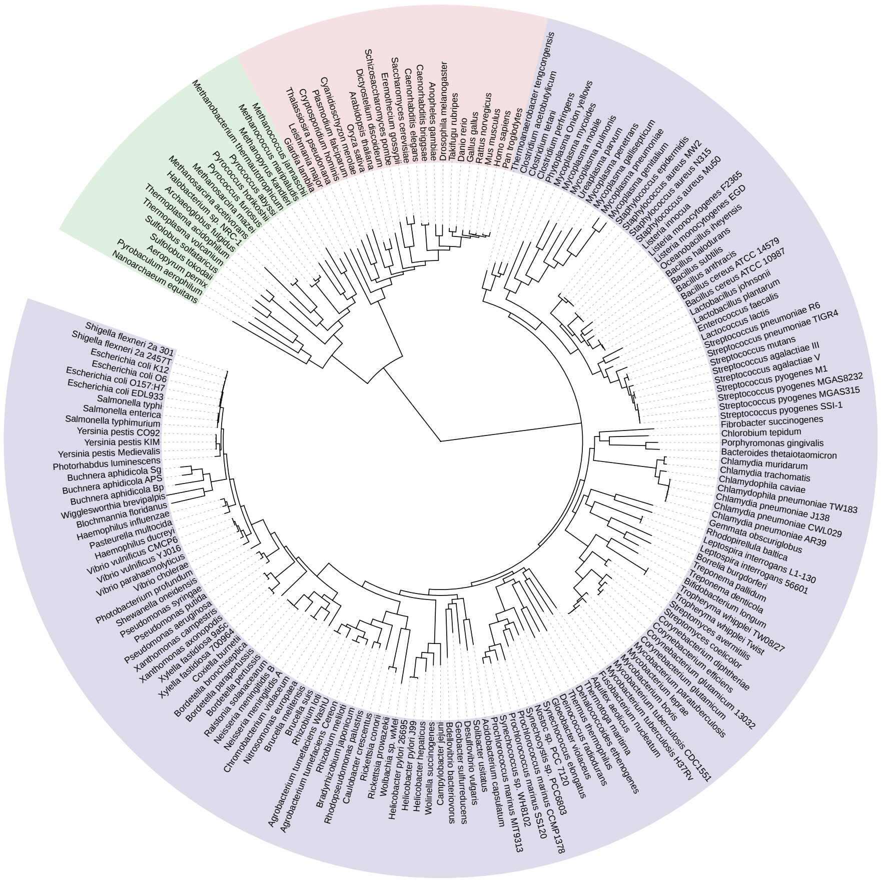
Phylogenetics and Phylogenomics

Bioinformatics Workshop for *M. tuberculosis*
Genomics and Phylogenomics

July 10-14, 2018 @The Philippine Genome Center



Ulas Karaoz, PhD

Ecology Department
Berkeley Lab

BERKELEY LAB
Lawrence Berkeley
National Laboratory
Bringing Science Solutions to the World

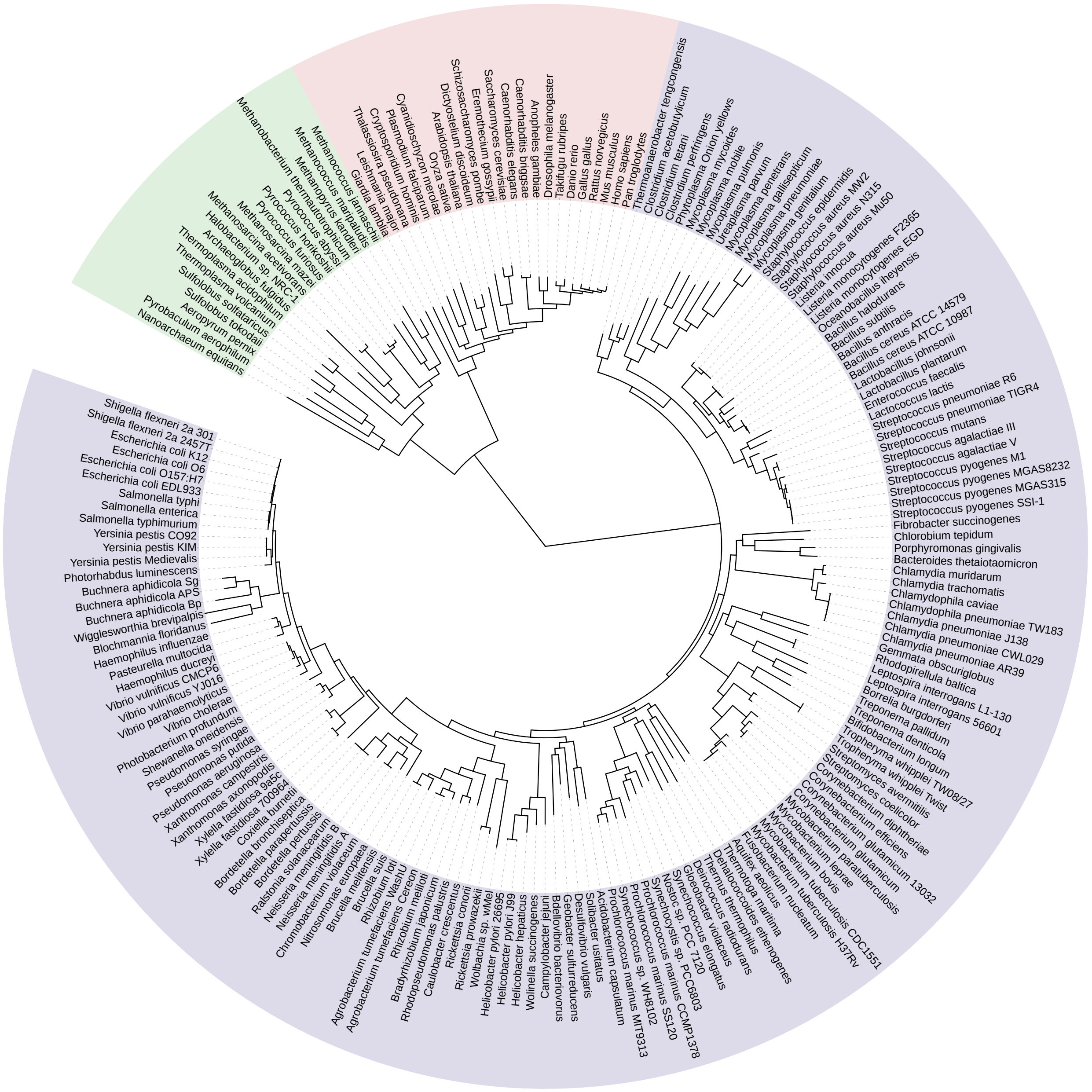
https://eesa.lbl.gov/profiles/ulas-karaoz, Email: ukaraoz@lbl.gov, Twitter: @ukaraoz

Learning Objectives

Learn:

- What a phylogenetic tree is,
- How to interpret phylogenetic tree,
- How to build phylogenetic trees

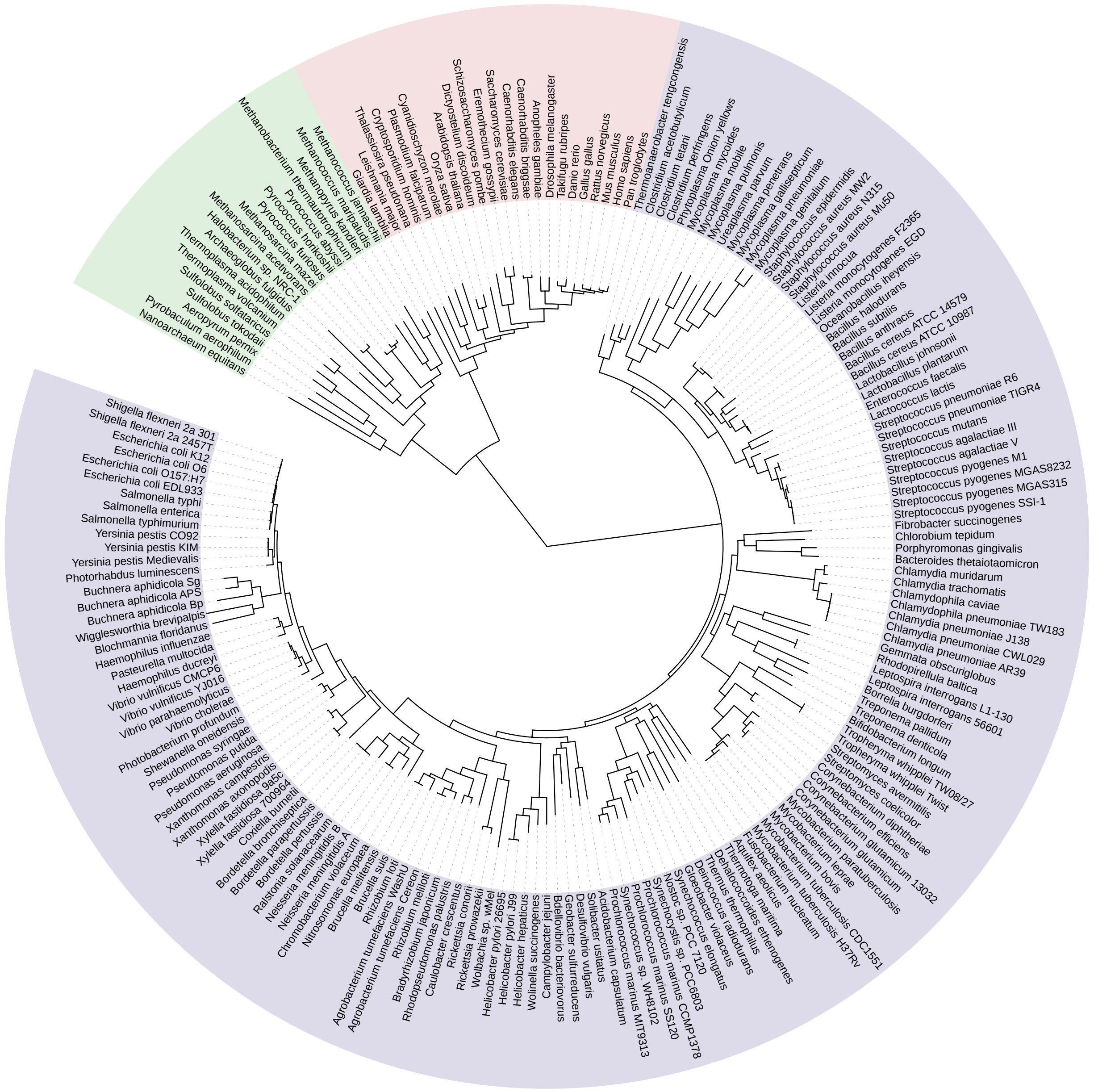
Phylogenetics



https://itol.embl.de/itol.cgi

- The study of evolutionary relationships among species, genes, organelles, or other biological entities.
- Usually involves molecular sequence data (DNA and proteins), but phylogenetic analysis can be performed using morphology (e.g. bone structure, flower structure, even language).
- The inferred evolutionary relationships are usually depicted as phylogenetic trees.

Phylogenetic Tree

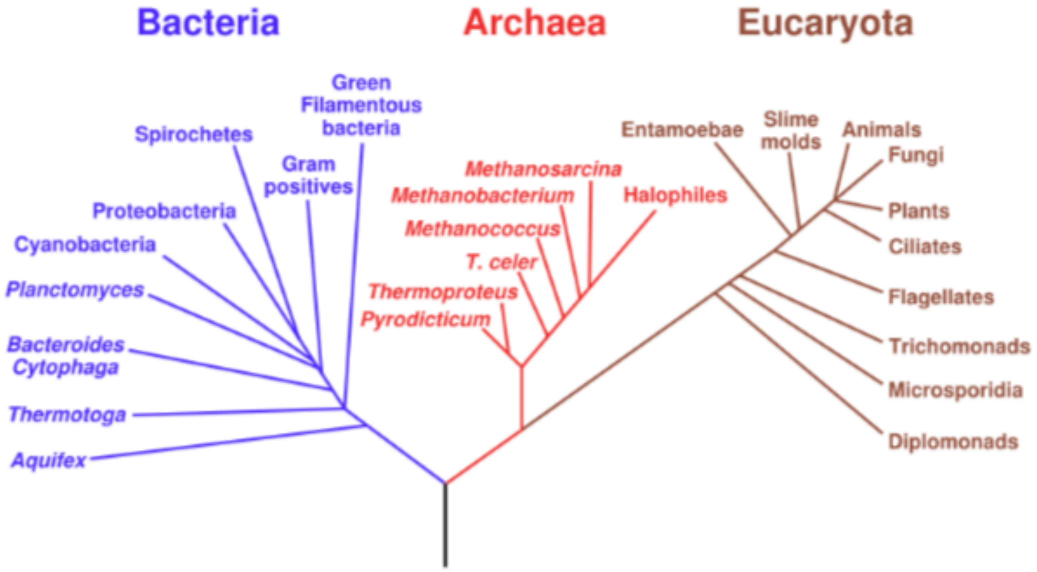


https://itol.embl.de/itol.cgi

- A phylogenetic tree graphically represents a hypothetical evolutionary process.
- It represents an estimated pedigree of the inherited relationships among molecules ("gene trees") or species ("species trees").
- Each node with descendants represents the most recent common ancestor of the descendants.
- The edge lengths in some trees correspond to time estimates.

Tree Terminology

- A phylogenetic tree is a structure used to model the evolutionary history of a group of sequences or organisms.
- Consists of nodes connected by branches.
- The terminal nodes are called leaves, or operational taxonomic units (OTUs) and represent sequences or species for which data was obtained. They usually represent living species
- Internal nodes represent hypothetical ancestral species.
- The ancestor of all the sequences or species in a given tree is called the root. Not all trees have a root.

