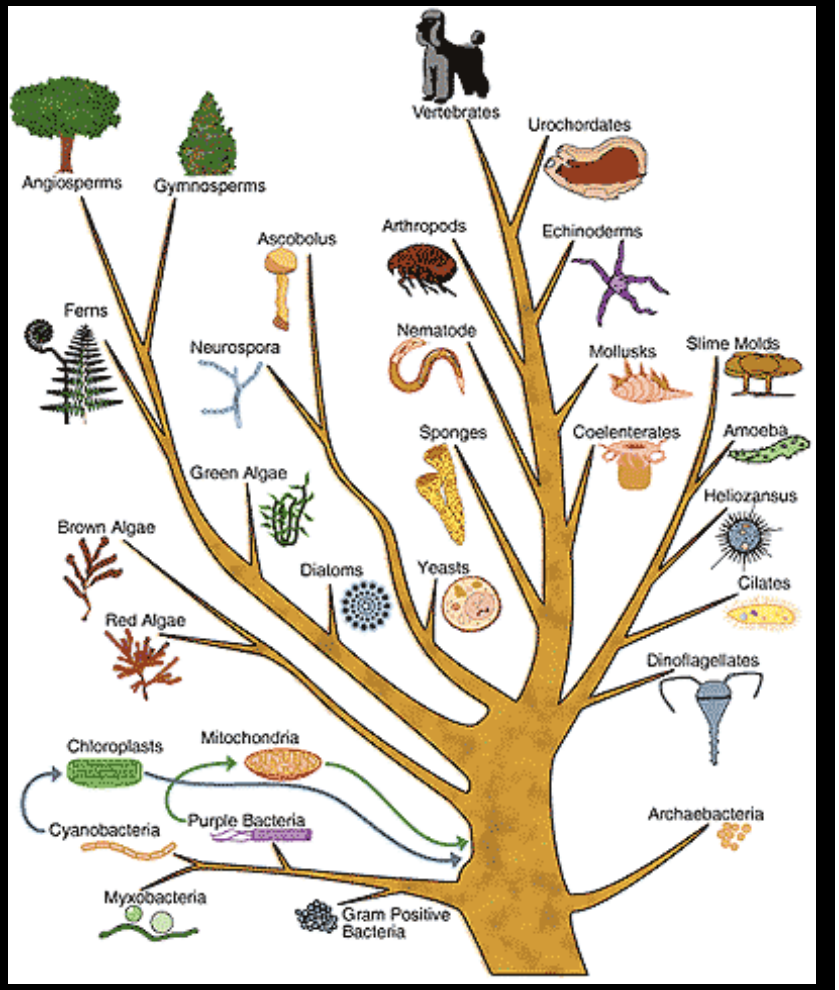
Phylogeny and molecular evolution

Estimation of phylogeny and selective pressure

- Foundation of comparative genomics and population genetics
- Relatively mature methodology frameworks
- Essential component or the ultimate purpose in many -omics studies
- Relate genotype with phenotype

Why reconstructing phylogenetic trees?

- Evolution tell many stories in biology
  - Origins and migrations of organisms
  - Natural selection acts on specific genes
  - Predicting functions of genes
  - Origins and spread of disease

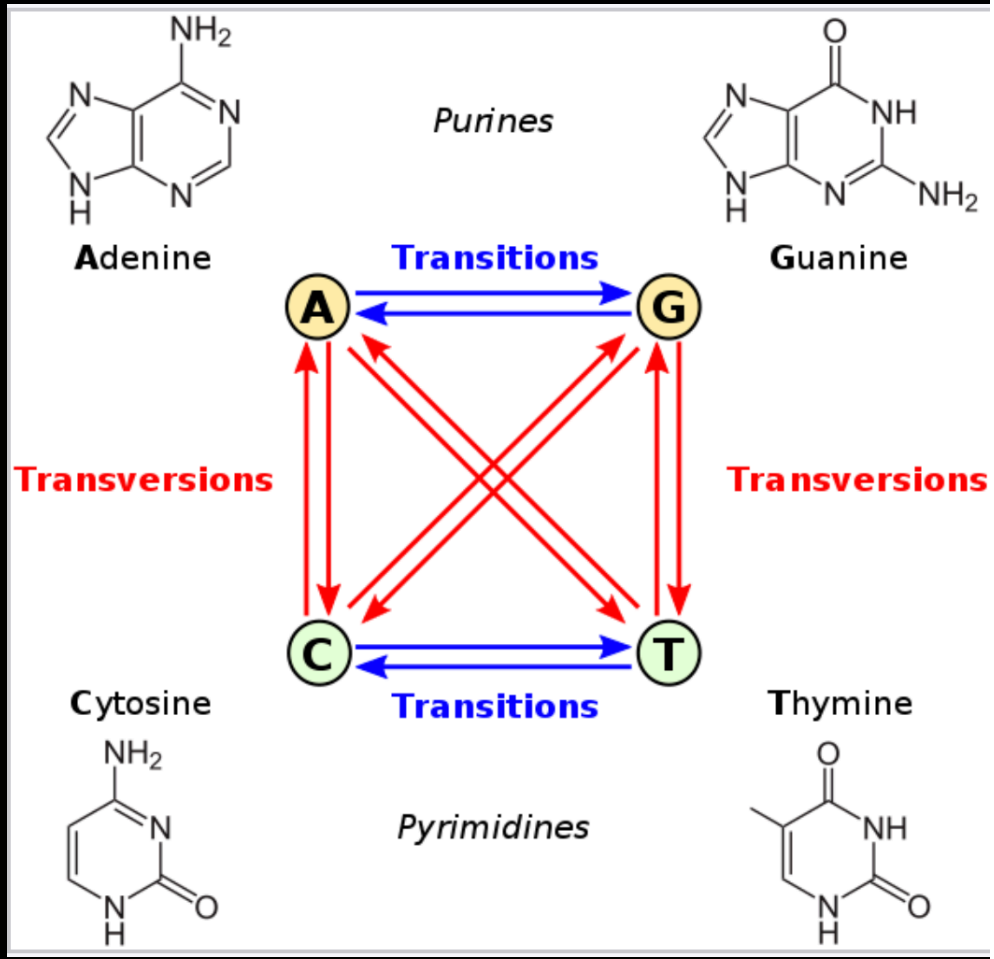


Sequence divergence, distance

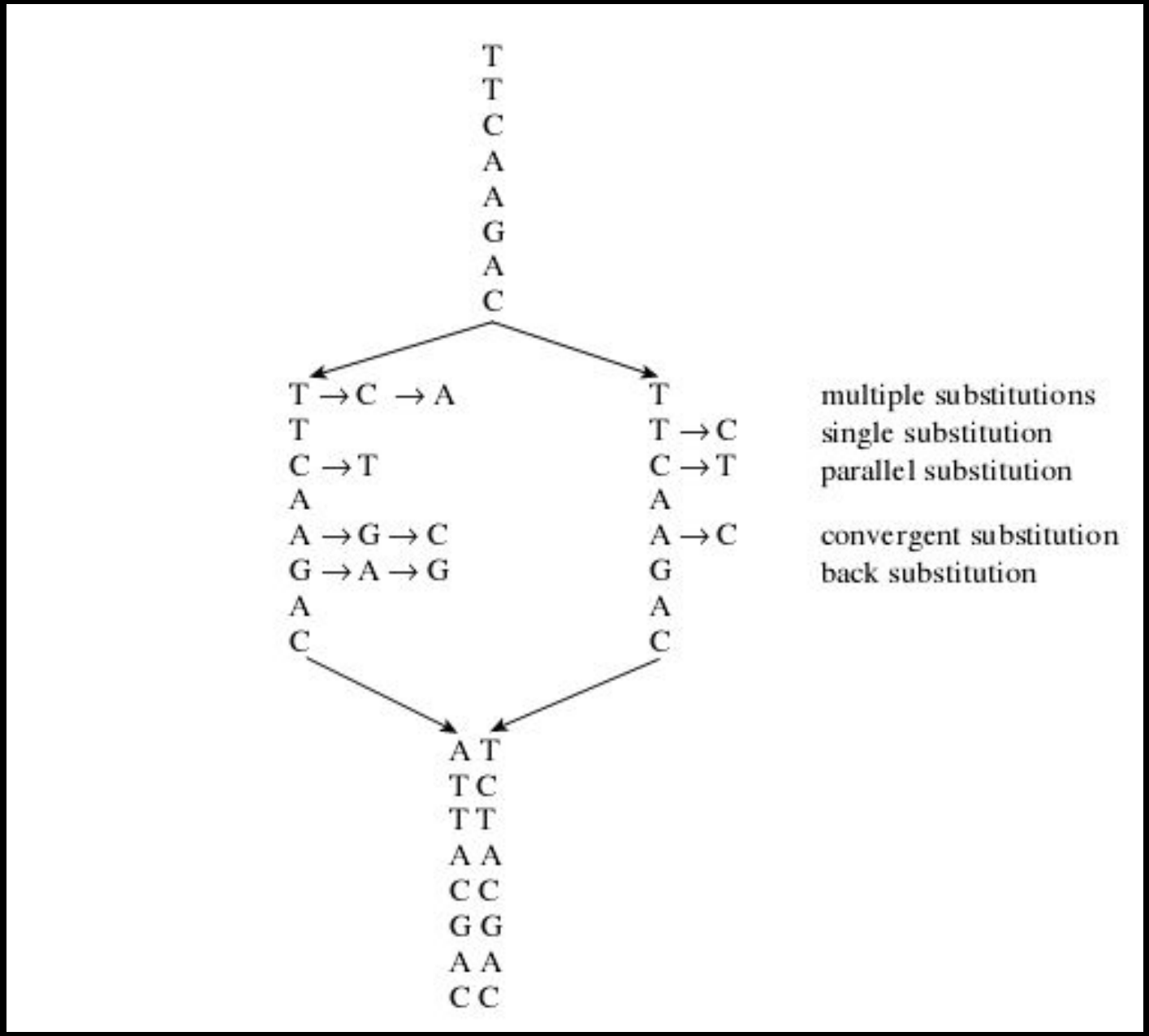
- The simplest measure of the extent of sequence divergence is the proportion (p) of nucleotide/amino acid sites, at which the two sequences are different

$$P = n_d / n$$

  - Where n_d is the number of different residues, and n is the number of total residues



