Functional and Comparative Genomics

ESCI upf.
School of International Studies

Comparative and Functional Genomics

Session 3

ESCI upf.
School of International Studies

Phylogenetic analysis

Gene family tree reconstruction.

Inference of gene duplication and other evolutionary events.

Detection of functional divergence.

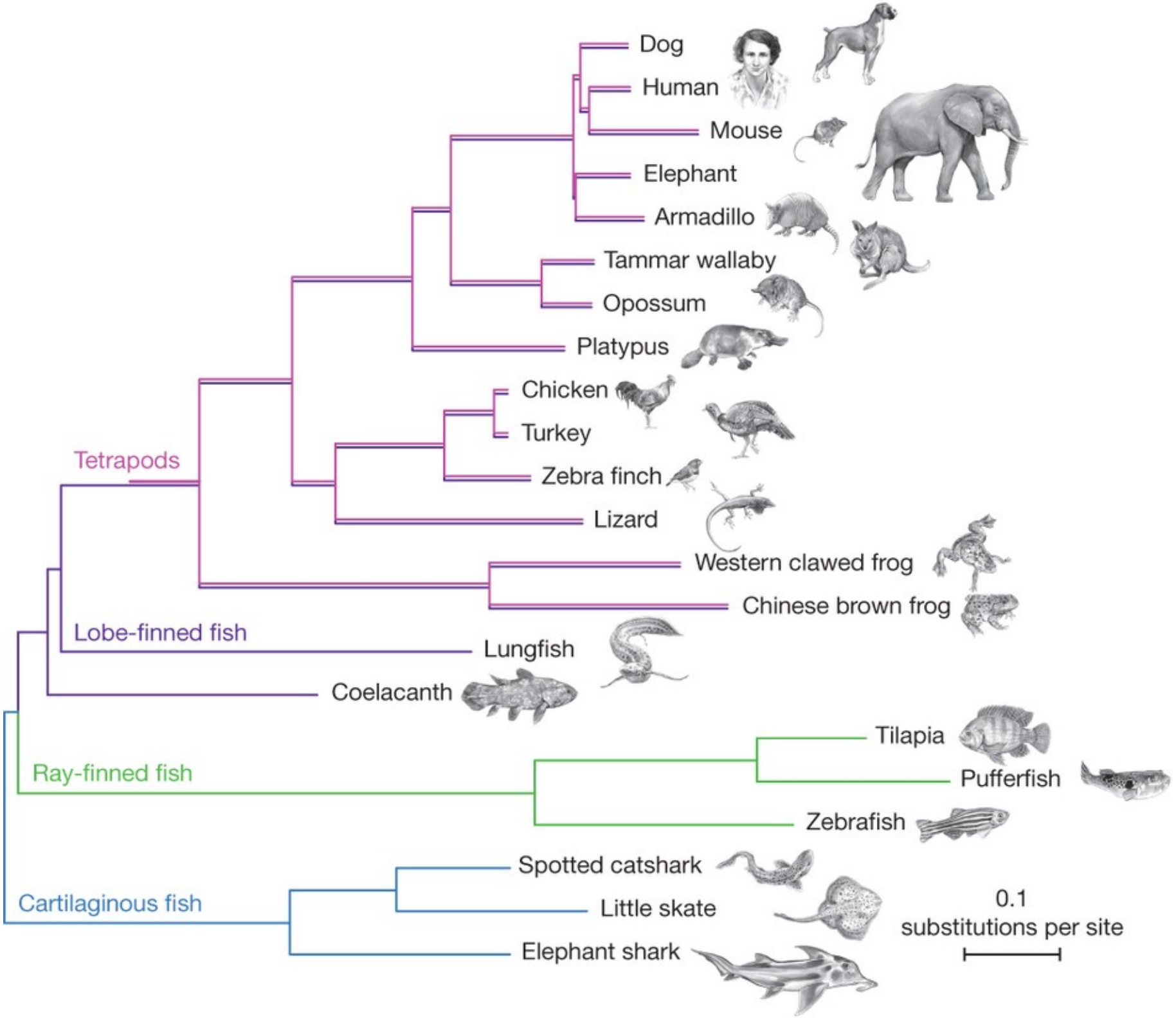
dn/ds analysis.

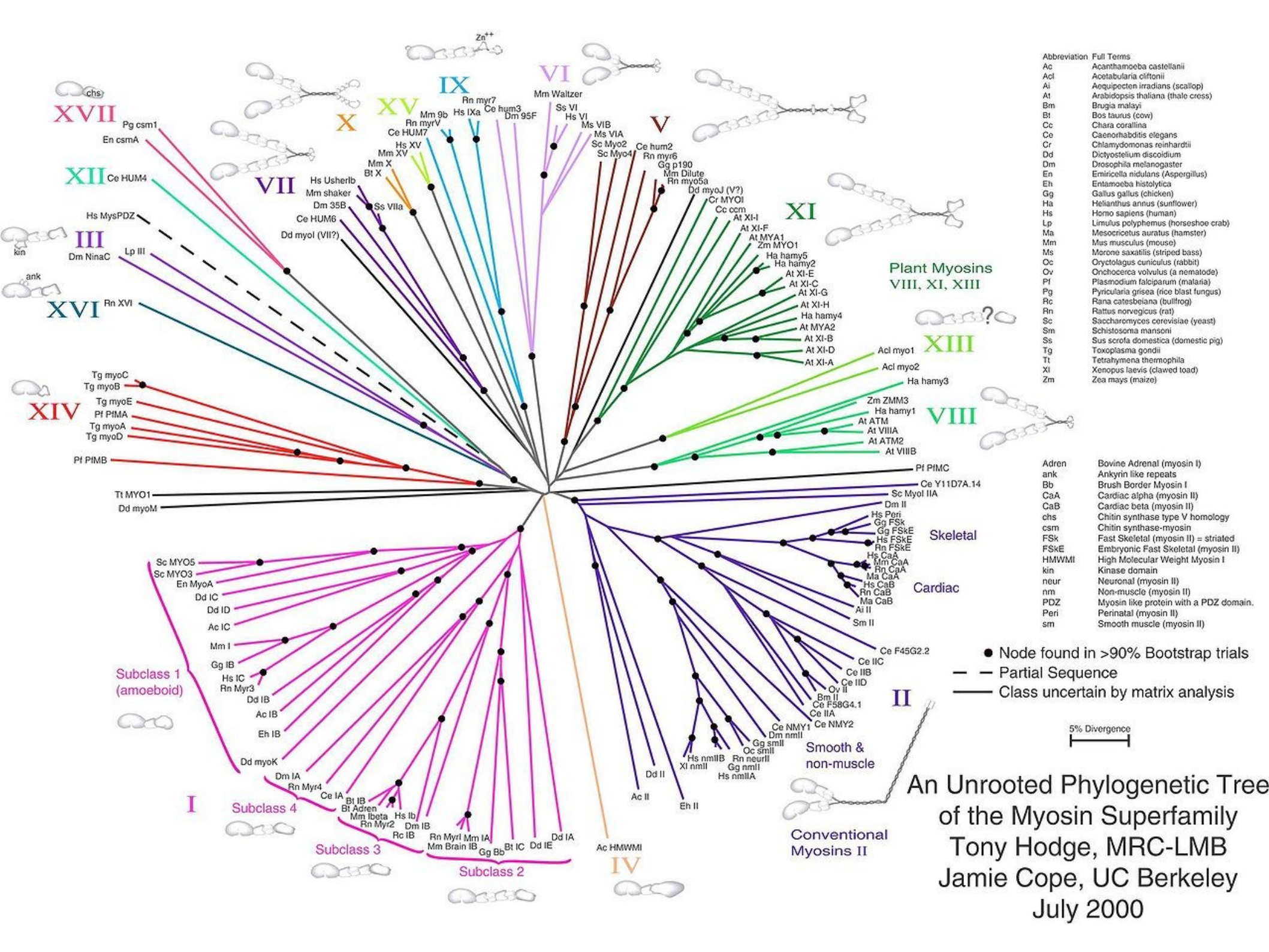
A phylogenetic tree

A branching diagram (**bipartite graph**) showing the inferred **evolutionary** relationships among various **biological species or other entities** (e.g sequences) **based on similarities and differences** in their physical and/or genetic characteristics.

Species tree

Gene tree


Dog
Human
Mouse
Elephant
Armadillo
Tammar wallaby
Opossum
Platypus
Chicken
Turkey
Zebra finch
Lizard
Western clawed frog
Chinese brown frog
Tetrapods
Lobe-finned fish
Lungfish
Coelacanth
Tilapia
Pufferfish
Zebrafish
Ray-finned fish
Spotted catshark
Little skate
Elephant shark
Cartilaginous fish
0.1
substitutions per site



Abbreviation	Full Terms
Ac	Acanthamoeba castellanii
Acl	Acetabularia cliftonii
Ai	Aequipecten irradians (scallop)
At	Arabidopsis thaliana (thale cress)
Bm	Brugia malayi
Bt	Bos taurus (cow)
Cc	Chara corallina
Ce	Caenorhabditis elegans
Cr	Chlamydomonas reinhardtii
Dd	Dictyostelium discoidium
Dm	Drosophila melanogaster
En	Emiricella nidulans (Aspergillus)
Eh	Entamoeba histolytica
Gg	Gallus gallus (chicken)
Ha	Helianthus annus (sunflower)
Hs	Homo sapiens (human)
Lp	Limulus polyphemus (horseshoe crab)
Ma	Mesocricetus auratus (hamster)
Mm	Mus musculus (mouse)
Ms	Morone saxatilis (striped bass)
Oc	Oryctolagus cuniculus (rabbit)
Ov	Onchocerca volvulus (a nematode)
Pf	Plasmodium falciparum (malaria)
Pg	Pyricularia grisea (rice blast fungus)
Rc	Rana catesbeiana (bullfrog)
Rn	Rattus norvegicus (rat)
Sc	Saccharomyces cerevisiae (yeast)
Sm	Schistosoma mansoni
Ss	Sus scrofa domestica (domestic pig)
Tg	Toxoplasma gondii
Tt	Tetrahymena thermophila
Xl	Xenopus laevis (clawed toad)
Zm	Zea mays (maize)

Adren	Bovine Adrenal (myosin I)
ank	Ankyrin like repeats
Bb	Brush Border Myosin I
CaA	Cardiac alpha (myosin II)
CaB	Cardiac beta (myosin II)
chs	Chitin synthase type V homology
csm	Chitin synthase-myosin
FSk	Fast Skeletal (myosin II) = striated
FSkE	Embryonic Fast Skeletal (myosin II)
HMWMI	High Molecular Weight Myosin I
kin	Kinase domain
neur	Neuronal (myosin II)
nm	Non-muscle (myosin II)
PDZ	Myosin like protein with a PDZ domain.
Peri	Perinatal (myosin II)
sm	Smooth muscle (myosin II)

● Node found in >90% Bootstrap trials

– – Partial Sequence

—— Class uncertain by matrix analysis

5% Divergence

An Unrooted Phylogenetic Tree of the Myosin Superfamily

Tony Hodge, MRC-LMB
Jamie Cope, UC Berkeley

July 2000

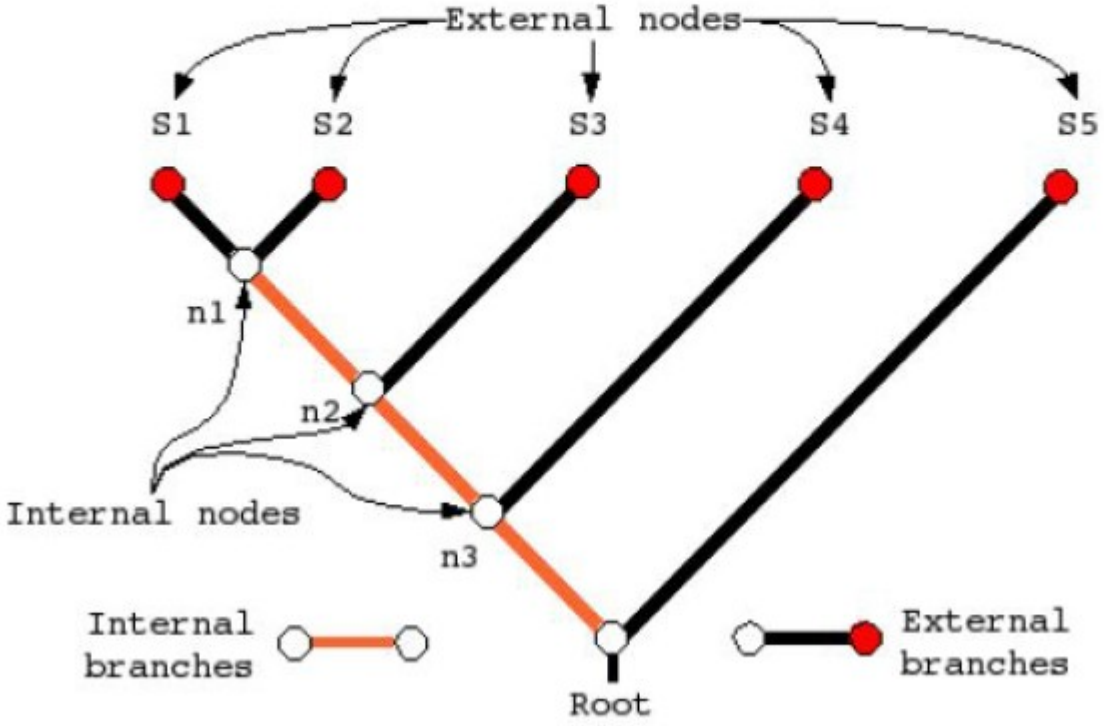
Why care about (phylogenetic) trees?

Biological systems are the result of the evolutionary process. Trees represent this process.

Assumptions on statistical tests usually assume “independence of data”, but no data in comparative biology is “independent”, as all organisms are related. Phylogenetic trees are the first step to detect and remove the effect of evolutionary relatedness.

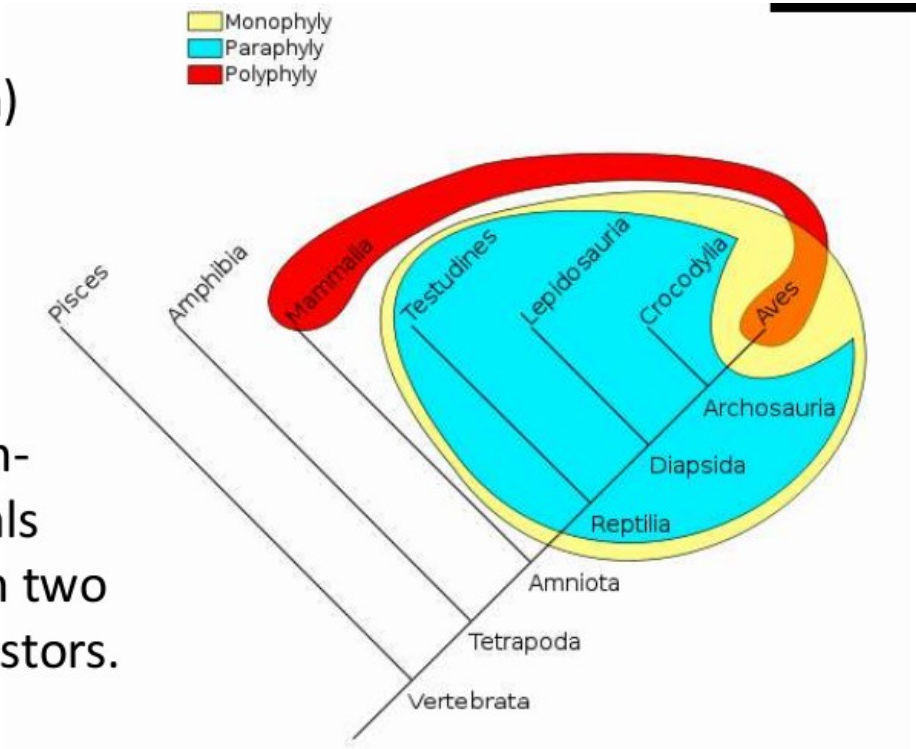
“Nothing make sense in biology if not in the light of evolution”

- **Nodes & branches.** Trees contain internal and external nodes and branches. In molecular phylogenetics, **external nodes** are sequences representing **genes, populations or species!**. Sometimes, **internal nodes** contain the ancestral information of the clustered species. A **branch** defines the relationship between sequences in terms of descent and ancestry.

