# Lecture 7: Evolutionary analysis, phylogenetic analysis

Torgeir R. Hvidsten

Professor
Norwegian University of Life Sciences

Guest lecturer
Umeå Plant Science Centre
Computational Life Science Cluster (CLiC)

Thanks to: Professor Sandra Baldauf at the Program in Systematic Biology, Evolutionary Biology Centre, Uppsala for many of the slides used here!

# This lecture

- Background: terminology, definitions and history
- Distance methods
- Discrete methods
- Why trees may lie

# Some Definitions

From Greek: phylon = race / tribe / class;  genesis = birth / origin

Phylogeny = evolutionary classification

Phylogenetics

study of <u>predicted</u> evolutionary relationships

we can (almost) never know for sure what really happened

we can not replay the past

we can only extrapolate back from the present

predict the past based on what we see now

phylogenetic “reconstruction”:

because we are trying to recover the past

Phylogeny = molecular archaeology

the clues left in genes, proteins (aa and nt substitutions)

~> random remnants of the past, like shards of broken pottery

not the best clues, often deeply flawed, but sometimes enough

# The first evolutionary trees - 1860's

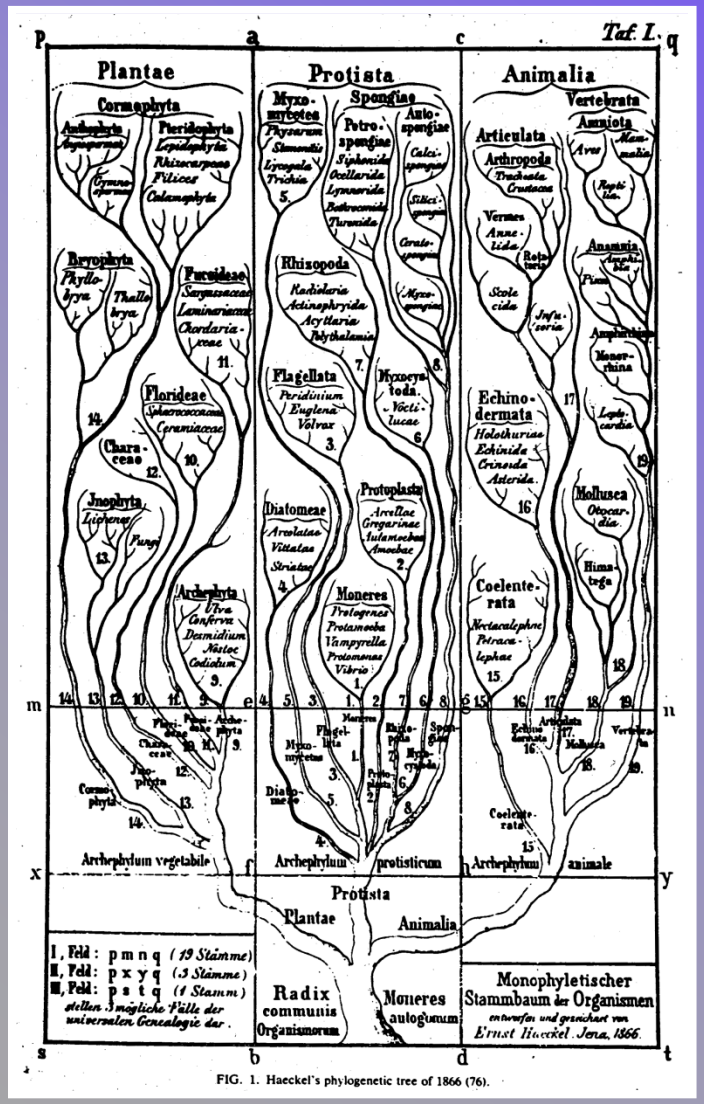
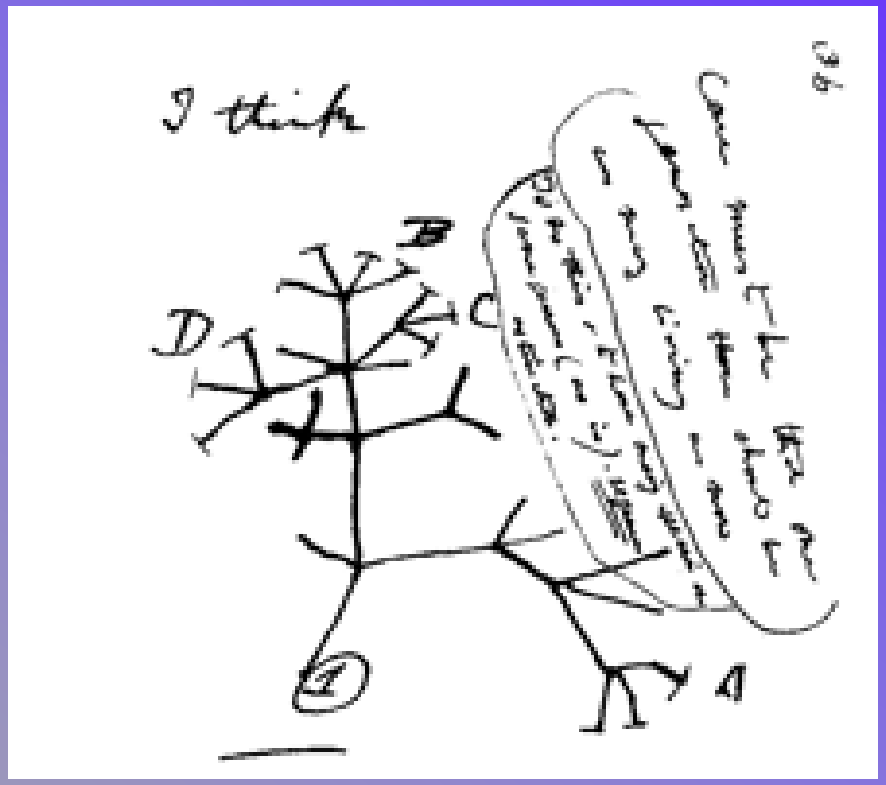


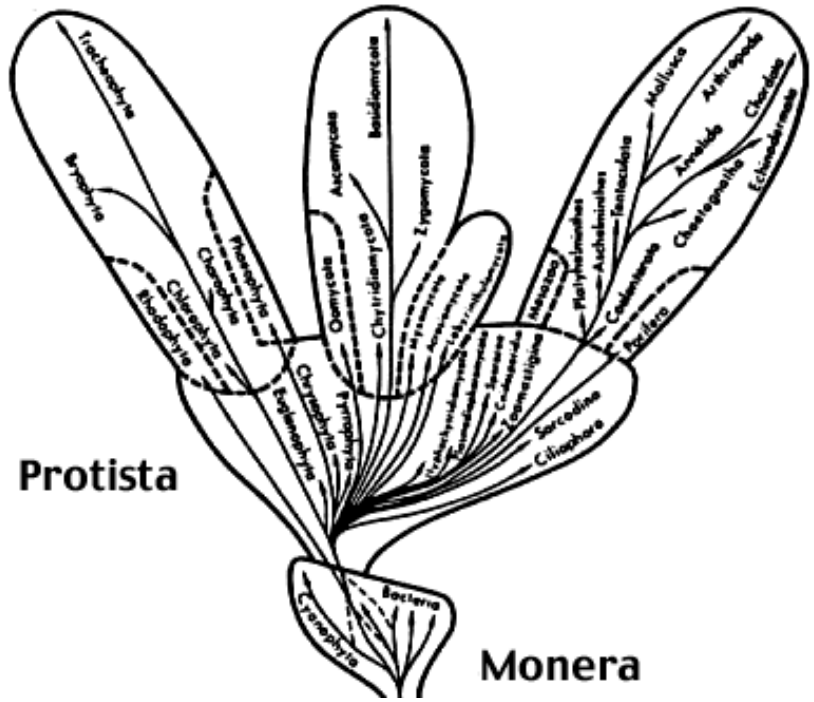

FIG. 1. Haeckel's phylogenetic tree of 1866 (76).

- Ernst Haeckel
  - first true trees of species




- Charles Darwin, Origin of Species
  - simple diagrams

# Phenetic Classification




Whittaker and Margulis, 1960s

"intuitive " phylogeny
based on overall similarity

"5 kingdom scheme"
Whittaker et al.
fungi elevated to 5th kingdom

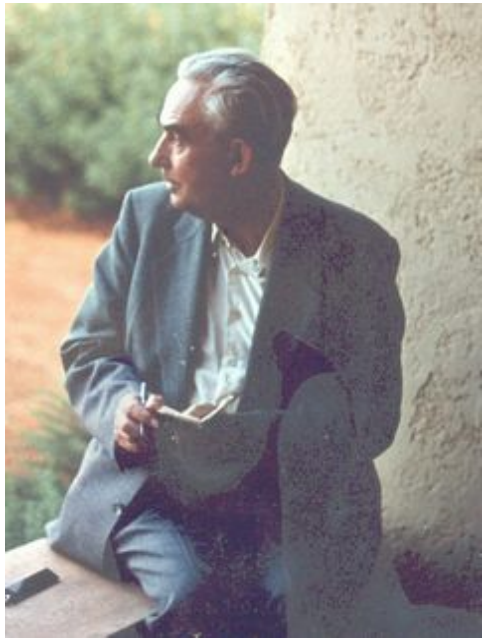
Lynn Margulis: popularized
(also endosymbiotic theory)
still in some textbooks

Improved microscopy (esp. electron microscopy)
=> new ultrastructural data: cytoskeleton, organelles, etc.

# 1960s: Cladistics

Hennig: formulated the rules of modern phylogenetic theory & practice
= cladistics: developed with morphogical characters
applies well with molecular data

Distinguished between ancestral similarities and derived similarities



Willi Hennig 1913-1976

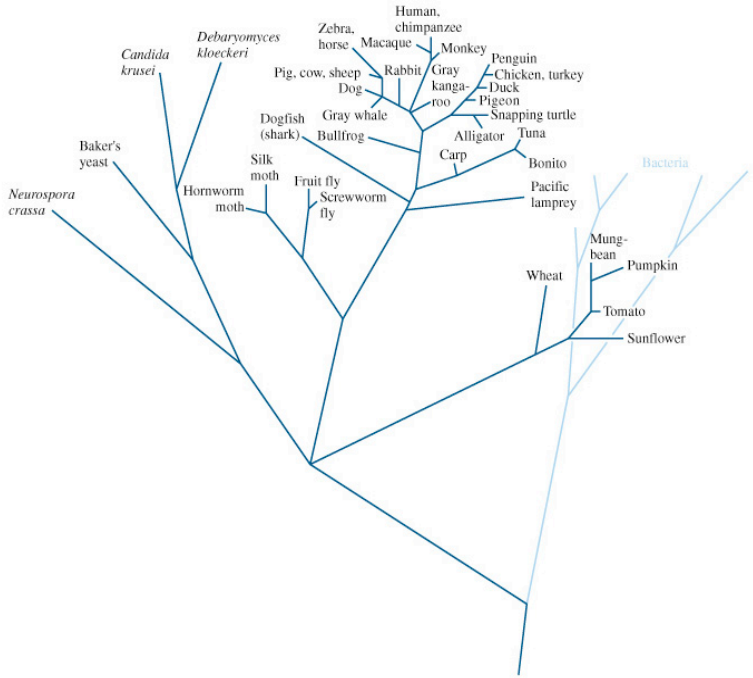
Ancestral characters (plesiomorphies – *"near"*)
Derived characters (apomorphies – "away")

Symplesiomorphides
shared primitive characters
= common heritage of all,
uninformative about unique relationships

Synapomorphies
shared derived characters
= unique heritage of subset of taxa,
define unique groups (clades)

# 1960s: Molecular Phylogeny

Margaret Dayhoff 1925-1983




Universal Tree based on ferredoxin sequences
Science (1966:) 152:363-366

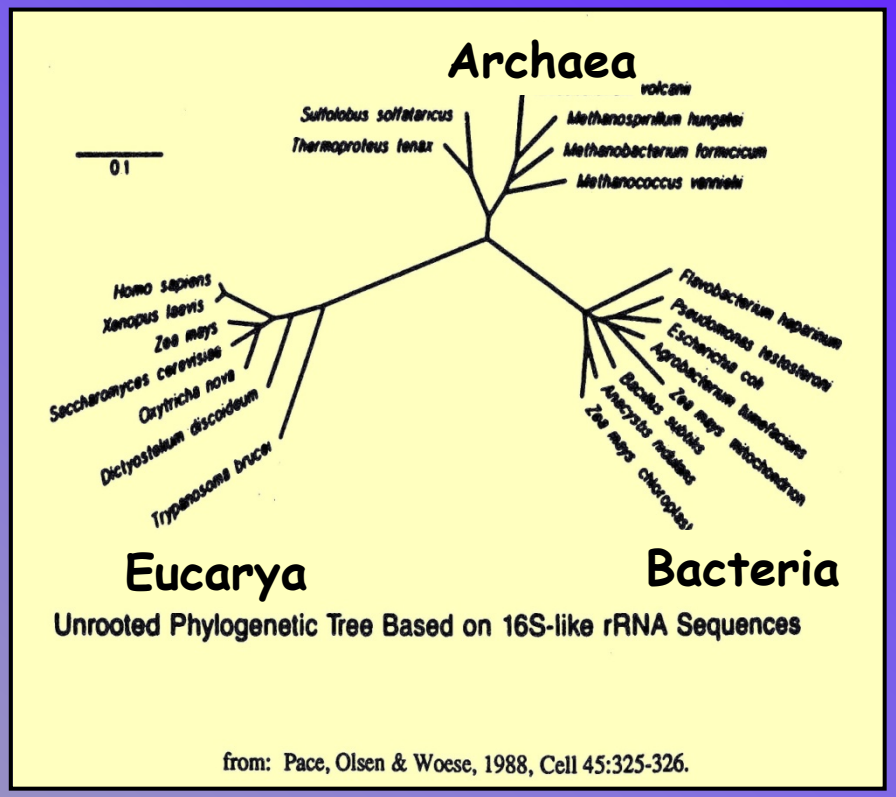
pioneered study of:
- protein evolution
- field of bioinformatics

first true universal evolutionary trees
used small proteins (~100 amino acids), sequenced “by hand”
no high-throughput automation, DNA sequence not invented yet

# 1980s: DNA Sequencing

Compared to protein sequencing
faster
easier
cheaper

More data from more and more
different organisms
-> bigger and better trees

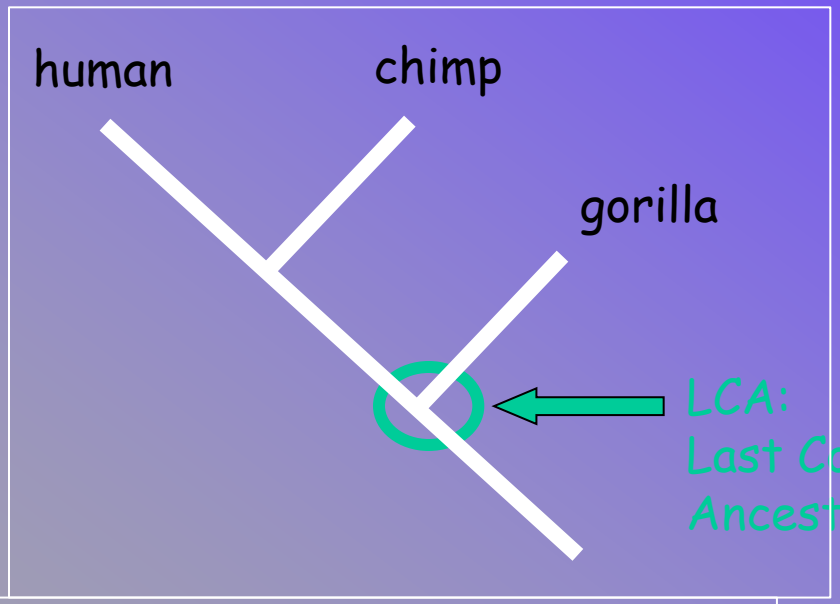



First tree of life including a wide variety of “bacteria”
indicated two fundamentally different kinds of bacteria
archaea = “third domain of life”

