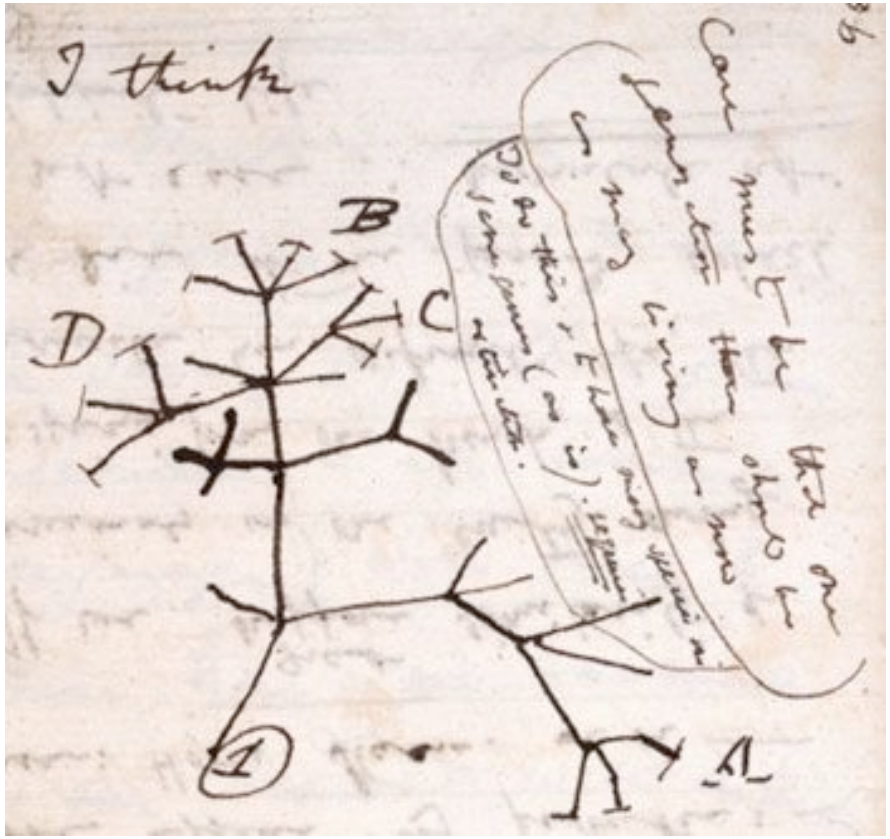
UBC Bioinformatics

Topic 9: Phylogenetic inference

Why do we need phylogenetics?

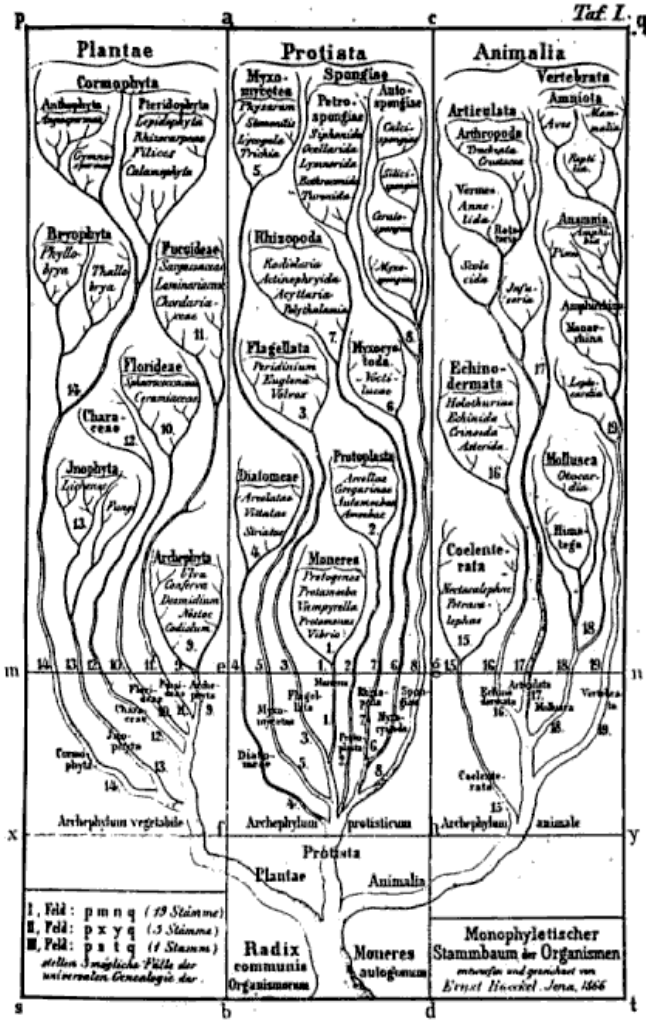
- Infer our evolutionary history
- Understand the diversity of life
- Determine how climate change may impact species distributions
- Determine how to develop drought/pest tolerant crops
- Development of new drugs and to synthesize materials
- Forensic scientists use phylogenetics in legal matters

Thinking about life

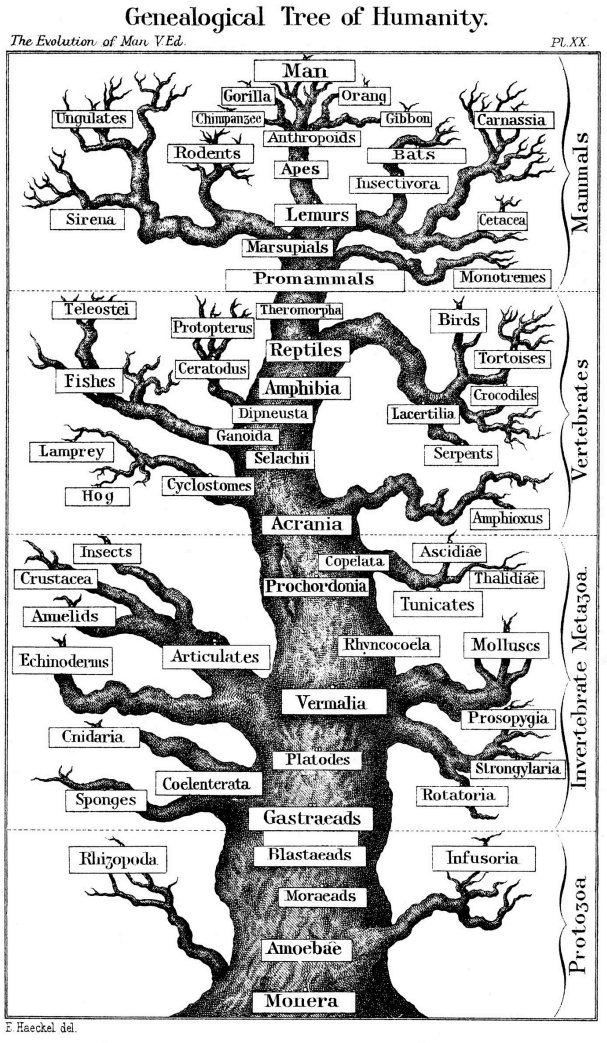


Darwin's "Transmutation of species" notebook (1837)

A phylogenetic view of biological diversification



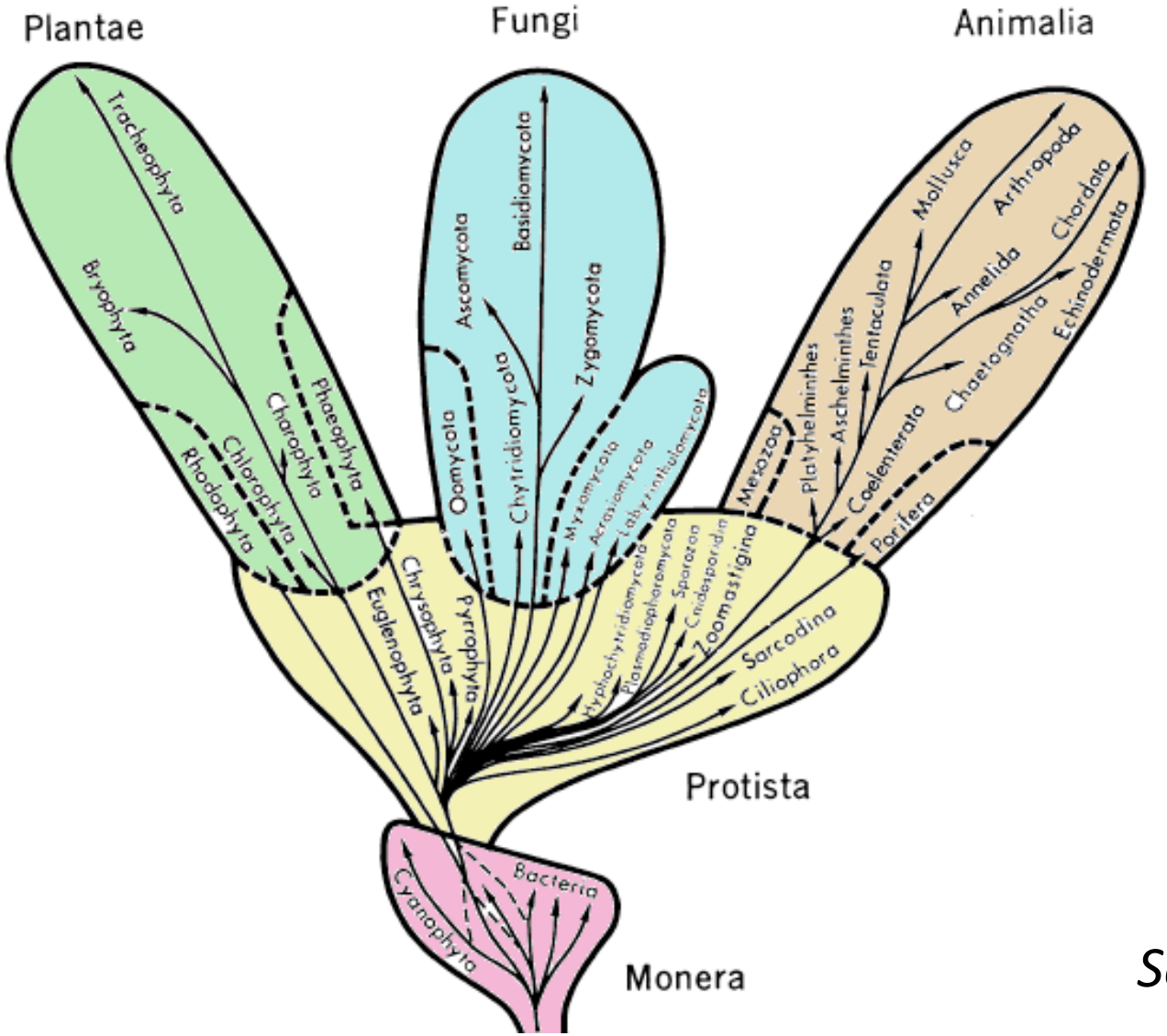
Haeckel's three kingdom system (1866)



Translation of Haeckel's Family Tree of the Human Race ("Anthropogenie" 1877)

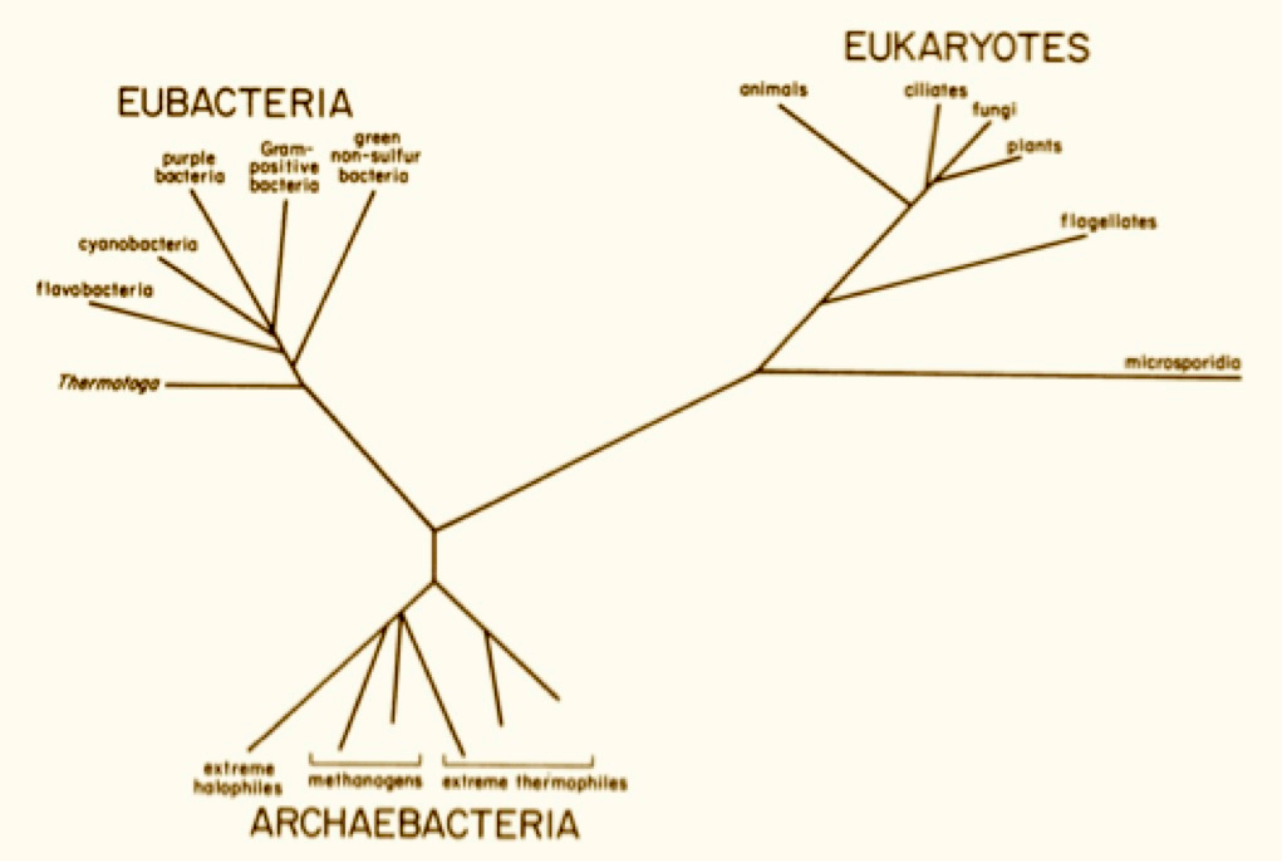
A phylogenetic view of biological diversification

