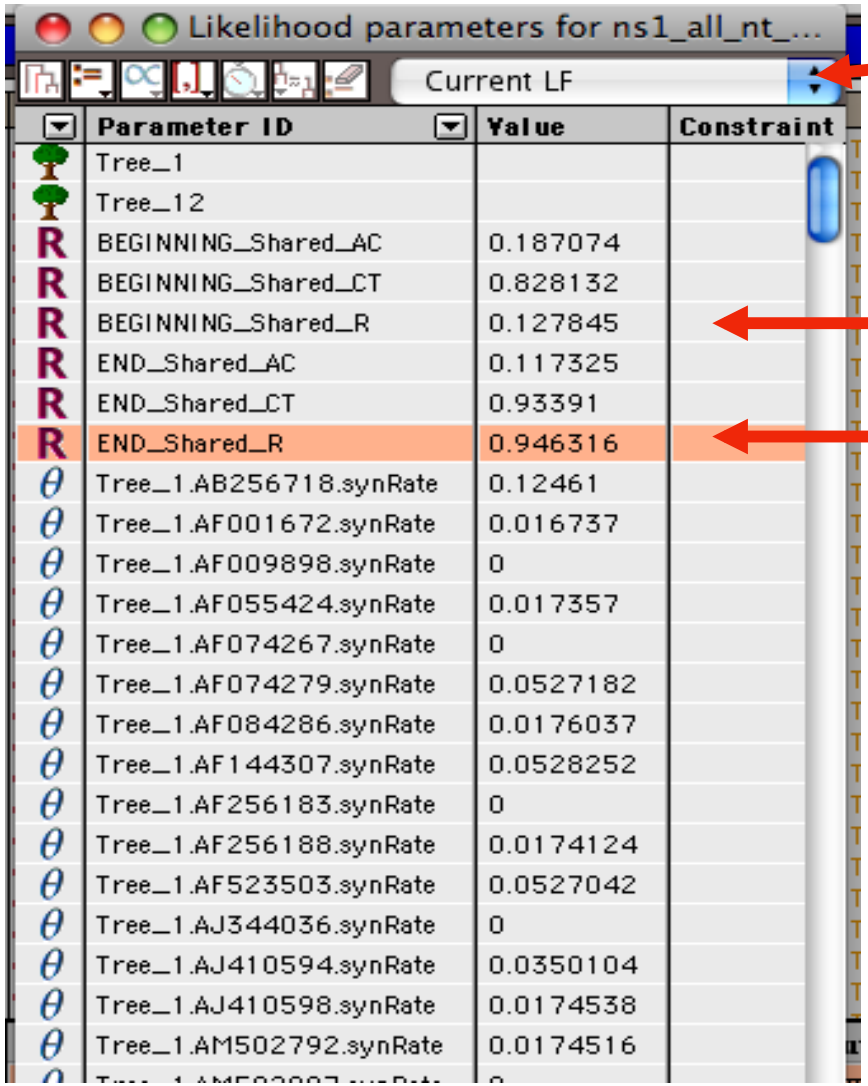


Example testing for dN/dS in two partitions of the data -- John's dataset



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

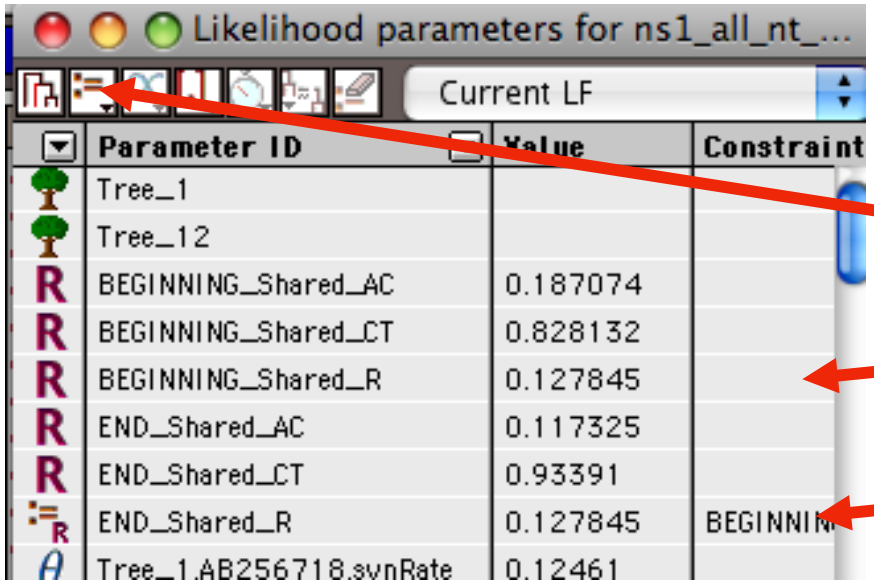
Example testing for dN/dS in two partitions of the data -- John's dataset