# Terminology

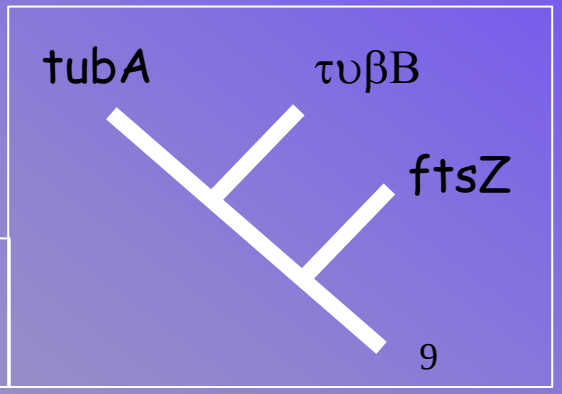
Node = a “divergence” or “splitting” events

Phylogenetic tree
= phylogram
= phylogeny
= evolutionary tree
= dendrogram
(“dendro” = tree)



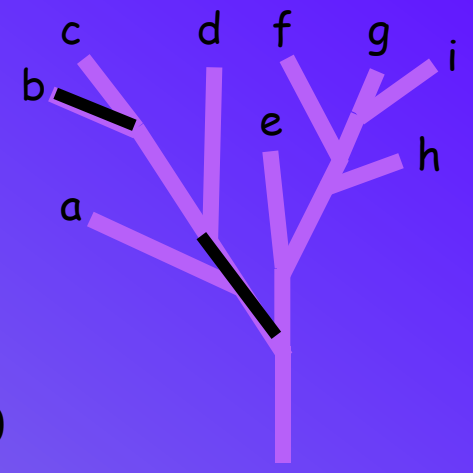

Tree
= branches, nodes

This is a species tree
divergences = speciation events



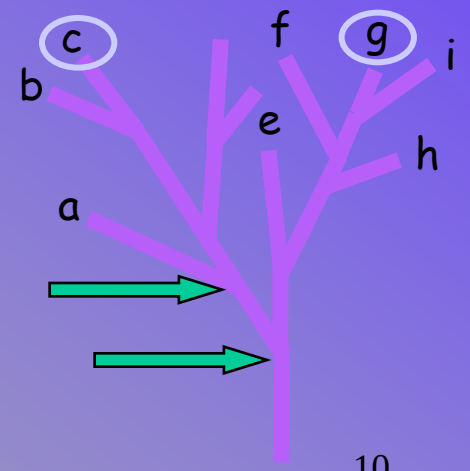

This is a gene tree
divergences = gene duplication events

# More Terminology




❖branches [“edges”] connect nodes
= internal (node to node) or terminal (node to terminal)

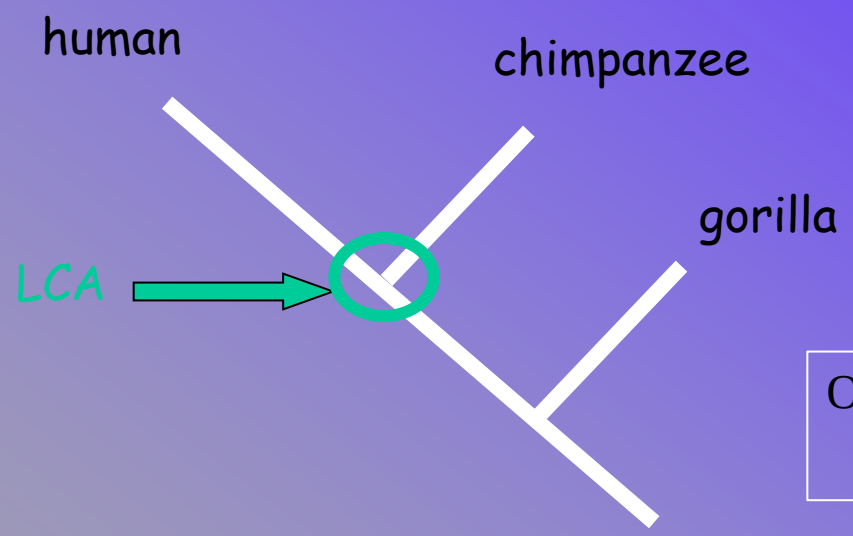
❖terminal nodes
- terminal nodes = [“leaves”] = operational taxonomic units, “OTUs”
- OTUs = organisms [“species tree”]
- OTUs = genes, proteins [“gene tree”]




❖internal node = point at which two branches diverge
- represent divergence events [“splittings”]

❖root = origin of the tree, or sub-tree
= point where everything started, corresponds to LCA

# LCA and Relatedness



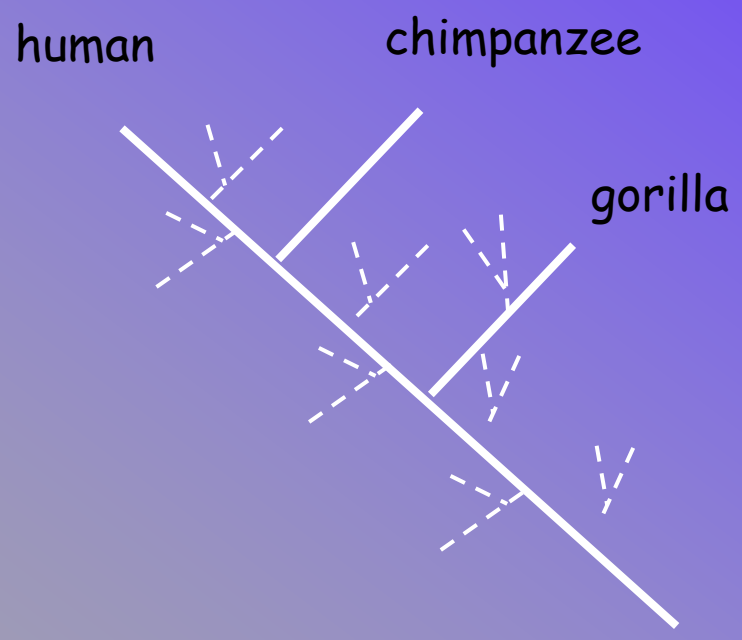

OTUs sharing more recent LCA,
are <u>more closely related</u>

Node = corresponds to last common ancestor (LCA) of diverging branches
= fossil, but mostly, hypothetical LCA

Human and chimp share more recent common ancestor with each other, i.e., they are more closely related to each other than either is to gorilla

LCA ≠ human or chimp (or even something in between),
LCA = something before, = equally ancestral to <u>both</u> lineages

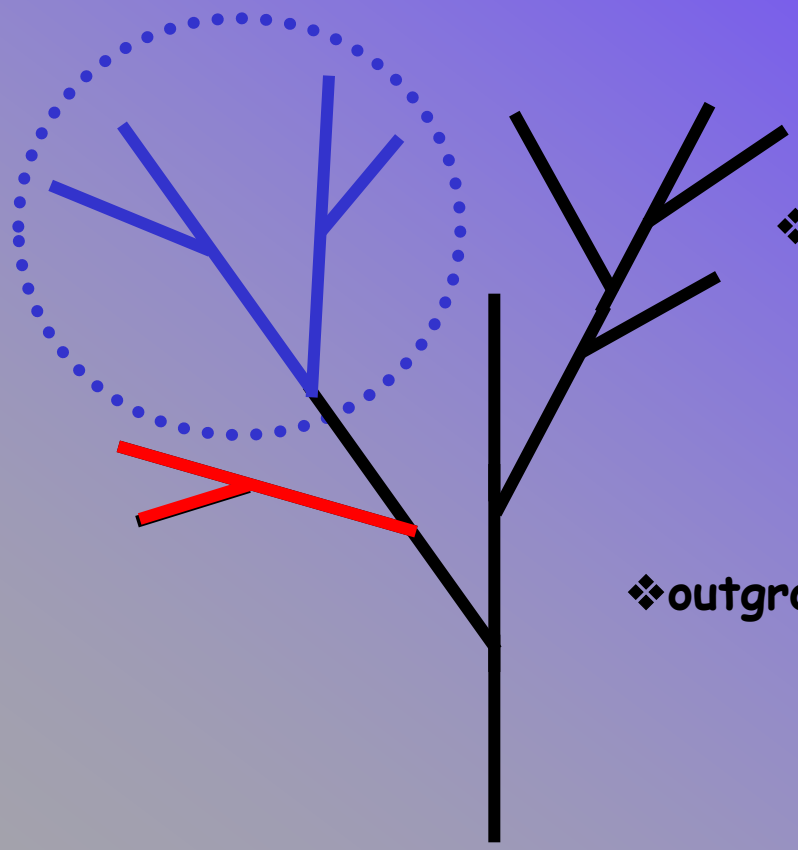
# A tree is usually only a fragment of the story




99.99999….% of all species that ever lived are extinct

true of genes as well?

# Clades



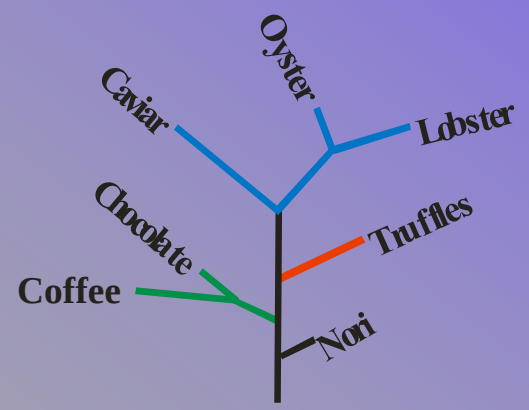
❖ clade (monophyletic group)
= complete group
= node plus all descendents
- share unique common ancestor,
and unique common history

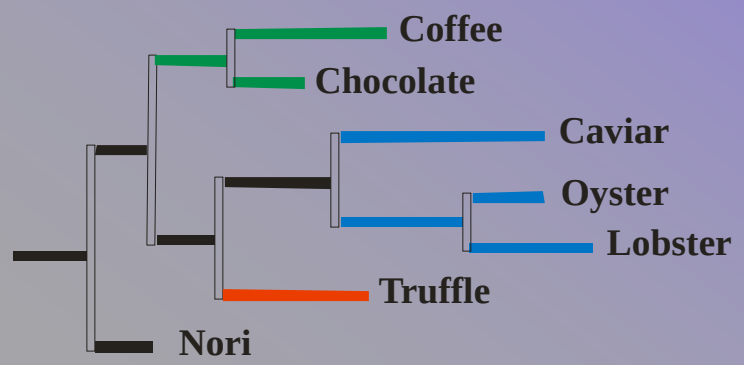
❖outgroup
- anything not in ingroup
(= group of interest)

❖ sister group
- closest outgroup to clade of interest
~> operational definition (true sister group probably extinct)
- operational sister group = closest outgroup available

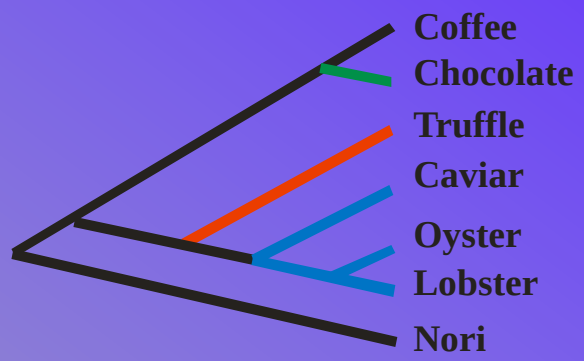
# Tree can be drawn with or without branch lengths (evolutionary distances)

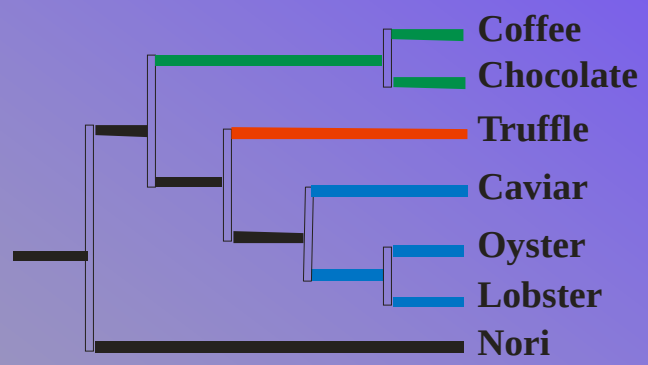
**Phylograms: relationships and distances**