Multiple substitution



Substitution matrix and genetic distance after correction

- For Jukes-Cantor model

$$\hat{d} = -\tfrac{3}{4}\log\left(1 - \tfrac{4}{3}\hat{p}\right),$$

Q =

	A	T	C	G
A	-	α	α	α
T	α	-	α	α
C	α	α	-	α
G	α	α	α	-

A G. C. T

- For Kimura 80 model

$$\hat{d} = -\tfrac{1}{2}\log(1 - 2S - V) - \tfrac{1}{4}\log(1 - 2V),$$

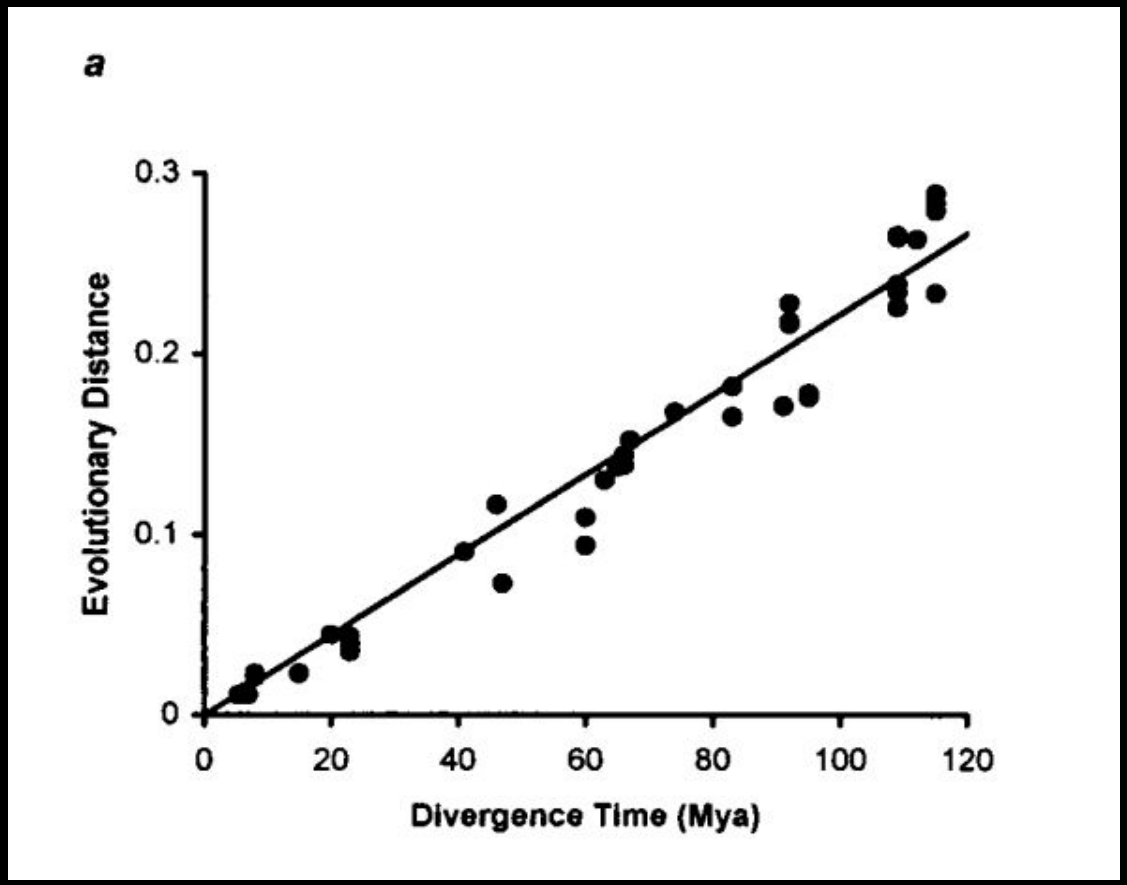
$$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$

- where S and V are the proportions of sites with transitional and transversional differences, respectively.

K represent the transition/transversion rate

Molecular Clock

- To describe a relatively constant rate of molecular evolution



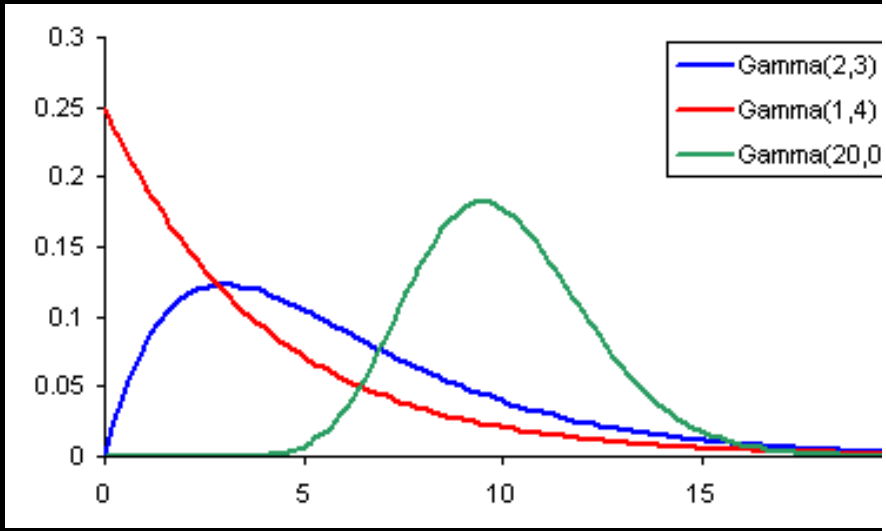
The ‘Sloppy’ Clock

- ‘Ticks’ are stochastic, not deterministic
  - Mutations happen randomly according to a Poisson distribution.
- Long divergence time or frequently mutating can result in the similar number of mutations

Good/strict clock

- Actually over-dispersed

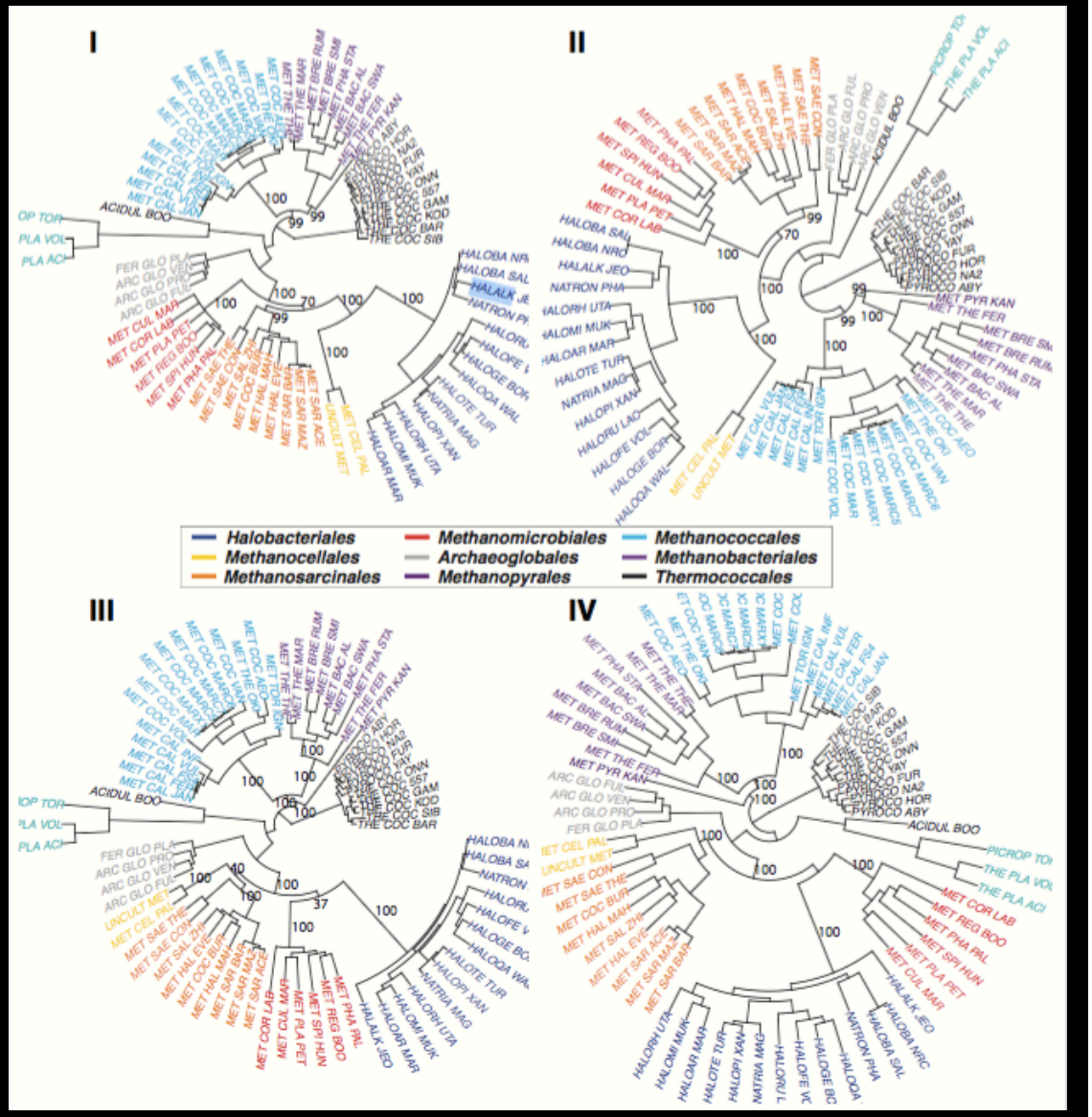
  Poisson Relaxed clock



Calculation of genetic distance

- Genetic distance measure the divergence between sequences/populations/species

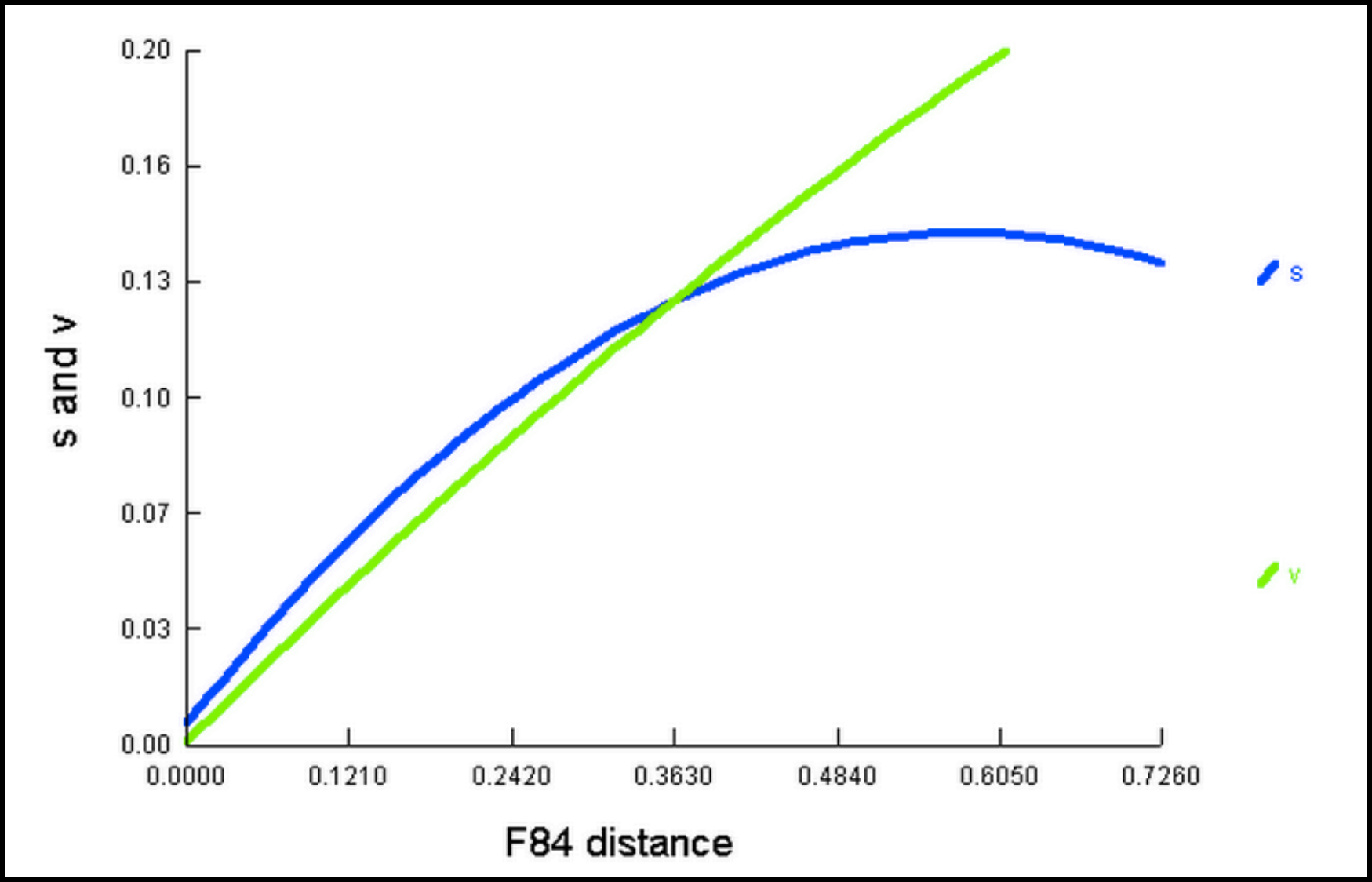
- Tools
  - Dnadist, Protdist in Phylip
  - YN00, 4DTV
    - fourfold degenerate transversion can tolerate any point mutation at the third position
  - Phyml, RAxML, Mrbayes
  - STRUCTURE
  - …



