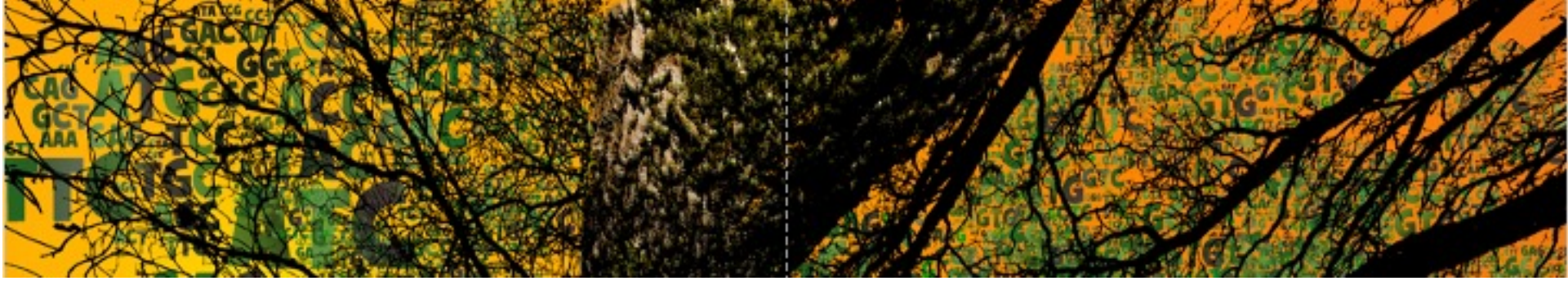


Natural selection and codon models

Maria Anisimova

Institute of Computational Life Sciences
Zurich University of Applied Sciences - ZHAW

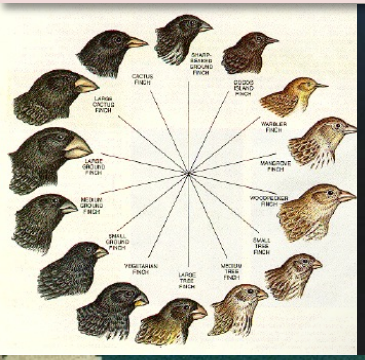
Life Sciences and Facility Management

Institute of Computational Life Sciences

11.05.2022, Lund

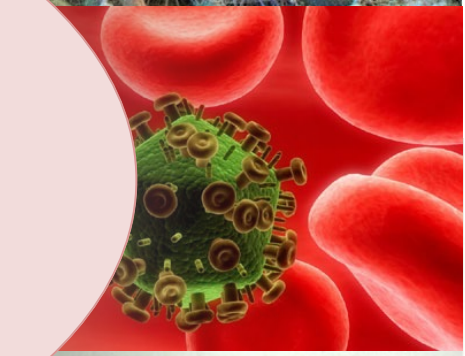
Swiss Institute of Bioinformatics

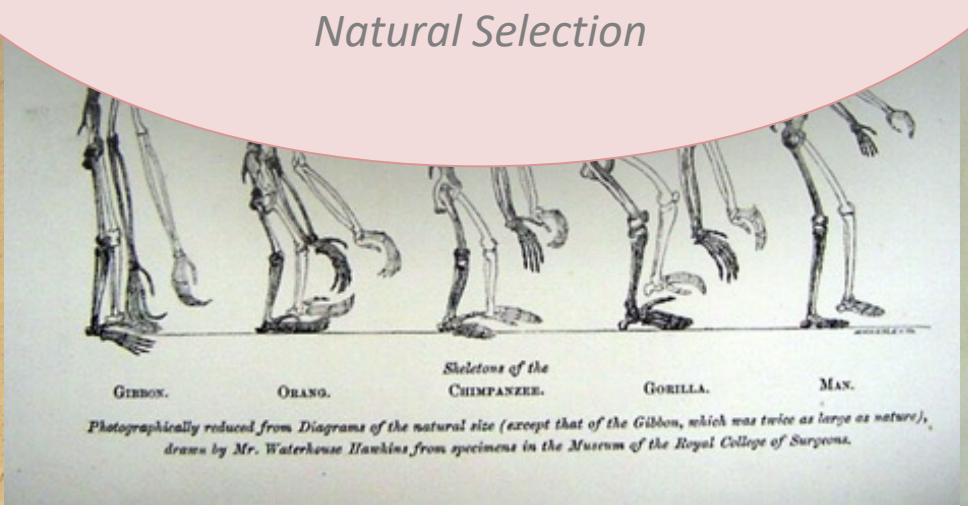
Why study natural selection

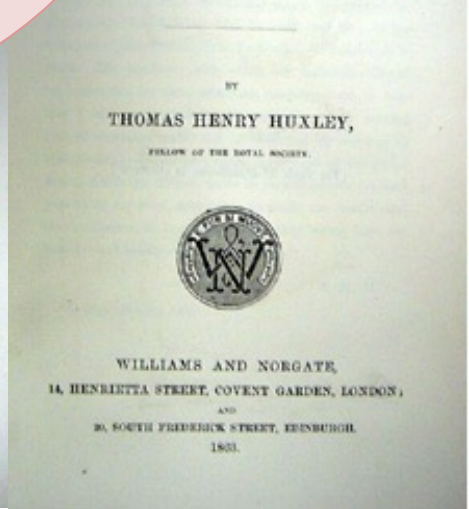


… from so simple a beginning
endless forms most beautiful
and most wonderful have been,
and are being evolved.

