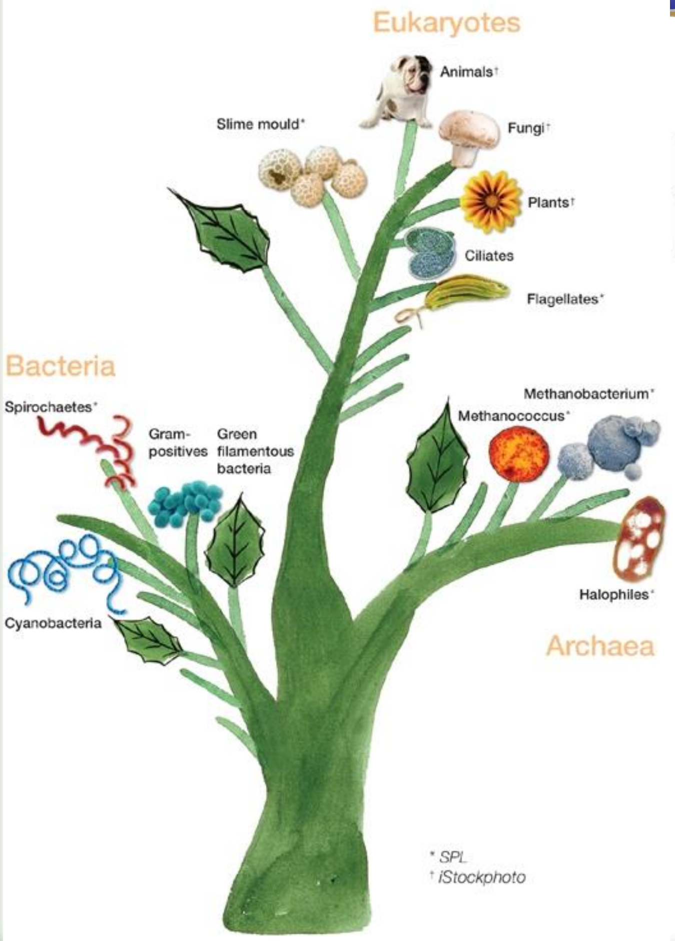
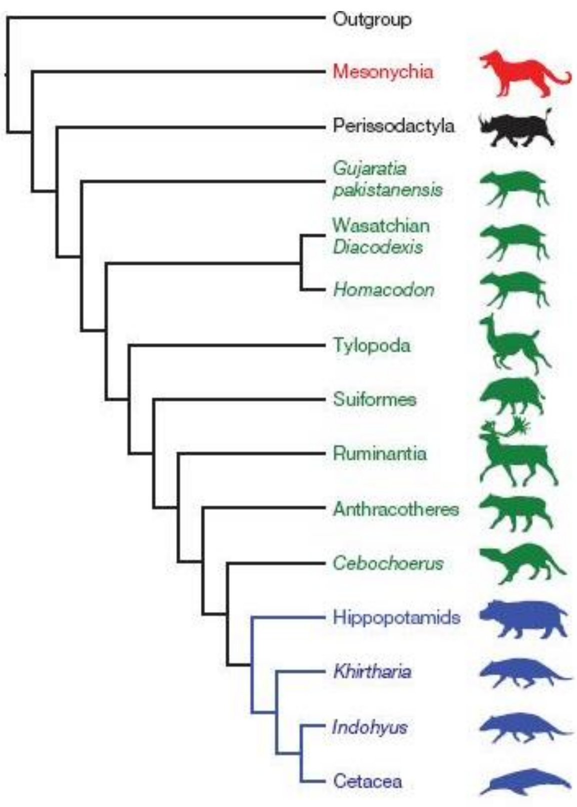
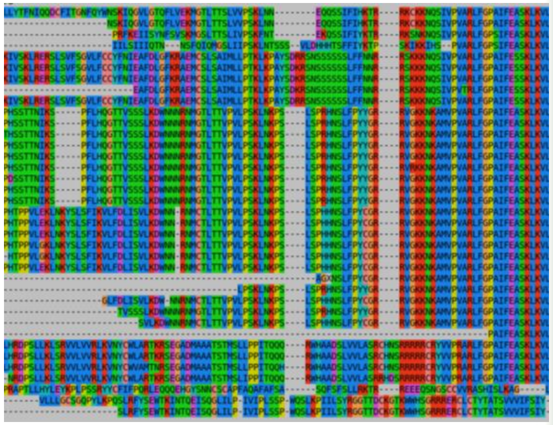
Evolution and Phylogenetic analysis