Li J. etc Genome Biol Evol. 2014 Dec

Easily saturated transition rate

So we go to transversion to estimate long genetic distance



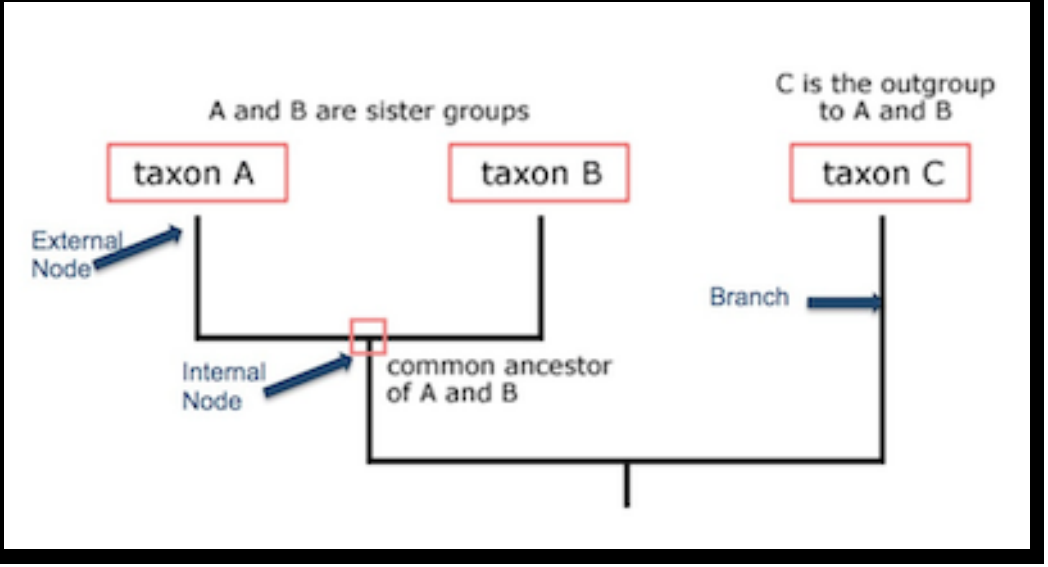
http://www.ccg.unam.mx/~vinuesa/Model_fitting_in_phylogenetics.html

Practice

- Calculate nucleotide distance between two sequences
- Calculate 4DTV between two divergent genome gene sets
- Visualize the distribution of 4DTV

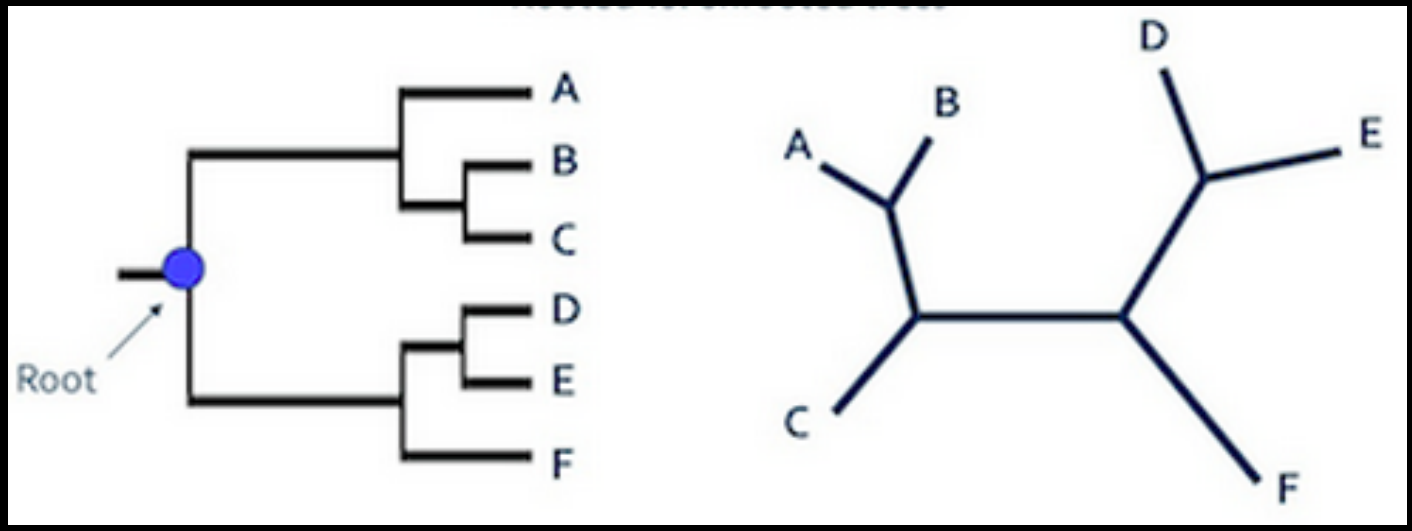
Basic terms in phylogeny

- **Topology** – structure and the relationship
- **Nodes** – DNA (RNA, mtDNA) sequences, proteins, **species** = taxonomic units (TUs)
- **Terminal (extant) nodes, leaves** – OTUs
- **Internal nodes** – unobserved ancestor sequences
- **Branches** – parent-child relations between two nodes
- **Clade** – a group of taxon
  With their common
  Ancestor and all the
  descendants



Rooted and unrooted trees

- Phylogenetic relationships of genes or organisms are usually presented in a tree-like form either with a root– rooted tree or without any root– unrooted tree.

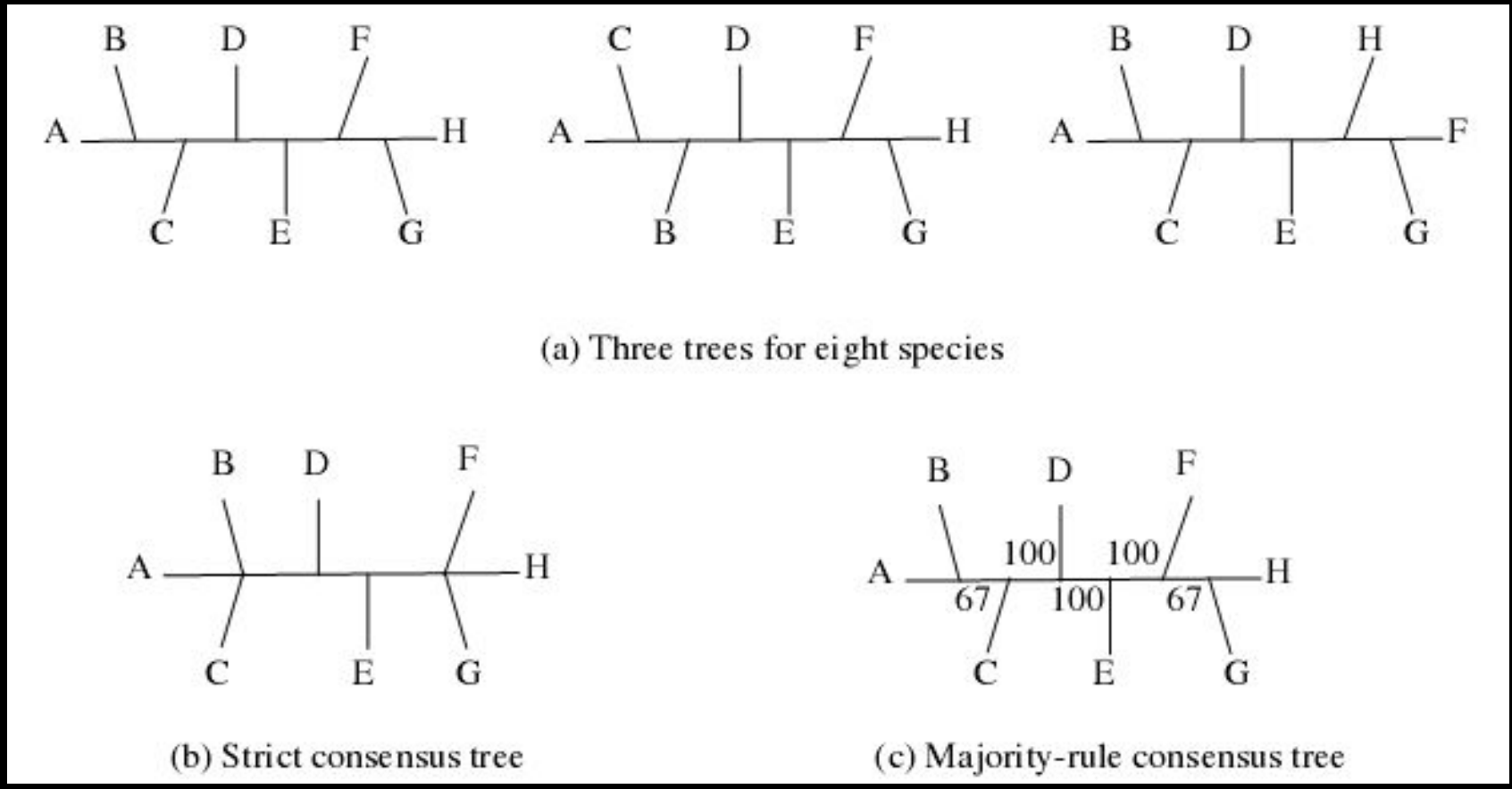


Consensus tree

- The strict consensus tree shows only those groups (nodes or clades) that are shared among all trees in the set, with polytomies(e.g. three-forked) representing nodes not supported by all trees.

- The majority-rule consensus tree shows nodes or clades that are supported by at least half of the trees in the set.