Safe 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

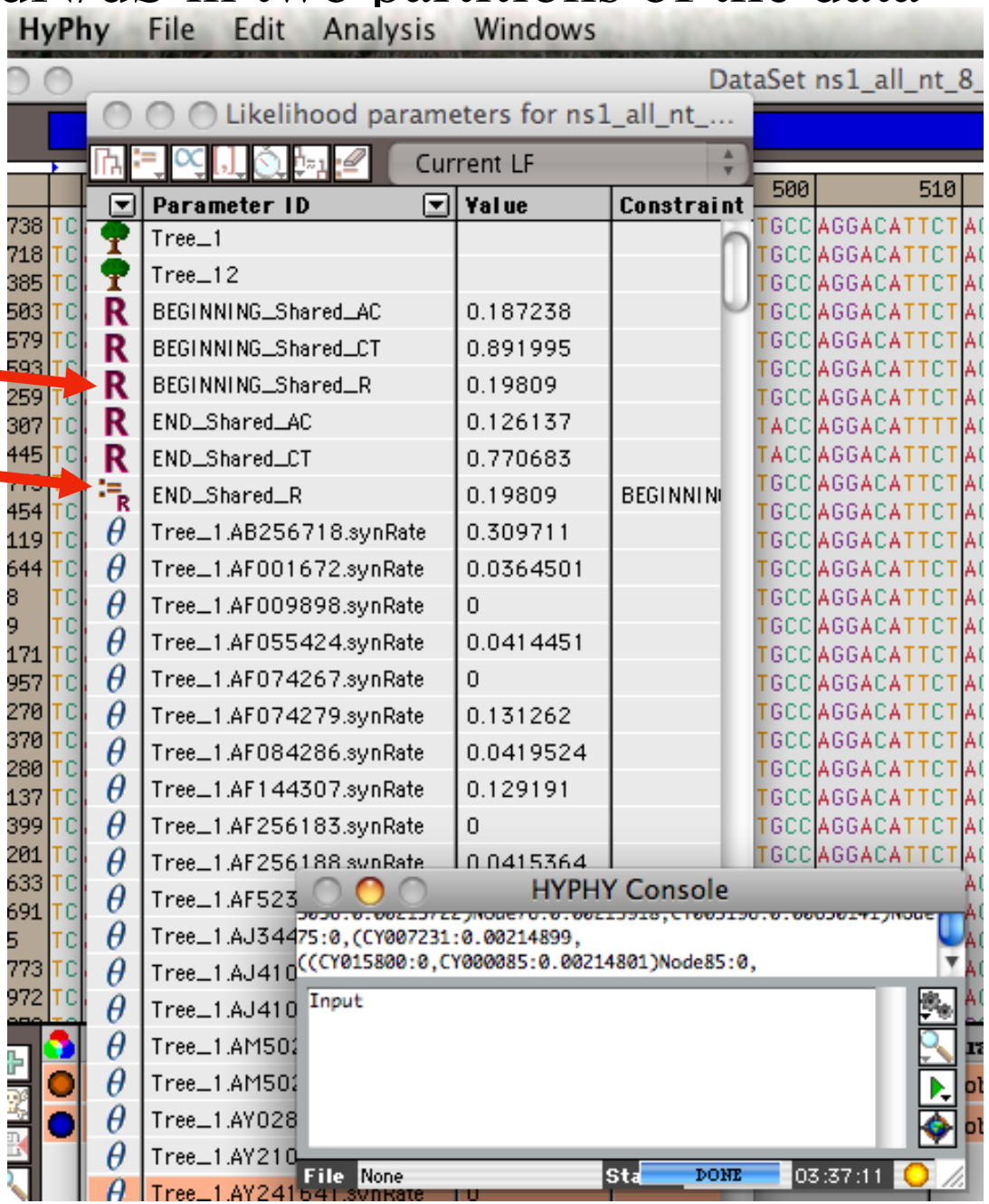


Set up null hypothesis, i.e.:

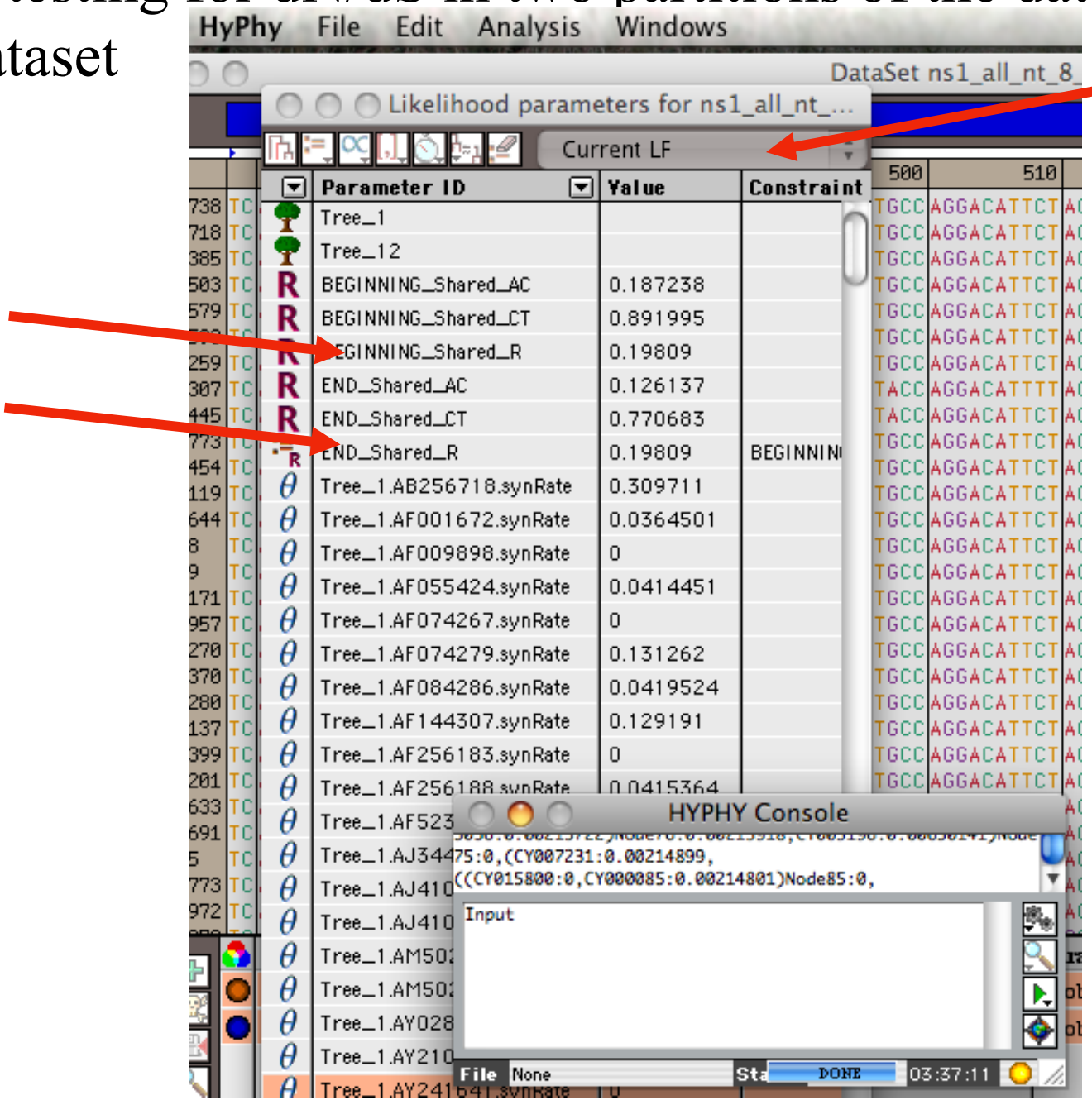
The two dN/dS are equal

(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