Kiatichai Faksri

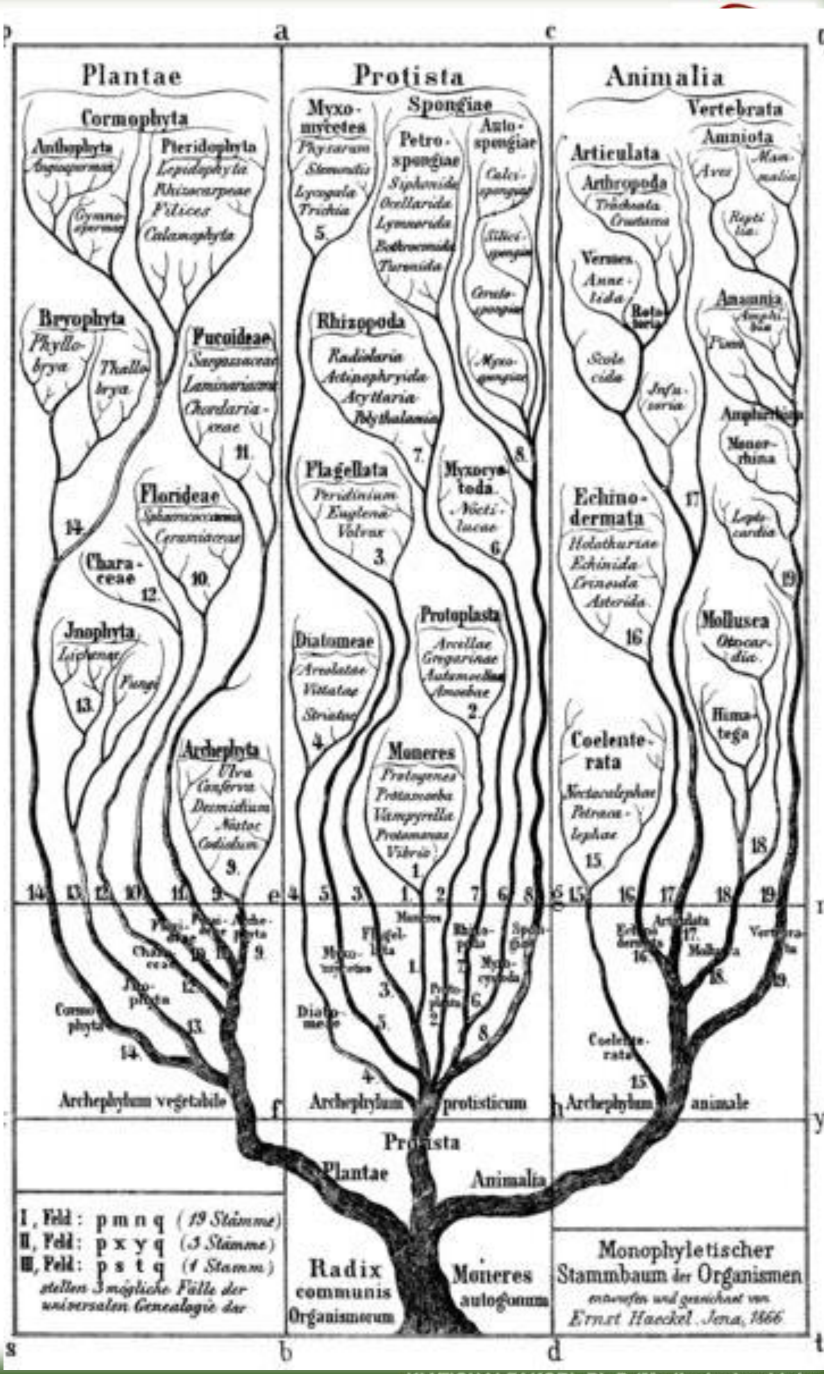
Ph.D, Medical Microbiology
Faculty of Medicine, KKU

Objectives

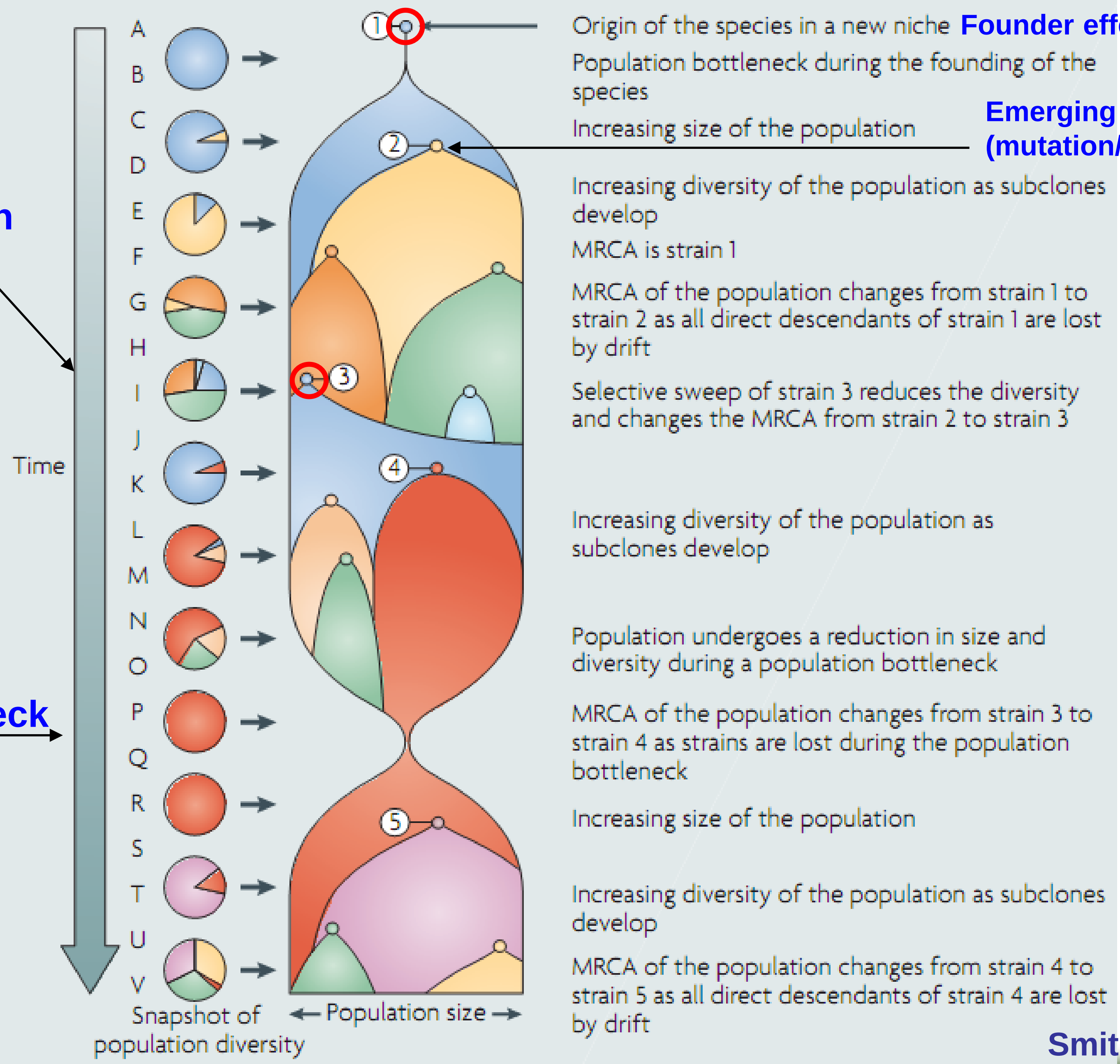
- 1. Evolution and introduction to phylogenetic analysis
- 2. Methods in phylogenetic analysis

