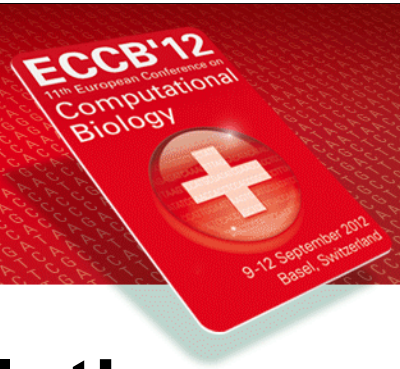
ECCB'12
11th European Conference on
Computational Biology
9-12 September 2012, Basel, Switzerland

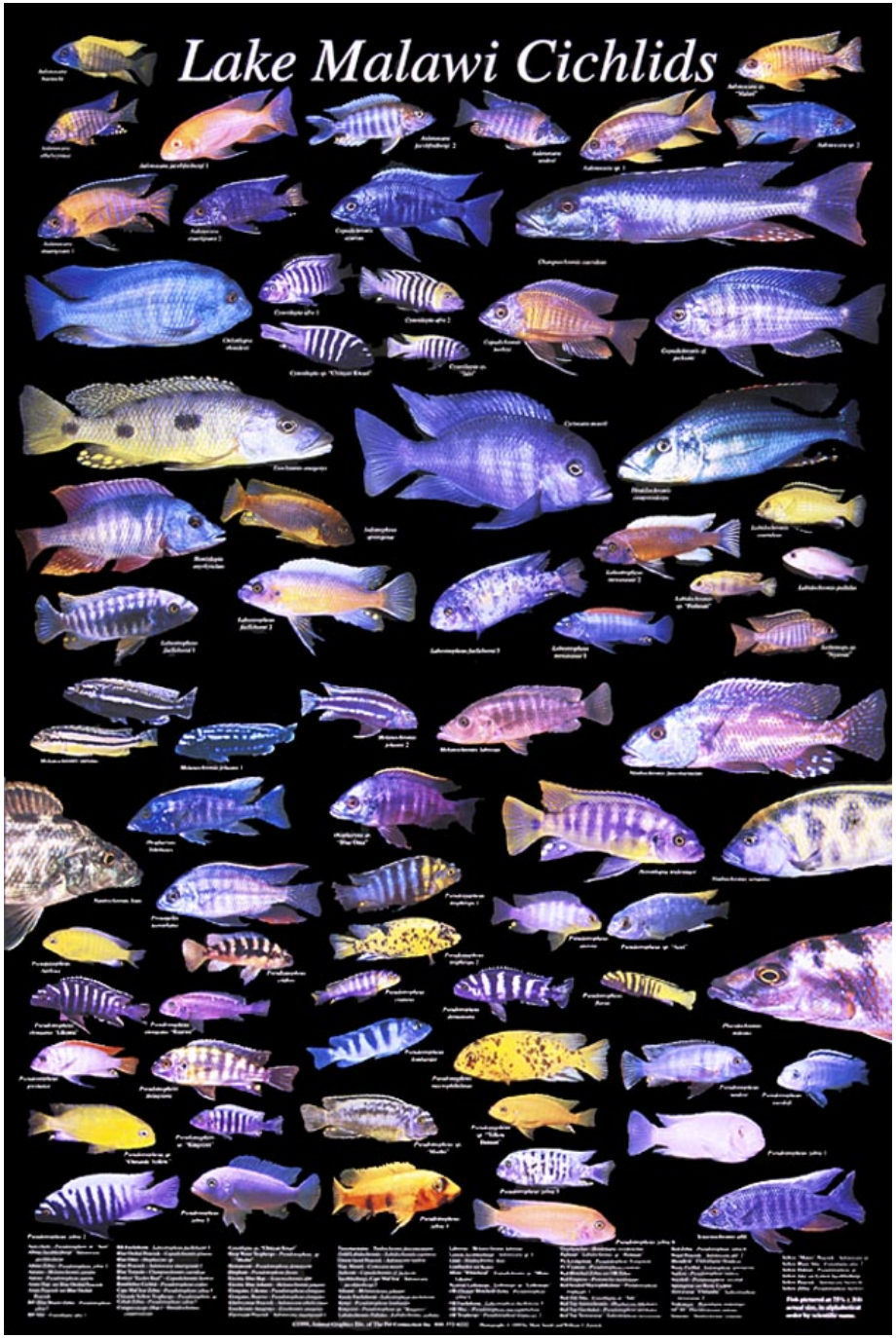


UCL

Detection of evolutionary shifts in protein sequences

Evolution: morphological differences

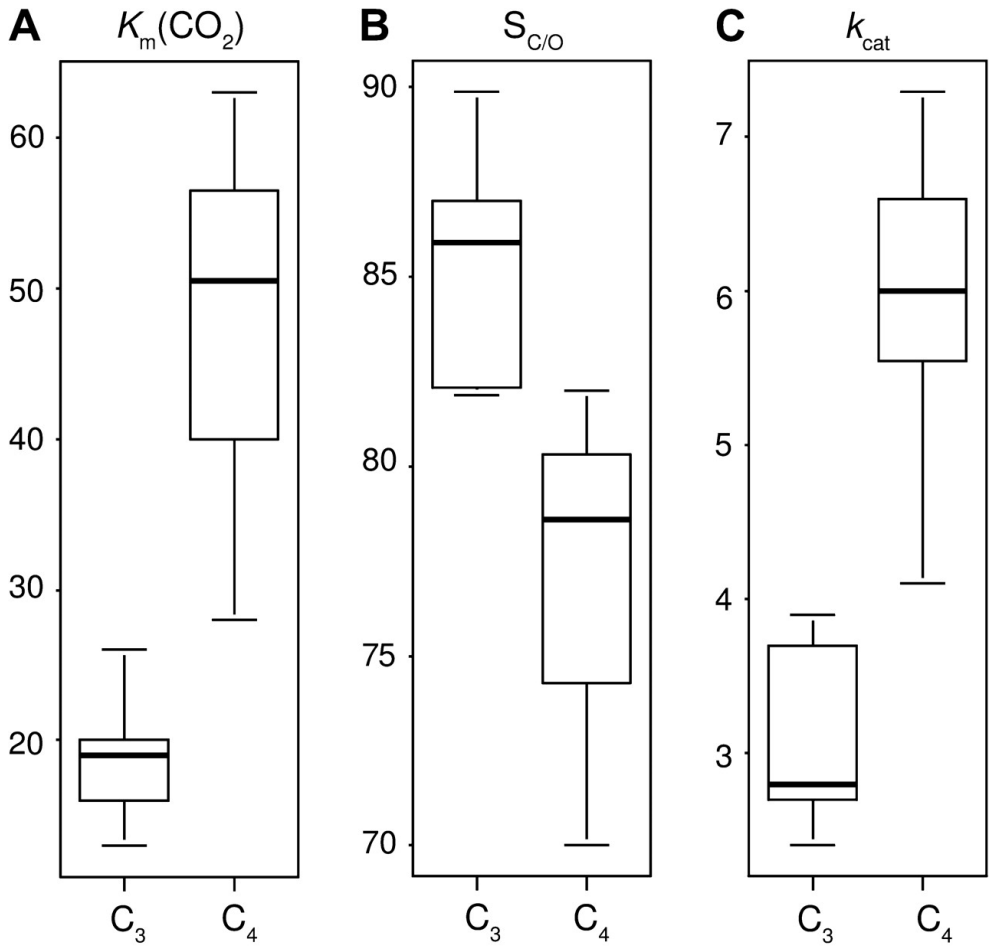
Can we explain the different colour patterns?



Evolution: catalytic activities

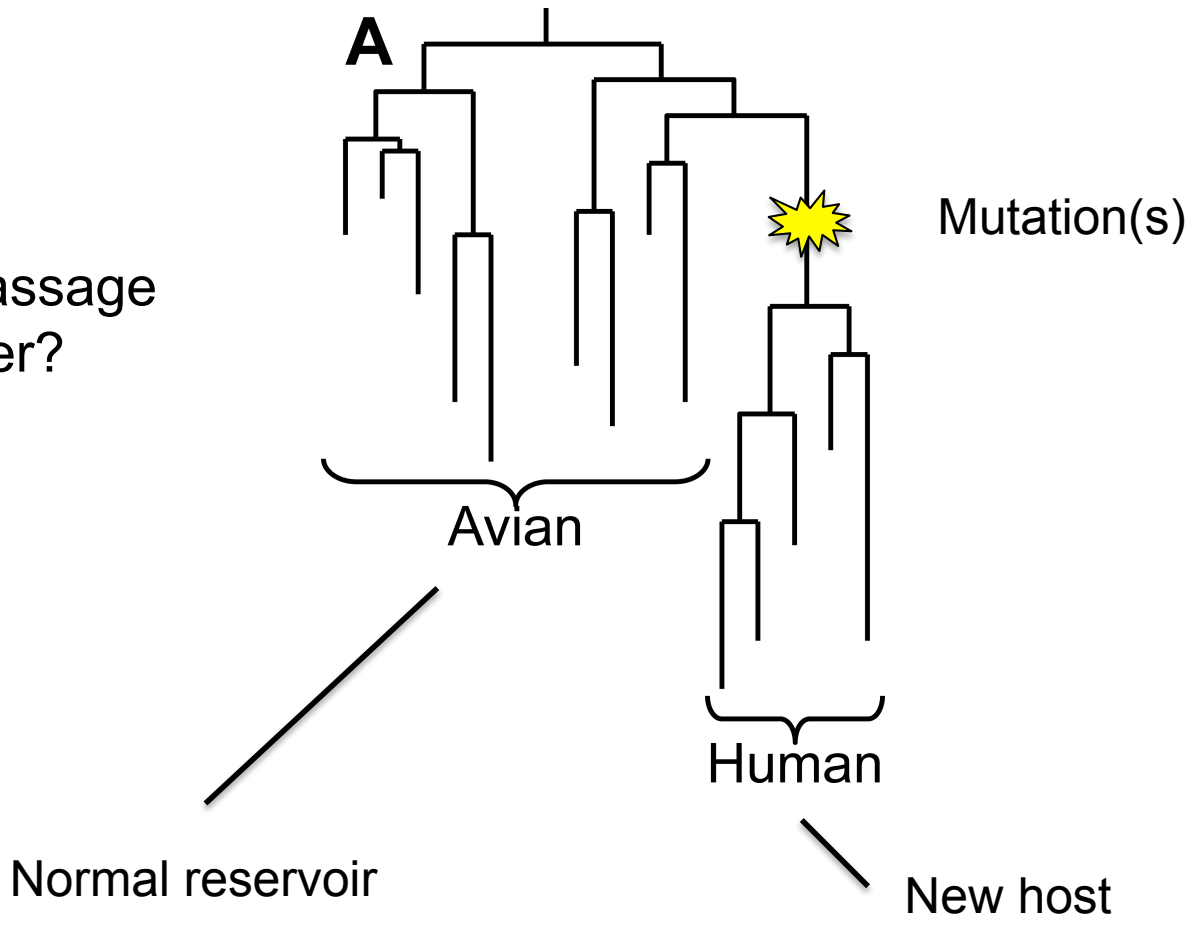
RubisCO enzyme from C3 and C4 forms in plants diverge by different kinetic parameters (affinity, specifity and activity)

Can we explain the different Kinetic parameters?



Evolution: host preferences

Can we explain the passage form an host to another?



Vectors of evolution

- Epigenetics (non genome modification, i.e. DNA methylation).

- Genetics (genome modification):
  - Changes in expression pattern or level.
  - Changes in non coding RNA (ncRNA).
  - Changes in proteins biochemistry:
    - => Change in the nucleotide sequence.
    - => Change in the amino acid sequence.

From genotype to phenotype

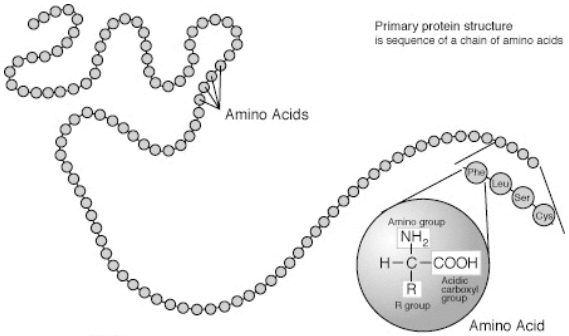
- Genotype (nucleotides, amino acids).

- Phenotype.

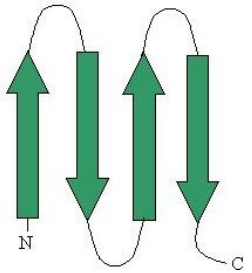
- Submitted to selective pressure.

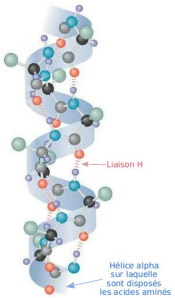
Proteins: 4 levels of organisation

- Primary structure:
  - amino acids sequence.

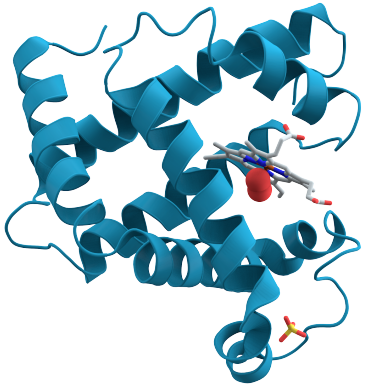


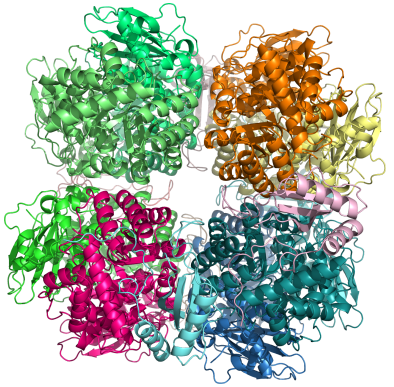
- Secondary structure
  - simple elements.





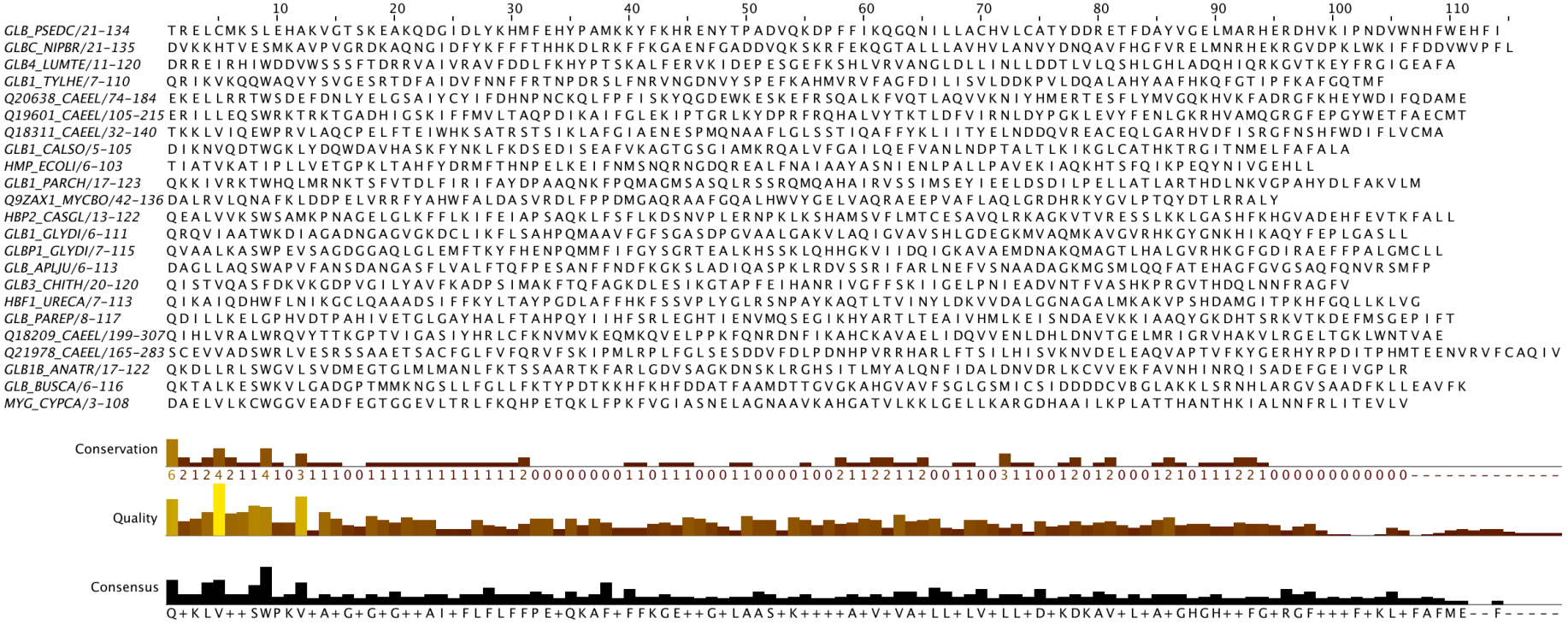
- Tertiary structure:
  - folded in 3D.





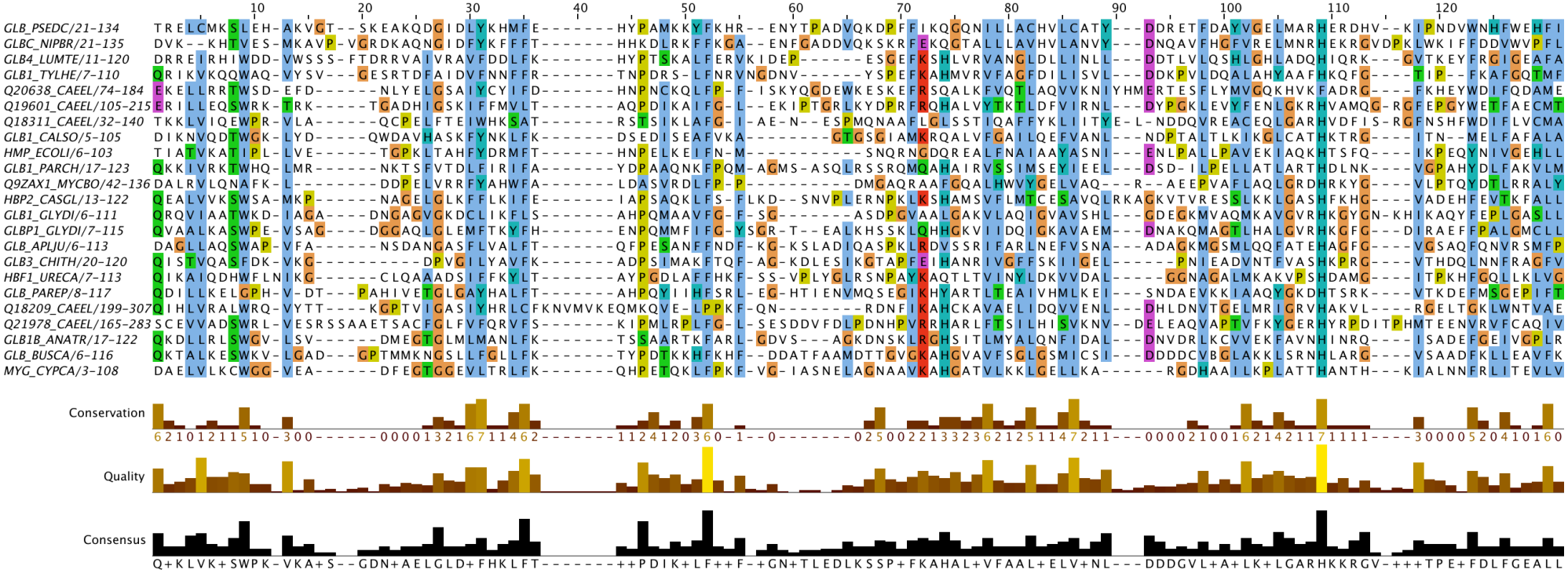
- Quarternary structure:
  - assembly of multiple monomeres.