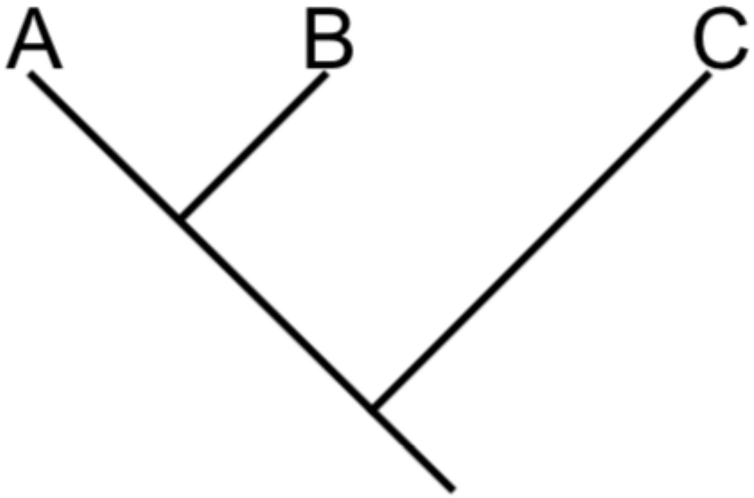


Phylogenetic tree built from rRNA gene sequences.

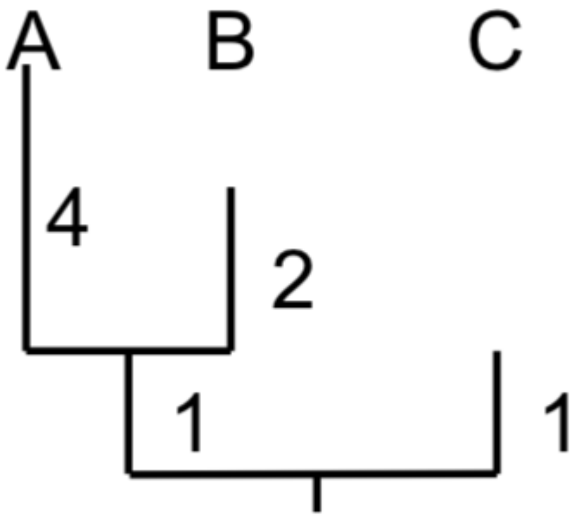
Fox and Woese (1977)