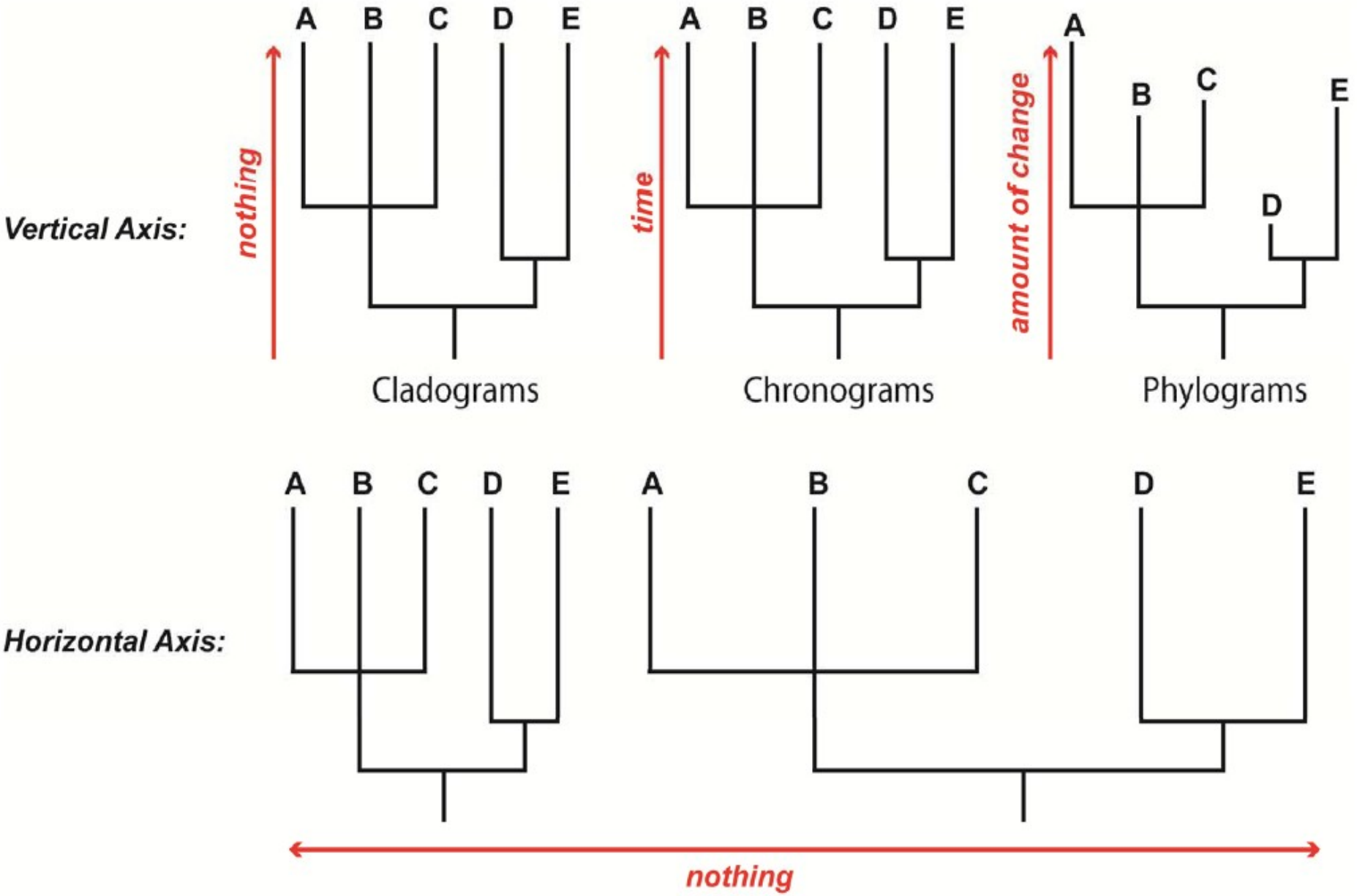


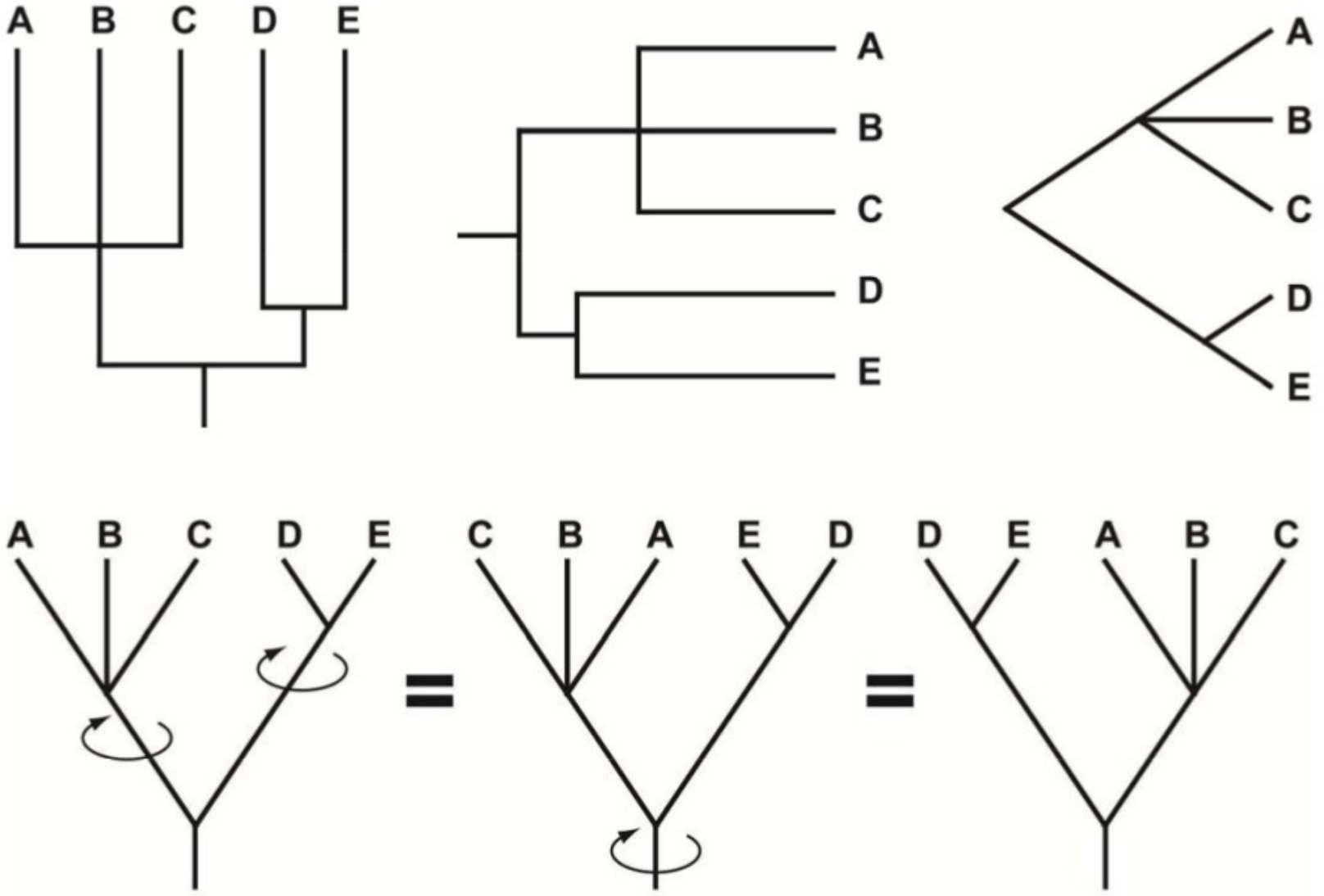
Paraphyletic group:
modern reptiles (cyan) contain not all descendants from a common ancestor.

Polyphyletic group: warm-blood animals (mammals and birds; red) are from two different common ancestors.



Monophyletic group (clade): birds and reptiles evolved from one common ancestor. (yellow)


Vertical Axis:
nothing
A B C D E
Cladograms
time
A B C D E
Chronograms
amount of change
A B C D E
Phylograms
Horizontal Axis:
A B C D E
A B C D E
nothing

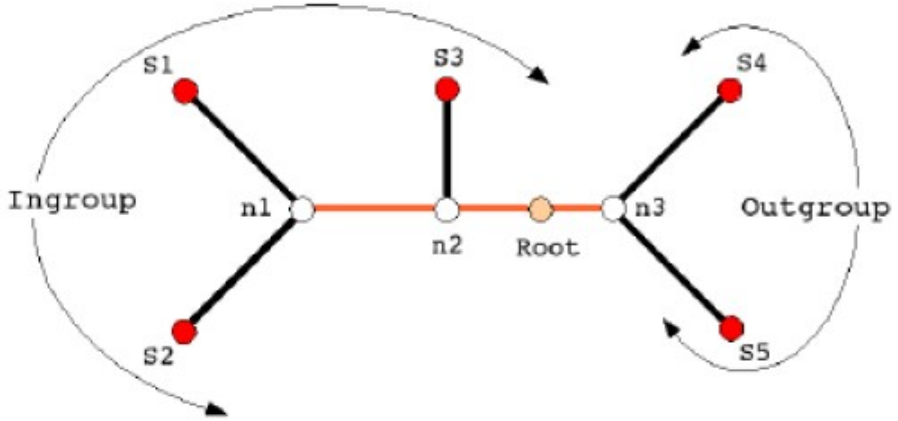


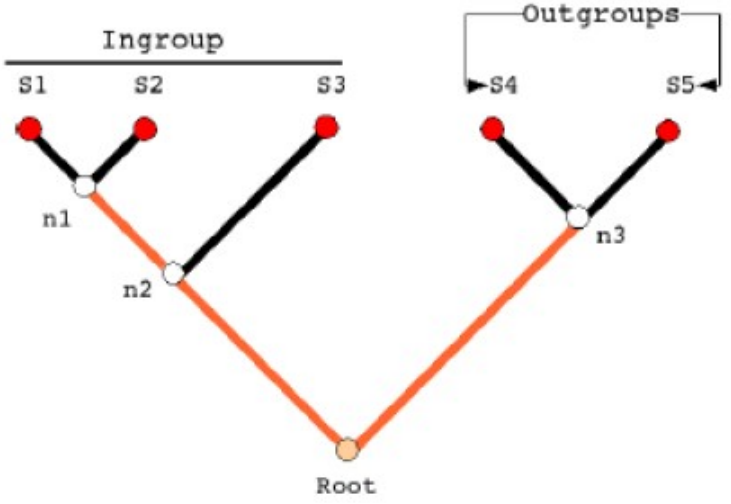
Newick standard: ((A,B,C),(D,E));

Trees can be rooted or un-rooted. Several strategies can be used for rooting (midpoint, outrgroup, etc)

Number of rooted trees with n OTUs: $(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!}$, for $n \geq 2$

Number of unrooted trees: $(2n-5)!! = \frac{(2n-5)!}{2^{n-3}(n-3)!}$, for $n \geq 3$





How do we reconstruct phylogenetic trees?

Phylogenetic trees can be based on anything that can tell us on similarities and differences. (e.g RFLPs, phenotypic characters, sequences).

Molecular Phylogenetics is nowadays the most widely used method

Sequences are compared by means of alignments and this constitutes the information used to make the tree.

→ quality of the tree depends on the quality of the underlying alignment

So, how to find the best tree?

Exhaustive search: make ALL trees first, and then see which one best fits the data (you need an optimality criterion)

Heuritisc search: Try to find a way to find an optimal tree (hopefully the best) without testing them all. You also need an optimality criterion and you are not guaranteed to find the best, but you save time.

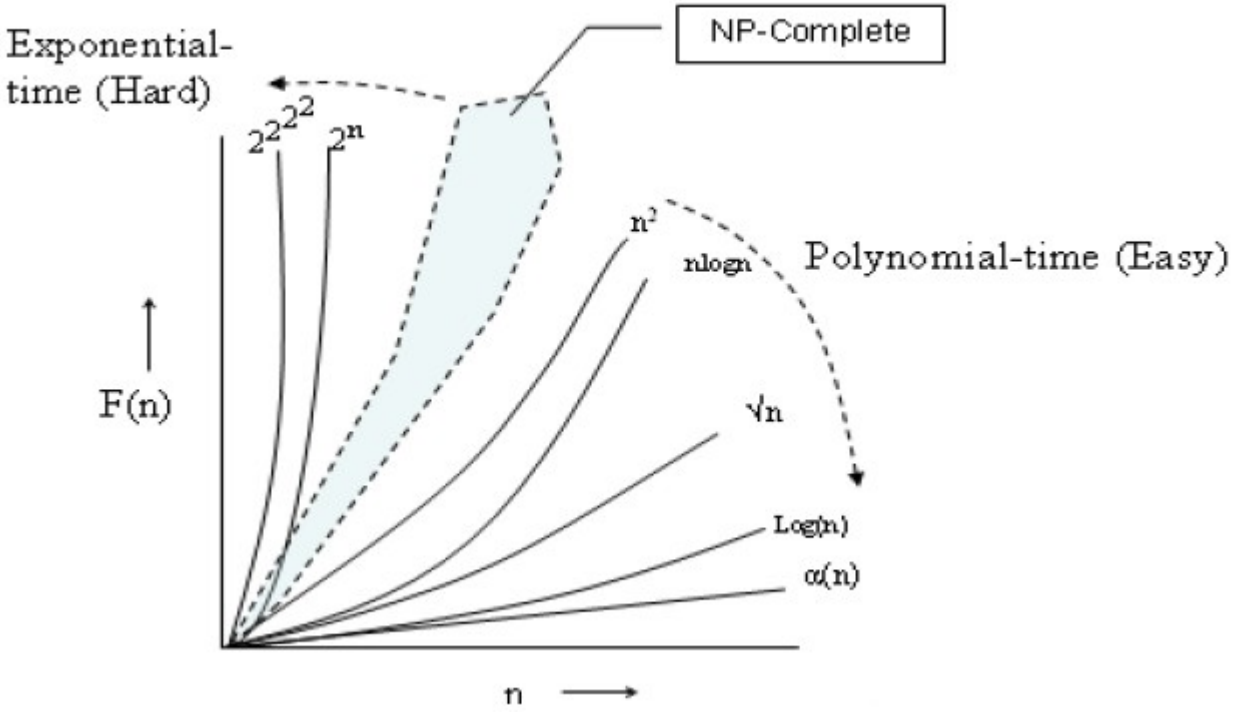
Number of taxa *T*	Number of unrooted bifurcating trees *B(T)*
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
22	3×10^{23}
50	3×10^{74}

Phylogenetic approaches:

Distance methods (NJ, UPGMA)

Maximum Parsiomy

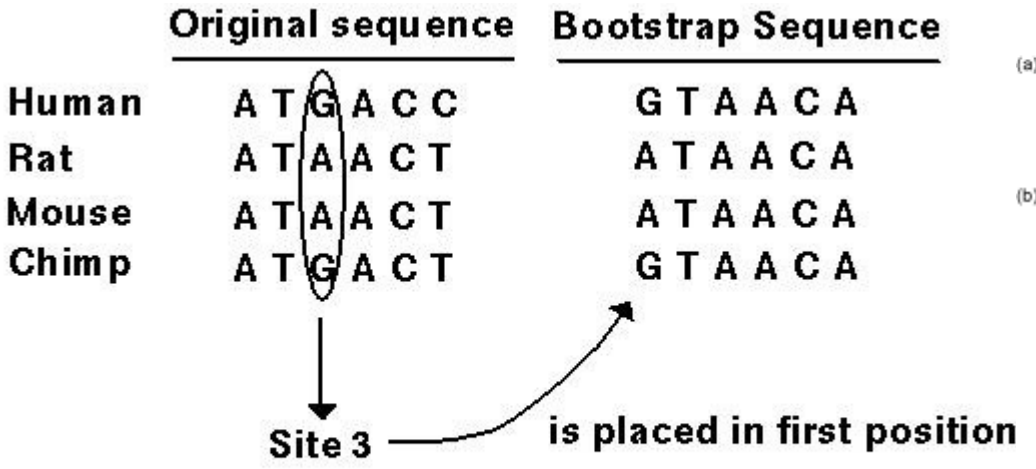
Probabilistic Methods (Maximum Likelihood and Bayesian Inference)



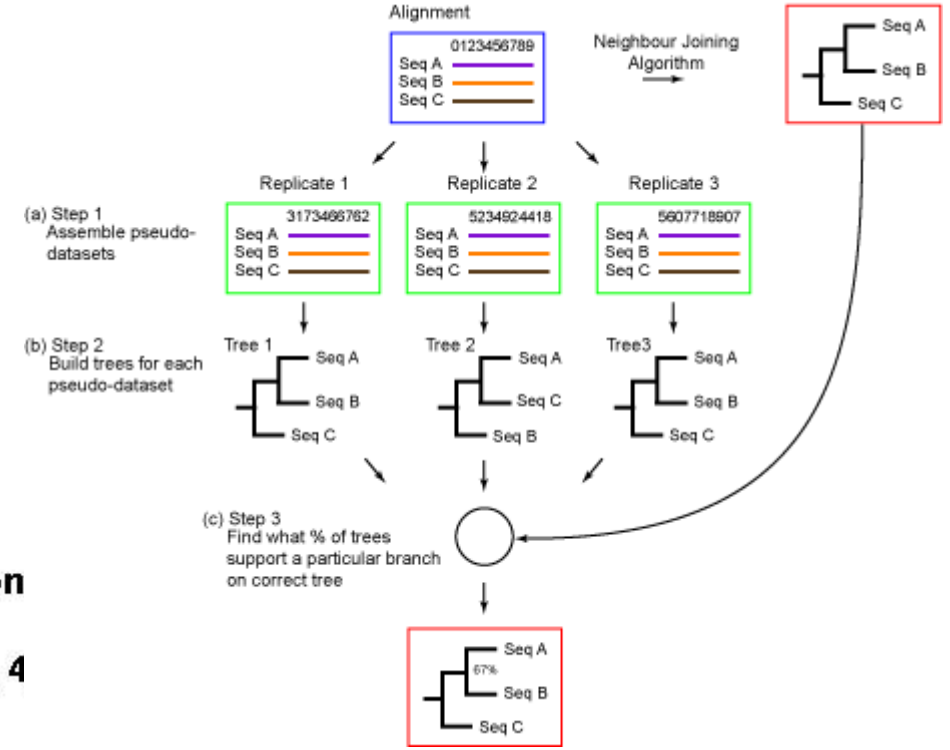
Distance-based methods: if there are no errors (in distances)then, the correct tree can be obtained in polynomial time. Otherwise, optimization problems are NP-hard.

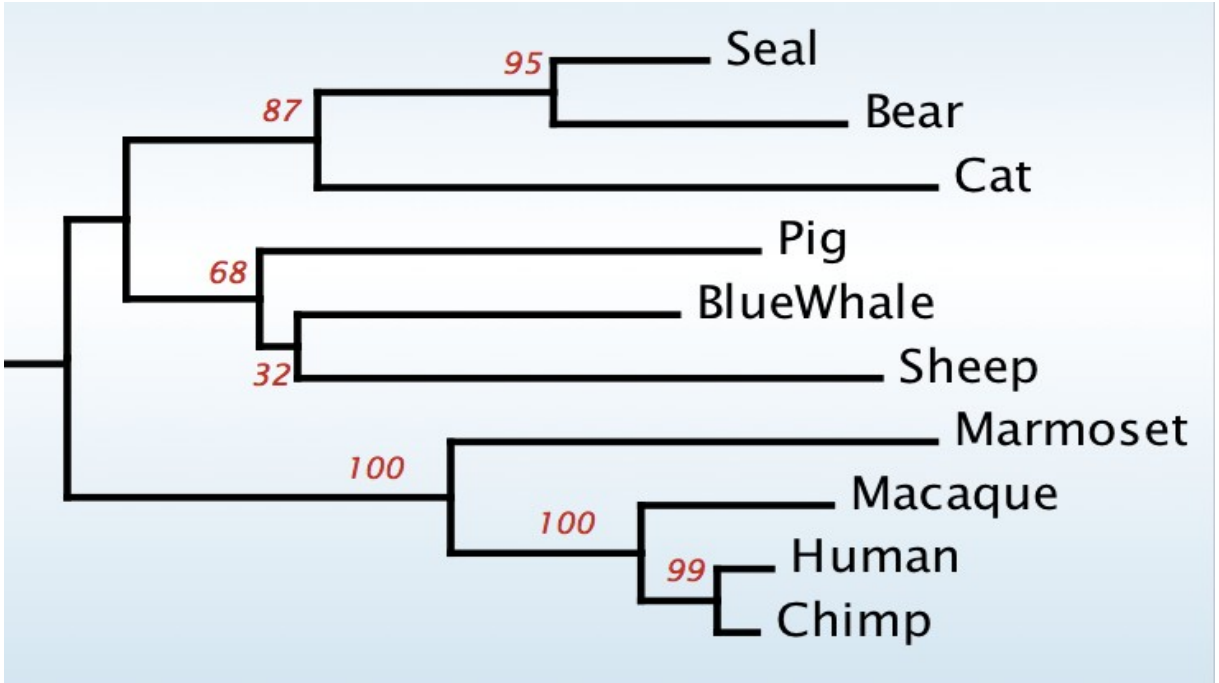
Maximum Parsimony, Probabilistic Methods: NP-hard

Confidence values: bootstrapping



(Then the next five randomly chosen sites: 2, 1, 1, 5, 4 are placed in the next five positions.)



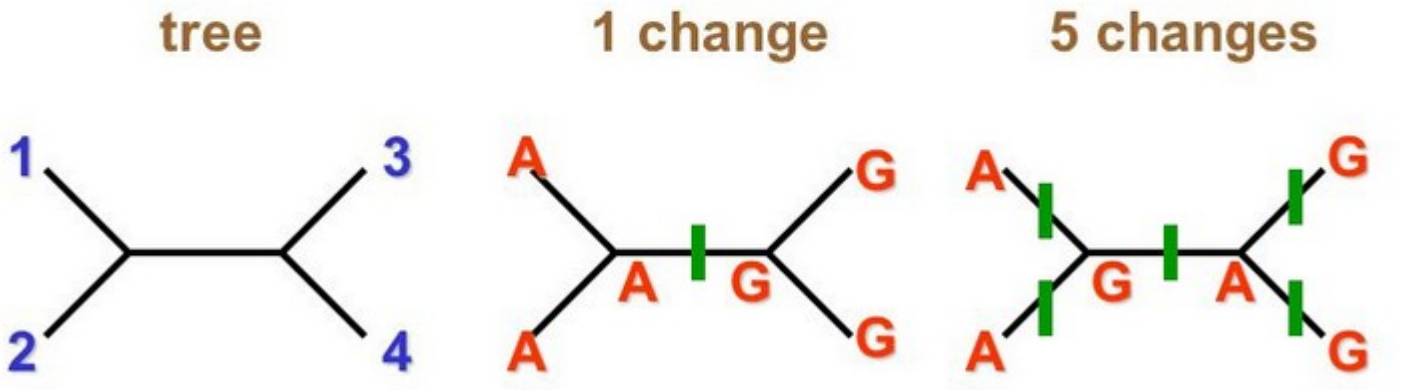


Maximum Parsimony

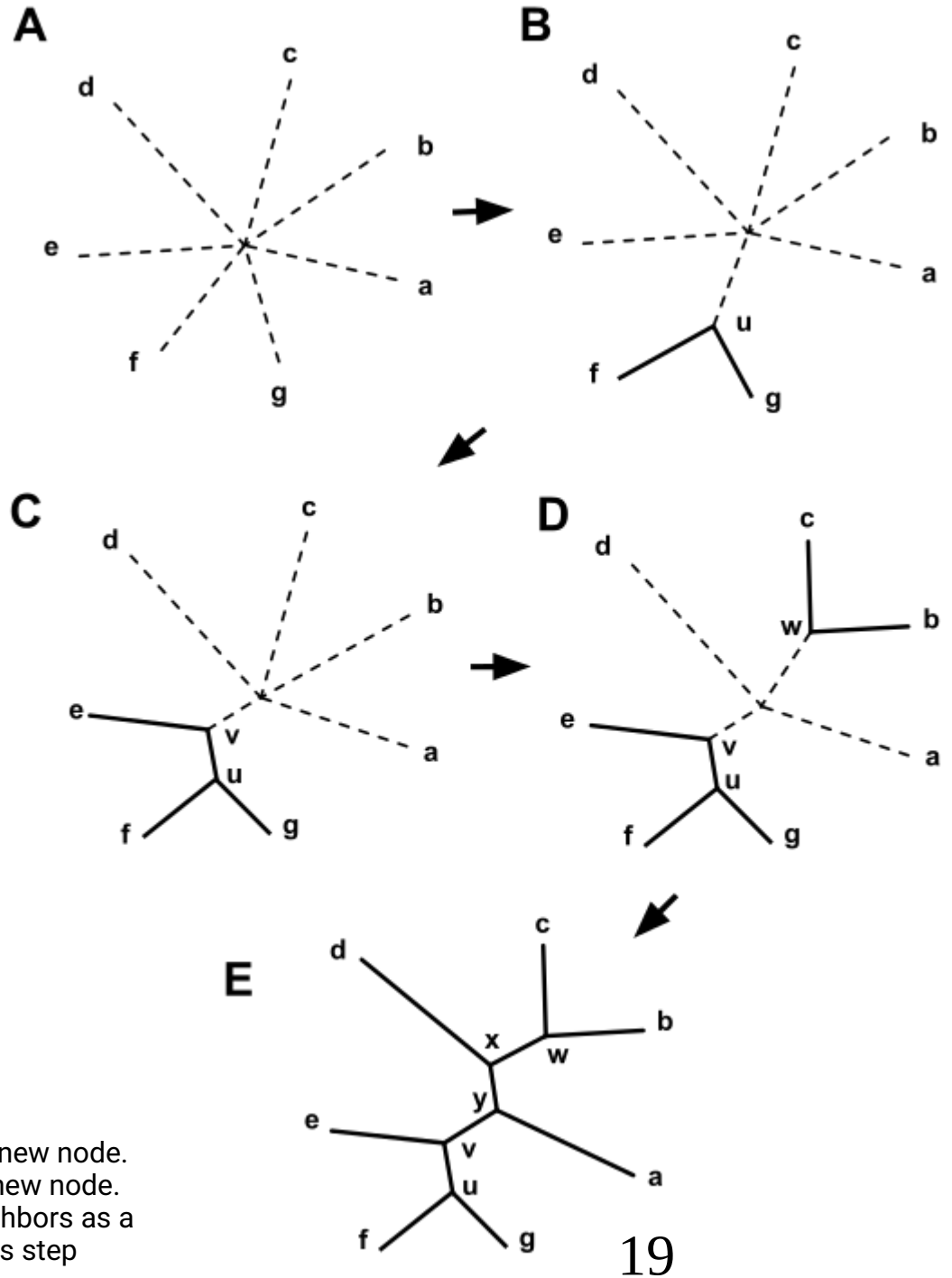
Finding the tree with that implies the minimal number of changes along its branches.