Phylogenetic tree

- Phylogenetic tree based on coancestral relationships
- It regards homology as evidence of common ancestry
- Distance between taxa reflects a decreasing number of homologous characters
- Constructed phylogenetic tree is not necessarily the same as actual evolutionary relationship



Box 3 | The most recent common ancestor



Origin of the species in a new niche

Population bottleneck during the founding of the species

Increasing size of the population

Increasing diversity of the population as subclones develop

MRCA is strain 1

MRCA of the population changes from strain 1 to strain 2 as all direct descendants of strain 1 are lost by drift

Selective sweep of strain 3 reduces the diversity and changes the MRCA from strain 2 to strain 3

Increasing diversity of the population as subclones develop

Population undergoes a reduction in size and diversity during a population bottleneck

MRCA of the population changes from strain 3 to strain 4 as strains are lost during the population bottleneck

Increasing size of the population

Increasing diversity of the population as subclones develop

MRCA of the population changes from strain 4 to strain 5 as all direct descendants of strain 4 are lost by drift

Founder effect

Emerging of new allele (mutation/ gene flow)

Selection

Bottle neck

Smith NH. et al., 2009

Terminology

- **Clades**: share a common ancestor that belongs to their own group
- **Monophyletic groups** (clades): contain taxa (taxonomic gr.) that are more closely related to each other than to any outside the group
- **Dendogram** = tree diagram that illustrate the arrangement of the clusters (cluster analysis) produced by hierarchical clustering (based on similarity)
- **Speciation =** new species that capable of making a living in a new way from the species which it arose
- **Homolog**: gene related to a second gene by descent from a common ancestral DNA sequence
  - **Ortholog** = genes in different species, evolved from a common ancestral gene by speciation, retain the same function
  - **Paralogs** = genes related by duplication within a genome, evolve new functions (within species)

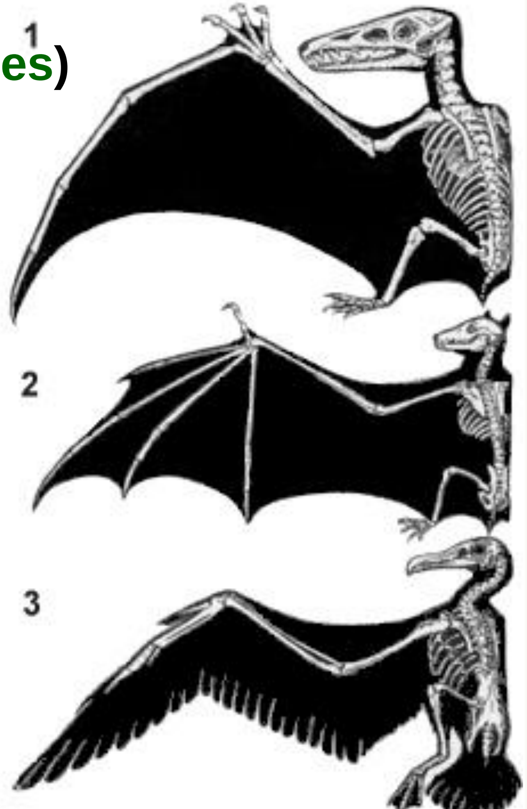
Why phylogenetic analysis?