Whittaker's five kingdom system



Science 1969

Carl Woese's three domain system (1977)



Analysis of ribosomal RNA sequences lead to Carl Woese to propose a three domain system (1977; PNAS 1990)

Basic components of phylogenetic analysis

- All phylogenetic analyses include four basic steps:

1) data acquisition

Basic components of phylogenetic analysis

- All phylogenetic analyses include four basic steps:

1) data acquisition
2) construction of character matrix (e.g. sequence alignment)

Basic components of phylogenetic analysis

- All phylogenetic analyses include four basic steps:

1) data acquisition
2) construction of character matrix (e.g. sequence alignment)
3) phylogeny estimation

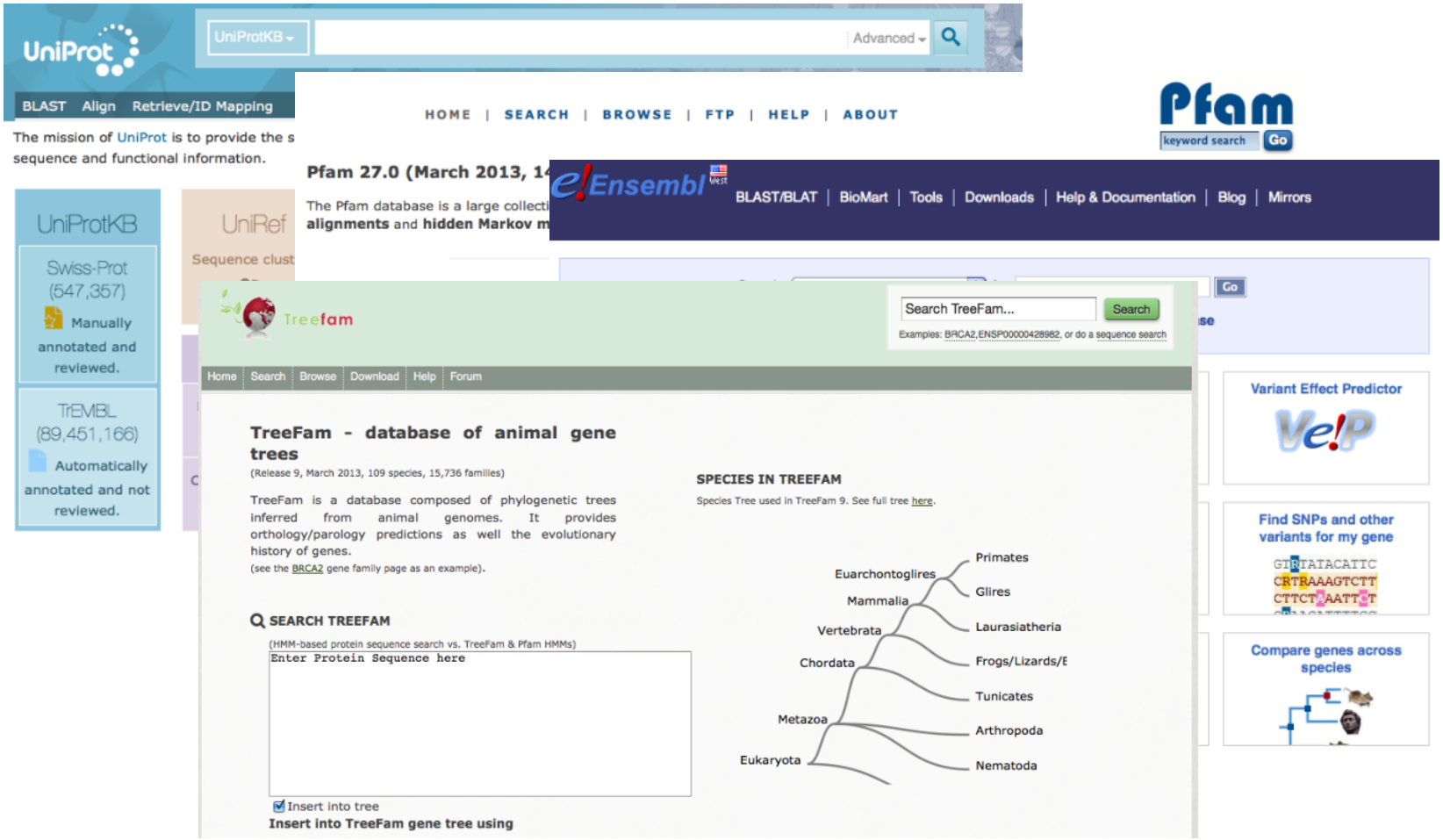
Basic components of phylogenetic analysis

- All phylogenetic analyses include four basic steps:

1) data acquisition
2) construction of character matrix (e.g. sequence alignment)
3) phylogeny estimation
4) interpretation of phylogenetic inference.

Sequence archives

Sequence archives



Constructing a distance matrix

Pairwise alignment – based on similarity or distance

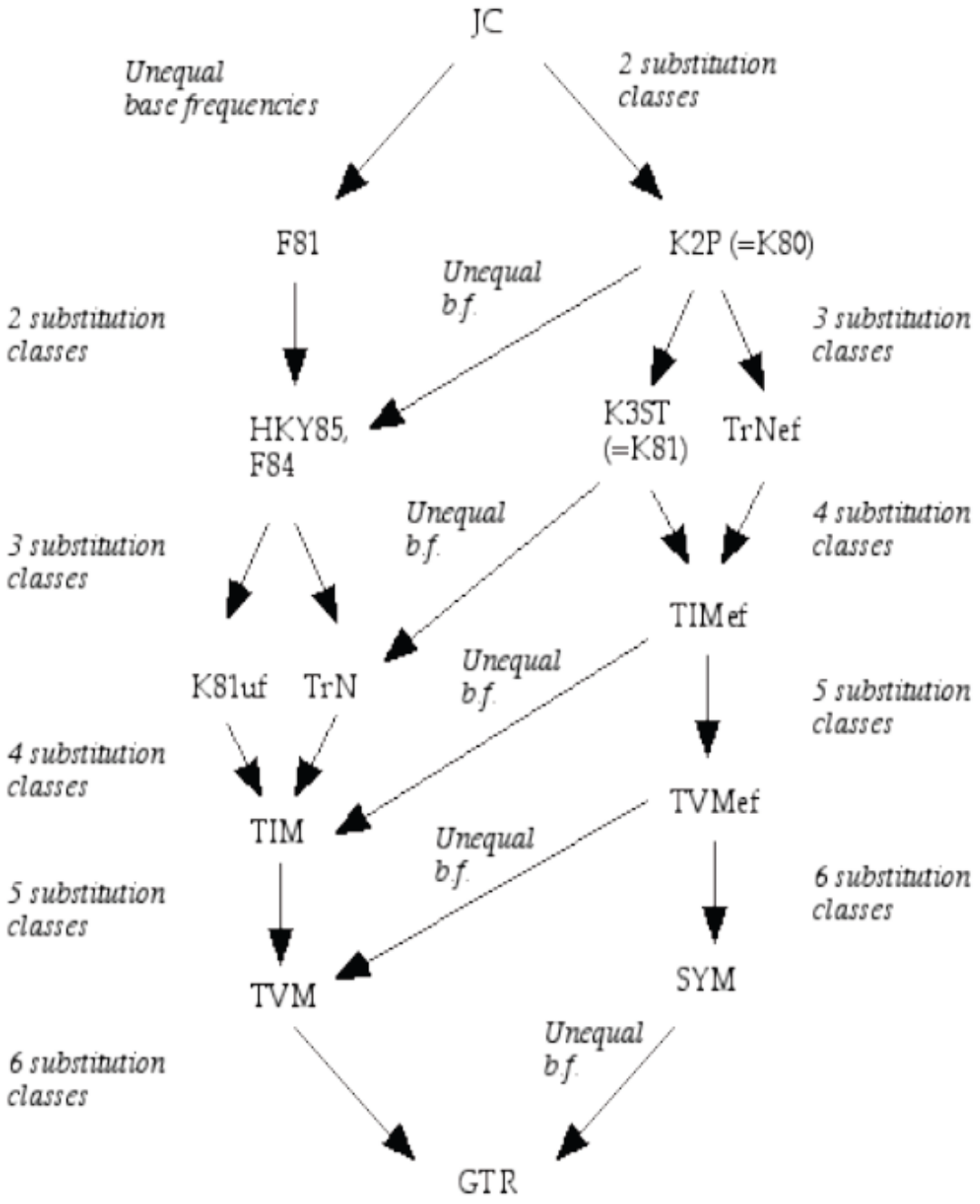
Edit distance – commonly, Hamming or Levenshtein distance between two strings

What about biological alphabets?
Are all changes equally likely?

Markov model of molecular evolution:

$P_{ij}(t) = e^{Qt}$

where Q is defined by one of the nucleotide substitution models shown to the right



Jukes-Cantor (1969) described simplest Markov model for nucleotide substitutions

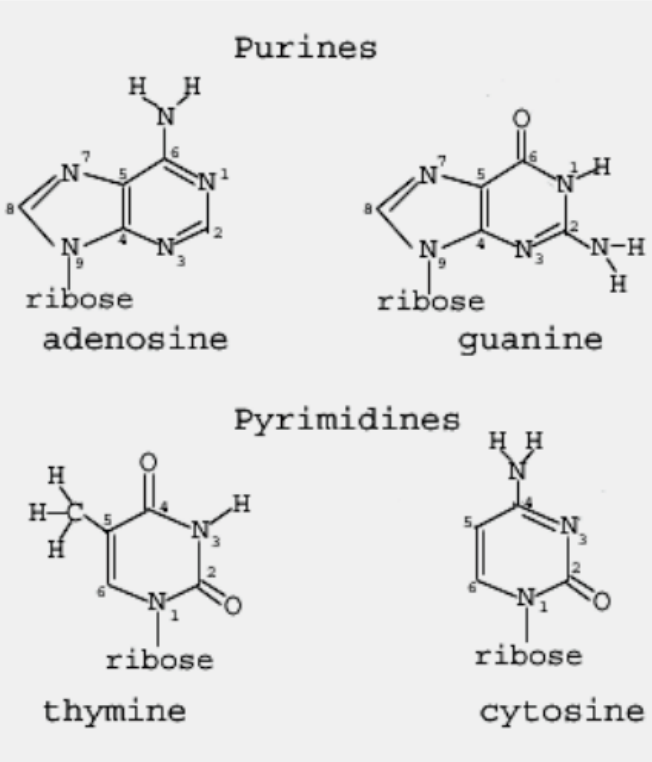
Pr(sequences differ at a position – I ≠ j) = 3/4 * (1- e-4/3t)
This implies equal probabilities for all substitutions (Qij = Qji)

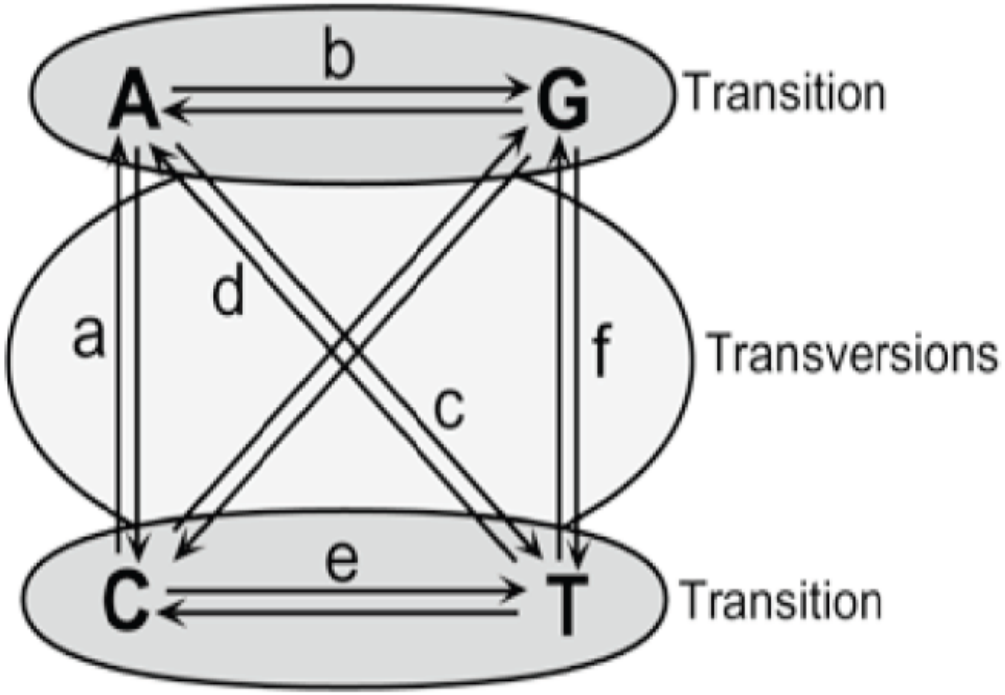
$$\mathcal{Q} = \begin{array}{c} \begin{matrix} T & C & A & G \end{matrix} \\ \begin{pmatrix} . & 1 & 1 & 1 \\ 1 & . & 1 & 1 \\ 1 & 1 & . & 1 \\ 1 & 1 & 1 & . \end{pmatrix} \end{array} \begin{matrix} T \\ C \\ A \\ G \end{matrix}$$

From

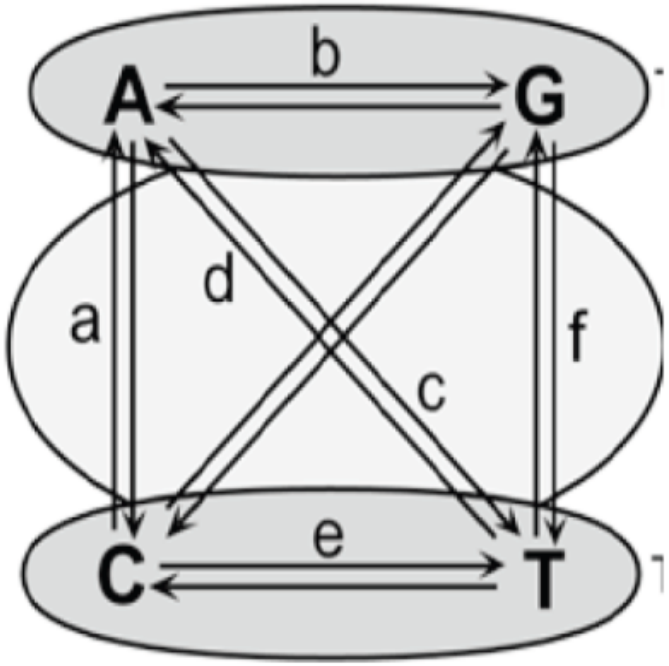
To

What about transition bias?