Taxon-1	ATATT
Taxon-2	ATCGT
Taxon-3	GCAGT
Taxon-4	GCCGT

For each site, the goal is to reconstruct the evolution of that site on a tree subject to the constraint of invoking the fewest possible evolutionary changes.



Neighbor Joining



Based on the current distances matrix calculate the matrix Q
2. Find the pair of taxa in Q with the lowest value.
Create a node on the tree that joins these closest neighbors.
3. Calculate the distance of each of the taxa in the pair to this new node.
4. Calculate the distance of all taxa outside of this pair to the new node.
5. Start the algorithm again, considering the pair of joined neighbors as a single taxon and using the distances calculated in the previous step

10. Statistical Methods

10.1. Maximum Likelihood

♣ The phylogenetic methods described **infered the history** (*or the set of histories*) that were **most consistent with a set of observed data**.
All the methods explained used **sequences as data** and give one or more **trees as phylogenetic hypotheses**. Then, they use the logic of:

$$P(H/D)$$

♠ **Maximum Likelihood** (*ML*)[28] methods (*or maximum probability*) computes **the probability of obtaining the data** (*the observed aligned sequences*) **given a defined hypothesis** (*the tree and the model of evolution*). That is:

$$P(D/H)$$

A coin example

The ML estimation of the heads probabilities of a coin that is tossed n times.

[28] ML was invented by Ronal A. Fisher [27]. Likelihood methods for phylogenies were introduced by Edwars and Cavalli-Sforza for gene frequency data [9]. Felsenstein showed how to compute ML for DNA sequences [24].

Likelihood

Given some data **(D)** a decission must be made about an adequate explanation (**H**, hypothesis)

D: alignment

H: Model of evolution, tree topology, branch lengths, parameters of the model

--> Each H will have a certain probability of producing the data P(D/H)

The best **H** is that of the greatest **P**

Important remark!!

The likelihood function **is not** the probability of a hypothesis being correct!!

The likelihood function is defined in terms of probability of producing the observed events not of the unknown parameters

Thus: the probability of observing the data has nothing to do with the probability that the underlying model is correct.

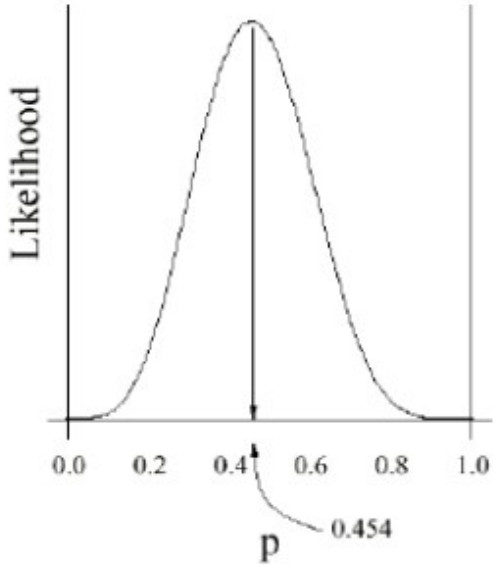
If tosses are all **independent**, and all have the same **unknown heads probability** p, then the observing sequence of tosses:

HHTTHTHHTTT

we can calculate the ML of these data as:

$$L = Prob(D/p) = pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p) = p^5(1-p)^6$$

Ploting L against p, we observe the probabilities of the same data (D) for different values of p.



Thus the ML or the maximum probability to observe the above sequence of events is at $p = 0.4545$,

Suppose we have:

- **Data**:

Sequence 1 **C C A T**

Sequence 2 **C C G T**

- **Model**:[29]

$$\pi = [0.1, 0.4, 0.2, 0.3]$$

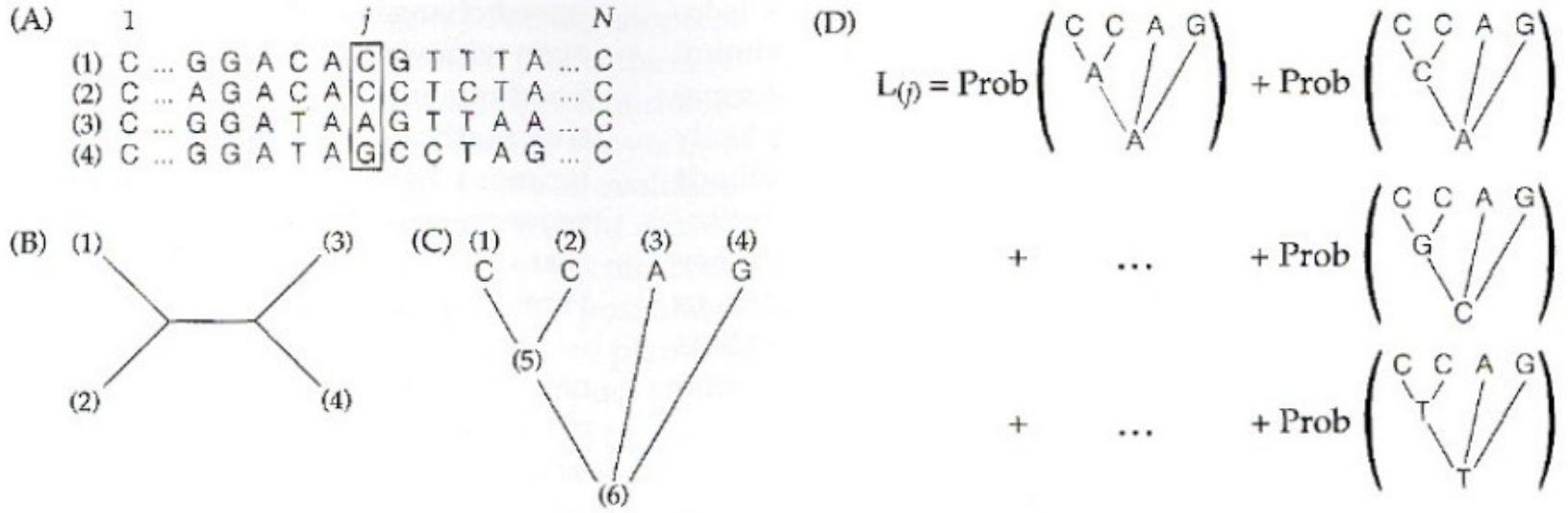
$$\boldsymbol{P} = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}$$

$$L_{(Seq.1 \to Seq.2)} = \pi_C P_{C \to C} \pi_C P_{C \to C} \pi_A P_{A \to G} \pi_T P_{T \to T}$$
$$0.4\text{x}0.983\text{x}0.4\text{x}0.983\text{x}0.1\text{x}0.007\text{x}0.3\text{x}0.979$$
$$= 0.0000300$$

$$lnL_{tree:Seq_1 \to Seq_2} = -10.414$$

[29] Note that the base composition sum one, but indeed the the rows of substitution matrix sum one. Why?

computation in a real problem



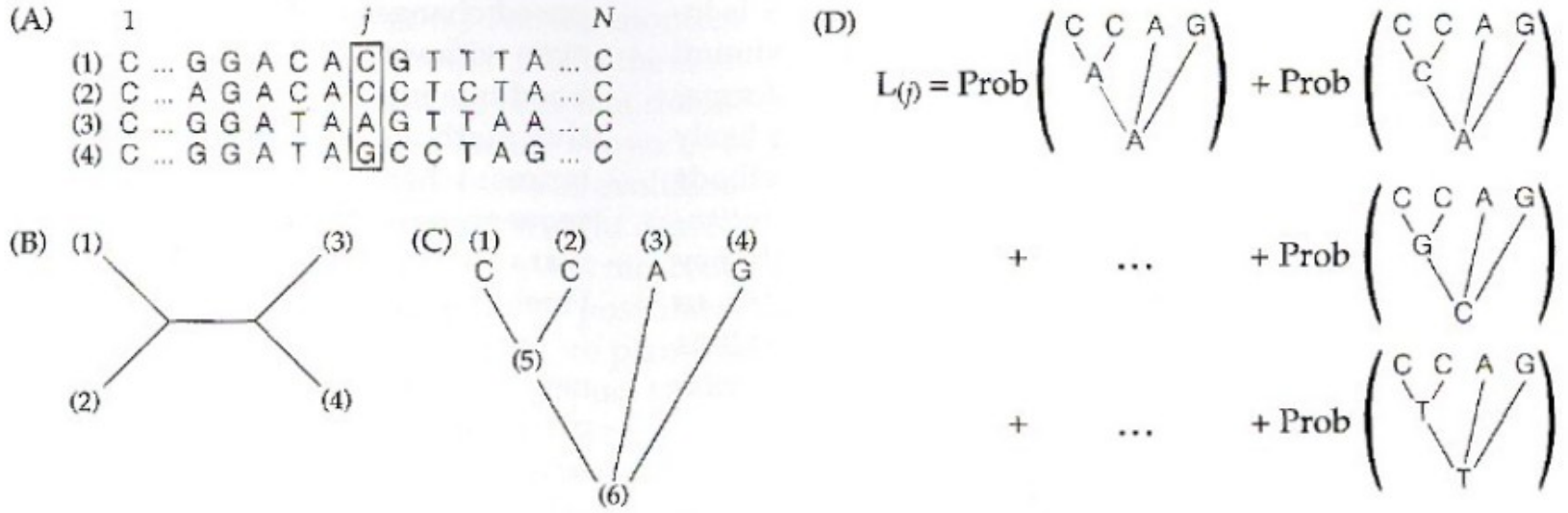
- Tree after rooting in an arbitrary node (reversible model).
- The likelihood for a particular site is the sum of the probabilities of every possible reconstruction of ancestral states given some model of base substitution.
- The likelihood of the tree is the product of the likelihood at each site.

$$L = L_{(1)} \cdot L_{(2)} \cdot \ldots \cdot L_{(N)} = \prod_{j=1}^{N} L_{(j)}$$

- The likelihood is reported as the sum of the log likelihhod of the full tree.

$$lnL = lnL_{(1)} + lnL_{(2)} + \ldots + lnL_{(N)} = \sum_{j=1}^{N} lnL_{(j)}$$

computation in a real problem



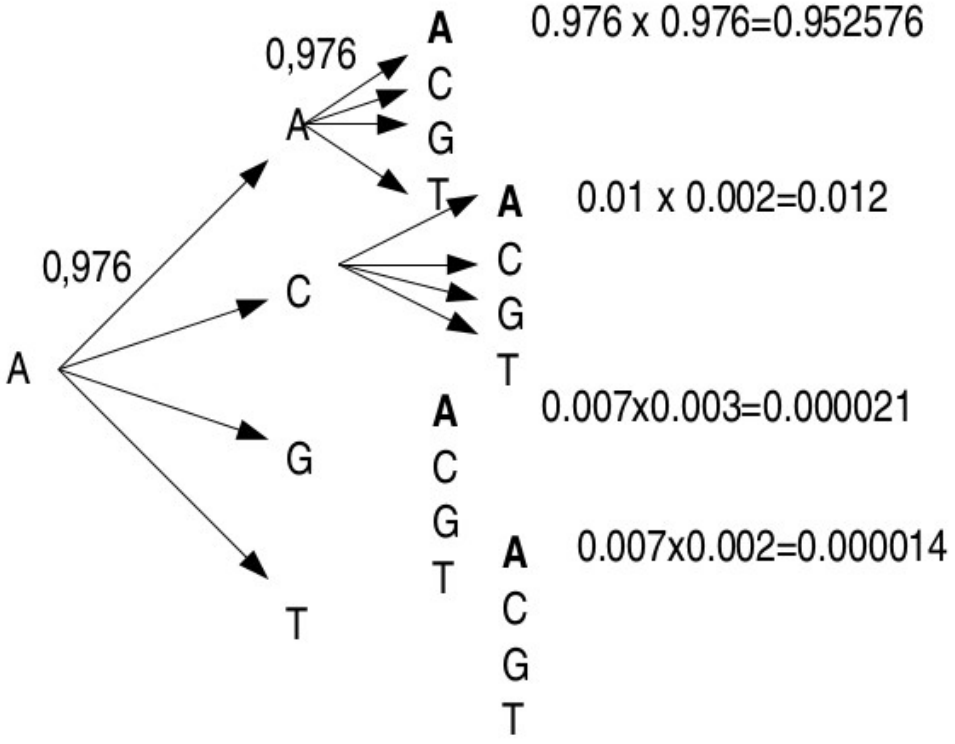
- Tree after rooting in an arbitrary node (reversible model).
- The likelihood for a particular site is the sum of the probabilities of every possible reconstruction of ancestral states given some model of base substitution.
- The likelihood of the tree is the product of the likelihood at each site.

$$L = L_{(1)} \cdot L_{(2)} \cdot \ldots \cdot L_{(N)} = \prod_{j=1}^{N} L_{(j)}$$

- The likelihood is reported as the sum of the log likelihhod of the full tree.

$$lnL = lnL_{(1)} + lnL_{(2)} + \ldots + lnL_{(N)} = \sum_{j=1}^{N} lnL_{(j)}$$

Ancestral= A
Observed= A
Time = 2x CED



P= A -> A = 0.976 (CED=1), 0.9646 (CED=2),.....

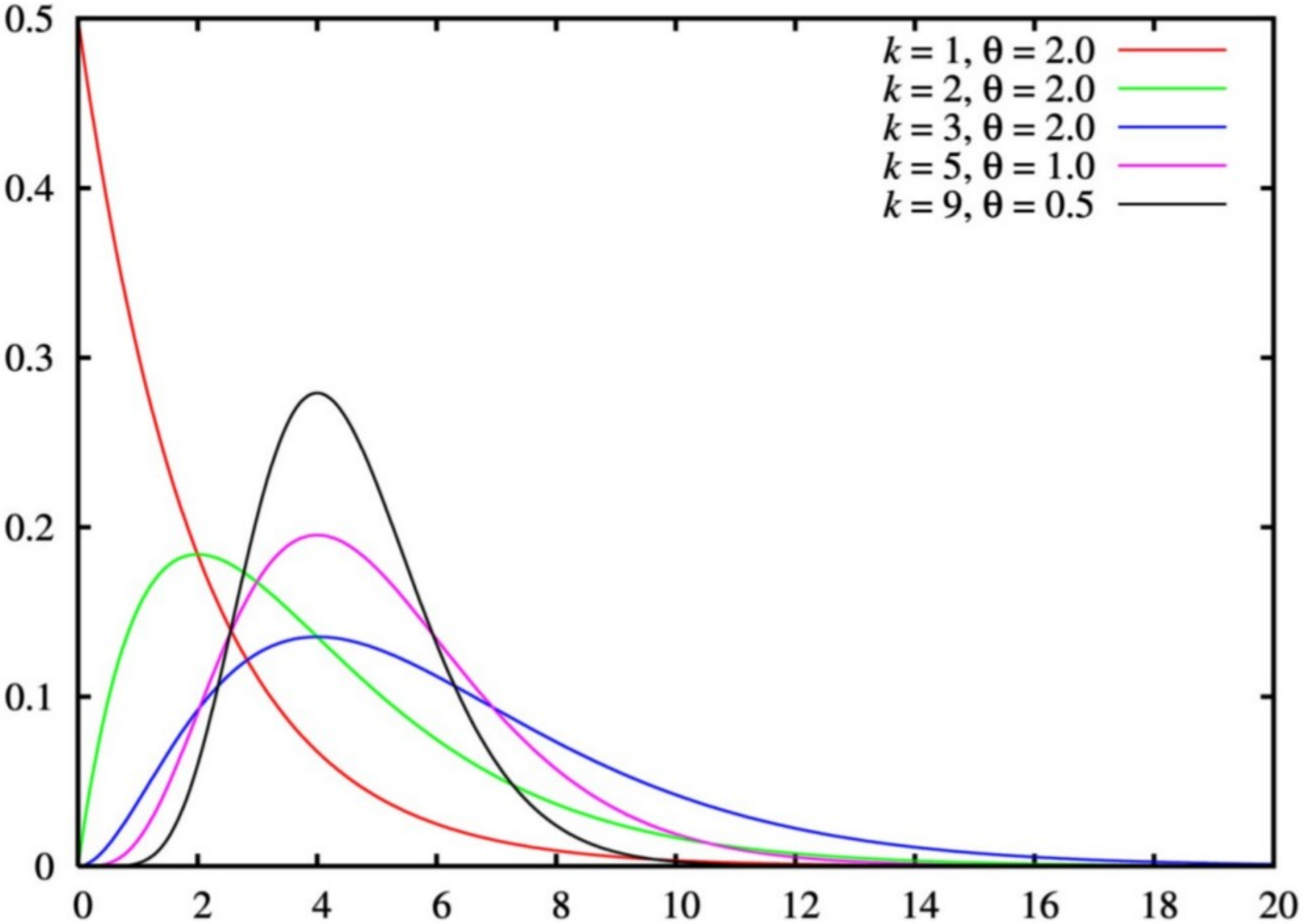
Variation of evolutionary rates over sites

So far we have assumed 1 evolutionary rate for all the sites is that necessary so?