Consensus tree-- example

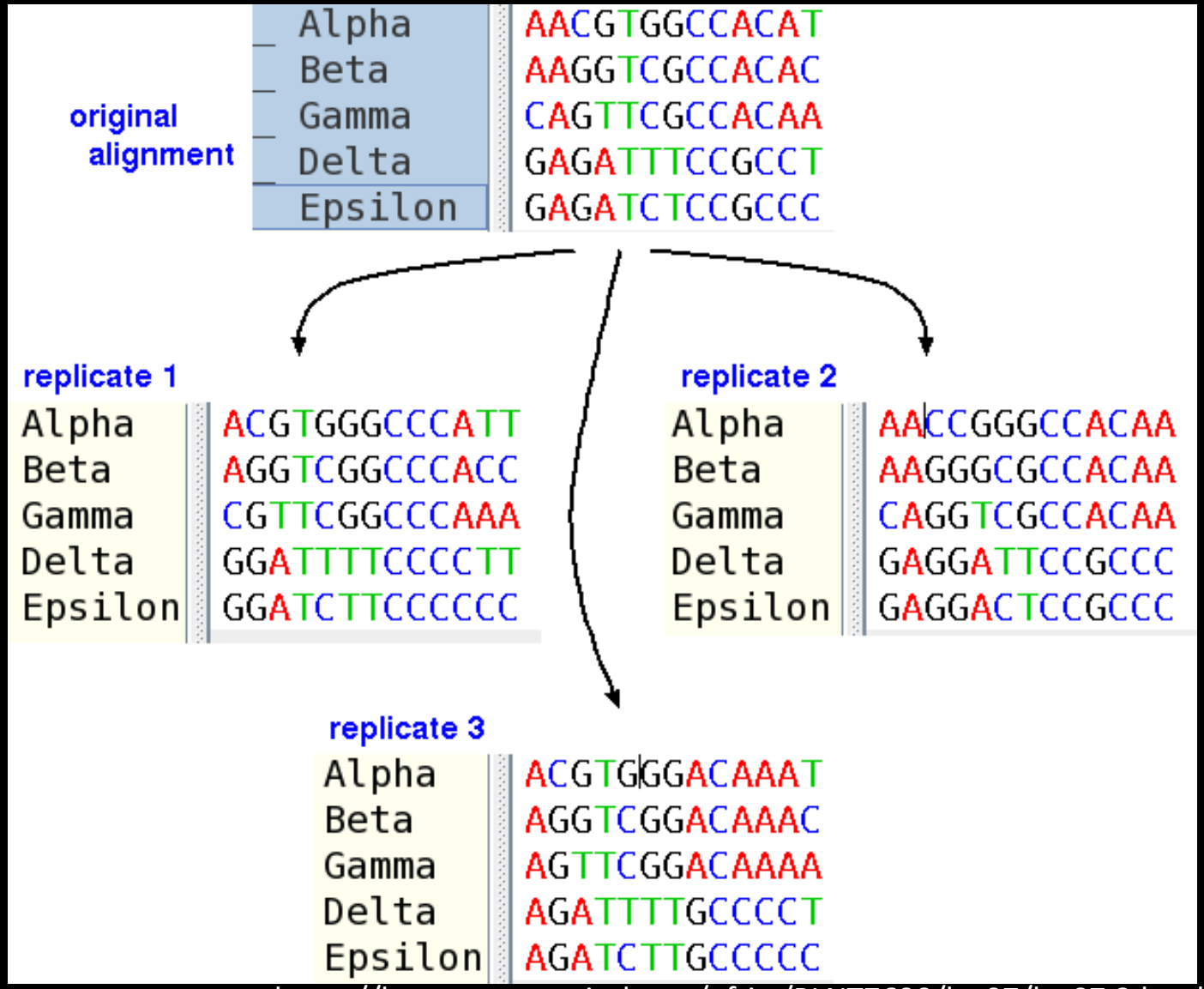


(a) Three trees for eight species

(b) Strict consensus tree

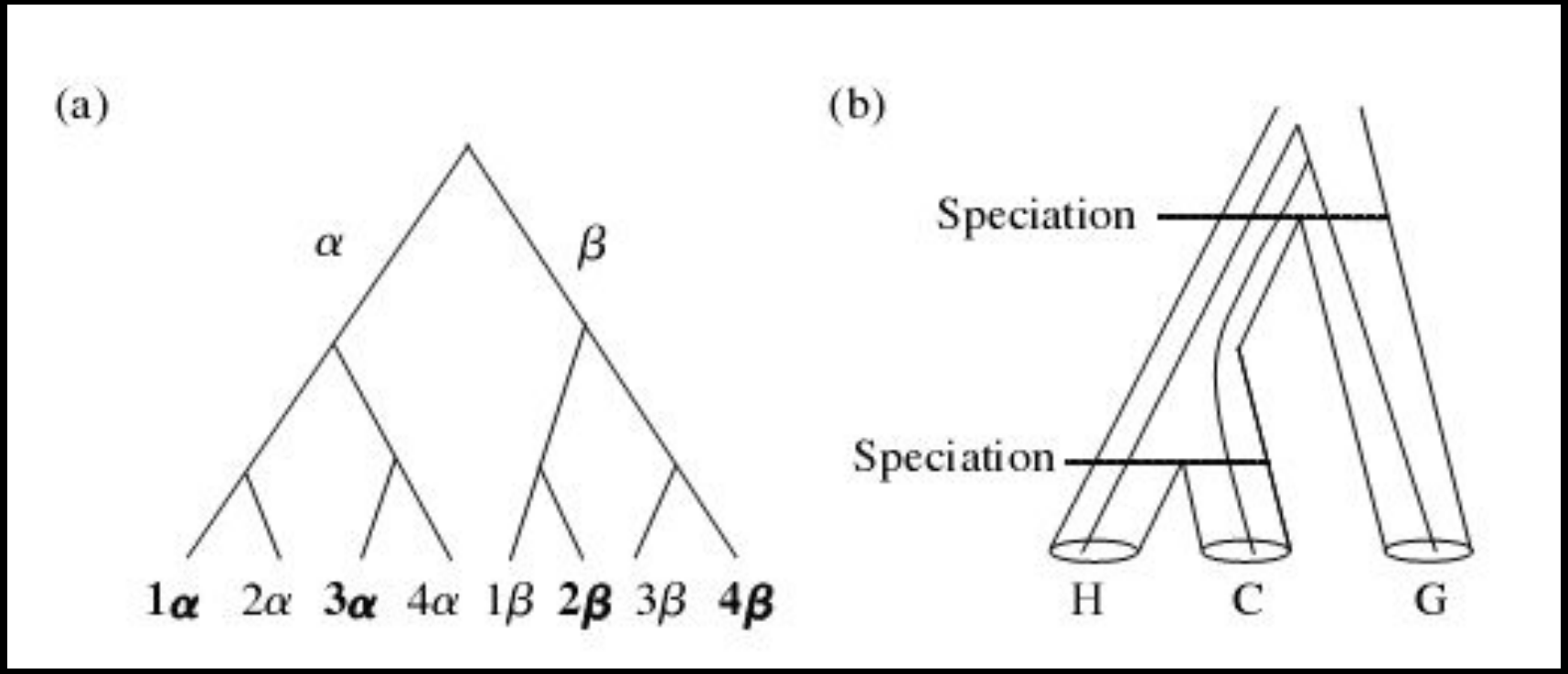
(c) Majority-rule consensus tree

Bootstrap procedure



https://home.cc.umanitoba.ca/~frist/PLNT7690/lec07/lec07.3.html

Discordance between gene trees and species trees-- example



A gene duplicated in the past, creating paralogous copies α and β, followed by divergences of species1,2,3,and 4. If we use gene sequences 1α,3α,2β, 4β for phylogeny reconstruction, the true gene tree is ((1α,3α),(2β,4β)),different from the species tree ((1,2),(3,4))

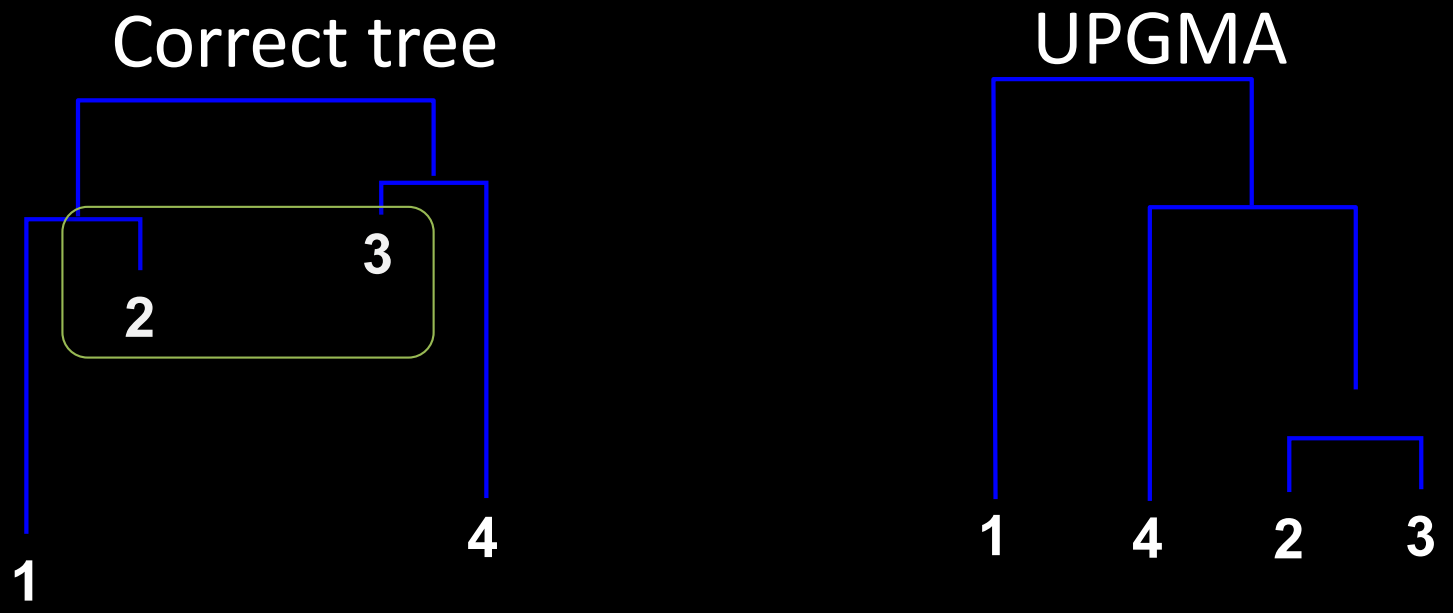
The species tree is ((human,chimpanzee), gorilla). However, due to ancestral polymorphism, the true gene tree is ((human,(chimpanzee, gorilla)).

Factors causing incongruence between gene tree and species tree

- Random estimation error (limited sequences) and systematic error (deficiencies of the tree-reconstruction method)
- Horizontal (lateral) gene transfer
- Gene duplication, especially followed by gene losses
- When species are closely related, ancestral polymorphism

Tree reconstruction method—UPGMA and its weakness

- UPGMA assumes a constant molecular clock: all species represented by the leaves in the tree are assumed to accumulate mutations (and thus evolve) at the same rate. This is a major pitfall of UPGMA.



