MCB 372

Positive, and purifying selection.
Neutral theory

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

the gradualist point of view

Evolution occurs within populations where the fittest organisms have a selective advantage. Over time the advantagous 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 (i.e. traits are fixed even though they do not provide an advantage)
- Gratuitous complexity (introns, split intein)
- Selfish genes (who/what is the subject of evolution??)
- Parasitism, altruism, Morons (Gene Transfer Agents)

Assignments:

- Read through chapter 9
- Work on your student project
- Analyze one dataset of your choice in MrBayes.

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

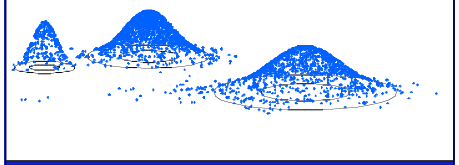


Figure generated using MCRobot program (Paul Lewis, 2001)

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:
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)

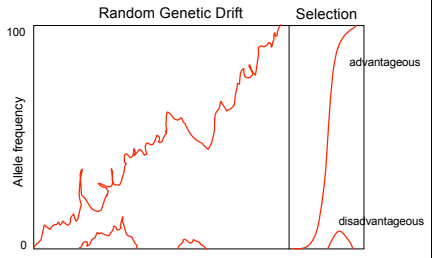
s>0

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

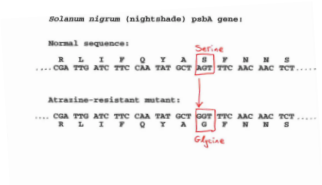
- A new allele (mutant) confers some increase in the **fitness** of the organism
- Selection acts to favour this allele
- Also called adaptive selection or Darwinian selection.

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)

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

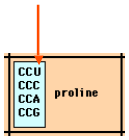
- Replacement of one nucleotide by another
- Synonymous (Doesn't change amino acid)
  - Rate sometimes indicated by Ks
  - Rate sometimes indicated by d_s
- Non-Synonymous (Changes Amino Acid)
  - Rate sometimes indicated by Ka
  - Rate sometimes indicated by d_n

(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	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		

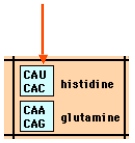
Genetic Code



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

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

dambe

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

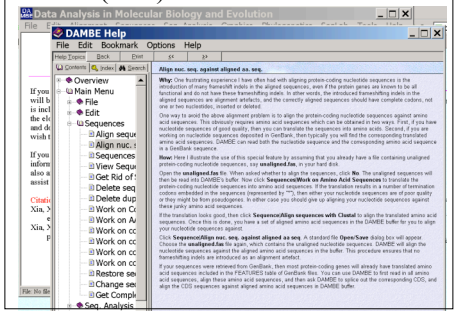
One is DAMBE (only for windows). This is a handy program for a lot of things, including reading a lot of different formats, calculating phylogenies, it even runs codeml (from PAML) for you.

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)



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 Wednesday, generate a codon based alignment with either *dambe* (on PCs only) or *tranalign* (Emboss, installed on cluster) 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 .

Nexus files:

This is the file format used by many popular programs like MacClade, Mesquite, ModelTest, MrBayes and PAUP*. Nexus file names often have a .nxs or .nex extension.

A formal description of the NEXUS format can be found in Maddison et al. (1997).

Conversion of an interleaved NEXUS file to a non-interleaved NEXUS file: execute the file in PAUP*, and export the file as non-interleaved NEXUS file. You can also type the commands:

export file=yourfile.nex format=nexus interleaved=no;
clustalw saves and reads Nexus sequence and tree files
(check on gap treatment and label as DNA or aa)