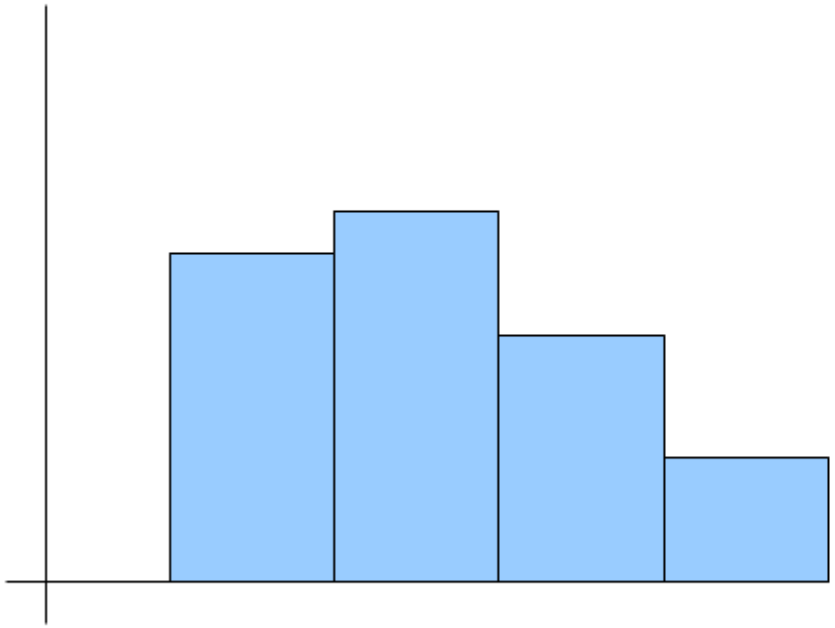
-> different sites, different evolutionary constrains

Approximation of rate-heterogenicity by a gamma-distribution

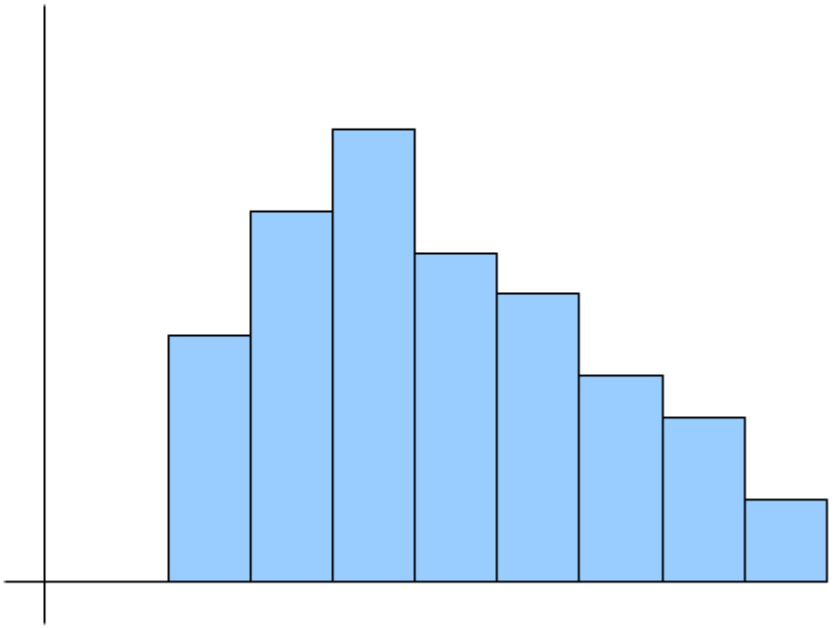
Parameters k (shape), Θ (scale). mean= k Θ



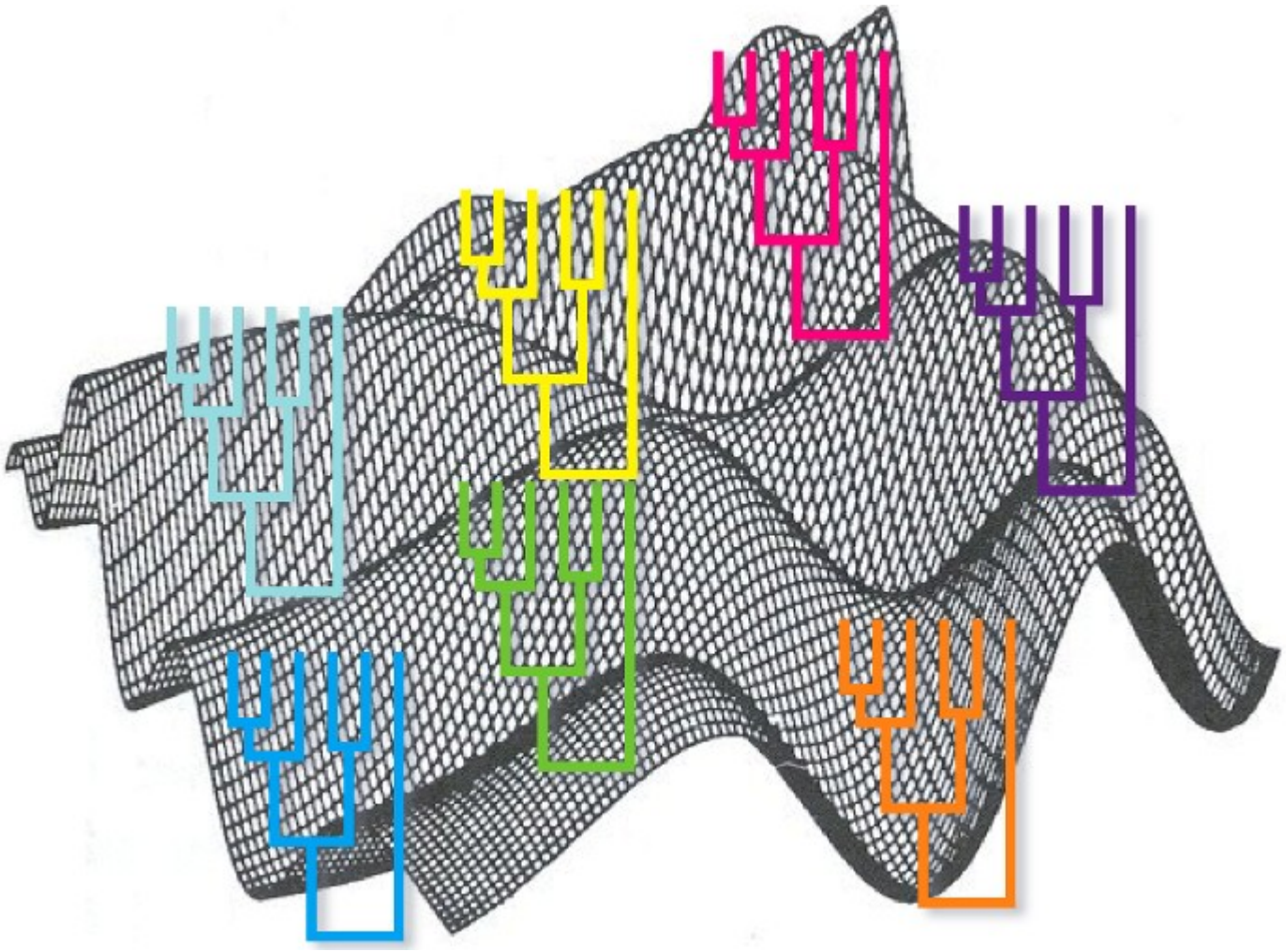
approximation by a discrete distribution



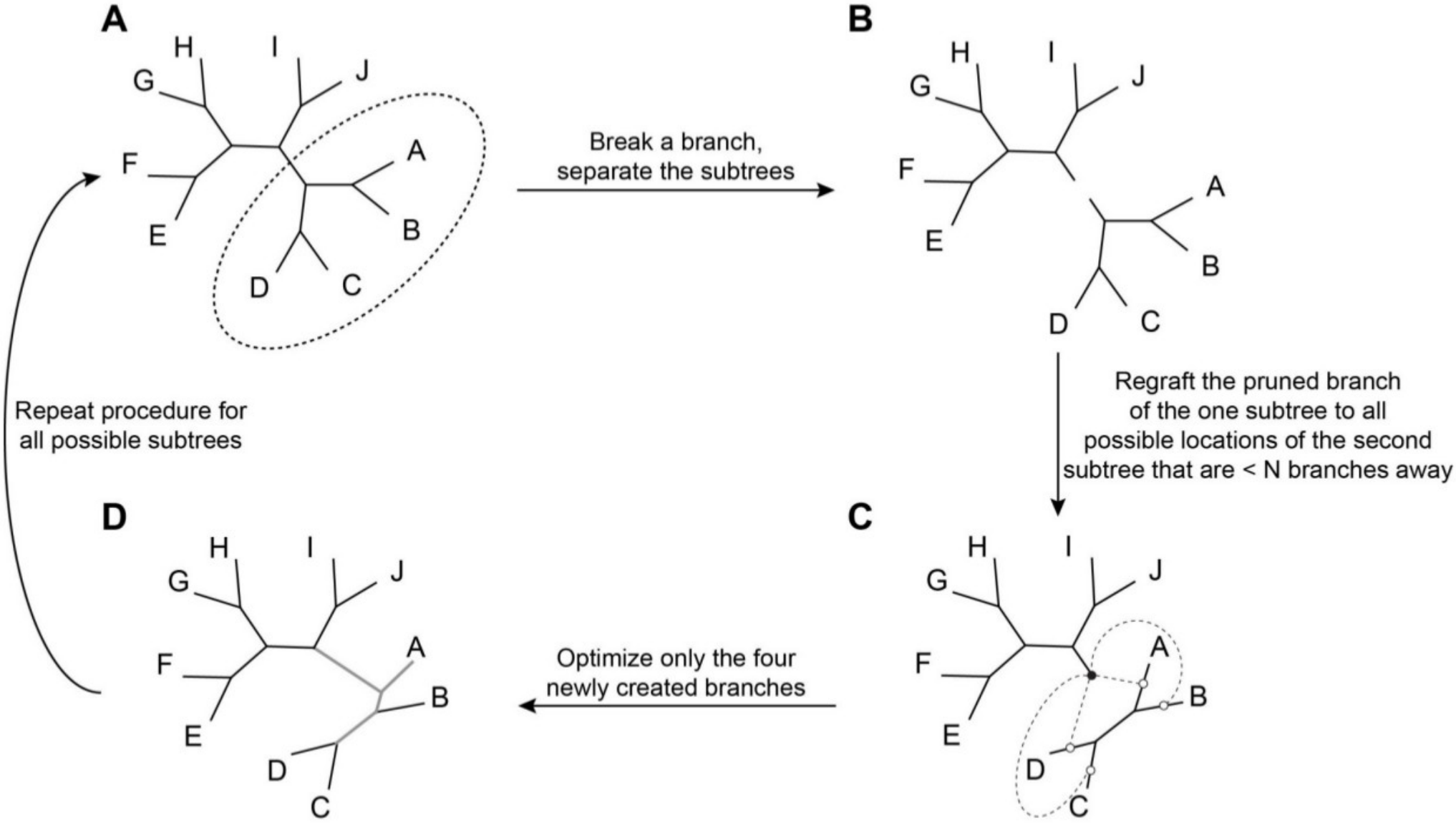
n=4



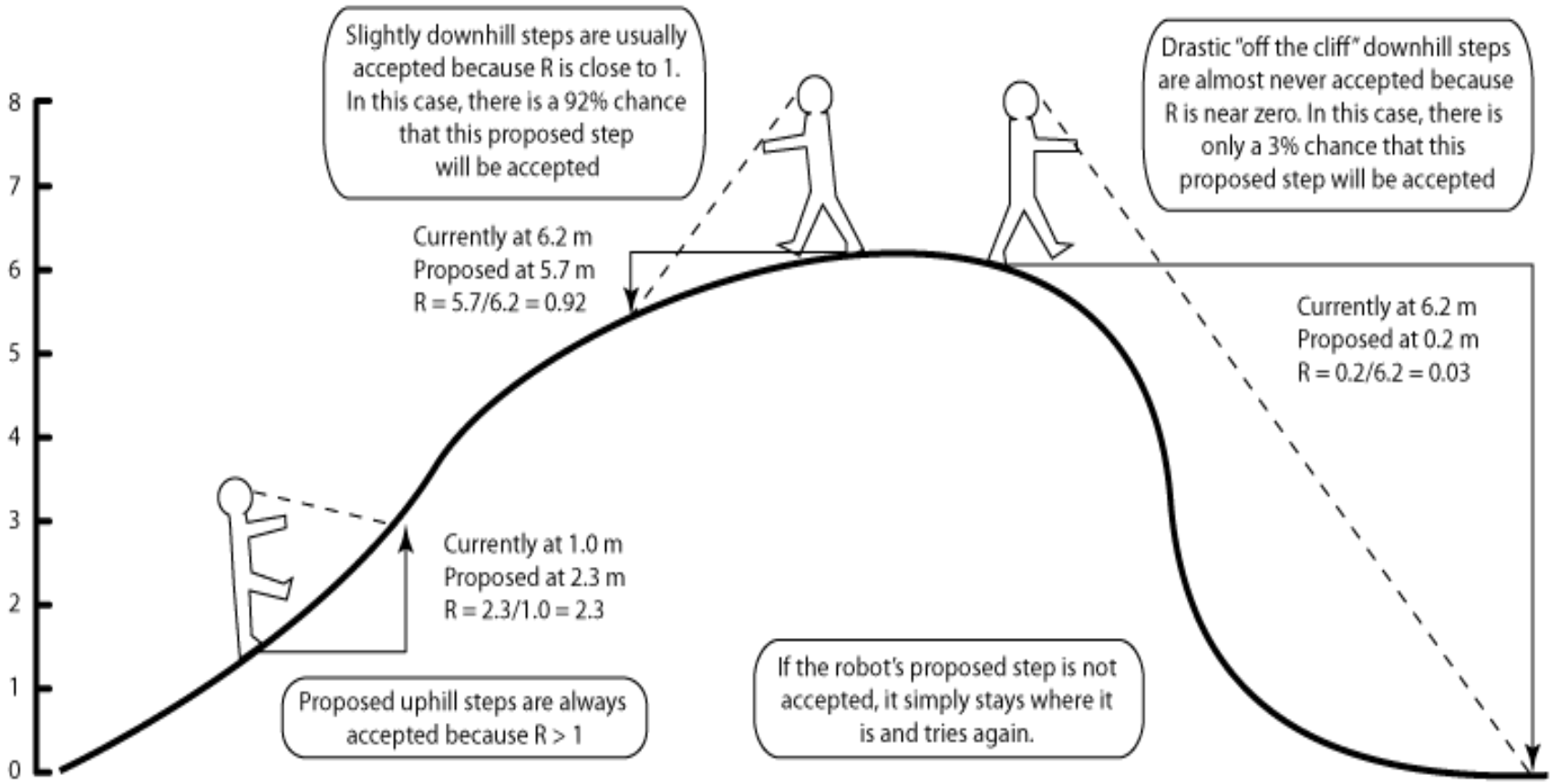
n=8



Heuristic Search



Heuristic searches don't guarantee discovery of optimal topology


Slightly downhill steps are usually accepted because R is close to 1. In this case, there is a 92% chance that this proposed step will be accepted
Drastic "off the cliff" downhill steps are almost never accepted because R is near zero. In this case, there is only a 3% chance that this proposed step will be accepted
Currently at 6.2 m
Proposed at 5.7 m
R = 5.7/6.2 = 0.92
Currently at 6.2 m
Proposed at 0.2 m
R = 0.2/6.2 = 0.03
Currently at 1.0 m
Proposed at 2.3 m
R = 2.3/1.0 = 2.3
Proposed uphill steps are always accepted because R > 1
If the robot's proposed step is not accepted, it simply stays where it is and tries again.
8
7
6
5
4
3
2
1
0

Bayesian Inference

♣ **Maximum Likelihood** will find the tree that is most likely to have produced the observed sequences, or formally $P(D/H)$ (the probability of seeing the data given the hypothesis).

♠ **A Bayesian approach** will give you the tree (or set of trees) that is most likely to be explained by the sequences, or formally $P(H/D)$ (the probability of the hypothesis being correct given the data).

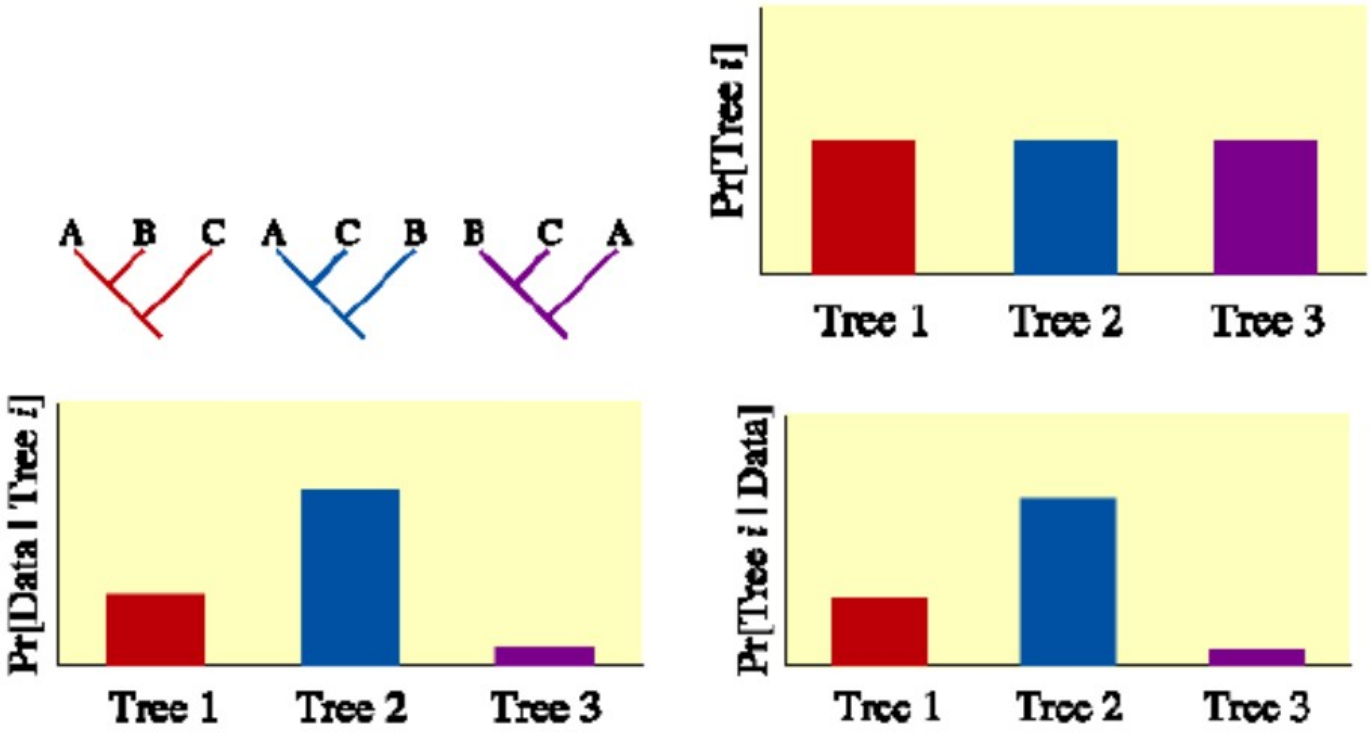
♢ **Bayes Theorem** provides a way to calculate the probability of a model (*tree topology and evolutionary model*) from the results it produces (*the aligned sequences we have*), what we call a **posterior probability**[31].

Thomas Bayes (1702-1761)

$$P(\theta/D) = \frac{P(\theta) \cdot P(D/\theta)}{P(D)}$$

The main components of Bayes analysis

- $P(\theta)$ The **prior probability** of a tree represents the probability of the tree before the observations have been made. Typically, all trees are considered equally probable.



- $P(D/\theta)$ The **likelihood** is proportional to the probability of the observations (data sets) conditional on the tree.