Charles Darwin, 1859
On the Origin of Species by Means of Natural Selection







Big question of biology

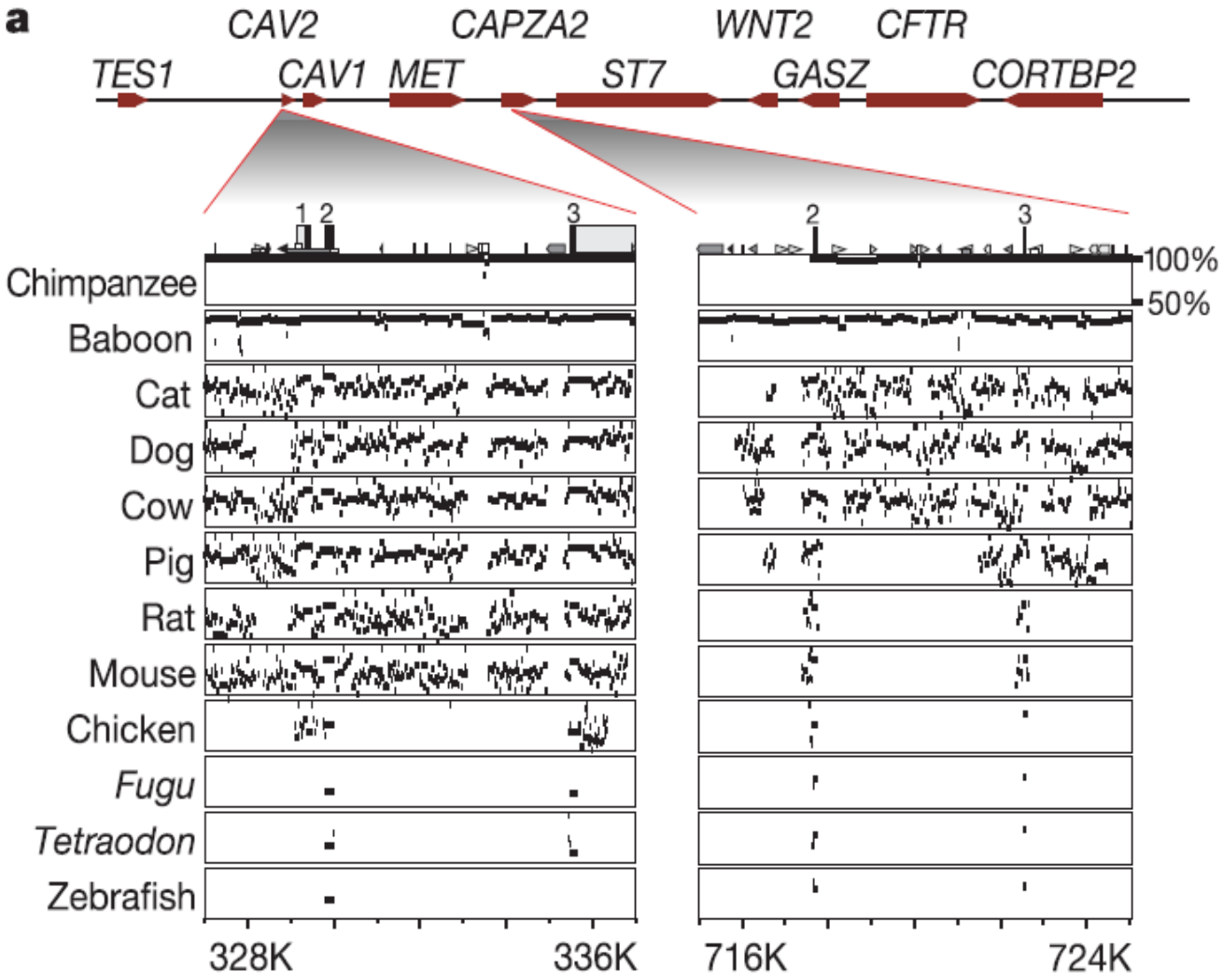
How does genotype…

… shape

Conservation

↓

function



Percentage identity when human is aligned with another species.
Close species are effective in identifying regulatory elements while distant species are effective in identifying coding regions.

Thomas, et al. 2003. *Nature*

High variability may also mean function: the variability may be driven by selection

Evolutionary biologists are more interested in positive selection because fixations of advantageous mutations in the genes or genomes are responsible for evolutionary innovations and species divergences.

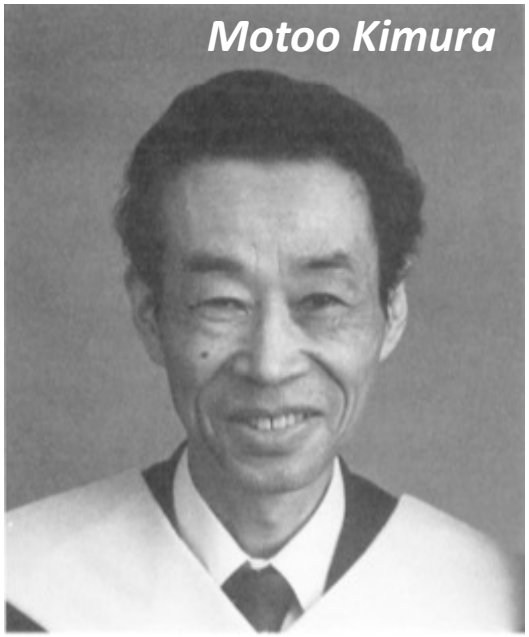
There are two main explanations for genetic variation observed within a population or between species:

Natural selection (survival of the fittest)
Mutation and drift (survival of the luckiest)

Gillespie, J.H. 1998. *Population genetics: a concise guide*. John Hopkins University Press, Baltimore.

Hartl, D.L., and A.G. Clark. 1997. *Principles of population genetics*. Sinauer Associates, Sunderland, Massachusetts.

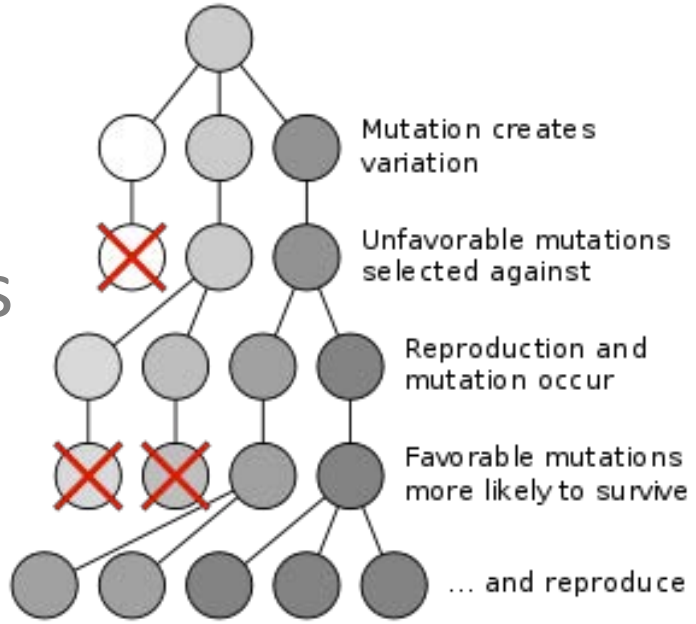
The neutral theory of molecular evolution



- Most mutations are deleterious
- Most changes: random fixation of neutral mutations
- The fate of alleles is determined by random genetic drift
- Substitution rate = neutral mutation rate (molecular clock)
- Selection may operate; but is too weak to influence
- Substitution = polymorphism
- Morphological traits evolve by natural selection

The neo-Darwinian theory of evolution

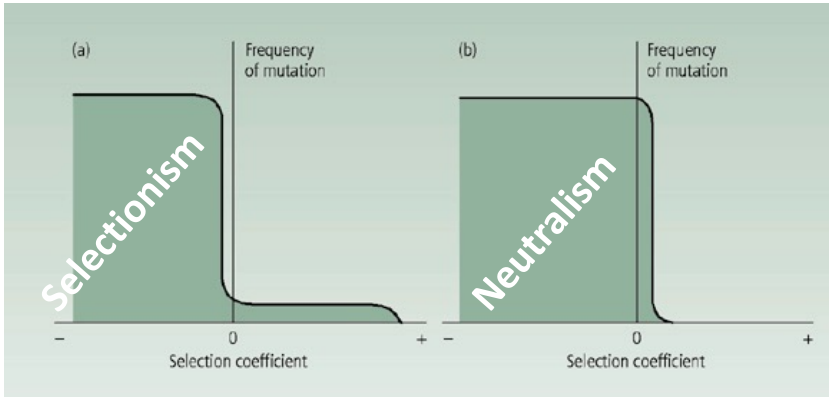
- Natural selection shapes the genetic makeup
- Most mutations are *deleterious,* removed by purifying selection
- Substitutions ≠ polymorphisms
- Substitutions are acquired by *positive selection*
- Polymorphisms are kept by *balancing selection*