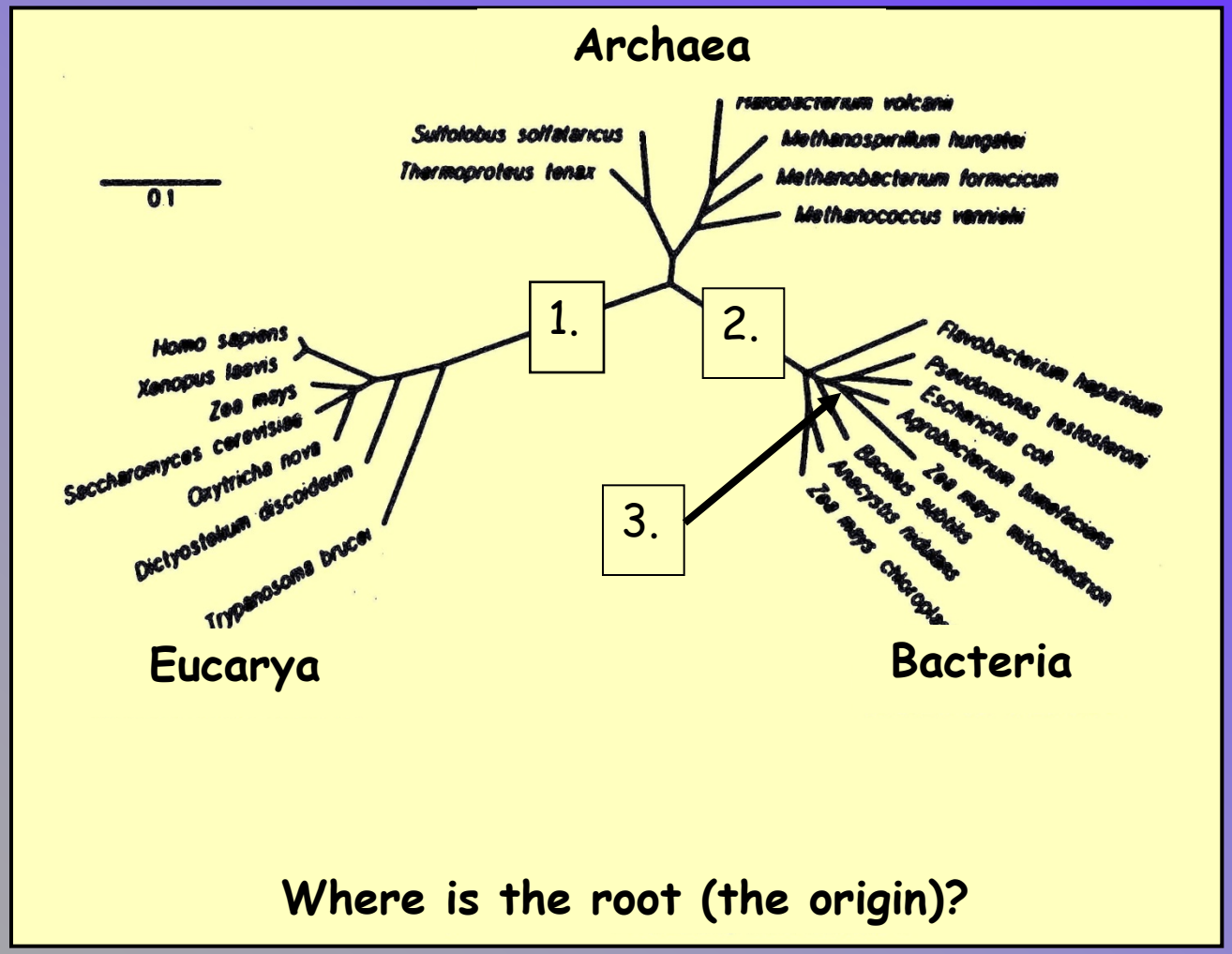



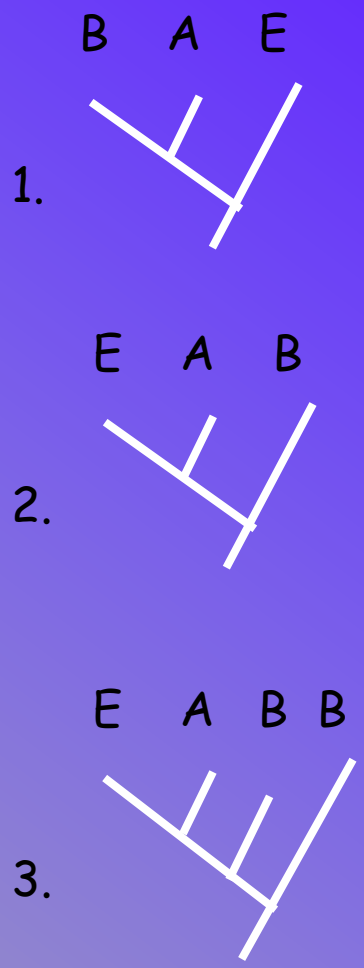

**Cladograms: relationships only**



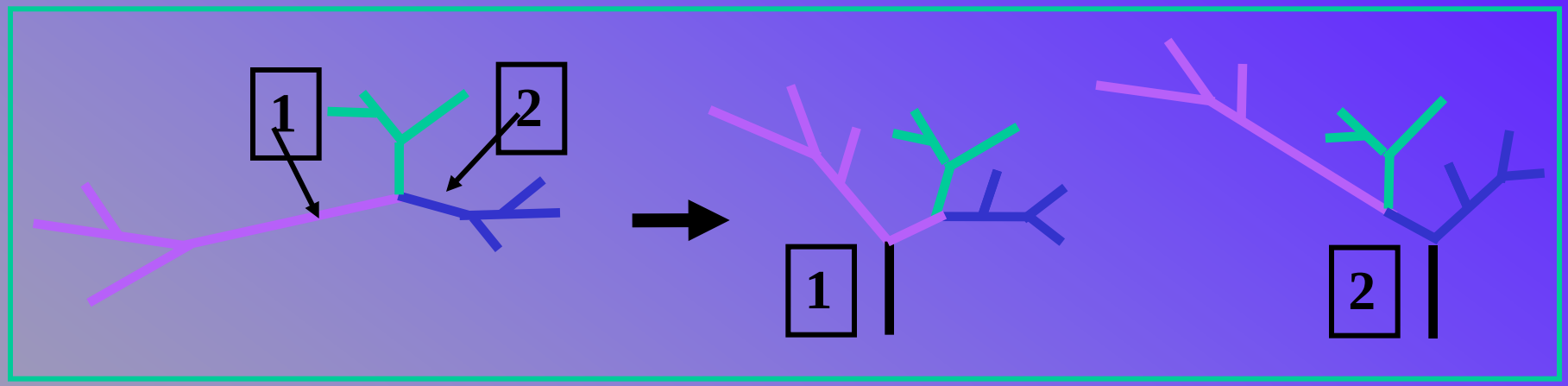



# Three Domains of Life



# Rooting Phylogenetic Trees:

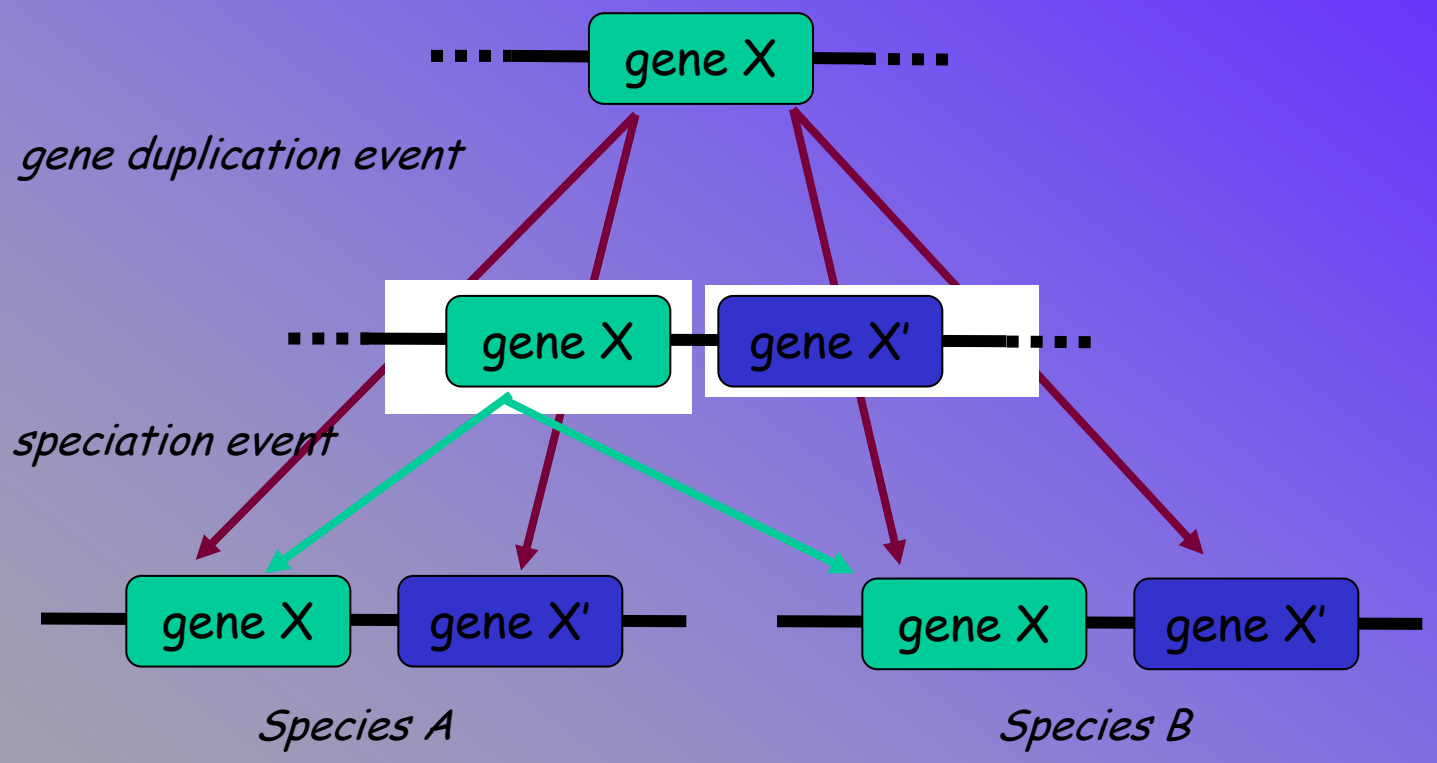



- root = oldest point in the tree
  if molecular clock (i.e. constant rate of evolution) -> root would be in the middle

- without a clock (i.e., in the real world) need external point of reference

  = outgroup, = anything not in your ingroup (= group of interest)
  for gene trees can use distant relative (paralogs)
  for species tree use sister group = closest relative to ingroup

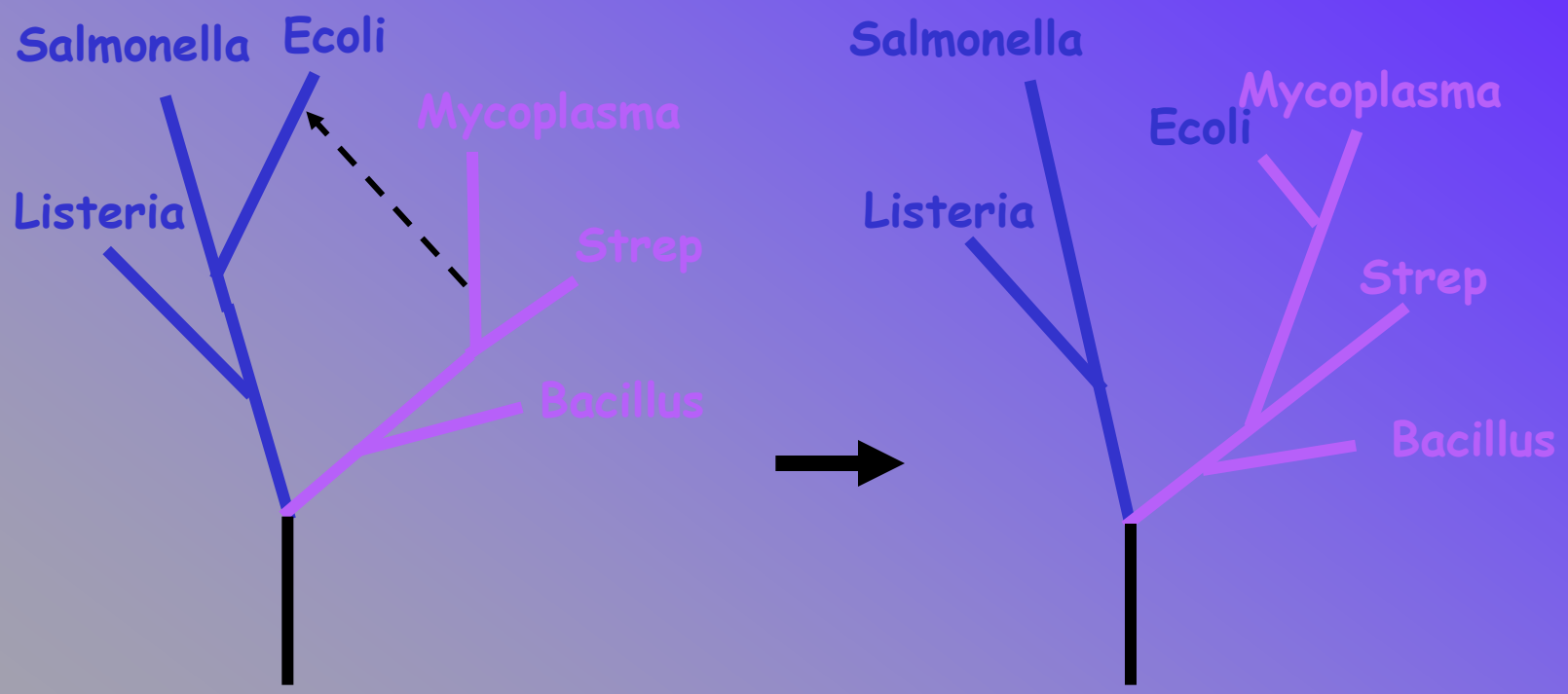
# Homologs / Orthologs / Paralogs




- all copies of gene X = orthologs

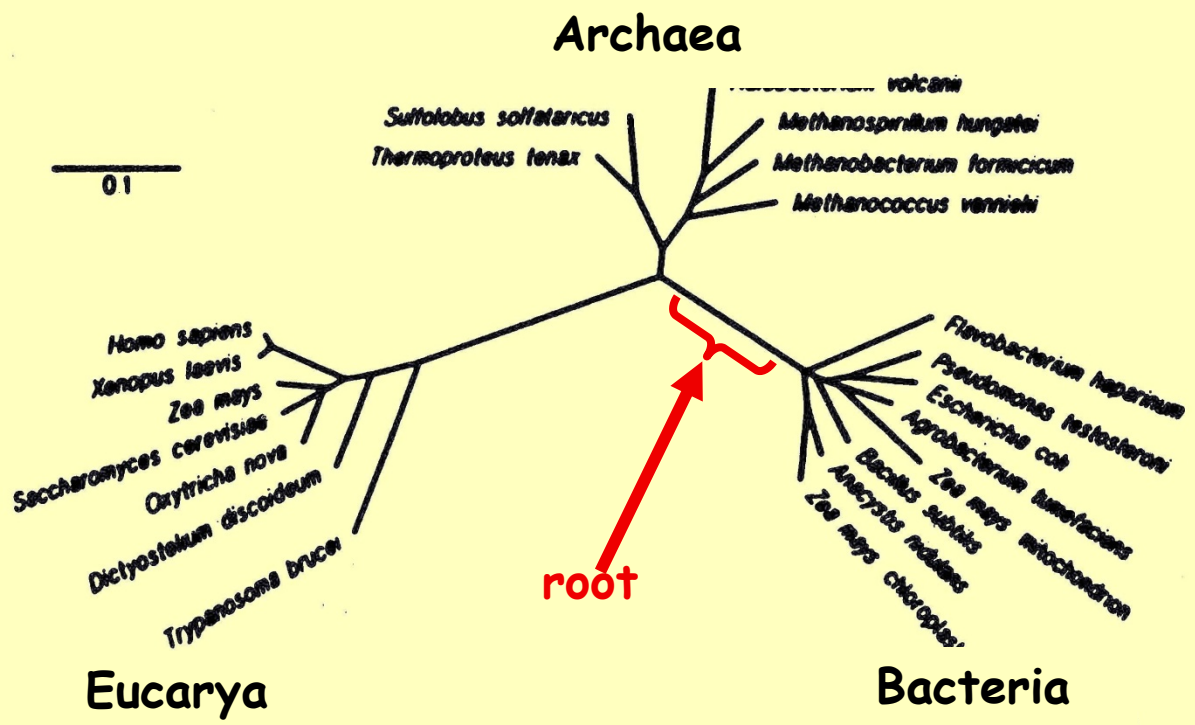
- genes X and X' are paralogs

# Laternal Gene Transfer -> Xenologs
# (xeno = foreign)




* very common in bacteria, especially for pathogenicity genes
important in bacterial evolution:
steal whole metabolic pathways from each other
important to us -> rapid spread of antibiotic resistance

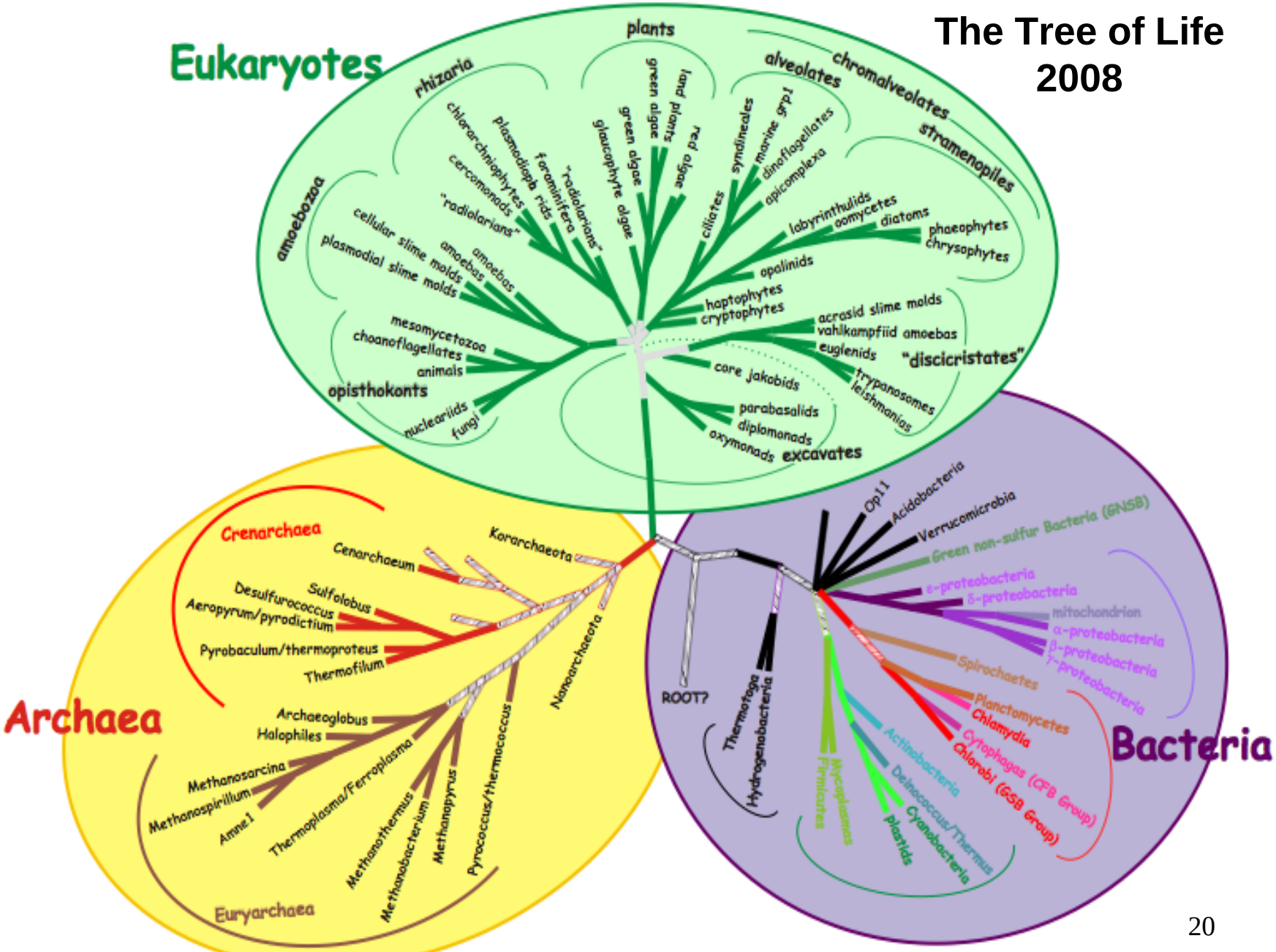
# First Molecular Trees (1988) -> Three Domains of Life




Unrooted Phylogenetic Tree Based on 16S-like rRNA Sequences

from: Pace, Olsen & Woese, 1988, Cell 45:325-326.