Cladograms vs Phylograms

Cladogram

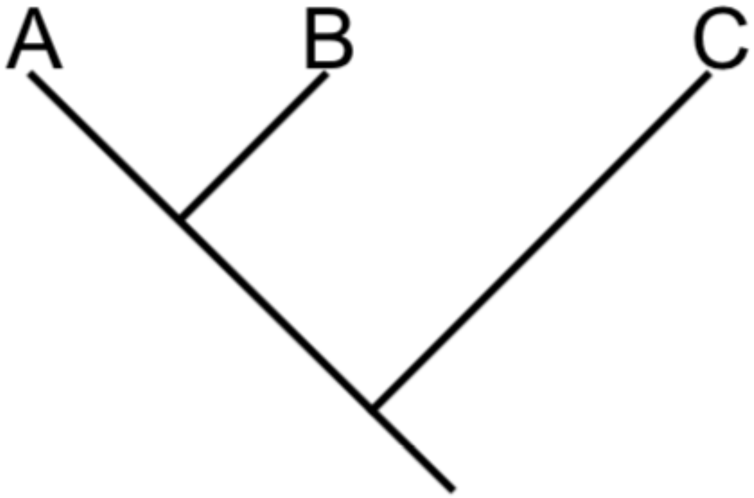


Phylogram

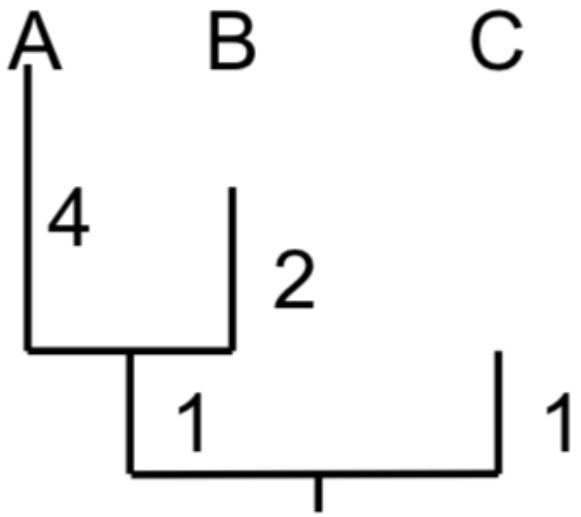


Cladograms vs Phylograms

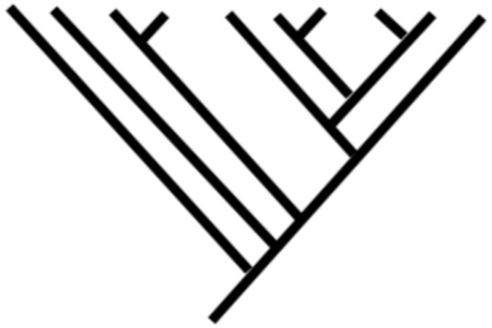
Cladogram



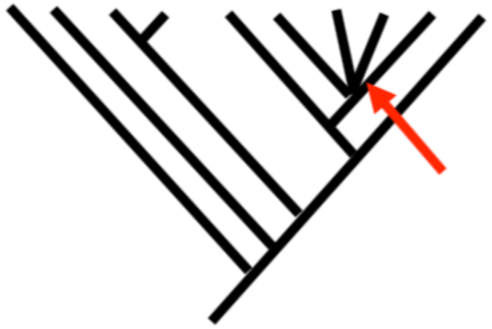
Phylogram



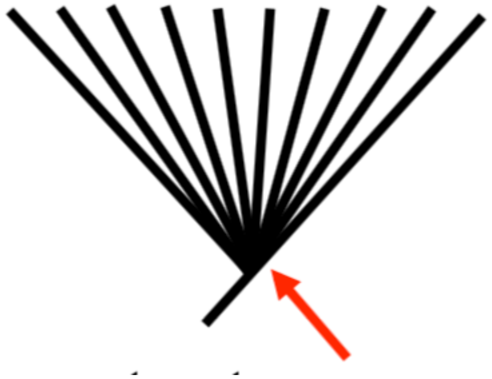
Fully resolved



Partially resolved

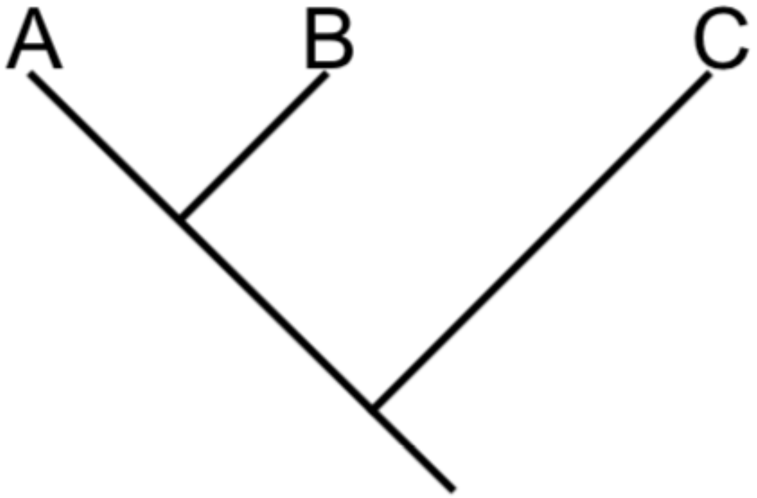


Star Tree

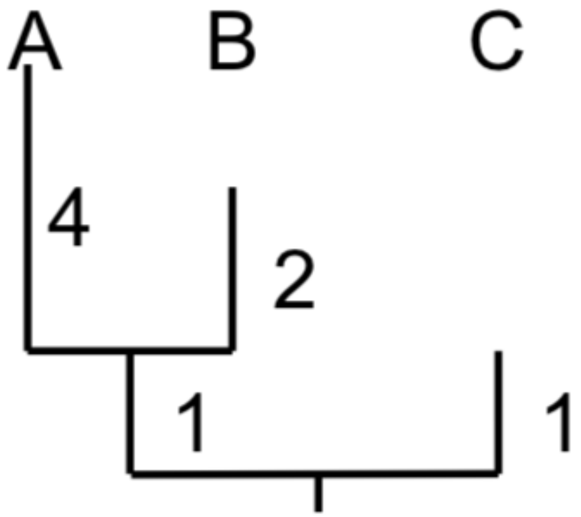


Cladograms vs Phylograms

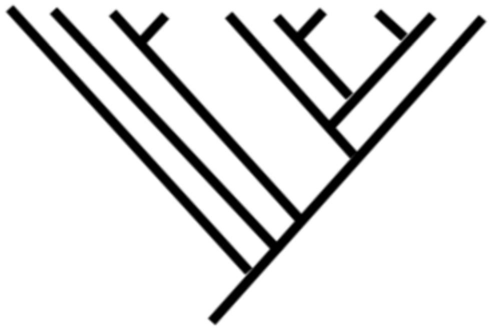
Cladogram



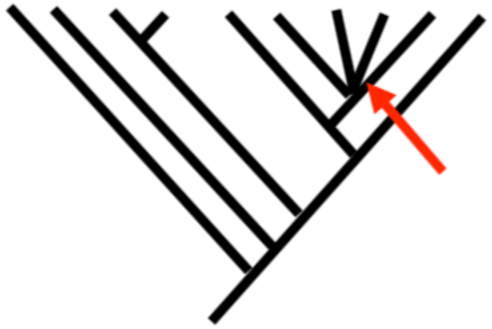
Phylogram



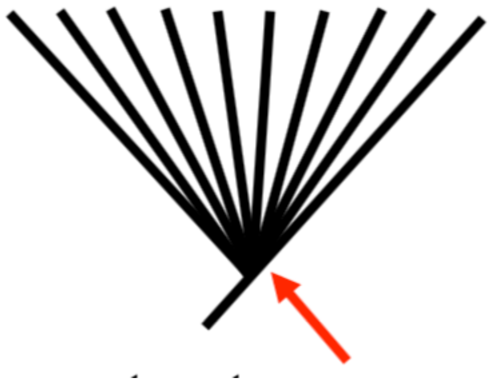
Fully resolved



Partially resolved



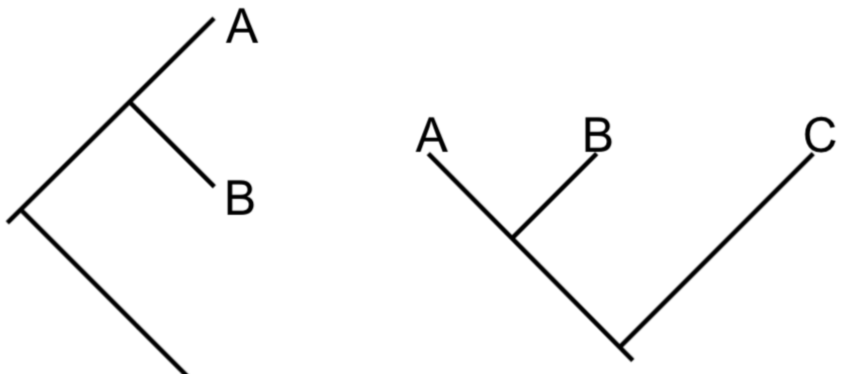
Star Tree



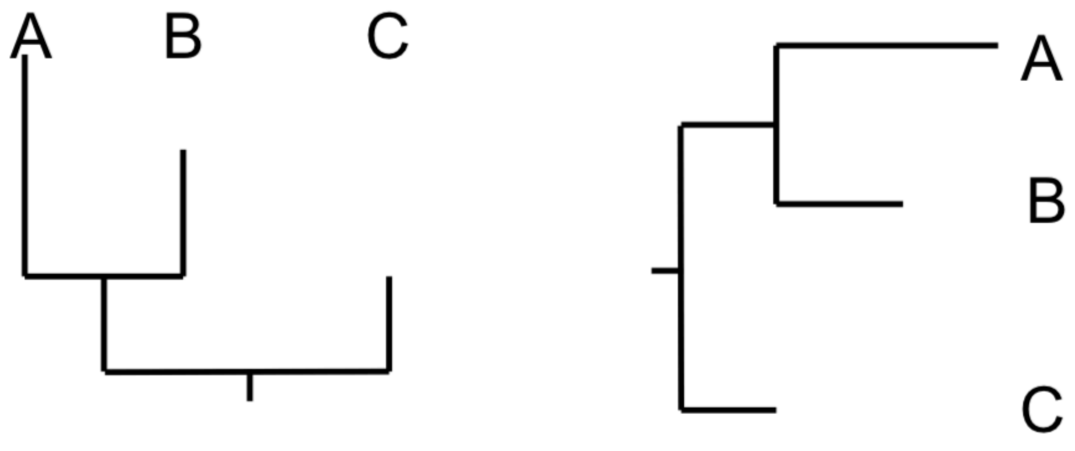
Polytomy

Tree Orientation & Rotation

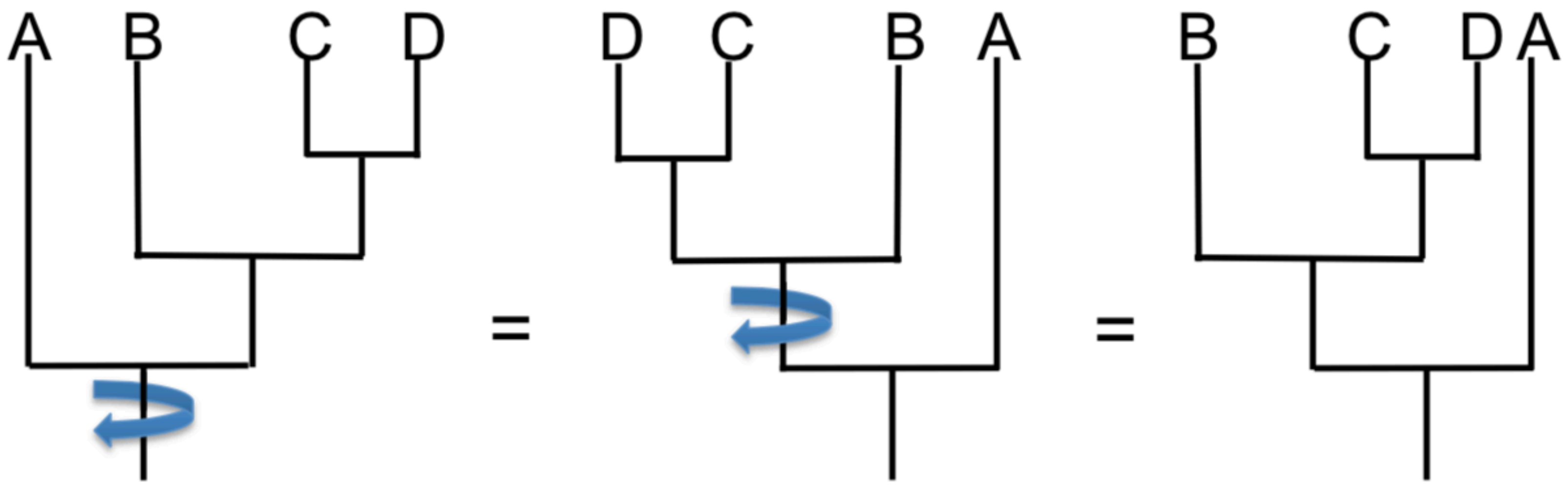
Identical Trees



Identical Trees

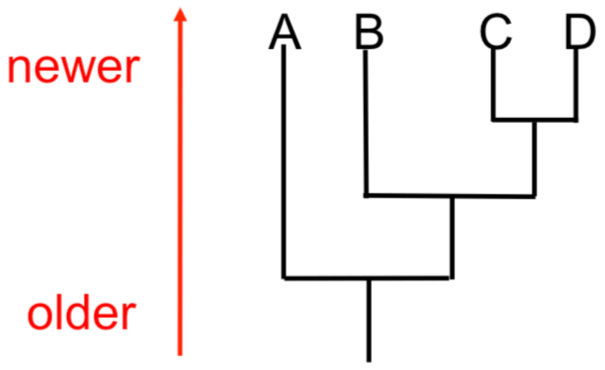


Identical Trees

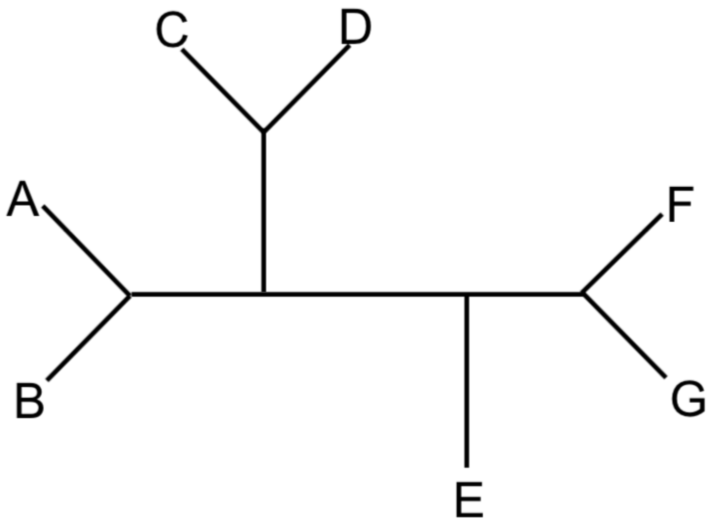


Rooting Trees

Rooting Trees

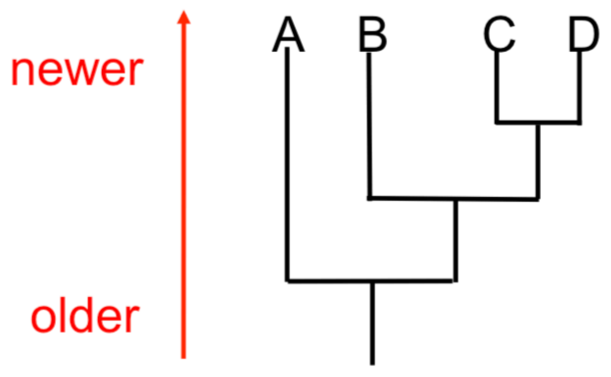


rooted tree



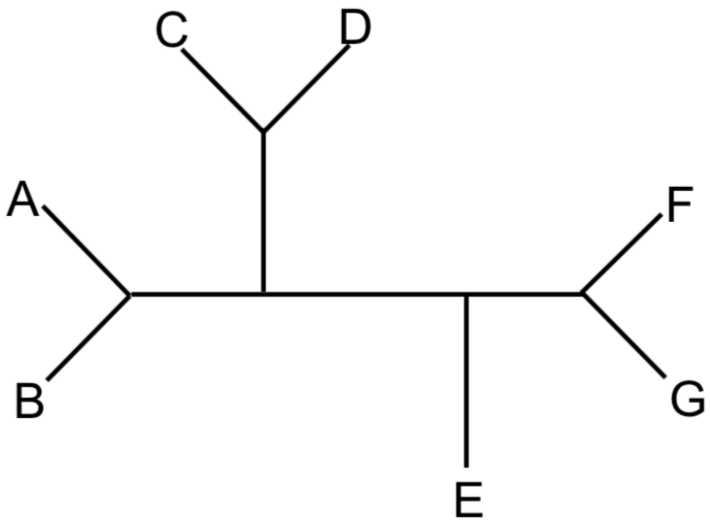
Rooting Trees

Rooted



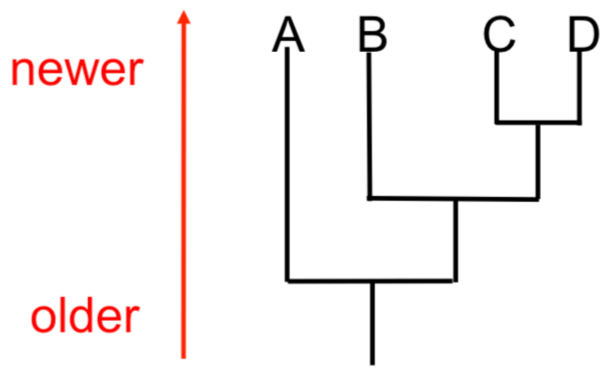
rooted tree

Unrooted



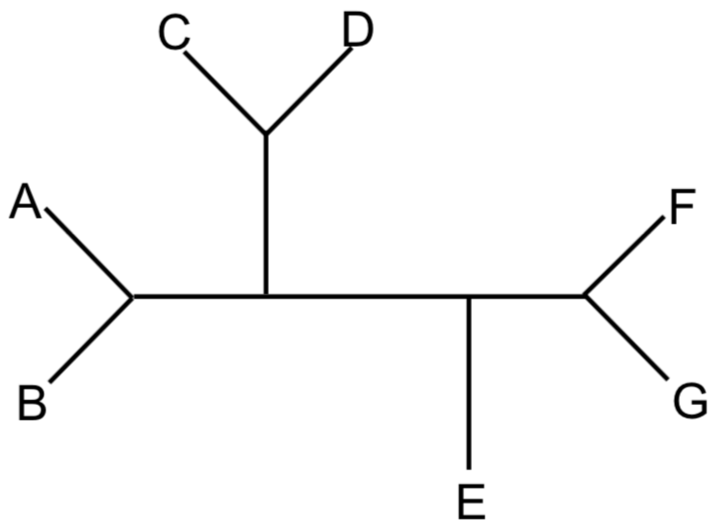
Rooting Trees

Rooted

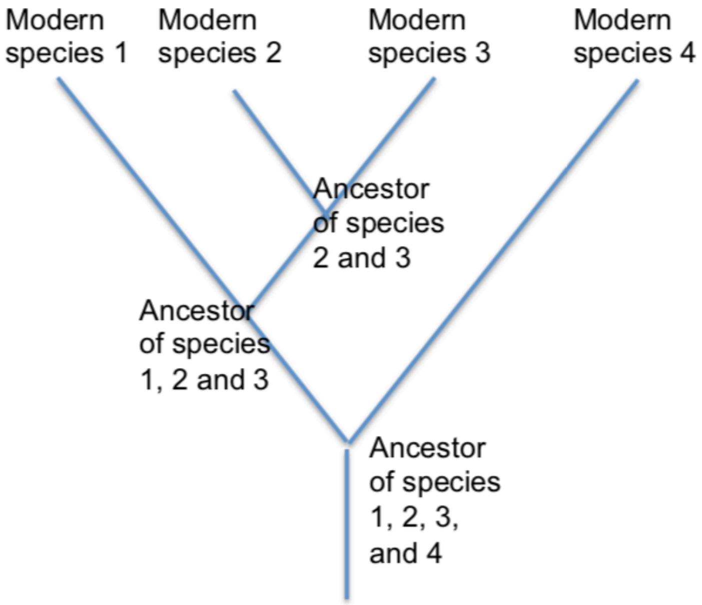


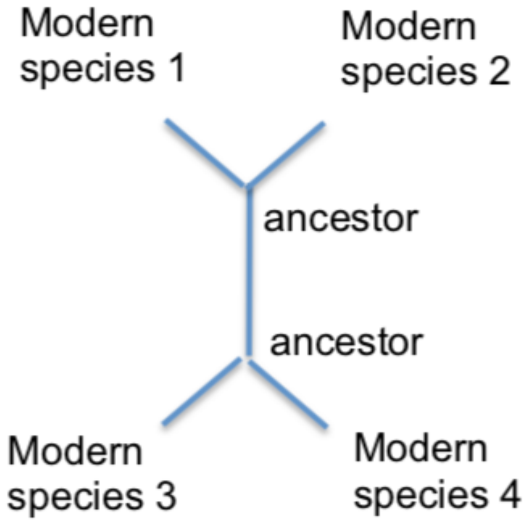
rooted tree

Unrooted



IMPORTANT: Only a rooted tree shows which ancestors led to which species


Modern species 1
Modern species 2
Modern species 3
Modern species 4
Ancestor of species 2 and 3
Ancestor of species 1, 2 and 3
Ancestor of species 1, 2, 3, and 4



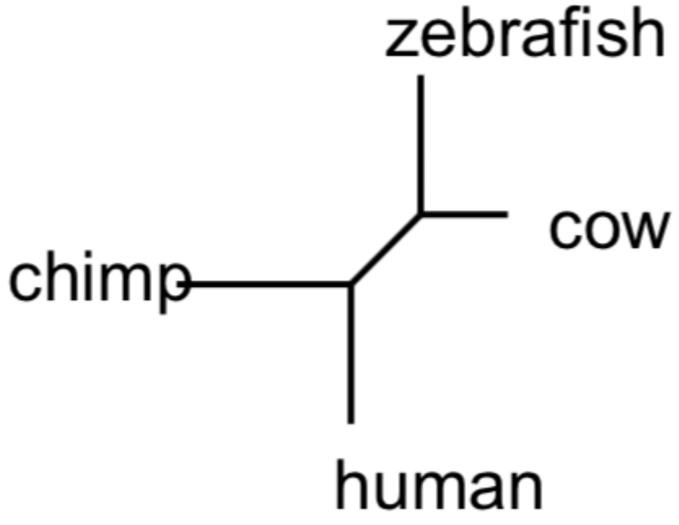
How do you root a tree?

How do you root a tree?

Outgroup: a sequence or species that is thought to be more distantly related to the **ingroup** members than they are to each other

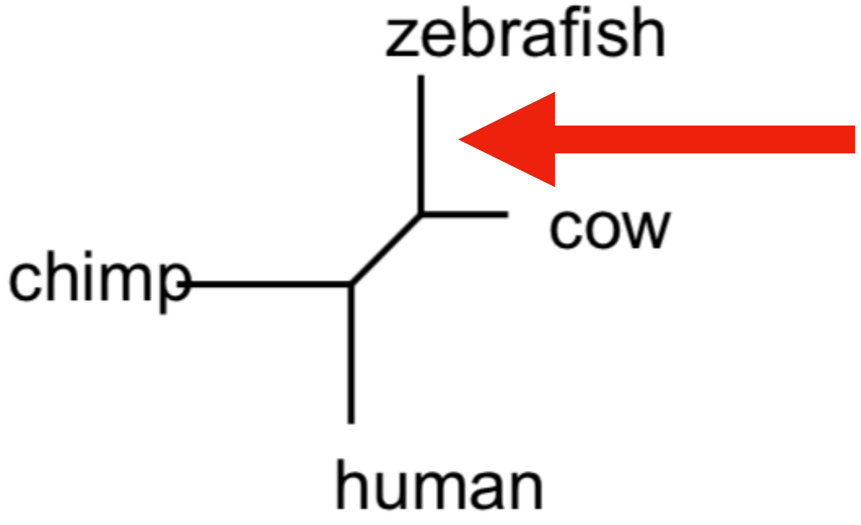
How do you root a tree?