- Determining the closest relatives of the organism and diversity
  - Novel organism (species)
  - Cluster analysis: outbreak of ID, genetic diseases
  - Map pathogen strain diversity for vaccines
  - Biodiversity studies
  - Understanding microbial ecologies
- Discovering the function of a gene
  - Orthologous/ paralogous genes
- Retracting the origin of a gene or organism
  - Understand evolutionary history
  - Recent common ancestor

Divergence versus Convergence evolution

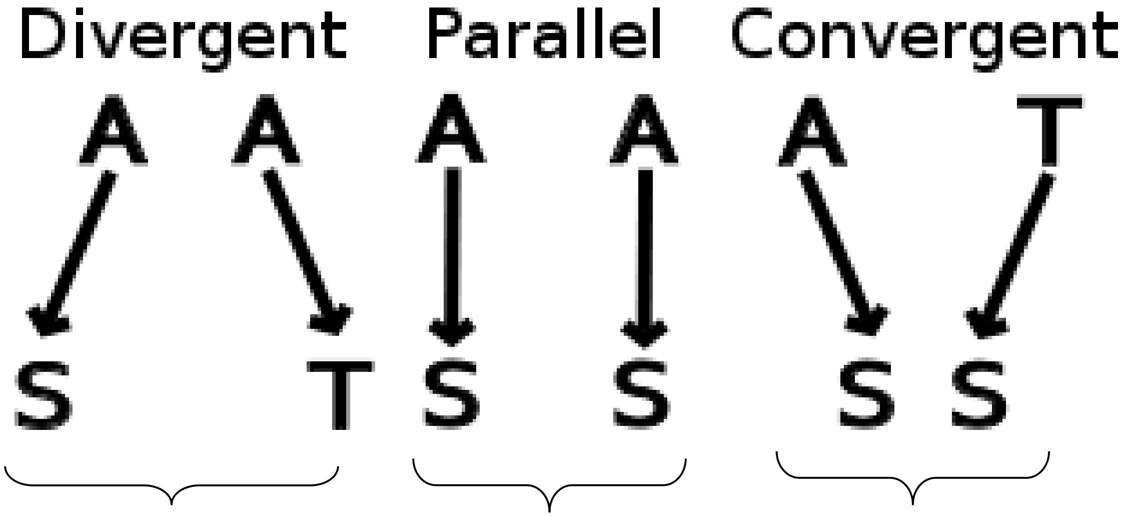
- Divergence evolution
  - Accumulation of differences (mutations) *between groups*, can lead to the formation of new species
  - Same species adapting to different pressure
  - e.g. organisms with 5 digit pentadactyle limbs
    : humans, bats, and whales (homologous structures)
- Convergence evolution
  - the acquisition of the same biological trait
    in unrelated lineages
  - Differnt species adapting to the same pressure
  - Analogous structures

Independently evolved genera of succulent plants



“fossil”

Divergent vs. Parallel vs. Convergent evolution



Same ancestor
Different descendant

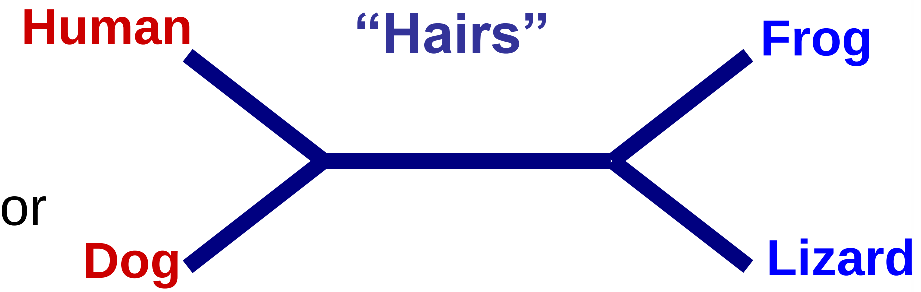
Same ancestor
Same descendant

Different ancestor
Same descendant

Homology versus Homoplasy

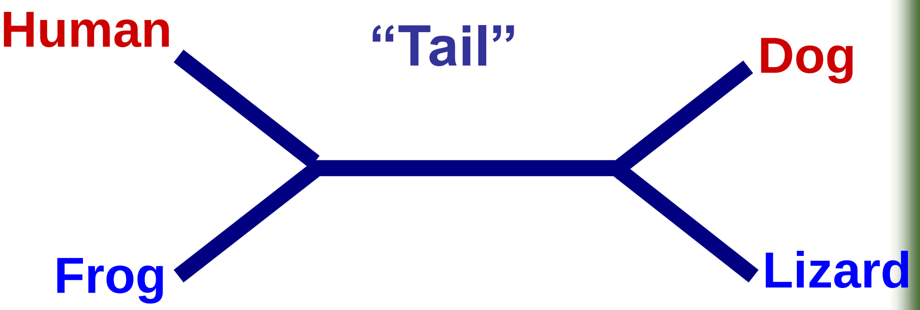
Homology =

- Similarity derived from a common ancestor
- Homologous characters = useful for phylogenetic tree construction