Phylogenetic inference: Neighbor joining (NJ)

- Finds a pair of leaves that are close to each other but far from other leaves: implicitly finds a pair of neighboring leaves
- Similar to UPGMA, merges clusters iteratively
- Finds two clusters that are closest to each other and farthest from the other clusters

Phylogenetic inference: maximum parsimony method

- Occam's Razor “Entities should not be multiplied unnecessarily."

Phylogenetic inference: maximum likelihood method

- The ML method estimates θ by maximizing the log likelihood, often using numerical optimization algorithms

$$\ell = \log(L) = \sum_{h=1}^{n} \log\{f(\mathbf{x}_h|\theta)\}.$$

where x_h denote the h_{th} column in the data matrix.

Tree reconstructing method

- UPGMA and Neighbor joining
  - distance matrix based
  - Fast
  - Perform poorly in high diverged relationships
- Maximum parsimony
  - No evolutionary model needed
  - Fast
  - Perform poorly in high diverged relationships
- Maximum likelihood method and Bayesian methods
  - Evolutionary model based and robust
  - Slow
  - Perform well at high diverged relationships

Software to reconstruct phylogenetic tree reconstruction

- **Treebest** (free)
  - NJ and ML
- **Phylip** (classic, free) and
  - NJ,MP and ML
- **Paup** (classic, not free)
  - NJ,MP and ML
- **Mega** (graphic interface, free)
  - NJ, MP and ML
- **Phyml** and **RAxML** (free)
  - the fastest and most widely used ML software
- **Mrbayes** and **PhyloBayes** (popular Bayesian tree reconstruction package)

Principle to reconstruct the phylogenomic tree

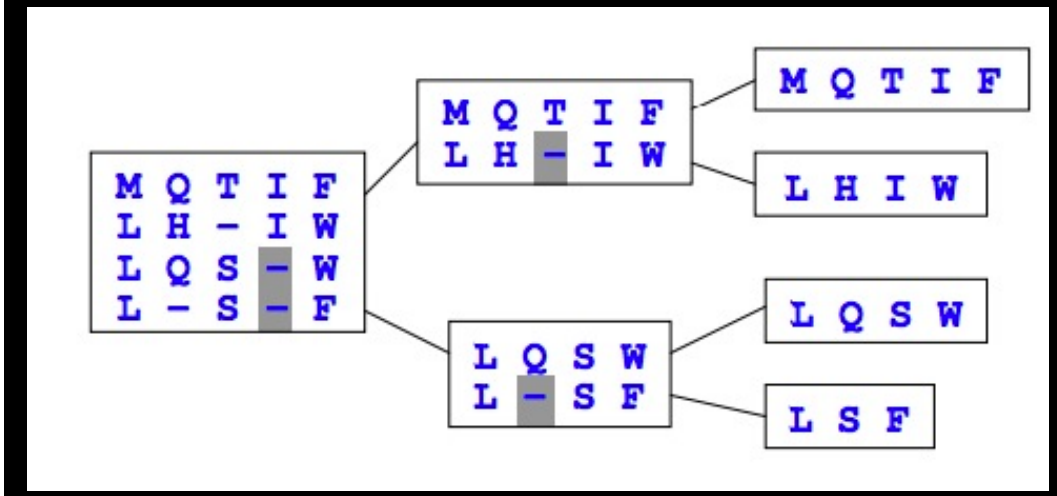
- Multiples universal genes
- Gene without significant recombination (HGT, etc) or linked selection (disequilibrium)
- Incorporate more species to interrupt the long-branch
- Protein sequences instead of DNA sequences for high divergence tree.
- Use likelihood ratio test or Bayesian method to select the best tree

First step in phylogeny reconstruction--multiple alignment using muscle

- Why muscle?

—One of highest accuracy

—Fastest multiple alignment software



Progressive alignment.

Sequences are assigned to the leaves of a binary tree. At each internal (i.e., non-leaf) node, the two child profiles will be merged. Indels introduced at each node are indicated by shaded background

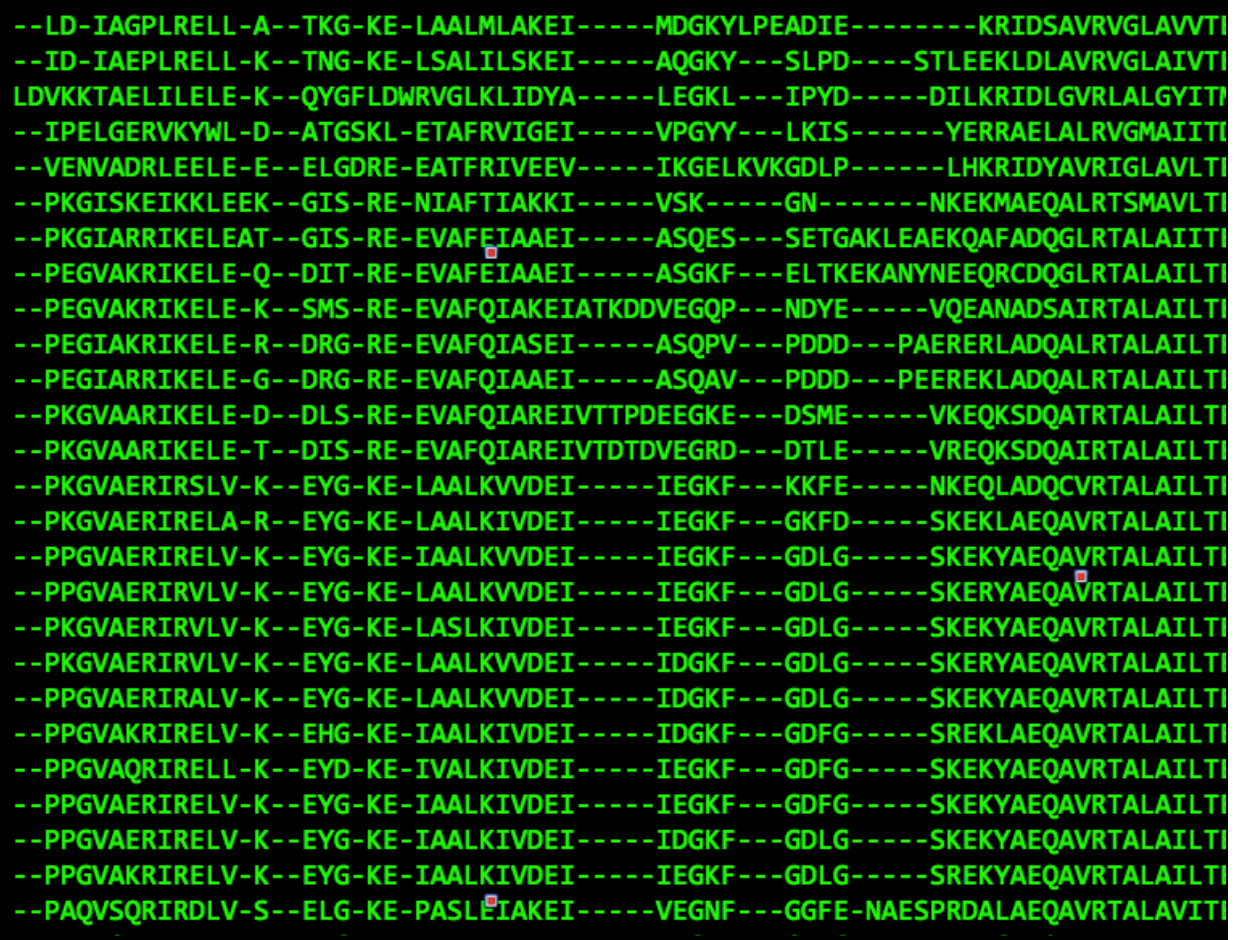
How to deal with divergent multiple alignment

Big impact on phylogeny reconstruction

Removing poor aligned regions would increase the accuracy of phylogeny reconstruction at the cost of losing certain information

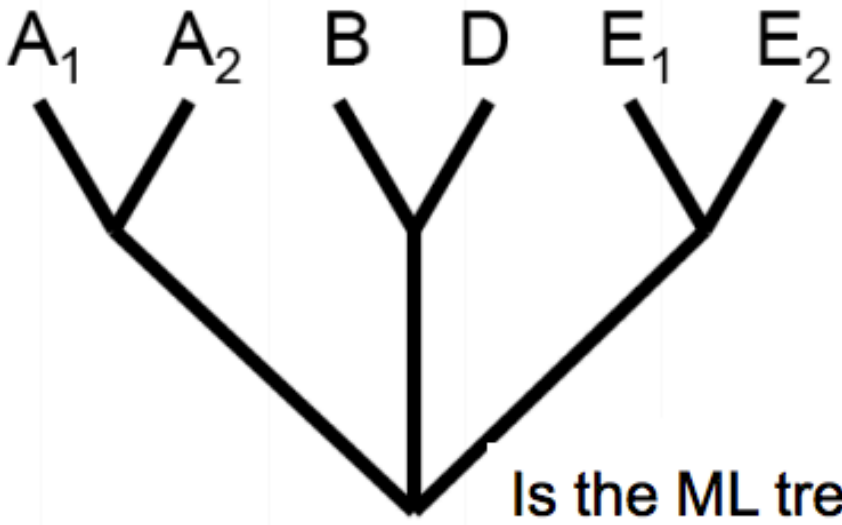
Manually removing problematic aligned regions.

Cleaning alignment using Gblocks, the most widely used alignment process soft



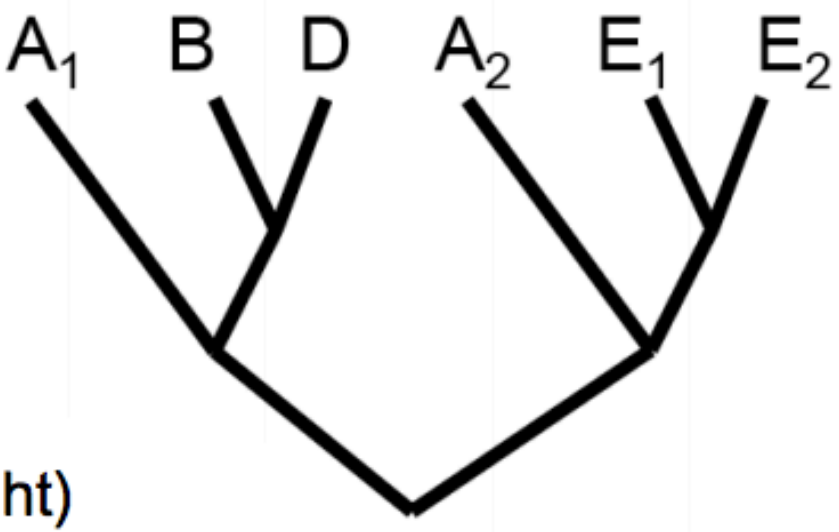
Incongruence test among phylogenetic trees