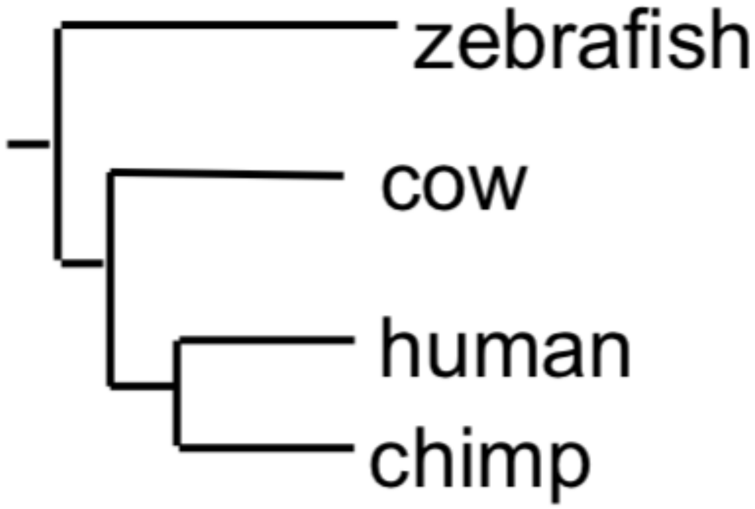
Outgroup: a sequence or species that is thought to be more distantly related to the **ingroup** members than they are to each other



How do you root a tree?

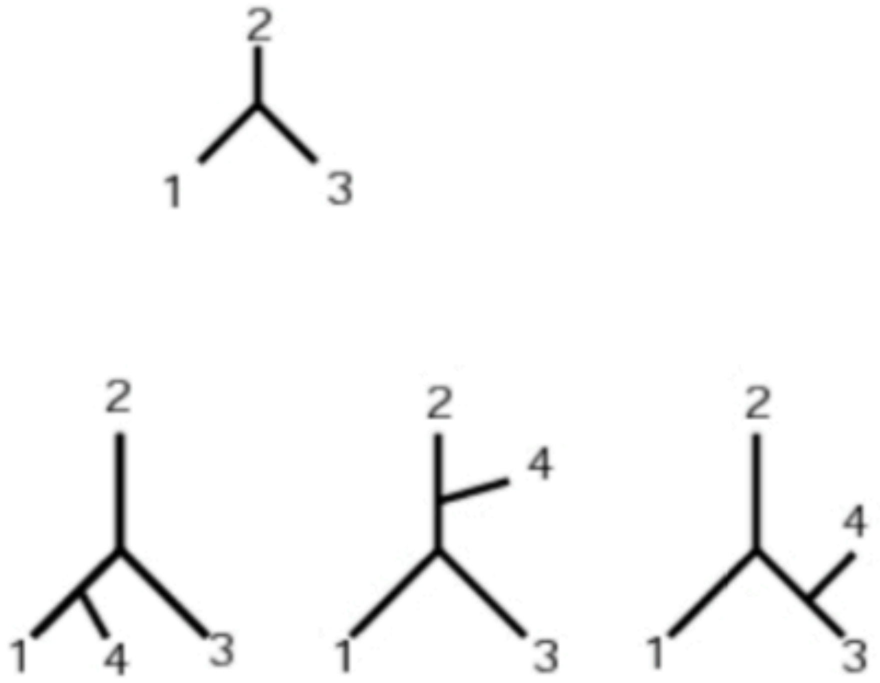
Outgroup: a sequence or species that is thought to be more distantly related to the **ingroup** members than they are to each other

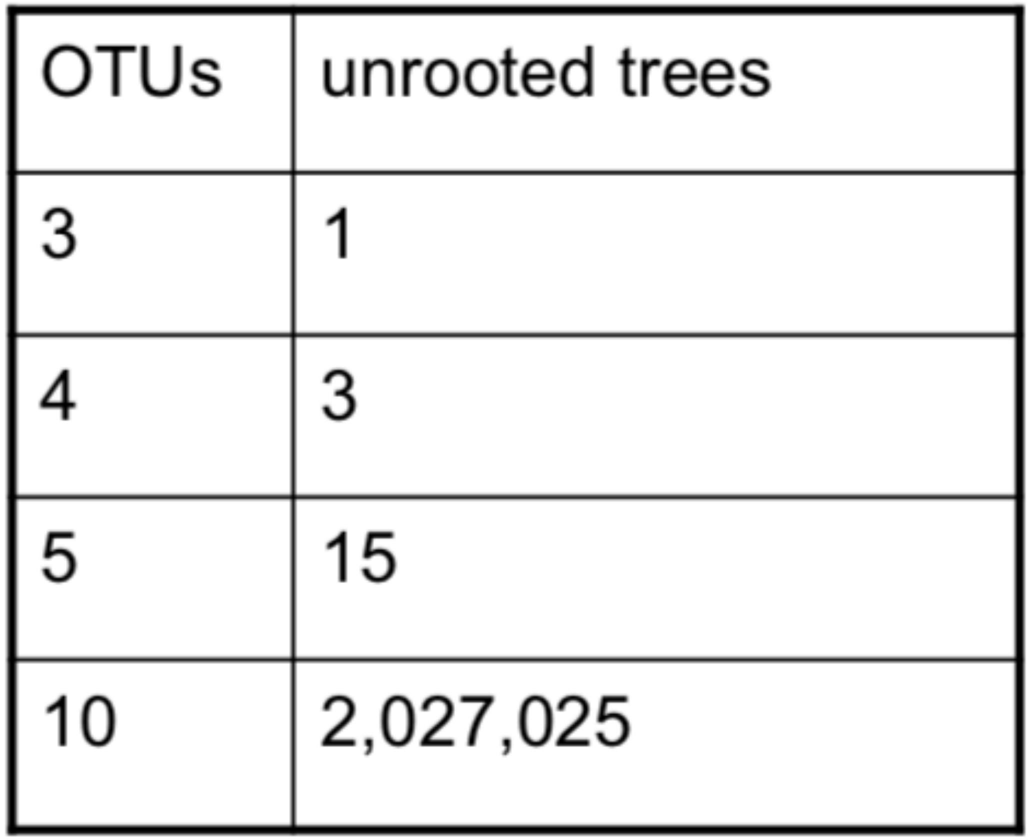




How Many Possible Trees?

Number of possible trees grows geometrically with number of species





OTUs	unrooted trees
3	1
4	3
5	15
10	2,027,025

How to build a tree?

A phylogenetic tree is a hypothesis about the evolutionary relationships between organisms.

well, until "time travel is invented"...

How to build a tree?

Building a tree **based on sequence** data requires?

How to build a tree?

Building a tree **based on sequence** data requires?

1. a multiple sequence alignment (MSA)

MSA column = homology

```
Species1 -TCAGGA-TGAAC--
Species2 ATCACGA-TGAACC-
Species3 ATCACGAATGATCC-
Species4 -TCACGAATGATCGC
Species5 -TCACGAATGATCGC
```

How to build a tree?

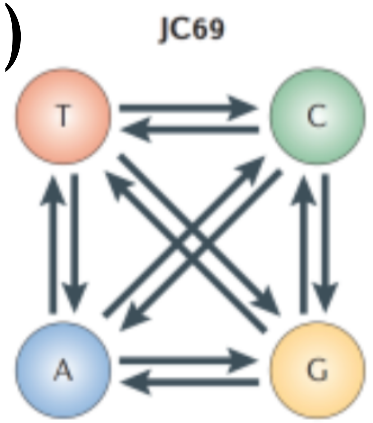
Building a tree **based on sequence** data requires?

1. a multiple sequence alignment (MSA)

MSA column = homology

```
Species1 -TCAGGA-TGAAC--
Species2 ATCACGA-TGAACC-
Species3 ATCACGAATGATCC-
Species4 -TCACGAATGATCGC
Species5 -TCACGAATGATCGC
```

2. a model of evolution (explicit or assumed)



How to build a tree?

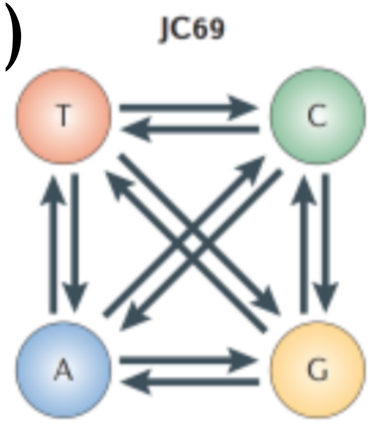
Building a tree **based on sequence** data requires?

1. a multiple sequence alignment (MSA)

MSA column = homology

```
Species1 -TCAGGA-TGAAC--
Species2 ATCACGA-TGAACC-
Species3 ATCACGAATGATCC-
Species4 -TCACGAATGATCGC
Species5 -TCACGAATGATCGC
```

2. a model of evolution (explicit or assumed)



3. a tree building algorithm

How to build a tree?

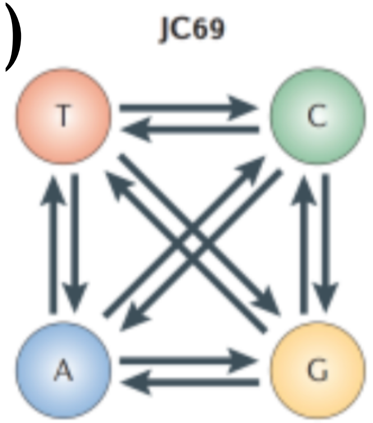
Building a tree **based on sequence** data requires?

1. a multiple sequence alignment (MSA)

MSA column = homology

```
Species1 -TCAGGA-TGAAC--
Species2 ATCACGA-TGAACC-
Species3 ATCACGAATGATCC-
Species4 -TCACGAATGATCGC
Species5 -TCACGAATGATCGC
```

2. a model of evolution (explicit or assumed)



3. a tree building algorithm
    - distance-based
    - character based
    - bayesian

Evolutionary Models

- To build accurate phylogenetic trees, we need to consider how sequences mutate. Observations of these mutations have been incorporated into evolutionary models.

- Evolutionary models are used to estimate the evolutionary distance between sequences (<u>branch lengths</u>).

Evolutionary Models

```
Seq1: ATGGCTTAAATTCCG
Seq2: ATGCCTTAAAATCCG
```

#differences/#positions: 2/15

What about multiple substitutions?

Poisson distance correction

What about rates of mutation across positions?

Codon wobble position

Gamma distance correction takes into account variation in mutation rates

I. Distance Methods

Main idea: Get a distance matrix

sequences

1 TTATTAA
2 AATTTAA
3 AAAAATA
4 AAAAAAT

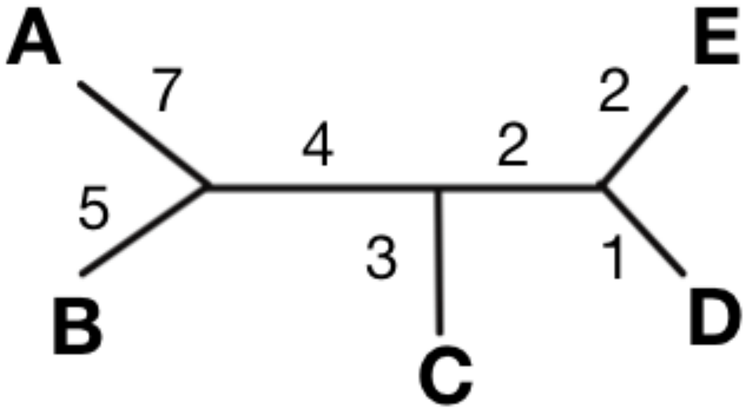
distance matrix

	A	B	C	D
A	0			
B	3	0		
C	5	4	0	
D	5	4	2	0

I. Distance Methods

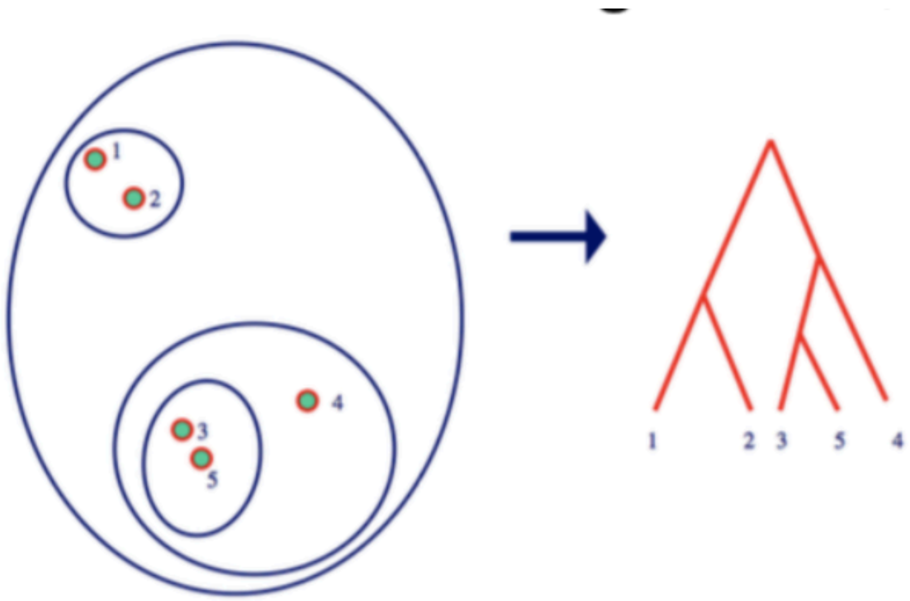
Main idea: Find a tree that is consistent with the distances

d	A	B	C	D	E
A	0				
B	12	0			
C	14	12	0		
D	14	12	6	0	
E	15	13	7	3	



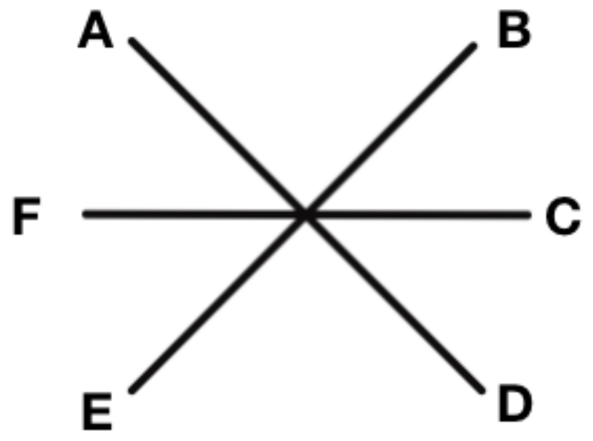
I. Distance Methods: UPGMA

- **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic mean.
- A distance-method heuristic.
- Clustering algorithm that involves building a tree rather than searching through all possible trees.
- Starts with most similar OTUs and builds a composite OTU.
- Distances to the composite OTU are calculated as arithmetic means.
- Of the composite and remaining OTUs, choose most similar, etc.