Protein sequences (globins)



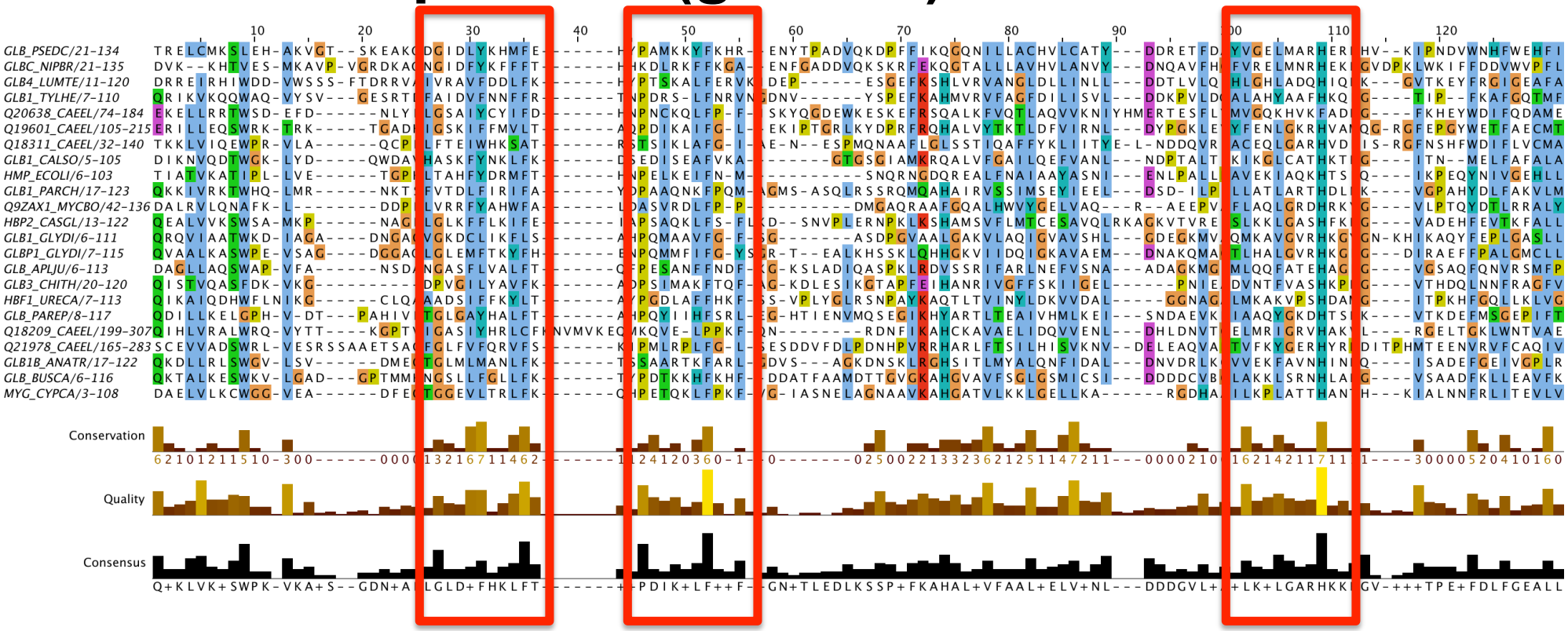
Group of highly divergent sequences

Protein sequences (globins)



Multiple aligner: ClustalW, MUSCLE, MAFFT, PROBCONS, PRANK

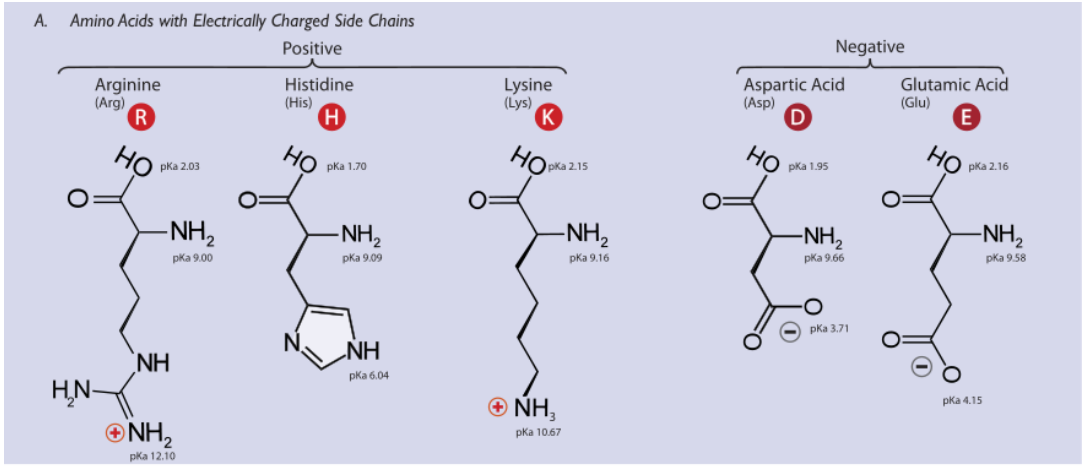
Protein sequences (globins)



Tools to select blocks: Gblocks, Guidance, M-coffee, Pagan

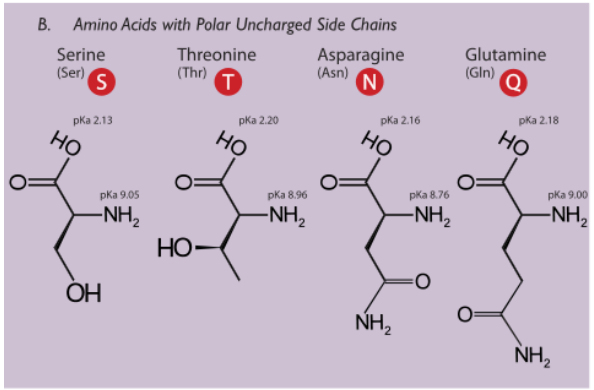
Amino acids have different physicochemical properties

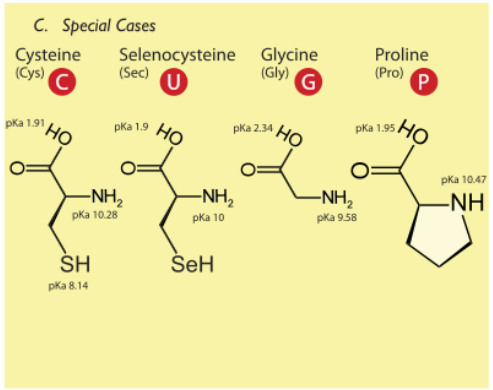
Charged: Acid (-)



Charged: Basic (+)

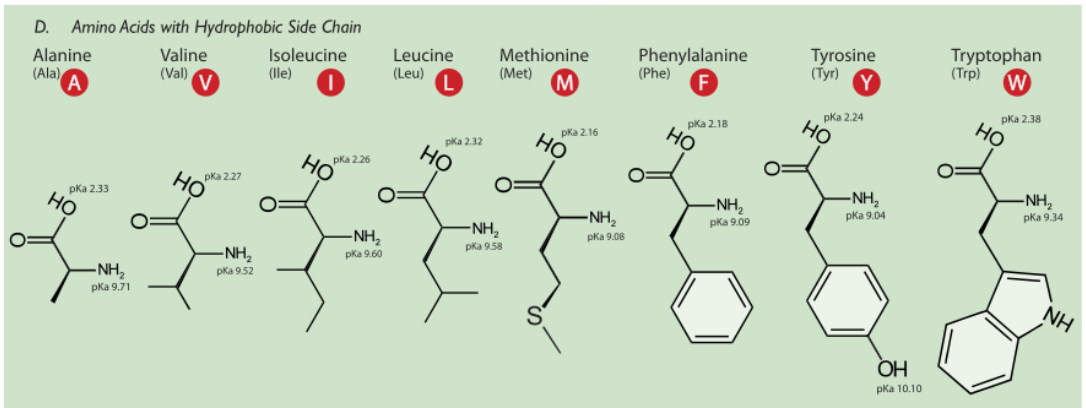
Polar





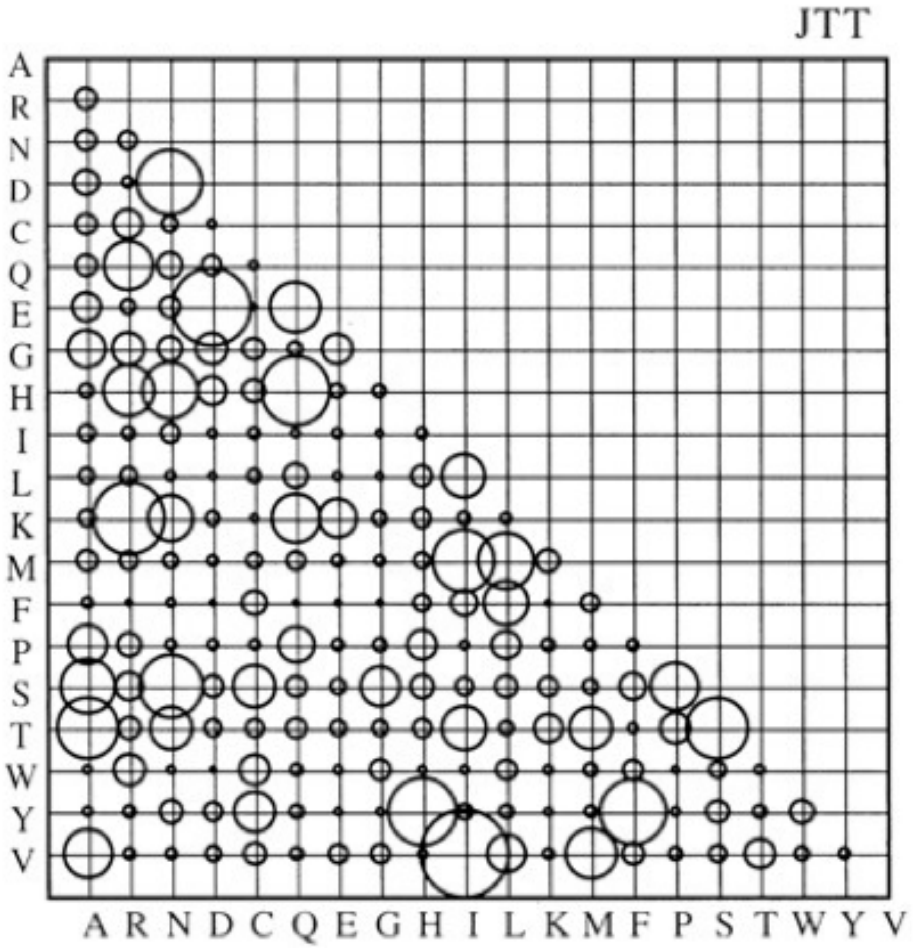
Special cases

Hydrophobic small



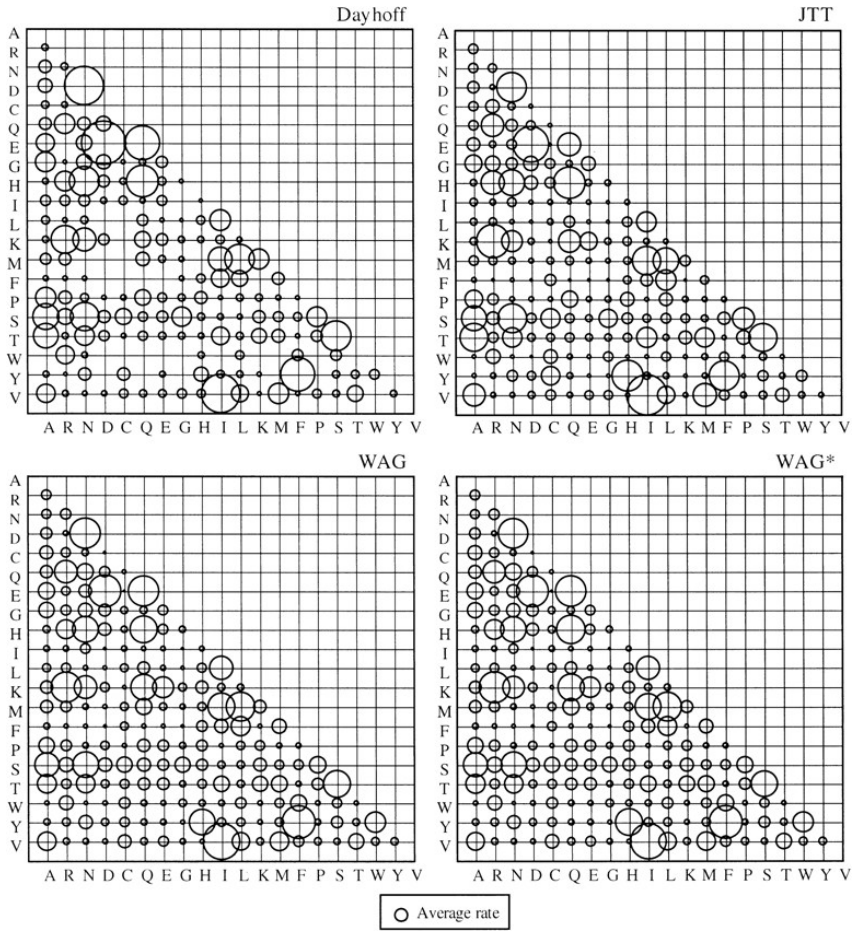
Hydrophobic big

Amino acids have different physicochemical properties
=> Replacements of amino acid are not equivalent



Schematic representations of amino acid replacement matrices.

Dayhoff (Dayhoff, Schwartz, Orcutt, 1978)



JTT (Jones, Taylor WR and Thornton, 1992)

WAG (Whelan and Goldman, 2001)

WAG (Whelan and Goldman, 2001)

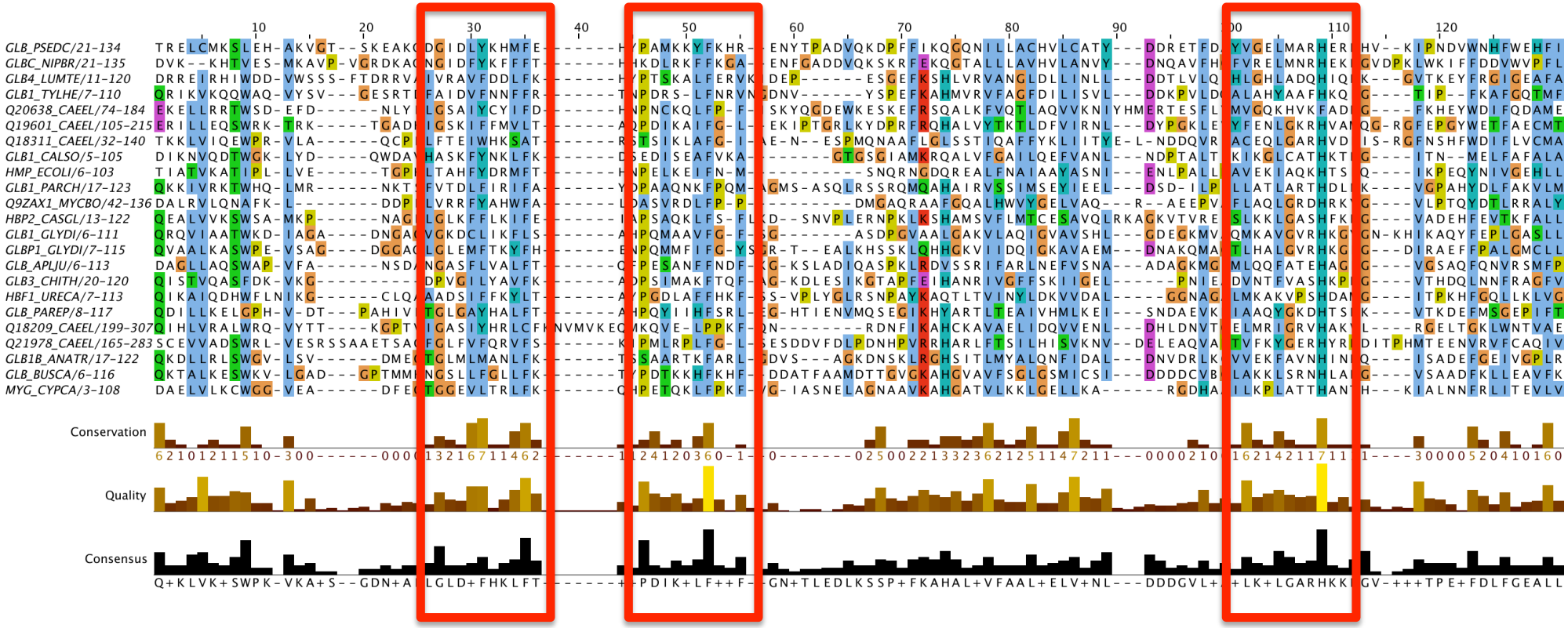
LG (Le and Gascuel, Mol Biol Evol 2008)
LG4M and LG4X (Le, Dang, Gascuel, Mol Biol Evol 2012)

Whelan S , Goldman N
Mol Biol Evol 2001;18:691-699

How to choose the best matrix?

- By experience (i.e. JTT/WAG for Vertebrates, mtREV24 for mitochondrial gene).
- Using tools: ProtTest, ReplacementMatrix

Positions don’t evolve at the same rate



Positions don't evolve at the same rate

Evolutionary rate
(number of mutation per site)

Slow:

- Presence of structural/ functional constraints

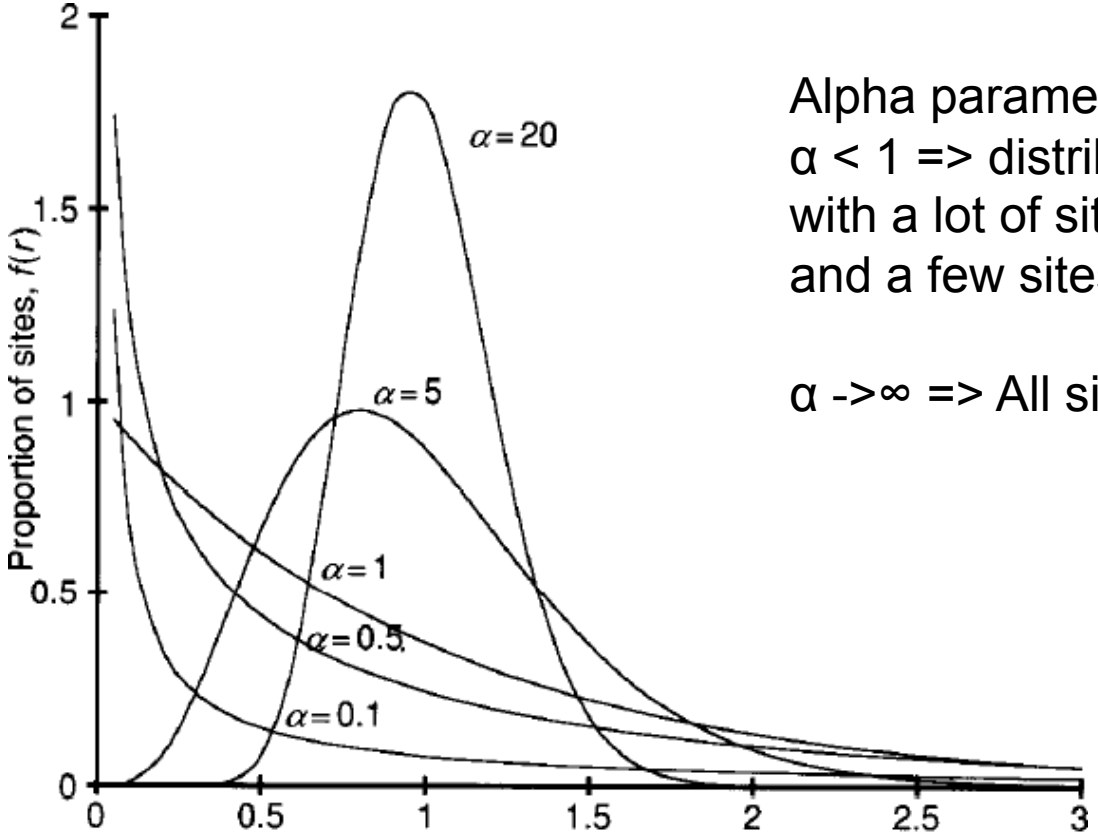
Fast:

- Absence of structural/ functional constraints
- Change in structural/ functional constraints

- Can be described by a **gamma distribution**.
  - Need to cluster sites into different categories of evolutionary rates (generally, 4 or 8 categories).