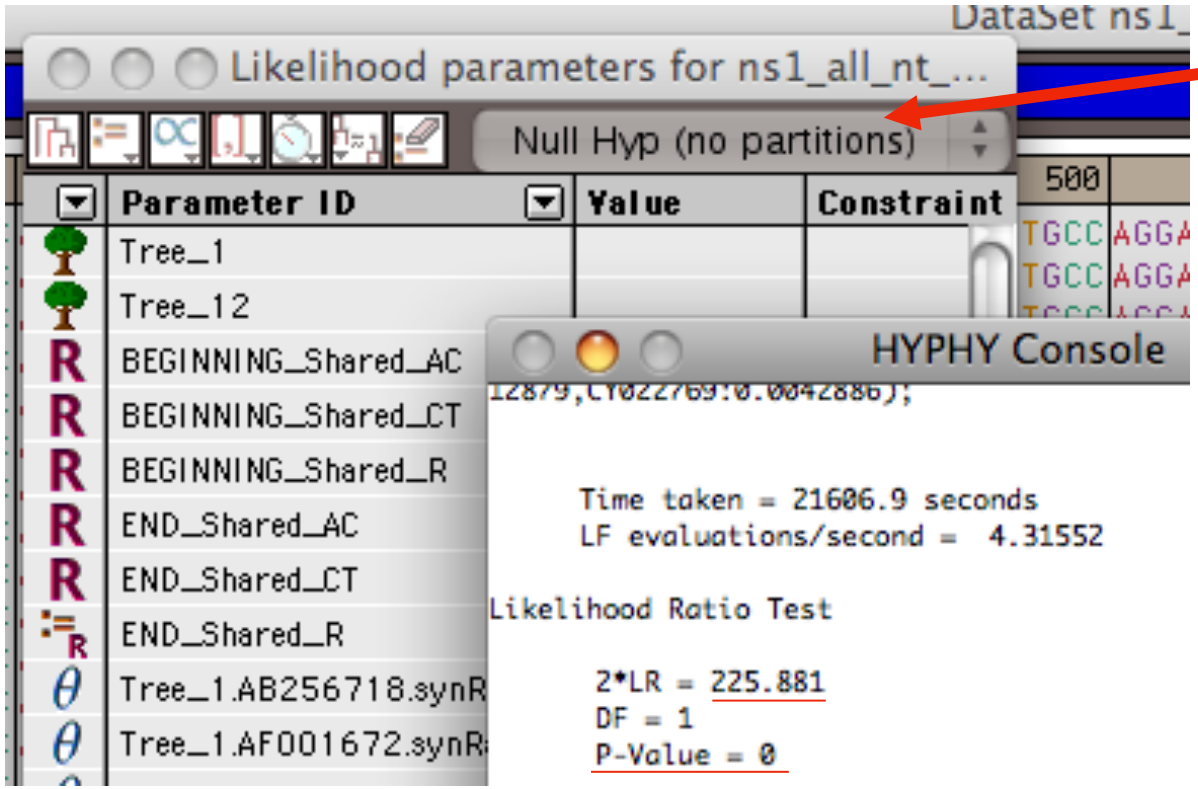


Example testing for dN/dS in two partitions of the data -- John's dataset



Name and save as Null-hyp.