The impact of the neutral theory

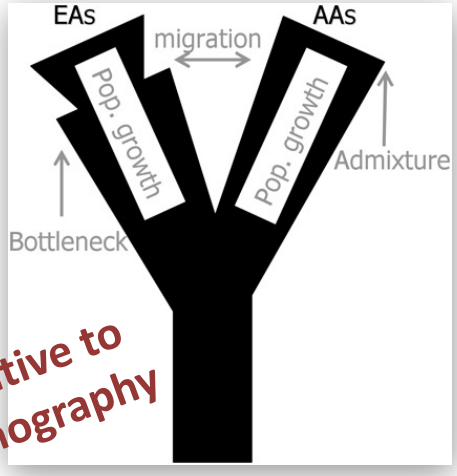
- Strengthened the connection between molecular biology and population genetics
- The neutral theory makes simple and testable predictions about what we should observe: provided *a falsifiable null hypothesis*
- Availability of such null hypothesis prompted the development of neutrality tests

***s* = selection coefficient**:
relative fitness of
mutant *a* vs. wild-type *A*.
Genotype fitness:
1 for *AA*, 1+*s* for *Aa*, 1+2*s* for *aa*
s > 0 positive selection
s < 0 negative selection



From Ridley (1996) Evolution

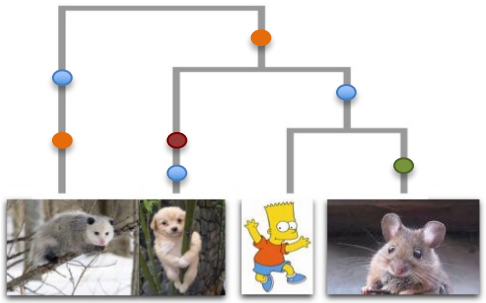
Neutrality and selection tests



- Mutational frequency spectrum
  (eg, Tajima's D, Tajima 1989)
- Population subdivision
- LD & haplotype structure
- Within/between species variability
  (HKA test, Hudson, Kreitman, Aguade 1987)

Account for codon structure:

- Within/between species variability
  (MK test, McDonald-Kreitman 1991)
- **Based on codon models**



Standard genetic code

The genetic code determines how random changes to the gene brought about by the process of mutation will impact the function of the encoded protein

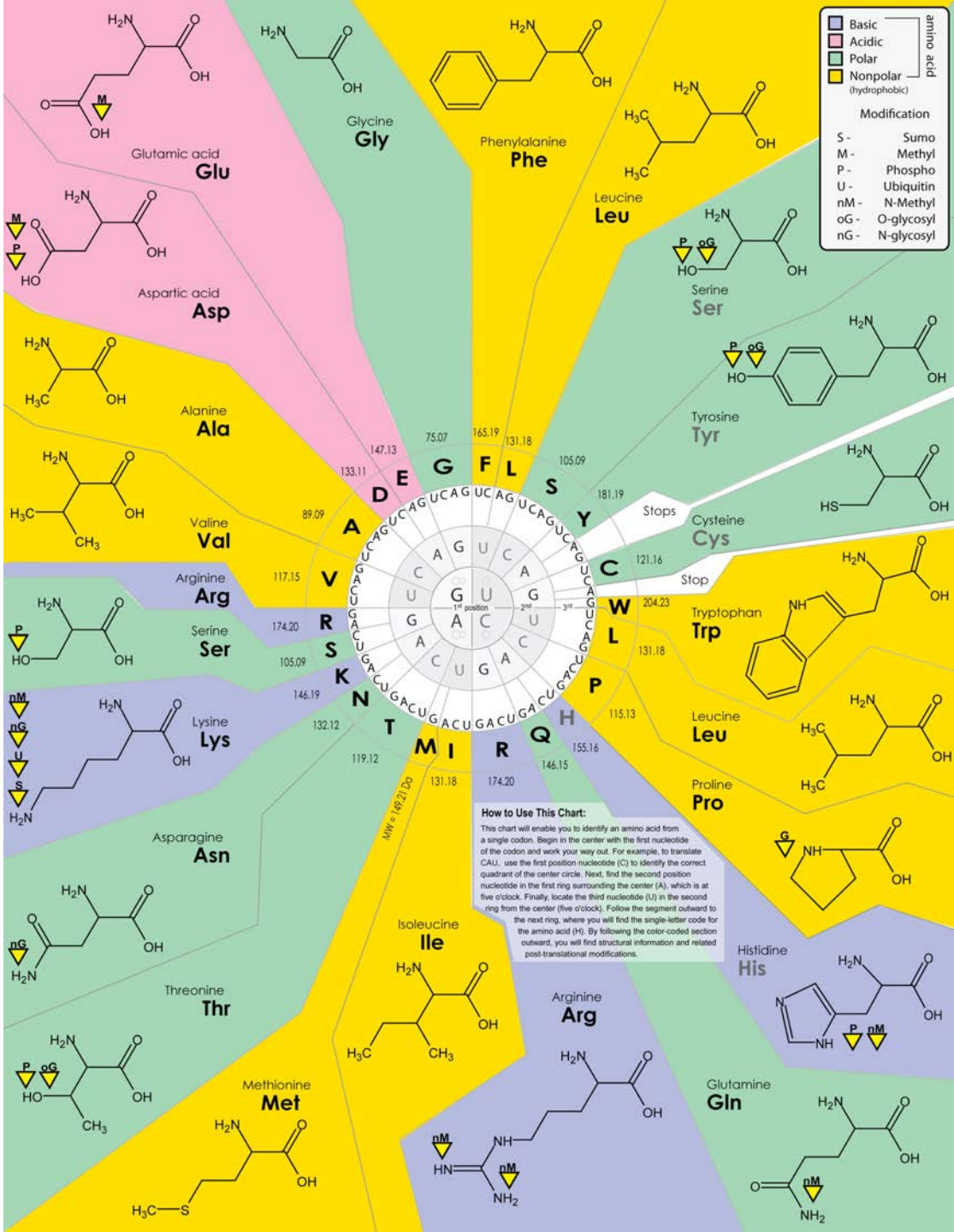
Types of codon changes

Synonymous (silent):