Positions don’t evolve at the same rate

% of sites



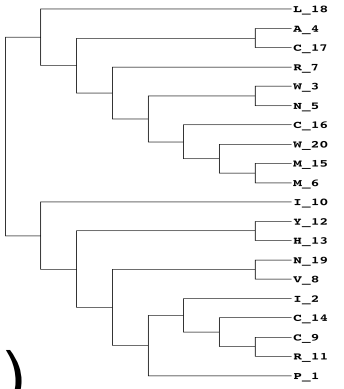
Substitution rate

Alpha parameter:
α < 1 => distribution shaped in **L form**, with a lot of sites evolving slowly and a few sites evolving fast.

α ->∞ => All sites evolve at the same rate

The among-site rate variation and its impact on phylogenetic analyses, Trends Ecol.Evol, Yang 1996

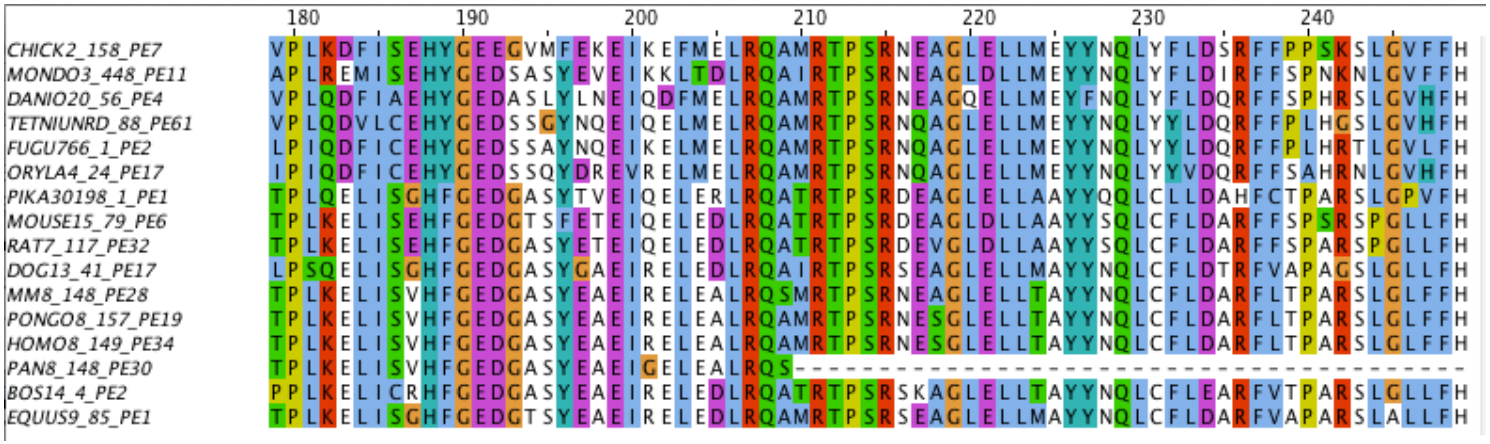
Phylogenetic tree building

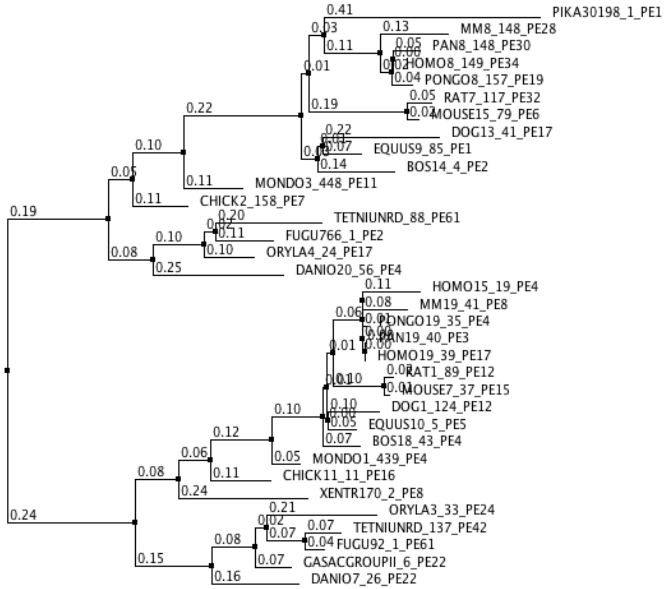


- Data: Multiple alignment of sequence.
- Matrix of replacment rate (i.e. JTT, WAG, LG).
- Use of gamma distribution. => Estimate alpha parameter and categorise sites.

- Method to build the tree:
  - Unweighted Pair Group Method with Arithmetic Mean (UPGMA)
  - Neighbour-Joinging (BioNJ)*
  - Maximum parsimony **
  - Minimum Evolution (ME) (FastME) **
  - Maximum likelihood (ML) (PhyML, RAxML) ***
  - Bayesian (MrBayes) ***

Results: multiple alignment and tree





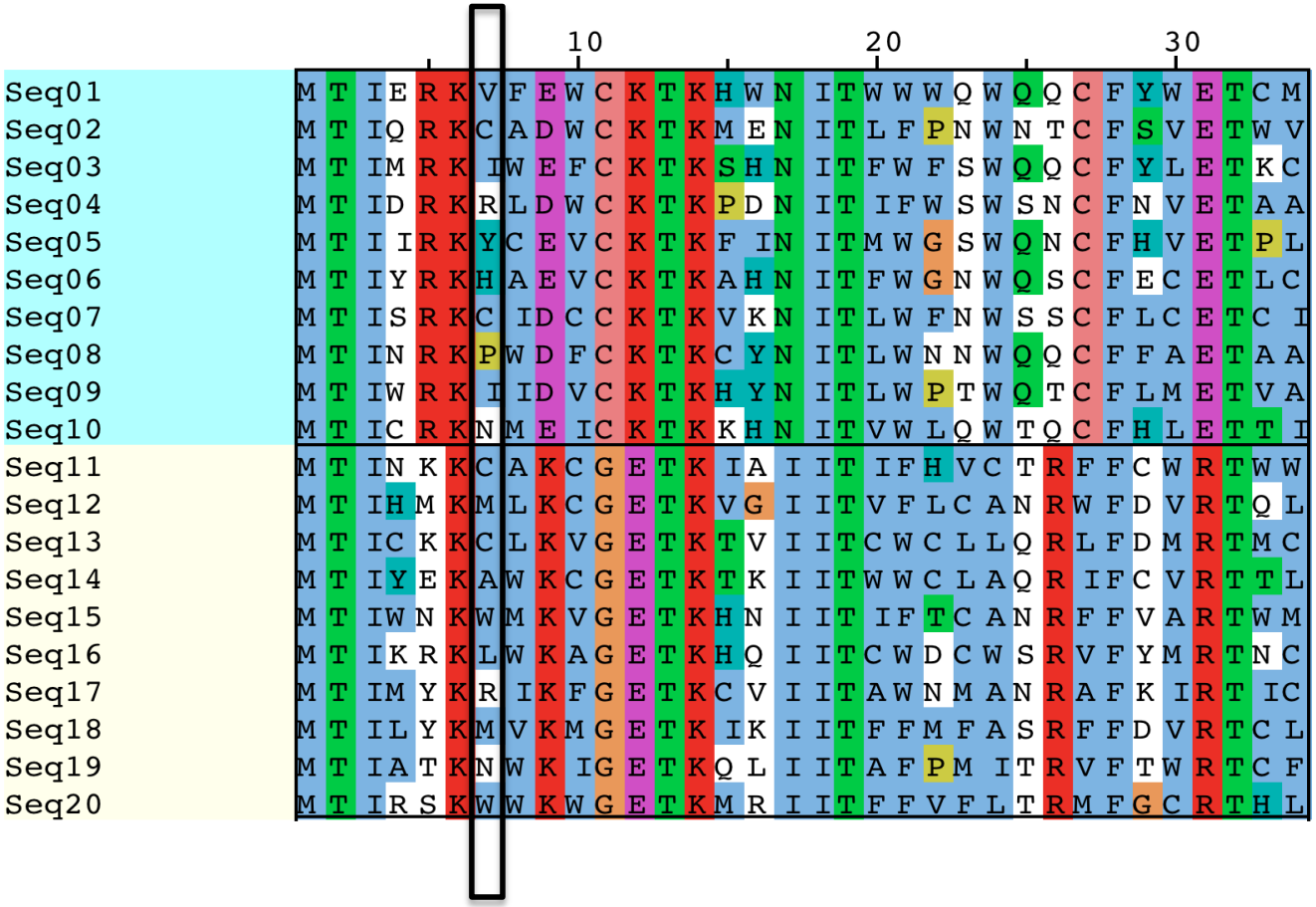
Ancestral Sequence Reconstruction

- Multiple Sequence Alignment
- Phylogenetic tree
- Substitution parameters (Matrix, Gamma distribution)

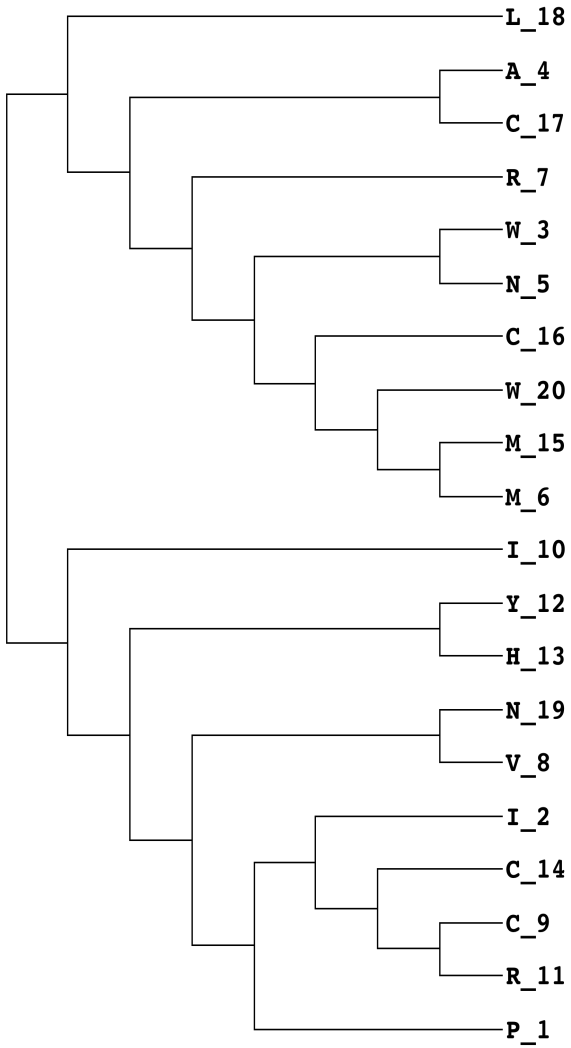
=> Infer each amino acid at each node.

=> Reconstruct ancestral sequences.

Ancestral Sequence Reconstruction

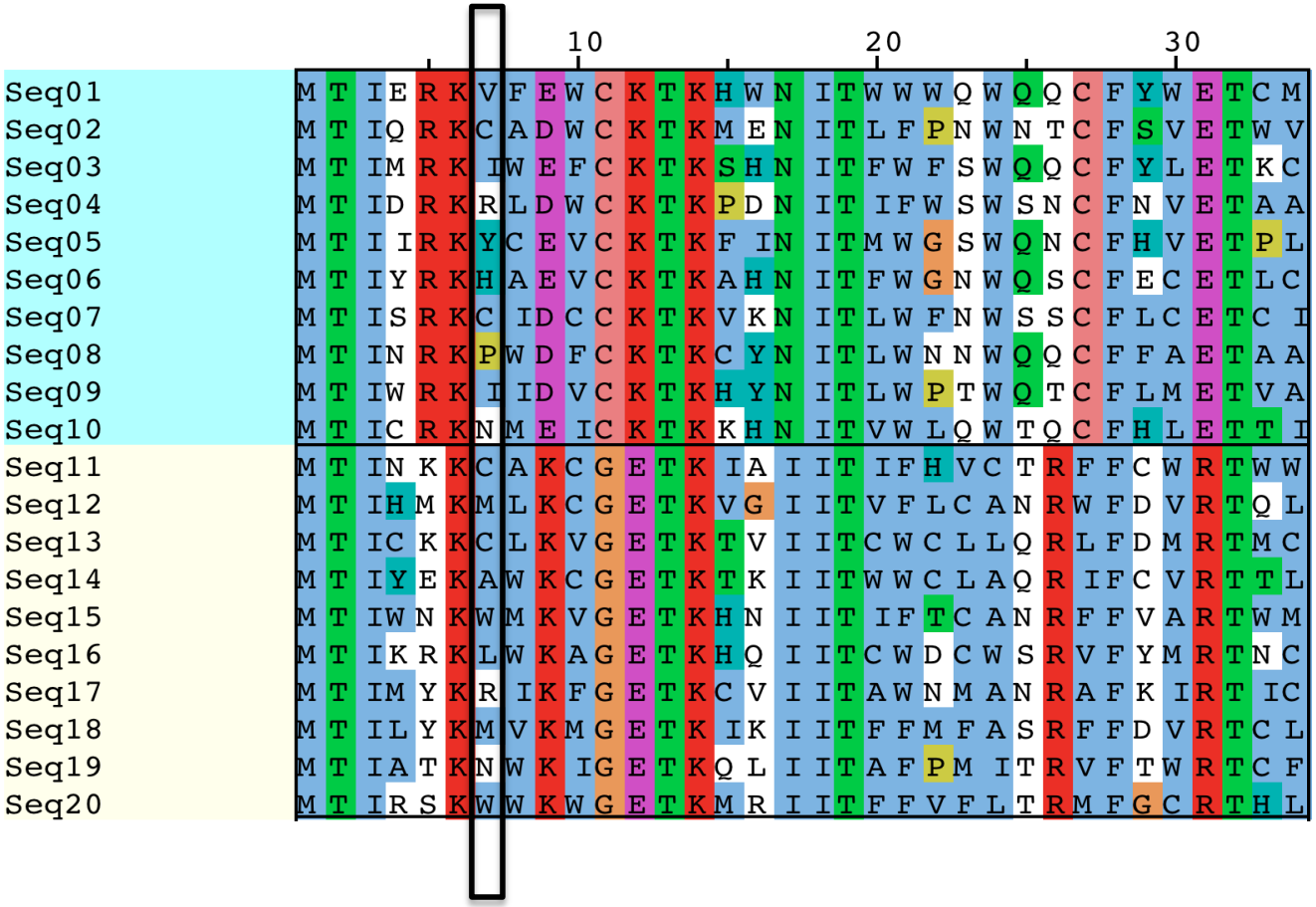


Amino acid at site 7

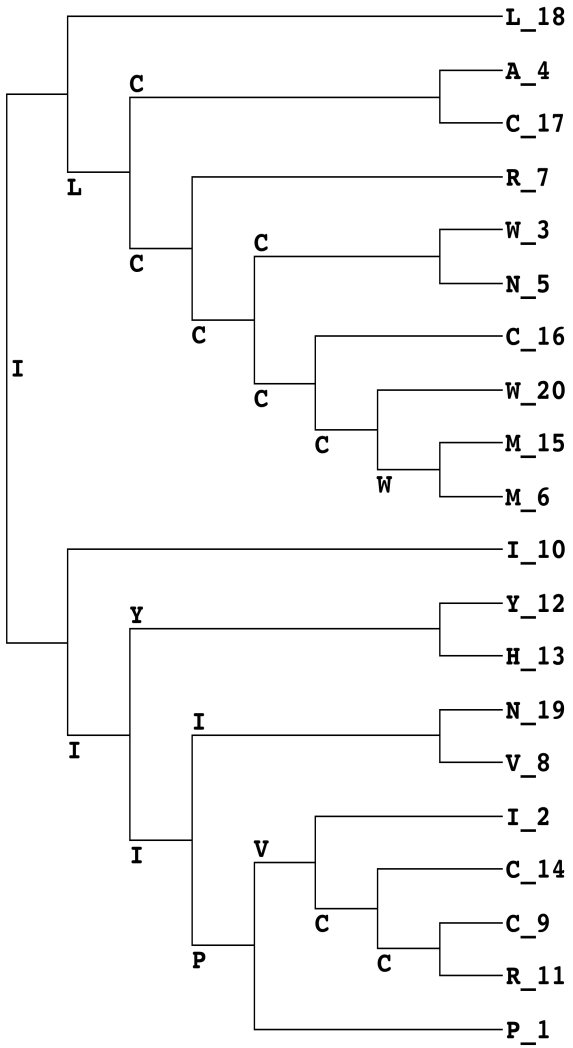


(Simulated alignment)