Posterior probability ~ likelihood x prior probability

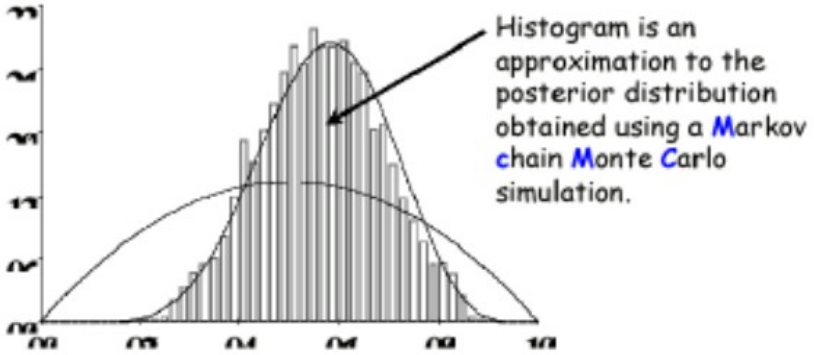
Usually flat distributions
uninformative
our beliefs before “seeing” the data

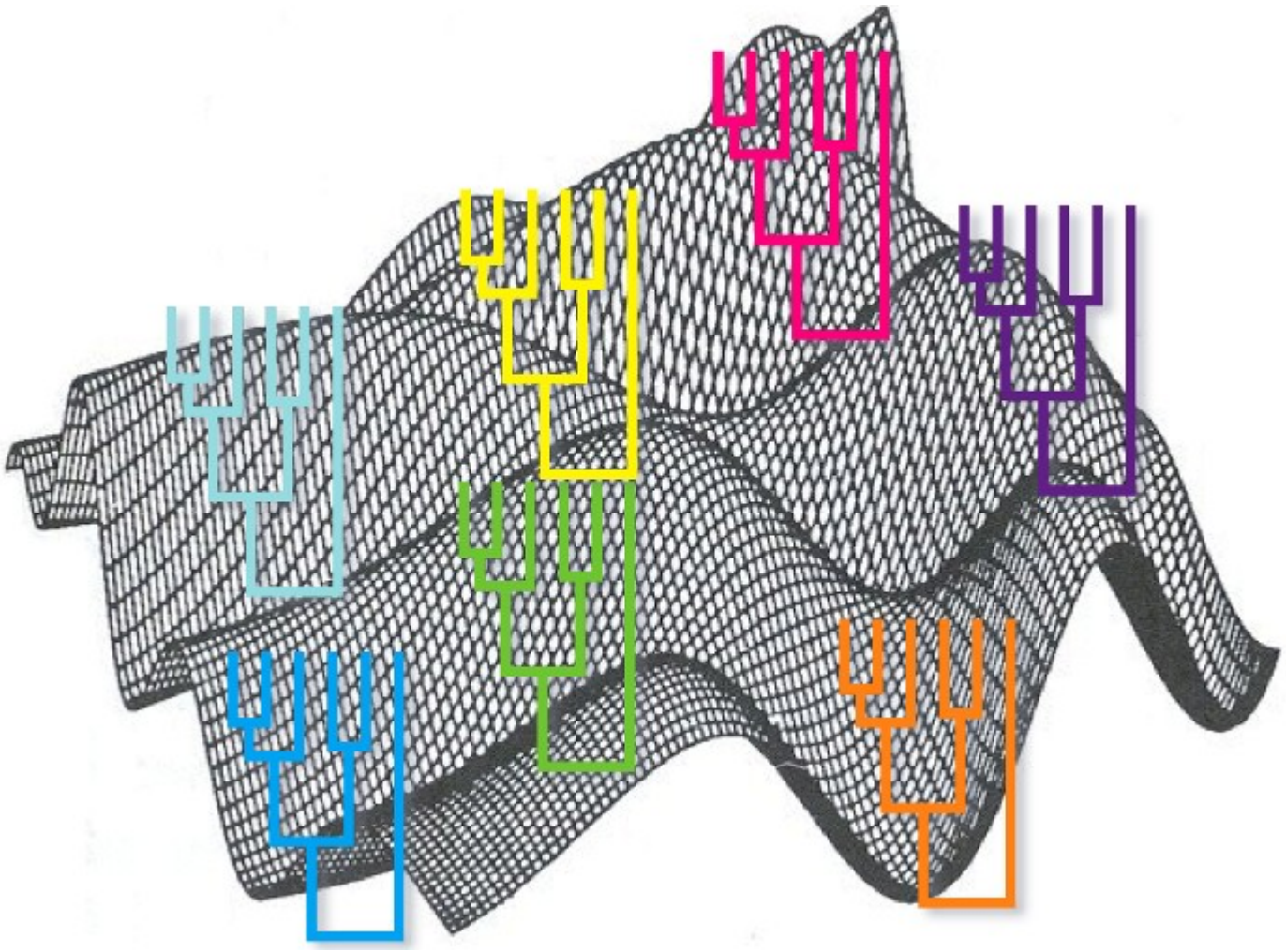
Therefore most of the differences in the posterior probability are due to differences in likelihood (-> strong connection to ML)

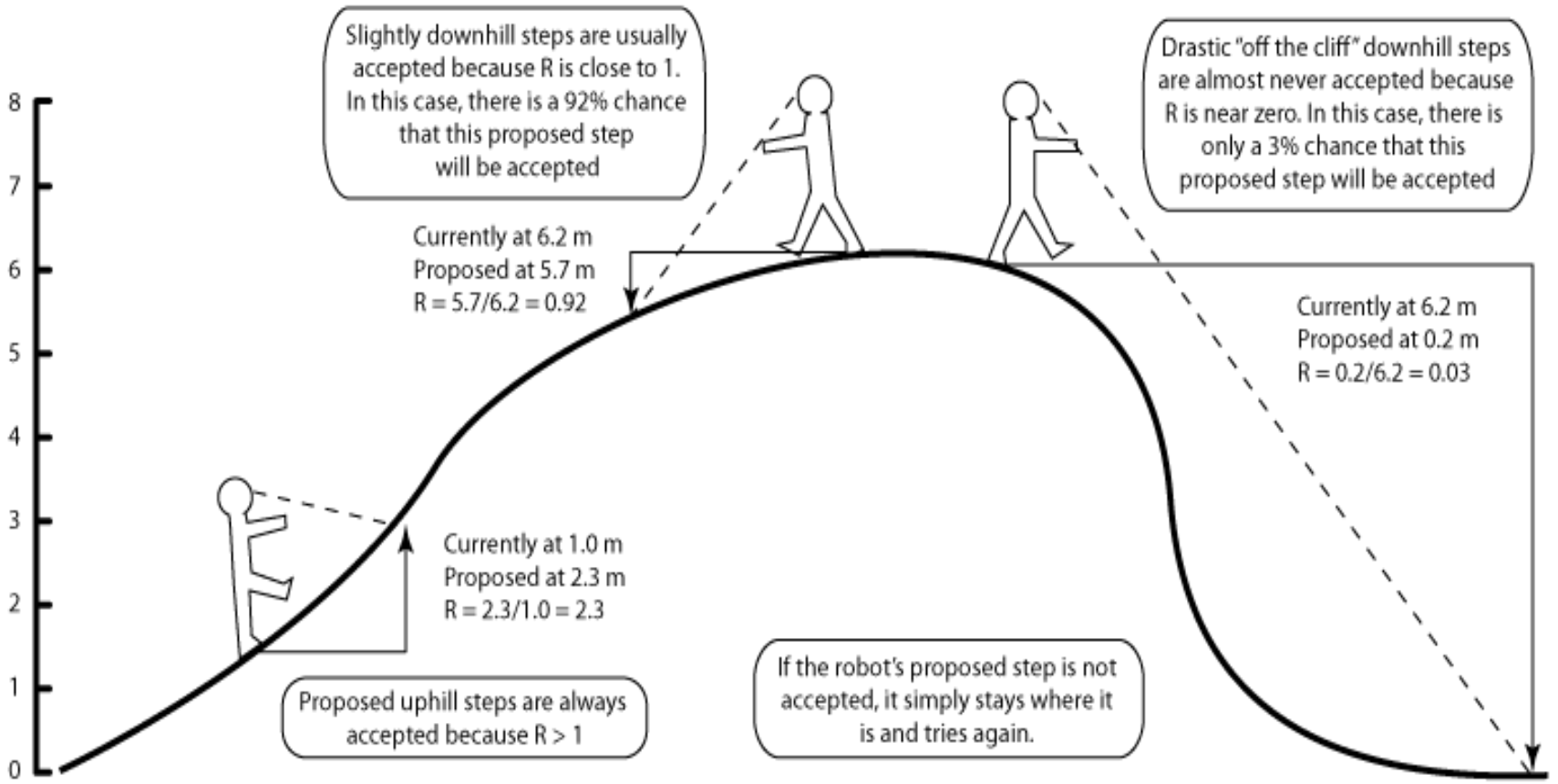
How to find the solution

There's no analytical solution for a Bayesian system. However, giving:

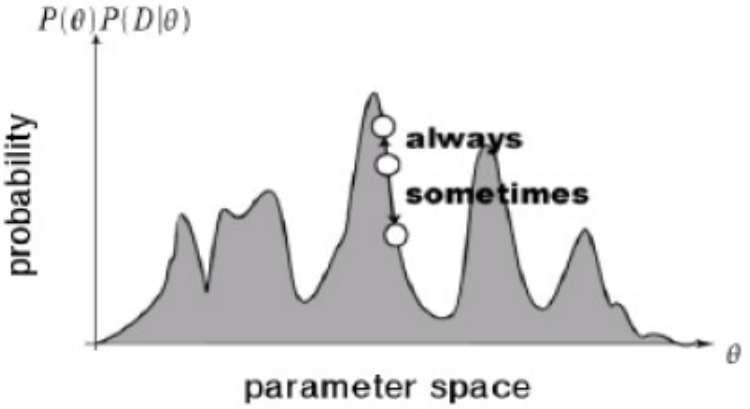
- **Data**: Sequence data,
- **Model**: The evolutionary model, base frequencies, among site rate variation parameters, a tree topology, branch lengths
- **Priors** distribution on the model parameters, and
- **A method** for calculating posterior distribution from prior distribution and data: **MCMC** technique[32]






Slightly downhill steps are usually accepted because R is close to 1. In this case, there is a 92% chance that this proposed step will be accepted
Drastic "off the cliff" downhill steps are almost never accepted because R is near zero. In this case, there is only a 3% chance that this proposed step will be accepted
Currently at 6.2 m
Proposed at 5.7 m
R = 5.7/6.2 = 0.92
Currently at 6.2 m
Proposed at 0.2 m
R = 0.2/6.2 = 0.03
Currently at 1.0 m
Proposed at 2.3 m
R = 2.3/1.0 = 2.3
Proposed uphill steps are always accepted because R > 1
If the robot's proposed step is not accepted, it simply stays where it is and tries again.
0
1
2
3
4
5
6
7
8

- Each step in a Markov chain **a random modification** of the tree topology, a branch length or a parameter in the substitution model (e.g. substitution rate ratio) is assayed.
- If the **posterior computed is larger** than that of the current tree topology and parameter values, the proposed step **is taken**.
- Steps downhill are not authomatic accepted, depending on the magnitude of the decrease.



- Using these rules, the **Markov chain visits regions** of the tree space **in proportion of their posterior**.
- Suppose you sample 100,000 trees and a particular clade appears in 74,695 of the sampled trees. The probability (giving the observed data) that the group is monophyletic is 0.746, because **MC visits trees in proportion to their posterior probabilities**.

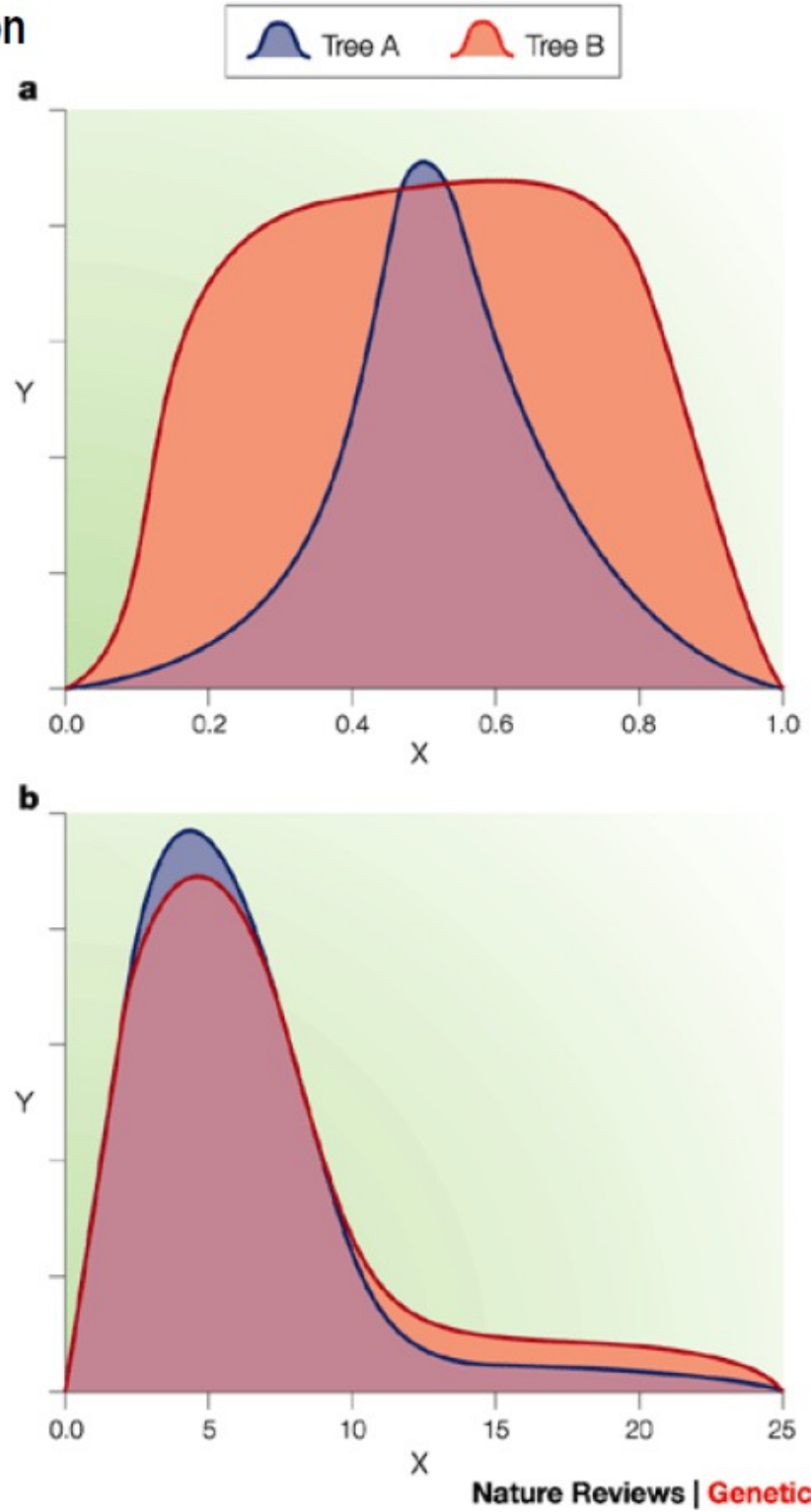
A fundamental difference:

ML commonly uses **Joint Estimation**: finding the highest point in the parameter landscape

Bayesian analyses measure the volume under the posterior probability surface, the parameters are integrated (marginalized) to obtain the marginal posterior probability of a tree (**Marginal estimation**)

Joint versus Marginal estimation

Likelihood



ML favours A
Bayesian favours B

Every part of the surface affects the results, so the prior distributions may be seriously considered.

ML favours A
Bayesian favours B

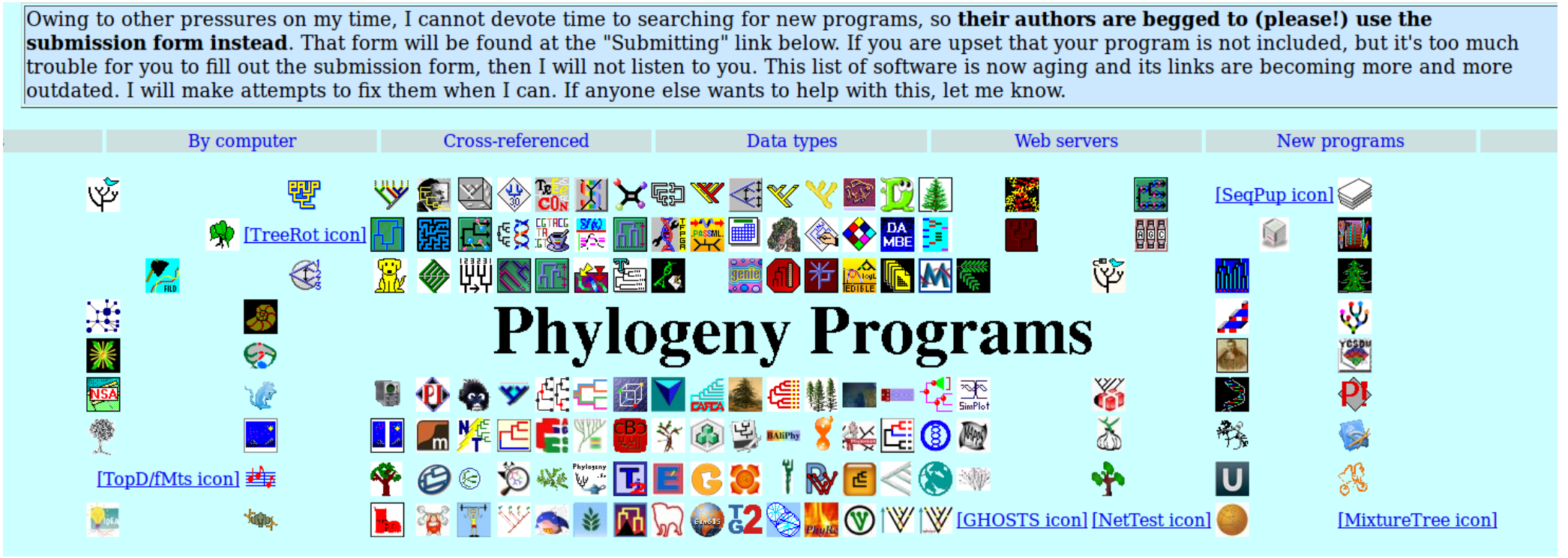
1) When there are few parameters and plenty of data, the likelihood and posterior landscapes become thin spires with high peaks:

---> The heigh of the peak is a good predictor of the integral over the whole surface..... ML ~ BI

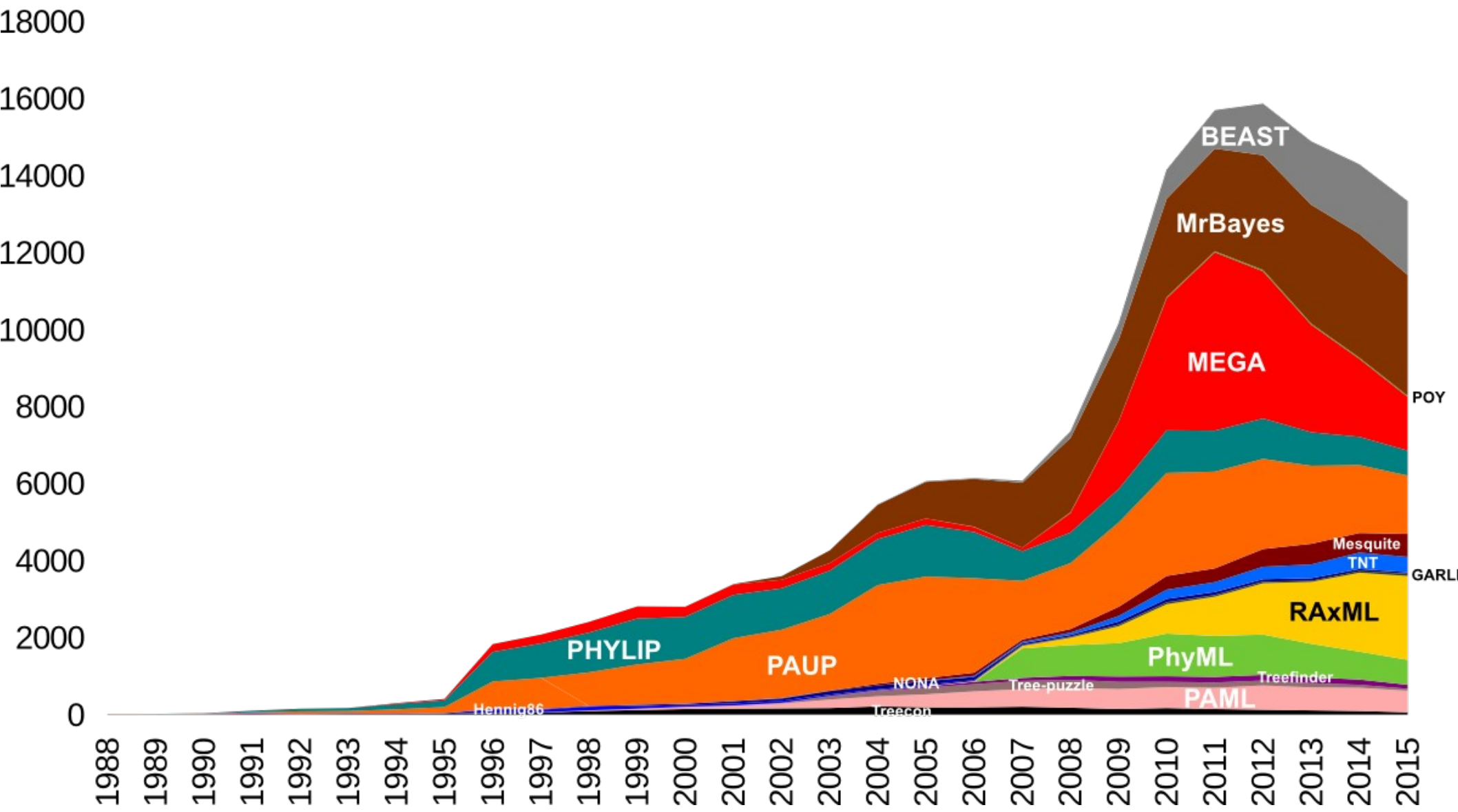
2) As the amount of data decreases relative to the number of parameters (complex models), the landscapes become soft hills, and consideration on the uncertainty of the parameters is necessary.