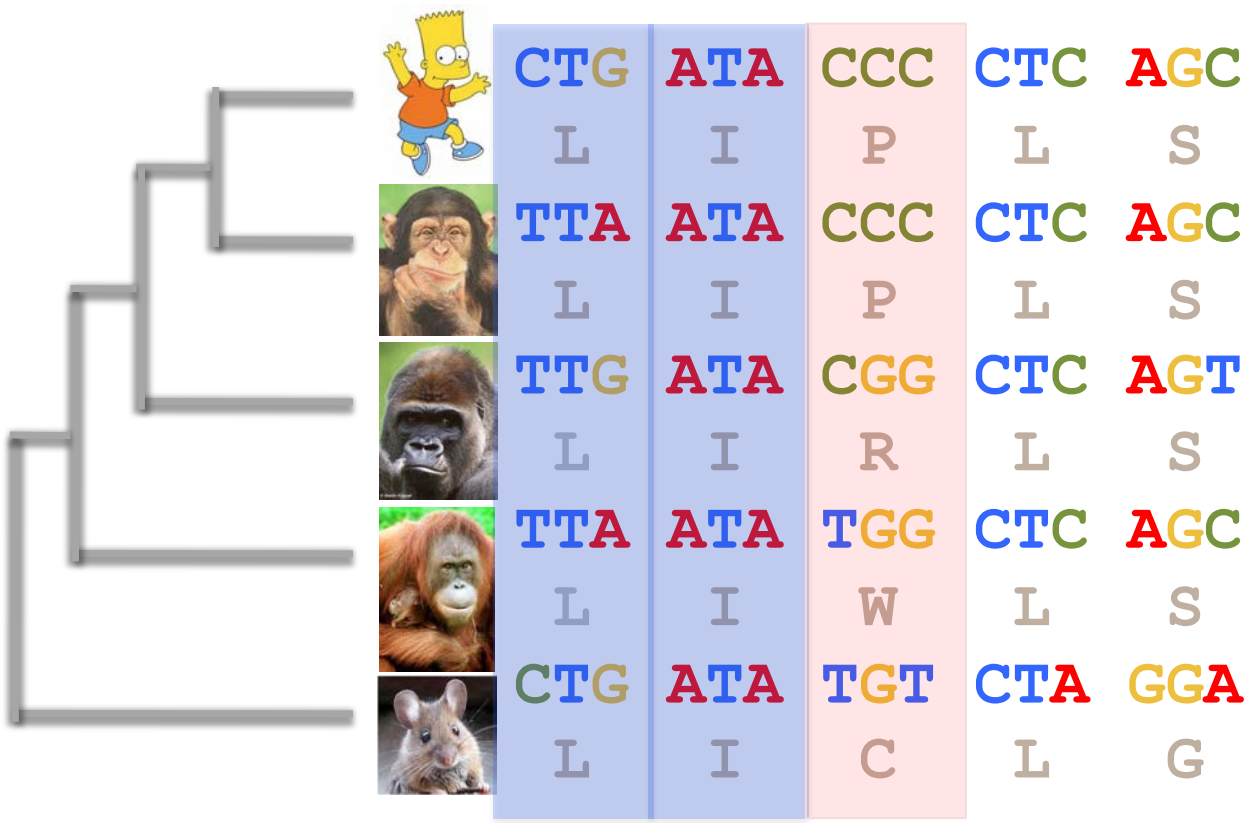
TT**C** (Phe) → TT**T** (Phe)

Nonsynonymous:

TT**C** (Phe) → TT**A** (Leu)



Measuring selection on the protein



synonymous rate: d_S nonsynonymous rate: d_N

$\omega = d_N/d_S > 1$ positive selection

$\omega < 1$ negative selection

Why not counts but rates?

Example:
Pairwise alignment of 500 codons
Observed differences:
5 synonymous differences
5 nonsynonymous differences

Conclusion: Neutral evolution?

Hint: Need to know how many sites are synonymous and how many are nonsynonymous

Evolution at the three codon positions

Relative proportion of different types of mutations in hypothetical protein coding sequence.				
	Expected number of changes (proportion)			
Type	**All 3 Positions**	**1st positions**	**2nd positions**	**3rd positions**
Total mutations	549 (100)	183 (100)	183 (100)	183 (100)
Synonymous	134 (25)	8 (4)	0 (0)	126 (69)
Nonsyonymous	392 (71)	166 (91)	176 (96)	57 (27)
nonsense	23 (4)	9 (5)	7 (4)	7 (4)

Modified from Li and Graur (1991). Note that we assume a hypothetical model where all codons are used equally and that all types of point mutations are equally likely.

Note: by framing the counting of sites in this way we are using a "mutational opportunity" definition of the sites Not everyone agrees that this is the best approach. For an alternative view see **Bierne and Eyre-Walker 2003 Genetics 168:1587-1597**.

Why not counts but rates?

Example:
Pairwise alignment of 500 codons (or 3x500 nt)

5 syn. differences, 25.5% syn. sites:
$S = 500 \times 3 \times 25.5\% = 382.5$, so $d_S = 5/382.5 = 0.013$

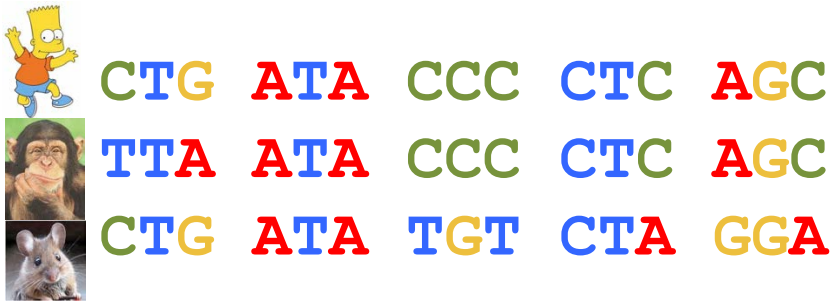
5 nonsyn. differences, 74.5% nonsyn. sites:
$N = 500 \times 3 \times 74.5\% = 1117.5$, so $d_N = 5/1117.5 = 0.0045$

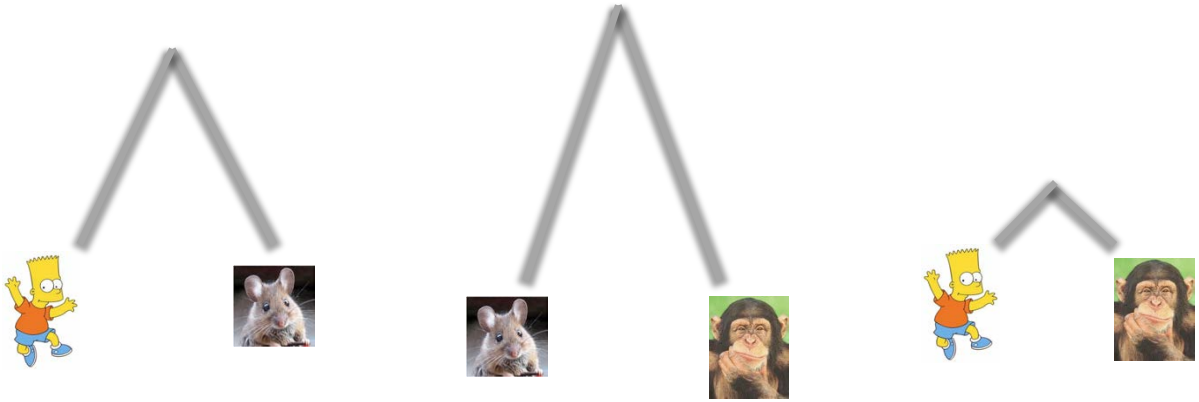
$d_N/d_S = 0.0045/0.013 = 0.35 < 1$

Conclusion: Purifying selection

Pairwise estimation of dN and dS

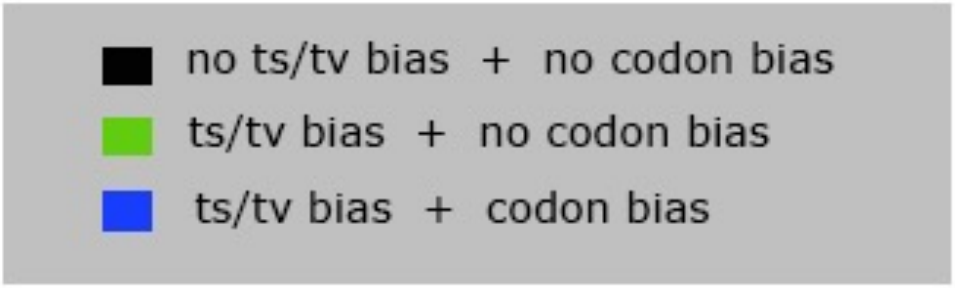
1. Count synonymous and nonsynonymous *sites* (*S* and *N*)
2. Count synonymous and nonsynonymous *differences*
3. Calculate the proportion of differences, then d_N and d_S
4. Correct for *multiple hits*