Ancestral Sequence Reconstruction



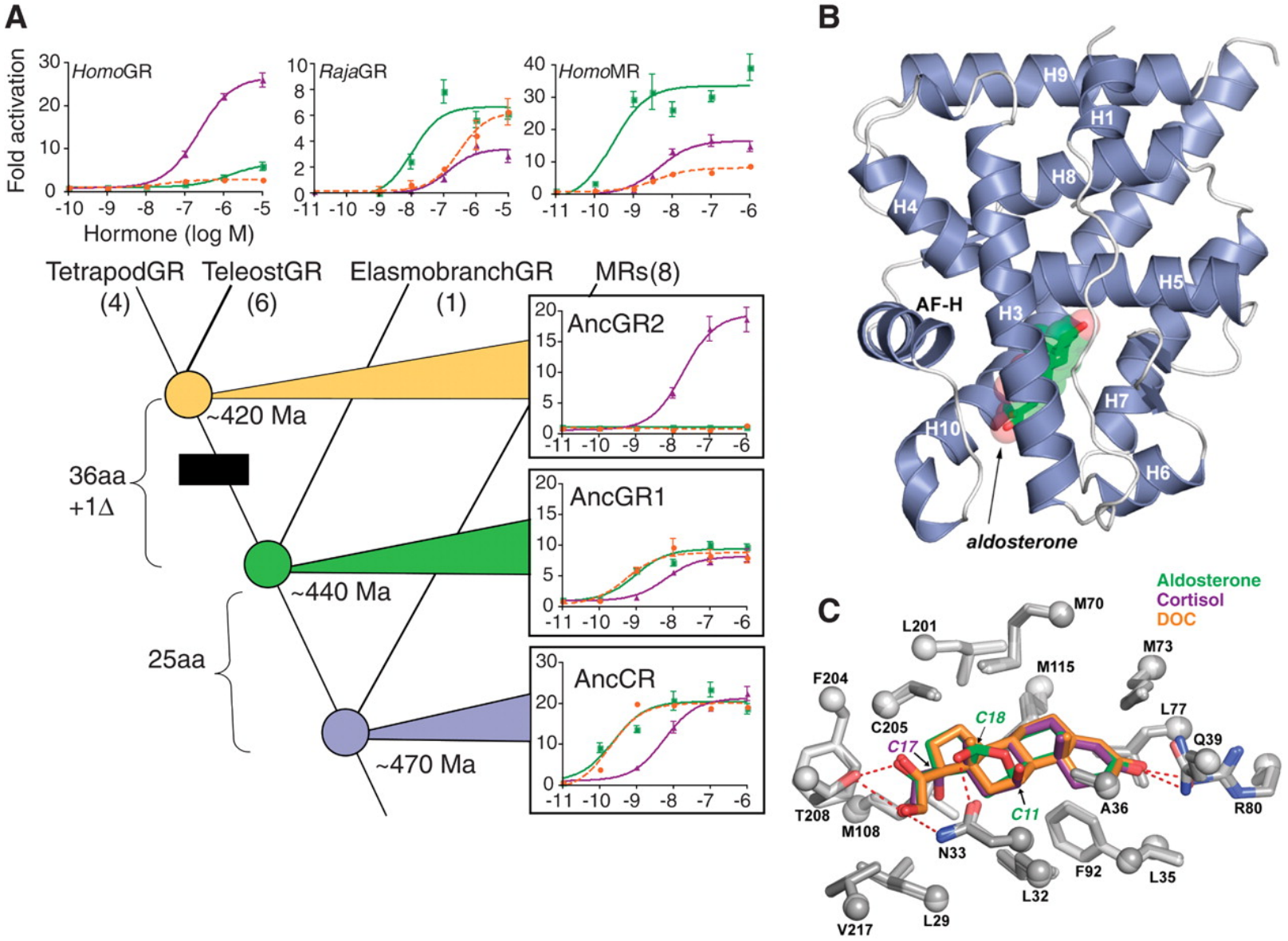
Amino acid at site 7



Control quality with Posterior Probalities

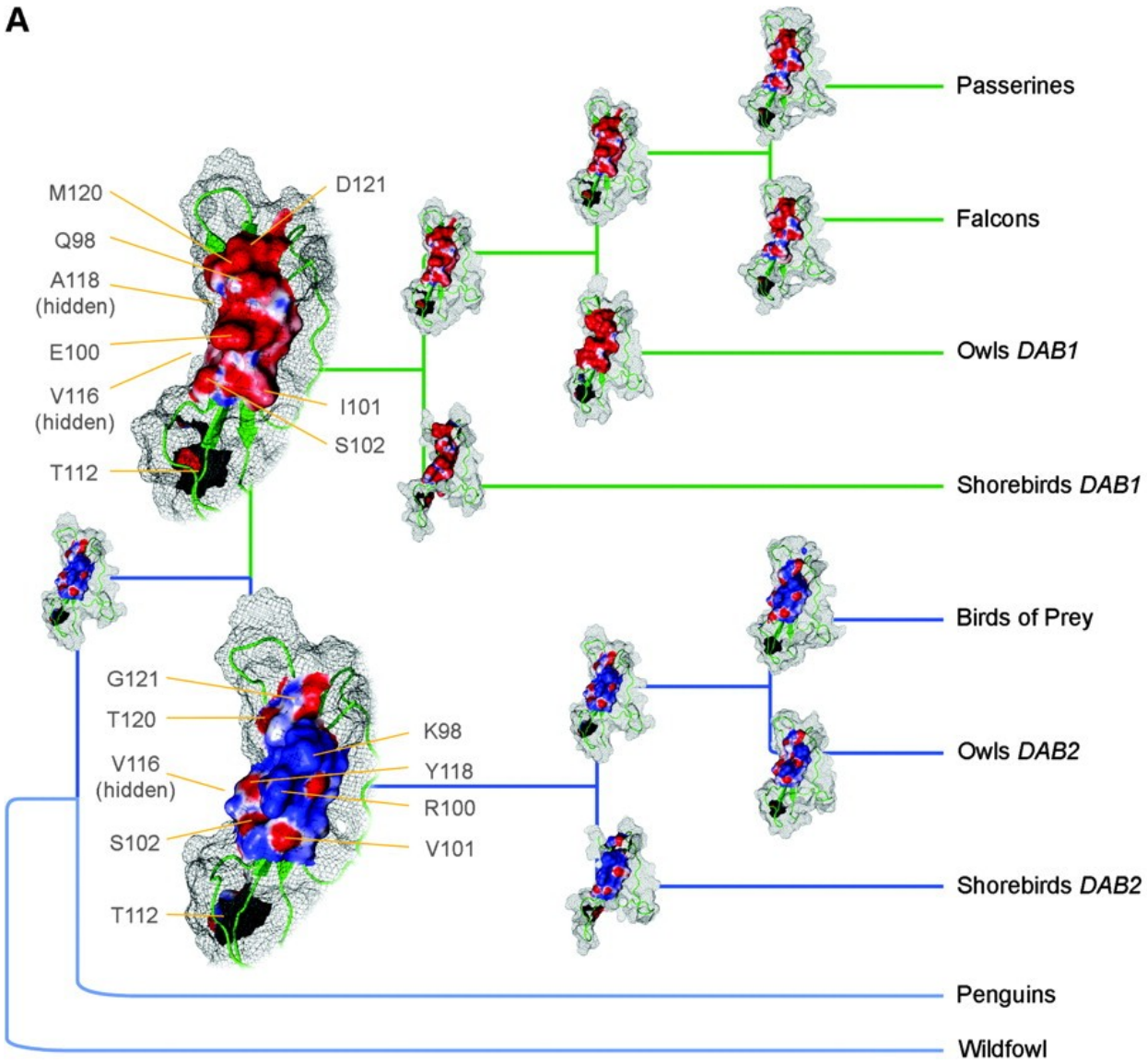
(Simulated alignment)

Ancestral Sequence Reconstruction



Ortlund et al., Science 2007

Ancestral Sequence Reconstruction

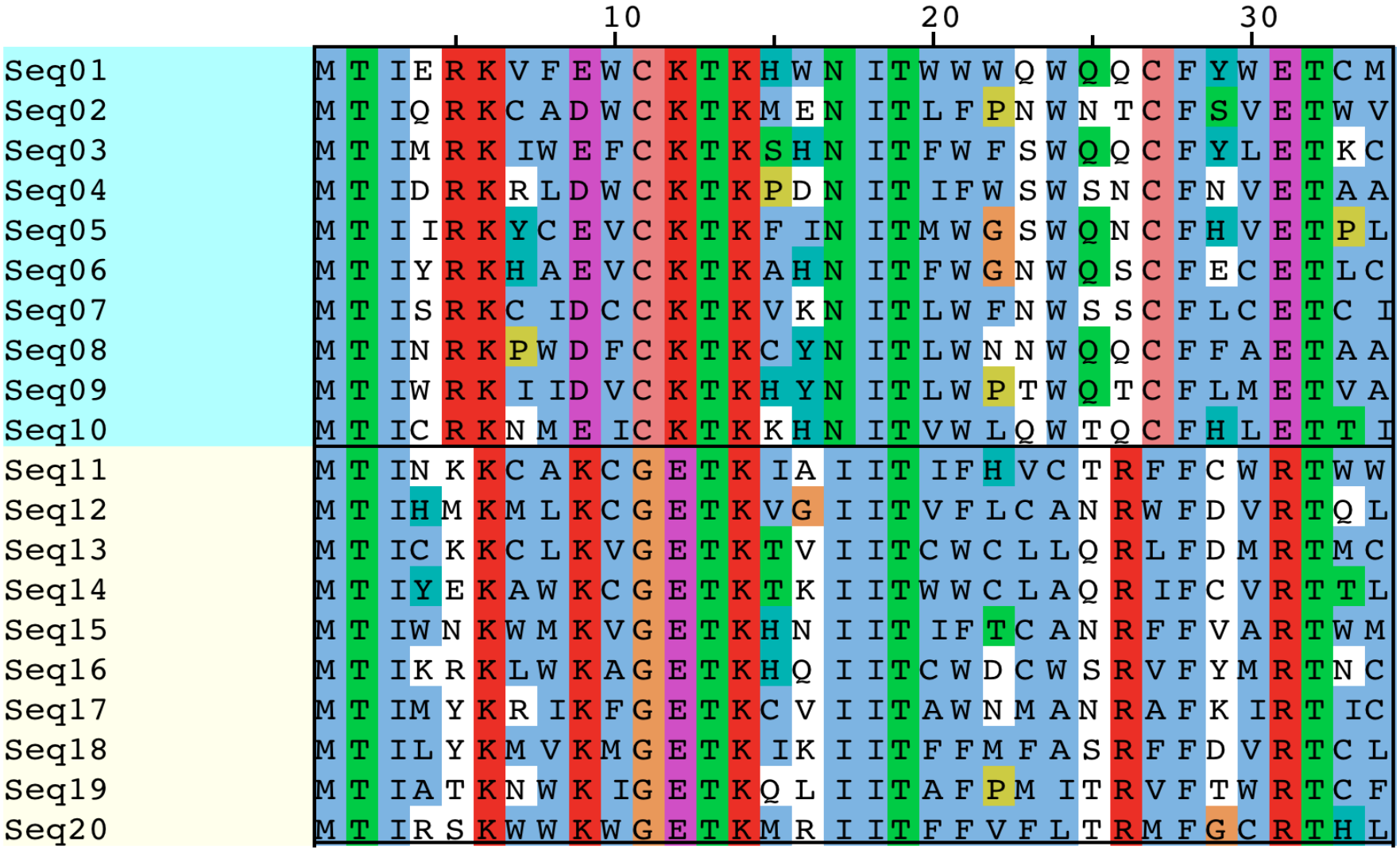


Burri et al., Mol Biol Evol 2010

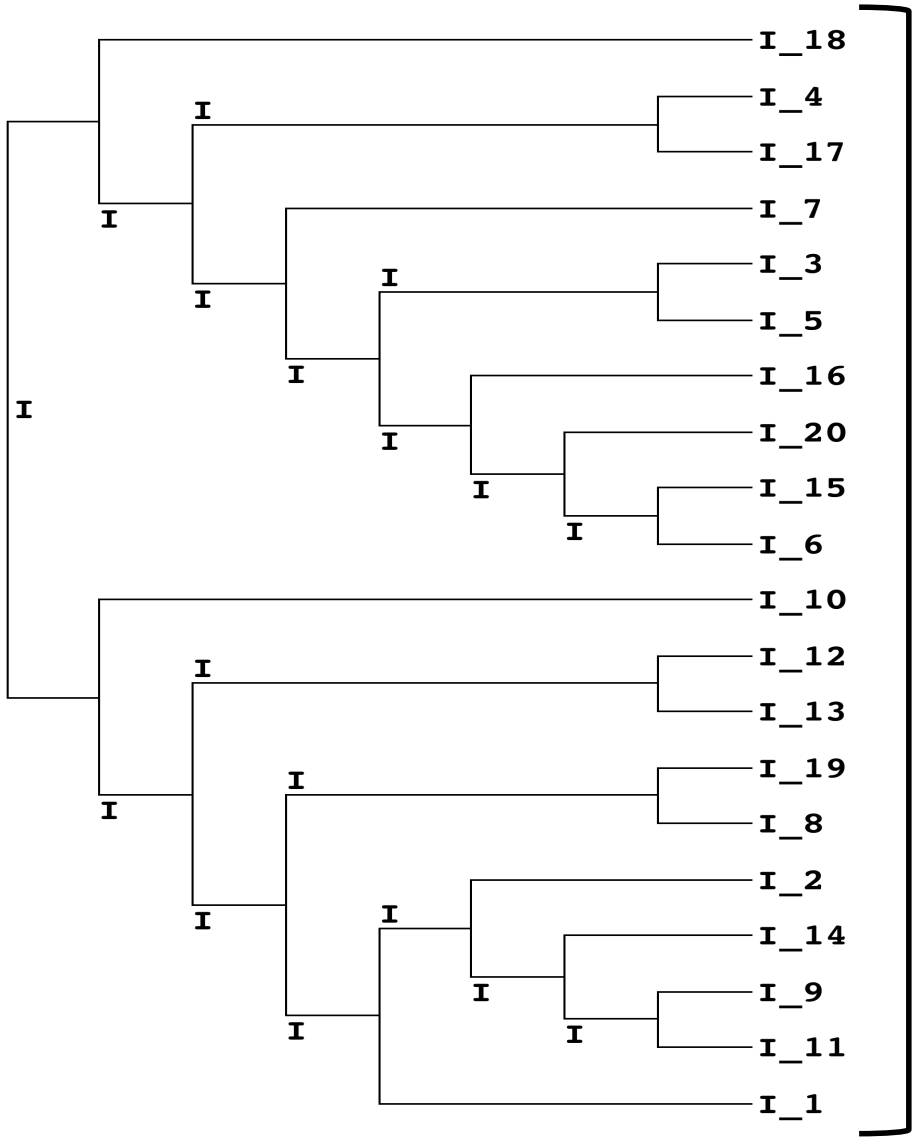
Why detecting events of adaptation in protein?

- Lot of mutations in proteins. Some are important, other are not.
- => How to find those that are important?