sample DNA file

```
#nexus
begin data;
dimensions ntax=10 nchar=705;
format datatype=dna interleave=yes gap=- missing=?;
matrix
Cow       ATGGCATATCCCATACAACTAGGATTCCAAGATGCAACATCACCAATCATAGAAGAACTA
Carp      ATGGCACACCCAACGCAACTAGGTTTCAAGGACGCGGCCATACCCGTTATAGAGGAACTT
Chicken ATGGCCAACCACTCCCAACTAGGCTTTCAAGACGCCTCATCCCCCATCATAGAAGAGCTC
Human     ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTACTTCCCCTATCATAGAAGACCTT
Loach     ATGGCACATCCCACACAATTAGGATTCCAAGACGCGGCCTCACCCCTAATAGAAGAACTT
Mouse     ATGGCCTACCCATTCCAACTTGGTCTACAAGACGCCACATCCCCTATTATAGAAGAGCTA
Rat       ATGGCTTACCCATTTCAACTTGGCTTACAAGACGCTACATCACCTATCATAGAAGAACTT
Seal      ATGGCATACCCCCTACAAATAGGCCTACAAGATGCAACCTCTCCCATTATAGAAGGAGTTA
Whale     ATGGCATATCCATTCCAACTAGGTTTCCAAGATGCAGCATCACCCATCATAGAAGAGCTC
Frog      ATGGCACACCCATCACAATTAGGTTTTCAAGACGCAGCCTCTCCAATTATAGAAGAATTA

Cow       CTTCACTTTCATGGACCACACGCTAATAATTGTCTTCTTAATTAGCTCATTAGTACTTTAC
Carp      CTTCACTTCCACGACCACGCATTAATAATTGTGCTCCTAATTAGCACTTTAGTTTTATAT
Chicken GTTGAATTCCACGACCACGCCCTGATAGTCGCACTAGCAATTTGCAGCTTAGTACTCTAC
Human     ATCACCTTTCATGATCACGCCCTCATAATCATTTCCTTATCTGCTTCCTAGTCCTGTAT
Loach     CTTCACTTCCATGACCACGCCCTAATAATTCATTTTGATTACGCCCTAGTACTTTAT
Mouse     ATAAATTTCCATGATCACACACTAATAATTGTTTTCCTAATTAGCTCCTTAGTCCTCTAT
Rat       ACAAACTTTCATGACCACACCCTAATAATTGTATTCCTCATCAGCTCCCTAGTACTTTAT
Seal      CTACACTTCCATGACCACACATTAATAATTGTGTTCCTAATTAGCTCATTAGTACTCTAC
Whale     CTACACTTTCACGATCATACACTAATAATCGTTTTCTAATTAGCTCTTTAGTTCTCTAC
Frog      CTTCACTTCCACGACCATACCCTCATAGCCGTTTTTCTTATTAGTACGCTAGTTCTTTAC

//
Frog      AACTGATCTTCATCAATACTA---GAAGCATCACTA------AGA
;
end;
```

sample aa file

```
#NEXUS

Begin data;
Dimensions ntax=10 nchar=234;
Format datatype=protein gap=- interleave;
Matrix
Cow       MAYPMQLGFQDATSPIMEELLHFHDHTLMIVFLISSLVLYIISLMLTTKLTHTSTMDAQE
Carp      MAHPTQLGFKDAAMPVMEELLHFHDHALMIVLLISTLVLYIITAMVSTKLTNKYILDSQE
Chicken MAHHSQLGFQDAASPIMEELVEFHDHALMVALAICSLVLYLLMLMEKLS-SNYVDAQE
Human     MAHAAQVGLQDATSPIMEELITFHDHALMIIFLICFLVLYALFLTLTTKLTNTNISDAQE
Loach     MAHPTQLGFQDAASPVMEELLHFHDHALMIVFLISALVLYVIITTVSTKLTNMYILDSQE
Mouse     MAYPFQLGLQDATSPIMEELMNFHDHTLMIVFLISSLVLYIISLMLTTKLTHTSTMDAQE
Rat       MAYPFQLGLQDATSPIMEELTNFHDHTLMIVFLISSLVLYIISLMLTTKLTHTSTMDAQE
Seal      MAYPLQMGLQDATSPIMEELLHFHDHTLMIVFLISSLVLYIISLMLTTKLTHTSTMDAQE
Whale     MAYPFQLGFQDAASPIMEELLHFHDHTLMIVFLISSLVLYITLMLTTKLTHTSTMDAQE
Frog      MAHPSQLGFQDAASPIMEELLHFHDHTLMAVFLISTLVLYIITIMTTKLTNTNLMDAQE

//
Loach     QTAFIASRPGVFYQCSEICGANHSFMPIVVEAVPLSHFENWSTLMLKDASLGS
Mouse     QATVTSNRPGLFYQQCSEICGSNHSFMPIVLEMVPLKYFENWSASMI-------
Rat       QATVTSNRPGLFYQQCSEICGSNHSFMPIVLEMVPLKYFENWSASMI-------
Seal      QPTLMTMRPGLYYQQCSEICGSNHSFMPIVLELVPLSHFEKWSTSML-------
Whale     QPTLMGTRPGLFYQQCSEICGSNHSFMPIVLELVPLKVFEKWSVSML-------
Frog      QTSFIATRPGVFYQQCSEICGANHSFMPIVVEAVPLTDFENWSSSML-EASL--
                :
End;
```

Another example is here

More information on Nexus files and PAUP and MrBayes commands are in the respective manuals:

http://paup.csit.fsu.edu/, manual here, tutorial here

http://mrbayes.csit.fsu.edu/, manual here Wikki

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.