# The Tree of Life 2008






adapted from Baldauf, *et al.*, in *Assembling the Tree of Life*, 2004

# Two General Categories of Phylogenetic Methods

❖ Distance methods

- sometimes referred to as “clustering” or algorithmic methods

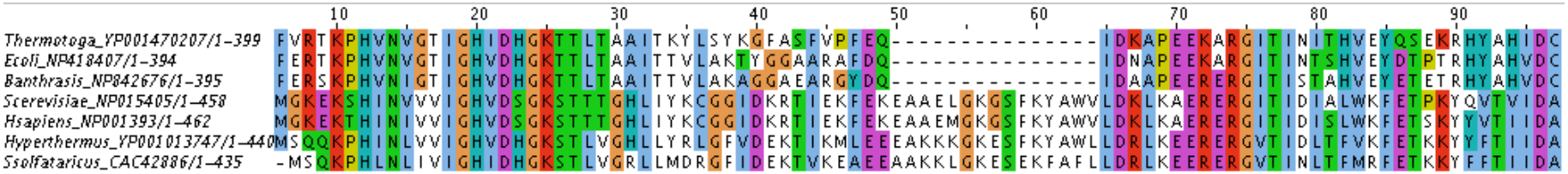
- calculate trees in two steps
    1. All data as single matrix of pairwise distances
    2. Distances assembled into tree,
        - most commonly using clustering algorithm
- fast, easy, reasonably accurate, good enough for many things
- methods: UPGMA (for clock-like evolution), <u>neighbor joining</u> (for reality)

❖ Discrete data (tree searching) methods

- each column in alignment = discrete data point
  =>hypothesis for each column of alignment
- look for the tree that best fits this collection of hypotheses
- much more details, better precision..., much slower
- methods: parsimony, maximum likelihood, bayesian inference

# Distance Methods

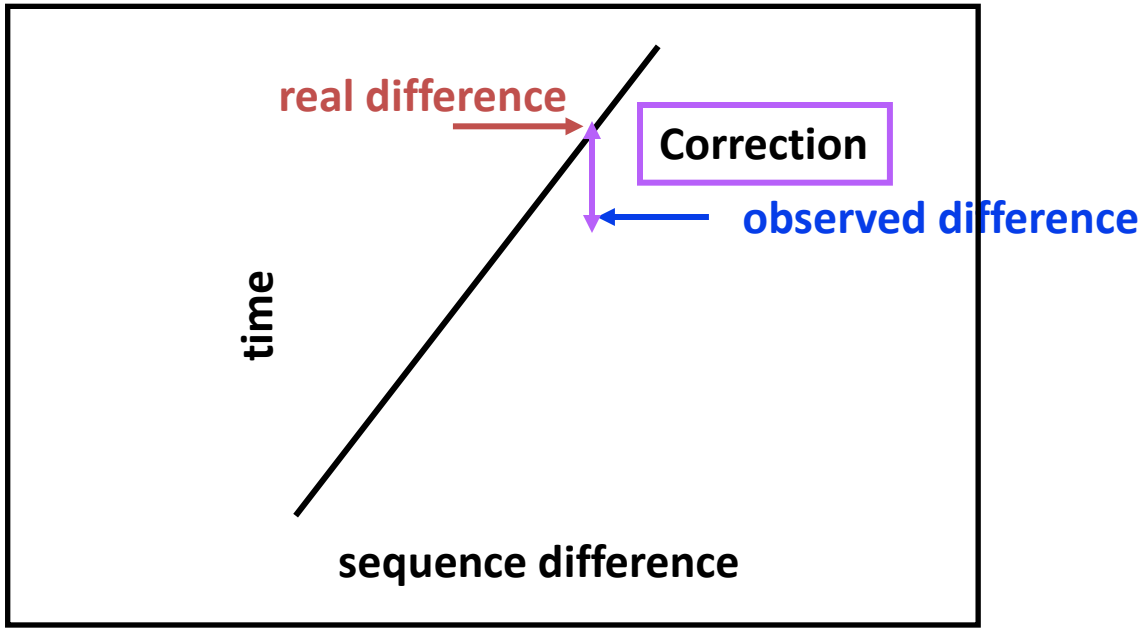
## Step 1: calculate a matrix of pairwise distances




| | Thermo | Ecoli | Bantha | Scere | Hsapi | Hyperth | Ssolf |
|---|---|---|---|---|---|---|---|
| Thermotoga | .000 | .245 | .325 | .731 | .727 | .786 | .786 |
| Ecoli | .245 | .000 | .333 | .739 | .733 | .778 | .780 |
| Banthrasis | .325 | .333 | .000 | .704 | .696 | .766 | .771 |
| Scerevisia | .731 | .739 | .704 | .000 | .143 | .400 | .431 |
| Hsapiens | .727 | .733 | .696 | .143 | .000 | .415 | .555 |
| Hypertherm | .786 | .778 | .766 | .400 | .415 | .000 | .222 |
| Ssolfatari | .786 | .780 | .771 | .431 | .555 | .222 | .000 |

# Distance Methods 1: Pairwise Distance Matrix

- All data reduced to single set of pairwise distances
  - therefore, important to accurately estimate distances
- Over short time, what you see is what you get
  Observed distance = true distance
- Over longer time "mutations on top of mutations" => hidden change
  simply counting differences under-estimate true distance




Over time, observed mutations ≠ true distance. Mutations still occur (distance still increasing) but no longer directly observable.

# Nucleotide Substitution Models

**Jukes-Cantor (JC)**
- **equal base sequencies**
- **all substitutions equal**

**K2P: Kimura 2-parameter**
- **equal base frequencies**
- **different rates for tsvstv**

**F81: Felsenstein 1981**
- **unequal base sequencies**
- **all substitutions equally likely**

**HKY85: Hasegawa et al., 1985**
- **unequal base sequencies**
- **different rates for ts vs tv**

**gtREV (GTR): General time reversible**
- **unequal base frequencies**
- **rate for each substitution type**

| | A | C | G | T |
|---|---|---|---|---|
| A | $q_{aa}$ | $\mu_{ac}\pi_c$ | $\mu_{ag}\pi_g$ | $\mu_{at}\pi_t$ |
| C | $\mu_{ca}\pi_a$ | $q_{cc}$ | $\mu_{cg}\pi_g$ | $\mu_{ct}\pi_t$ |
| G | $\mu_{ga}\pi_a$ | $\mu_{gc}\pi_c$ | $q_{gg}$ | $\mu_{gt}\pi_t$ |
| T | $\mu_{ta}\pi_a$ | $\mu_{tc}\pi_c$ | $\mu_{tg}\pi_t$ | $q_{tt}$ |

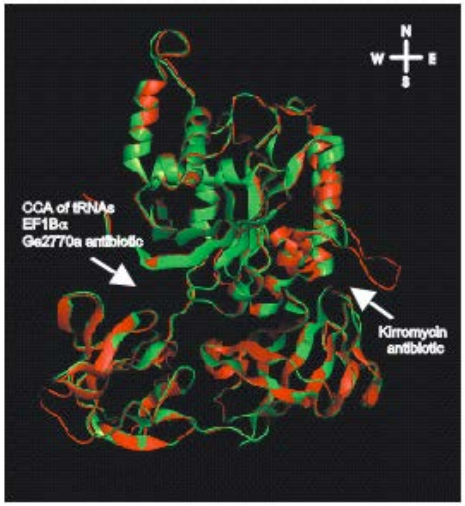
$2\beta$

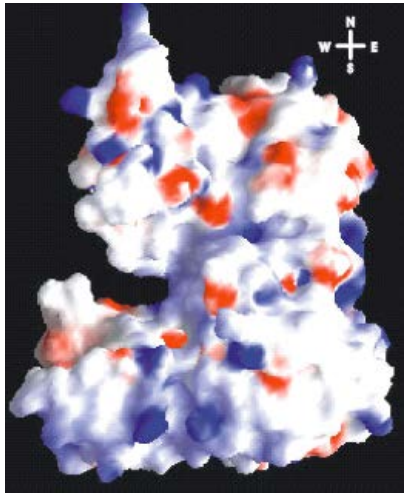
# Not all sites evolve according to the same rules

```
                                   10        20        30        40        50        60        70        80        90
Thermotoga_YP001470207/1-399   FVRTKPHVNVGTIGHIDHGKTTLTAAITKYLSYKGFASFVPFEQ---------------IDKAPEEKARGITINITHVEYQSEKRHYAHIDC
Ecoli_NP418407/1-394           FERTKPHVNVGTIGHVDHGKTTLTAAITTVLAKTYGGAARAFDQ---------------IDNAPEEKARGITINTSHVEYDTPTRHYAHVDC
Banthrasis_NP842676/1-395      FERSKPHVNIGTIGHVDHGKTTLTAAITTVLAKAGGAEARGYDQ---------------IDAAPEERERGITISTAHVEYETETRHYAHVDC
Scerevisiae_NP015405/1-458     MGKEKSHINVVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAAELGKGSFKYAWVLDKLKAERERGITIDIALWKFETPKYQVTVIDA
Hsapiens_NP001393/1-462        MGKEKTHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDISLWKFETSKYYVTIIDA
Hyperthermus_YP001013747/1-440MSQQKPHINLVVIGHVDHGKSTLVGHLLYRLGFVDEKTIKMLEEEAKKKGKESFKYAWLLDRLKEERERGVTIDLTFVKFETKKYYFTIIDA
Ssolfataricus_CAC42886/1-435 -MSQKPHLNLIVIGHIDHGKSTLVGRLLMDRGFIDEKTVKEAEEAAKKLGKESEKFAFLLDRLKEERERGVTINLTFMRFETKKYFFTIIDA
```

Different positions in a sequence can evolve at very different rate

Some sites change a lot
Others unchanged



# Distance Matrix Methods: Step 2 - Tree Building

- 1. UPGMA (unweighted pairgroup method)
  - group most similar sequences first
  - only works if there is a molecular clock, which there isn't
  - simple,fast, ~> highly inaccurate
  - no one uses this anymore!

- 2. neighborjoining method (NJ)
  - group sequences stepwise to minimize tree length
  - much more accurate, nearly as fast, now
  - progressively pair sequences

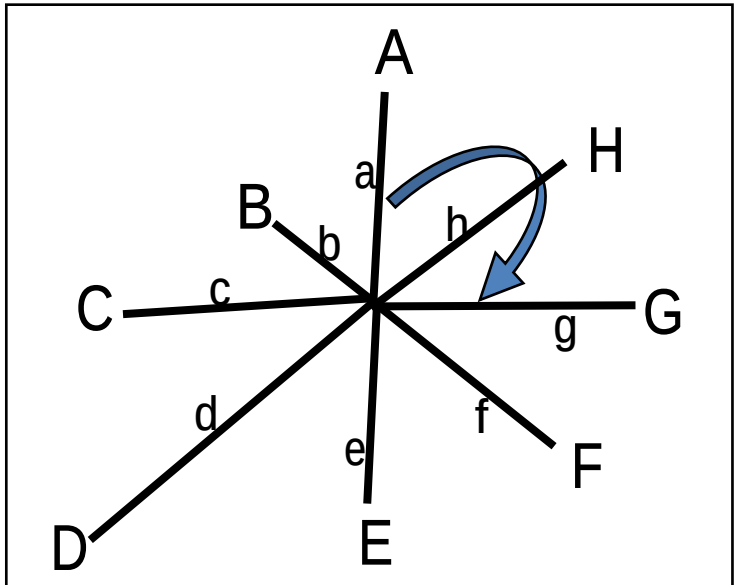
- Both take distance matrix and turn it into a tree
  - independent of method used to derive the matrix

# Neighborjoining Distance Method (NJ)

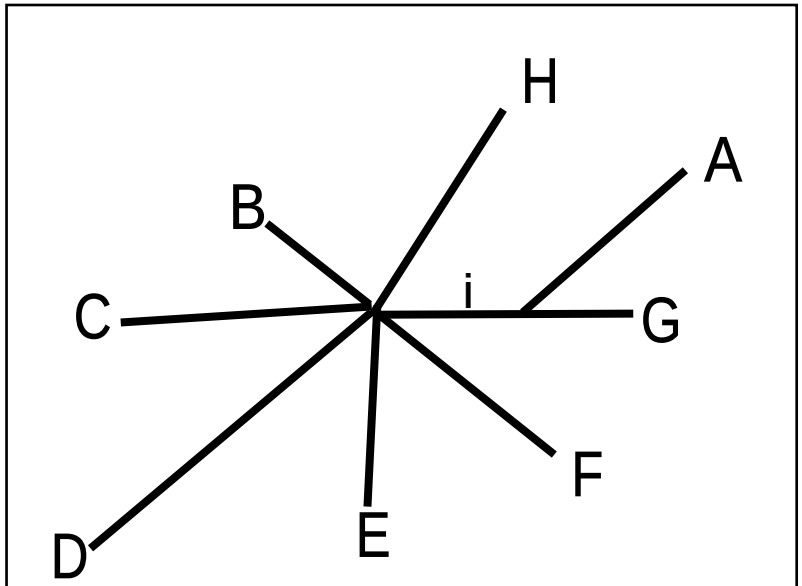
**group sequences stepwise to minimize tree length (L = sum of branches)**

**start with star phylogeny (fully unresolved tree = longest possible)**

- **progressively pair sequences**
- **select pairing that shortens the tree the most (L' = L-i)**
- **recalculate the distances, repeat→ fully resolved tree**




**L (length) = a+b+c+d+e+f+g+h**




**L (length) = a+b+c+d+e+f+g+h-i**

# Evaluating Trees: Bootstrap Analysis

- a method for calculating the reliability of different parts of the tree
- "random sampling with replacement"

1. create multiple pseudo-datasets from the real dataset by repeatedly drawing sites from the real-dataset (with replacement)
   - pseudo-dataset have the same size as the real dataset
   - but some sites are present multiple times, others absent
   - repeat x times (1000 minimum)
2. calculate phylogenetic tree for each pseudo-dataset
3. reliability score: how many pseudo-trees contain clade (node) x

- advantages

  it works: tested in lab with populations of viruses:

  - simulate evolution, sequence -> tree, bootstrap (*Hillis & Bull, 1993)*

  can use with any phylogenetic method

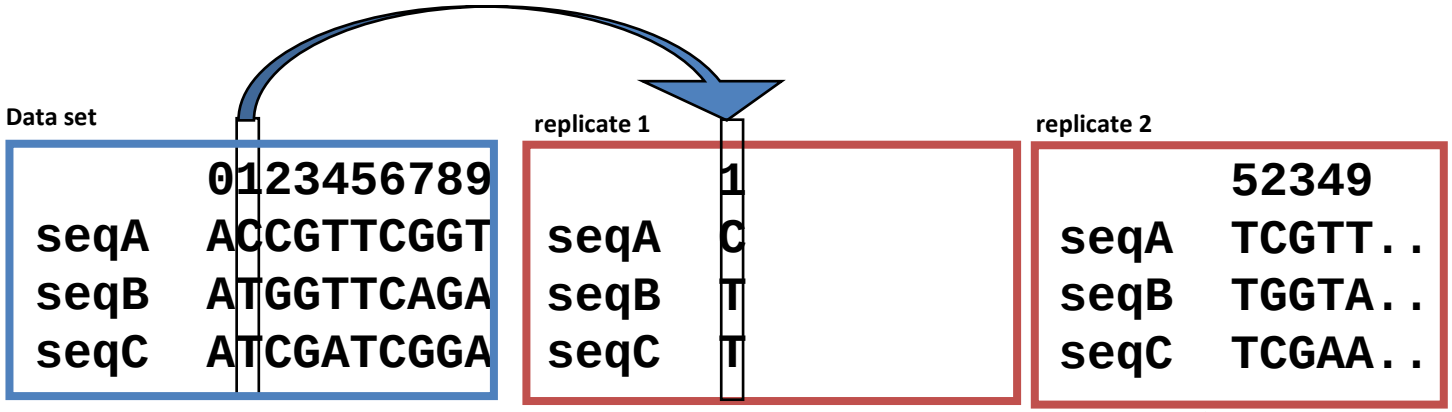
  - well understood

# Bootstrap Analysis

**Step 1**

**a. build pseudodata sets**

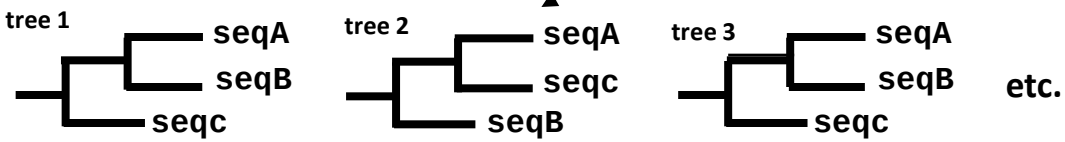
**b. repeat x 1000**




**Step 2**

**build trees for each**

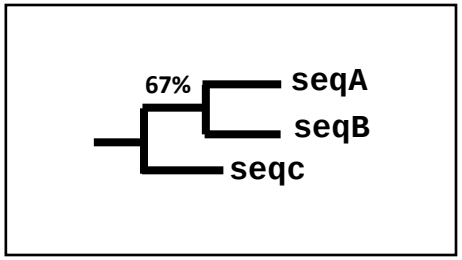
**(= 1000 trees)**




**Step 3**

**tabulate results**

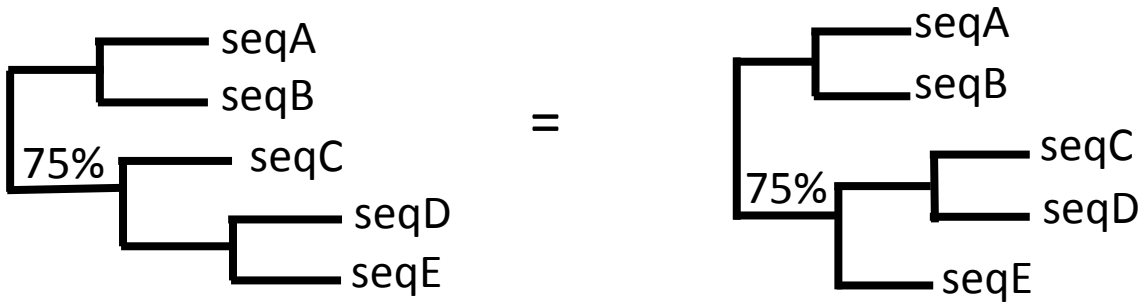
**(strict consensus tree)**



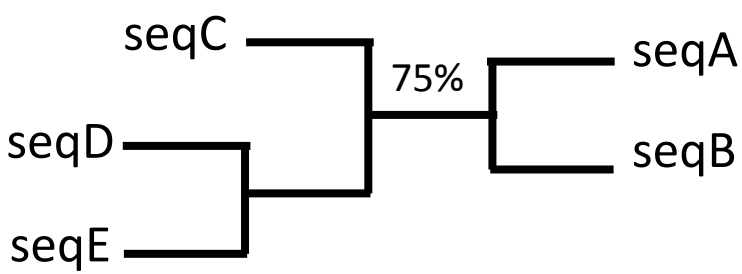

**bootstrap consensus tree**

# Bootstrap: rule 1

❖1. Bootstrap (BP) values=support for a clade (a single branch in the tree)
no statement about relationships within that clade




❖Each bootstrap divides tree in half
bootstrap value = equal support for each half



# Bootstrap: rules 2-3

- 2. theoretically, only BP > 95% = significant

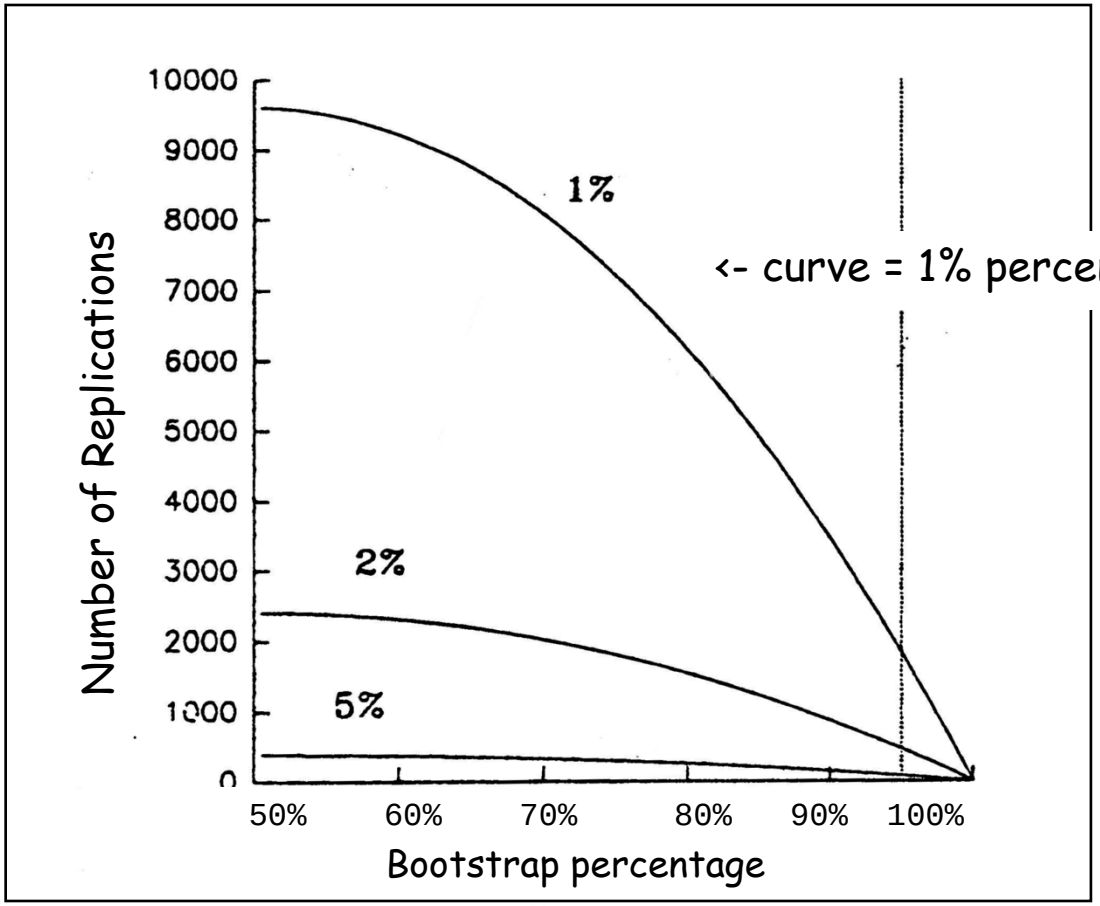
  experimental evidence: ~> BP>70% = robust
  at least for molecular data

  Hillis& Bull, 1993, Systematic Biology, 42:182

- 3. what if BP > 90% for clade of interest, but <50% for others

  - count yourself lucky!
  - trees don't have to be fully resolved to be useful
  - don't expect 100% BP for every branch on your tree

# Bootstrap rule 4: More Is Better



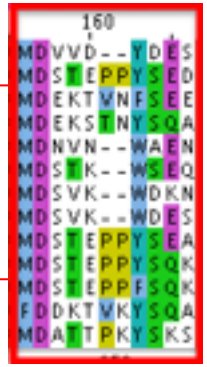

<- curve = 1% percent variance in p-value

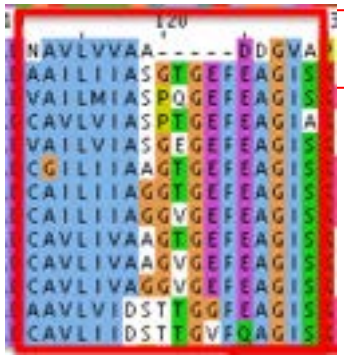
ref: Hedges 1998

- 100 bootstrap replicates
  60% bootstrap = +/- <5%
  100% bootstrap = +/- 1%
- 1000 bootstraps
  60% = +/- ~4%
  100% = +/- 1%

# A Tree is Only As Good as the Alignment Its Based On

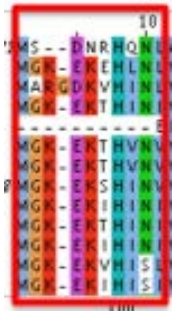
Delete regions of uncertain alignment (= uncertain homology)
there are other ways to align this region
hard to know which is correct





Low sequence similarity ≠ uncertain homology

Also:
delete regions with incomplete sequence for >1 OTU
(otherwise more data for some OTUs than others)



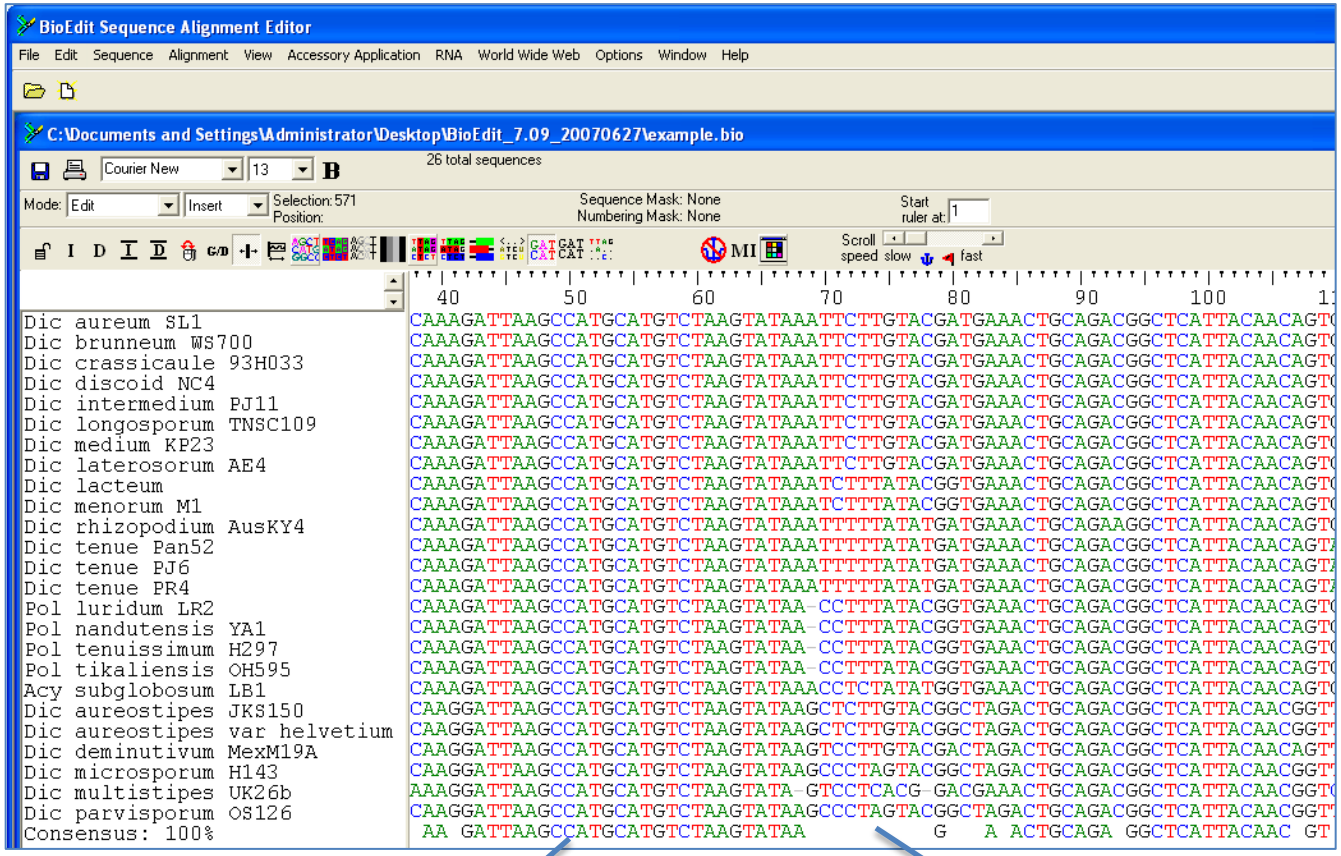
OK

not OK

Delete large indels.

# Defining regions of certain homology: consensus sequences

100% consensus means all sequences have same character at this position




100% identical for all sequences

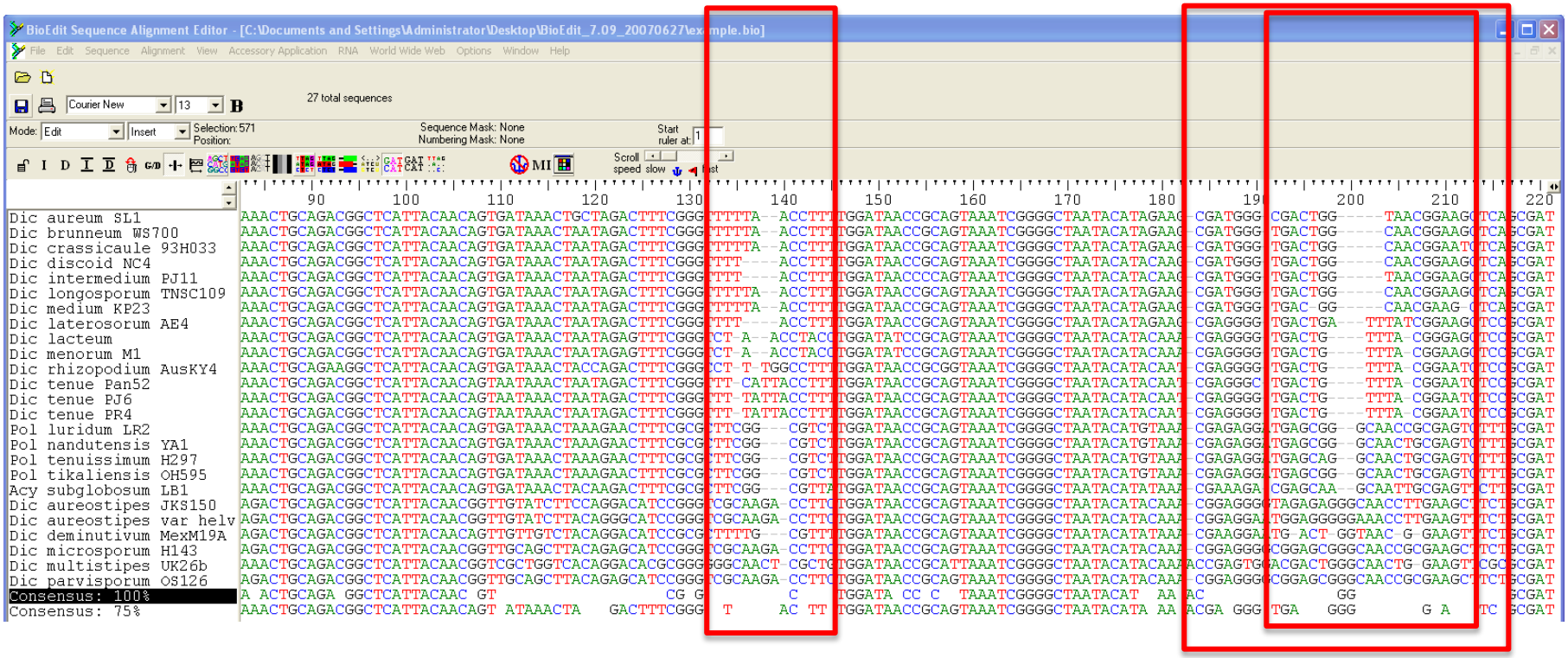
not 100% identical

# Defining regions of certain homology: consensus sequences

100% consensus too “stringent”,

more common to use ~75% (but depends on the data set)

more distantly related sequences/organisms, may require lower stringency




Regions to delete: gaps, and surrounding regions of “uncertain” alignment/homology