- Evaluate whether two topology fit the data at the same level



$\ln L_1 = -5073.75$

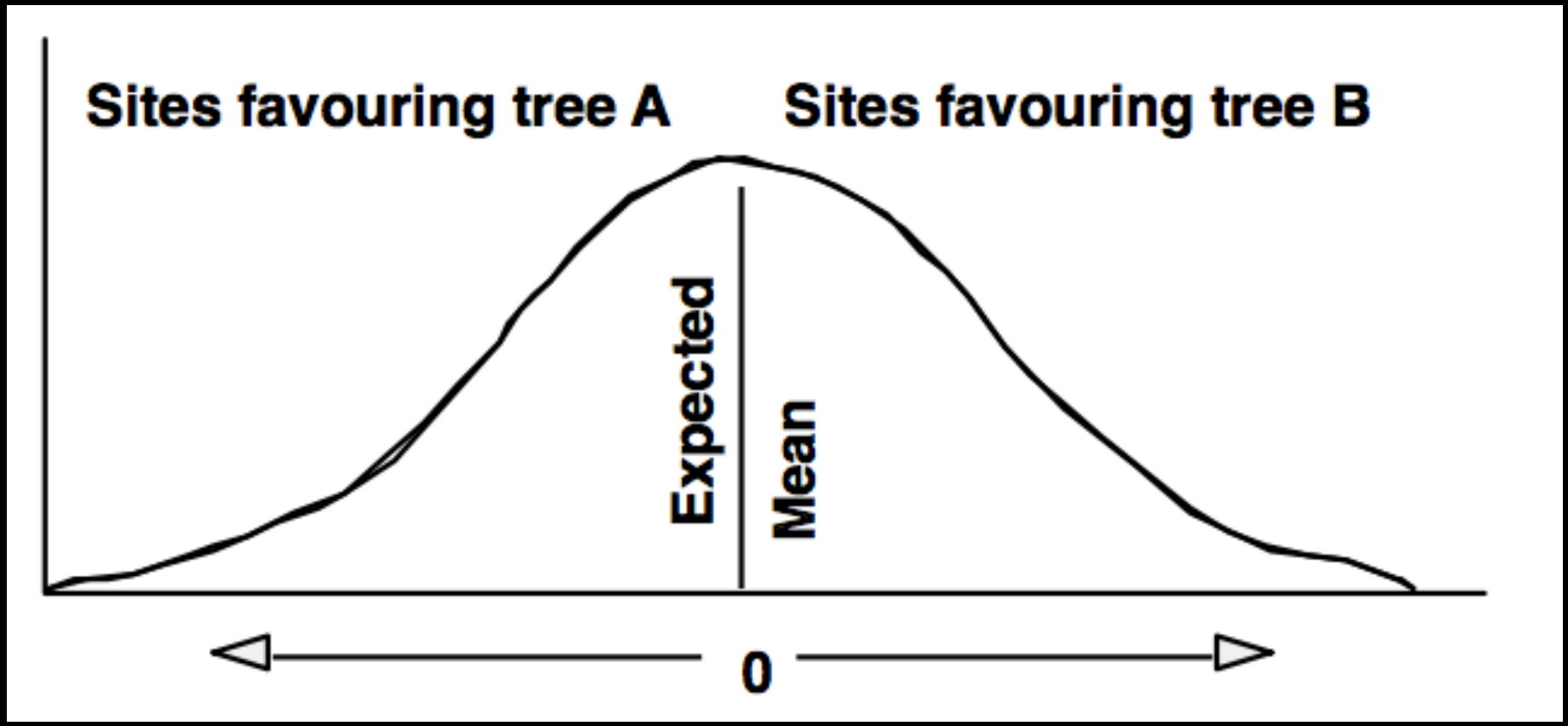
Is the ML tree (right) significantly better than the accepted phylogeny (left)?



$\ln L_{ML} = -5069.85$

Approximately unbiased (AU) test

- Bootstrap based likelihood test



Null hypothesis: Site likelihood difference between two trees is expected to be 0

Deduce gene duplication, loss and HGT

- **Theoretical frame work:**
  - Phylogenetic reconciliation

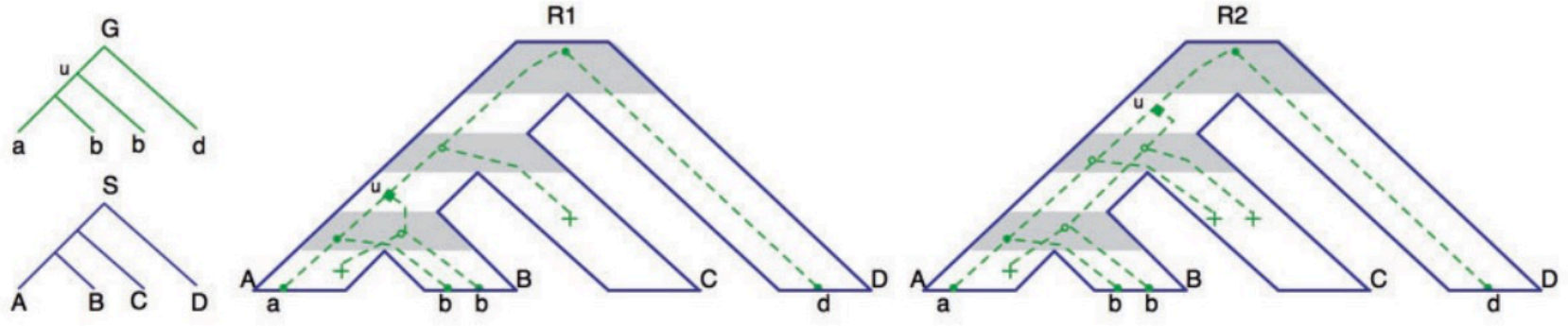


Figure I: Drawing representation of a reconciliation. A species tree S and a gene tree G, where each lower case letter denotes an observed gene of an extant species (gene 'a' belongs to species 'A', etc). RI and R2 are two reconciliations that embed G into S. A grayed zone (resp. tube) corresponds to a vertex (resp. branch) of S and G is depicted with dotted lines. Nodes of the embedded tree represent: duplication (lozenge), loss (+), speciation either present in G (filled circle) or not (open circle). Observe that node *u* of G is a duplication for RI and R2, although located on different branches of S.

BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 5. 392–400

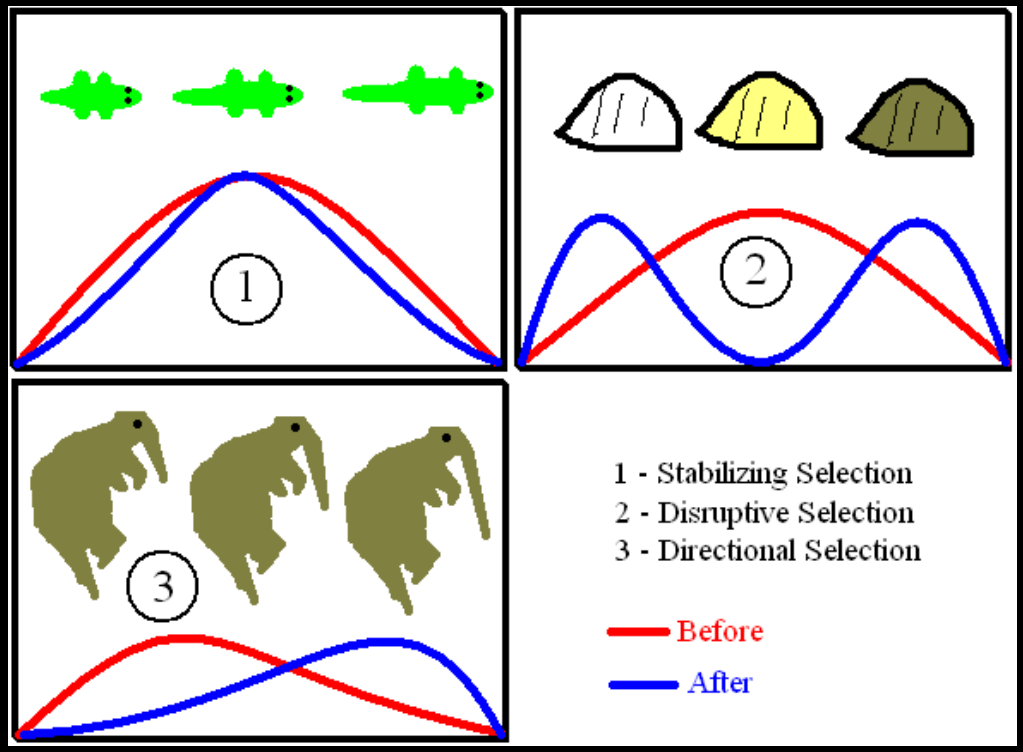
- Software: RANGER-DTL, ANGST, etc

Practice

- Do multiple alignment using protein sequences and convert the alignment into DNA.
- Reconstruct the low divergent tree using NJ method and visualize it (download and install Mega X)
- Filter low quality regions in the alignment in divergent sequences and reconstruct the phylogeny using ML method
- Given one CDS alignment and two topologies, we build the phylogenetic tree using these topologies (estimate branch length and other parameters)
- Compare these two topologies using AU-test and see whether they are significantly different
- Deduce gene duplication, loss and HGT

• Type of natural selection

- Negative directional (purifying/stabilizing) selection
- Positive directional selection
- Balancing (diversifying/disruptive) selection