Kimura (1980) modified the JC model



$$\mathcal{Q} = \begin{array}{c} \begin{array}{cccc} \mathsf{T} & \mathsf{C} & \mathsf{A} & \mathsf{G} \end{array} \\ \begin{pmatrix} . & \kappa & 1 & 1 \\ \kappa & . & 1 & 1 \\ 1 & 1 & . & \kappa \\ 1 & 1 & \kappa & . \end{pmatrix} \end{array} \begin{array}{c} \mathsf{T} \\ \mathsf{C} \\ \mathsf{A} \\ \mathsf{G} \end{array}$$

What about variation in nucleotide frequencies? – Felsenstein (1981)

$$Q = \begin{array}{c} \begin{matrix} \text{T} & \text{C} & \text{A} & \text{G} \end{matrix} \\ \begin{pmatrix} . & \pi_C & \pi_A & \pi_G \\ \pi_T & . & \pi_A & \pi_G \\ \pi_T & \pi_C & . & \pi_G \\ \pi_T & \pi_C & \pi_A & . \end{pmatrix} \end{array} \begin{matrix} \text{T} \\ \text{C} \\ \text{A} \\ \text{G} \end{matrix}$$

Hasegawa, Kishino, and Yano (1985) put these together

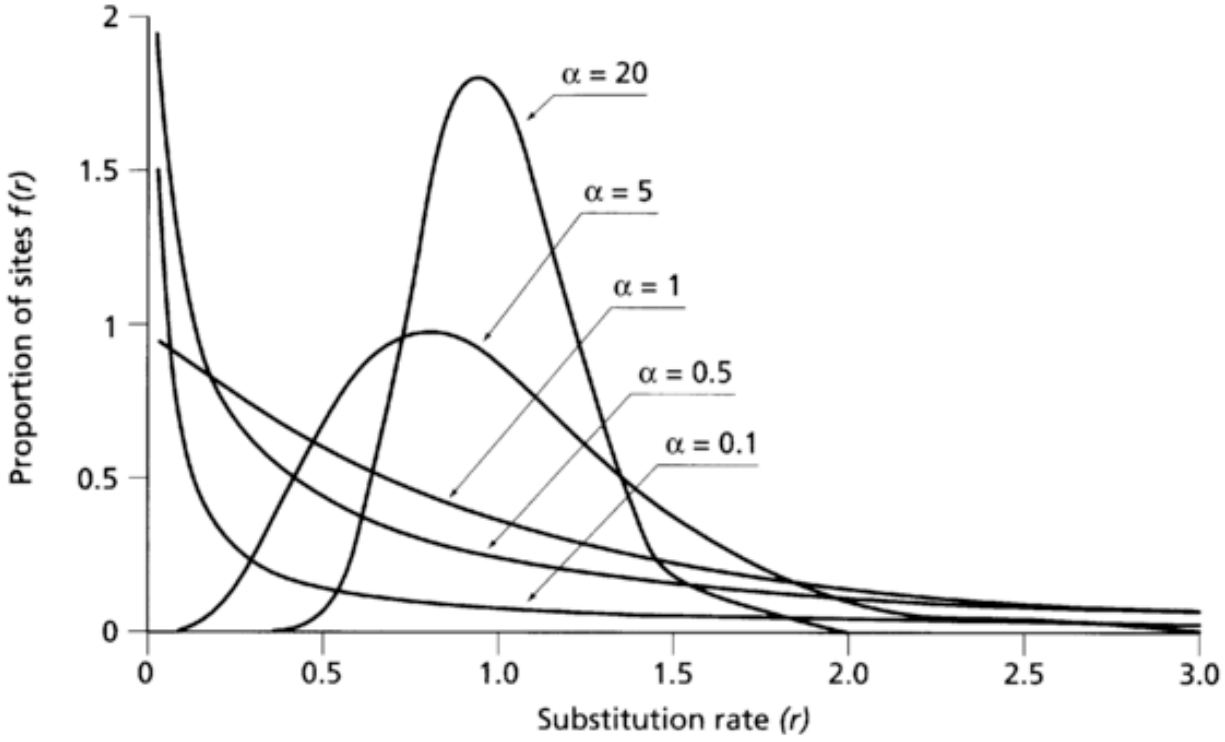
$$Q = \begin{array}{c} \begin{array}{cccc} \text{T} & \text{C} & \text{A} & \text{G} \end{array} \\ \begin{pmatrix} . & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & . & \pi_A & \pi_G \\ \pi_T & \pi_C & . & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & . \end{pmatrix} \end{array} \begin{array}{l} \text{T} \\ \text{C} \\ \text{A} \\ \text{G} \end{array}$$

General time reversible (GTR) model – Tavaré 1986

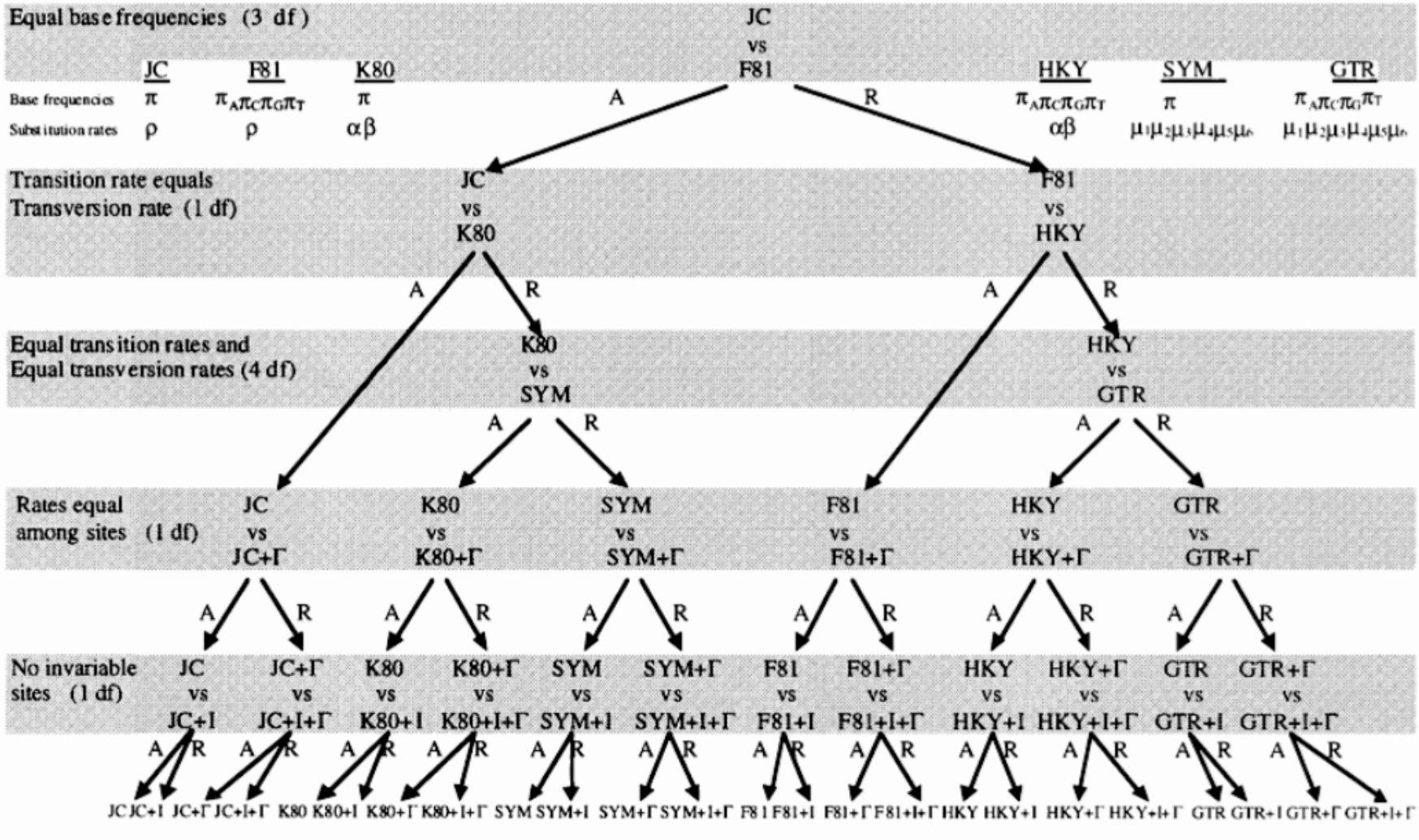
$$Q = \begin{array}{c} \begin{matrix} \text{T} & \text{C} & \text{A} & \text{G} \end{matrix} \\ \begin{pmatrix} . & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & . & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & . & \pi_G \\ c\pi_T & e\pi_C & \pi_A & . \end{pmatrix} \end{array} \begin{matrix} \text{T} \\ \text{C} \\ \text{A} \\ \text{G} \end{matrix}$$

Can we relax the assumption of equal rates across all sites?

- Allow a certain proportion of invariable sites (I)
- Model a distribution of rates - gamma distribution



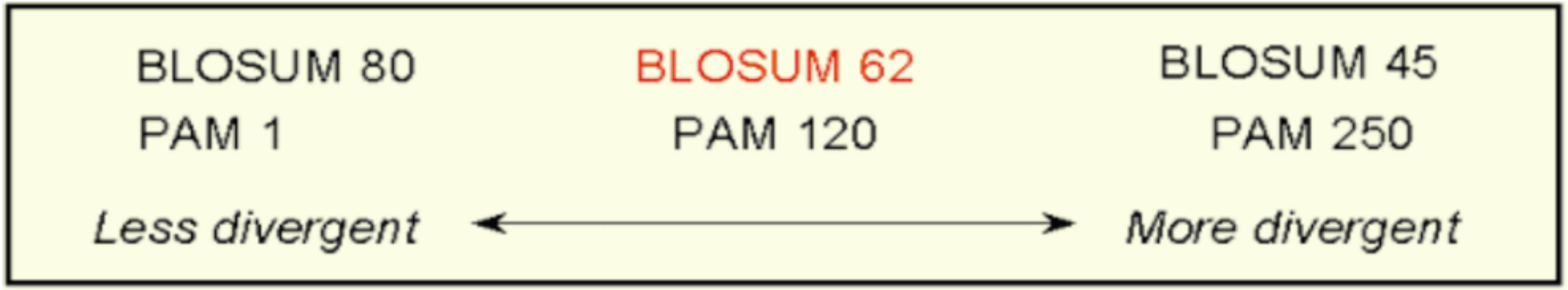
Modeltest (Posada and Crandall 1998)



Empirical estimation of AA transition frequencies

PAM (Point Accepted Mutation), JTT (Jones, Taylor, Thornton) and WAG (Whelan and Nick Goldman) matrices based on sequence comparisons in global alignments

BLOSUM (BLOcks of Amino Acid SUbstitution Matrix) matrices, based pairwise sequence comparisons, used for BLAST



BLOSUM62 substitution matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

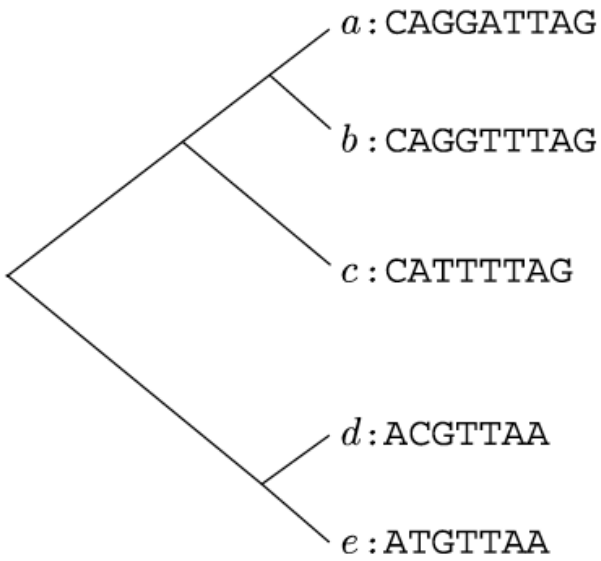
Progressive alignment

a: CAGGATTAG
b: CAGGTTTAG
c: CATTTTAG
d: ACGTTAA
e: ATGTTAA

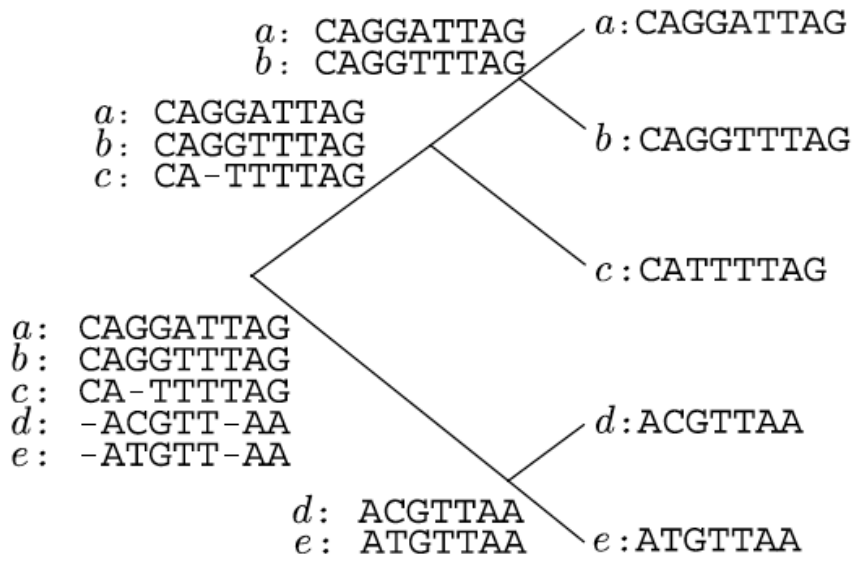
Input

	a	b	c	d	e
a	0	1	3	4	4
b	1	0	2	4	4
c	3	2	0	5	5
d	4	4	5	0	1
e	4	4	5	1	0

Distance matrix



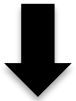
Guide tree



Progressive alignment

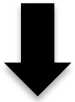
Iterative alignment

Distance matrix based on the the number of shared seeds



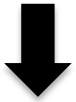
Construct guide tree

Progressive alignment



Re-construct guide tree

Re-alignment



Iterative refinement (based on likelihood)