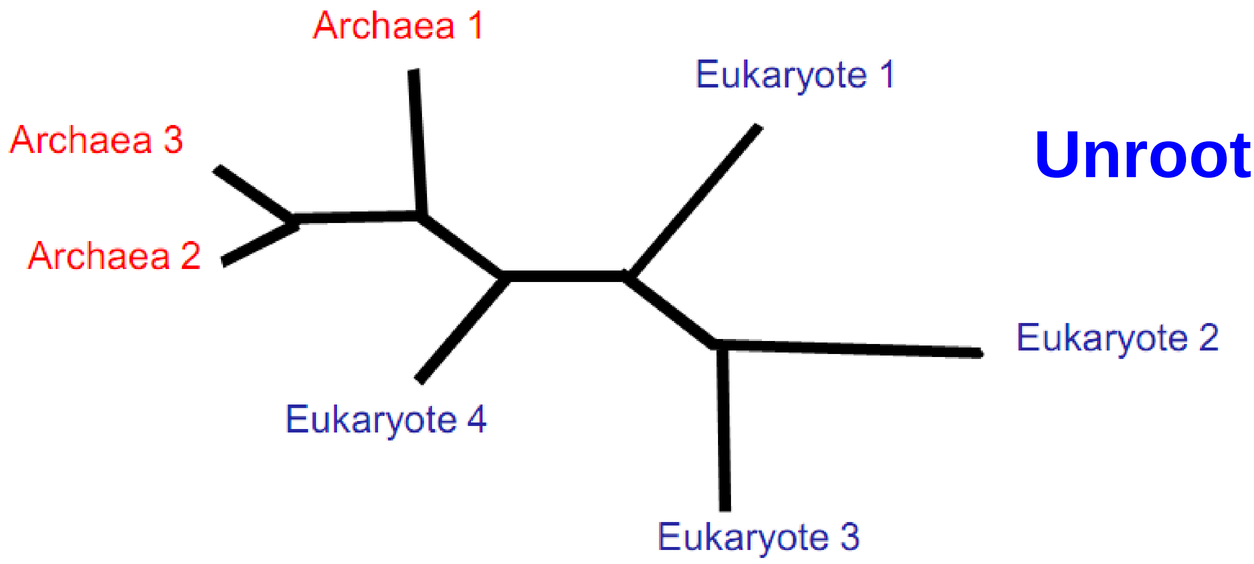
Homoplasy =

- Similarity due to independent acquisitions of the same or superficially similar characteristics
- Homoplasic characters = misleading picture of phylogeny



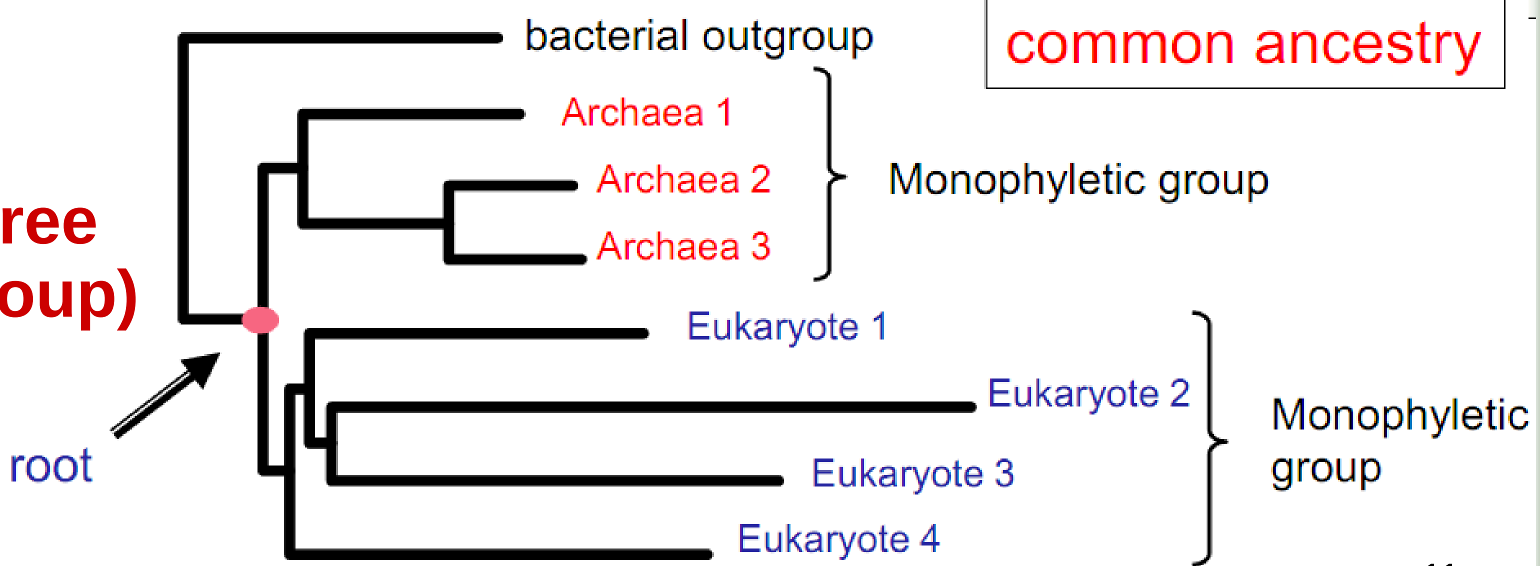
Simon Harris, Wellcome Trust Sanger Institute

Rooted versus unrooted tree



Unrooted tree

The root defines common ancestry



Rooted tree (by outgroup)

