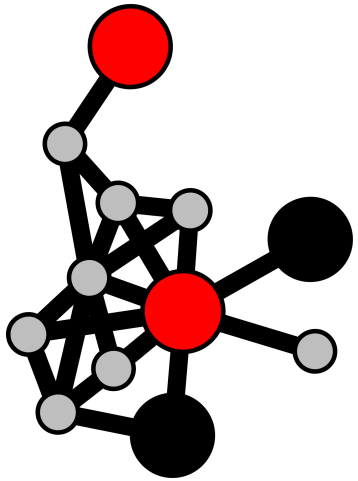
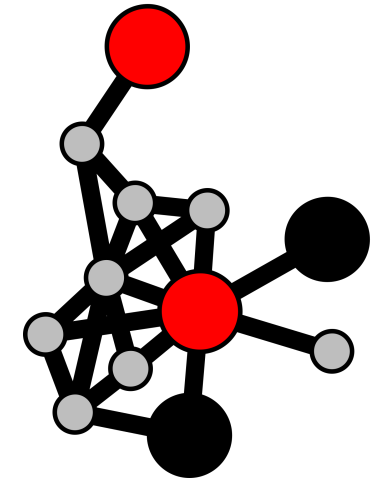
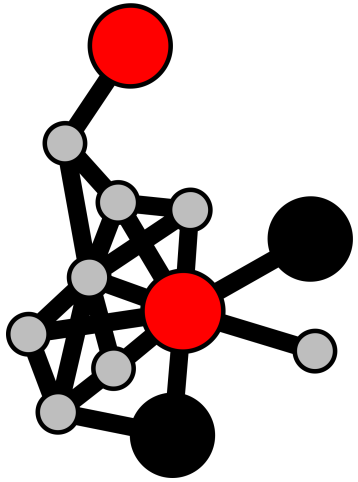
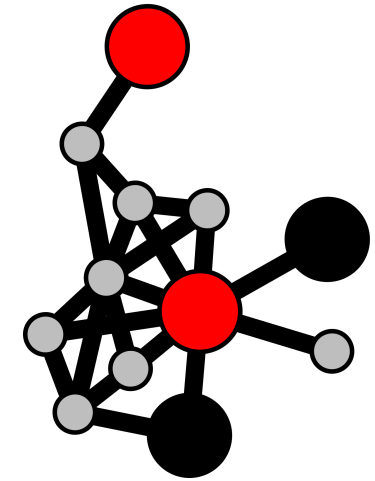




Comparative Genomics

Lesson 1: Basics of dN/dS and using PAML





Synonymous vs. Nonsynonymous mutations

- **Synonymous mutation**: one that results in a change in the nucleotide sequence, but does not result in a change in the encoded peptide.
  - Provides a metric of neutral evolution of the sequence
- **Nonsynonymous mutation**: one that results in a change in the nucleotide sequence, **and** also results in a change in amino acid sequence
  - Provides of metric for the conservation of protein sequence, in the context of the neutral changes estimated by dS (the rate of synonymous substitutions)

Silent vs nonsilent mutations

1st letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	3rd letter
U	UUU Phe UUC UUA Leu UUG	UCU UCC Ser UCA UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G
C	CUU CUC Leu CUA CUG	CCU CCC Pro CCA CCG	CAU His CAC CAA Gln CAG	CGU CGC Arg CGA CGG	U C A G
A	AUU AUC Ile AUA AUG Met	ACU ACC Thr ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU GUC Val GUA GUG	GCU GCC Ala GCA GCG	GAU Asp GAC GAA Glu GAG	GGU GGC Gly GGA GGG	U C A G

http://biology.kenyon.edu/courses/biol114/Chap05/Chapter05.html

dN/dS

Silent, or synonymous, mutations → **dS** = synonymous mutations per possible synonymous sites

Nonsilent, or nonsynonymous, mutations → **dN** = nonsynonymous mutations per possible nonsynonymous sites