Alignment methods (condensed)

Progressive

- ClustalW
- T-Coffee
- LALIGN
- PSAlign

Slow, inaccurate

Iterative

- MUSCLE
- MAFFT
- DIALIGN

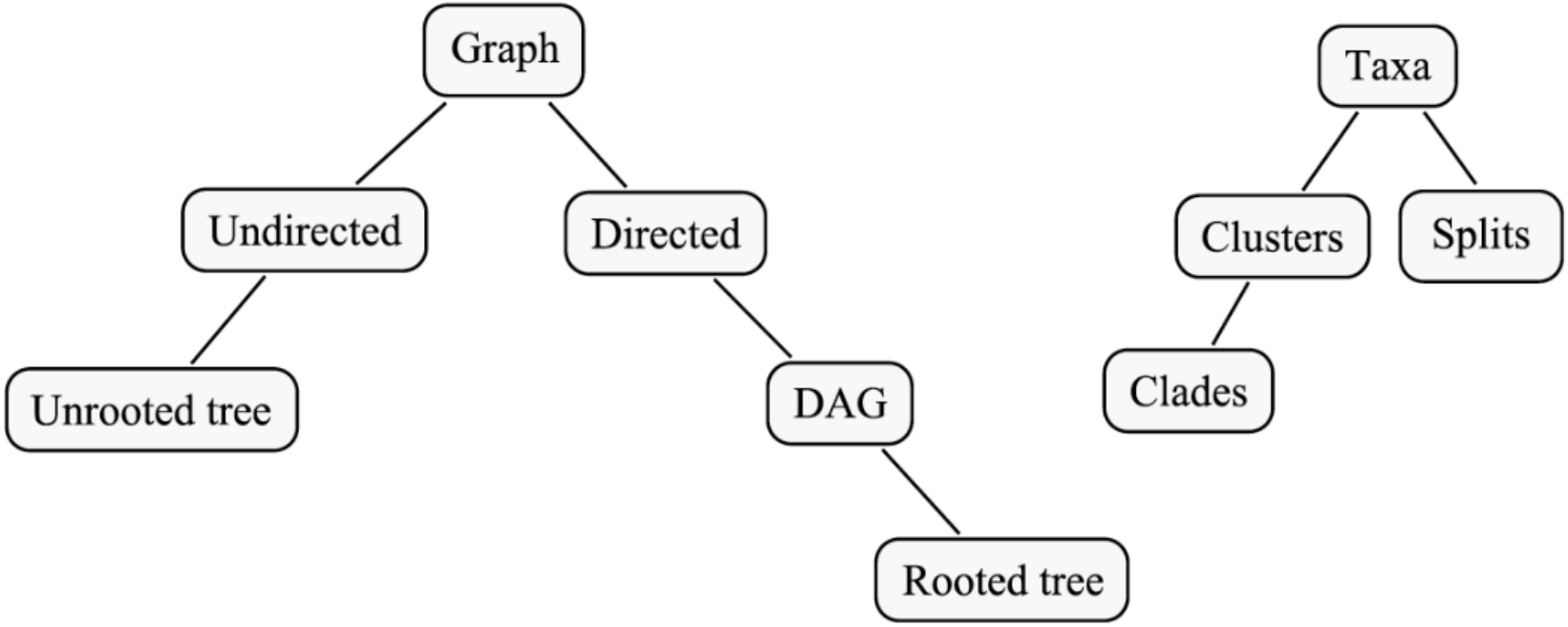
Fast, moderate-sized data

Other

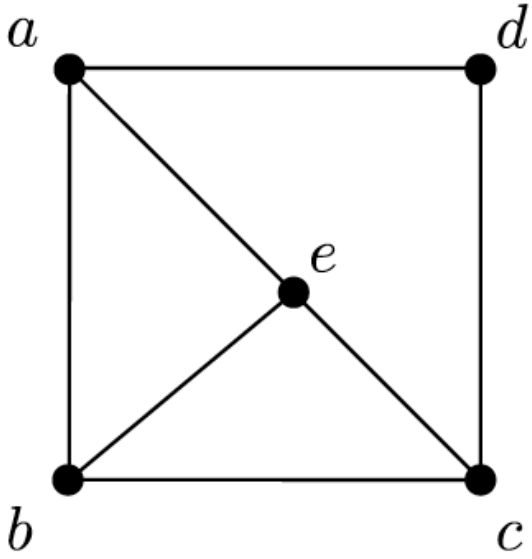
- Saté
- MAFFT
- Prank

Large data, accurate

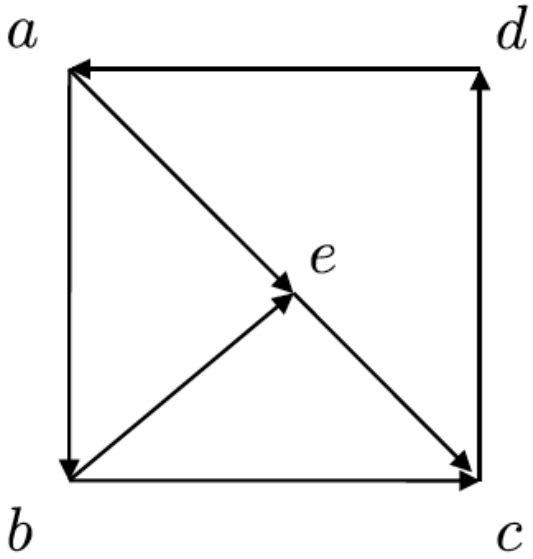
Basic graph types



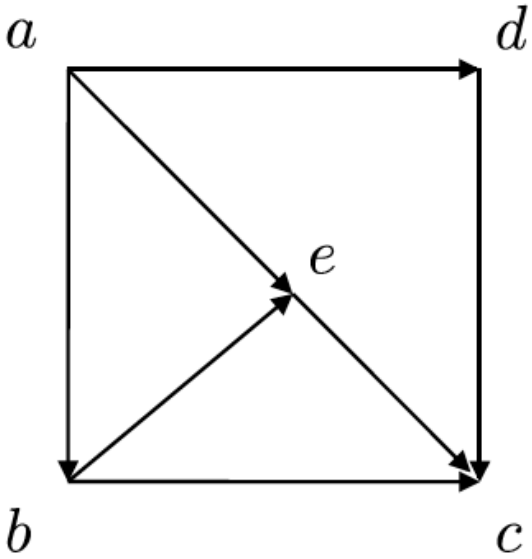
Basic graph types



Undirected graph

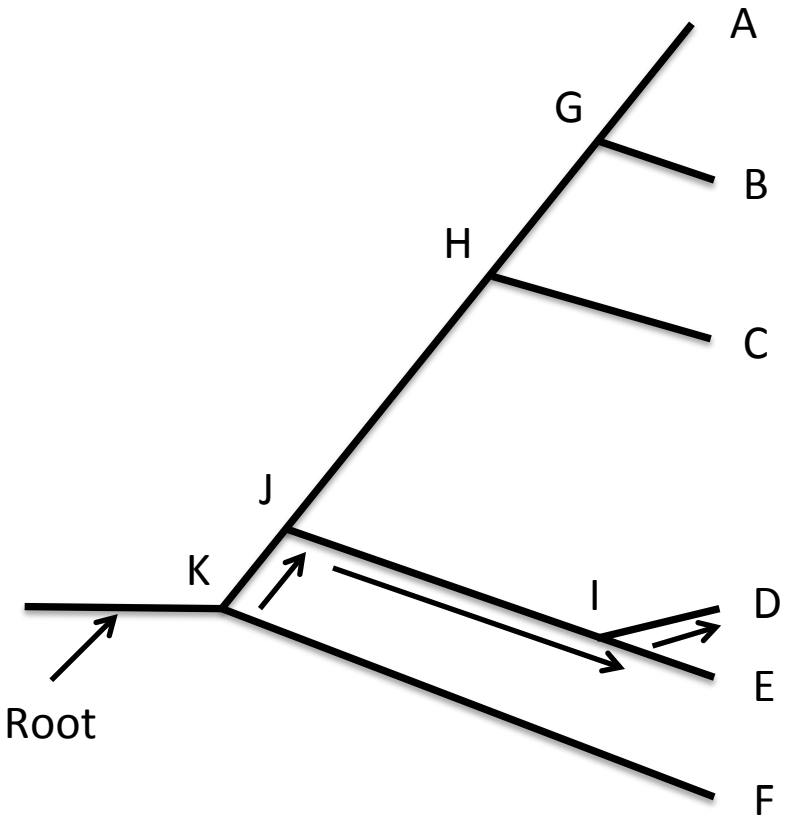


Directed graph

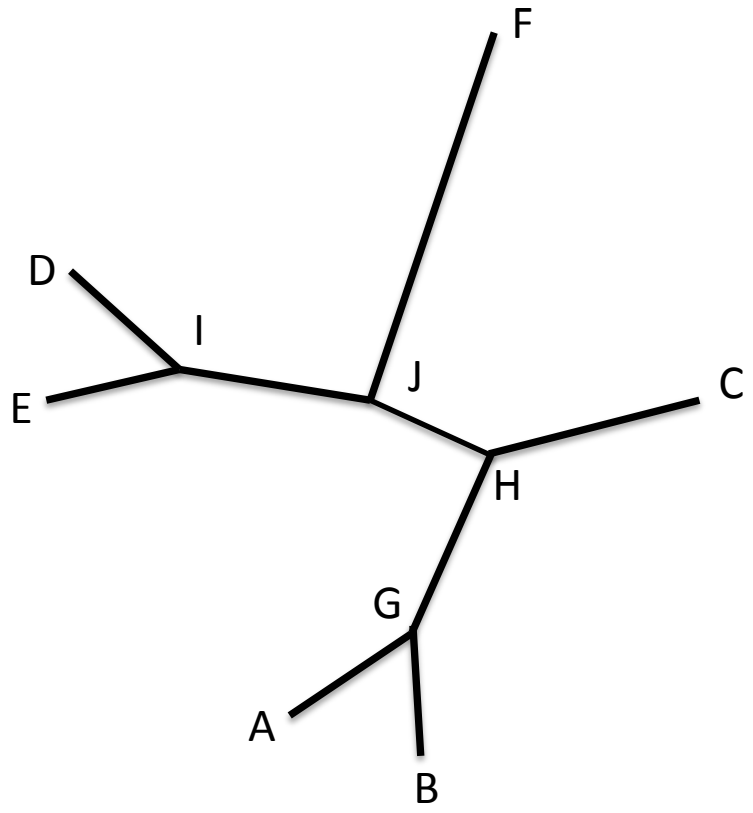


DAG

Basic tree types



Rooted tree



Unrooted tree

Phylogeny inference

There are four primary types of phylogenetic inference:

- **Distance based** – minimize distance between clusters
- **Parsimony** – minimum number of mutations
- **Maximum Likelihood** – maximizes likelihood of obtaining data, under a model of evolution
- **Bayesian methods** – compute posterior distribution of trees based on the data, model of evolution, and assumed prior dist. of trees

Support obtained from posterior distribution, or bootstrapping

Maximum Parsimony

- **Maximum Parsimony (MP)** – aims to find the tree with the fewest changes

- Calculate the minimum number of changes at each position along the branches to explain the observed terminal nodes

- ***Parsimony length*** – sum of scores for all positions

Maximum Parsimony

Caveats:

Always underestimates the real divergence between distantly related taxa

There is no way to correct for multiple substitutions, this may lead to long branch attraction*

*Substitution models can account for increased prob. change on long branches

Distance-based methods

Unweighted pair group method with arithmetic means (UPGMA) – Clustering by searching for the smallest distance in the pairwise matrix

- repeats by finding distance to new clusters
- averages distance between original clusters
- assumes the rate is the same on all branches

Distance-based methods

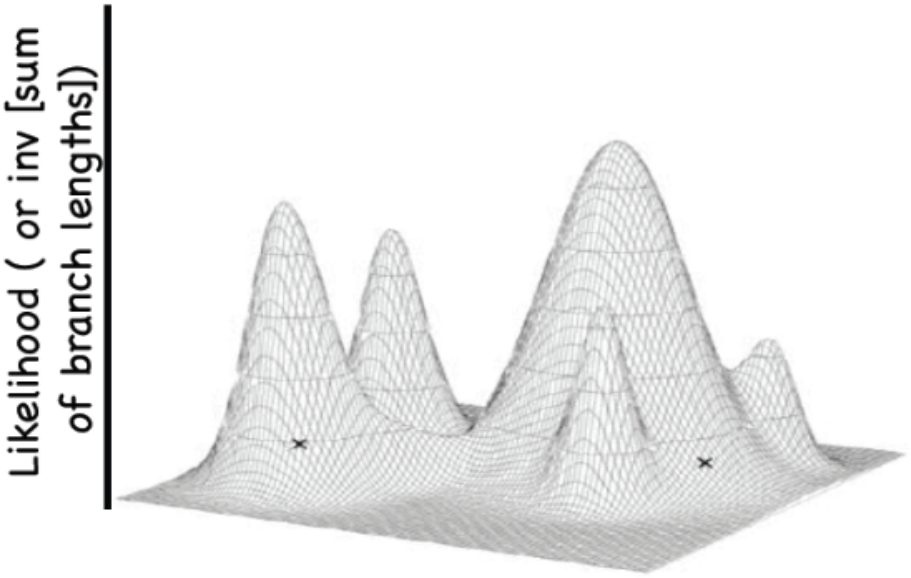
Neighbor-Joining (NJ) – tree constructed by sequentially finding pairs