Simon Harris, Wellcome Trust Sanger Institute

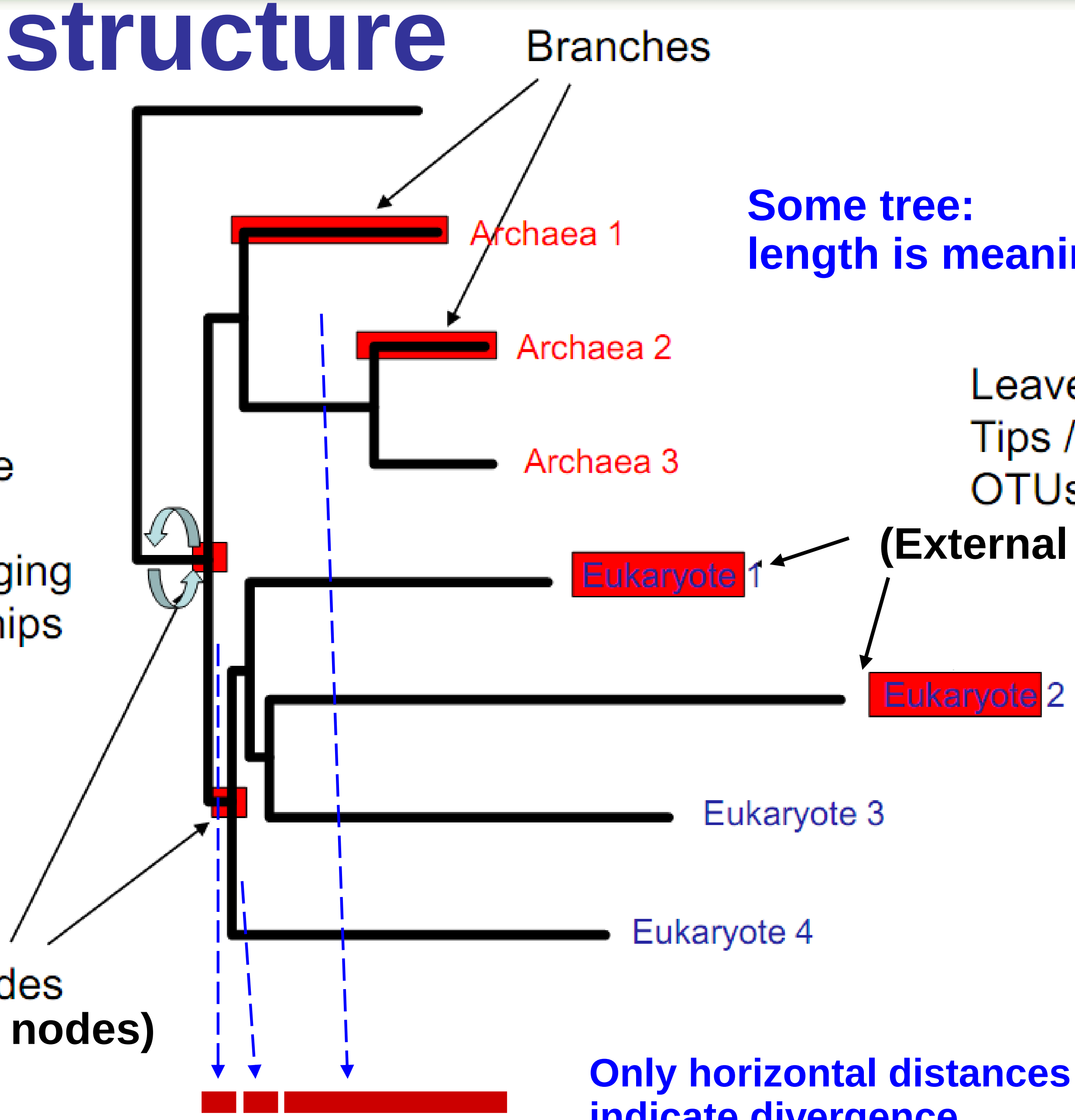
Faculty of Medicine, KKU

Number of OTUs (Taxa) and Number of Tree

Number of OTUs	Number of unrooted trees	Number of rooted trees
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	954	10,395
8	10,395	135,135
9	135,135	34,459,425
10	34,459,425	2.13E+15
15	2.13E+15	8.00E+21

So, if apply for long sequence data set,
it may take a very long time for analysis