Amino acids: detecting sites with different evolutionary pattern

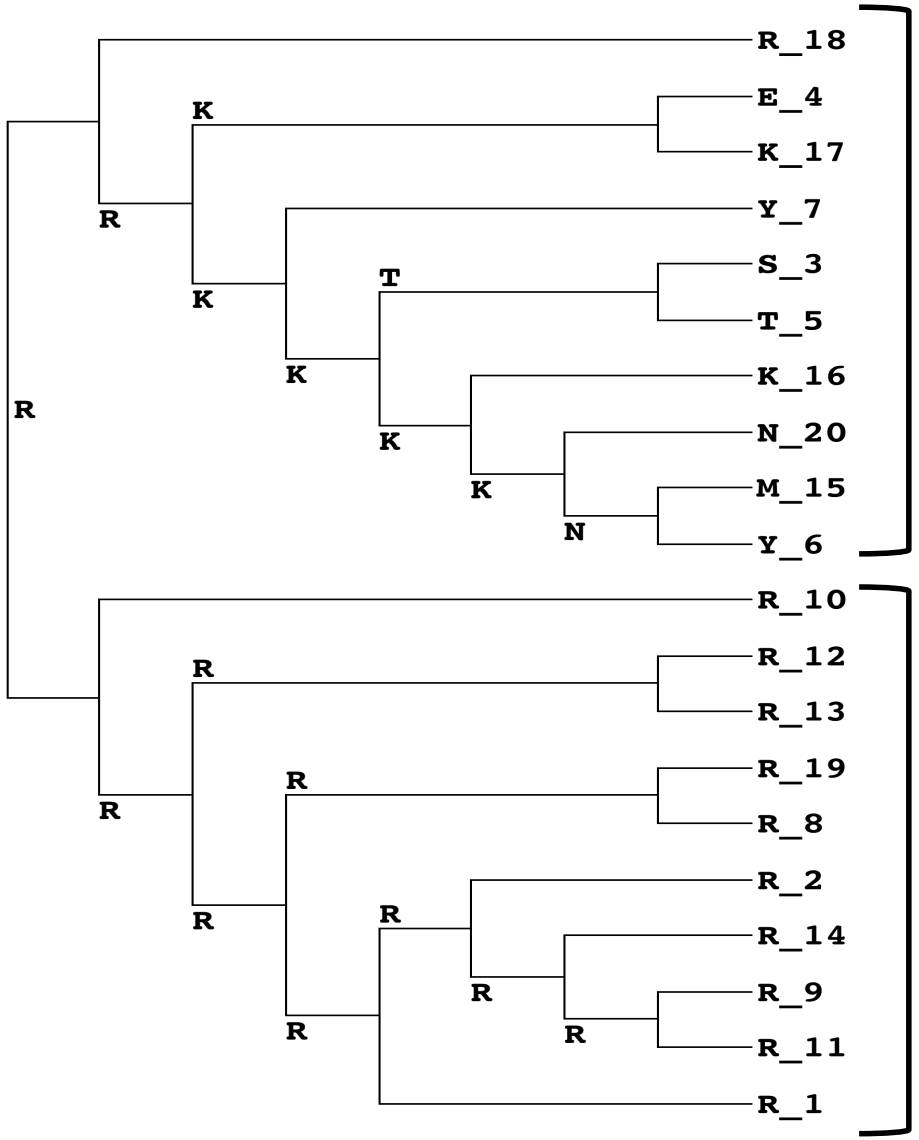


Amino acids: detecting sites with different evolutionary pattern



Strictly conserved
for Isoleucine

Amino acids: detecting sites with different evolutionary pattern

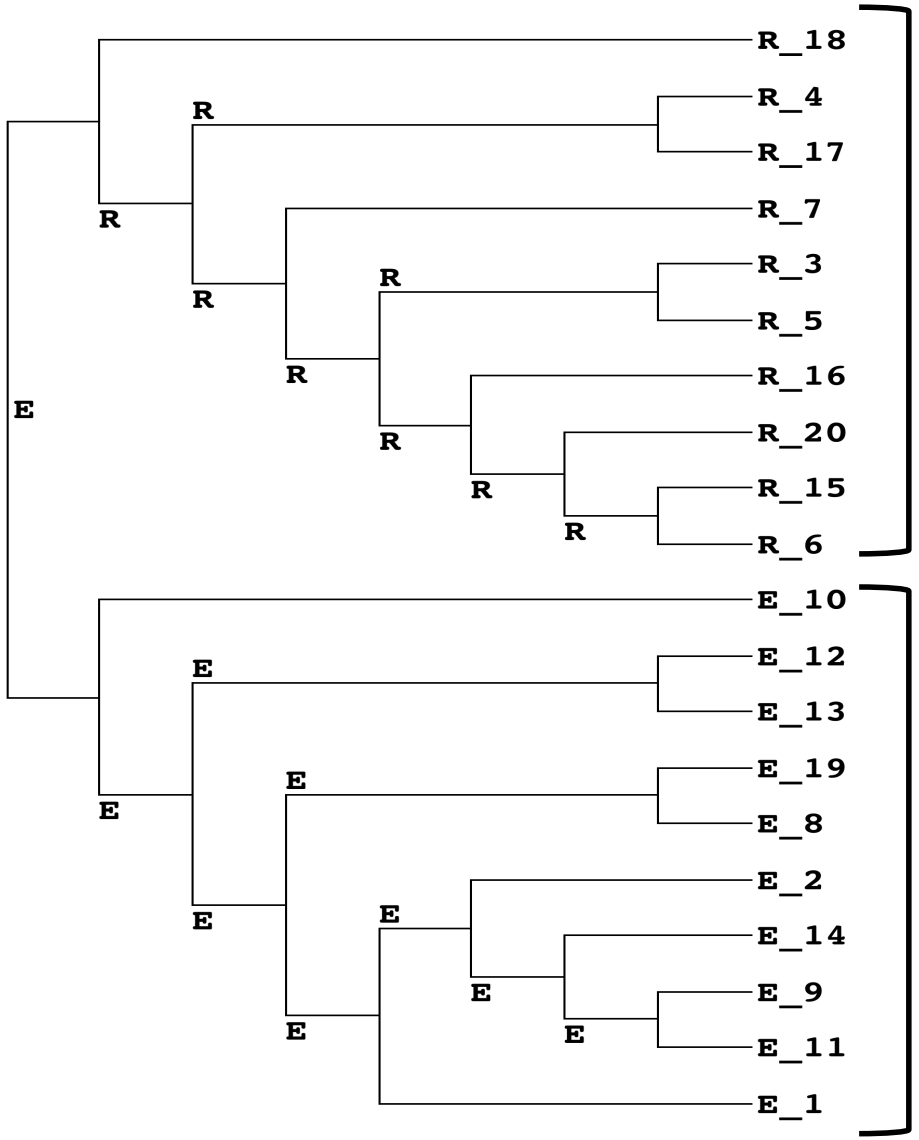


Highly divergent

- Type I of functional divergence
- Heterotachy
- Covarion-like

Highly conserved

Amino acids: detecting sites with different evolutionary pattern

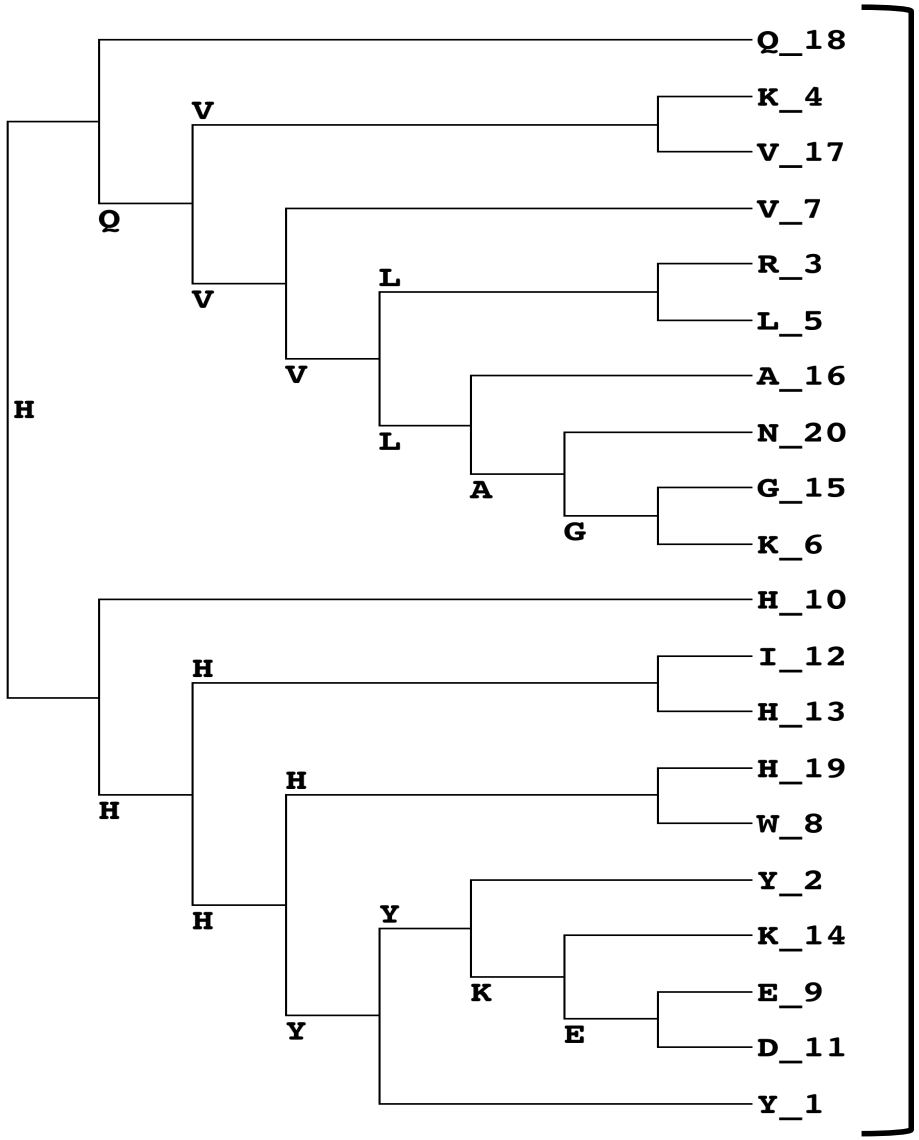


Conserved for basic residues

- Type II of functional divergence
- Constant-but-different

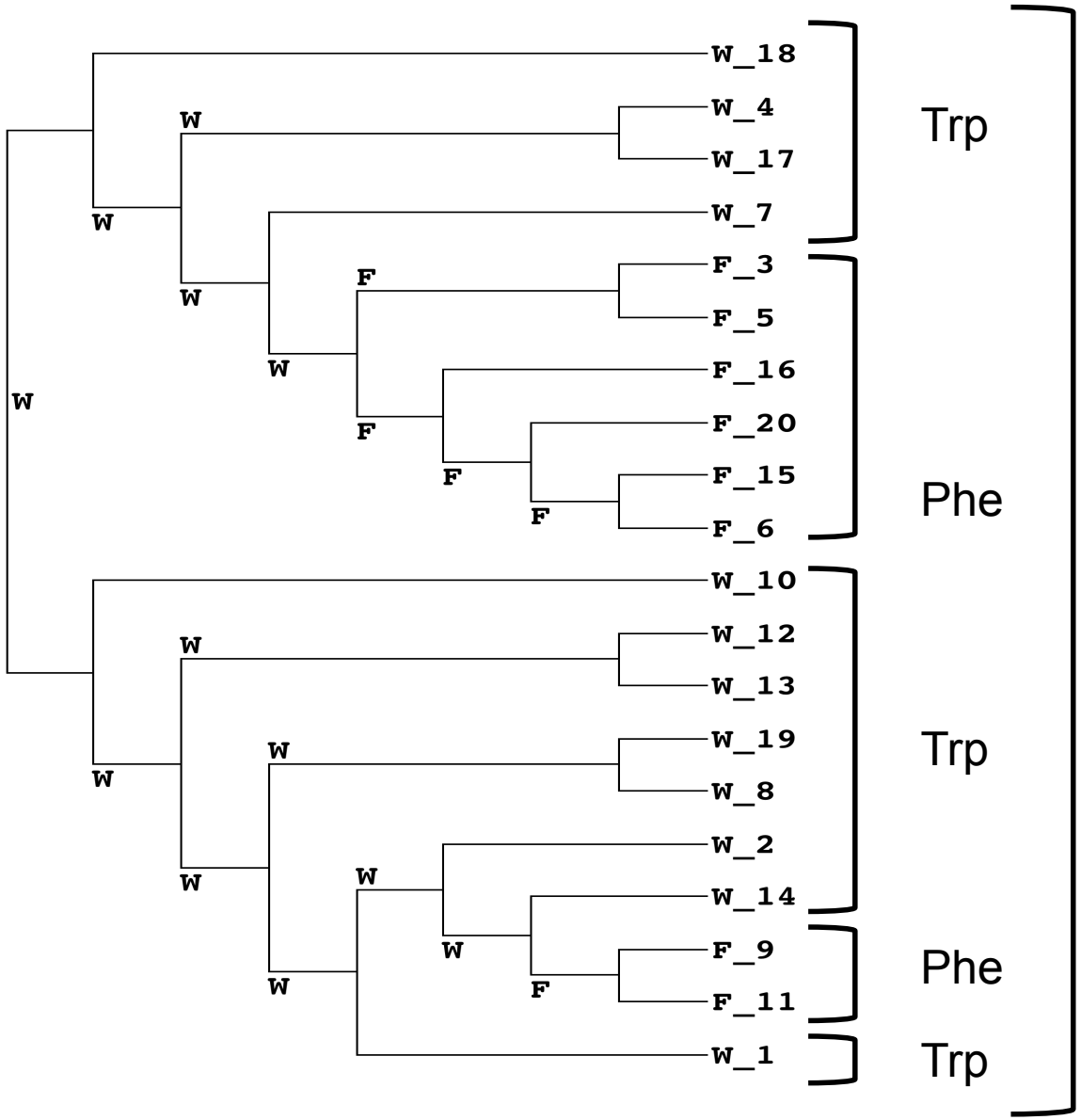
Conserved for acidic residues

Amino acids: detecting sites with different evolutionary pattern



Highly divergent

Amino acids: detecting sites with different evolutionary pattern



Conserved for big hydrophobic residues

Summary of changes

Group 1	Group 2	Type	Comment
Slow evolution (highly conserved)	Slow evolution (highly conserved)	Type 0	Conserved for structural and/or biochemical properties
Fast evolution (highly divergent)	Fast evolution (highly divergent)	Type 0	Either not important, or adapted in each species.
Slow evolution (highly conserved)	Fast evolution (highly divergent)	Type I	Change in evolutionary rate. Either a residue is recruited for a new function, or a position presents relaxation in selection pressures.
Fast evolution (highly divergent)	Slow evolution (highly conserved)	Type I	
Slow evolution (highly conserved)	Slow evolution, but another type (highly conserved)	Type II	A residue is recruited for a new function.

Fixation of mutations in the genome

- Have a deleterious effect on the fitness:
  - Not likely to bo fixed.

  => **Negative (purifying) selection** (Darwin theory)
- Have a positive effect on the fitness:
  - Likely to be fixed.

  => **Positive selection** (Darwin theory)
- Have no effect on the fitness:
  - May or may not be fixed (random process)

  ⇒ **Neutral evolution** (Kimura theory, early 70s)

Fixation of mutation in the genome

- Difficulty:

  Differentiate between

  NEUTRAL EVOLUTION (Most cases)

  and

  POSITIVE SELECTION (A few cases)

Statistical measure of selection / adaptation

- Null hypothesis:
  - All amino acids / nucleotides are fixed under neutral evolution.

- Alternative hypothesis:
  - A subset of amino acids / nucleotides are fixed under positive selection.

- Test: Do we have more amino acids under positive selection than expected?

Amino acids: detecting sites with different evolutionary pattern

- Various tools exist:
  - **DIVERGE** (Gu *et al.*, Bioinformatics 2002): Graphical User Interface. Works only in Windows.
  - **BADASP** (Edwards and Shields, Bioinformatics 2005): Easy to use. Doesn't provide stastitical measure.
  - **TDG09** (Tamuri et al. Plos Comp Biol 2009): Provide likelihood inference per site. Can be used on convergent data.
  - **FUNDI** (Gaston *et al.*, Bioinformatics. 2011): Mix Type I and Type II. Use phylogenetic tool to estimate parameters.

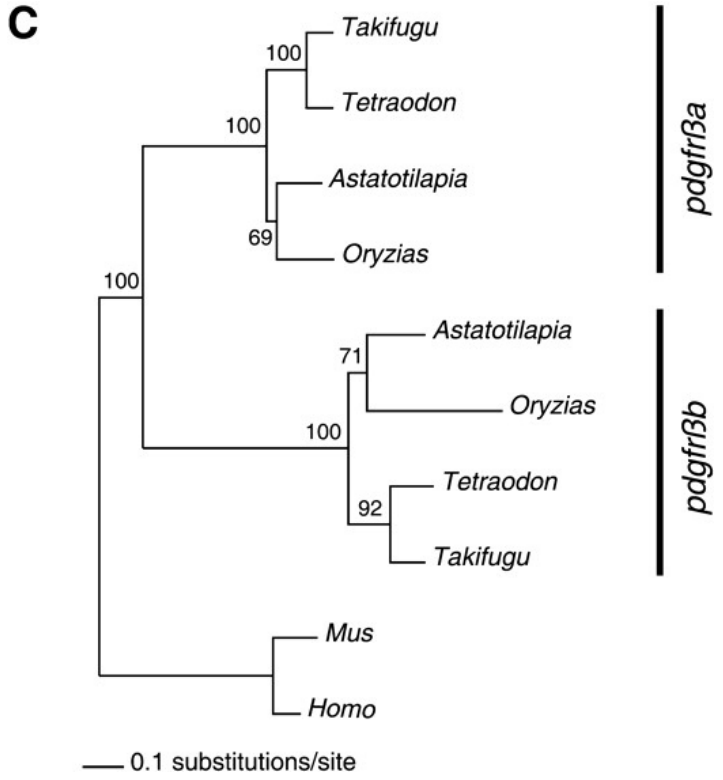
Other tools:

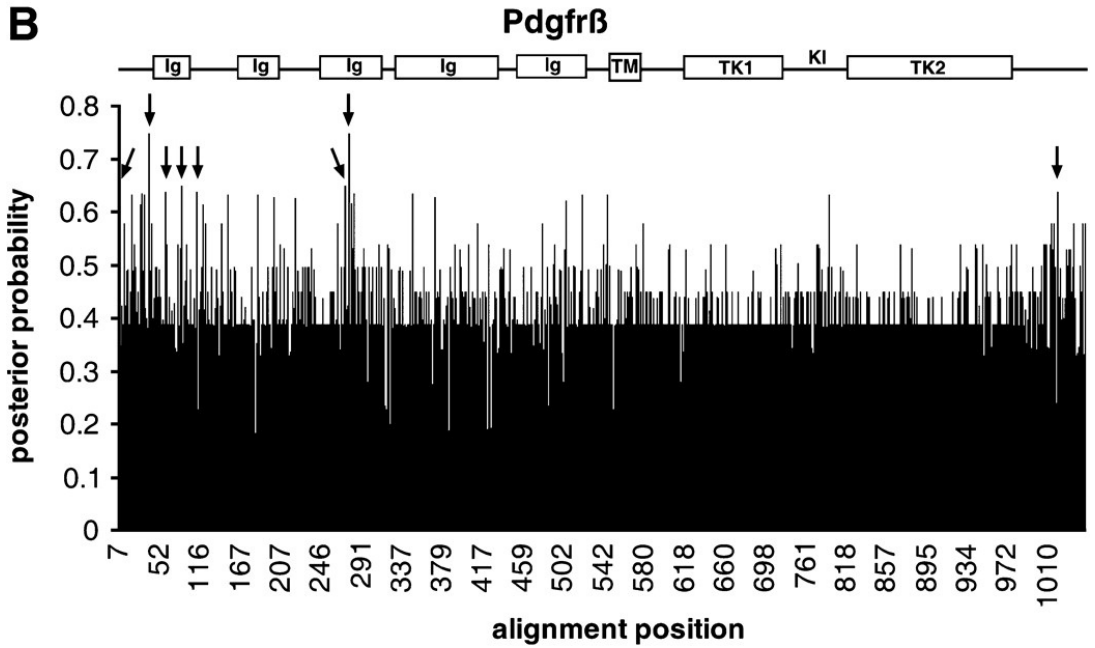
- SPEER (Chakrabarti *et al.,* Mol Biol. 2007)
- CHECKCOV (Pupko and Galtier, Proc Biol Sci. 2002)
- SHIFT-FINDER (Pontarotti's lab, unpublished)

Parameters to be estimated

- Alpha shape parameter of the gamma distribution
- Amino acid frequencies
- Branch size

Asymmetric Evolution in Two Fish-Specifically Duplicated Receptor Tyrosine Kinase Paralogons Involved in Teleost Coloration





DIVERGE, used to detect Type I of functional divergence

R
R
R
R
E
F
W
C

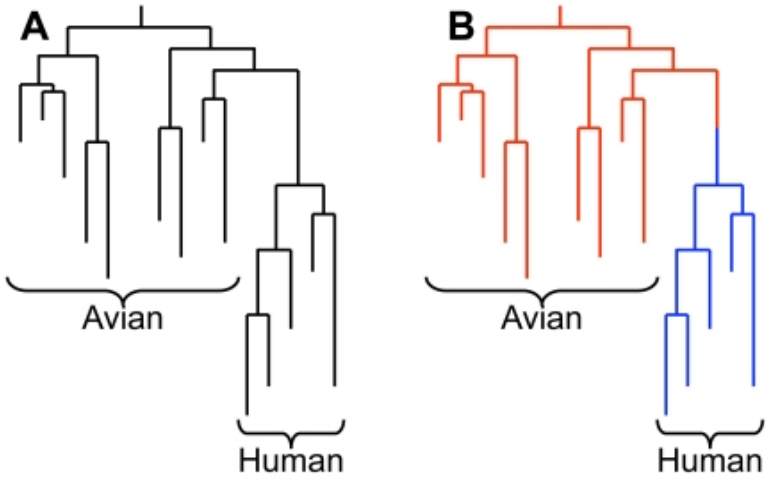
Braasch *et al*., Mol Biol Evol 2006

TDG09 Algorithm: Amino Acids

- Identifying changes in selective constraints: host shifts in influenza.
  Tamuri AU, Dos Reis M, Hay AJ, Goldstein RA.
  PLoS Comput Biol. 2009 Nov;5(11):e1000564.

- Sitewise non-homogeneous phylogenetic model that explicitly takes into account **differences in the equilibrium frequencies of amino acids** in different hosts and locations.

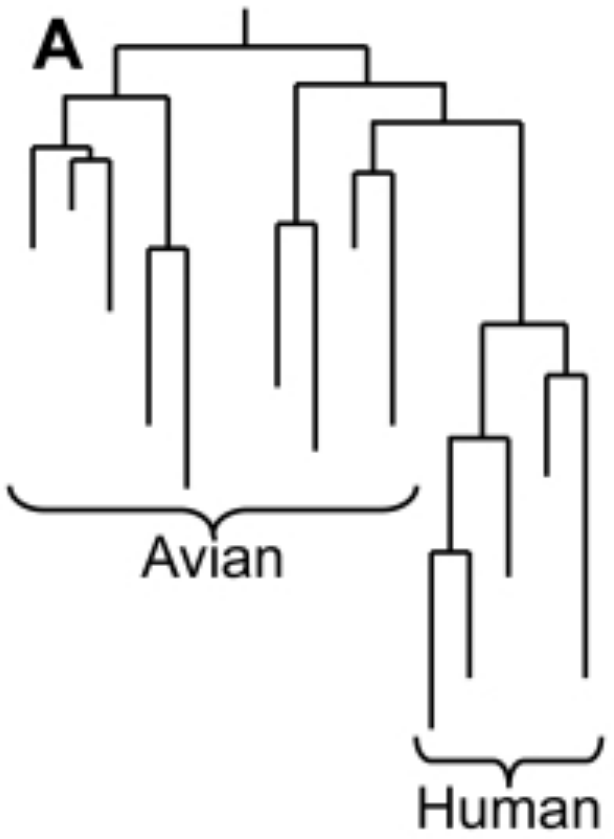
Homogeneous assumption



Non-homogeneous assumption

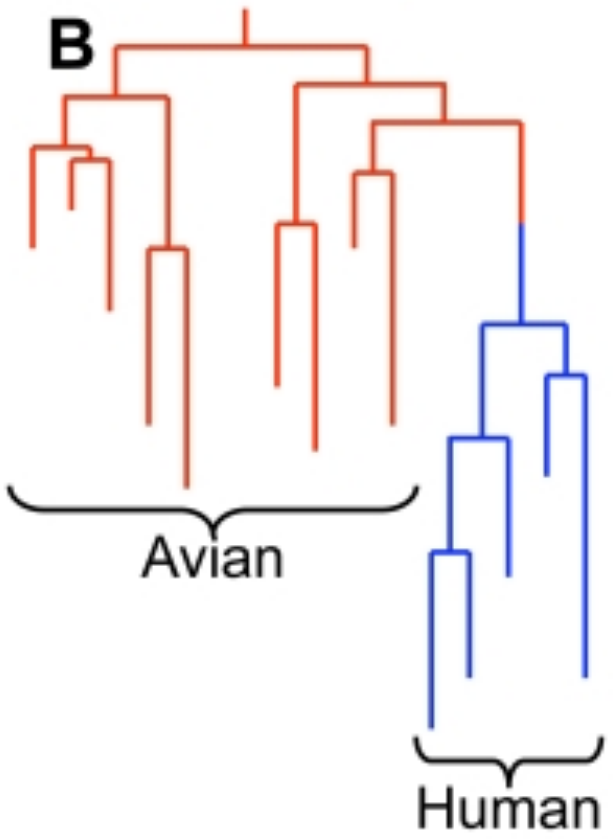
TDG09 Algorithm: Amino Acids

R0



Substitution rate
=
identical throughout the tree

R1



R2

Substitution rate
=
different throughout the tree

Practical:

http://beta.cathdb.info

⇒About

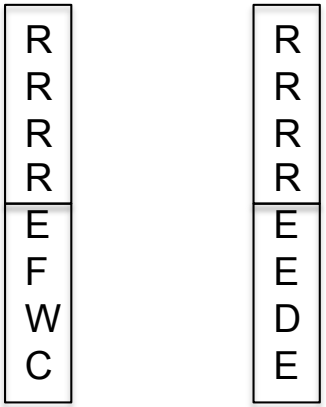
⇒Tutorials

⇒Detecting sites under functional divergence

⇒Part I on amino acids dataset

- Protein: nucleotide to amino acid
- Measuring positive selection using the dN/dS ratio.