I. Distance Methods: Neighbor Joining

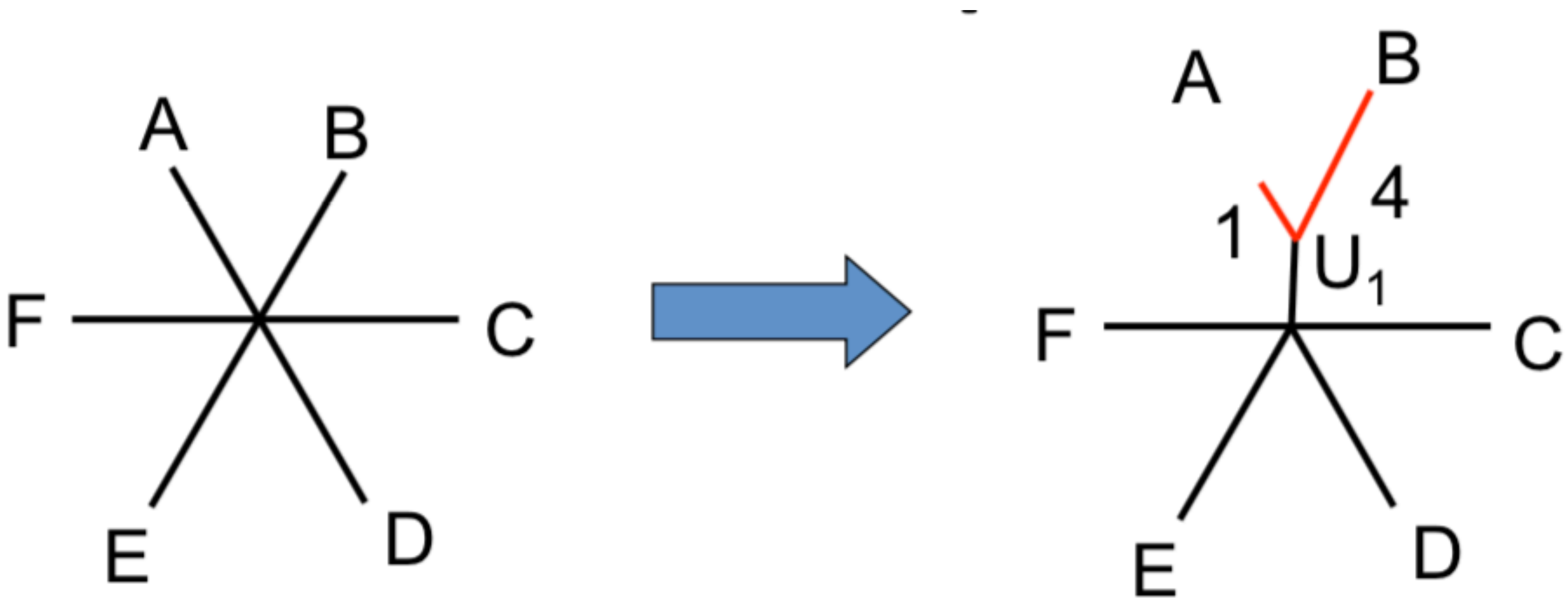
- Also clustering based heuristic
- UPGMA: inspect just closest neighbors
- NJ: takes into account average distances to other leaves as well.



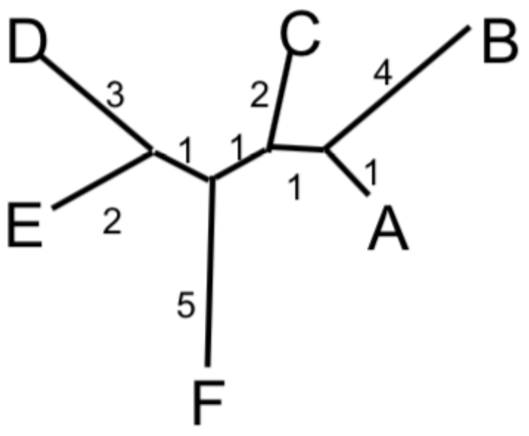
d	A	B	C	D	E	F
A	0					
B	5	0				
C	4	7	0			
D	7	10	7	0		
E	6	9	6	5	0	
F	8	11	8	9	8	0

d	B	C	D	E	F
B	-13				
C	-11.5	-11.5			
D	-10	-10	-10.5		
E	-10	-10	-10.5	-13	
F	-10.5	-10.5	-11	-11.5	-11.5

I. Distance Methods: Neighbor Joining



Result



- You get an unrooted tree with branch lengths.
- Optimality is NOT guaranteed
- Despite its naivety, works not bad in practice

I. Distance Methods: Take Home

- "quick and dirty",easy computation
- NJ superior to UPGMA
- throws out and "ignores" lots of information
- Next: character based, which are generally regarded as more credible

Character Based Tree Building Methods

Character Based Tree Building Methods

- Character-based: use sequences/MSA directly to infer a tree

Character Based Tree Building Methods

- Character-based: use sequences/MSA directly to infer a tree
- also called discrete methods

Character Based Tree Building Methods

- Character-based: use sequences/MSA directly to infer a tree
- also called discrete methods
- Two methods you should know about:

Character Based Tree Building Methods

- Character-based: use sequences/MSA directly to infer a tree
- also called discrete methods
- Two methods you should know about:

  1. **Maximum parsimony (MP):**

Character Based Tree Building Methods

- Character-based: use sequences/MSA directly to infer a tree
- also called discrete methods
- Two methods you should know about:

  1. **Maximum parsimony (MP):**

     minimize the number of evolutionary steps

Character Based Tree Building Methods

- Character-based: use sequences/MSA directly to infer a tree
- also called discrete methods
- Two methods you should know about:

  1. **Maximum parsimony (MP):**

     minimize the number of evolutionary steps

  2. **Maximum likelihood (ML):**

Character Based Tree Building Methods

- Character-based: use sequences/MSA directly to infer a tree
- also called discrete methods
- Two methods you should know about:

1. **Maximum parsimony (MP):**

   minimize the number of evolutionary steps

2. **Maximum likelihood (ML):**

   mostly likely tree given the data and a probabilistic model of sequence evolution

Tree Building: Maximum Parsimony

Occam's razor for tree building:

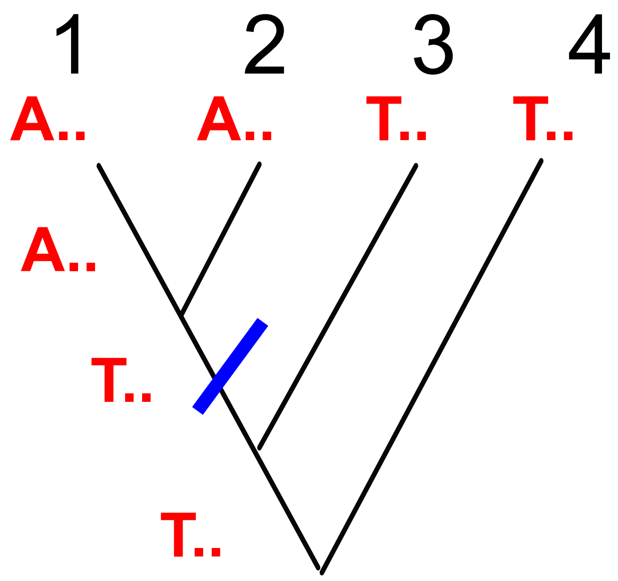
parsimony criterion = choose the simplest possible hypothesis

Consider position 1:

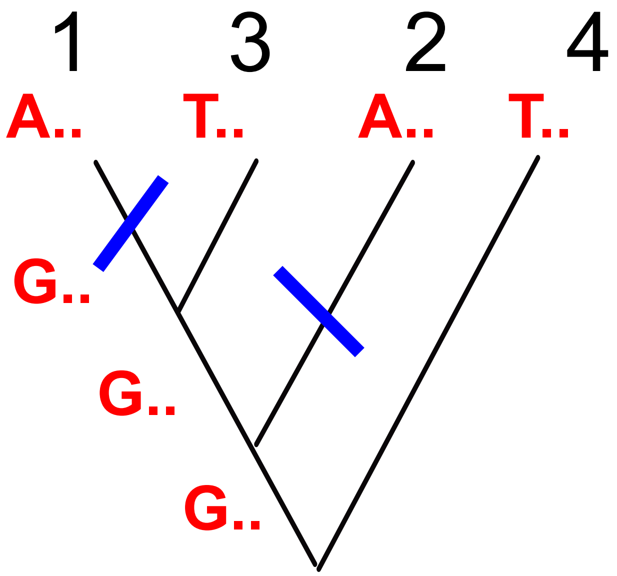
positions

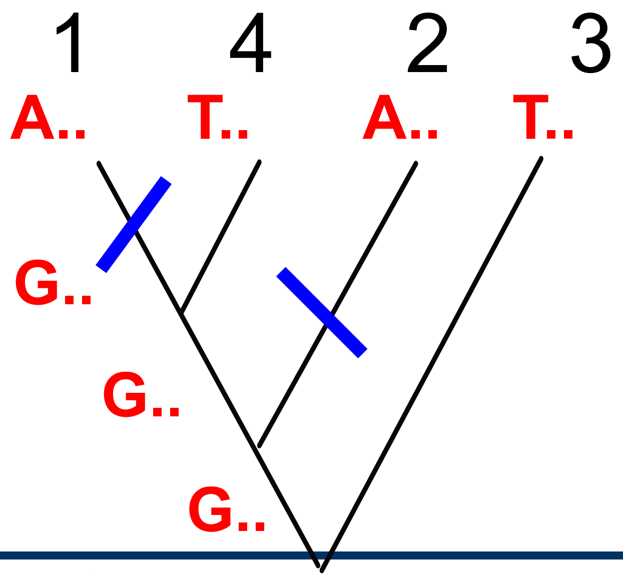
sequences

1	AAA
2	ATA
3	TAT
4	TTT



homoplasy

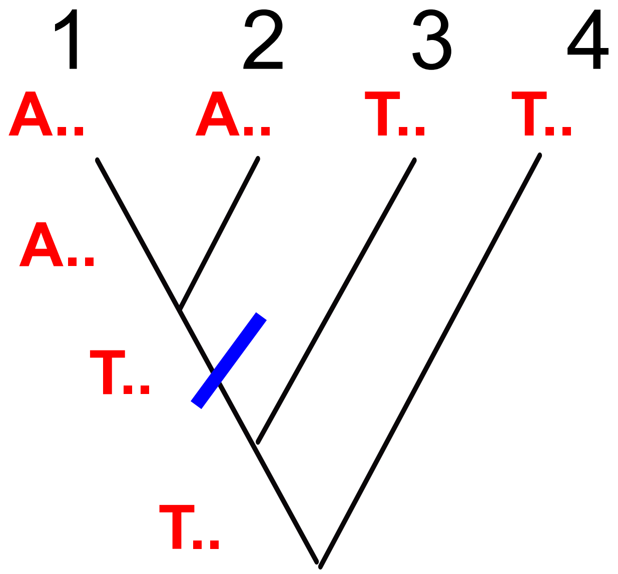




Tree Building: Maximum Parsimony

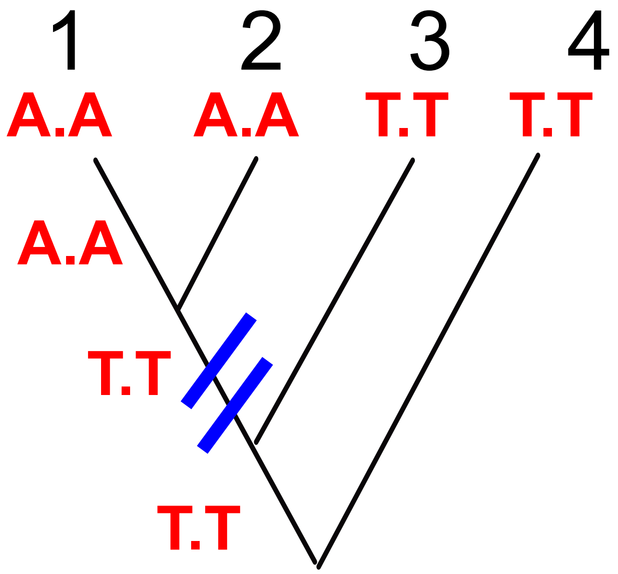
Consider position 3:

sequences	positions
1	AAA
2	ATA
3	TAT
4	TTT



Consider positions 1 and 3:

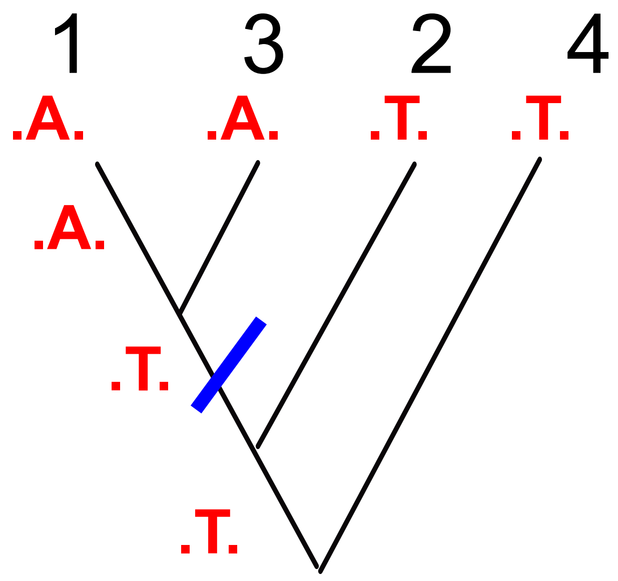
sequences	positions
1	AAA
2	ATA
3	TAT
4	TTT