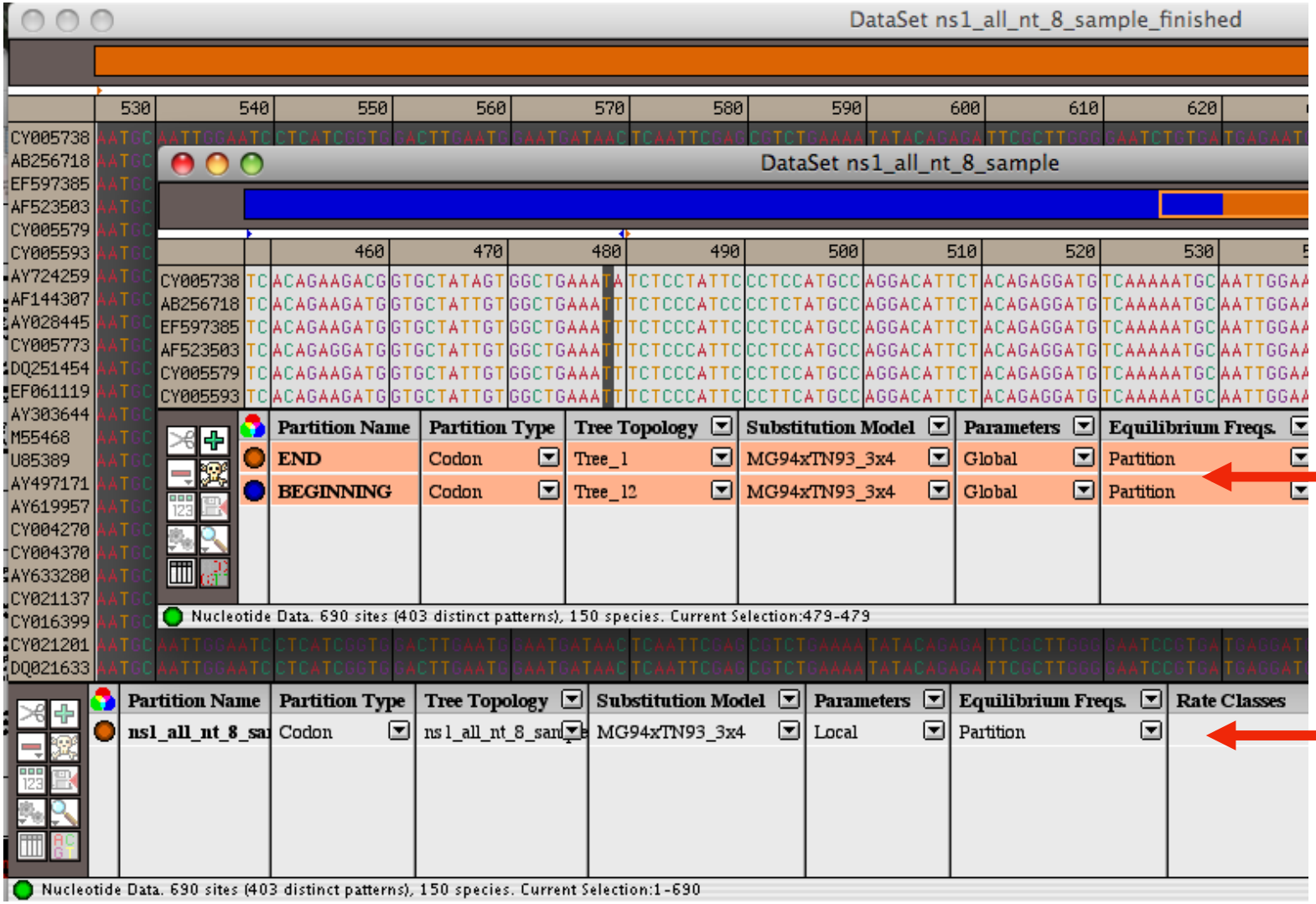
Example testing for dN/dS in two partitions of the data -- John's dataset



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:



Model 1

Model 2

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