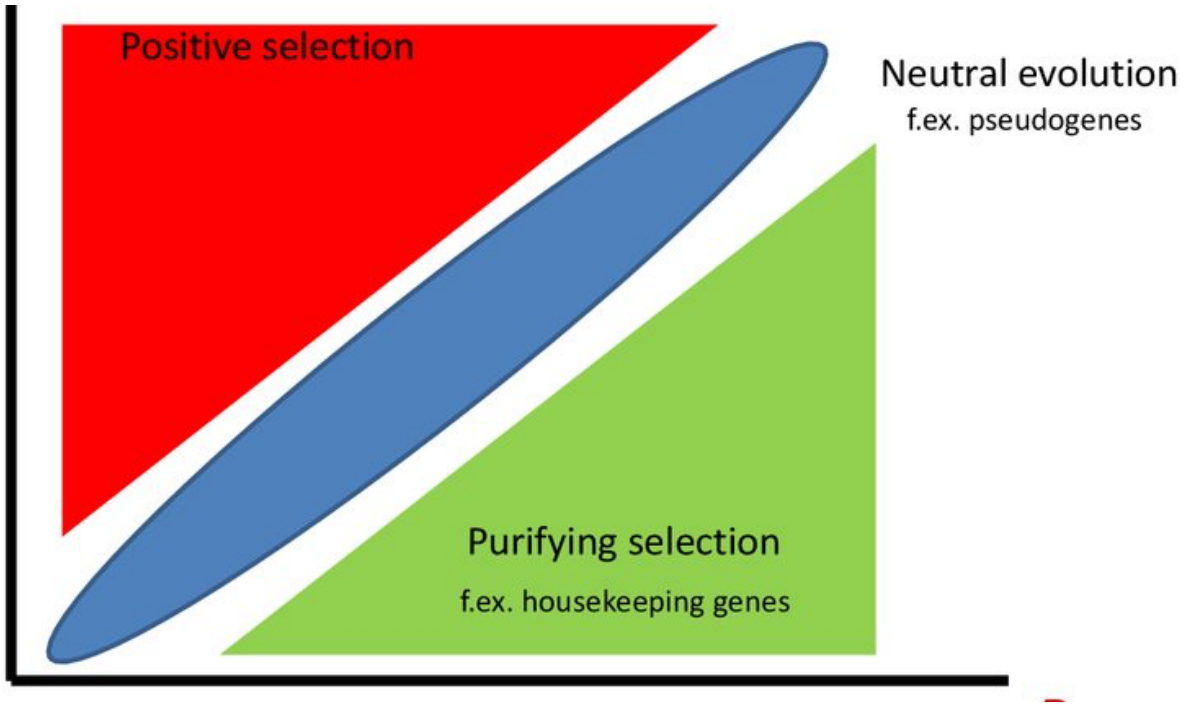
dN/dS → infer conservation or divergence of a particular peptide sequence given the amount of neutral (synonymous) genetic drive that has occurred