- Type I
- Type II



- dN/dS helps to identify the constraints.

Protein: nucleotide to amino acid

Inverse table (compressed using IUPAC notation)

AA	Codons
Ala/A	GCU, GCC, GCA, GCG
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG
Asn/N	AAU, AAC
Asp/D	GAU, GAC
Cys/C	UGU, UGC
Gln/Q	CAA, CAG
Glu/E	GAA, GAG
Gly/G	GGU, GGC, GGA, GGG
His/H	CAU, CAC
Ile/I	AUU, AUC, AUA
START	AUG

AA	Codons
Leu/L	UUA, UUG, CUU, CUC, CUA, CUG
Lys/K	AAA, AAG
Met/M	AUG
Phe/F	UUU, UUC
Pro/P	CCU, CCC, CCA, CCG
Ser/S	UCU, UCC, UCA, UCG, AGU, AGC
Thr/T	ACU, ACC, ACA, ACG
Trp/W	UGG
Tyr/Y	UAU, UAC
Val/V	GUU, GUC, GUA, GUG
STOP	UAA, UGA, UAG

How to detect Positive Selection?

- Make use of the degenerate genetic code.

- Assume that dS (synonymous) substitutions are neutral.

- Assume that dN (non-synonymous) substitution are either neutral (fixed under random drift) or fixed under positive selection.

How to detect Positive Selection?

Species 1	A	L	P	H	Y
	GCC	CTT	CCT	CAT	TAT
Species 2	A	R	P	H	Y
	GCC	CGT	CCT	CAT	TAC

How to detect Positive Selection?

Species 1	A	L	P	H	Y
	GCC	CTT	CCT	CAT	TAT
Species 2	A	R	P	H	Y
	GCC	CGT	CCT	CAT	TAC

15 nucleotides: 11 non-synonymous sites and 4 synonymous sites
1 synonymous substitutions (S) and 1 non-synonymous substitution (N)

$$dN = \frac{\text{number of non-synonymous substitutions}}{\text{non-synonymous sites}} = 1 / 11 = 0.09$$

$$dS = \frac{\text{number of synonymous substitutions}}{\text{synonymous sites}} = 1 / 4 = 0.25$$

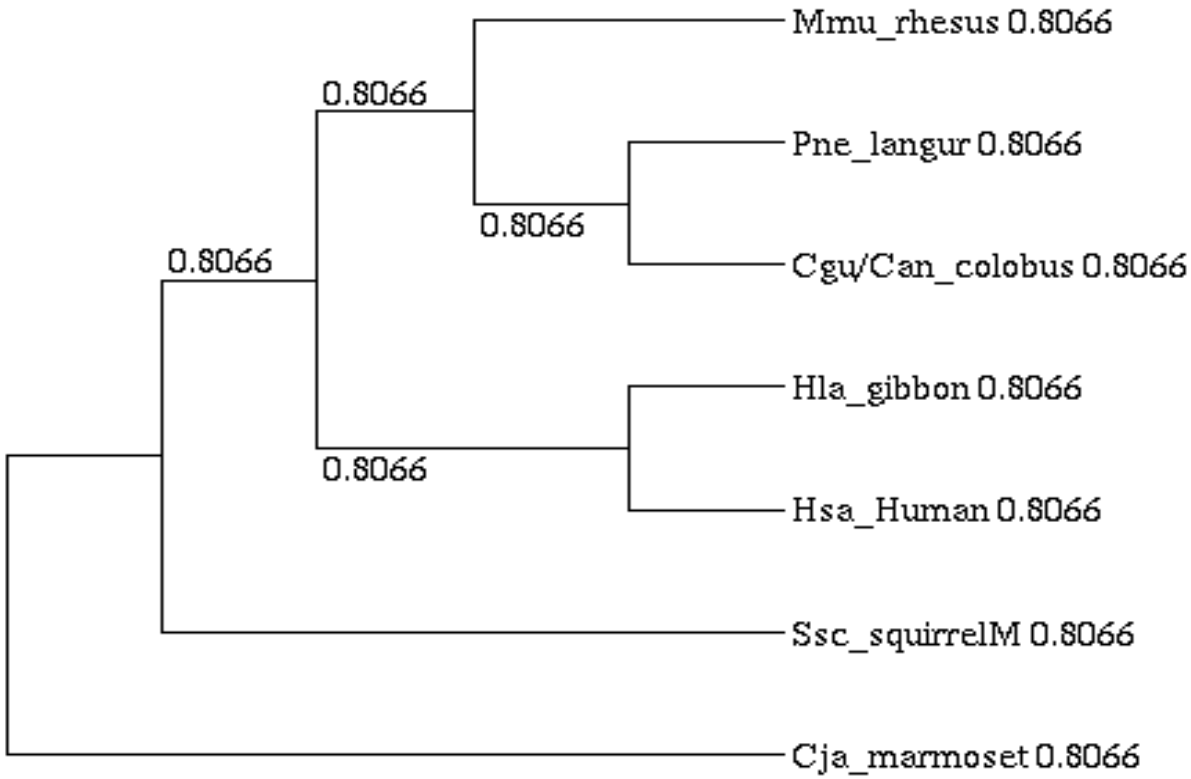
$dN/dS\ (\omega) = 0.09 / 0.25 = 0.36$

$dN/dS < 1$ -> purifying selection
$dN/dS = 1$ -> neutral evolution
$dN/dS > 1$ -> positive selection !!!!!!!!!

Branch models

(Yang 1998; Yang and Nielsen 1998)

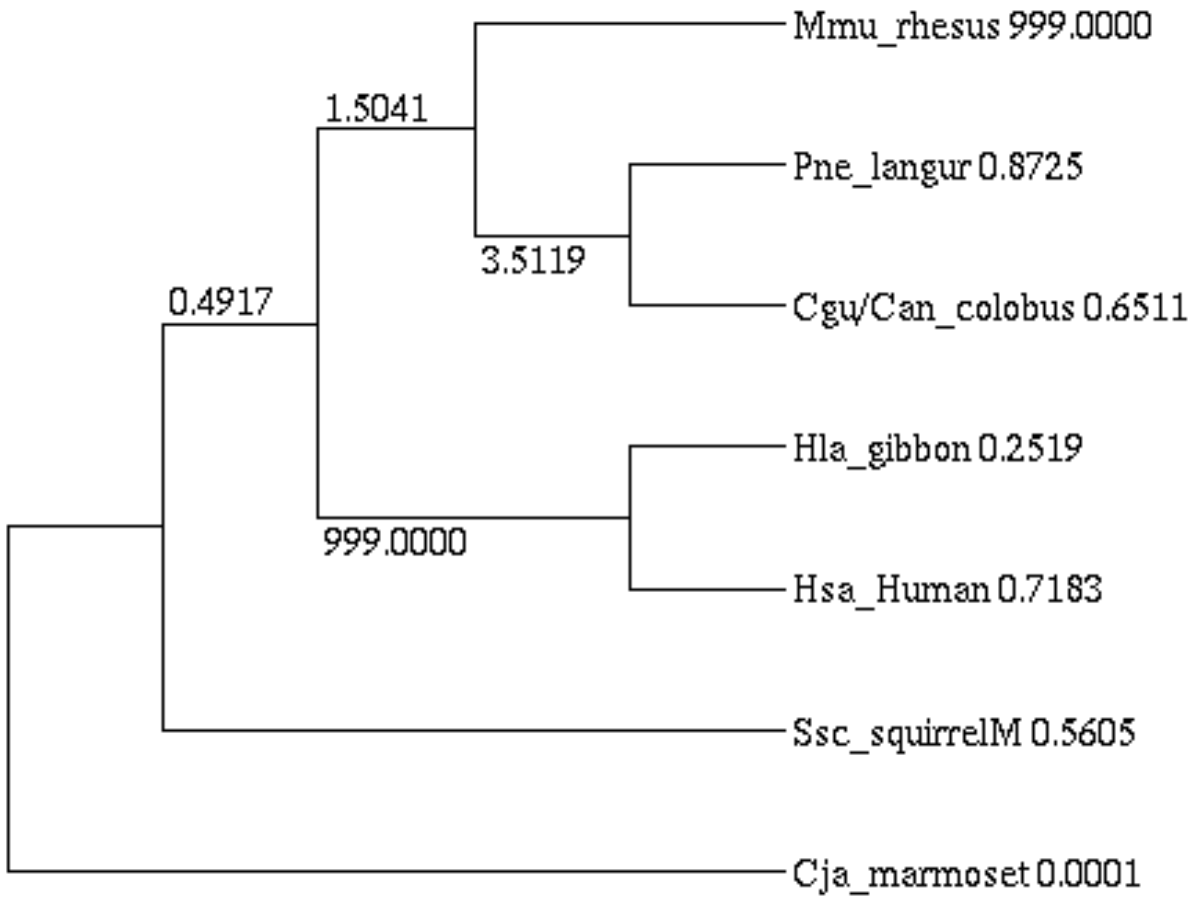
- One ratio model: estimate different dN and different dS, but keep the same ratio on all branches: null model
- All branches on the tree are under the same degree of selective pressure (i.e. same constraints).



Branch models

(Yang 1998; Yang and Nielsen 1998)

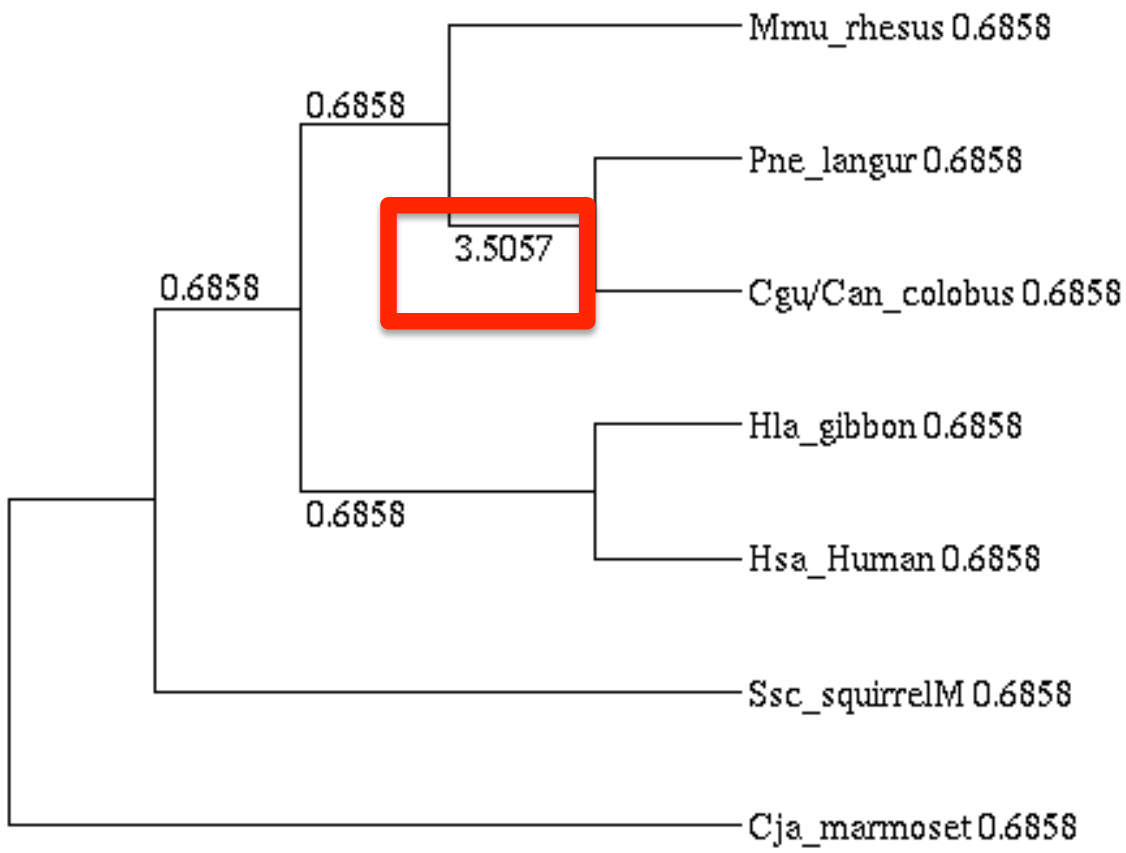
- Estimate all dN/dS ratio for <u>each branch</u> on the phylogeny



Branch models

(Yang 1998; Yang and Nielsen 1998)

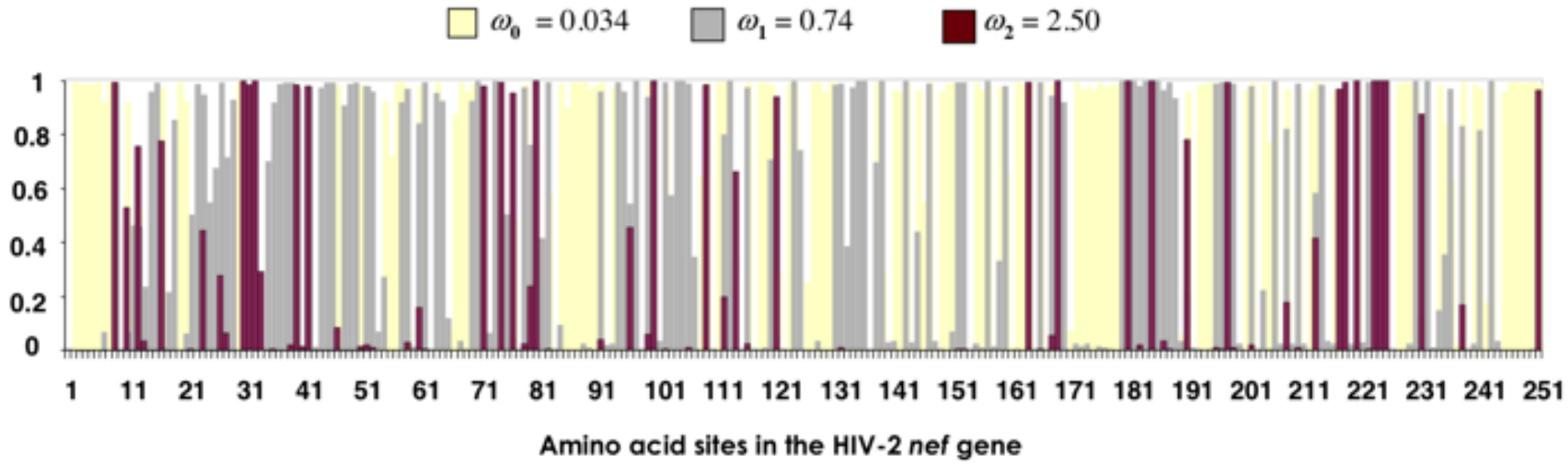
- Estimate different dN/dS ratio on one or more branches



Site models

(Nielsen and Yang 1998; Yang et al. 2000, Yang et al. 2005)

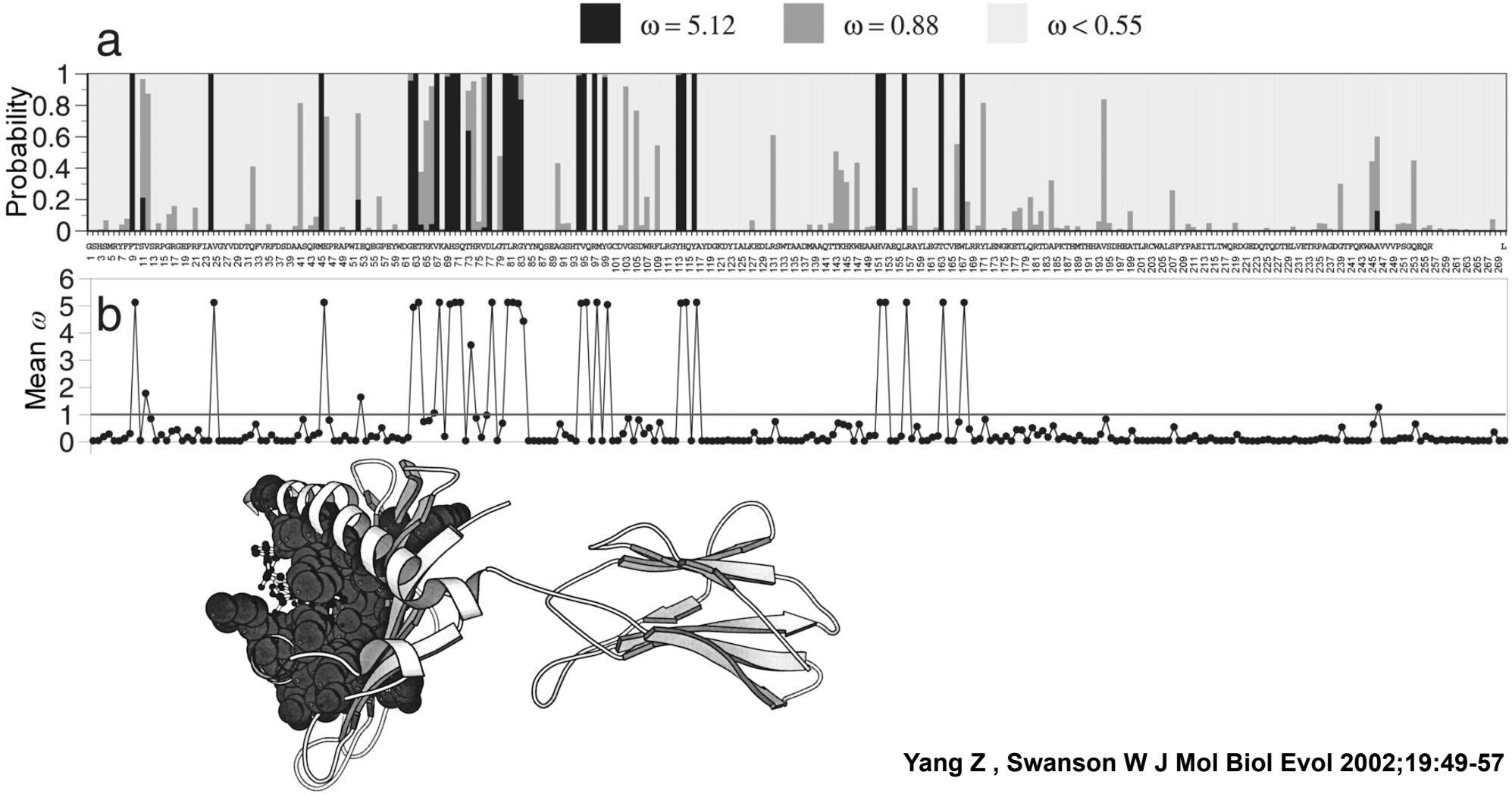
- Estimate all dN/dS (w) at each site (codons) into three categories (or more):
  ω0 (yellow) = negative selection
  ω1 (grey) = neutral evolution (or nearly-neutral)
  ω2 (rouge) = positive selection