Numerous counting methods of increasing sophistication

1. Perler, F. et al. 1980. Cell 20: 555-566
2. Miyata, T. & T. Yasunaga. 1980. *JME* 16:23-36
3. Li, W.-H., C.-I. Wu, & C.-C. Luo. 1985. MBE 2:150-174
4. Nei, M. & T. Gojobori. 1986. *MBE* 3: 418-426
5. Li, W.-H. 1993. *JME* 36:96-99
6. Pamilo & Bianchi 1993 *MBE* 10:271-281
7. Ina, Y. 1995. *JME* 40:190-226
8. Comeron, J. M. 1995. *JME* 41:1152-1159
9. Moriyama, E. N. & F. R. Powell, 1997. *JME* 45:378-391
10. Yang, Z., and R. Nielsen. 2000. *MBE* 17:32-43.



Human & orangutan α2-globin genes: 142 codons

Method/Model	κ	S	N	d_N	d_S	d_N/d_S
NG86	1	109.4	316.6	0.0095	0.0569	0.168
Ina95	2.1	119.3	299.9	0.0101	0.0523	0.193
YN00	6.1	61.7	367.3	0.0083	0.1065	0.078
ML (GY94)						
(1) ML Fequal, $\kappa = 1$	1	108.5	317.5	0.0093	0.0557	0.167
(2) ML Fequal, κ estimated	3.0	124.6	301.4	0.0099	0.0480	0.206
(7) ML F61, $\kappa = 1$ fixed	1	58.3	367.7	0.0082	0.1145	0.072
(8) ML F61, κ estimated	5.3	55.3	370.7	0.0082	0.1237	0.066

Base frequencies at 3rd position:
T = 9%, C = 52%, A = 1%, G = 37%

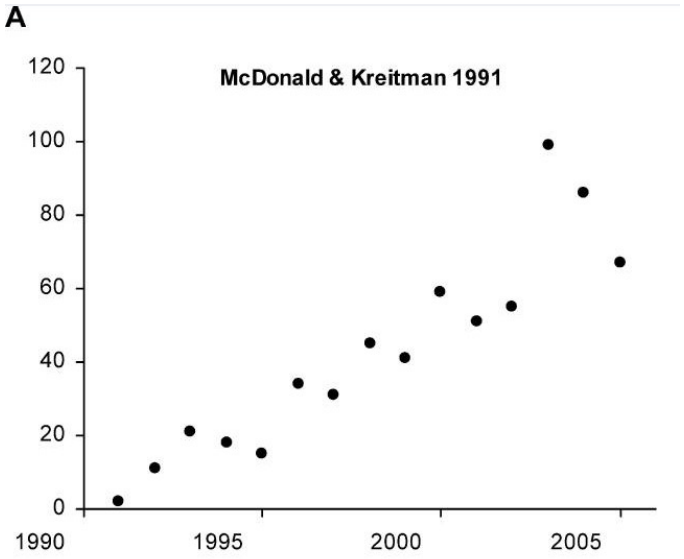
(Yang & Bielawski 2000. *TREE* 15:496–503)

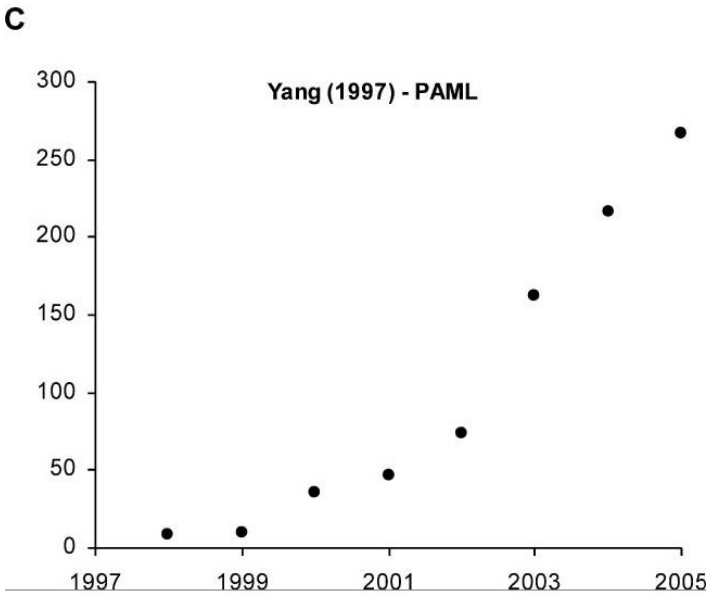
Why Markov codon models

- Take phylogeny into account
- Estimate evolutionary parameters
- Correct for multiple hits
- Account for all possible evolutionary pathways between codons and weight them based on a model

Markov codon models: a success story

- Rigorous statistical framework for hypothesis testing
- Explicitly incorporates evolutionary parameters
- Extensively tested in simulation and on real data:
  low false positive rate and good power
  (e.g., Anisimova *et al.* 2001-2003; Anisimova & Yang 2007)





Types of codon substitution models

- Branch models to test positive selection on lineages on the tree
  (Yang 1998. *Mol. Biol. Evol.* 15:568-573)

- Site models to test positive selection affecting individual sites
  (Nielsen & Yang. 1998. *Genetics* 148:929-936;
  Yang, *et al.* 2000. *Genetics* 155:431-449)

- Branch-site models to detect positive selection at a few sites on a particular lineage
  (Yang & Nielsen. 2002. *Mol. Biol. Evol.* 19:908-917;
  Yang, *et al.* 2005. *Mol. Biol. Evol.* 22:1107-1118)

Markov model of codon evolution

Instantaneous substitution matrix $Q = \{q_{ij}\}$:

MG-type model	Type of change	GY-type model
0	2 or 3 nt changes	0
f_x^p	Synonymous transversion	π_j
κf_x^p	Synonymous transition	$\kappa \pi_j$
ωf_x^p	Nonsynonymous transversion	$\omega \pi_j$
$\omega \kappa f_x^p$	Nonsynonymous transition	$\omega \kappa \pi_j$

ω = d_N/d_S (selection on protein)

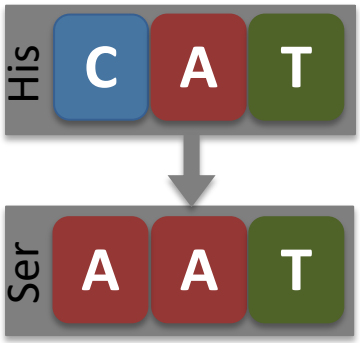
κ = transition/transversion ratio

π_j = frequency of codon j

f_x^p = frequency of nucleotide x at codon position p

Defining instantaneous rates

There are many ways to define instantaneous rates:



Exchangeabilities based on	MG-type frequencies	GY-type frequencies
HKY85	$\omega\,\kappa\,f_A^1$	$\omega\,\kappa\,\pi_{AAT}$
GTR	$\omega\,r_{C\to A}\,f_A^1$	$\omega\,r_{C\to A}\,\pi_{AAT}$
Codon-based	$R_{CAT\to AAT}\,f_A^1$	$R_{CAT\to AAT}\,\pi_{AAT}$

Modeling codon frequencies

All codon models assume reversibility and stationarity

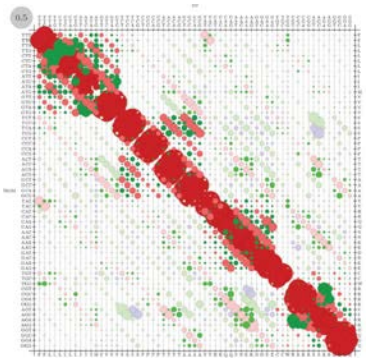
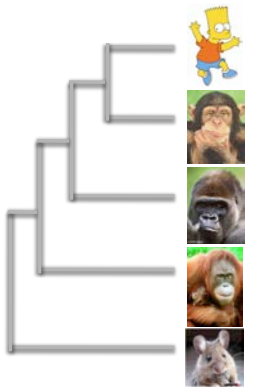
Codon frequencies $\{\pi_j\}$ are the same at any time

Model	His C A T	Ser A A T
Fequal	1/61	1/61
F1x4	$f_C f_A f_T$	$\left(f_A\right)^2 f_T$
F3x4	$f_C^1 f_A^2 f_T^3$	$f_A^1 f_A^2 f_T^3$
F61	π_{CAT}	π_{AAT}

Likelihood function over phylogeny

Transition probability matrix over time t: $\boldsymbol{P(t) = e^{Qt}}$

Using P(t) a likelihood L(Data) can be constructed:

Pr( |  )

Data

Model

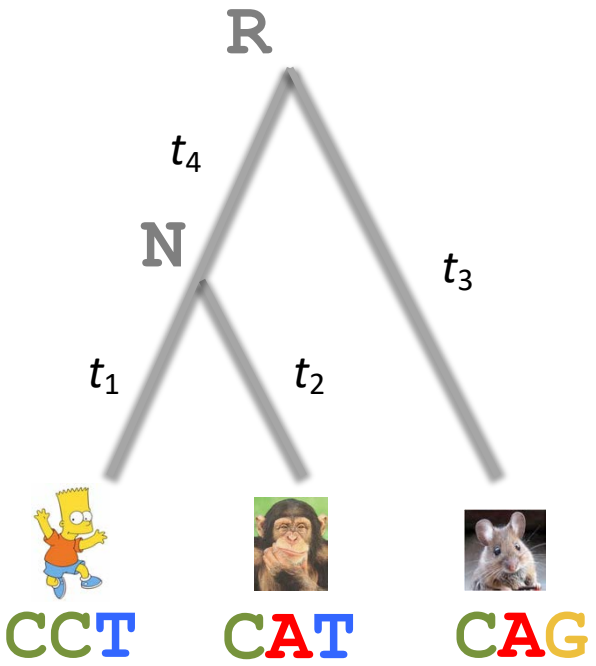
Tree with branch lengths

Parameters optimized by maximum likelihood

Likelihood function over phylogeny

For each site compute the likelihood:

$$L_h = L\begin{pmatrix} CCT \\ CAT \\ CAG \end{pmatrix} = \sum_R \pi_R p_{R\to CAG}(t_3) \sum_N p_{R\to N}(t_4) p_{N\to CCT}(t_1) p_{N\to CAT}(t_2)$$



Unrooted tree – arbitrary root

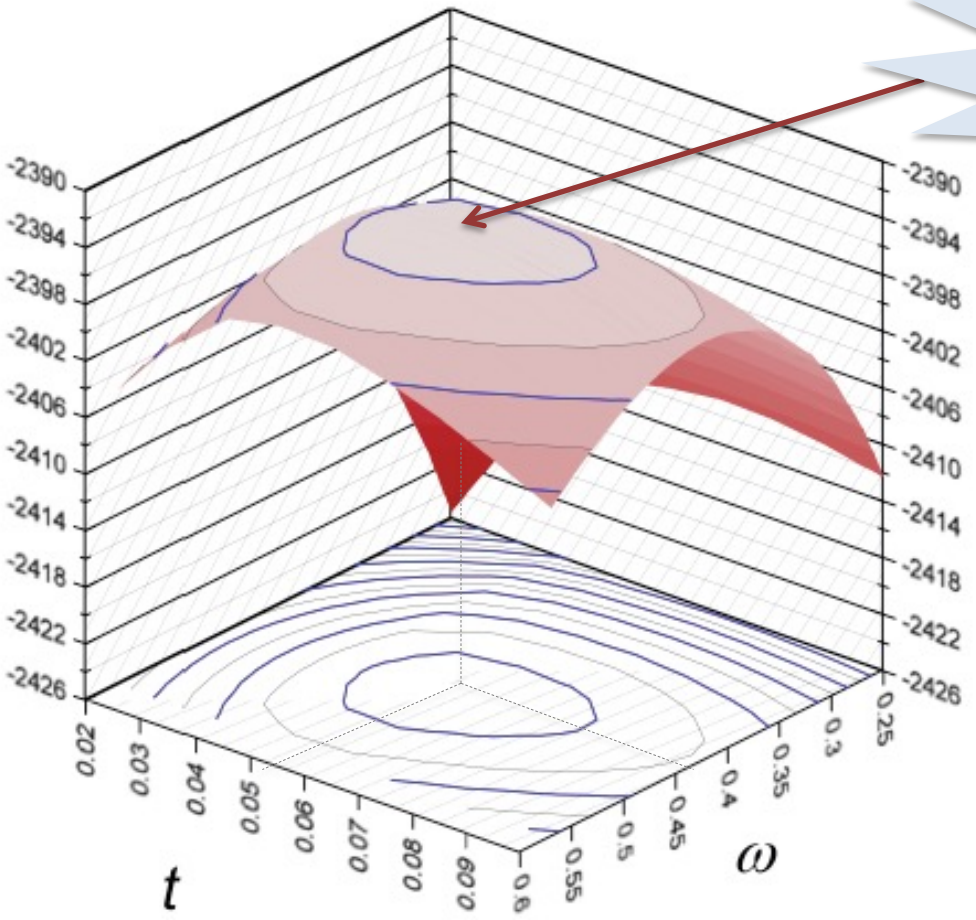
Compute total likelihood assuming independent & identical distribution (i.i.d.) for all sites:

$$L = L_1 \times L_2 \times \ldots \times L_n = \prod_{h=1}^{n} L_h$$

Log-likelihood is optimized (for convenience):

$$\ell = \ln L = \ln L_1 + \ln L_2 + \ldots + \ln L_n = \sum_{h=1}^{n} \ln L_h$$

ML parameter estimation



Numerical optimization by hill-climbing

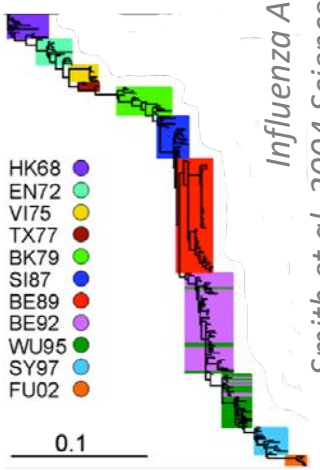
Example ML estimation for acetylcholine α receptor from human and mouse