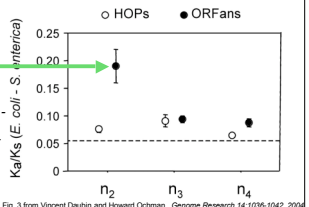
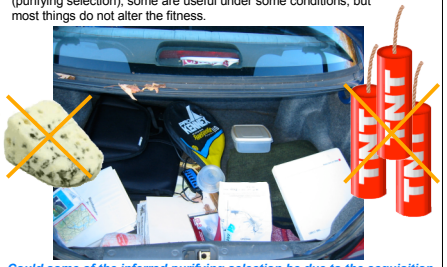


Fig. 3 from Vincent Daubin and Howard Ochman, *Genome Research 14:1036-1042, 2004*

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.



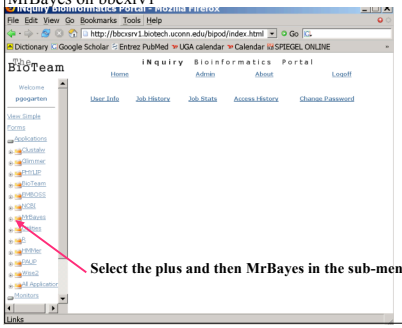
Could some of the inferred purifying selection be due to the acquisition of novel detrimental characteristics (e.g., protein toxicity)?

MrBayes on bbcxrv1

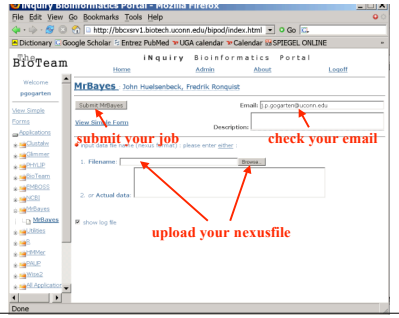
Create the nexus file on your computer.
It will help to have MrBayes installed locally, this way you can check that you don't have any typos in the MrBayes block.

Direct your browser to
http://bbcxsrv1.biotech.uconn.edu/bipod/index.html

MrBayes on bbcxrv1



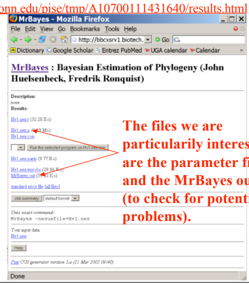
MrBayes on bbcxrv1



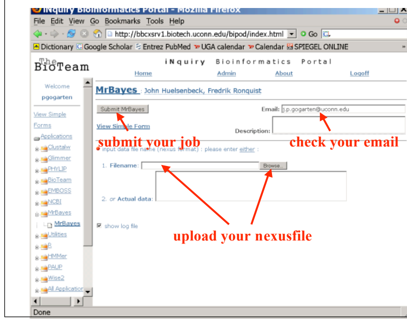
MrBayes on bbcxrv1

You will receive the results per email, and you will receive the link of a web page that lists all the output files. In this case: http://bbcxsrv1.biotech.uconn.edu/pise/tmp/A10700111431640/results.html

You can save the files from your browser, or open the email attachments. .



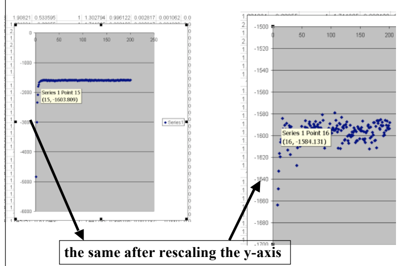
MrBayes on bbcxrv1



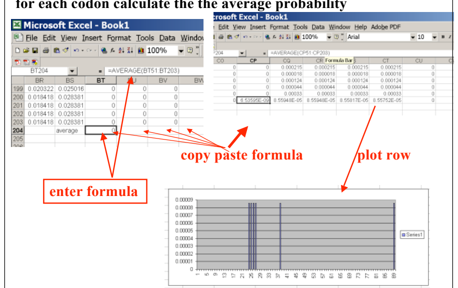
MrBayes analyzing the *.nex.p file

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 exercise 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



for each codon calculate the the average probability



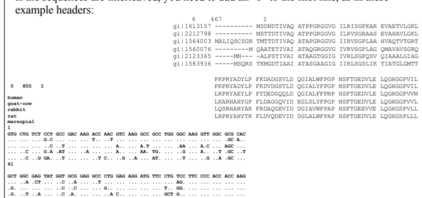
MrBayes on bbcxrv1

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:



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.

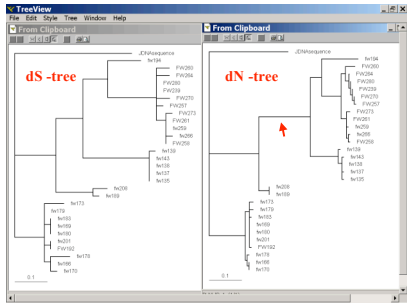
The output is written, as directed by the control file, into a file called Hv1.branch.codeml_out

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

else

hy-phy (hypothesis testing using phylogenetics) does very well in analyzing selection pressures.

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

hy-phy

Results of an anaylsis using the SLAC approach