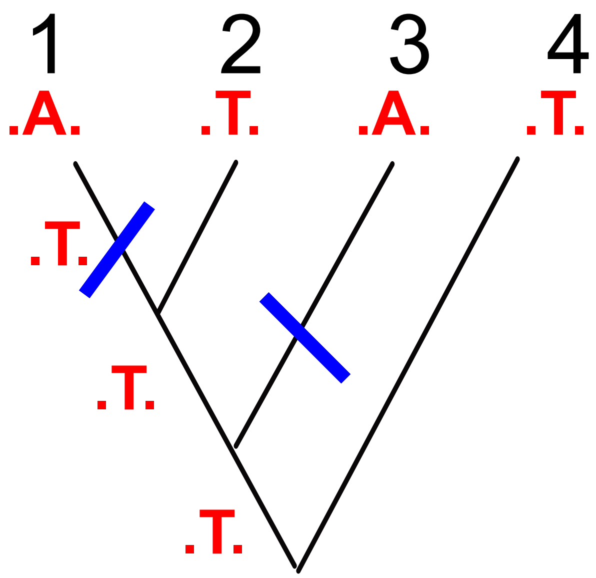


Tree Building: Maximum Parsimony

Now, consider position 2:

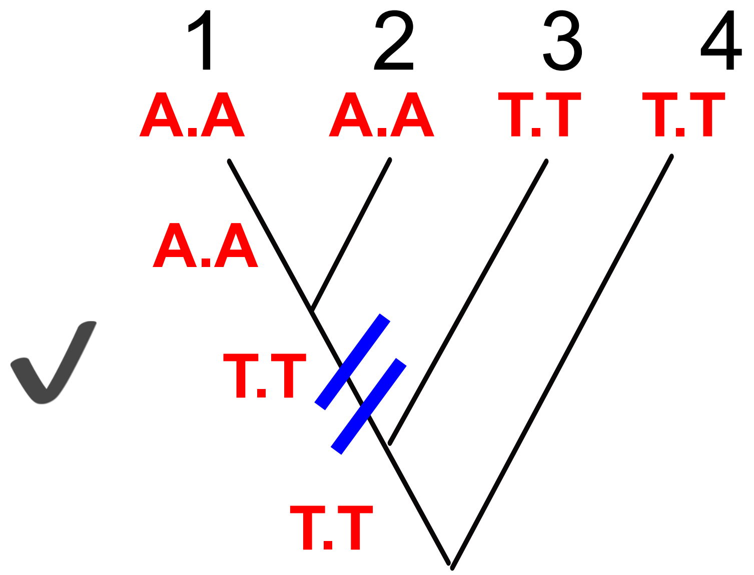
sequences	positions
1	AAA
2	ATA
3	TAT
4	TTT

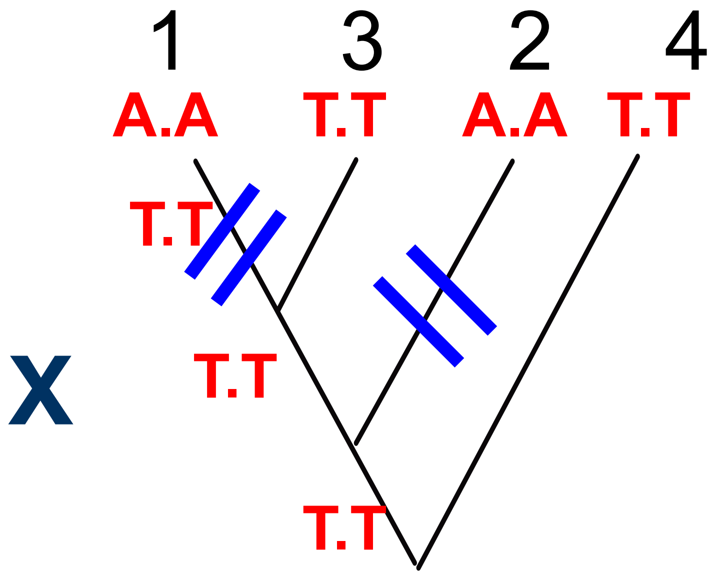




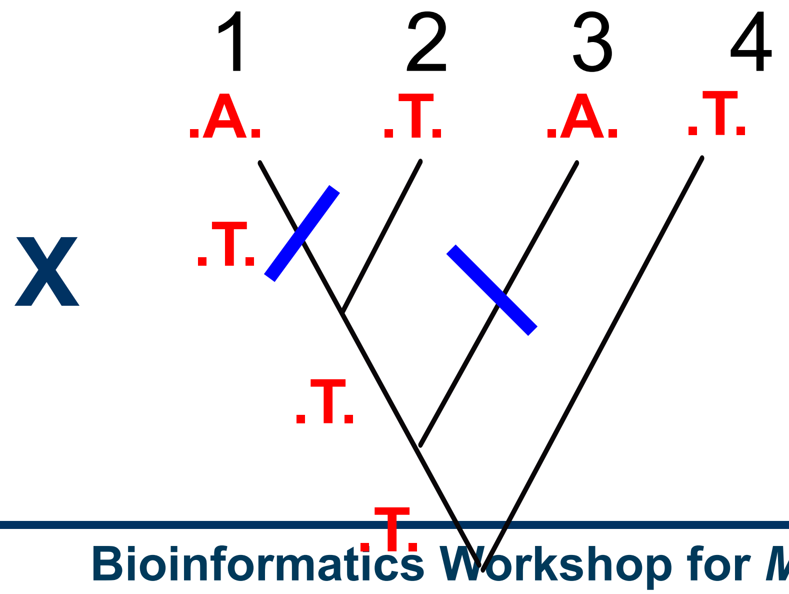
Tree Building: Maximum Parsimony

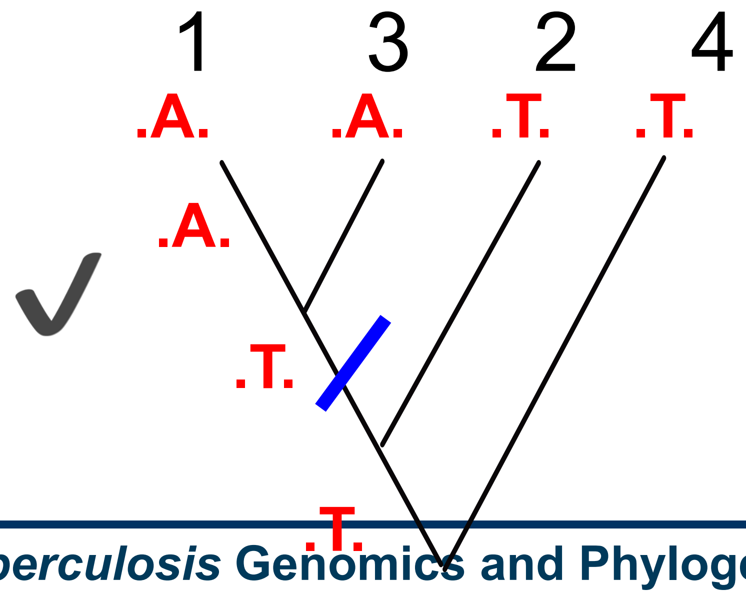
For positions 1 and 3, the most parsimonious tree is:





For position 2, the most parsimonious tree is:

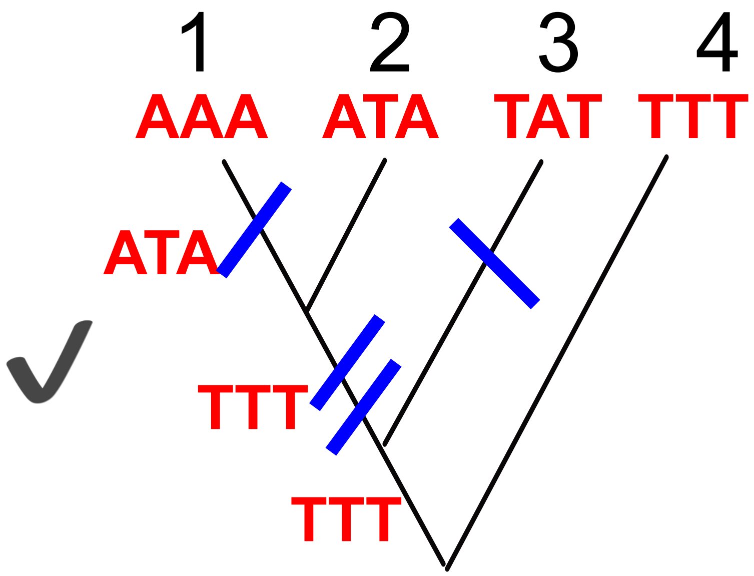




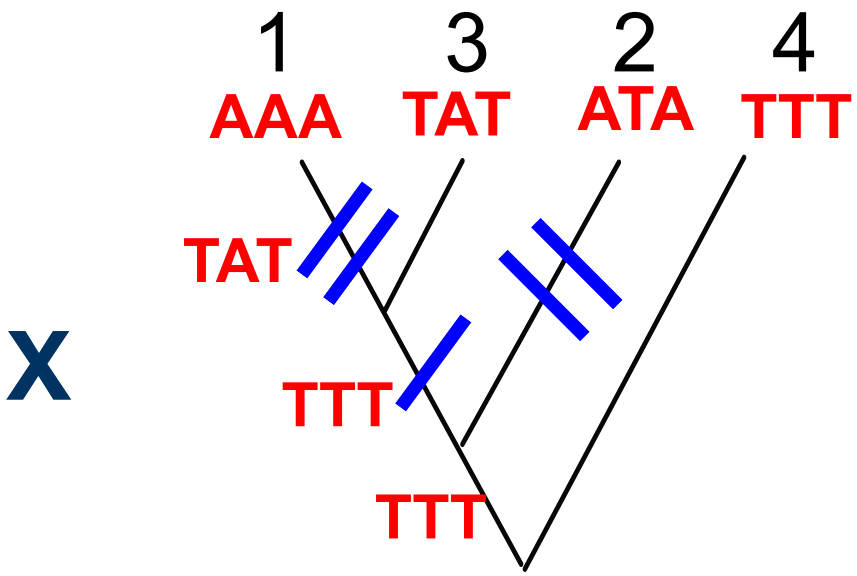
Tree Building: Maximum Parsimony

We had a conflict, how do we choose the simplest tree for all positions?

Choose the one with the least **homoplasies** (=shortest).



#mutations: 4



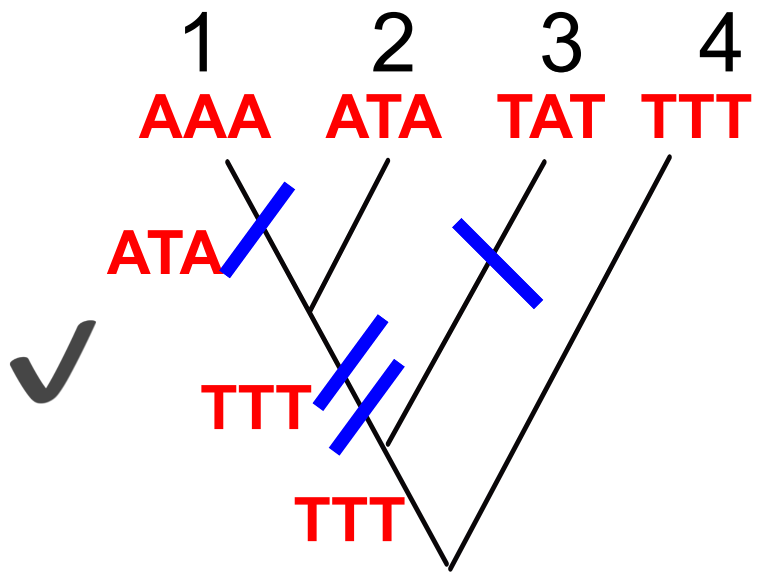
#mutations: 5

Tree Building: Maximum Parsimony

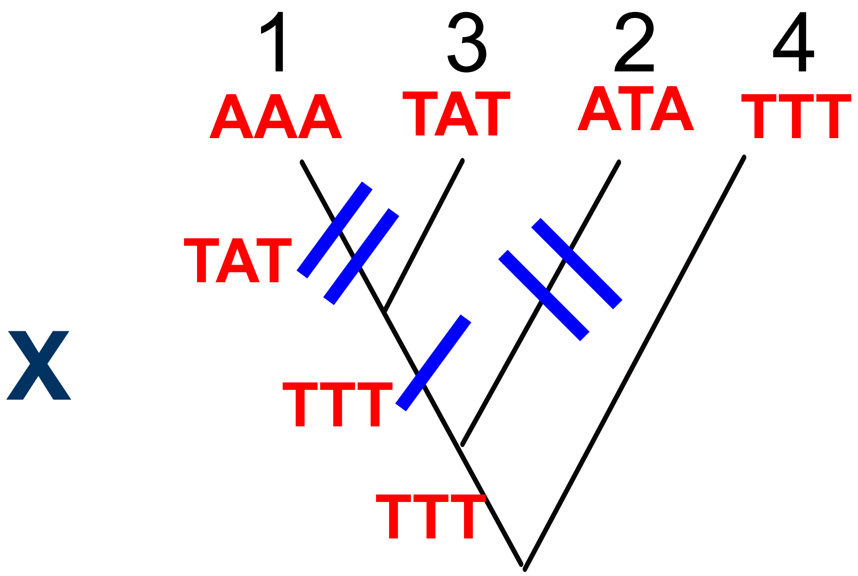
We had a conflict, how do we choose the simplest tree for all positions?