Estimating selection using dN/dS

dN/dS = 1 → = **neutral selection of entire sequence**

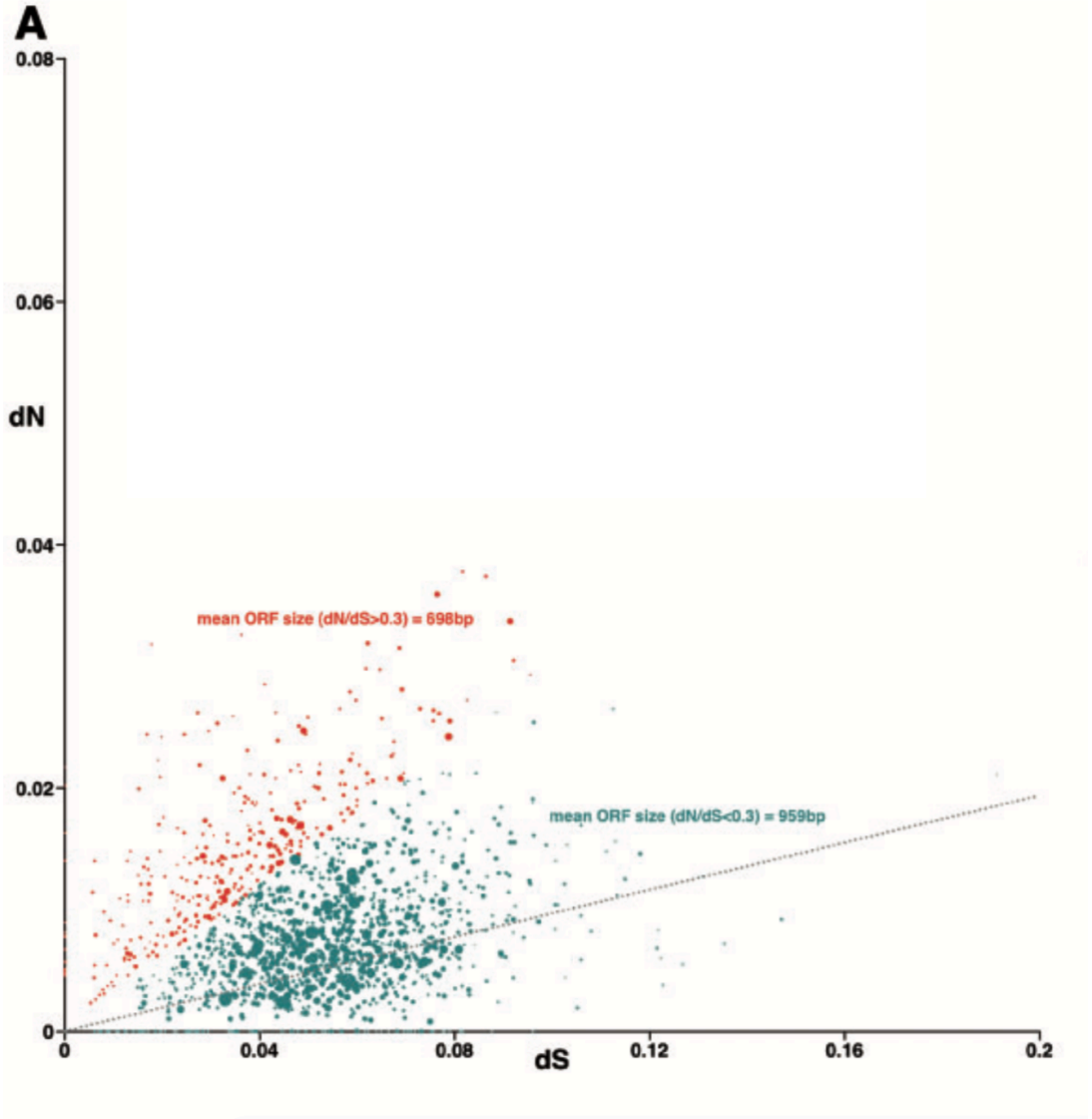
High dN/dS (> 1) = high rate of nonsynonymous substitutions, given a certain rate of synonymouns substitutions = **positive selection and adaptive evolution**

Low dN/dS (< 1) = low rate of nonsynonymous substitutions, given a certain rate of synonymouns substitutions = **purifying selection**



Taken from Max Planck Institute: IMPRS seminar summer term 2020

Estimating selection using dN/dS



Oakeson et al., 2014; Genome Biol. Evol. 6(1):76–93

In reality:

High dN/dS (> 0.2/0.3) = high rate of nonsynonymous substitutions, given a certain rate of synonymouns substitutions = **neutral or positive selection**

Low dN/dS (< 0.2) = low rate of nonsynonymous substitutions, given a certain rate of synonymouns substitutions = **purifying selection**

Limitations of dN/dS

- Not appropriate for non-coding sequences (rRNA, tRNA, tmRNA, other types of small regulatory RNAs)
- Mutational saturation can lead to back-mutations that underestimate divergence (recommended dS < 1)
- Near-identical sequences with only a few SNPs can lead to over-esimtation of dN/dS → (recommended dS > 0.01)
- “Masking” of positions or domains under high positive or purifying selection, if the entirety of the sequence is conserved

PAML (Phylogenetic Analysis by Maximum Likelihood)

A program package by Ziheng Yang

(Demonstration by Joseph Bielawski)

Next four slides (including this one) from PAML website

What does PAML do?