Site models

(Nielsen and Yang 1998; Yang et al. 2000, Yang et al. 2005)

Posterior probabilities of site classes for sites along the MHC class I gene under the random-sites model M8



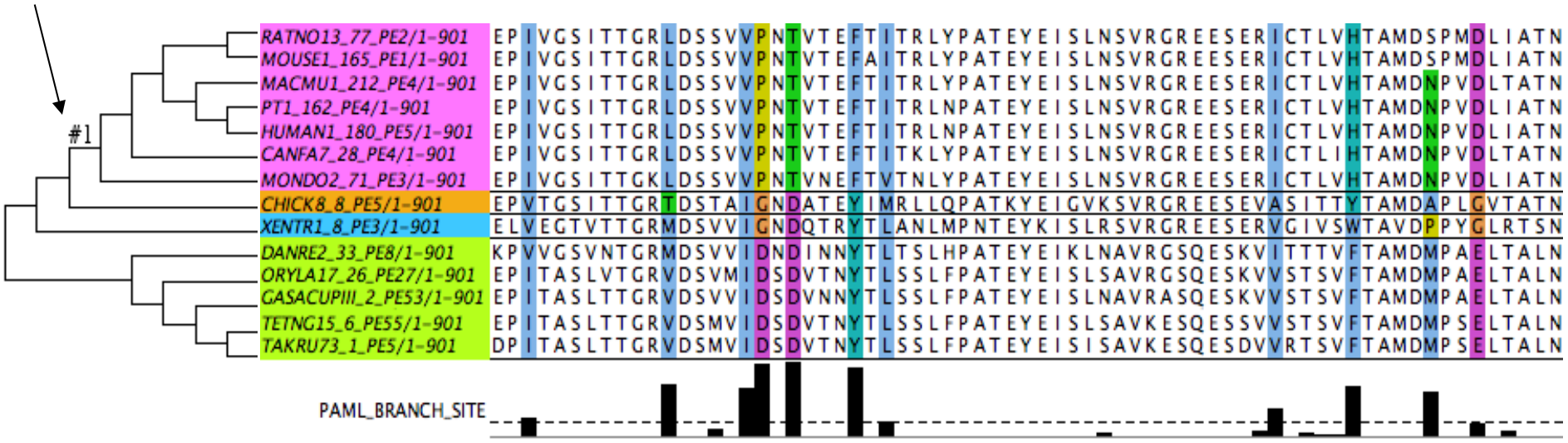
Yang Z , Swanson W J Mol Biol Evol 2002;19:49-57

Branch-site models

(Yang and Nielsen 2002, Zhang *et al*. 2005)

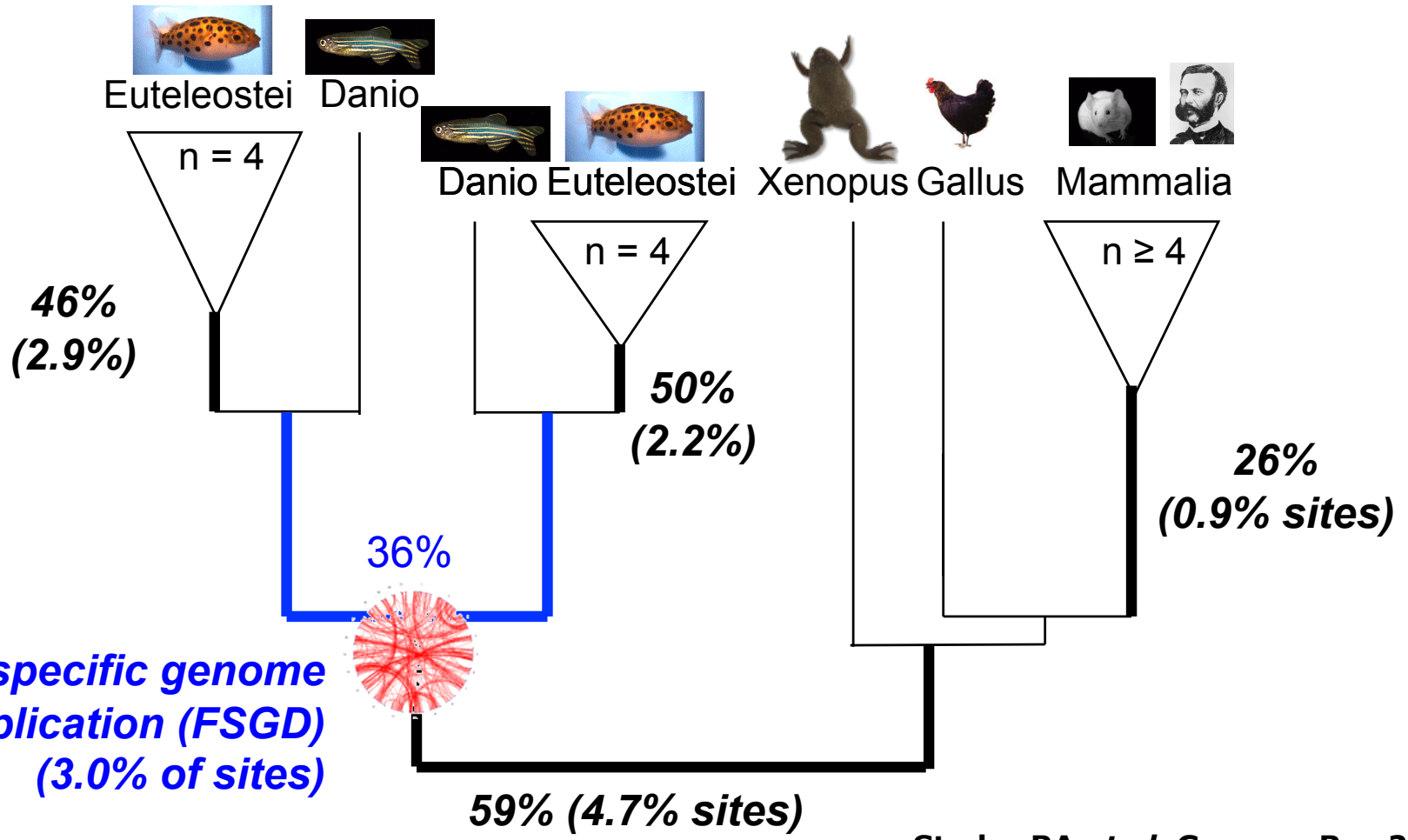
- Detect positive selection that affects only a **few sites** on pre-specified lineages.
- ***foreground* branches** = branches under test for positive selection.
- ***background* branches** = all other branches**.**
- LRT: branch-site model A is the alternative model (**ω2 > 1**), while the simpler null model is **model A but with ω2 = 1 fixed**.

Branch leading
to Mammals



We can estimate ancient dN/dS.

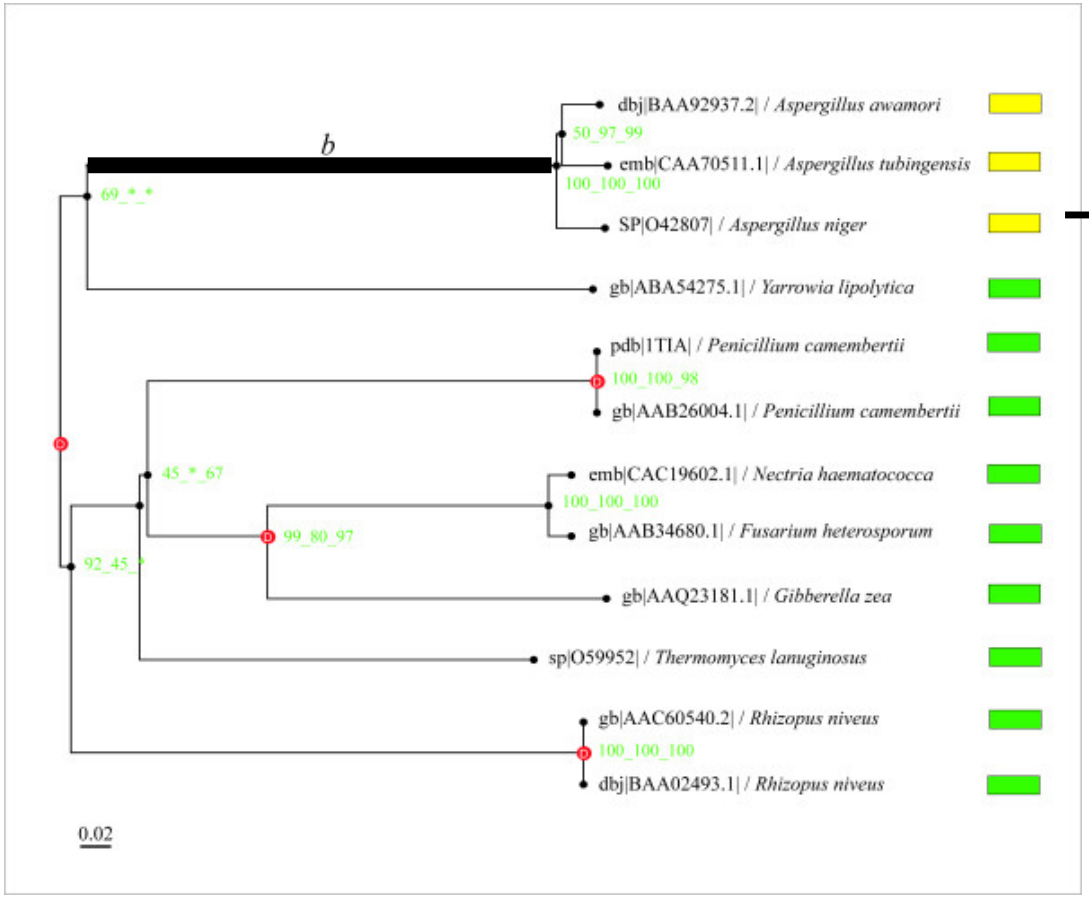
- Analyses done on 884 genes families (2’673 branches).
- Percentages of branches with sites under positive selection:



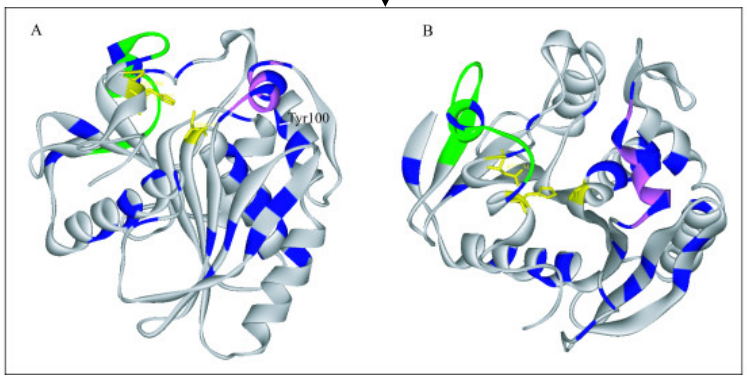
Studer RA *et al.* Genome Res 2008

Example of positive selection in enzyme

Lipase / feruloyl esterase A



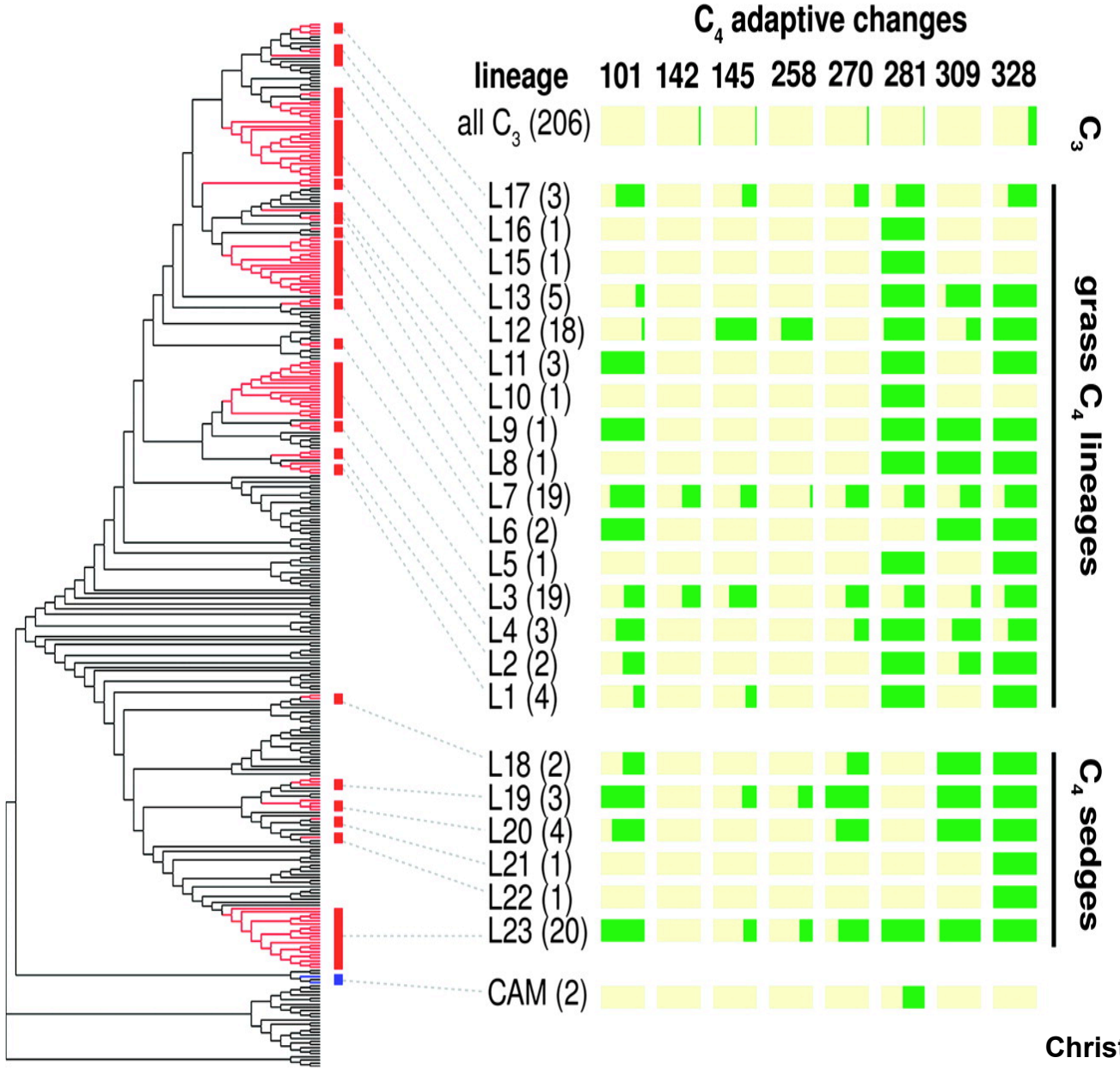
SITES IN 3D???



Positively selected sites in blue

Levasseur A *et al.* BMC Evol Biol 2006

Positive selection detected in RubisCO



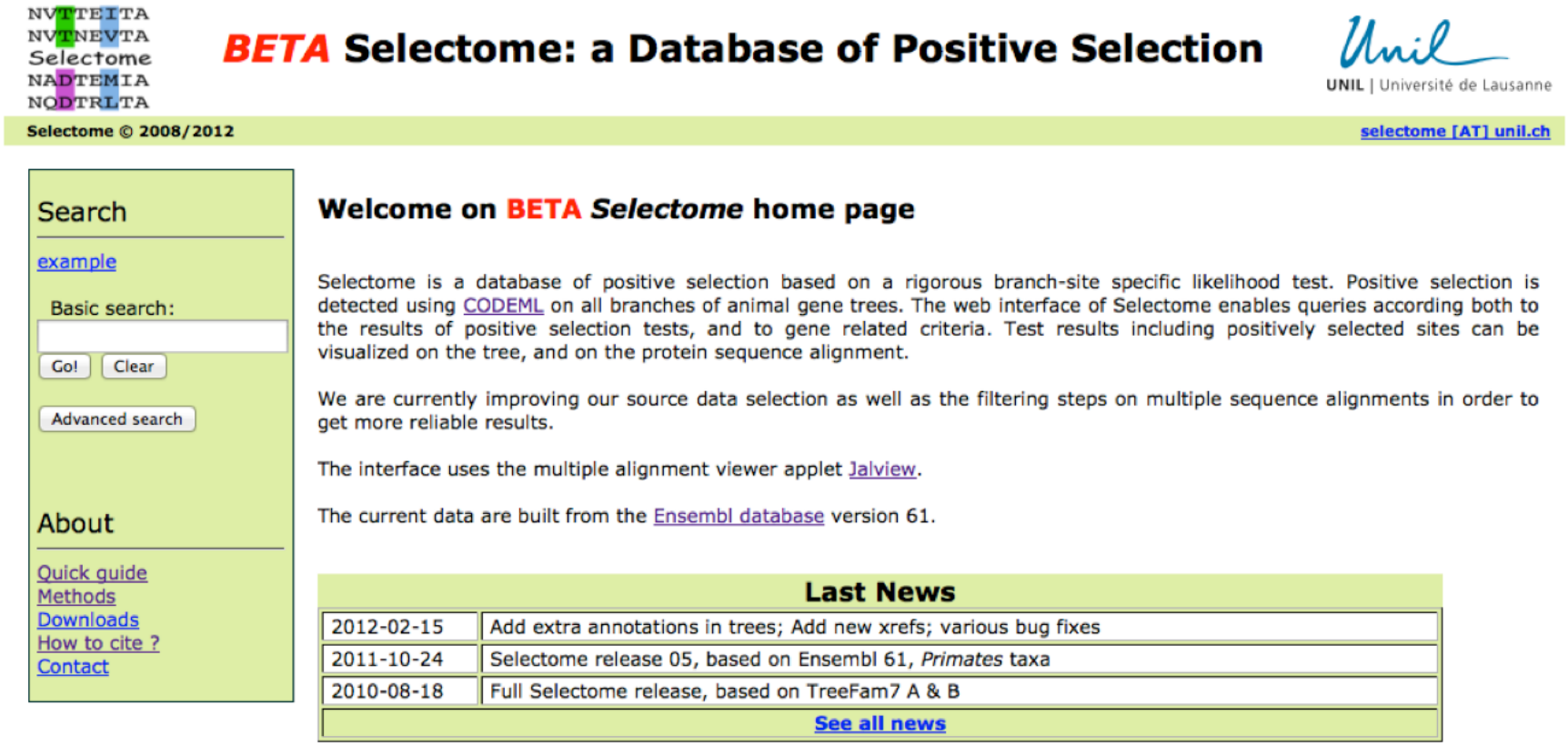
C4 HAS A HIGHER TURN-OVER THAN C3.