Choose the one with the least **homoplasies** (=shortest).

Homoplasy: When "look-alikes" are unrelated



#mutations: 4



#mutations: 5

Tree Building: Maximum Parsimony

- There might be multiple MP trees for a given alignment.

- "Branch-and-bound" algorithm can find the best tree without considering all the trees.

- When "branch-and-bound" is too slow, heuristic methods are used.

Character Based Tree Building Methods

Character Based Tree Building Methods

- Character-based: use sequences/MSA directly to infer a tree
- also called discrete methods
- Two methods you should know about:

1. Maximum parsimony (MP):

   minimize the number of evolutionary steps

2. **Maximum likelihood (ML):**

   mostly likely tree given the data and a probabilistic model of sequence evolution

Tree Building: Maximum Likelihood

- Find the tree that maximizes the probability of observing the data (sequences).

- Requires: MSA and a model of sequence evolution

Tree Building: Maximum Likelihood

- Requires searching an enormous number of trees.
- Computationally most intensive.
- ML tree dependent on the model of evolution used.
- RAxML:
  - Randomized Axelerated Maximum Likelihood
  - parallelized implementations exist
- FastTree: approximate Maximum Likelihood

Bootstrapping: How do you "defend" your tree?

- "sampling with replacement"
- Can be used with distance (UPGMA, NJ) and character (MP, ML) methods
- Bootstrap support > 70% typically considered strong support for a node.

Bayesian Trees

Bayes Theorem: Motivation

"Prosecutor's Fallacy": A medical diagnosis example

- You wake up one day, with spots on your face. You are worried you might have smallpox!
- Your doctor looks up into his medical book and reasons: "I looked it up and 90% of people with validated smallpox have spots on their face"
- Should you get worried?

adapted from Stone J, 2013

Bayes Theorem: Motivation

"Prosecutor's Fallacy": A medical diagnosis example

- You wake up one day, with spots on your face. You are worried you might have smallpox!
- Your doctor looks up into his medical book and reasons: "I looked it up and 90% of people with validated smallpox have spots on their face"
- Should you get worried?

Doctor says: $p(\text{spots}|\text{smallpox}) = 0.9$

adapted from Stone J, 2013

Bayes Theorem: Motivation

"Prosecutor's Fallacy": A medical diagnosis example

- You wake up one day, with spots on your face. You are worried you might have smallpox!
- Your doctor looks up into his medical book and reasons: "I looked it up and 90% of people with validated smallpox have spots on their face"
- Should you get worried?

Doctor says: $p(\text{spots}|\text{smallpox}) = 0.9$

You need: $p(\text{smallpox}|\text{spots}) = ?$

adapted from Stone J, 2013

Bayes Theorem: Motivation

"Prosecutor's Fallacy": A medical diagnosis example

- You wake up one day, with spots on your face. You are worried you might have smallpox!
- Your doctor looks up into his medical book and reasons: "I looked it up and 90% of people with validated smallpox have spots on their face"
- Should you get worried?

Doctor says: $p(\text{spots}|\text{smallpox}) = 0.9$

You need: $p(\text{smallpox}|\text{spots}) = ?$

$p(\text{smallpox}) = ?$ 10^{-3}

$p(\text{spots}) = ?$ 10^{-1}

$$p(\text{smallpox}|\text{spots}) = \frac{0.9 \text{ X } 10^{-3}}{10^{-1}} = 0.009$$

adapted from Stone J, 2013

Bayesian Trees

Bayes Theorem

"likelihood"

$$p(\text{hypothesis}|\text{evidence}) = \frac{p(\text{evidence}|\text{hypothesis}) \ \text{X} \ p(\text{hypothesis})}{p(\text{evidence})}$$

"posterior"

"prior"

"counted OR estimated"

"computed"

$$p(\text{smallpox}|\text{spots}) = \frac{p(\text{spots}|\text{smallpox}) \ \text{X} \ p\ (\text{smallpox})}{p\ (\text{spots})}$$

"what you want to know"

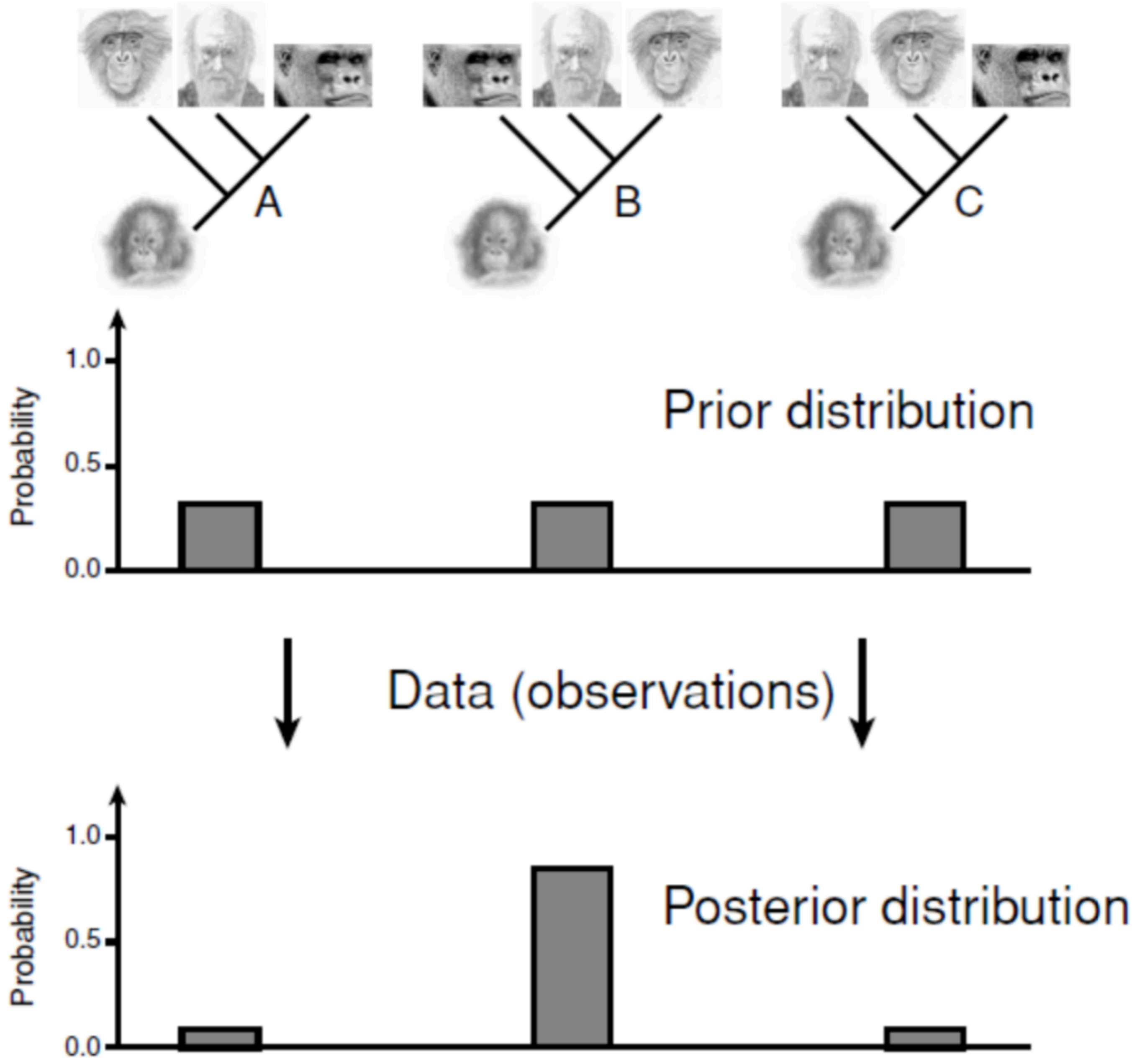
"assumed, often intelligibly"

Bayesian Trees

To build a bayesian tree, you need:

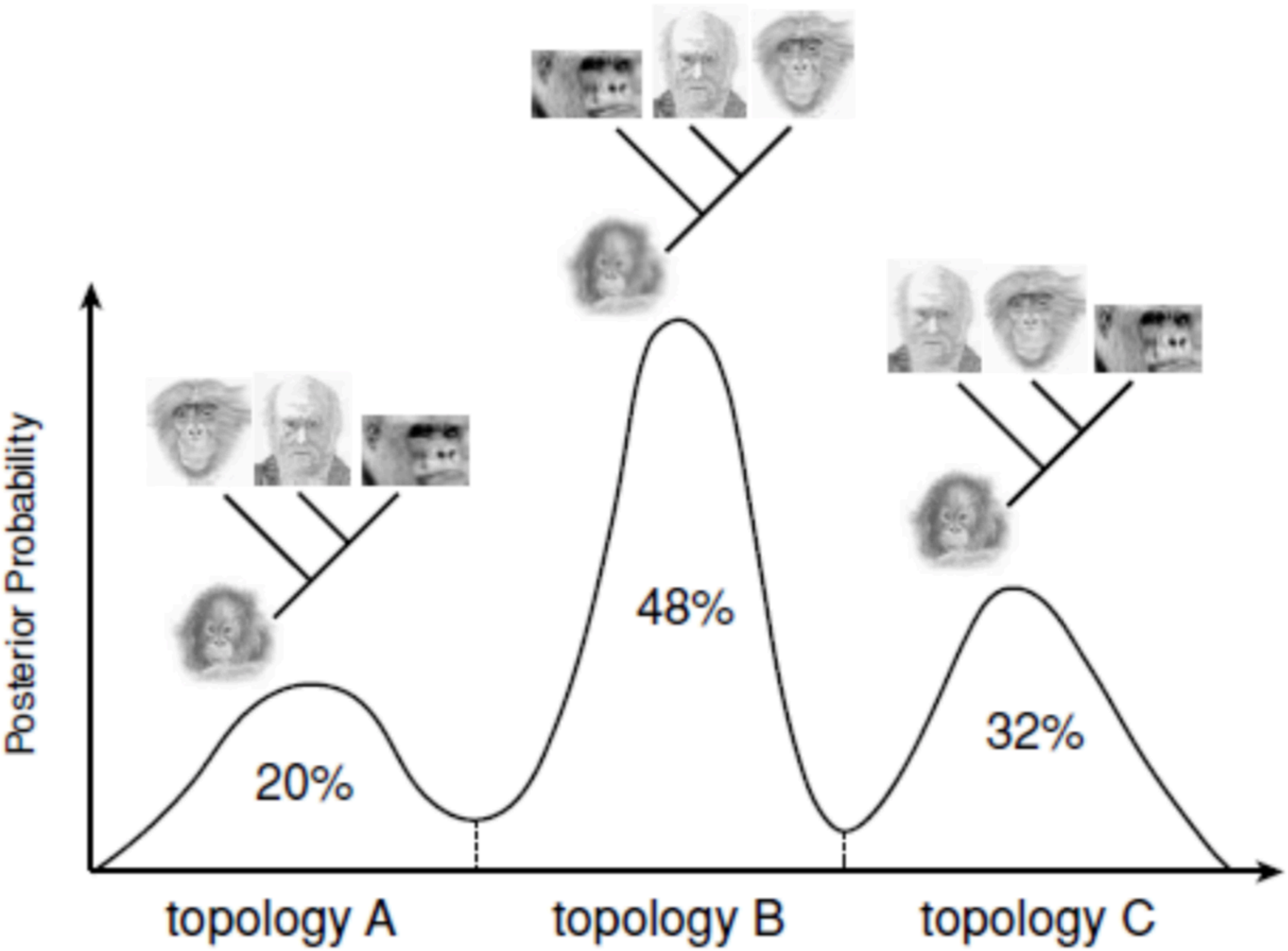
- phylogenetic tree (T)
- data (say an Multiple Sequence Alignment, X)
- Tree prior options, for instance:
    - epidemiology
    - "ignorance"

How do you build a Bayesian Tree?



The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, Cambridge University Press 2009.