Tree structure



Some tree:
length is meaningless

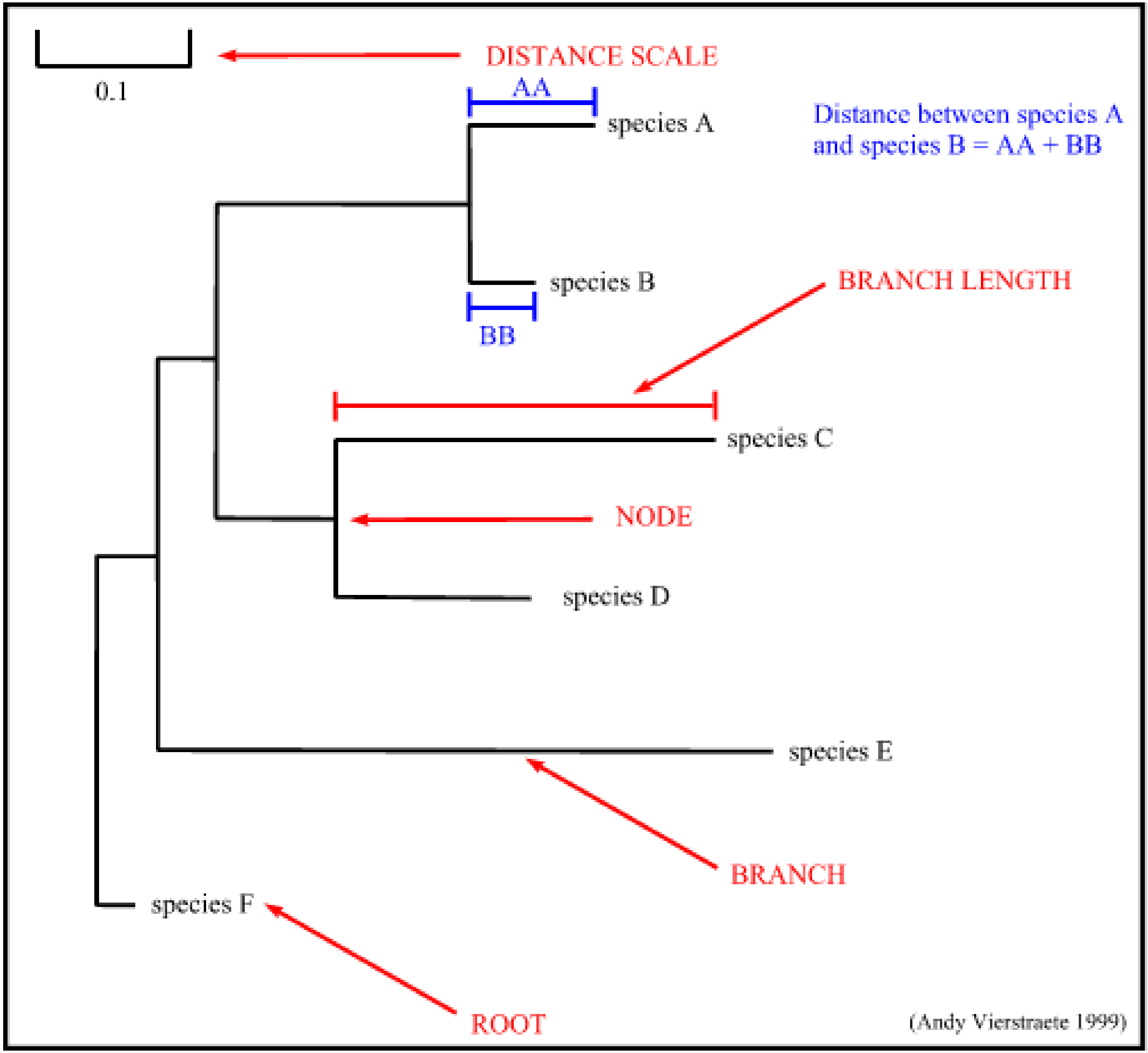
Leaves / Tips / OTUs
(External nodes)

Nodes can be freely rotated without changing the relationships shown

Nodes
(Internal nodes)

Only horizontal distances indicate divergence

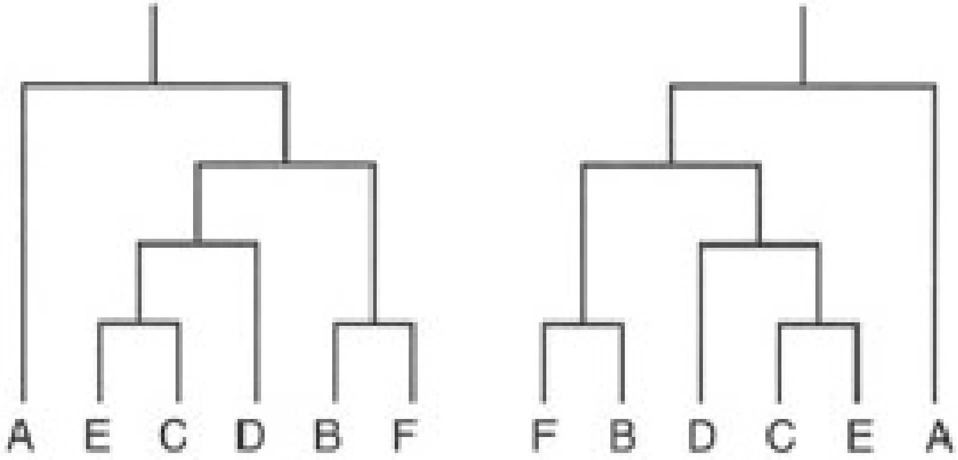
Simon Harris, Wellcome Trust Sanger Institute