- doesn't attempt to cluster based on similarity
- minimizes length of all internal branches (and tree)
- starts with a star-like tree, adds a branch, recalculates tree

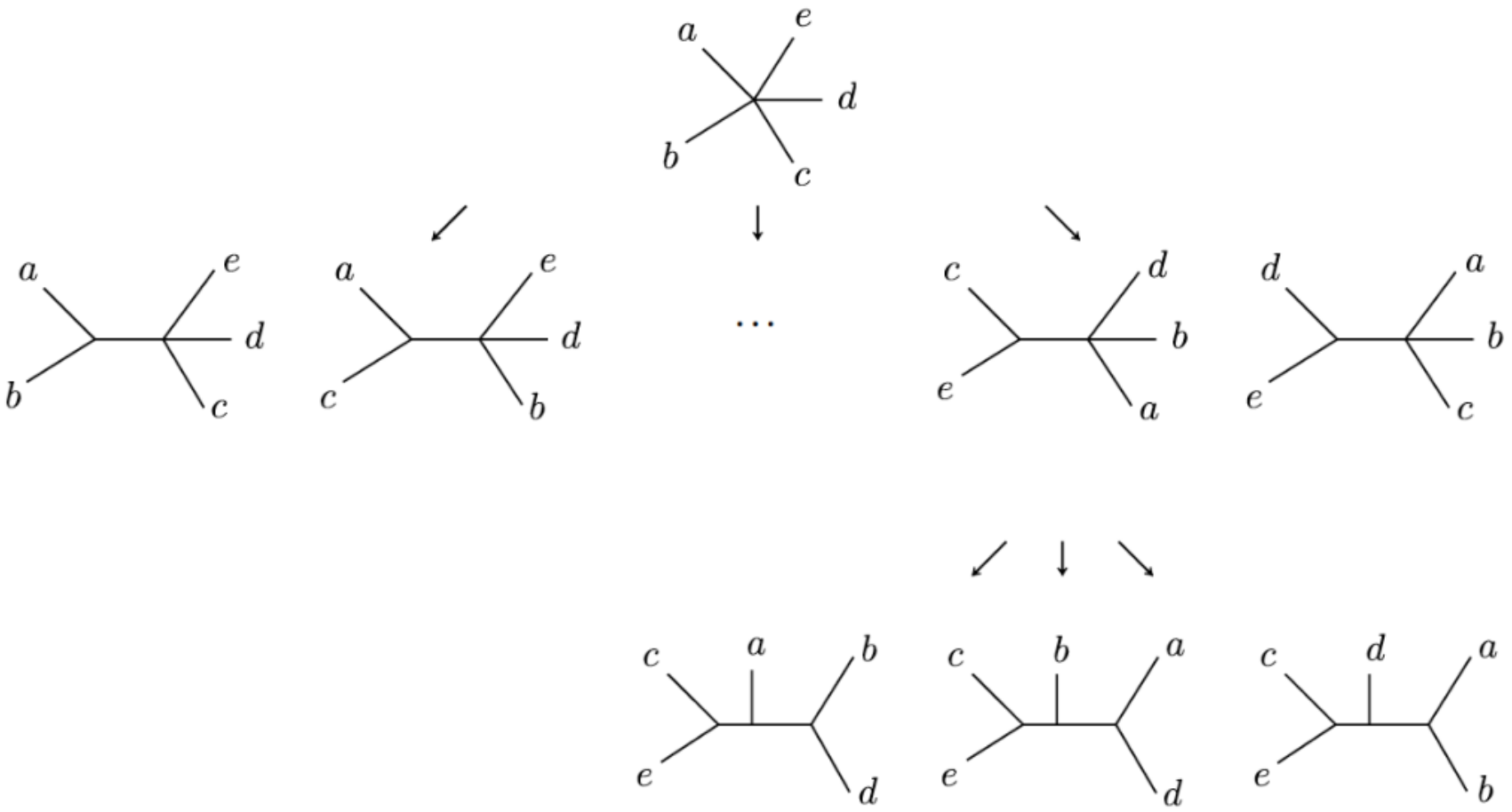
How to assess confidence in the resulting tree?

Tree search problem

Number of taxa	Number of trees
10	2×10^{6}
22	3×10^{23}
50	3×10^{74}
100	2×10^{182}
1,000	$2 \times 10^{2,860}$
10,000	$8 \times 10^{38,658}$
100,000	$1 \times 10^{486,663}$
1,000,000	$1 \times 10^{5,866,723}$
10,000,000	$5 \times 10^{68,667,340}$



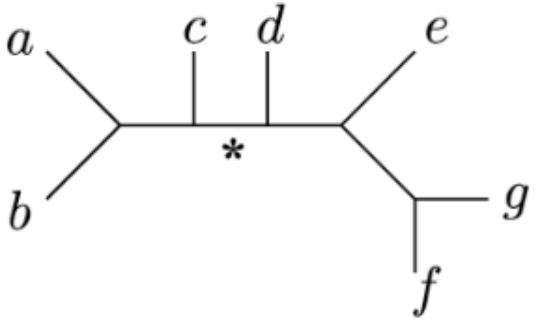
Star-like decomposition



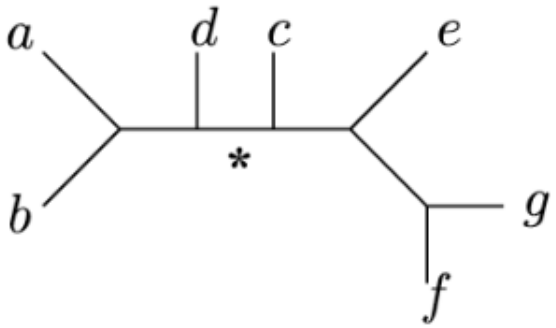
Branch swapping can refine trees initially found in heuristic search

- **Nearest-Neighbor Interchange (NNI)**
- **Subtree Pruning & Regrafting (SPR)**
- **Tree Bisection & Reconnection (TBR)**

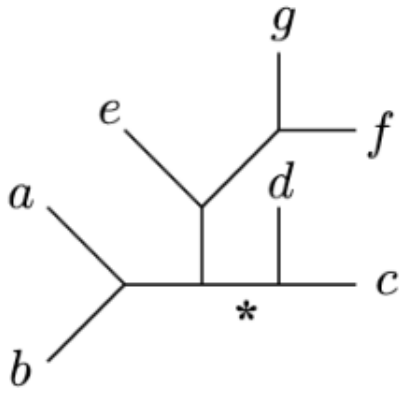
Nearest neighbor interchange (NNI)



Phylogenetic tree

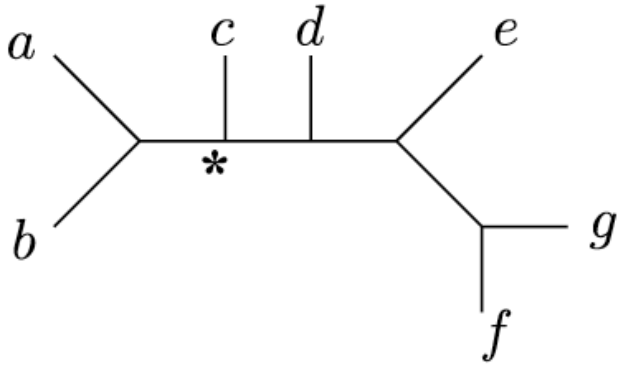


One NNI tree

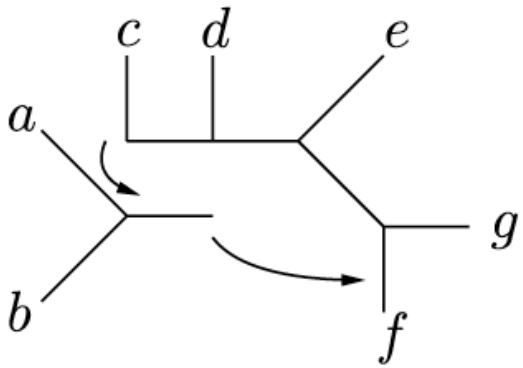


The other NNI tree

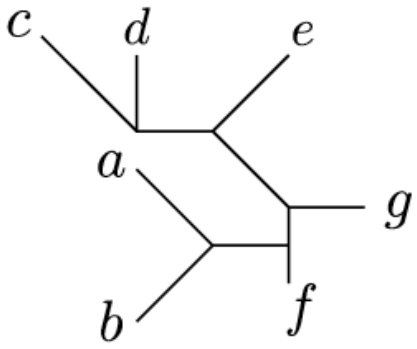
Subtree prune and regraft (SPR)



Phylogenetic tree

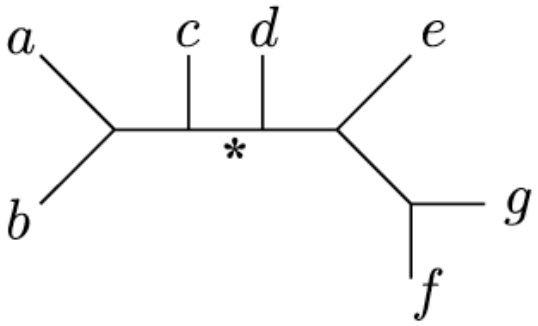


Subtree prune…

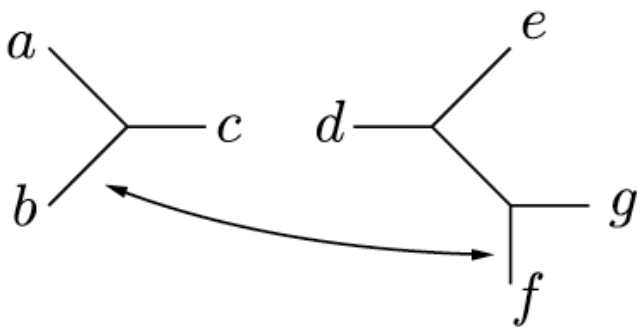


Regraft

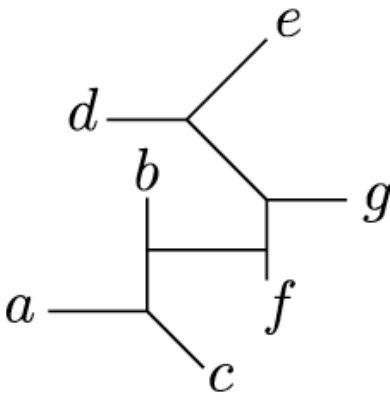
Tree bisection and reconnection (TBR)



Tree bisection…



Reconnection choice…



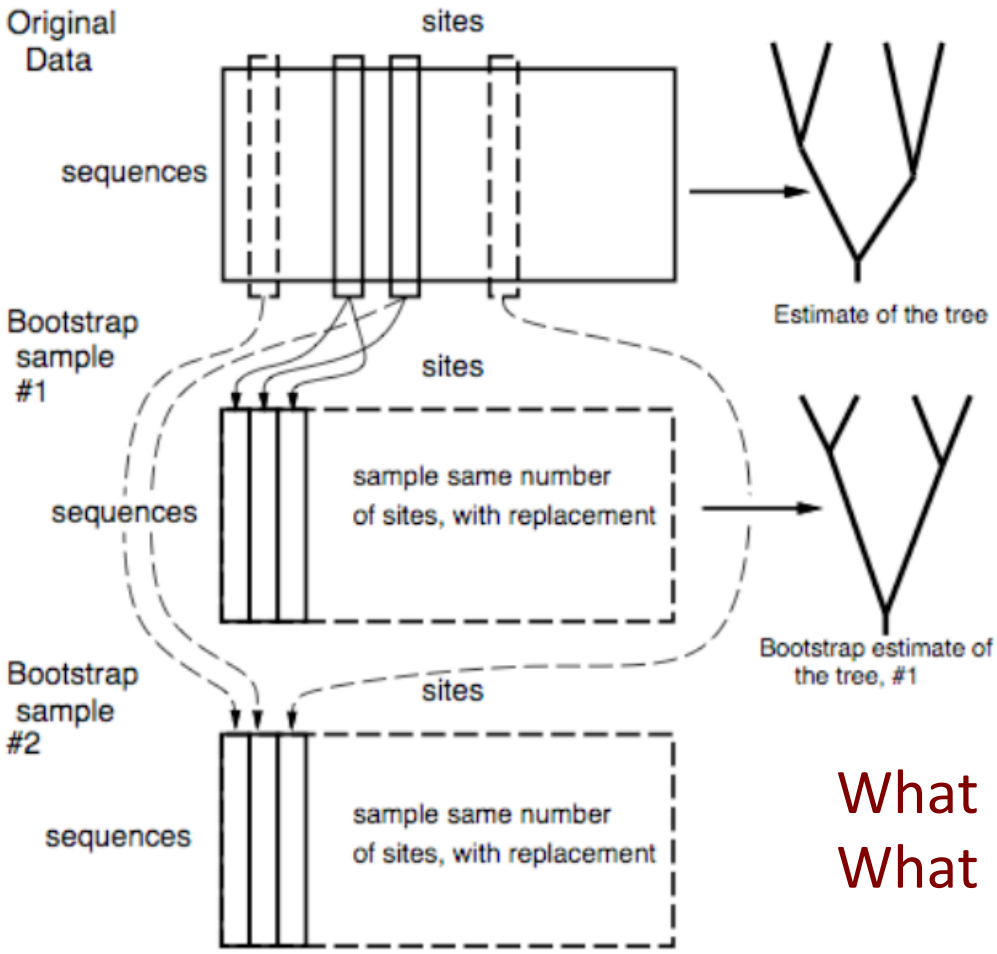
And reconnection

Which topology is correct?

$$NNI(\mathbf{T}) \subseteq SPR(\mathbf{T}) \subseteq TBR(\mathbf{T})$$

- The size of NNI(**T**) grows linearly with **n**
- The size of SPR(**T**) grows quadratically
- The size of TBR(**T**) depends not only on **n**, but also on the actual topology of the phylogenetic trees considered

Nonparametric bootstrapping



What if the tree is wrong?
What if the model is wrong?