Christin PA *et al.* Mol Biol Evol 2008

Databases of positive selection

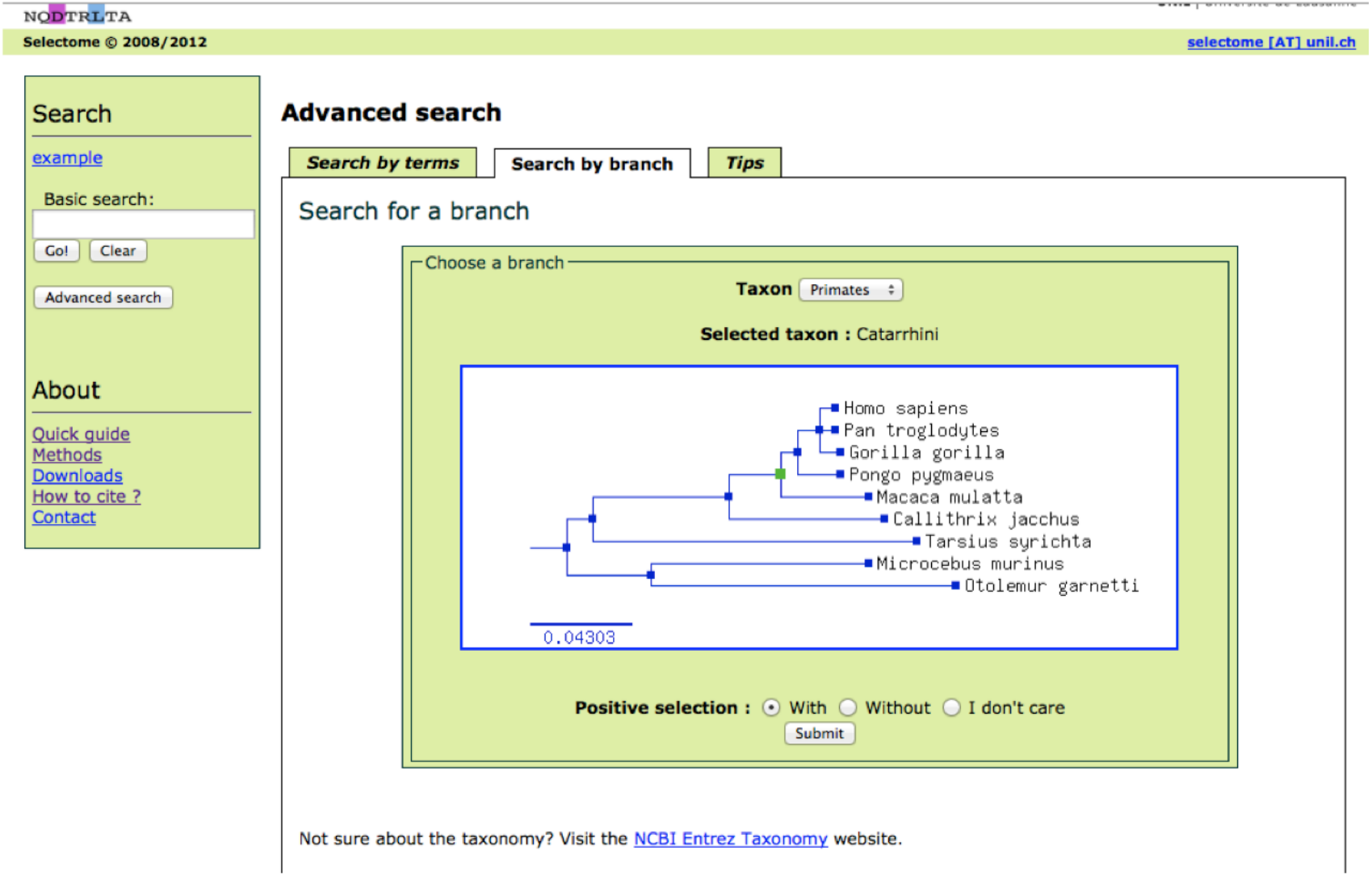
Selectome: database of positive selection

http://selectome.unil.ch/



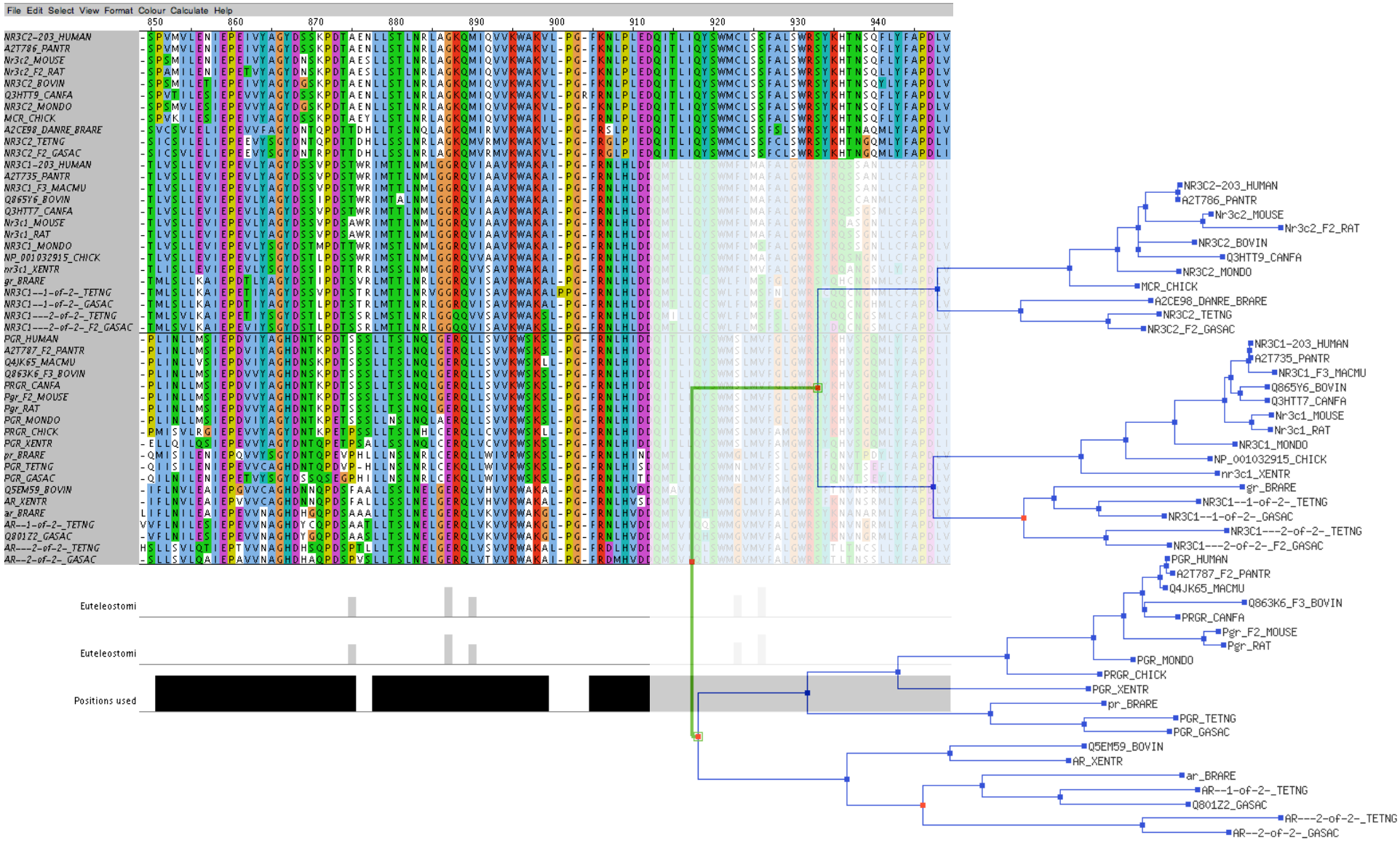
Selectome: database of positive selection

http://selectome.unil.ch/



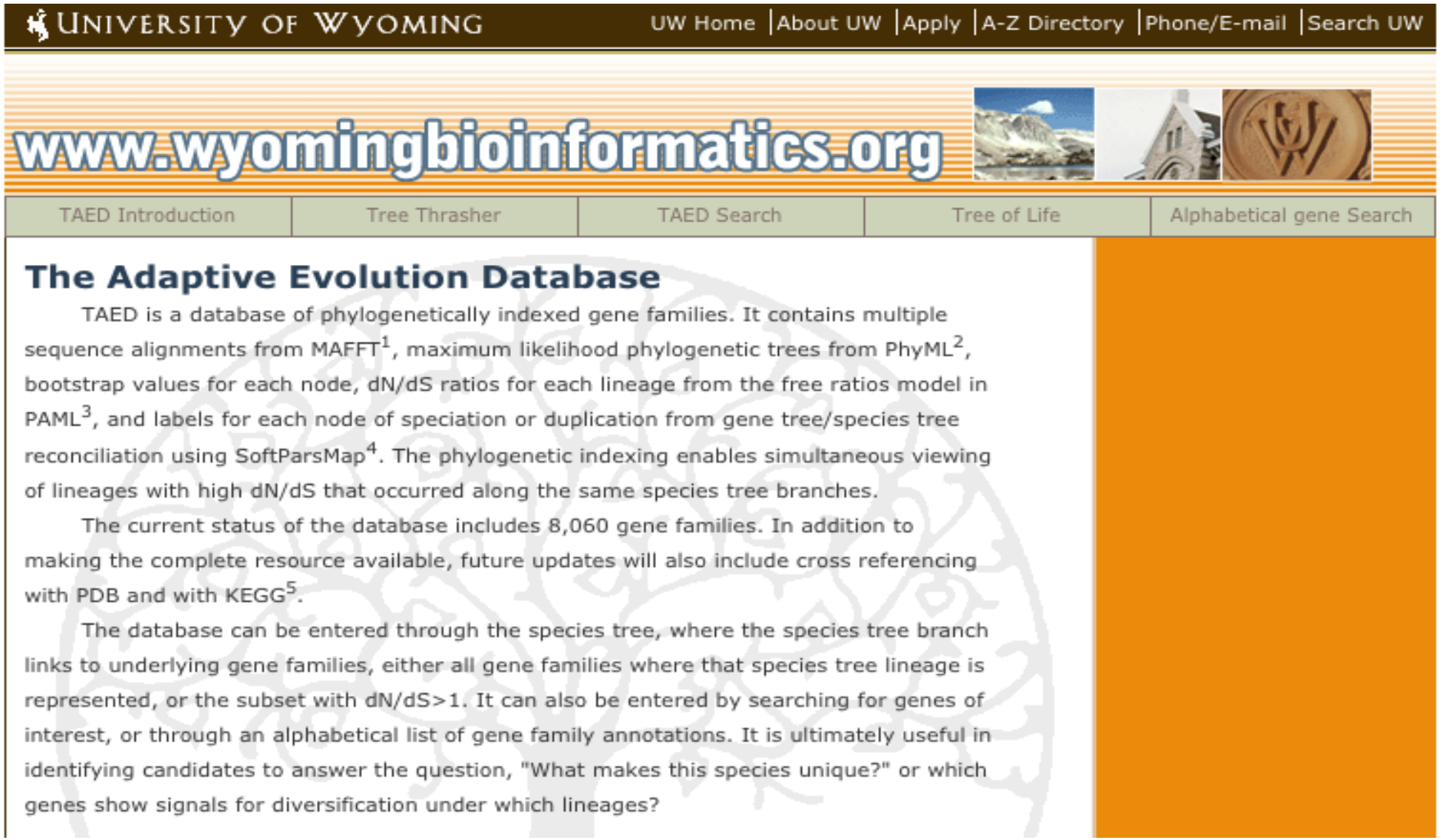
Selectome: database of positive selection

http://selectome.unil.ch/



The Adaptive Evolution Database (TAED)

http://www.wyomingbioinformatics.org/TAED/



UNIVERSITY OF WYOMING | UW Home | About UW | Apply | A-Z Directory | Phone/E-mail | Search UW

www.wyomingbioinformatics.org

TAED Introduction | Tree Thrasher | TAED Search | Tree of Life | Alphabetical gene Search

The Adaptive Evolution Database

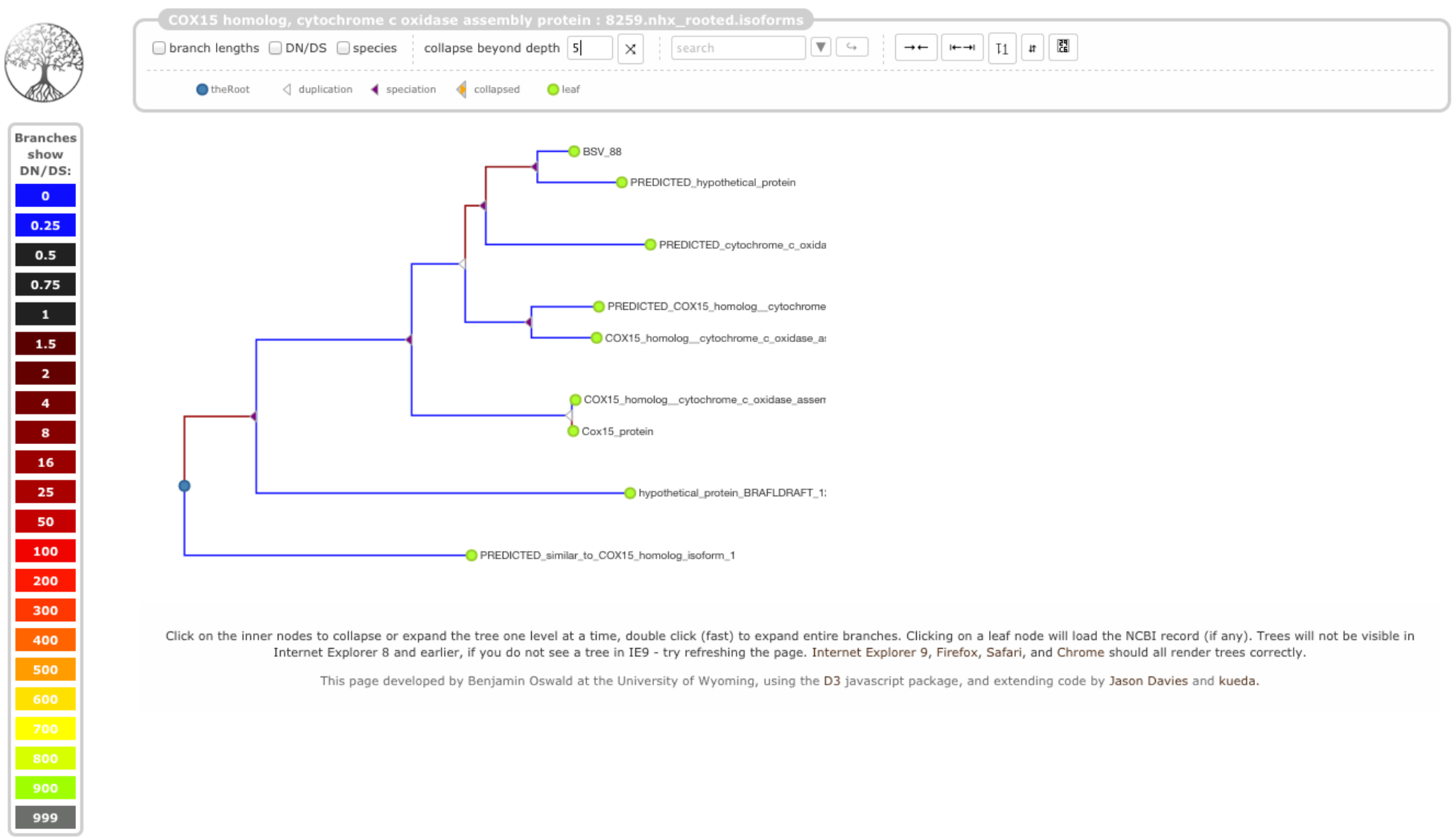
TAED is a database of phylogenetically indexed gene families. It contains multiple sequence alignments from MAFFT[1], maximum likelihood phylogenetic trees from PhyML[2], bootstrap values for each node, dN/dS ratios for each lineage from the free ratios model in PAML[3], and labels for each node of speciation or duplication from gene tree/species tree reconciliation using SoftParsMap[4]. The phylogenetic indexing enables simultaneous viewing of lineages with high dN/dS that occurred along the same species tree branches.

The current status of the database includes 8,060 gene families. In addition to making the complete resource available, future updates will also include cross referencing with PDB and with KEGG[5].

The database can be entered through the species tree, where the species tree branch links to underlying gene families, either all gene families where that species tree lineage is represented, or the subset with dN/dS>1. It can also be entered by searching for genes of interest, or through an alphabetical list of gene family annotations. It is ultimately useful in identifying candidates to answer the question, "What makes this species unique?" or which genes show signals for diversification under which lineages?

The Adaptive Evolution Database (TAED)

http://www.wyomingbioinformatics.org/TAED/



Some advices

- Garbage in / garbage out

  => Sequences of good quality !
- Multiple alignment with best methods
- Remove badly aligned blocks

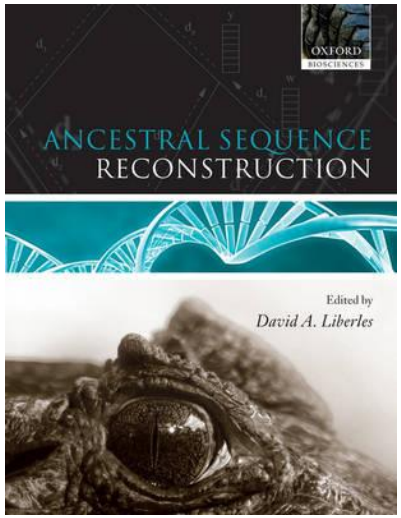
- Select most appropriate model of substitution (ie WAG or LG).
- Build trees with Maximum Likelihood / Bayesian.

Some advices

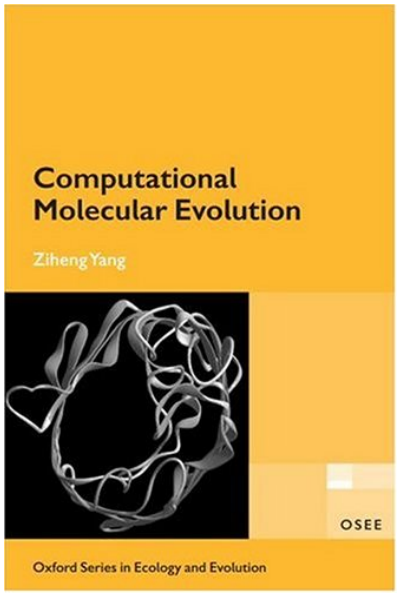
- Prior biological hypothesis:
  ⇒ Apply test on the branch of interest.
- No prior hypothesis:
  ⇒ Apply tests on different branches.
  ⇒ Apply a False Discovery Rate (FDR) *a posteriori*.
- Always visualise your result!

Recommended books

- Ancestral Sequence Reconstruction, by David A. Liberles



- Computational Molecular Evolution, by Ziheng Yang



Practical:

http://beta.cathdb.info

⇒About

⇒Tutorials

⇒Detecting sites under functional divergence

⇒Part II on nucleotides dataset