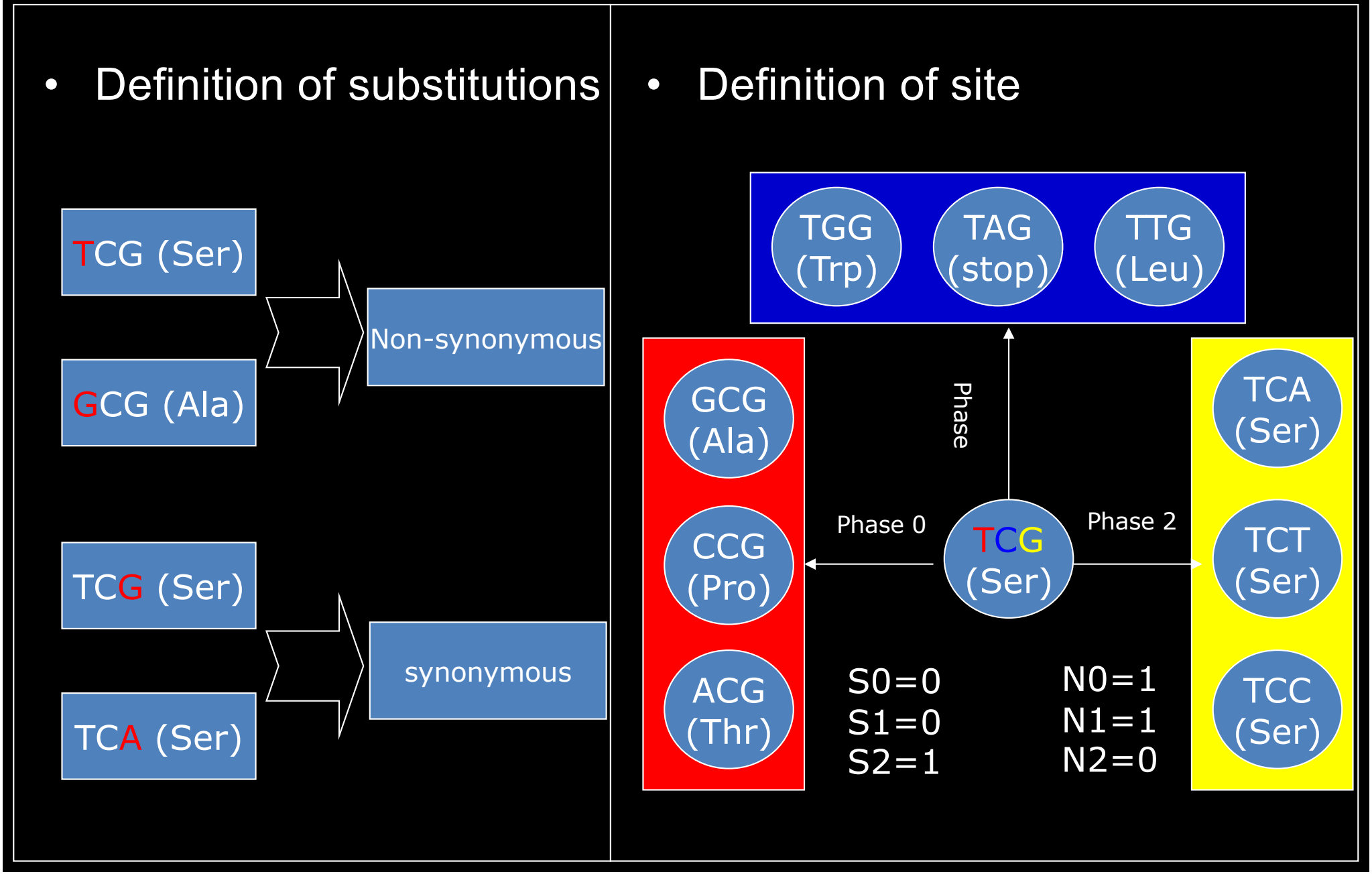
Method to detect positive selection

- Ka/Ks (dN/dS)
- Phylogenetic footprinting;
- McDonald-Kreitman test
- LD and Haplotype structure methods
- Frequency spectrum methods --Tajima's D test, Fu & Li's test, Fst

Comparative genomics or population

Population

• Definition in Ka (dN) and Ks (dS) calculation



dN/dS (ω), a diagnoser for selective pressure

- dN (Ka)– Nonsynonymous substitution rate per nonsynonymous site
- dS (Ks) – Synonymous substitution rate per synonymous site

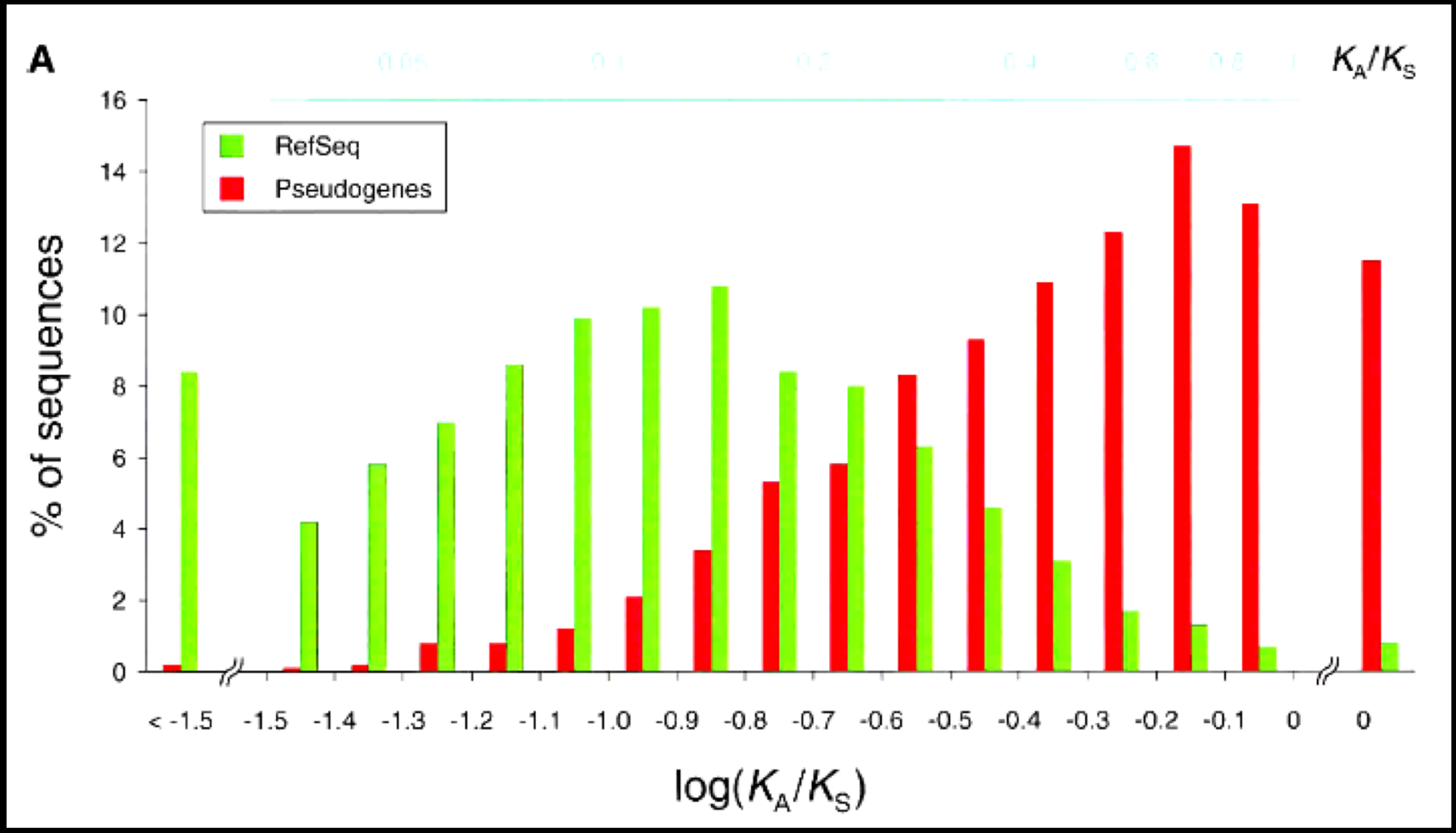
- (I) Diagnosing the form of sequence evolution
- dN/dS < 1, negative selection
- dN/dS ~ 1, neutral evolution
- dN/dS > 1, positive selection

- (II) Estimate the divergence time

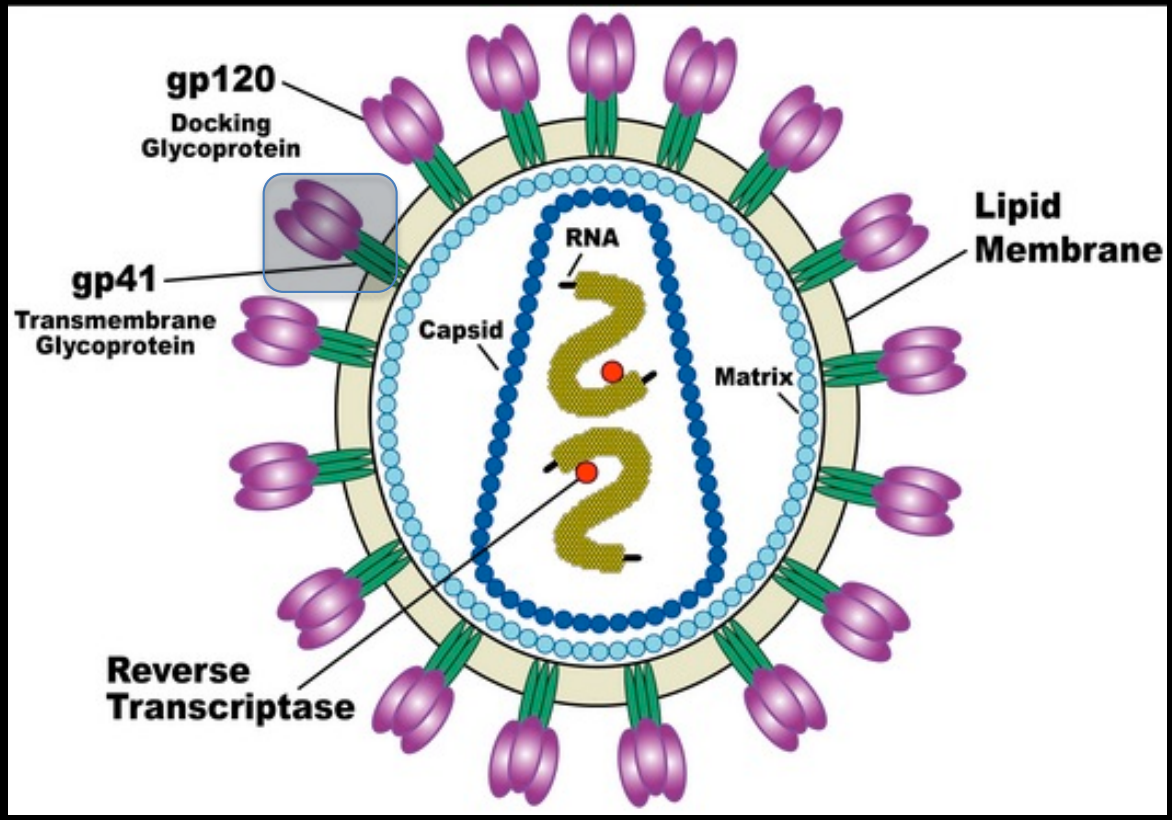
- Ks is approximately neutral distance which could help us estimate the divergence time. Actually, Ka performs better for highly divergent sequences.