Features include:

- estimating synonymous and nonsynonymous rates
- testing hypotheses concerning d_N/d_S rate ratios
- various amino acid-based likelihood analysis
- ancestral sequence reconstruction (DNA, codon, or AAs)
- various clock models
- simulating nucleotide, codon, or AA sequence data sets
- and more ……

Programs in the package

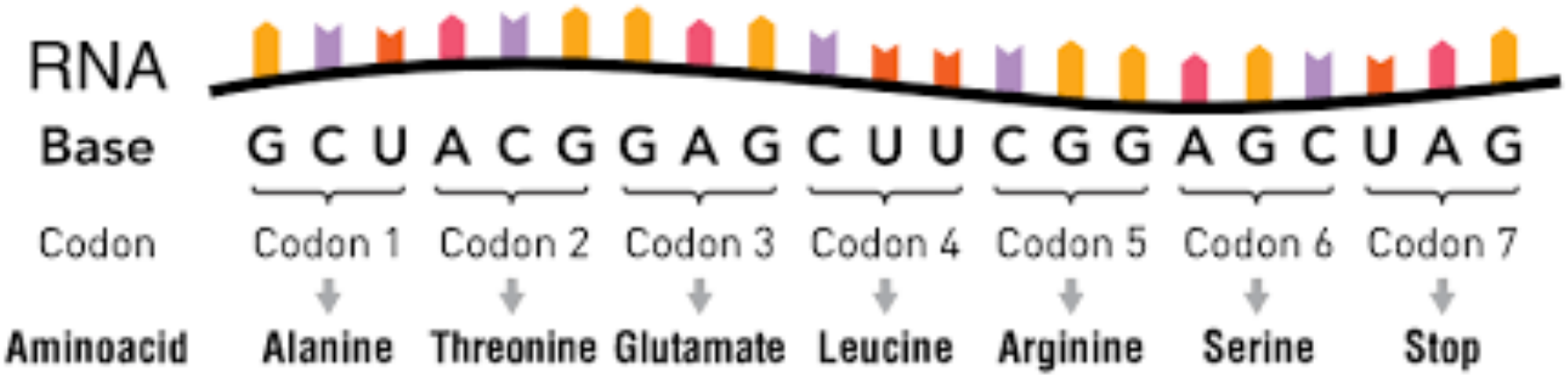
baseml	for bases
basemlg	continuous gamma for bases
codeml	aaml for amino acids & codonml for codons
evolver	simulation, tree distances
yn00	d_N and d_S by Yang & Nielsen (2000)
chi2	chi square table
pamp	parsimony (Yang and Kumar 1996)
mcmctree	Bayesian MCMC divergence time estiamtion, under soft bounds (Yang & Rannala 2006)