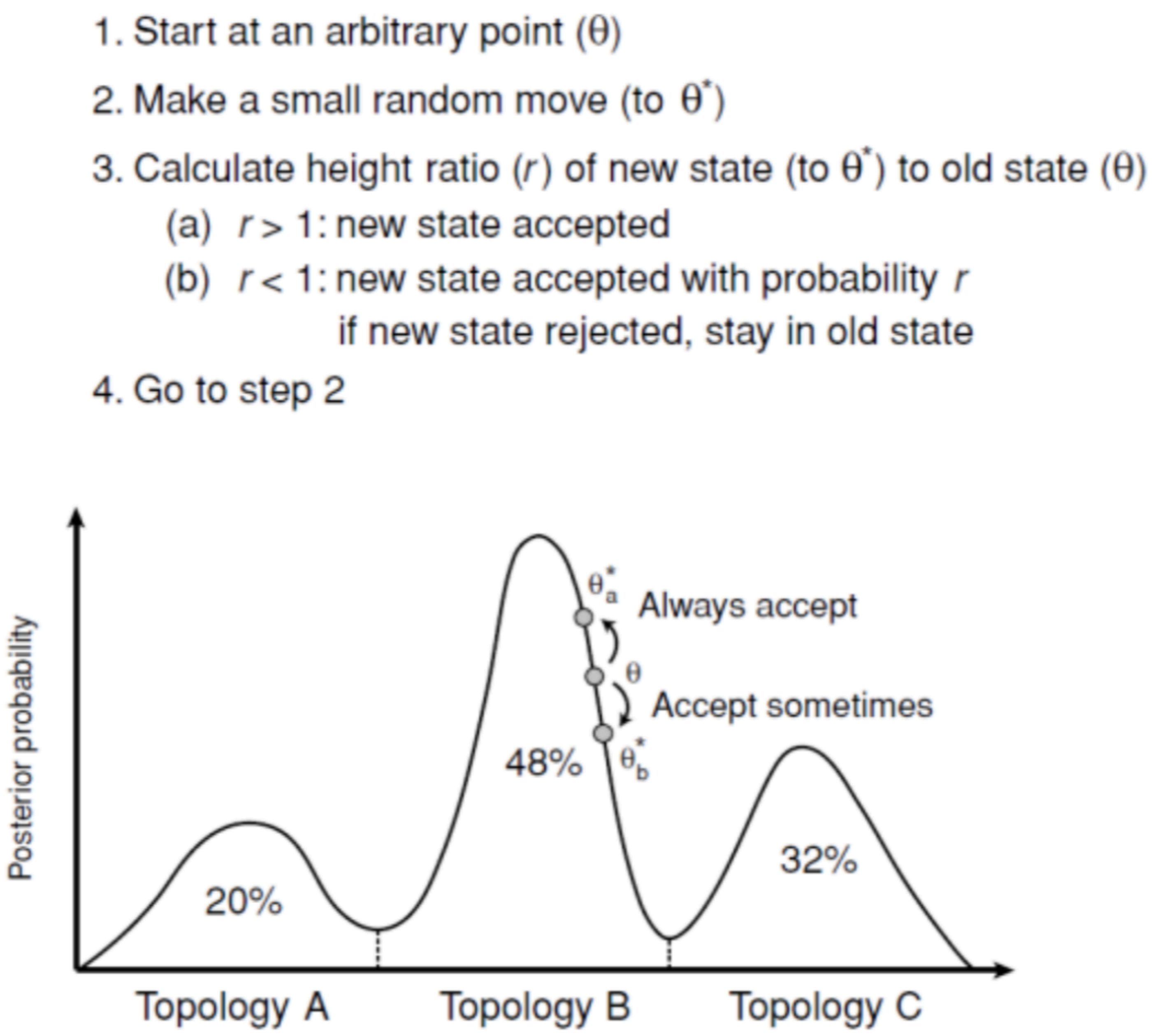
How do you build a Bayesian Tree?



The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, Cambridge University Press 2009.

How do you build a Bayesian Tree?

Markov Chain Monte Carlo (MCMC) Sampling



1. Start at an arbitrary point (θ)
2. Make a small random move (to θ^*)
3. Calculate height ratio (r) of new state (to θ^*) to old state (θ)
   - (a) $r > 1$: new state accepted
   - (b) $r < 1$: new state accepted with probability r
     if new state rejected, stay in old state
4. Go to step 2

The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, Cambridge University Press 2009.

Phylogenomics

"Whole genome phylogenies"

Whole Genome Phylogenies

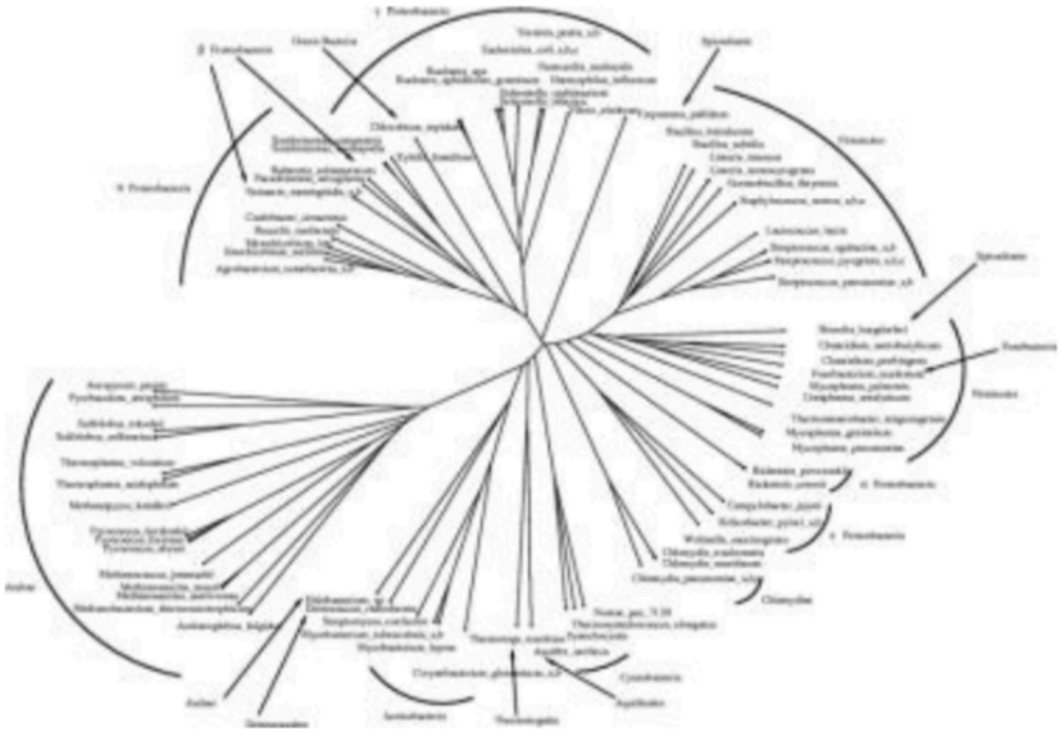
Advantages:

- ultimate in resolution
- robust

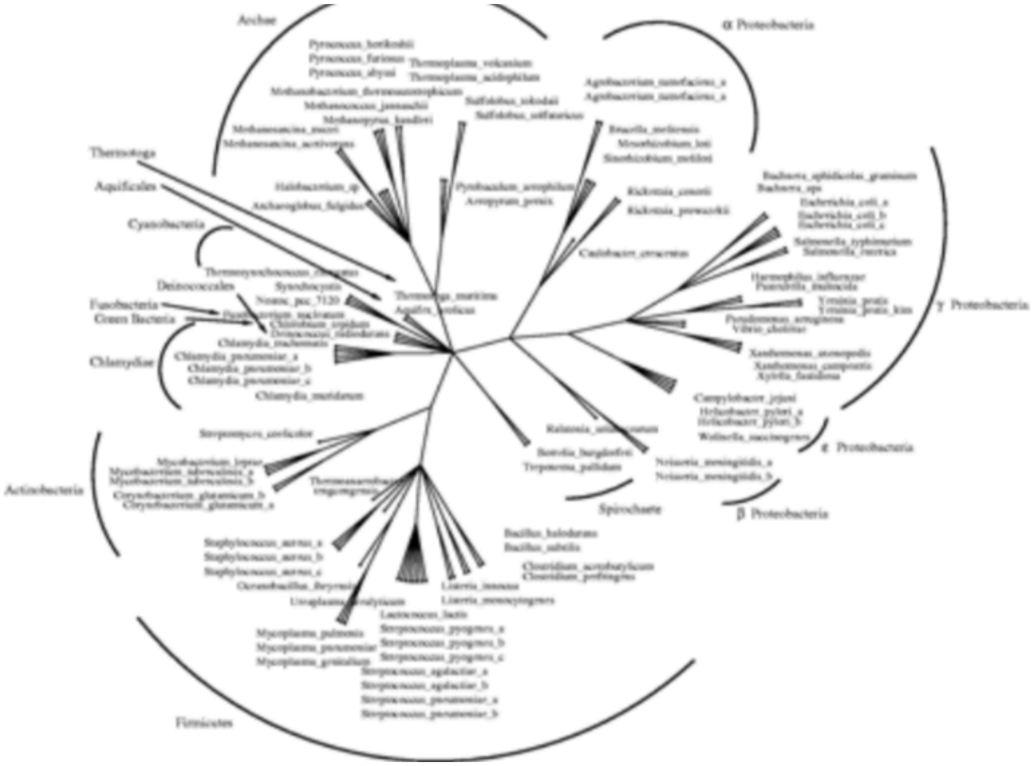
Confounding factors (not unique whole genome phylogenies):

- horizontal gene transfer
- recombination

"Phylogenetic" Tree



"Phylogenomic" Tree



Henz S R et al. Bioinformatics 2005;21:2329-2335

Whole Genome Phylogenies

How to build it?

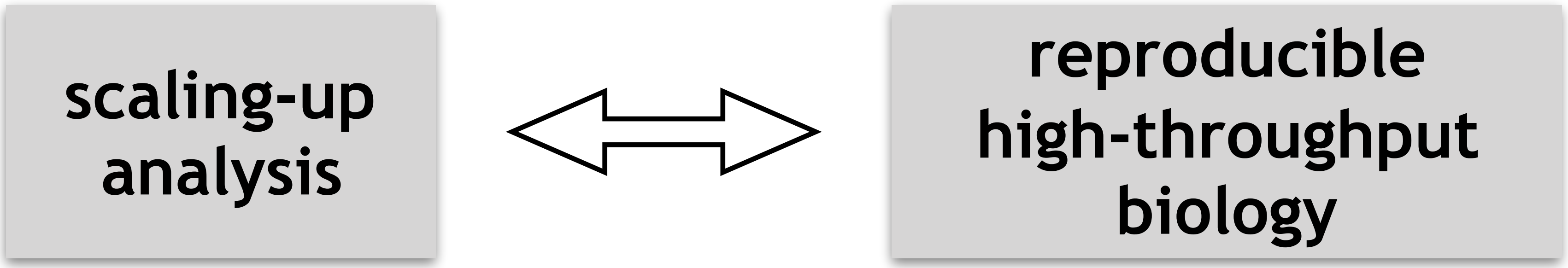
- You just started yesterday!

Tips & Tricks

- Reduce your chance of including "bad" SNPs (i.e. paralogs from genomes not reflected in the reference)
- Homoplastic SNPs (e.g. from recombination): will confound phylogenetic inference
- Choice of reference is very important and can be difficult

Reproducibility and Analysis Scale up

Reproducibility and Analysis Scale up



scaling-up analysis

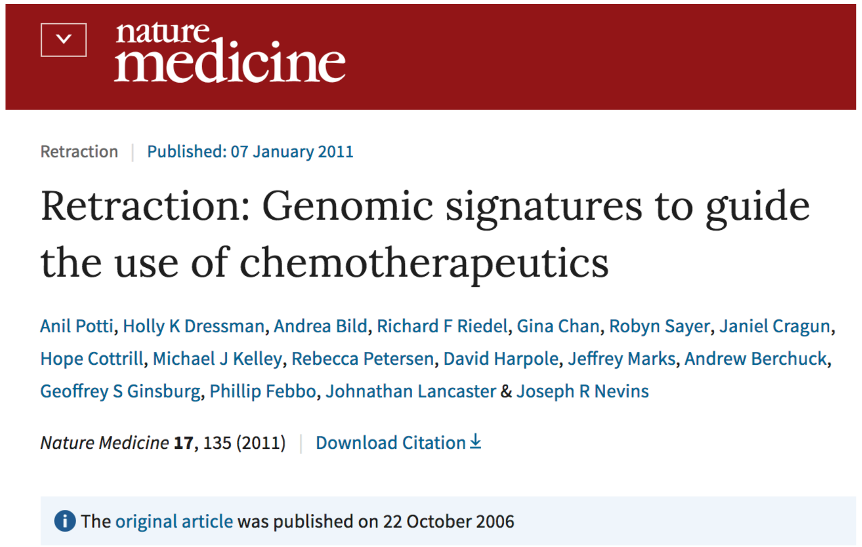
reproducible high-throughput biology

We have a problem of reproducibility in high-throughput biology.

- technically, easy to solve
- in reality, quite hard to address: people's opinions/habits/resistance to change...
- cultural differences btw. computational and wet-lab/clinical scientists
- conflict of interest regarding journals and scientific publication process
- review process is inadequate
- curriculum & training
- will likely take a generational change to completely iron it out

Reproducibility and Analysis Scale up

"*Forensic Bioinformatics*": not fun and waste of humanity's resources!



nature medicine

Retraction | Published: 07 January 2011

Retraction: Genomic signatures to guide the use of chemotherapeutics

Anil Potti, Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S Ginsburg, Phillip Febbo, Johnathan Lancaster & Joseph R Nevins

Nature Medicine **17**, 135 (2011) | Download Citation

The original article was published on 22 October 2006

- despite non-disputable computational evidence involving:
  - "mislabeling",
  - "training/testing set" issues due to blind use of black box software
  - "problems in gene names (for instance, off by one errors),

  the journal and people responsible for running the clinical trials based upon the study were literally "a wall" for years.
- took 5 years to convince them and...
- not because of irreproducibility (which was obvious from day 1) but scientific misconduct that became apparent later on
- the irreproducibility related issues were very easy to detect

bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified/IOM/duke_historical_perspective_3_29_11.pdf

Regarding Drs. Baggerly and Coombes suggestion that the pemetrexed sensitivity signature was reversed, it was believed that their conclusion was based on the use of a published heat map for the pemetrexed signature and then working back to identify cell lines that could produce that heat map. Then, when they examined these cell lines they found that **there was a reversal of the labels for sensitive and resistant**. This possible reversal of sensitivity/resistant labels could have been the result of a reversal of the training set as suggested by Drs. Baggerly and Coombes, which would have indeed negated the predictor.

...

that the pemetrexed signature was reversed, and that many, **if not all, of a collection of ovarian cancer samples were incorrectly labeled.** This Duke web page had been used for the sharing of data by the investigators involved in developing the new manuscript, as well as preparing for the review.

...

In contrast to the two approaches, when **principal components are built solely from the training set data** (cell lines) and no methods are used to normalize the gene expression profiles with the tumor samples, the predicted probabilities are poorly distributed such that the model has no capacity to discriminate the responders from non-responders in tumor samples (described in a manuscript submitted to Clinical Cancer Research

...

...In evaluating these errors, it is clear that there were inadequate processes and data systems being used to assure data provenance and the ability of others to replicate the studies.