# Discrete Data Methods

- - start with tree
- - fit the data to the tree
- - measure goodness of fit

- parsimony, maximum likelihood, bayesian inference
  - each measures goodness of fit in slightly different ways

- parsimony
  - measures steps (mutations)
  - best tree = least number of steps (shortest = simplest)
  - Occum's razor, simplest solution most likely correct

- likelihood
  - measure likelihood of data given the tree
  - best tree = one with maximum (=highest) likelihood
  - readily accommodates complex models (substitution weighting)
    - same models as distance (JC, K2P, HKY, etc.)
    - (unlike parsimony)

- bayesian inference
  - best tree = most probably tree given the data (posterior probability)
  - modifies the model as the search proceeds
  - algorithm learns and improves itself

# Bayes' Theorem

Posterior Likelihood Prior

- $$P(T|D) = \frac{P(D|T) * P(T)}{P(D)}$$

Normalization constant

## Discrete Data Methods

- - start with tree
- - fit the data to the tree
- - measure goodness of fit

- calculations (measure of tree quality) ~straightforward
  - challenge is finding the right tree(s)

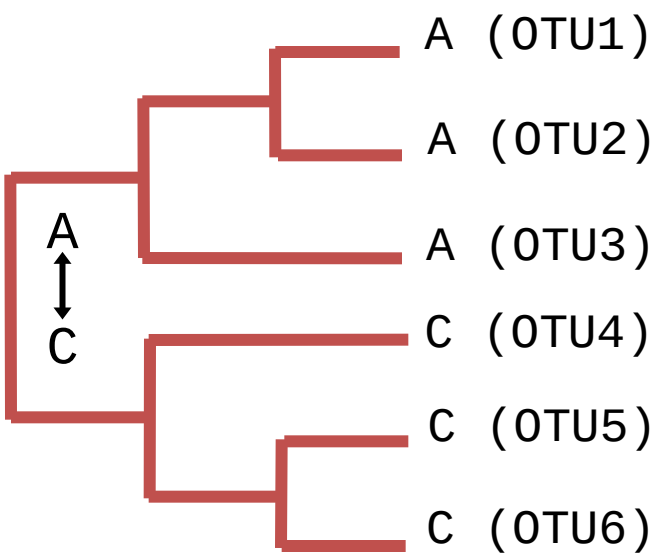
- in a ideal world, examine all possible trees
  (universe of all possible trees for set of OTUs
  = tree space)
  - take each tree, fit data to tree, best fit tree wins

- problem: number of possible trees for n OUT = $n^{n-2}$
  - # possible trees increases rapidly with # OTUs
  - ~20 OTUs: # possible trees > # stars in universe
  - exhaustive search impossible > 14 OTUs

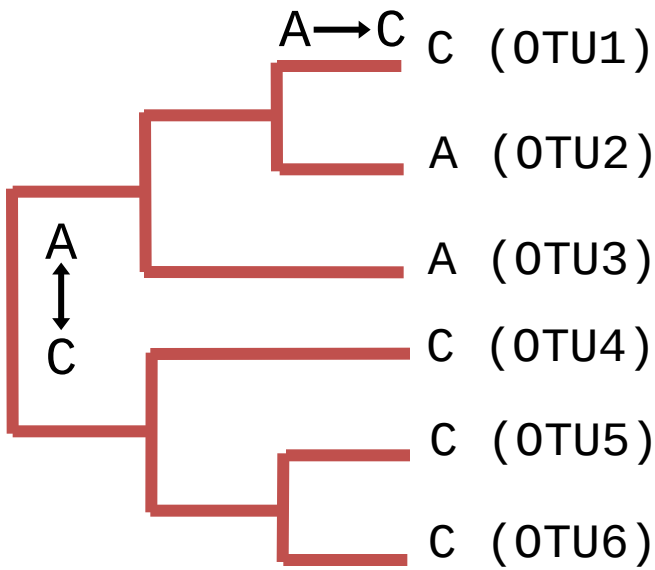
# Measuring Goodness of Fit: Parismony

- parsimony measures tree fitness in "steps" (mutation events)
  - sum for each position (column) in alignment separately

Tree 1: alignment position 1



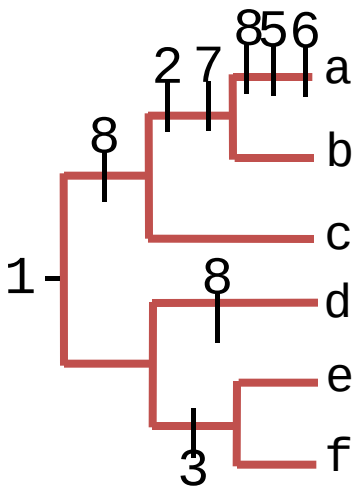

Tree 1: alignment position 2



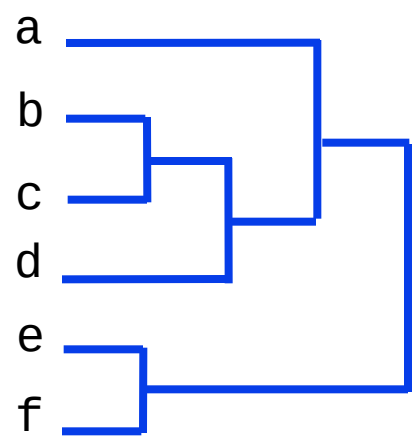

- total number of steps = length of tree for given alignment
  repeat for all trees
  tree requiring fewest number of changes = best tree
  Occum's razor - the simplest solution is most likely correct

# A Parsimony Problem

Tests two alternative Trees
identify one requiring the Least Number of Changes
(= simplest hypothesis)




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| seq-a | C | C | C | C | A | A | A | C |
| seq-b | C | C | C | C | C | C | A | A |
| seq-c | C | A | C | C | C | C | C | A |
| seq-d | A | A | C | C | C | C | C | A |
| seq-e | A | A | A | C | C | C | C | C |
| seq-f | A | A | A | C | C | C | C | C |

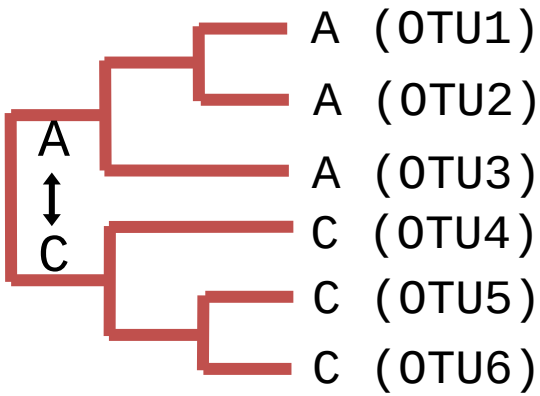



Tree A
9 steps

Tree B
11 steps

# Maximum Likelihood

- essentially = parsimony, but with weighting
- weights =same as distance models (JC, K2P, etc.)



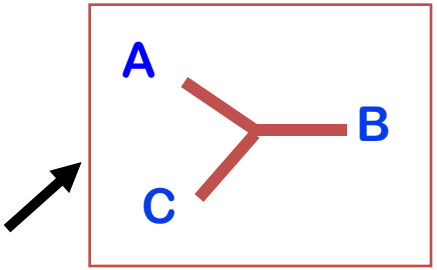

parsimony = 1 step

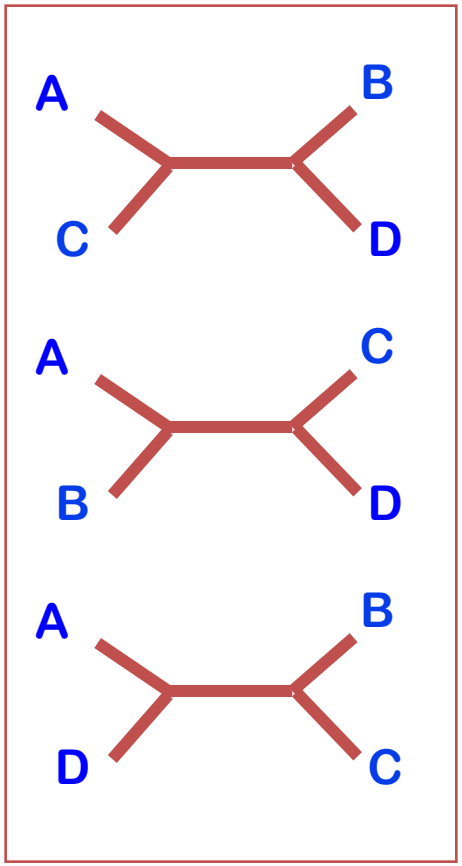
likelihood = 1 x weight

- Likelihood with all changes weighted equally => parsimony
- Likelihood = slower, but more accurate
  - more likely to find true tree in messy data

# All Discrete Data Methods Start with the Tree

- Ideally - generate all possible trees
  - measure fit of the data to the tree
    - best fit = correct tree
      (most likely to be)



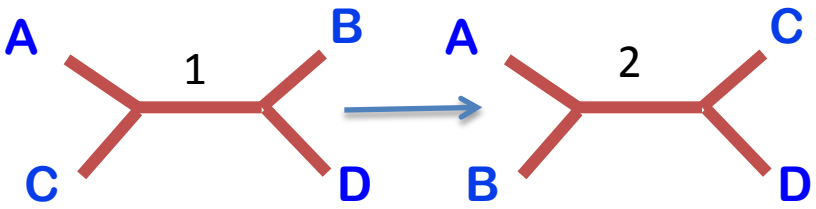



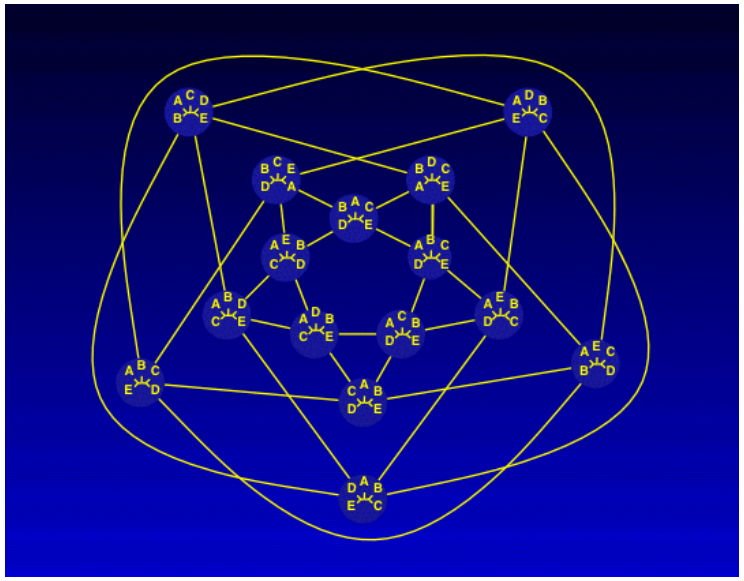

- 3 OTUs -> 1 possible tree
- 4 OTUs -> 3 possible tree
- 5 OTUs -> 15 possible tree
- x OTUs -> $x^{x-2}$ possible tree
- 15 OTUs > # stars in the universe

- >14 OTUs, exact solution not possible
  - need short cuts - heuristics, intelligent search
  - need an intelligent way to search tree space

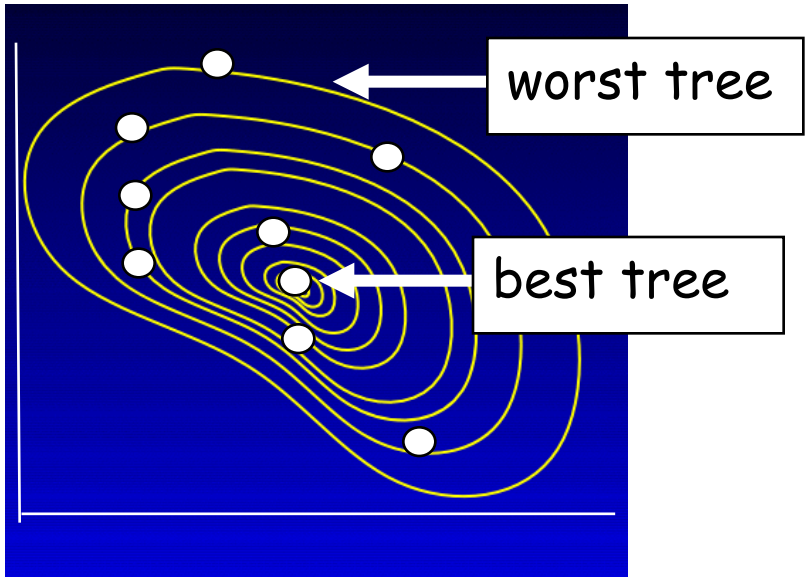
# Tree Space = Universe of All Possible Trees for a set of OTUs

- All trees within tree space are related to each other



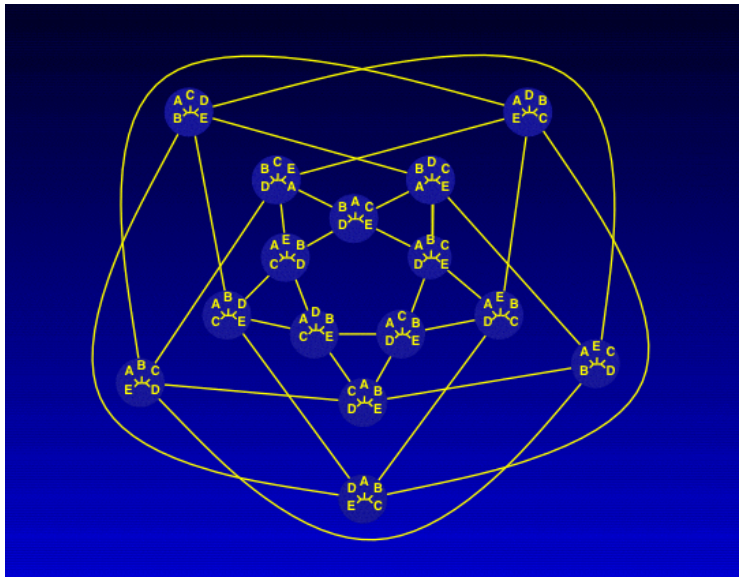
Tree space for 5 OTUs
All trees connected by single rearrangement of branches




Trees as a landscape

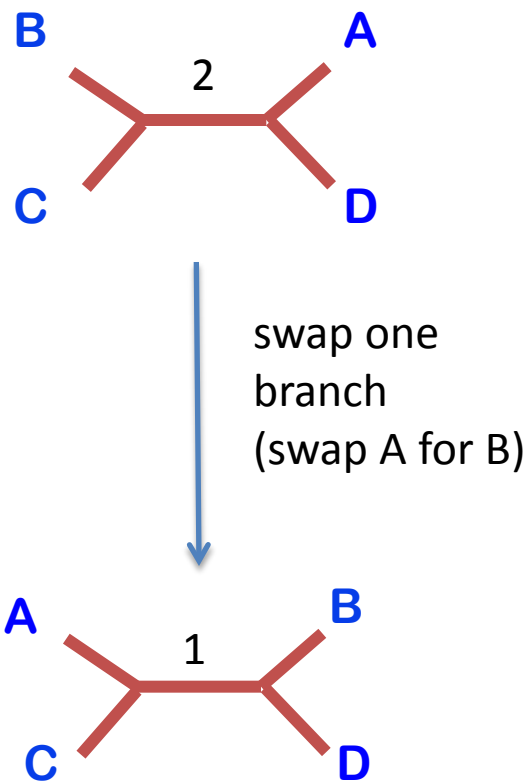
# Searching Tree Space

- All trees within tree space are related (connected)