Running PAML programs

1. Sequence data file
2. Tree file
3. Control file (*.ctl)

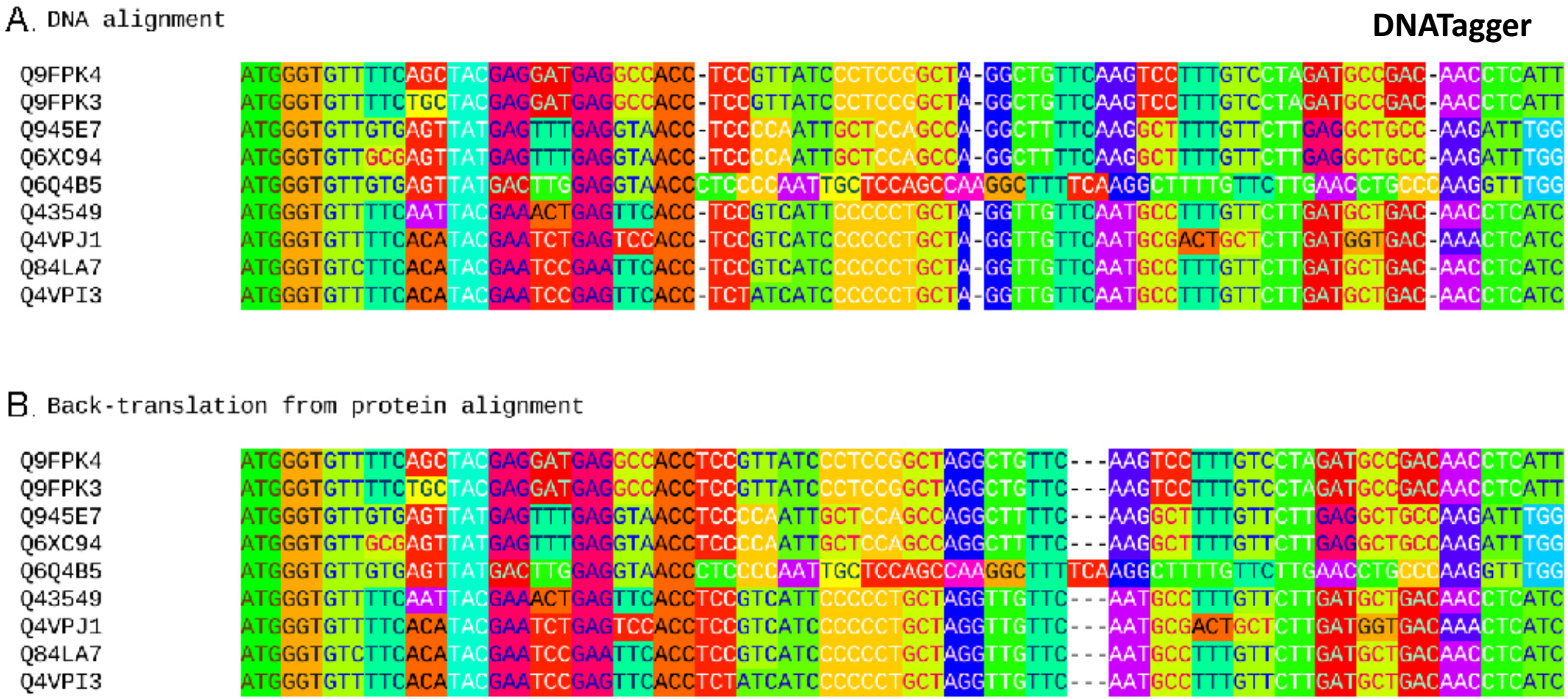
Making a codon alignment

Needs a protein alignment, and associated nucleotide sequences



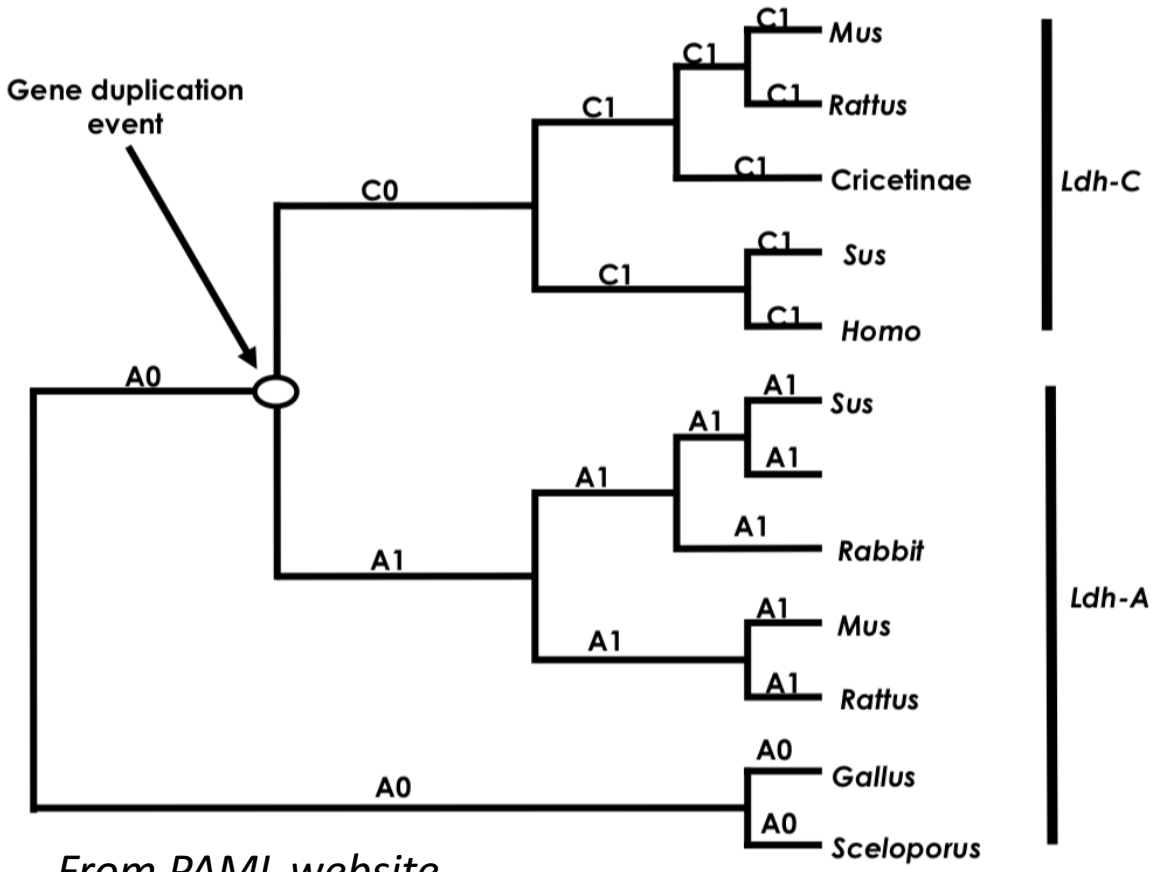
pal2nal.pl → codon alignment

Codon alignment



Sherrer and Basso, 2008, Genetics and Molecular Research