Likelihood methods

- In statistics we usually estimate the probability of a hypothesis given data:

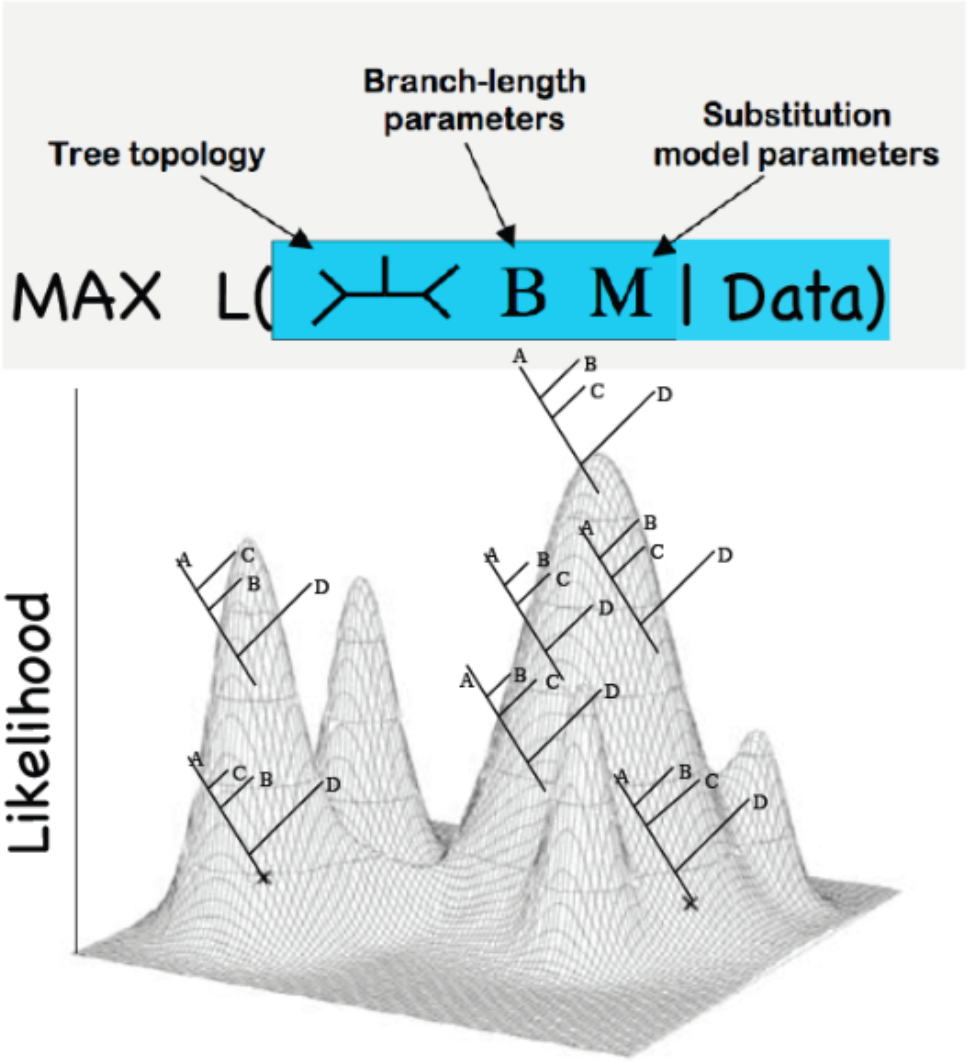
  Pr (Hypothesis|Data)

- But Likelihoods measure the probability of the data given a hypothesis:

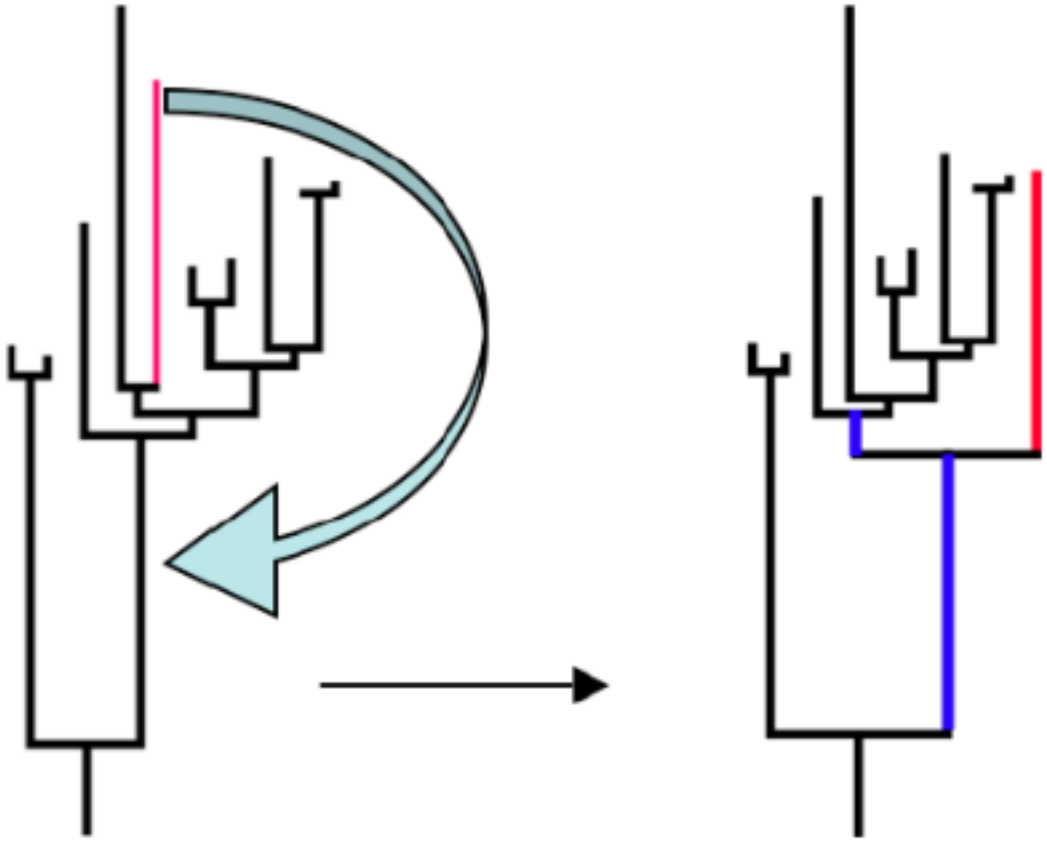
  L (Tree|Data) = Pr (Data|Tree)

- Any estimate of **Pr (Data|Hypothesis)** requires a model: e.g. the binomial for coin flips, and a substitution model for molec. phylogenetics

Likelihoods in phylogenetic inference



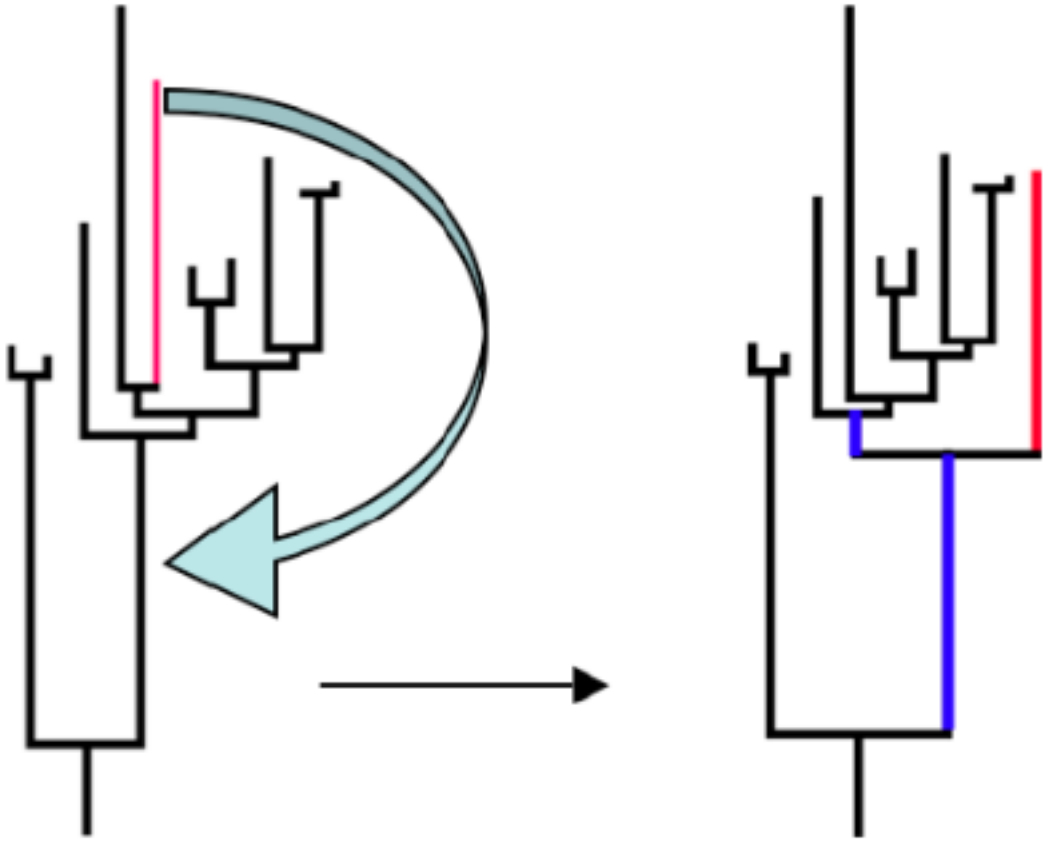
Approximate likelihood



Garli only reoptimizes Branch lengths in the old and new neighborhood of pruned and regrafted branch

Zwickl 2007

RAxML employs a similar approach with SPR swapping on a MP tree



Bootstrap analysis with random addition MP tree for each round is highly recommended…??

Stamatakis

Bayesian inference

X = Data

Θ = Model parameters

Posterior distribution

Prior distribution

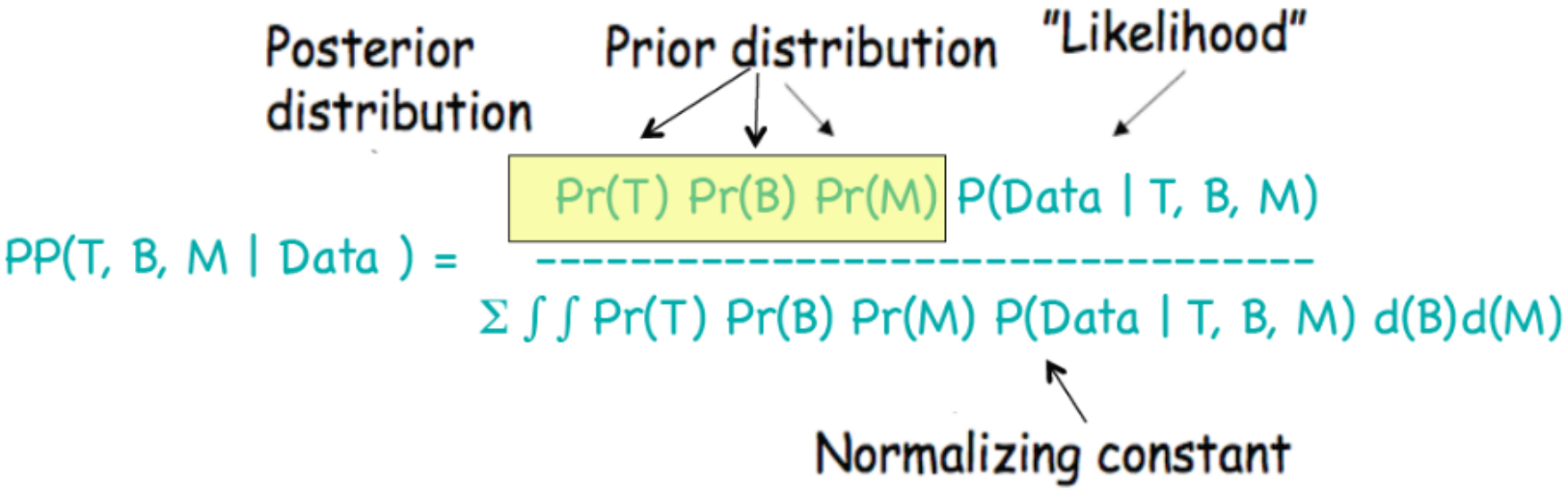
Likelihood

$$f(\theta \mid X) = \frac{f(\theta) f(X \mid \theta)}{\int f(\theta) f(X \mid \theta) d\theta}$$

Normalizing constant

Bayesian inference

"Model Parameters" may include tree topology, branch lengths and substitution model parameters



Ronquist 2007

Bayesian inference

Pros:

- Can incorporate prior information (divergence times, subs. rates across genes)
- Can impose parameter constraints with a prior
- Allows inference across complex tree and parameter space (rather than maximization)