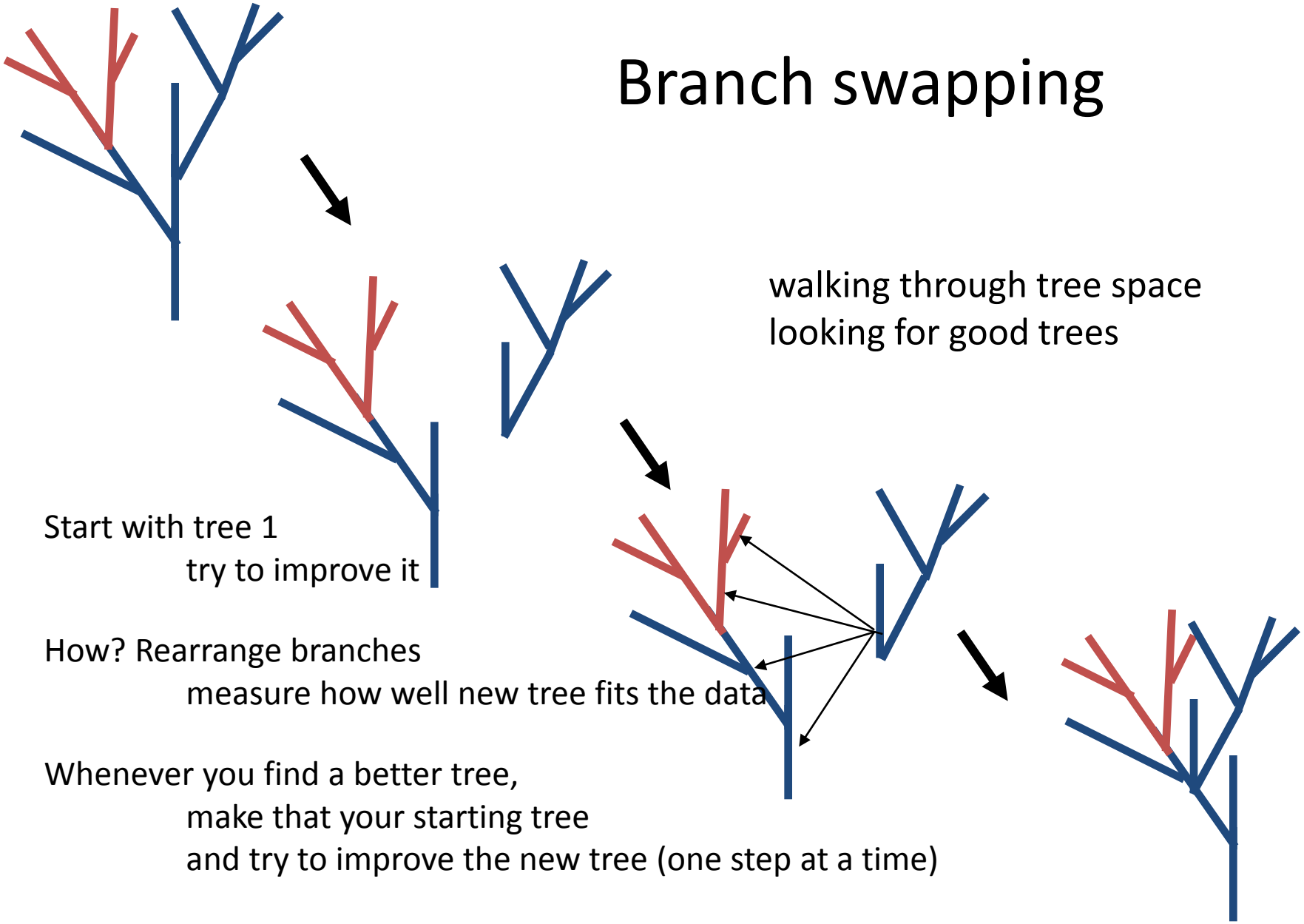


Tree space for 5 OTUs

all trees related by single rearrangement of branches



# Branch swapping



walking through tree space
looking for good trees

Start with tree 1
try to improve it

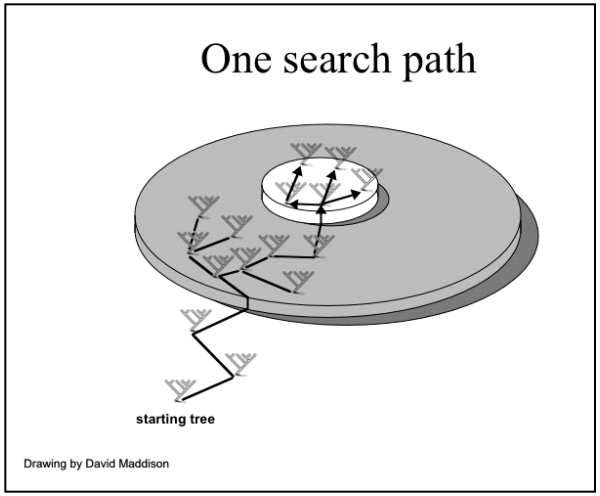
How? Rearrange branches
measure how well new tree fits the data

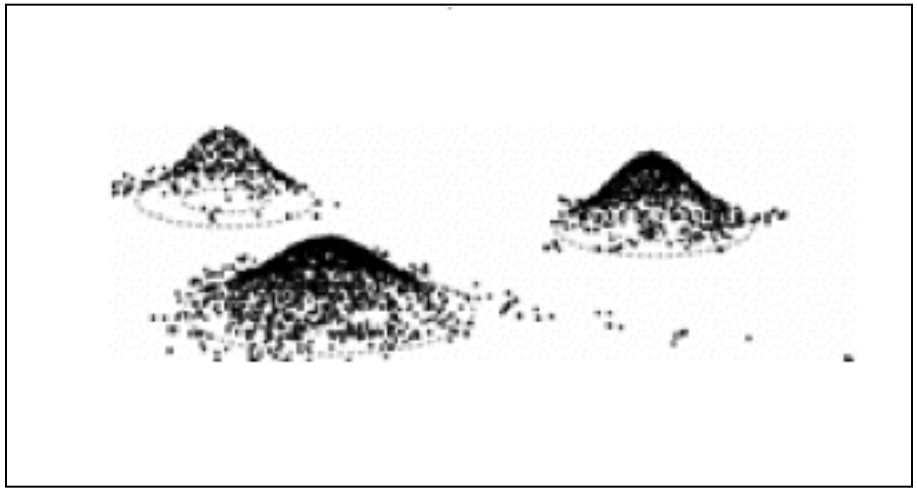
Whenever you find a better tree,
make that your starting tree
and try to improve the new tree (one step at a time)

Continue until you can't find anything better

# Complex Tree Space

- Branch swapping would be easy, if tree space were simple
  - but, tree space can often be very complex
  - multiple sets of pretty good trees (tree islands)
  - correct tree is on one of these, but which one?

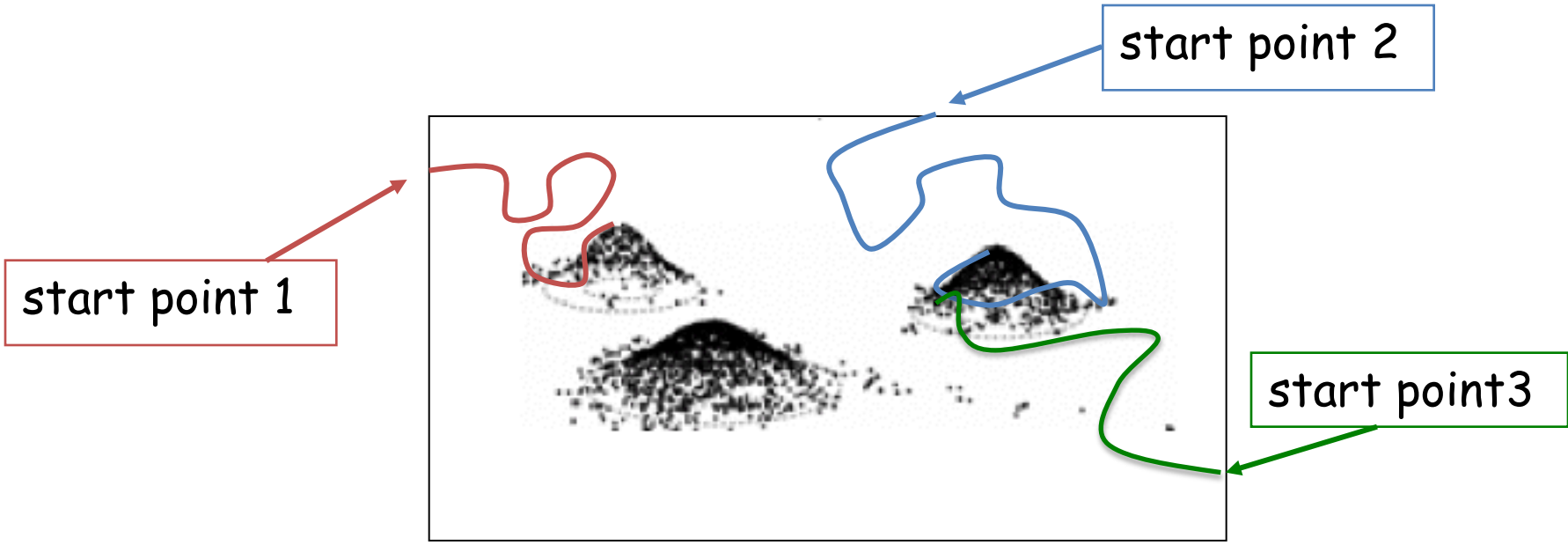


- branch swapping algorithms are "greedy"
  - once on path upwards can only go up
  - algorithms only accept better trees
- problem: how to avoid getting stuck on a sub-optimal island

# How to avoid getting stuck on sub-optimal islands (hills)

- solution: multiple independent starts
  every random start -> one path through tree space




- usually run 100's, 1000s or even 10000s of random starts

# Bayesian Inference

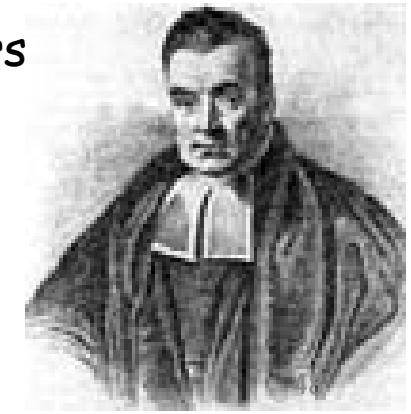
Posterior probability of phylogeny
probability of a tree conditioned on the observations.

Examine universe of possible trees (tree space)
and all possible parameters for evolutionary model
identifies combinations of branching patterns + model parameters
that give highest likelihood trees

In a sense, maximum likelihood with learning

- adjusts model as search progresses
  - better trees -> better estimates of model parameters

- Bayesian Inference invented in 1600's by Thomas Bayes
  - rediscovered in late 1990's
  - formally applied to phylogeny in 2000
  - MrBayes (2002) first widely useful implementation

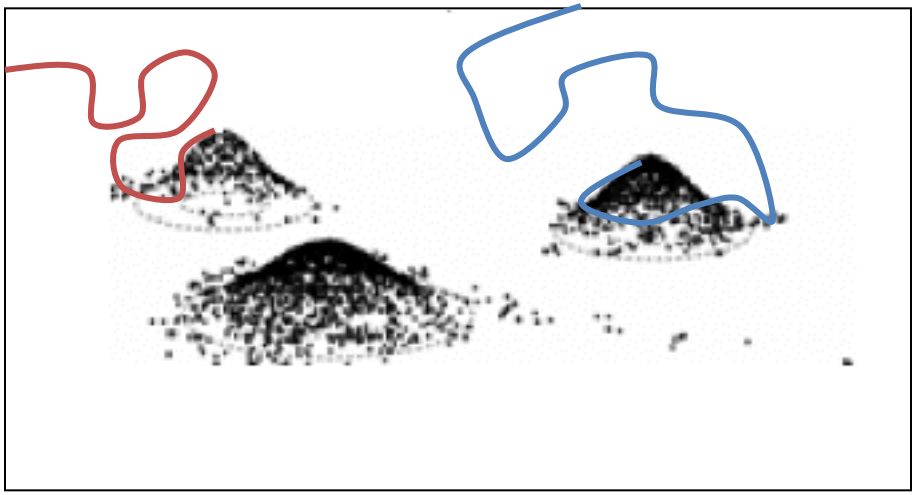


# BI – Searching Tree Space

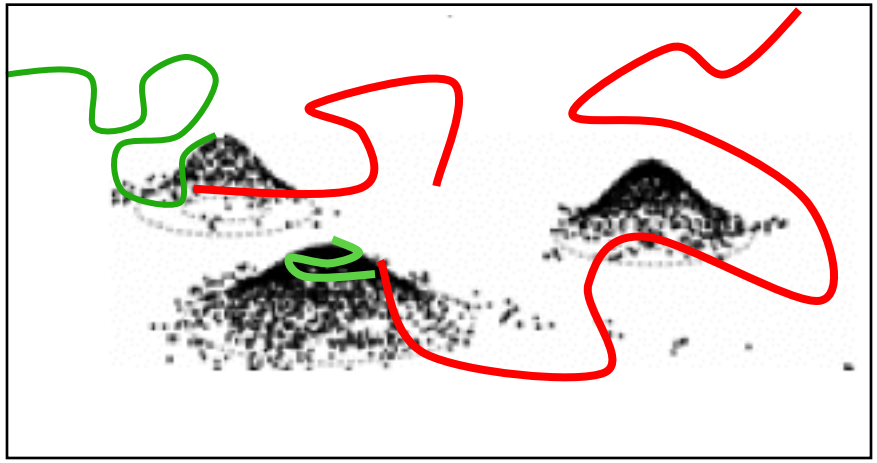
- MrBayes: MCMCMC algorithm to search tree space
  - (Metropolis-Hasting Coupled Markov Chain Monte-Carlo)

- four searches run in parallel (chains)
  - each chain = independent random walk through tree space

- But chains are not equal
  - 1 conservative (cold) chain, conservative rearrangements only (slow, step by step search)
  - 3 "heated" chains, multiple simultaneous rearrangements => large jumps through tree space



# Searching Tree Space with MCMC



- Most importantly, 4 chains talk to each other
  - heated chains mostly find bad trees
  - but occasionally may stumble across a new tree island
- when heated chain finds better tree
  - transforms into a "cold" chain
    - and old cold chain becomes "hot"
- hot chains essentially = random walk through tree space
    - avoids problem of "greedy" algorithm
- When is search "complete"? No improvement for long time…

# Why trees may lie?
# (where do trees go wrong)

Bad data

sequences aren't homologous (mixing orthologs, paralogs, xenologs/ horiz. gene transfer)
incorrect alignment, using misaligned regions of the alignment
too little difference between sequences (not enough data)
too much difference between sequences (too much homoplasy/convergent evolution)

Bad analyses

incorrect models: too much correction, not enough correction, incorrect model parameters
incorrect methods: UPGMA, unweighted parsimony for distantly related sequences

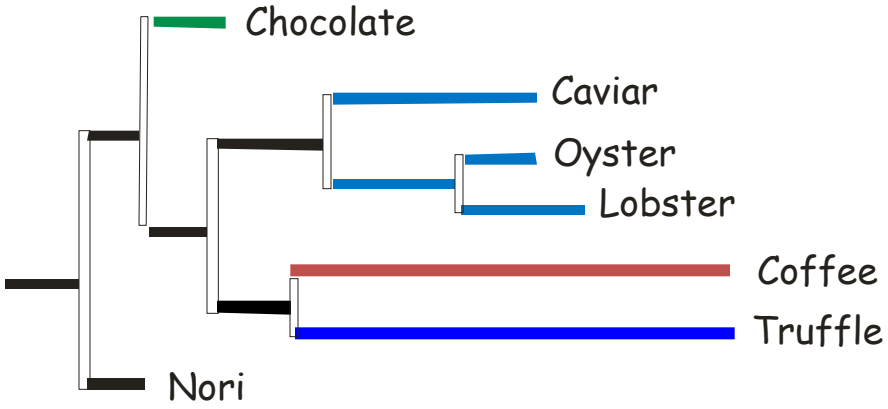
Over interpreting weak trees

BP < 70% : means it could be wrong, other hypotheses not ruled out
BP < 50% : means over half of BP replicates => something else!

Phylogenetic artefacts

some problems are extremely difficult
sequences very distantly related
sequences evolving at very different rates in different species

# Long Branch Attraction (The Felsenstein Zone)




Isolated long branches tend to attract each other

Rapidly evolving lineages are inferred to be closely related, regardless of their true evolutionary relationships

Two random sets of character states are more likely to resemble one another than either is to resemble any of the non-randomly associated sets of states among the other taxa

# What causes long branches?

A. Fast evolution

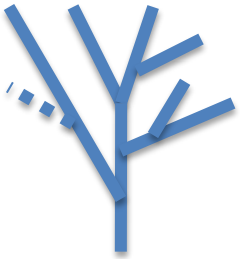
- increased selection

  parasites (evolutionary arms race), changing environment, …..