Steps in analyses with codon models

Ideally:

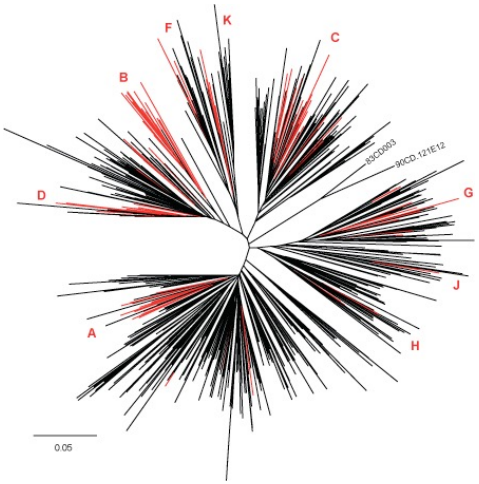
- Infer multiple sequence alignment (directly on codons?)
- Infer phylogeny (using codon models?)
- Analyse data with codon models

Selection affects the shape of tree



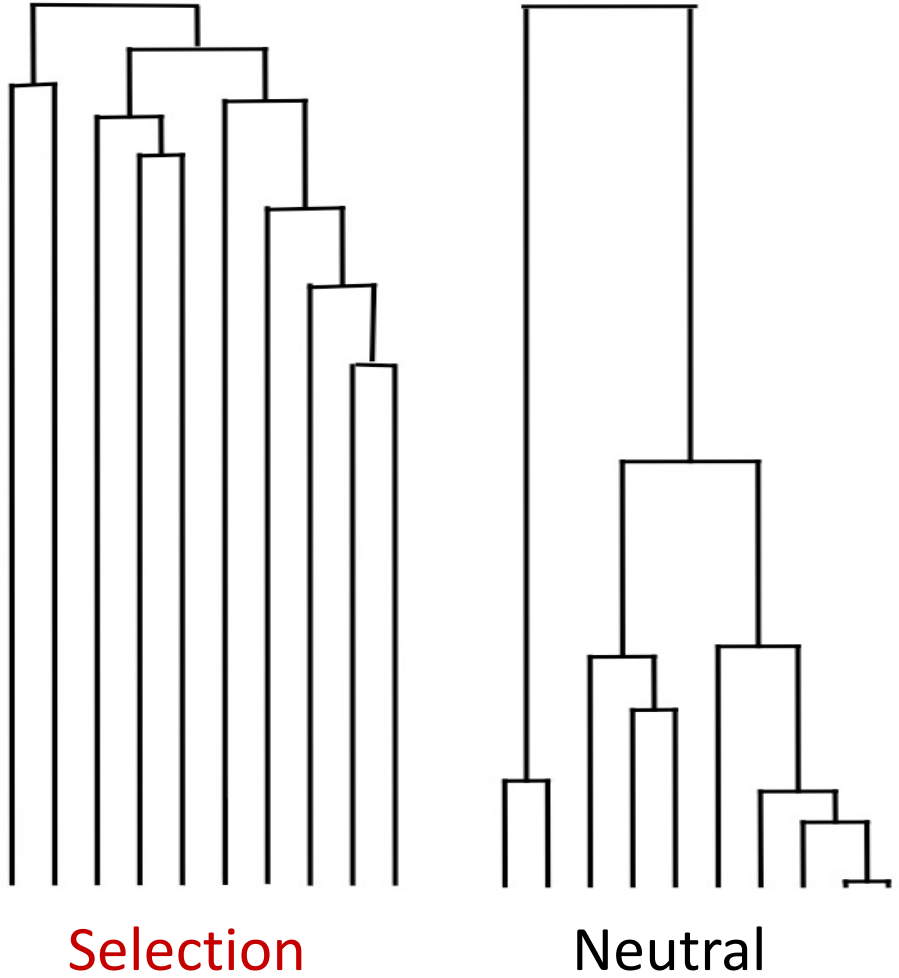
Influenza A

Smith et al. 2004 Science



HIV-1 subtypes

Archer, Robertson 2007, AIDS



Selection

Neutral

CodonPhyML : maximum likelihood tree inference

Hundreds of codon models

- Parametric, empirical, semi-parametric
- Comparable likelihoods across AA, DNA, codon data