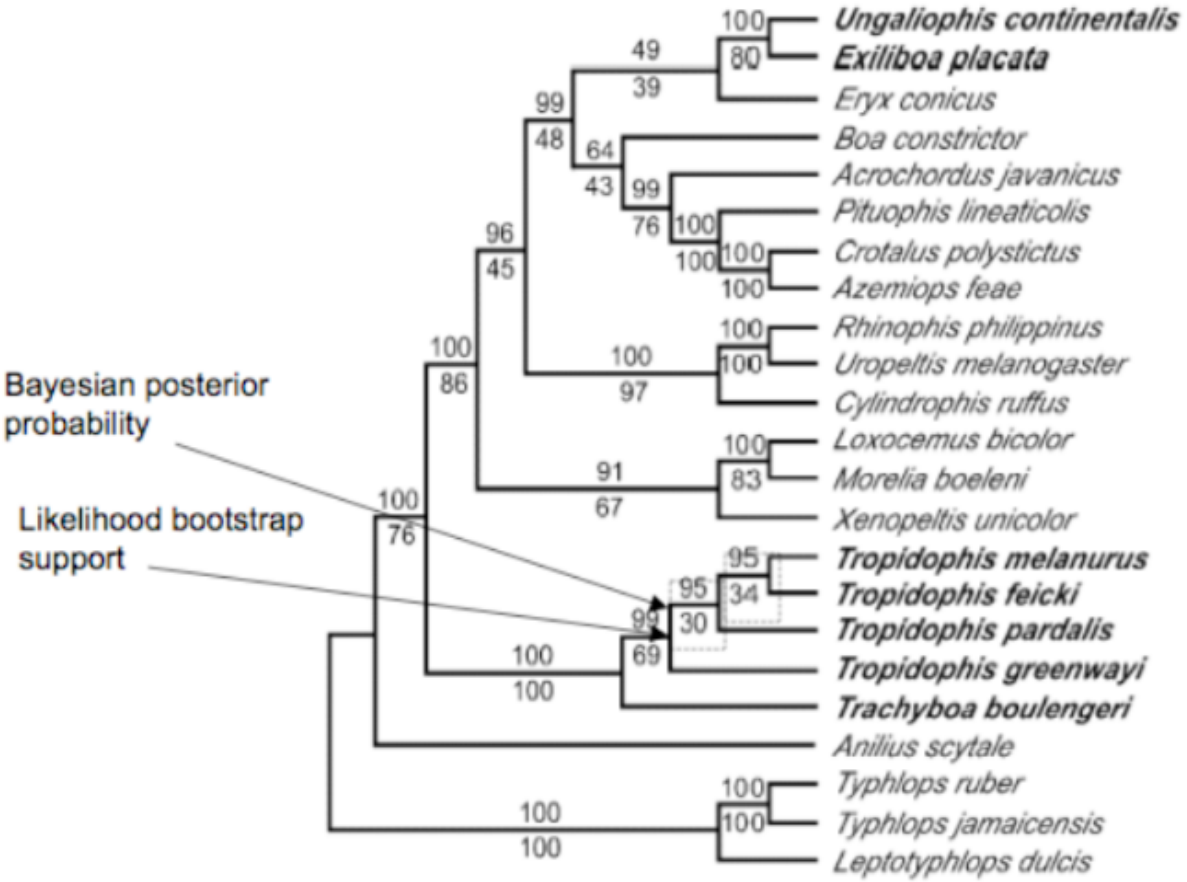
  - this simplifies searches for large trees

Bayesian inference

Minuses:

- Must incorporate information about parameters and that is not always available (e.g., influence of branch length prior and subs. model on posterior probability)
- Numerical methods for inferring posterior probabilities are difficult to diagnose (convergence in MrBayes)

How to assess posterior probability and ML bootstrap support?



KH (Kishino-Hasegawa), SH (Shamodaira-Hasegawa), AU (Approximately Unbiased) Tests - Implementations

- Propose a set of hypothesized trees and substitution model
- Estimate likelihoods for each tree:

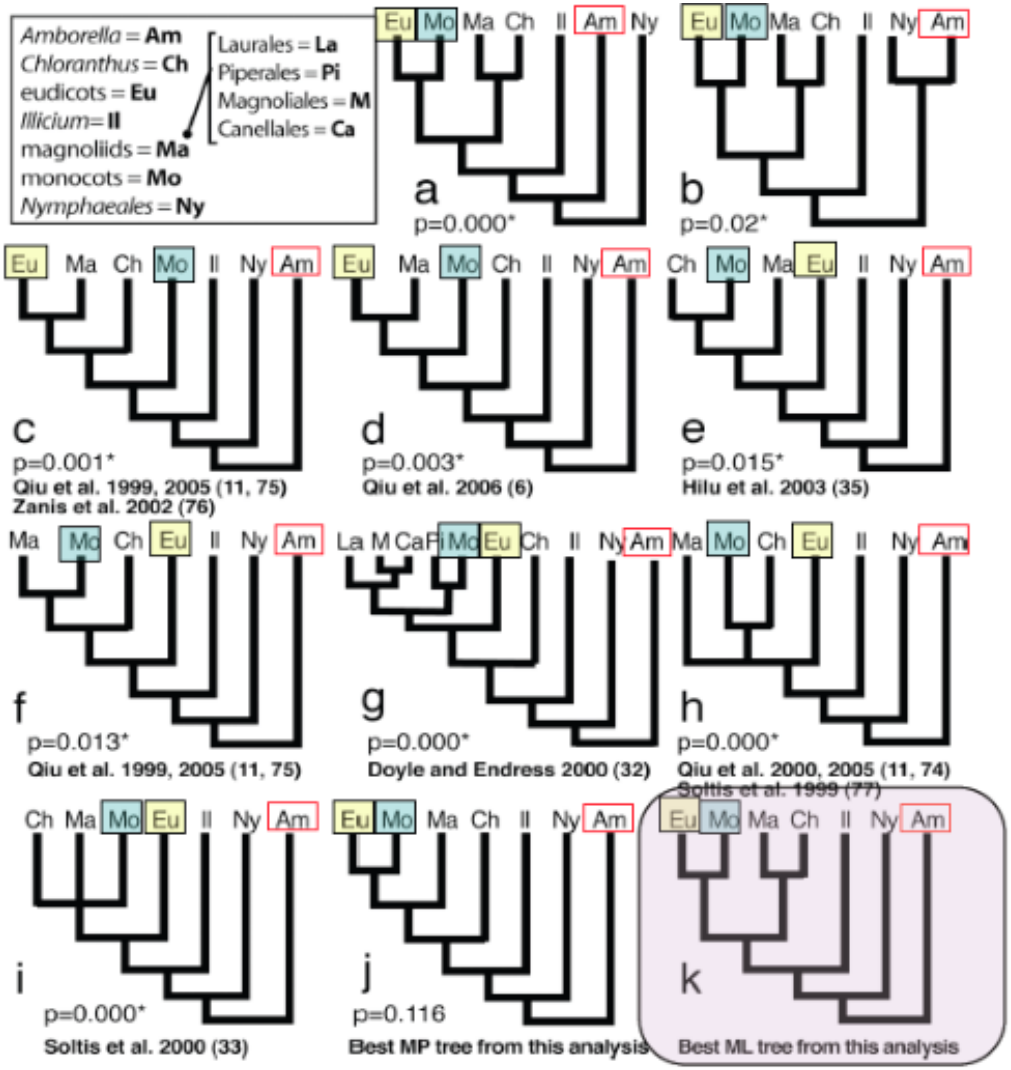
  Max L (B, M | Tree, Data)

KH test compares pairs of trees by assessing difference in likelihood

SH and AU tests bootstrap resample the original data matix thousands of times

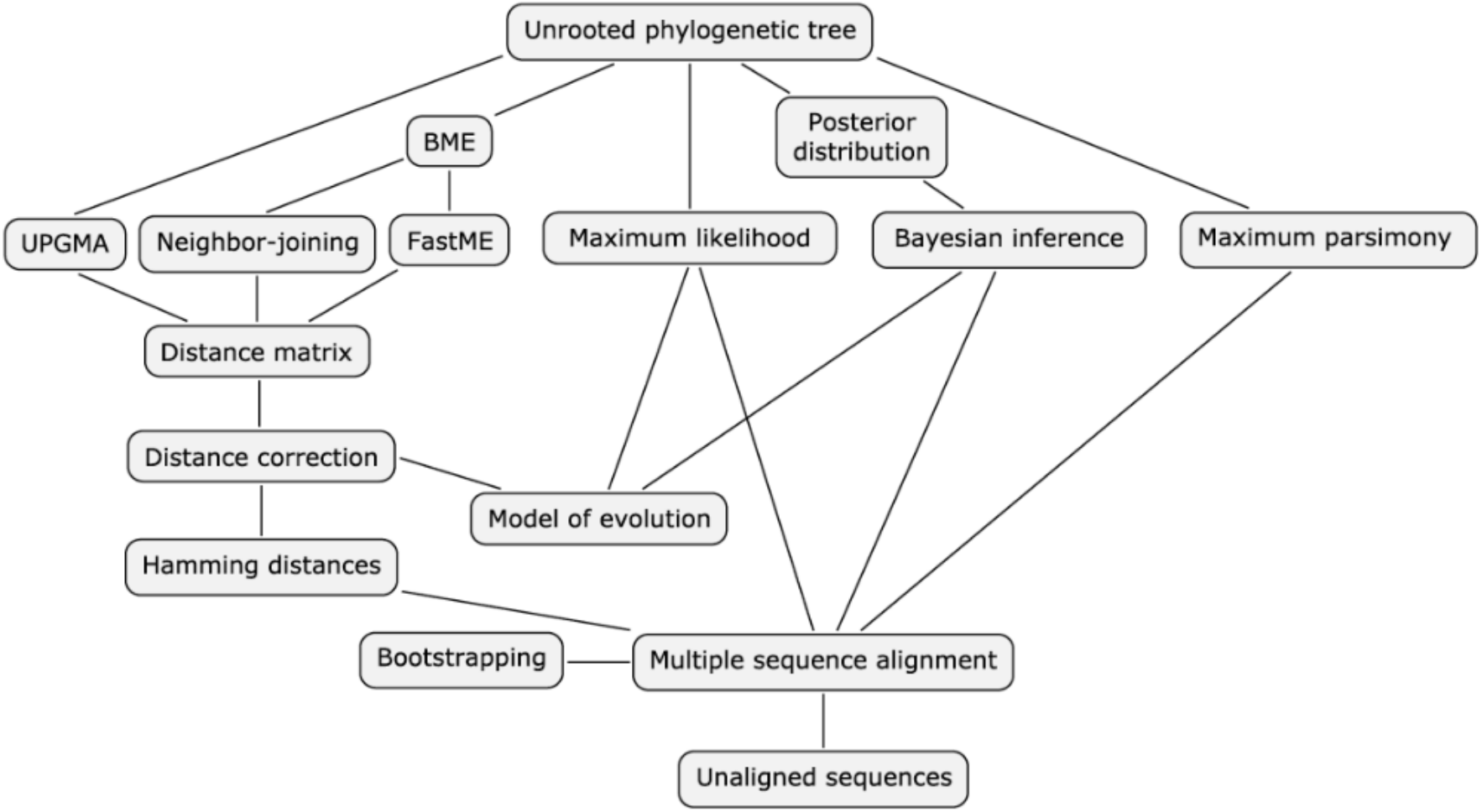
Likelihoods compared for best tree vs. alternative trees

SH, AU tests of alternative topologies

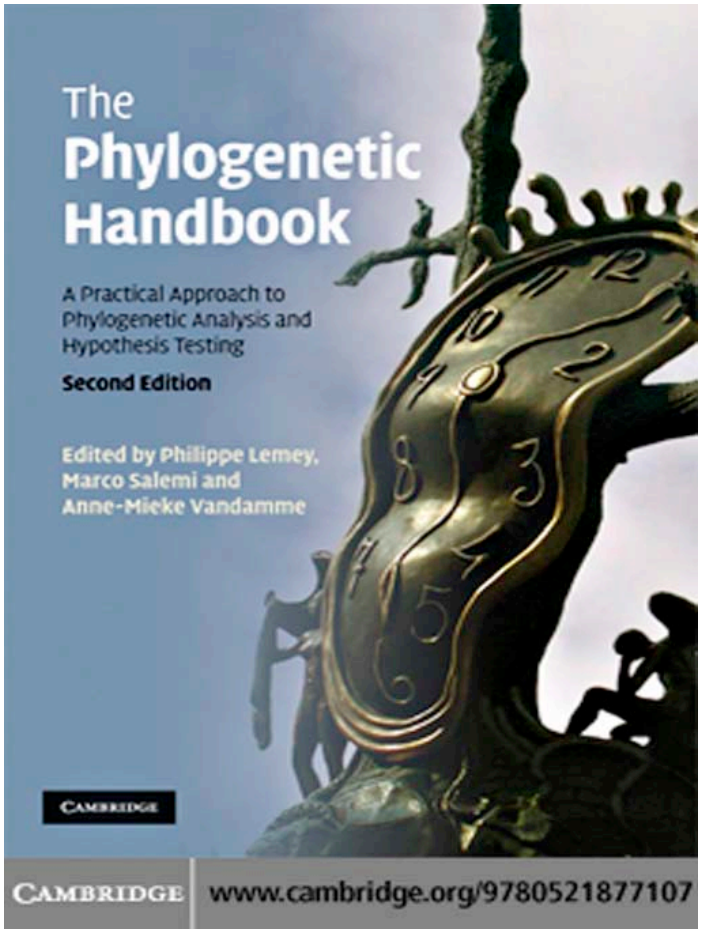


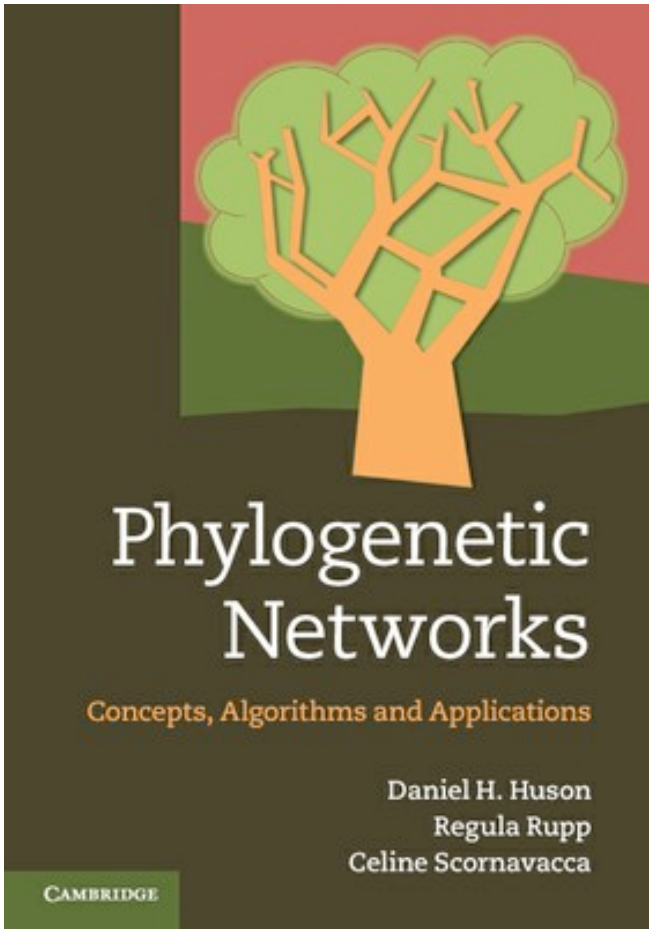
Janesen et al. PNAS 2007

High level view of phylogeny inference



Resources


The
Phylogenetic
Handbook
A Practical Approach to
Phylogenetic Analysis and
Hypothesis Testing
Second Edition
Edited by Philippe Lemey,
Marco Salemi and
Anne-Mieke Vandamme
CAMBRIDGE
CAMBRIDGE
www.cambridge.org/9780521877107