- relaxed selection

  founder effects, loss of function

  gene duplications -> partial loss of function



B. Species without close relatives (“isolated branches”)

- close relatives unknown or extinct
- close relatives existant, but not included in analysis

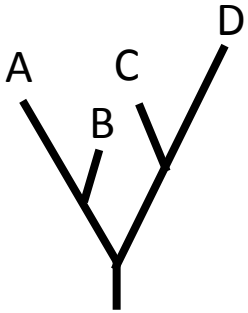
C. Bad evolutionary methods/models

- incorrect model e.g., overweighting simple mutations

# LBAs and BPs

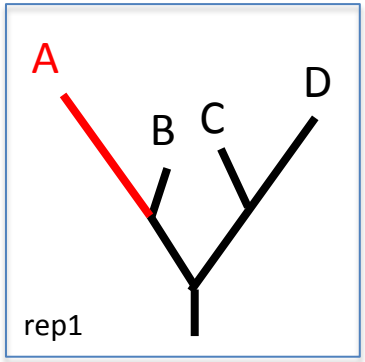
**Data set**

| | |
|---|---|
| seqA | ACCGTTCGGT |
| seqB | ATGGTTCAGA |
| seqC | ATGGATCGGA |
| seqD | ACCGACCGGA |



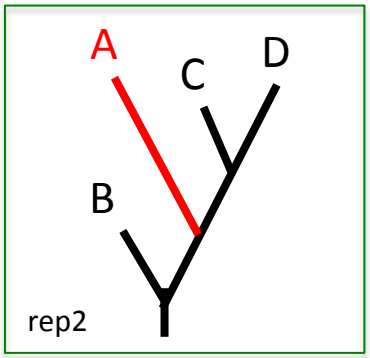

**BP rep 1**

| | |
|---|---|
| seqA | AGTTTCGGTA |
| seqB | AGTTTCAGAA |
| seqC | AGAATCGGAA |
| seqD | AGAACCGGAA |



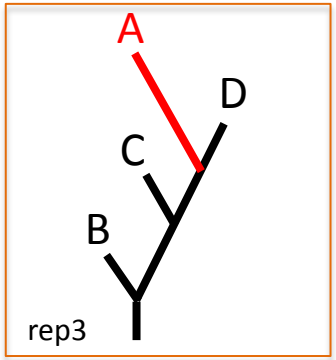

**BP rep 2**

| | |
|---|---|
| seqA | CTCCGCTTTC |
| seqB | TTCGGTTATT |
| seqC | TTCCGTAATT |
| seqD | CTCGGCTATT |

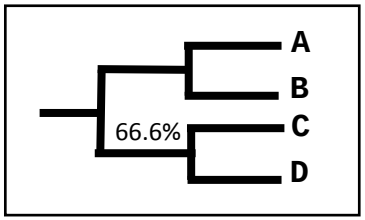



**BP rep 3**

| | |
|---|---|
| seqA | ACCGCTCGGT |
| SeqB | ATTGCTCAGA |
| seqC | ATTGCTCGGA |
| seqD | ACCGCTCGGA |




bootstrap consensus tree:



# Solutions to LBA problems

A. remove the “offending” branch

(if you don’t need it)

Hampl et al. (2009) *Proc Natl Acad Sci USA*

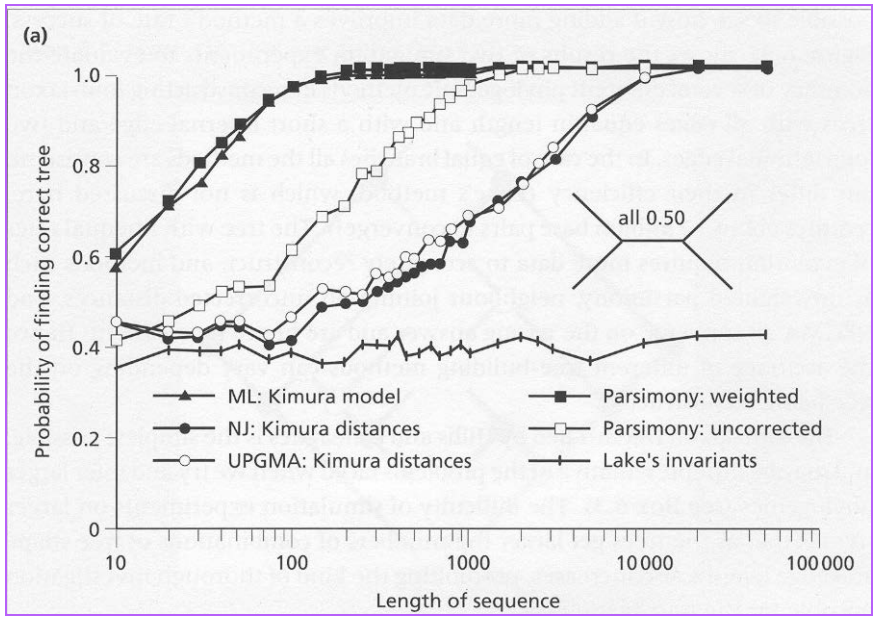
B. more data

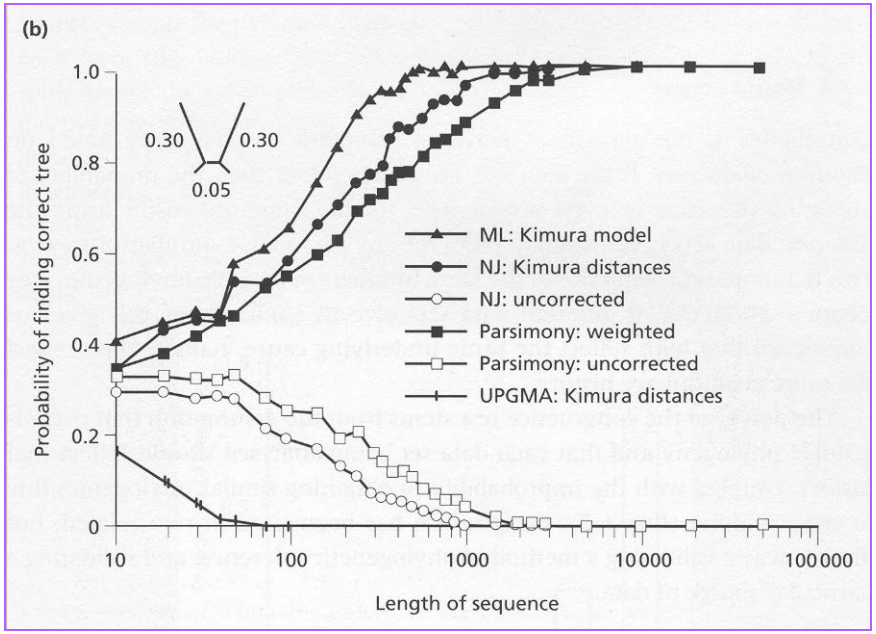
- given infinite data, most methods give the true tree

C. better evolutionary model

- give a perfect model, all methods give the true tree
- perfect model = time machine

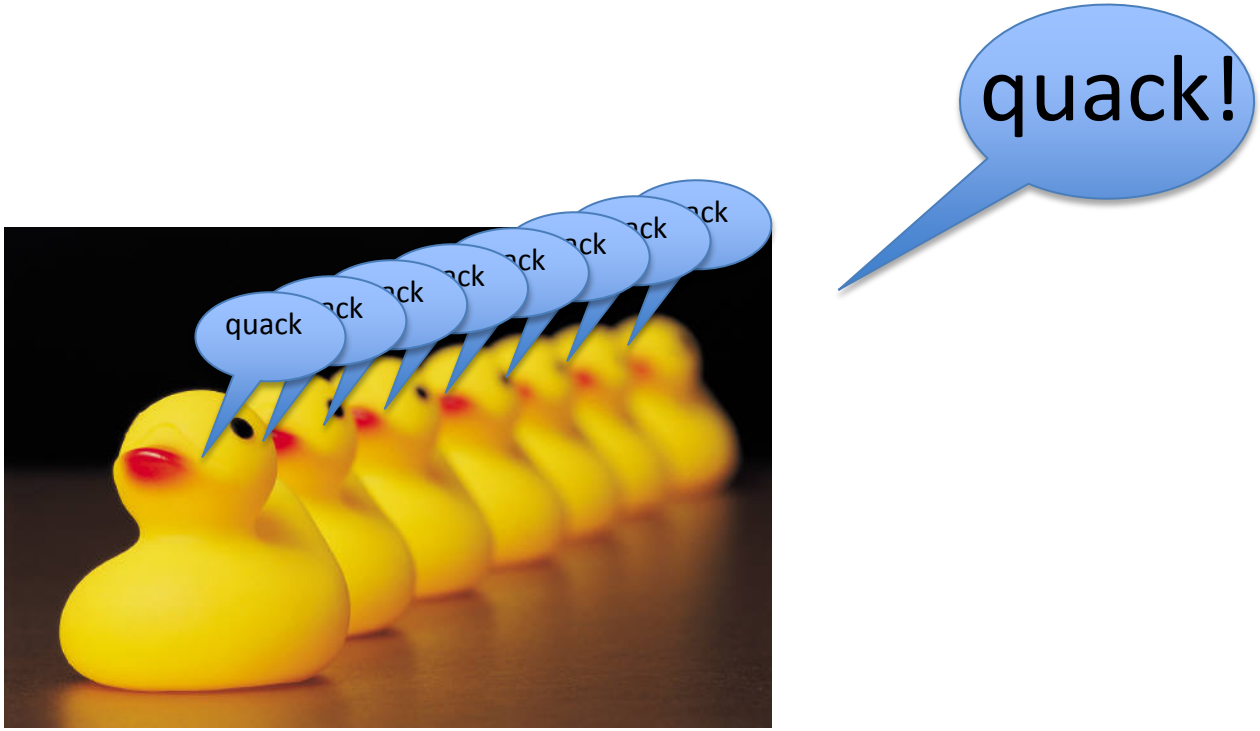
# LBA - Which Phylogenetic Method is Best?




The accuracy of several different phylogenetic methods in reconstructing two four-taxon trees with (a) all edges equal in length and (b) with a short internal edge and two long terminal edges. In each graph the proportion of analyses that recovered the correct tree is plotted against the length of the simulated sequences. From Huelsenbeck et al. (1996).

# Combining data



# Three schools of thought

1. Always combine everything: “total evidence school”
   all the data = most comprehensive approach

   This assumes there’s no such things as “bad data”
   bad data = data inappropriate for the question
   e.g., species trees with laterally transferred genes

2. Never combine data:
   instead: use consensus – agreement among trees

   Pros: congruence/consistency = strongest form of proof in evolutionary study
   Cons: consensus can not discover anything new
   single gene trees – poor resolution of many branches, especially deep ones
   only combining gives enough information to resolve all branches

3. Conditional combination
   test the data for congruence, only combine congruent data

# Combining vs consensus

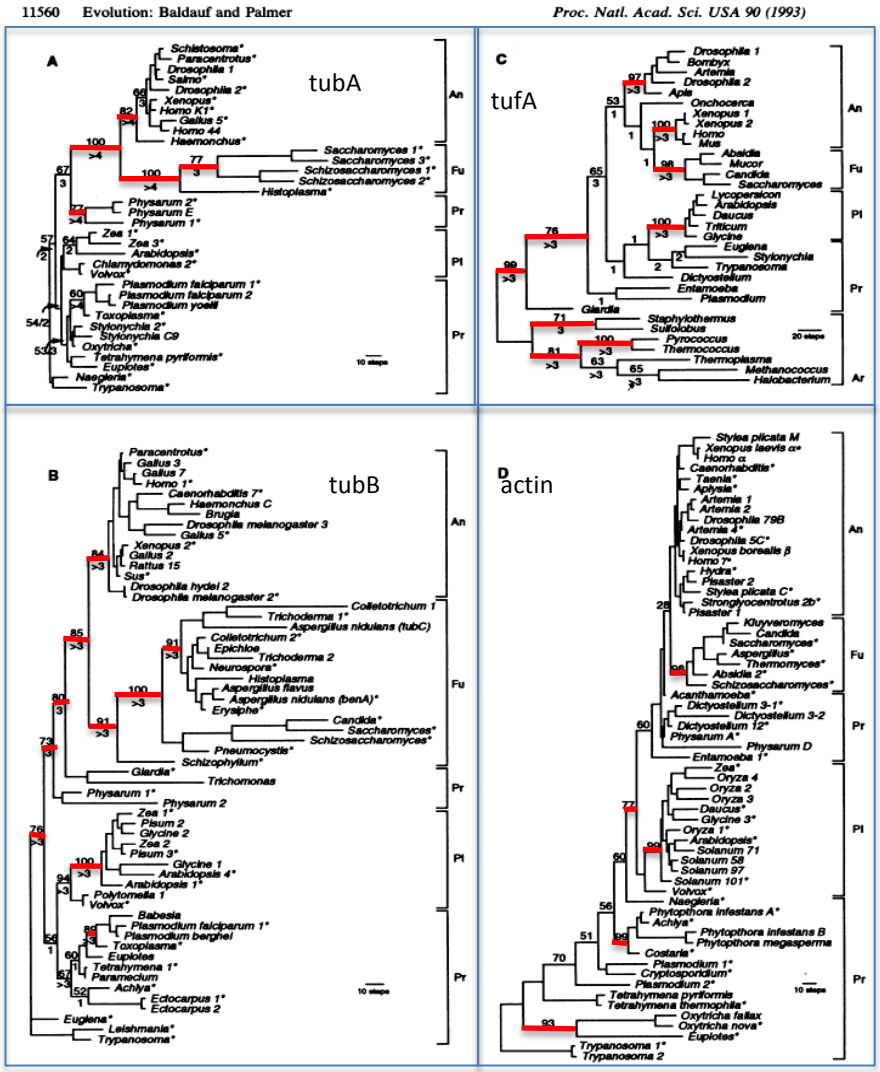
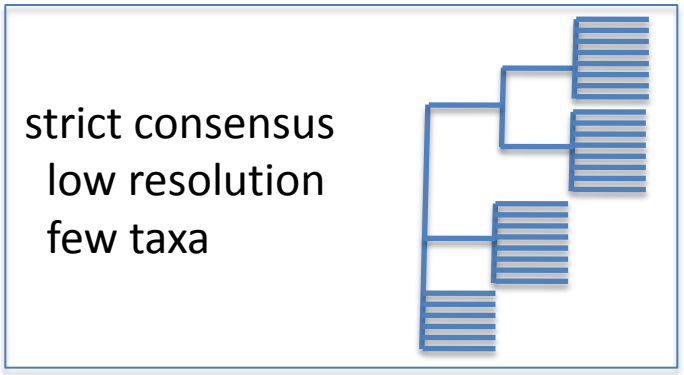
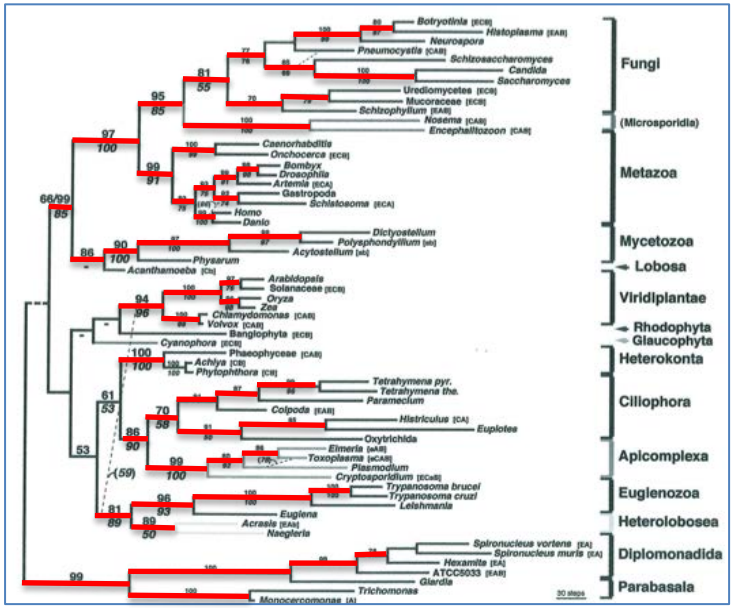



FIG. 2. (Figure legend appears at the bottom of the opposite page.)

Individual trees for 4 different proteins




combined sequence tree

all OTUs, most branches BP>80%