Phylogentic tree input - parenthetic (Newick) format

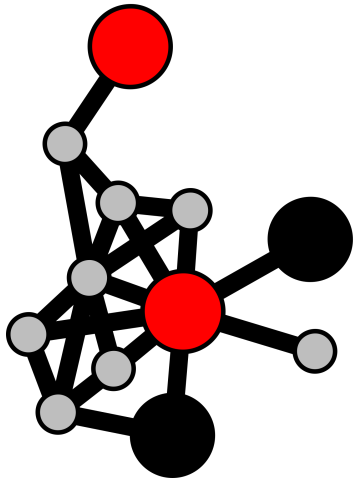
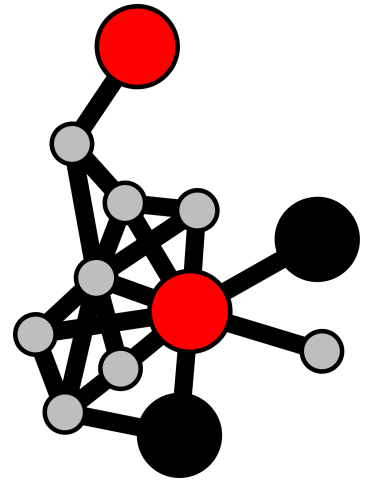


From PAML website

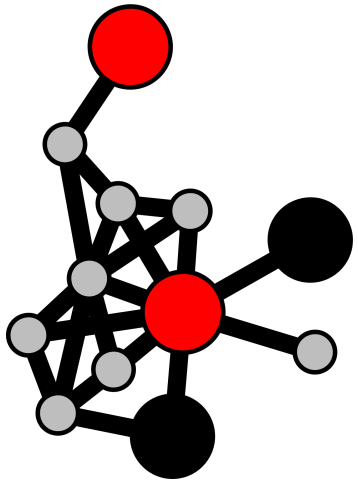
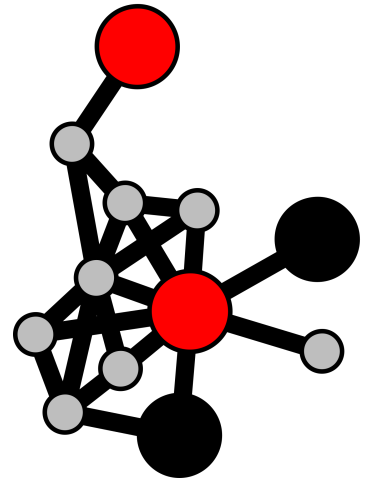
Input to codeml: **codeml.ctl** file – contains locations of input and output files, as well as parameters