--> marginal estimation (BI) is more reliable

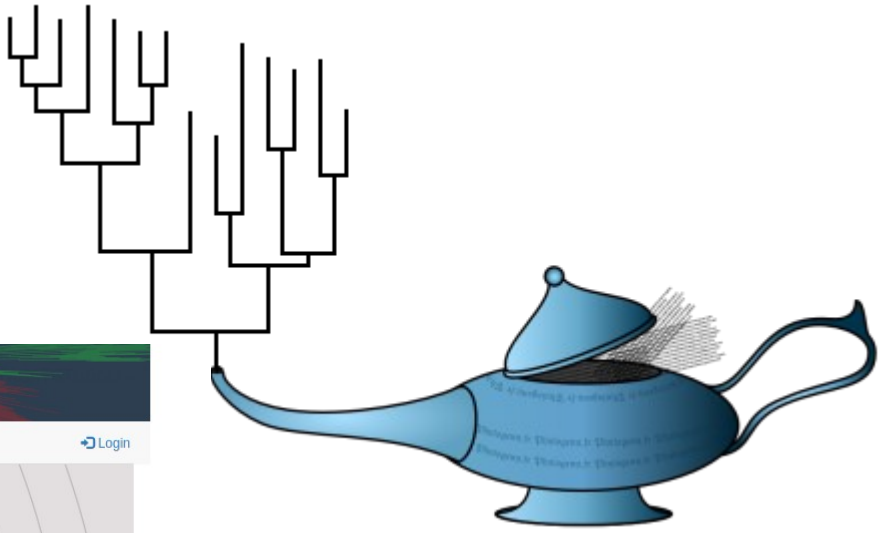
http://evolution.genetics.washington.edu/phylip/software.html



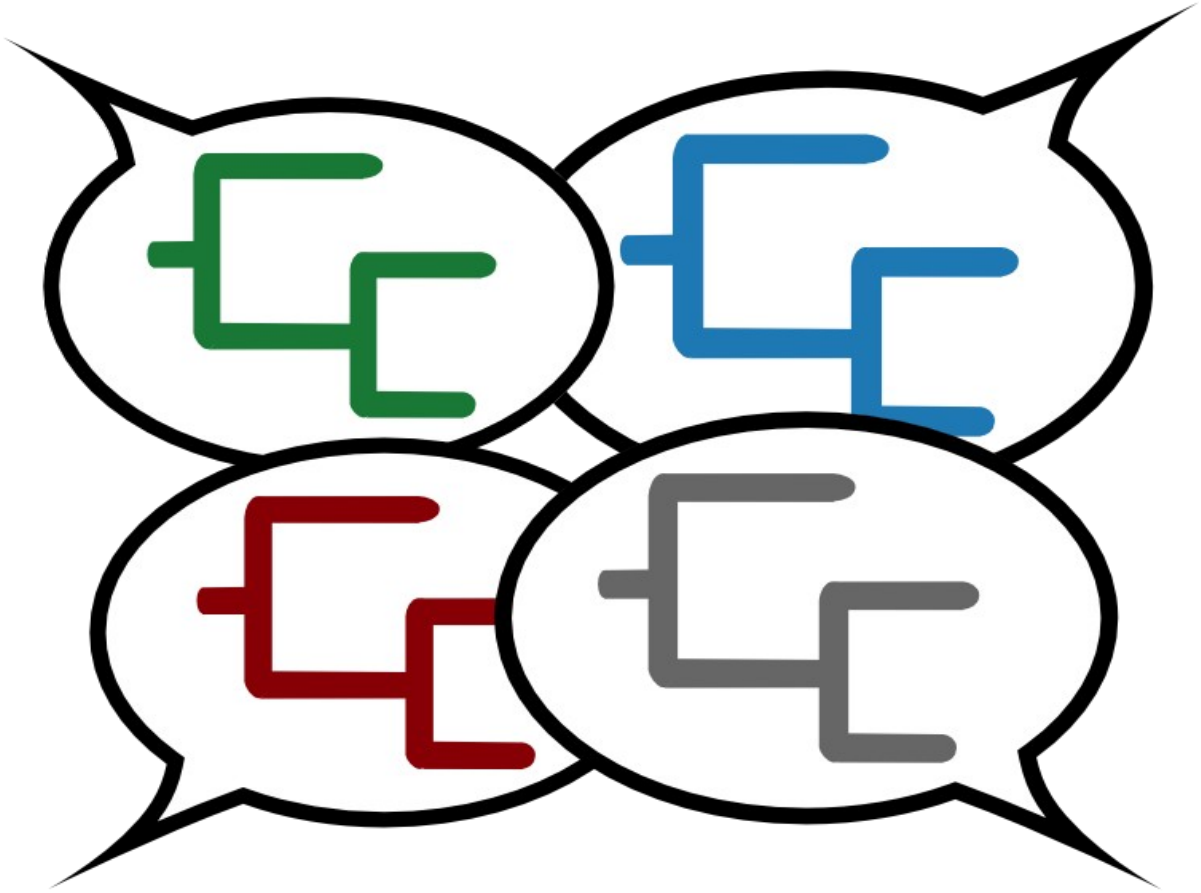
https://en.wikipedia.org/wiki/List_of_phylogenetics_software


18000
16000
14000
12000
10000
8000
6000
4000
2000
0
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
BEAST
MrBayes
MEGA
POY
Mesquite
TNT
GARLI
RAxML
PhyML
Treefinder
PAML
PHYLIP
PAUP
NONA
Tree-puzzle
Hennig86
Treecon

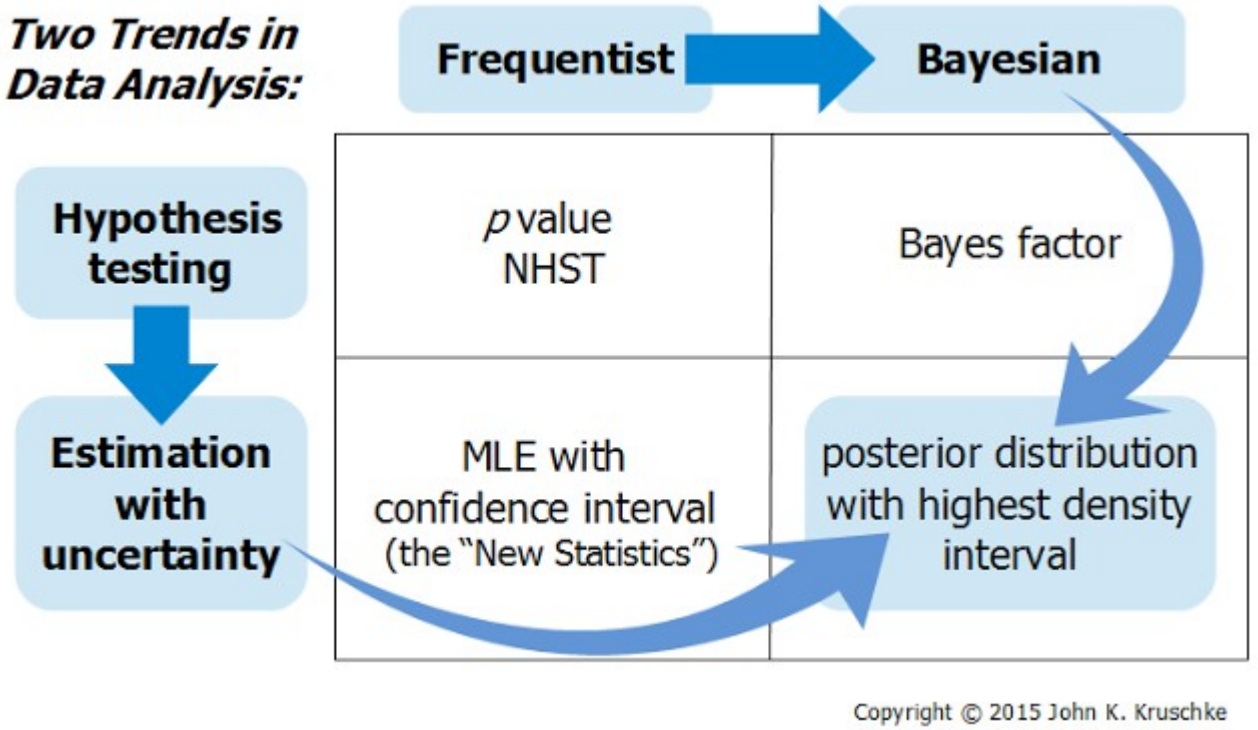
ngphylogeny.fr



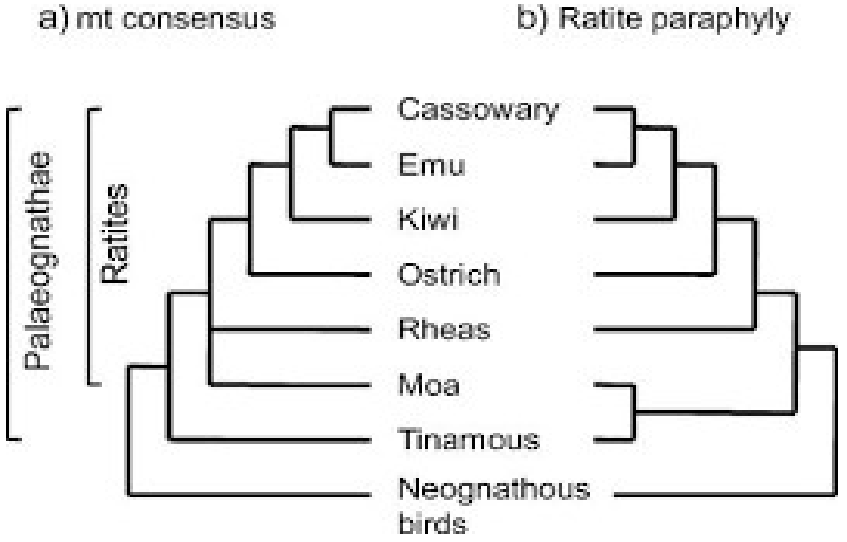
Hypothesis testing on gene family trees



Probabilistic methods render themselves for testing

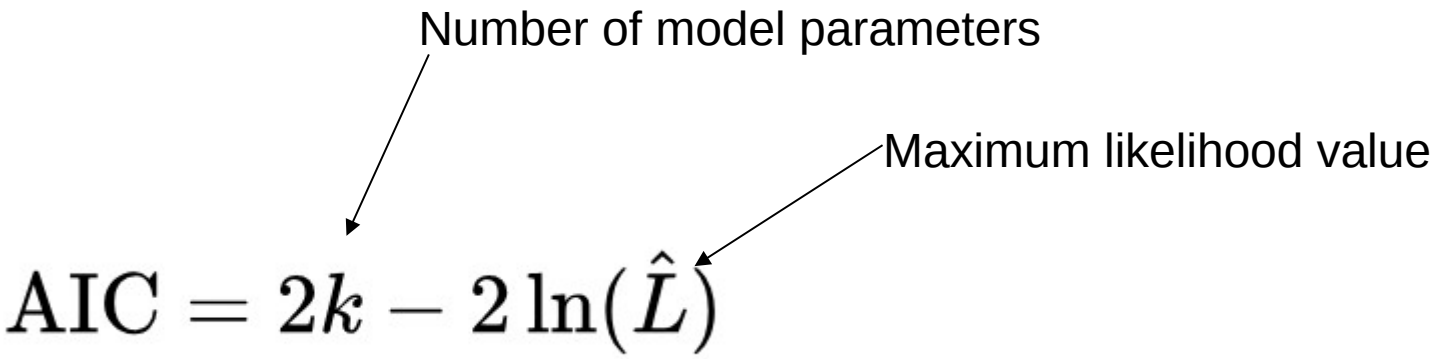


Comparing (testing) alternative topologies



Likelihood 1

Likelihood 2



The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

Given a set of model, we can use AIC to choose the one with the minimum value, achieves Better likelihood with the least number of parameters

Inferring evolutionary events

Speciations/Duplications.

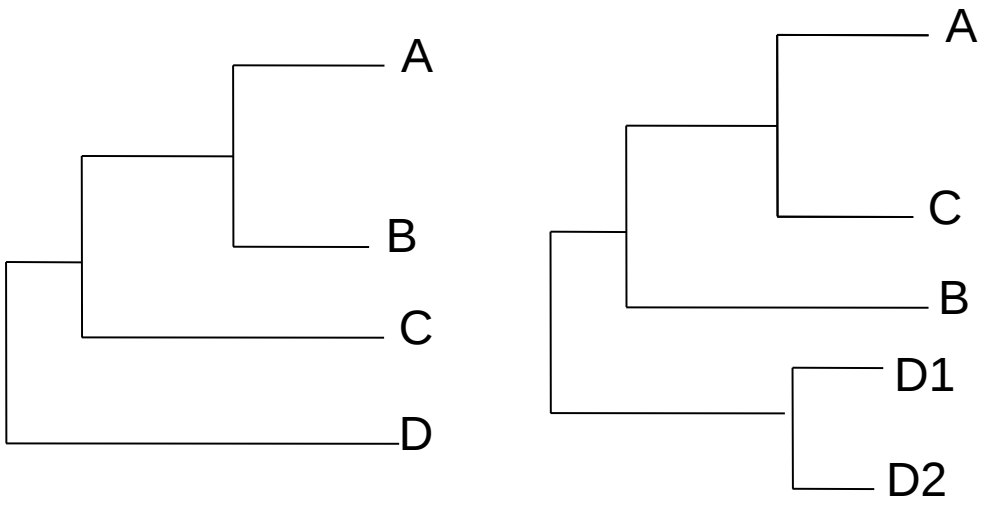
1- Reconciliation algorithms

2- Species overlap

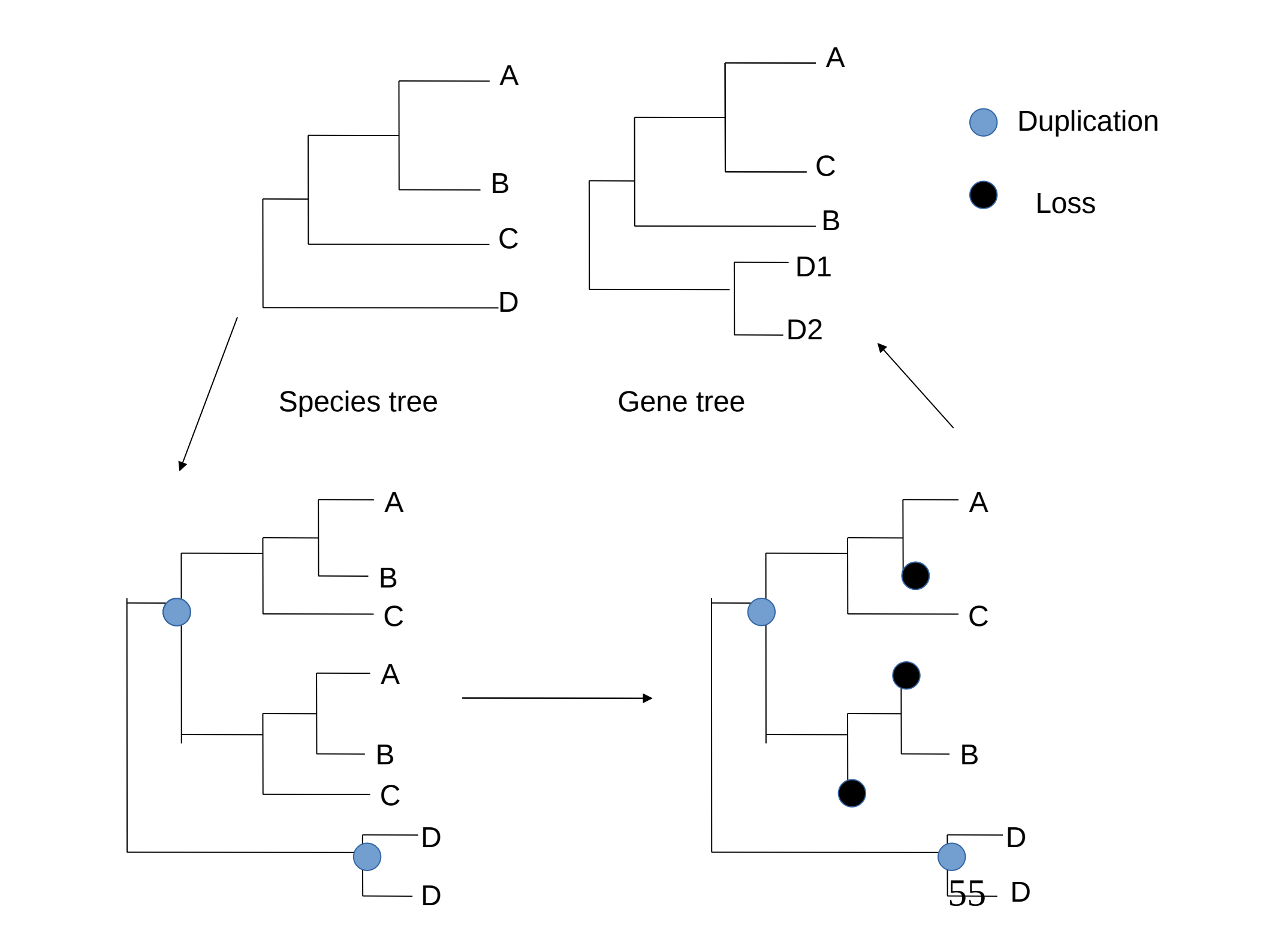
Reconciliation algorithm.

(Hard reconciliation) Resolve any incogurence between gene tree and species tree by introducing the minimal number of gene duplicatios and losses.

(Soft reconciliation) Allow incongruences below a given support value



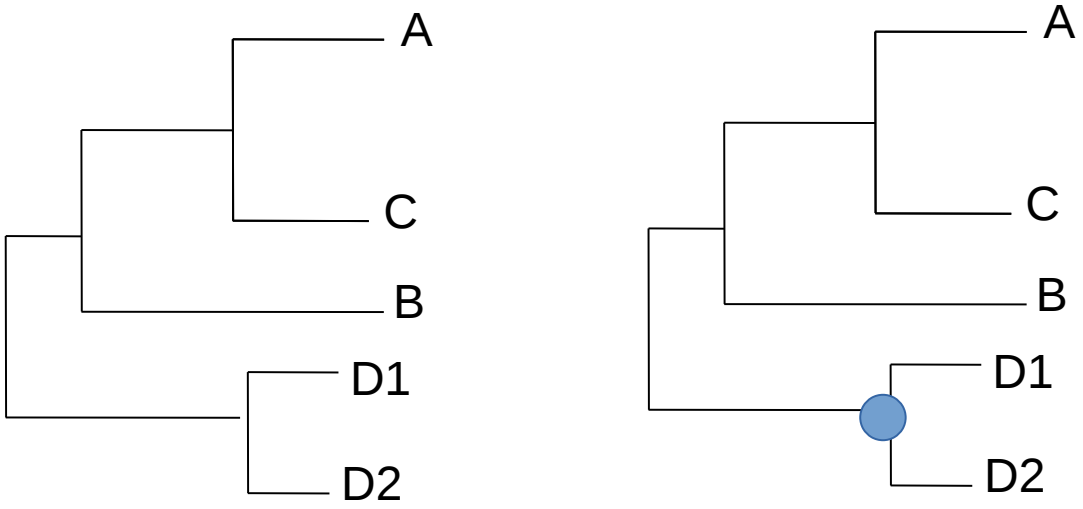
Species tree Gene tree


A
B
C
D
Species tree
A
C
B
D1
D2
Gene tree
Duplication
Loss
A
B
C
A
B
C
D
D
A
C
B
D
D

Species ovelap algorithm.