Phylogenetic
Networks
Concepts, Algorithms and Applications
Daniel H. Huson
Regula Rupp
Celine Scornavacca
CAMBRIDGE

Practical session

- Exercise list is online:

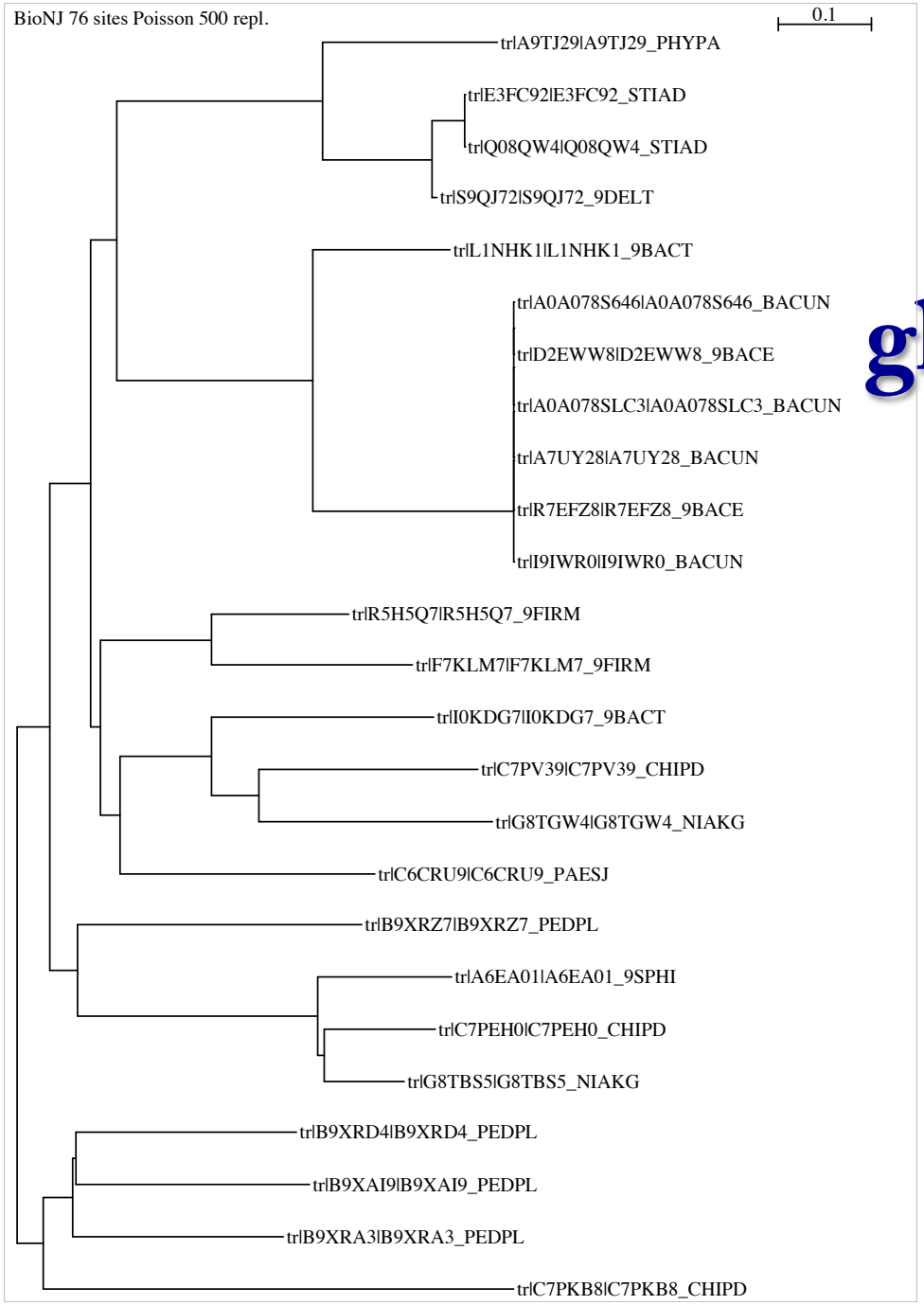
github.com/UBCBio525/Bio525D/Day8

Review of yesterday's findings

- Species identification – *Physcomitrella patens*
- BLAST hit summary
- Enrichment findings

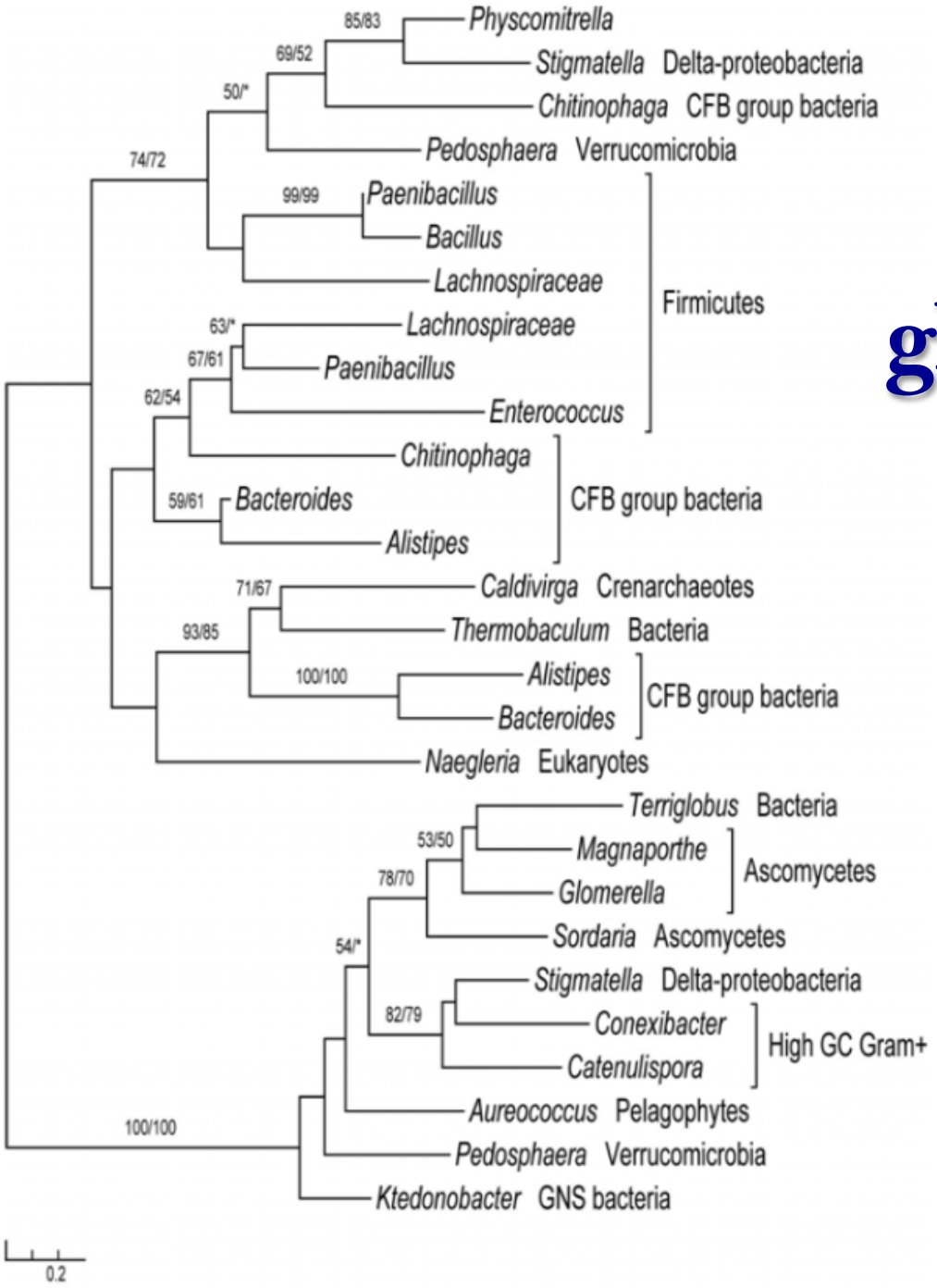
- Evolutionary hypothesis?

- Results of phylogenetic inference….

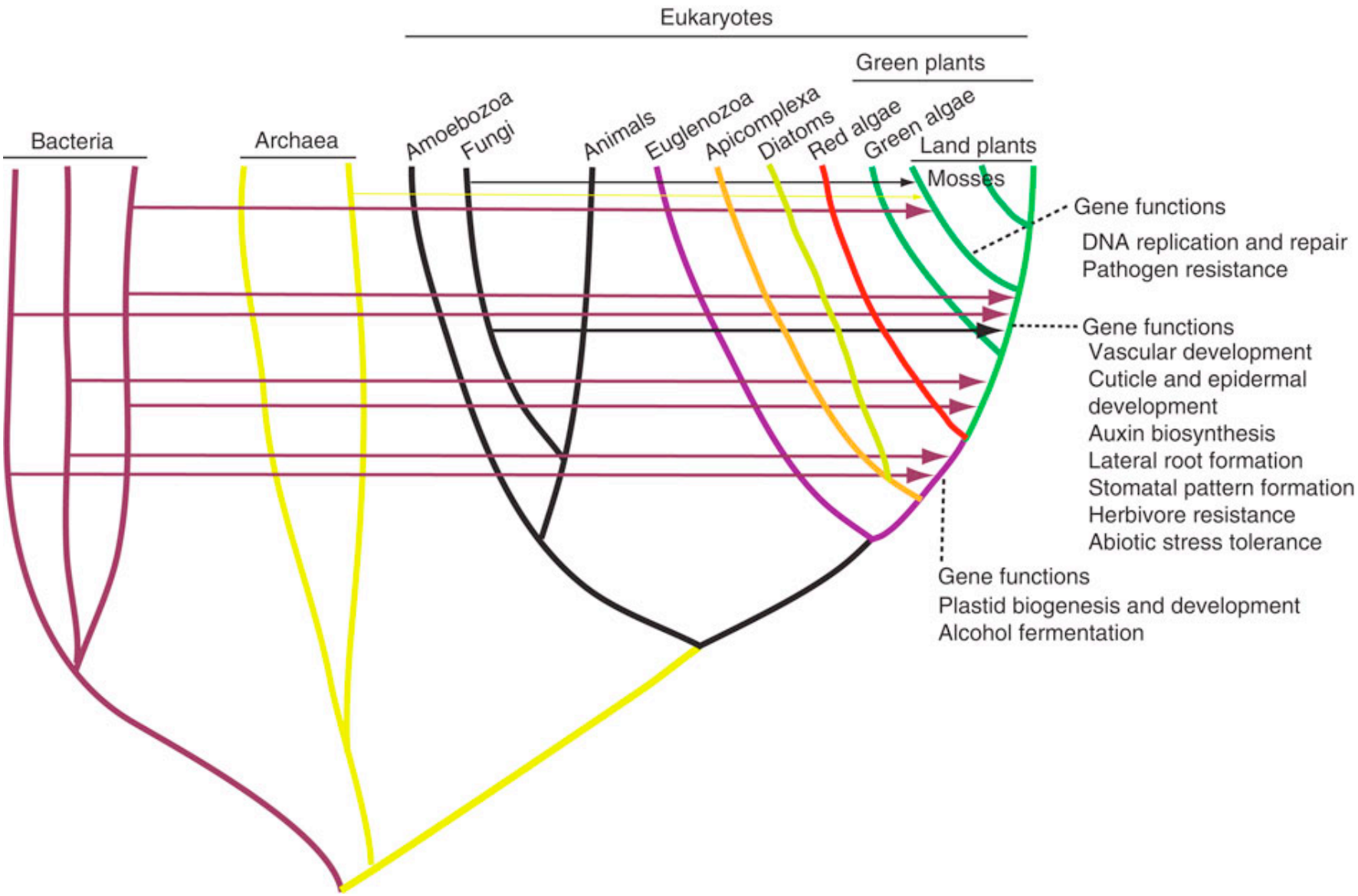


Phylogeny of glycoside hydrolase

Phylogeny of glycoside hydrolase



Yue et al. 2012 *Nature Comm.*



Yue et al. 2012 *Nature Comm.*

Reference (and data source)

nature COMMUNICATIONS

OPEN

ARTICLE

Received 10 Jun 2012 | Accepted 20 Sep 2012 | Published 23 Oct 2012

DOI: 10.1038/ncomms2148

Widespread impact of horizontal gene transfer on plant colonization of land

Jipei Yue[1,2], Xiangyang Hu[1,3], Hang Sun[1], Yongping Yang[1,3] & Jinling Huang[2]