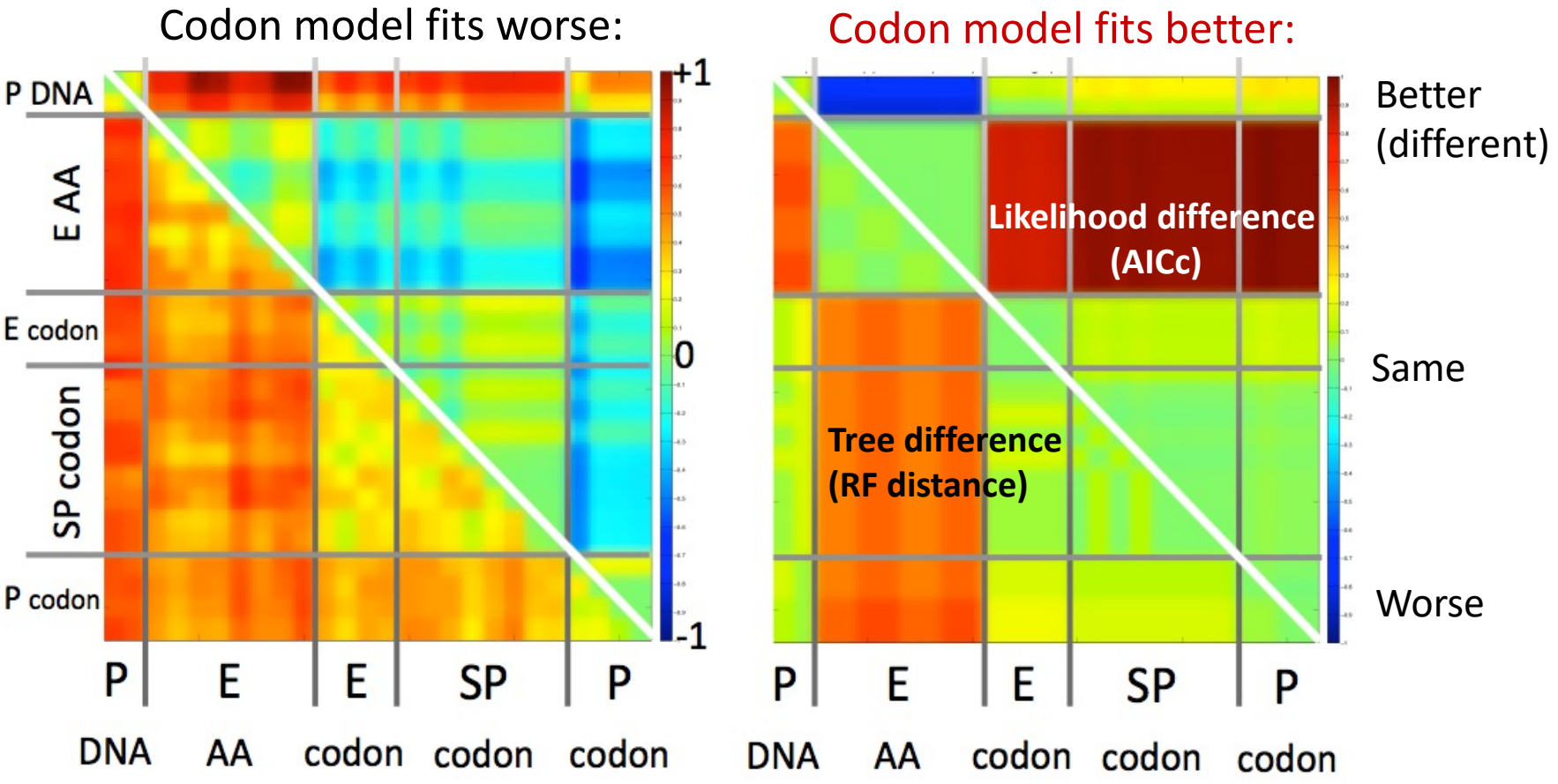
High performance computing

Anisimova, Gascuel 2006 Syst Biol
Guindon et al. 2010 Syst Biol
Anisimova et al. 2011 Syst Biol
Gil et al. 2013 Mol Biol Evol

CodonPhyML:
Model & tree comparison on real data

Model types: DNA, AA, codon

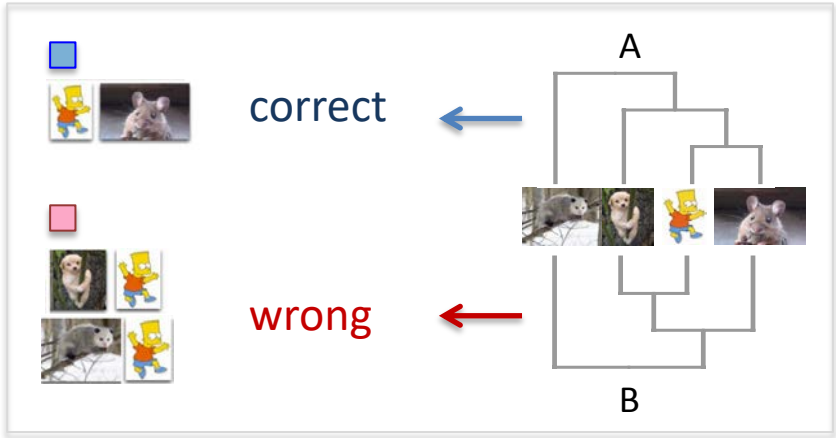
E = empirical, **SP** = semi-parametric, **P** = parametric

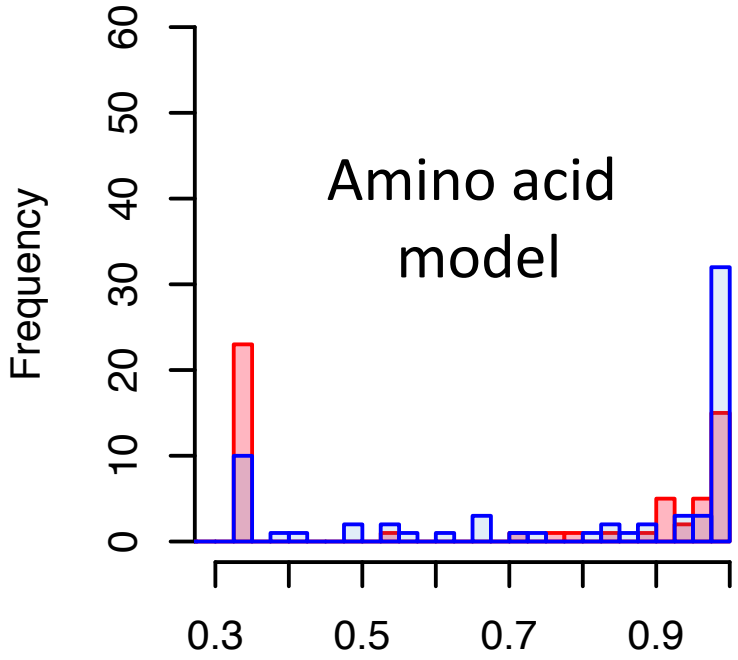


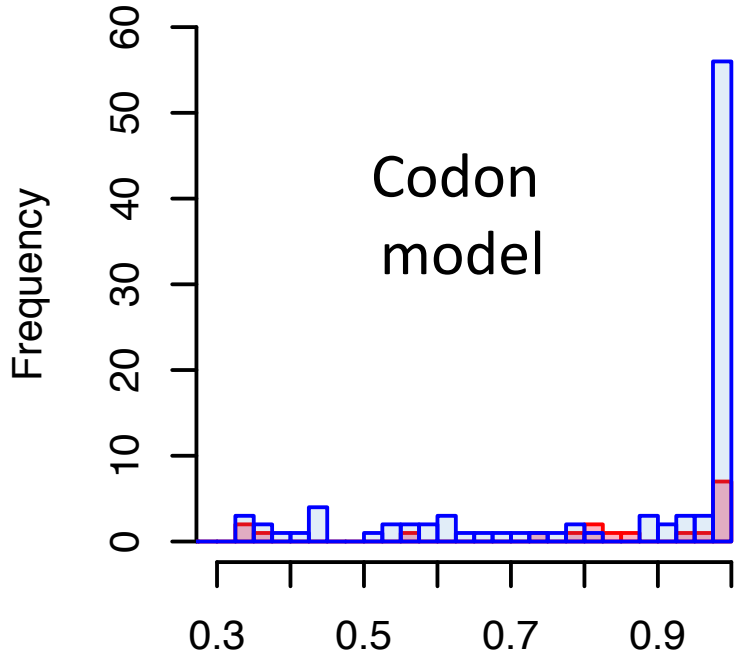
Gil et al 2013, Mol Biol Evol

CodonPhyML: evaluating inferred splits

22 mammalian species
72 protein orthologs







aBayes support (Anisimova et al 2011, Syst Biol)

Unpublished, Anisimova

http://sourceforge.net/projects/codonphyml

Summary Files Reviews Support Wiki Discussion Donate Code Bugs

codonPhyML

anisimova, laduplessis, mgil_, mszanetti, stefanzoller

6 Recommendations
85 Downloads (This Week)
Last Update: 4 hours ago

Download source code

Browse All Files

Tweet 0 +1 0 Like 0



Description

codonPhyML uses Markovian codon models of evolution in phylogeny reconstruction. Given a set of species characterized by their DNA sequences as input, codonPhyML will return the phylogenetic tree that best describes their evolutionary relationship. Our paper describing codonPhyML has been accepted for publication in the journal "Molecular Biology and Evolution". For more details, follow the link: http://mbe.oxfordjournals.org/content/early/2013/02/23/molbev.mst034.short.

Exercises with PAML (codeml)

Focus of exercise #1:

ML estimation with one ω-ratio model M0