codeml-2.ctl

```
     seqfile = zinc_transporter_ZitB.codonalign.fna * sequence data filename
     treefile = gtotree_out3.tree
     outfile = mlc           * main result file name

       noisy = 0  * 0,1,2,3,9: how much rubbish on the screen
     verbose = 1  * 0: concise; 1: detailed, 2: too much
     runmode = -2  * 0: user tree;  1: semi-automatic;  2: automatic
                 * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

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

*       ndata = 10
       clock = 0  * 0:no clock, 1:clock; 2:local clock; 3:CombinedAnalysis
      aaDist = 0  * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a
  aaRatefile = dat/jones.dat  * only used for aa seqs with model=empirical(_F)
                 * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own

       model = 2
                 * models for codons:
                     * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
                 * models for AAs or codon-translated AAs:
                     * 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
                     * 6:FromCodon, 7:AAClasses, 8:REVaa_0, 9:REVaa(nr=189)

     NSsites = 0  * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
                 * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
                 * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
                 * 13:3normal>0

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

    fix_kappa = 0  * 1: kappa fixed, 0: kappa to be estimated
        kappa = 2  * initial or fixed kappa
    fix_omega = 0  * 1: omega or omega_1 fixed, 0: estimate
        omega = .4 * initial or fixed omega, for codons or codon-based AAs

    fix_alpha = 1  * 0: estimate gamma shape parameter; 1: fix it at alpha
        alpha = 0. * initial or fixed alpha, 0:infinity (constant rate)
       Malpha = 0  * different alphas for genes
        ncatG = 8  * # of categories in dG of NSsites models

       getSE = 0  * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 1  * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)

   Small_Diff = .5e-6
    cleandata = 1  * remove sites with ambiguity data (1:yes, 0:no)?
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
       method = 0  * Optimization method 0: simultaneous; 1: one branch a time

* Genetic codes: 0:universal, 1:mammalian mt., 2:yeast mt., 3:mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of GENEBANK.
```

Onto to the Jupyter binder tutorial!

Counts of possible synonymous sites for each gene (S)

	1	2	3	4	5
	Pro	Phe	Gly	Leu	Phe
Seq 1	CCC	UUU	GGG	UUA	UUU
Seq 2	CCC	UUC	GAG	CUA	GUA
	Pro	Phe	Ala	Leu	Val

Calculate potential synonymous sites (S) for each codon

A fourfold degenerate site counts as S = 1 (N = 0)

A non-degenerate site counts as S = 0 (N = 1)

A two fold degenerate site counts as S = 1/3 (N = 2/3)

1. Proline S = 0 + 0 + 1 = 1
2. Phenylalanine S = 0 + 0 + 1/3 = 1/3
3. For Glycine S = 0 + 0 + 1 = 1, for Alanine S = 0 + 0 + 1
   Take the average: S=1
4. Leucine for UUA, S = 1/3 + 0 + 1/3 = 2/3
   for CUA, S = 1/3 + 0 + 1 = 4/3
   Take the average of these: S = 1 for codon 4
5. Phenylalanine for UUU, S = 1/3
   for guanine, S = 1
   Take average: S = 2/3

For whole sequence, S = 1 + 1/3 + 1 + 1 + 2/3 = 4

N = total number of sites: S = 15 - 4 = 11

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

Slide from Max Planck Institute: IMPRS seminar summer term 2020

Counts of synonymous changes

	1	2	3	4	5
	Pro	Phe	Gly	Leu	Phe
Seq 1	CCC	UUU	GGG	UUA	UUU
Seq 2	CCC	UUC	GAG	CUA	GUA
	Pro	Phe	Ala	Leu	Val

Calculate S_d and N_d for each codon.

1. $S_d = 0$, $N_d = 0$
2. $S_d = 1$, $N_d = 0$
3. $S_d = 0$, $N_d = 1$
4. $S_d = 1$, $N_d = 0$
5. this could happen in two ways

UUU --> GUU --> GUA

$N_d = 1$ $S_d = 1$ Route 1: $S_d = 1$, $N_d = 1$

UUU --> UUA --> GUA

$N_d = 1$ $N_d = 1$ Route 2: $S_d = 0$, $N_d = 2$

Take average of these two:

$S_d = 0.5$, $N_d = 1.5$

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

Total $S_d = 2.5$ Total $N_d = 2.5$

$S_d / S = 2.5/4 = 0.625$ $N_d / N = 2.5/11 = 0.227$

dN/dS = 0.363

Slide from Max Planck Institute: IMPRS seminar summer term 2020