- Software:
- PAML
- KaKs_Calculator
- …

dN/dS as a selective sieve

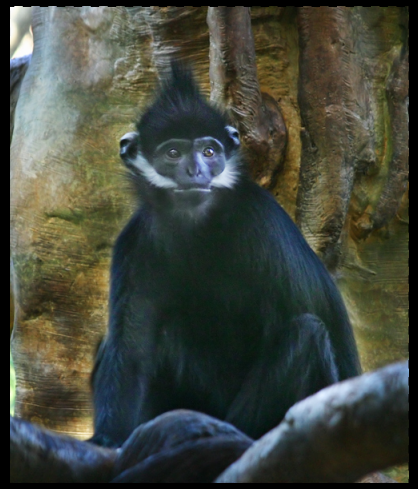


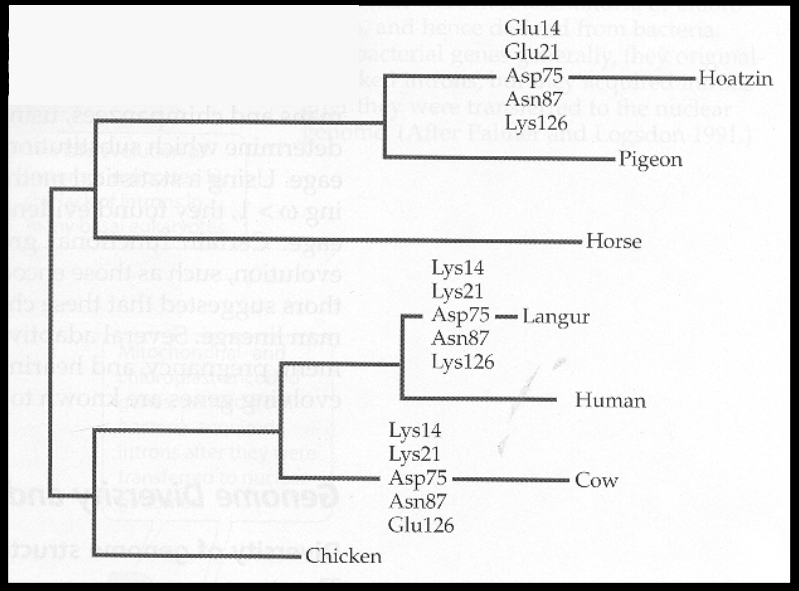
Adaptive evolution in HIV envelope gene



• Lysozyme enzyme

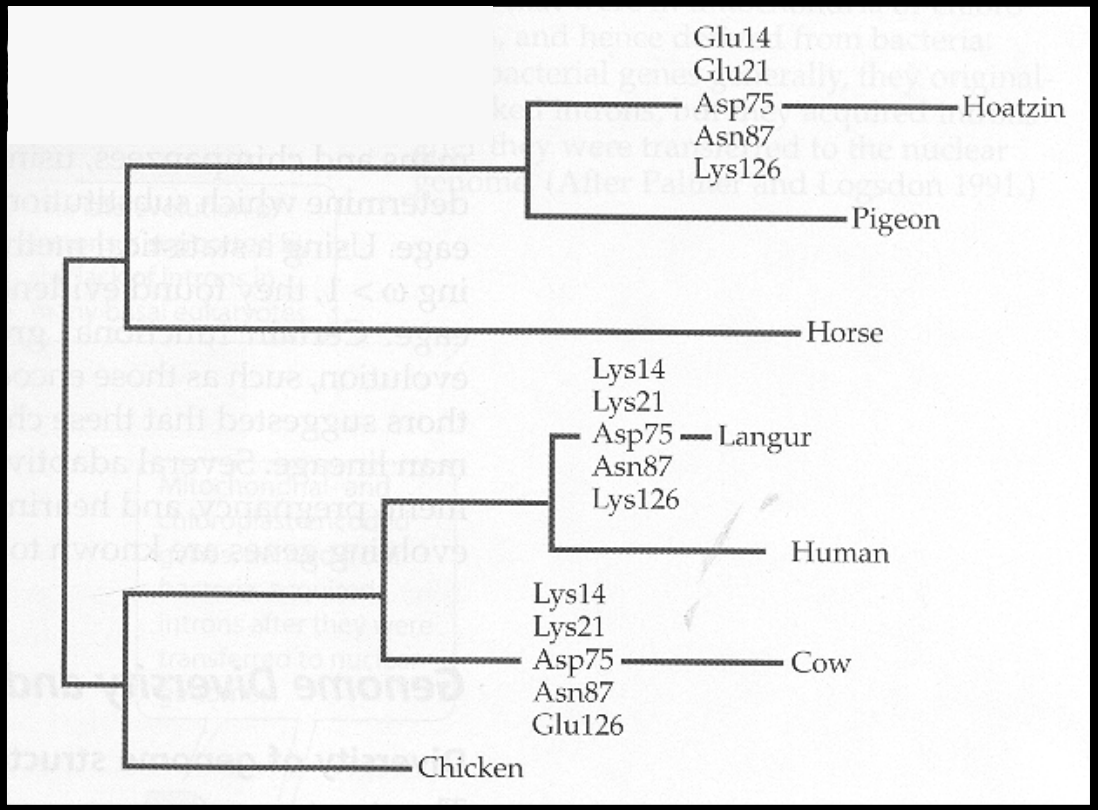
- Breaks down bacterial cell walls
- Ruminants, colobine monkeys, and the hoatzin all have high levels of cellulose-digesting bacteria in the gut
- All three lineages exhibit accelerated nonsynonymous evolution of the lysozyme gene, and convergent amino acid sequences





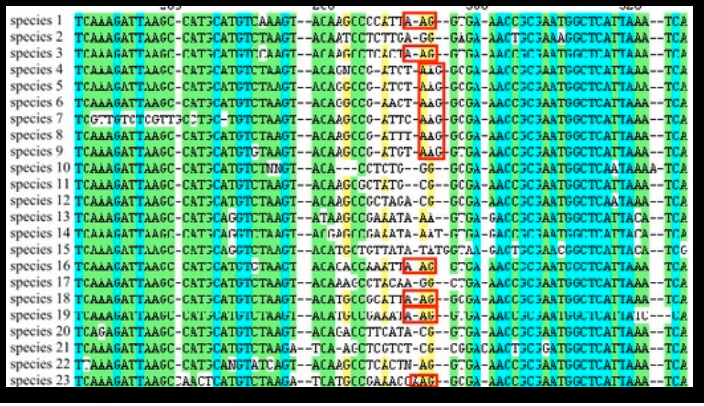
Models for dN/dS (ω) estimation

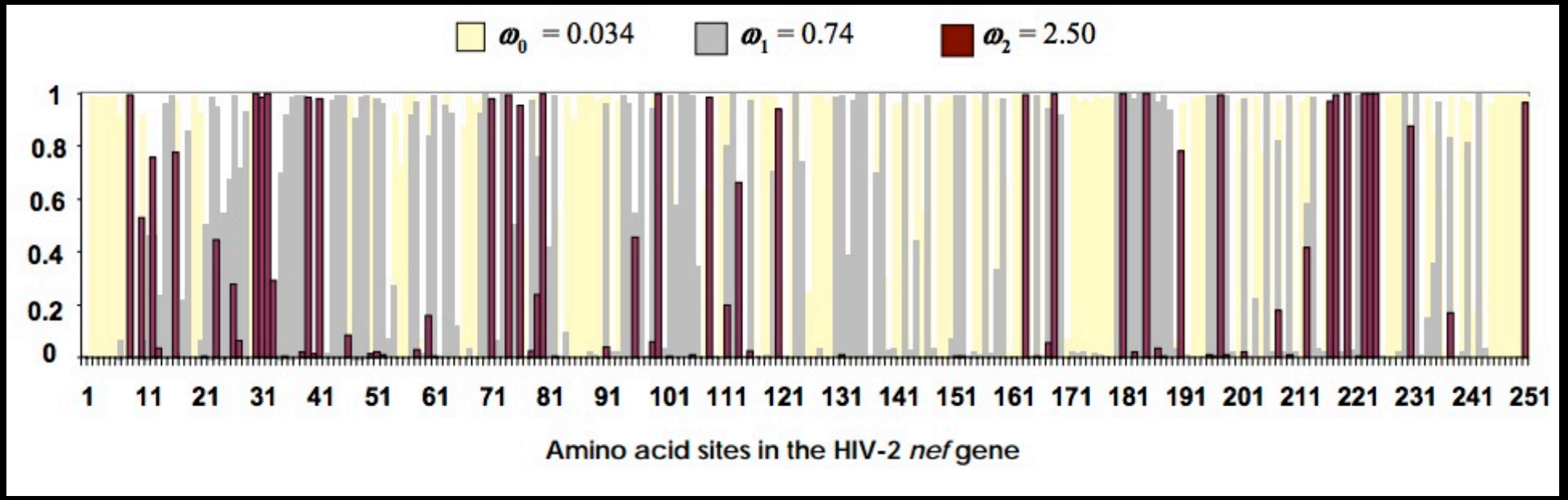
- Universal dN/dS in a phylogenetic tree
- Site model
- Branch model
- Branch-site model



Site model for dN/dS estimation

- 3 category of sites (4 free parameters)
  - $\Omega < 1$, Proportion p_0
  - $\Omega = 1$, Proportion p_1
  - $\Omega > 1$, Proportion p_2





Results from site model (HIV)

Bayes Empirical Bayes (BEB) analysis (Yang, Wong & Nielsen 2005. Mol. Biol. Evol. 22:1107-1118)
Positively selected sites (*: P>95%; **: P>99%)
(amino acids refer to 1st sequence: U68496)

	Pr(w>1)	post mean +- SE for w
1 V	0.548	2.293 +- 1.292
9 S	0.750	2.975 +- 1.428
22 S	0.812	3.142 +- 1.329
24 E	0.607	2.453 +- 1.297
26 N	0.904	3.408 +- 1.184
28 T	0.999**	3.729 +- 1.024
31 N	0.604	2.481 +- 1.356
39 H	0.668	2.674 +- 1.364
40 F	0.512	2.212 +- 1.318
51 I	0.872	3.277 +- 1.190
66 E	0.998**	3.727 +- 1.026
68 N	0.632	2.547 +- 1.334
69 N	0.825	3.133 +- 1.247
76 E	0.686	2.695 +- 1.313
83 I	0.808	3.078 +- 1.263
87 V	0.987*	3.696 +- 1.062

Result of dNdS from branch model

```
kappa (ts/tv) =  4.56120

w (dN/dS) for branches:  0.68581 3.50573

dN & dS for each branch

 branch          t       N       S   dN/dS      dN      dS  N*dN  S*dS

   8..9      0.070   282.2   107.8  0.6858  0.0207  0.0302   5.8   3.3
   9..1      0.026   282.2   107.8  0.6858  0.0076  0.0110   2.1   1.2
   9..2      0.039   282.2   107.8  0.6858  0.0115  0.0168   3.2   1.8
   8..10     0.044   282.2   107.8  0.6858  0.0130  0.0189   3.7   2.0
  10..11     0.079   282.2   107.8  3.5057  0.0328  0.0094   9.3   1.0
  11..3      0.044   282.2   107.8  0.6858  0.0130  0.0189   3.7   2.0
  11..4      0.052   282.2   107.8  0.6858  0.0154  0.0225   4.4   2.4
  10..5      0.019   282.2   107.8  0.6858  0.0058  0.0084   1.6   0.9
   8..12     0.121   282.2   107.8  0.6858  0.0359  0.0523  10.1   5.6
  12..6      0.041   282.2   107.8  0.6858  0.0121  0.0177   3.4   1.9
  12..7      0.024   282.2   107.8  0.6858  0.0070  0.0103   2.0   1.1

tree length for dN:       0.1749
tree length for dS:       0.2164

dS tree:
((Hsa_Human: 0.011030, Hla_gibbon: 0.016794): 0.030198, ((Cgu/Can_colobus: 0.018929, Pne_langur: 0.022497): 0.009367, Mmu_rhesus: 0.008406): 0.018
929, (Ssc_squirrelM: 0.017701, Cja_marmoset: 0.010258): 0.052331);
dN tree:
((Hsa_Human: 0.007565, Hla_gibbon: 0.011517): 0.020710, ((Cgu/Can_colobus: 0.012981, Pne_langur: 0.015429): 0.032838, Mmu_rhesus: 0.005765): 0.012
982, (Ssc_squirrelM: 0.012140, Cja_marmoset: 0.007035): 0.035889);

w ratios as labels for TreeView:
((Hsa_Human #0.6858 , Hla_gibbon #0.6858 ) #0.6858 , ((Cgu/Can_colobus #0.6858 , Pne_langur #0.6858 ) #3.5057 , Mmu_rhesus #0.6858 ) #0.6858 , (Ss
c_squirrelM #0.6858 , Cja_marmoset #0.6858 ) #0.6858 );
```

Practice

- Calculate global selective pressure on paired sequences using KaKs_Calculator
- Deduce the positive selected sites on HIV envelop gene (site model)
- Calculate branch-specific selective pressure on lysozyme data (branch model)