It does not require a species-tree but needs to know the species to which The genes belong
In essence can be seen as a reconciliation with an unresolved species tree

For every node in the gene tree evaluate whether the daughter partitions share any species. If the overlap (number of species shared over total number of species) is higher than the given threshold. Inpute a duplication at that node.

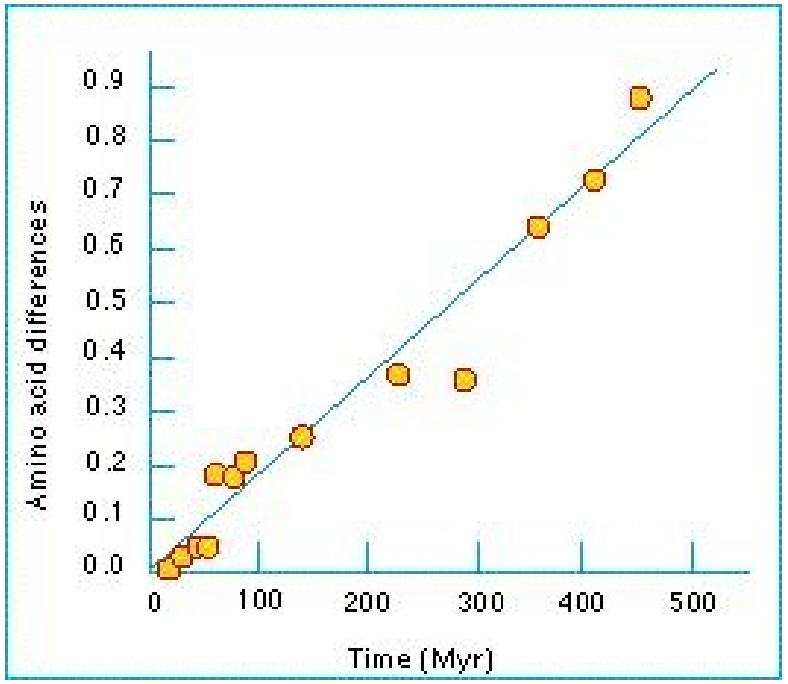


Gene tree

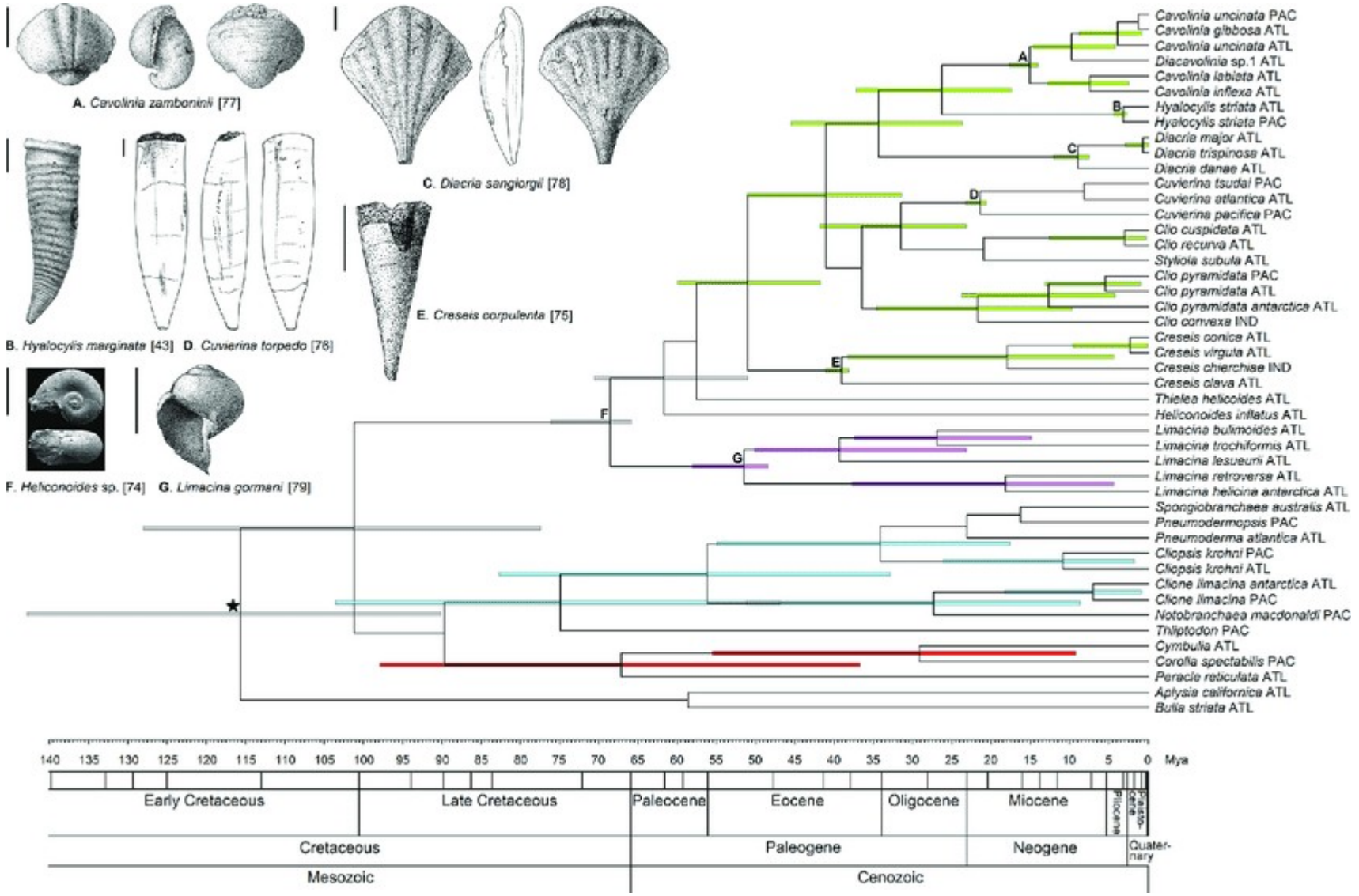
Inferring relative timing of speciations and duplications

A **molecular clock** is said to exist when substitutions acumulate linearly with time. This assumption is violated most of the time, particularly at long evolutionary distances.

For a constant mutation rate, neutral substitutions are expected to behave more clock-like than non-neutral substitutions, that is why they are generally used as a **proxy** for time



Fossils can be used to calibrate timing of speciation nodes in a phylogeny,
But fossil dating is also subject to errors

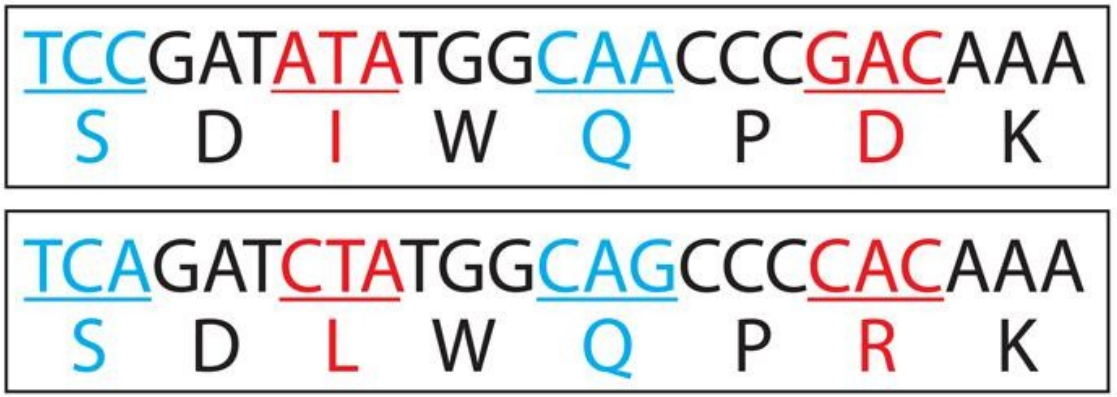


What about speciations and duplication nodes in a gene tree?

Genes do not have equal evolutionary rates across a genome

Synonymous substitutions still used as a proxy. However variations in GC content, Local variations in mutation rates, or repair can introduce noise

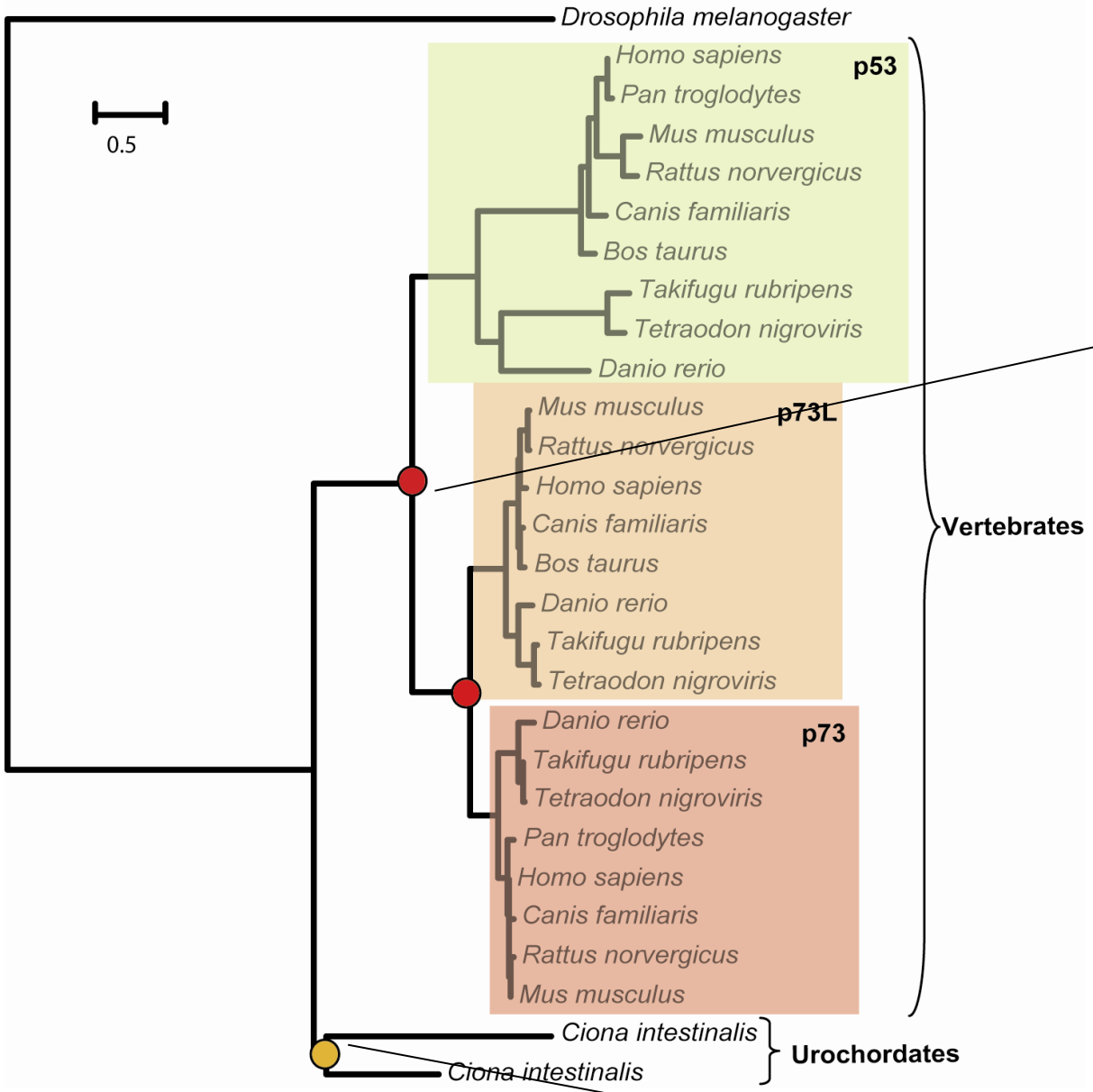
A Nonsynonymous / Synonymous substitution

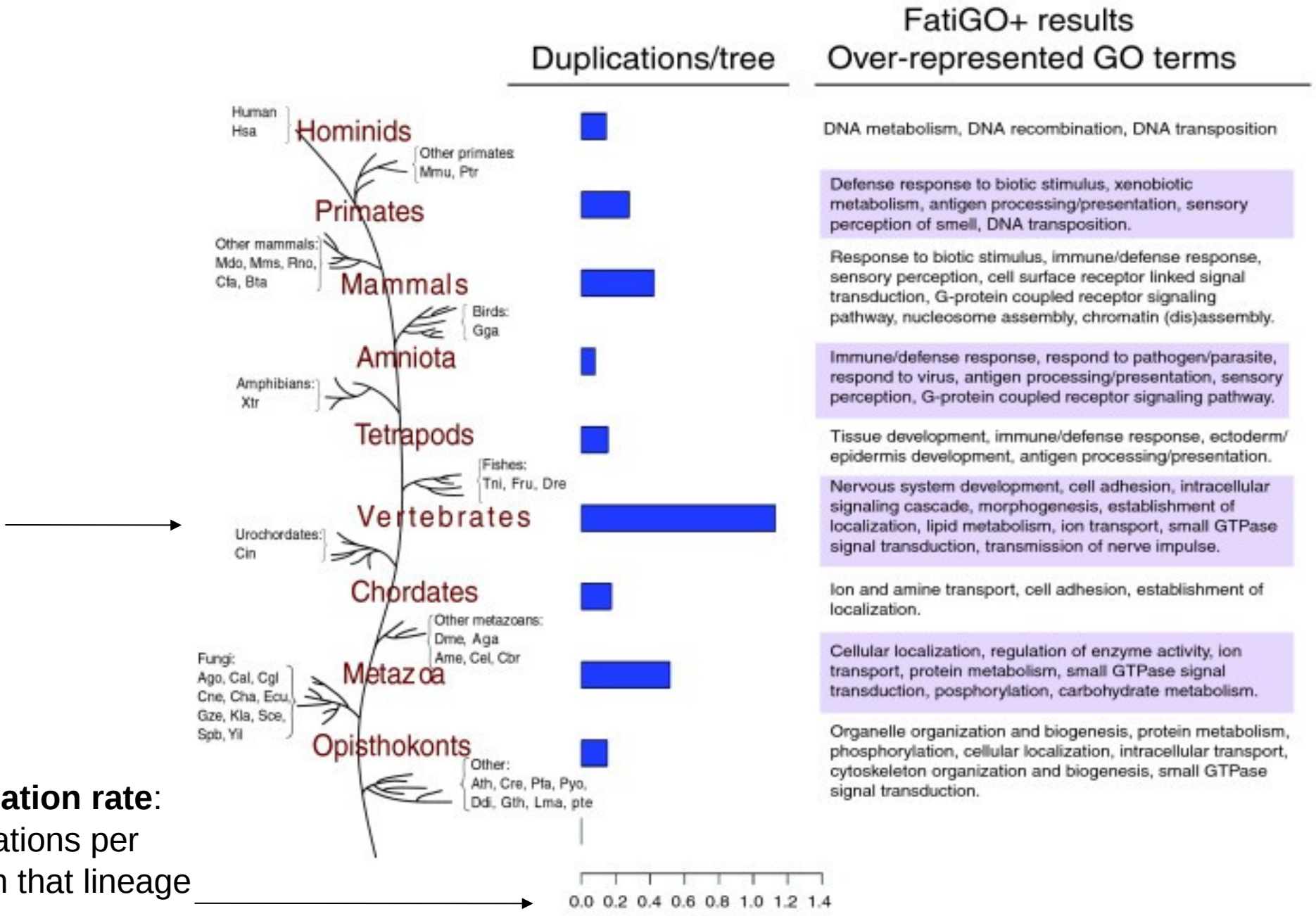


B Radical / Conservative substitution

\+ positive
\- negative
N neutral

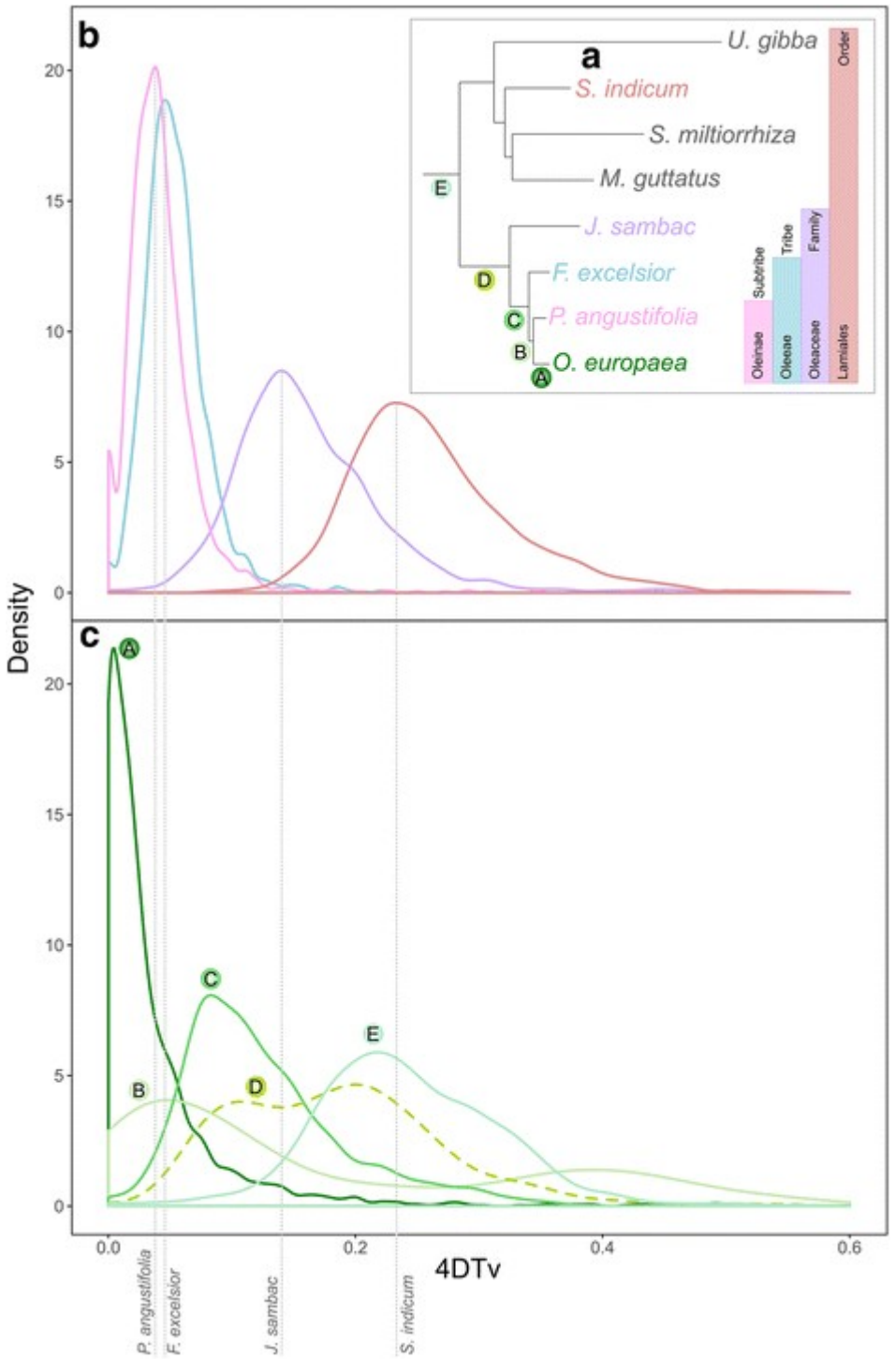
Ks = rate of synonymous substitutions, 4Dtv = rate at **four fold degenerate sites**


Drosophila melanogaster
0.5
Homo sapiens
Pan troglodytes
Mus musculus
Rattus norvergicus
Canis familiaris
Bos taurus
Takifugu rubripens
Tetraodon nigroviris
Danio rerio
p53
Mus musculus
Rattus norvergicus
Homo sapiens
Canis familiaris
Bos taurus
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
p73L
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
Pan troglodytes
Homo sapiens
Canis familiaris
Rattus norvergicus
Mus musculus
p73
Vertebrates
Ciona intestinalis
Ciona intestinalis
Urochordates
Duplication at the base of vertebrates
Ciona-specific duplication



Duplication rate: Duplications per gene in that lineage

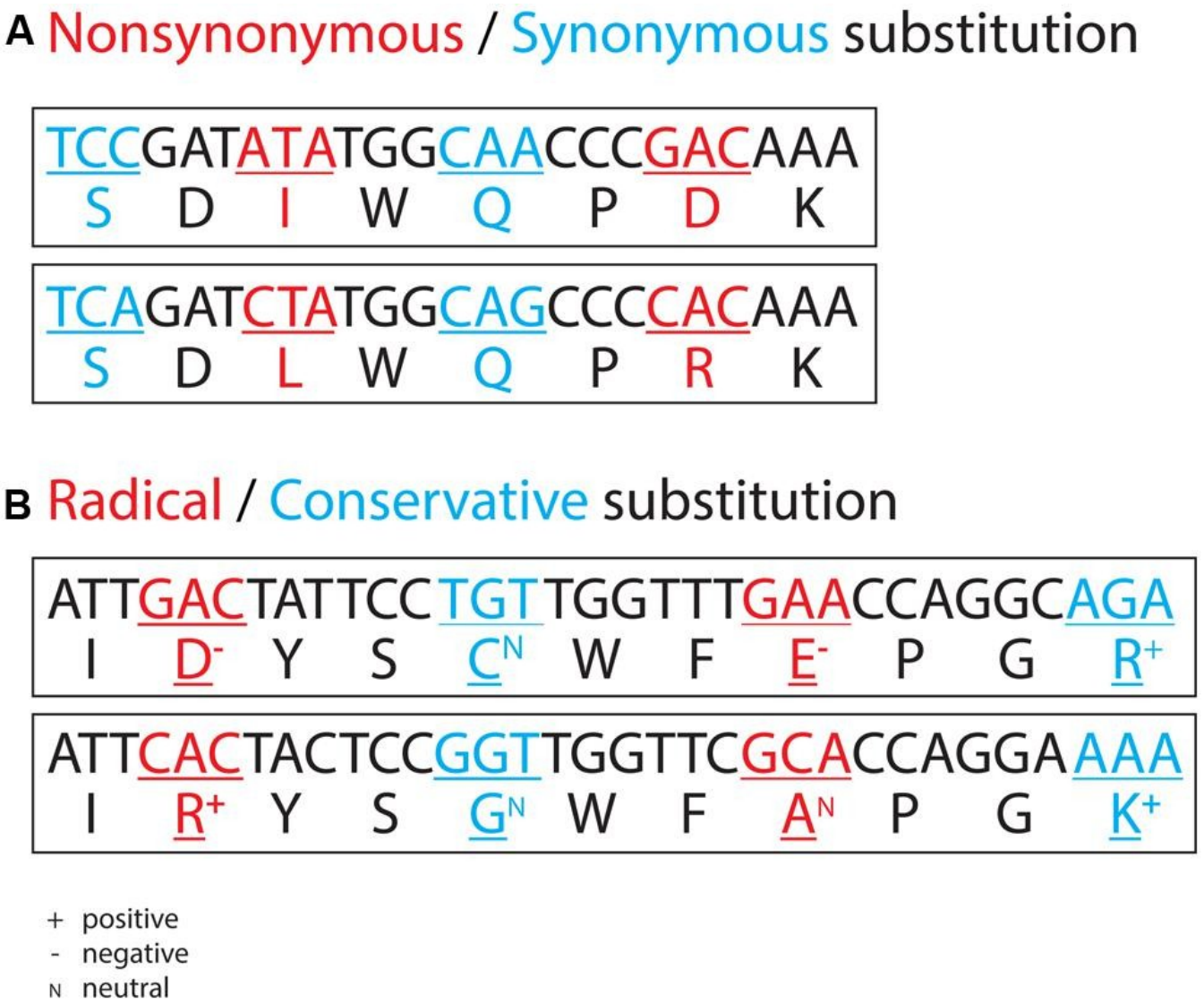
Huerta-Cepas et. al. (2007)



Orthologs

Paralogs

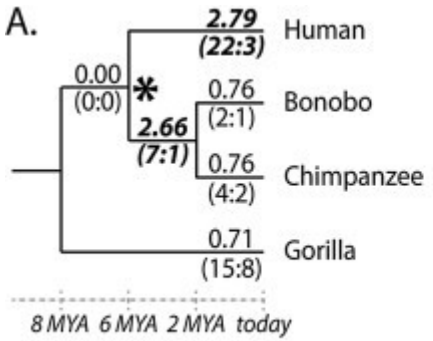
If non-synonymous mutations are the result of selection, can we use them to detect positive selection, functional shifts, etc?



dN/dS ( )

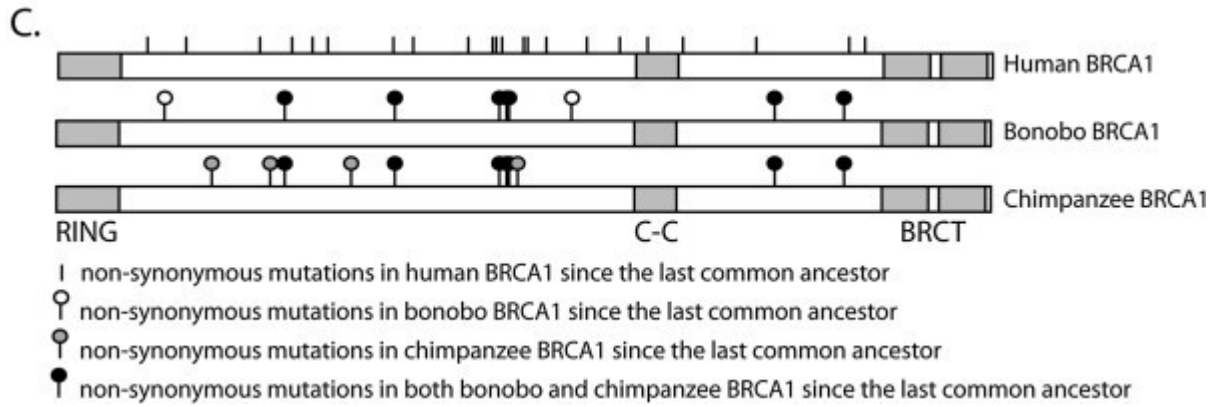
The ratio of non-synonymous to synonymous substitutions (dN/dS) is a useful measure of the strength and mode of natural selection acting on protein-coding genes.

Popular ML programs (i.e. PAML) can compute dN/dS ratios per branch, and per site

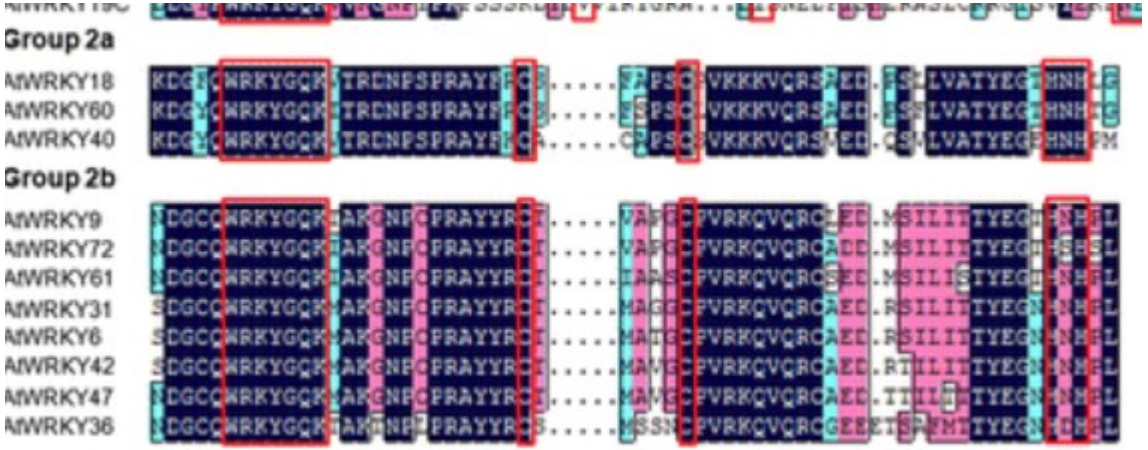


B.

Gene	kb	N	S	N/kb	S/kb	N/kb / S/kb
BRCA1	5.589	22	3	3.9	0.5	7.8
BACH1	2.208	0	3	0.0	1.4	0.0
BARD1	2.331	1	1	0.4	0.4	1.0
CtIP	2.691	5	7	1.9	2.6	0.7
PALB2	3.558	7	5	2.0	1.4	1.4
RAD50	3.936	1	3	0.3	0.8	0.4
BRCA2	10.254	7	8	0.7	0.8	0.9



Detecting radical amino acid changes in a phylogeny can be indicative of functional shift (Program DIVERGE2)

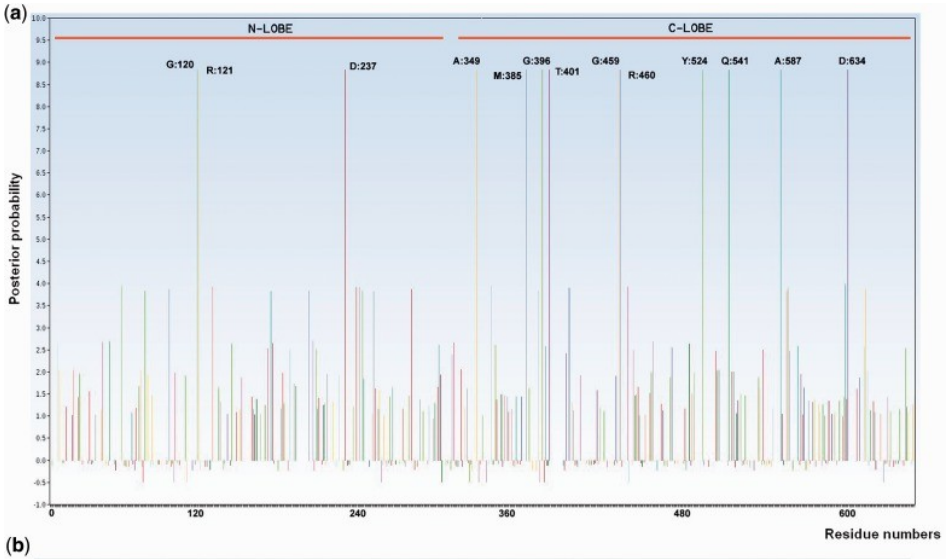


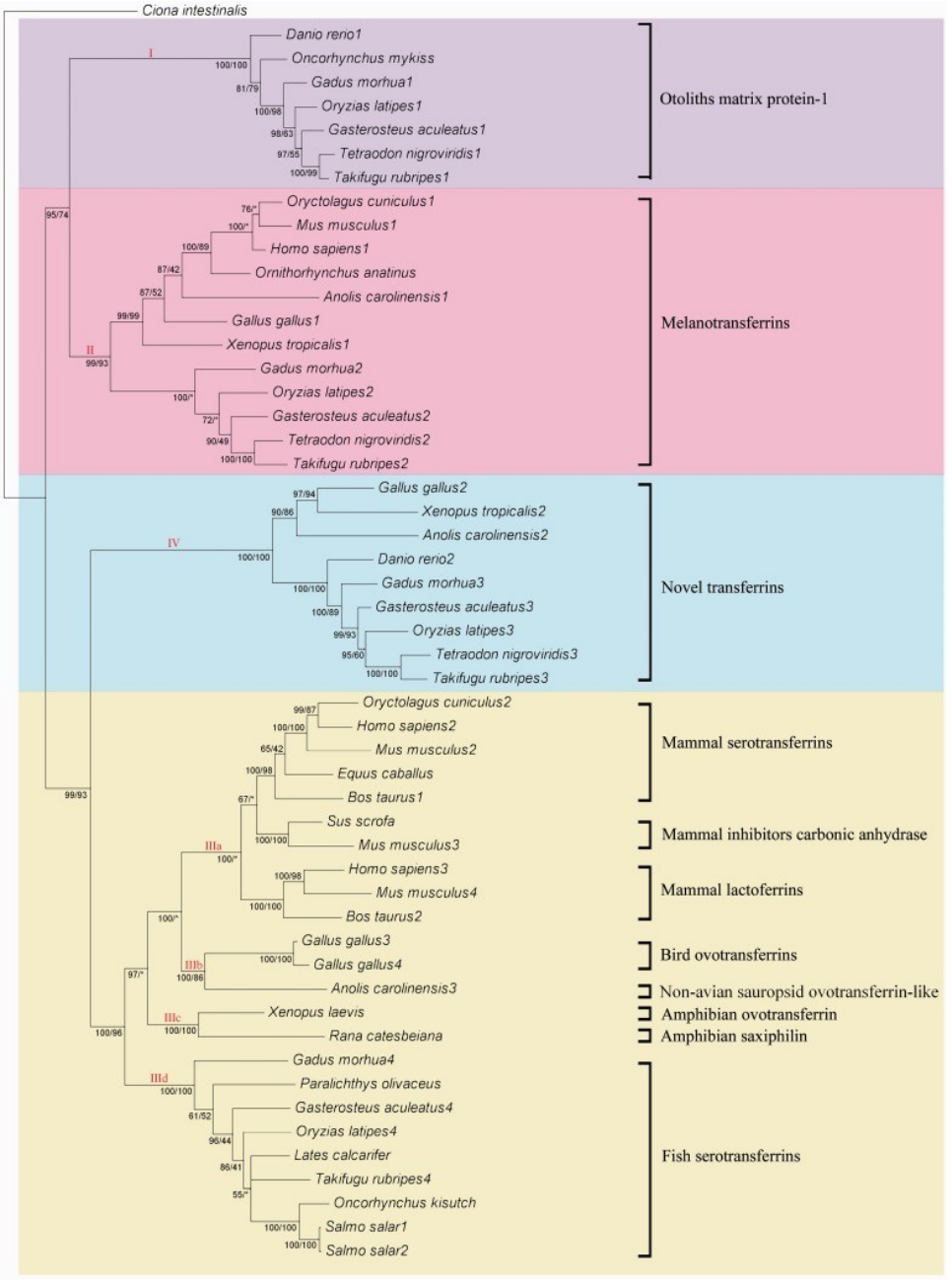
Mol Biol Evol. 2013 Apr; 30(4): 894–905.
Published online 2012 Dec 19. doi: 10.1093/molbev/mss325
PMCID: PMC3603312
PMID: 23258311

Phylogenetic Analyses Uncover a Novel Clade of Transferrin in Nonmammalian Vertebrates

Hirzahida Mohd-Padil,[1] Adura Mohd-Adnan,[*,1] and Toni Gabaldón[*,2,3]

Author information ► Copyright and License information ► Disclaimer





Ancestral Sequence Ressurection: directly testing for ancestral protein functions

Published online 5 September 2002 | Nature | doi:10.1038/news020902-7

News

Triassic reptile saw red

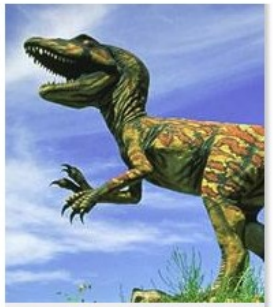
Resurrected protein suggests that crocodiles' ancestors roamed at night.

Helen Pearson

A reptile from the Triassic period may have done its stalking at night. So suggest scientists who have resurrected a 240-million-year-old eye protein that sees dim light[1].

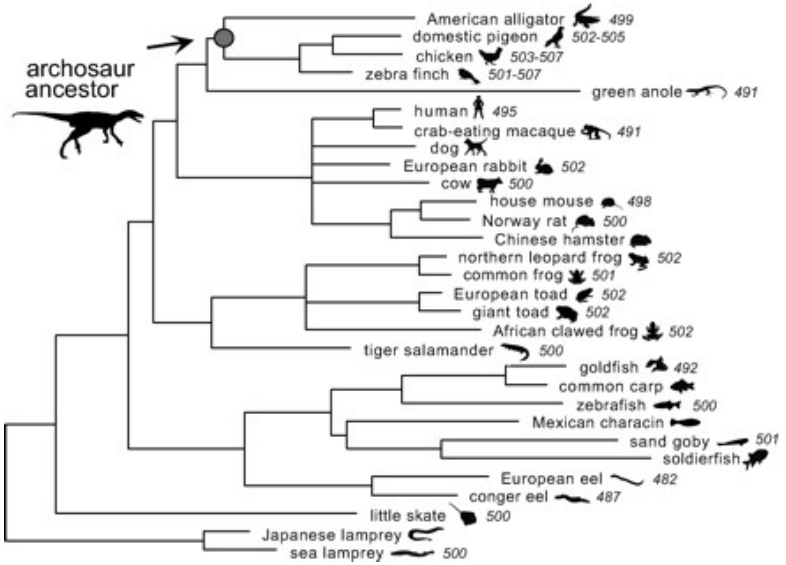
Such a molecule may have been found in the eyes of the earliest archosaurs, which were predecessors of the dinosaurs. Similar proteins, called rhodopsins, perceive low levels of light in humans and other animals.

Thomas Sakmar of Rockefeller University in New York and his colleagues used a computer program to extrapolate the DNA sequence of the ancient rhodopsin from known sequences in alligator, birds, frogs and fish.



Gene reconstruction gives researchers a dim view of the distant past.

© SPL



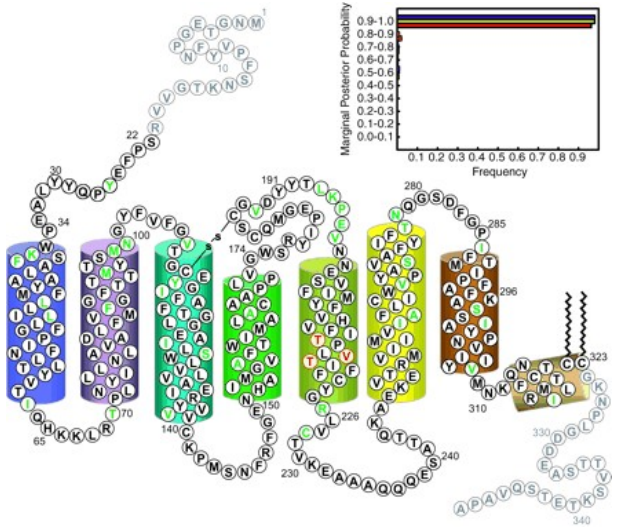
Recreating a Functional Ancestral Archosaur Visual Pigment

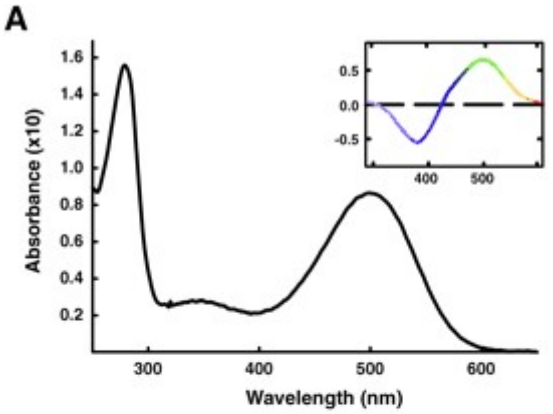
FREE

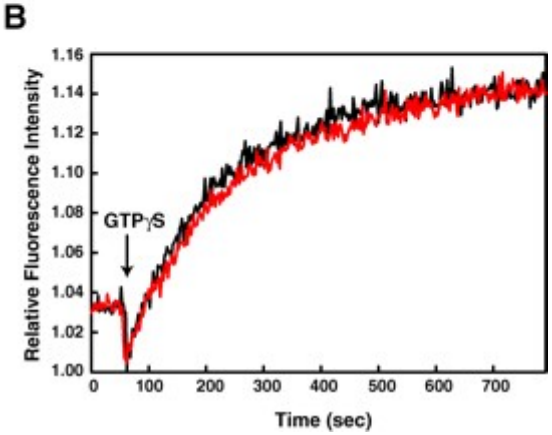
Belinda S. W. Chang, Karolina Jönsson, Manija A. Kazmi, Michael J. Donoghue, Thomas P. Sakmar

Molecular Biology and Evolution, Volume 19, Issue 9, 1 September 2002, Pages 1483–1489, https://doi.org/10.1093/oxfordjournals.molbev.a004211

Published: 01 September 2002 **Article history** ▾







THANKS