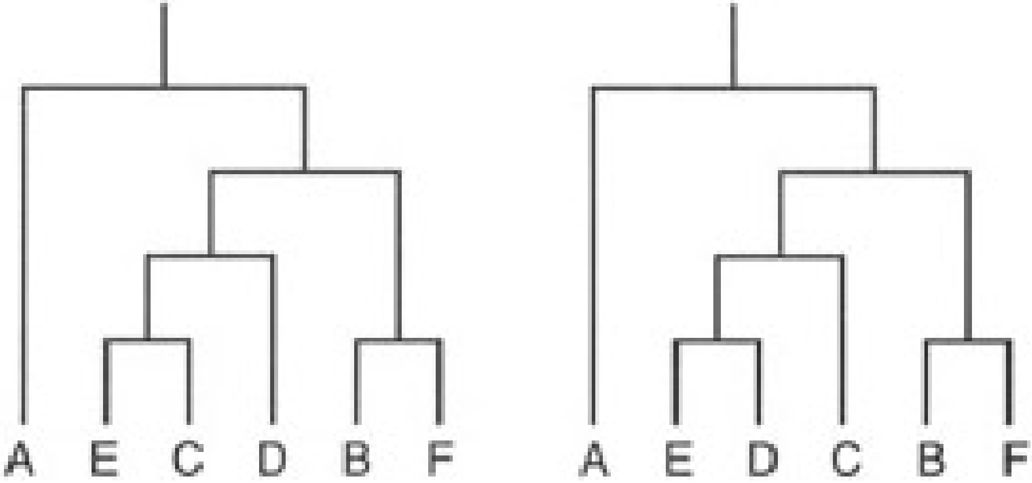
(Andy Vierstraete 1999)

Tree topology

(a) Is the following pair of trees identical in topology?



(b) Is the following pair of trees identical in topology?



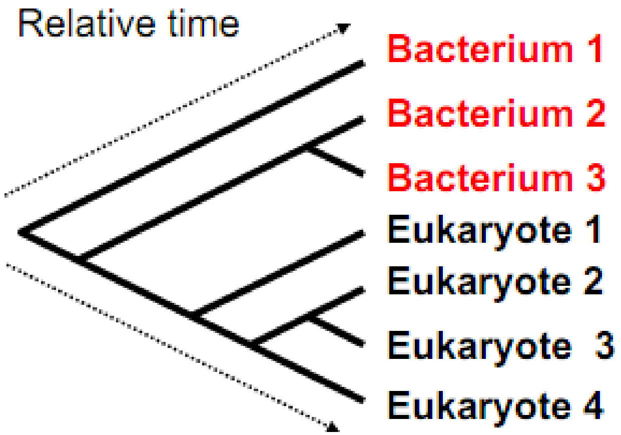
So……..

- Root: origin of evolution
- Leaves: current organisms, spp., groups
- Branches: relationship between organisms, spp.
- Branch length: evolutionary time

Cladograms versus Phylograms

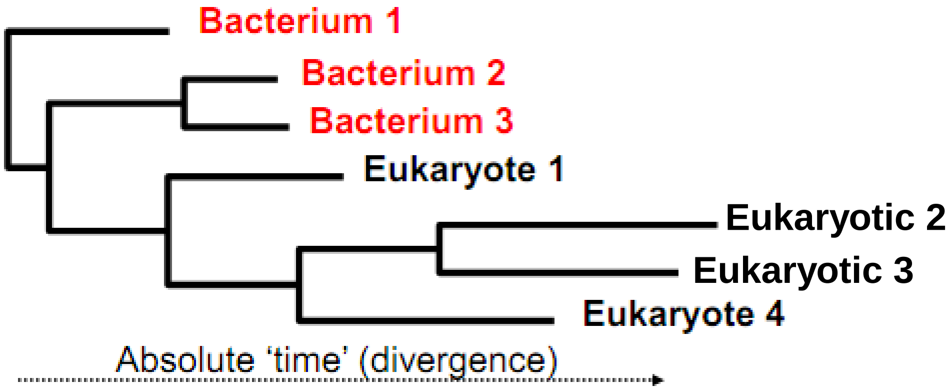
Cladogrames =

- branch lengths are meaningless



Phylograms =

- Provide branch lengths



Simon Harris, Wellcome Trust Sanger Institute

Bootstrapping

- The same dataset >> many tree shapes (relationship): What is the most correct one?

- Statistical method to measures the accuracy of sampling distribution

  = Characters (sites/ sequences) random resampling with replacement methods (e.g. 1000 times (replica))

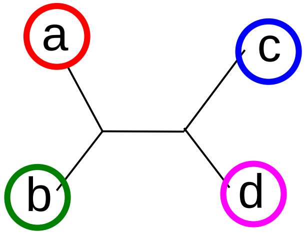
- Frequency of occurrence of groups in the results support the accuracy of that groups
- Showing how often that relationships occurred in the replicate analyses

- Assess quality or reliability of a reconstructed tree

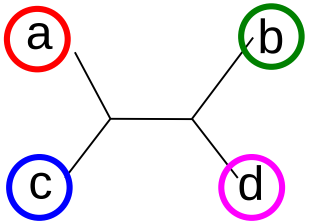
Bootstrapping

	1234567
a	ATATAAA
b	ATTATAA
c	TAAAATA
d	TATAAAT

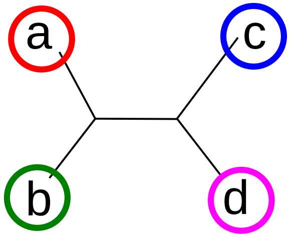
	1224567
a	ATTTAAA
b	ATTATAA
c	TAAAATA
d	TAAAAAT



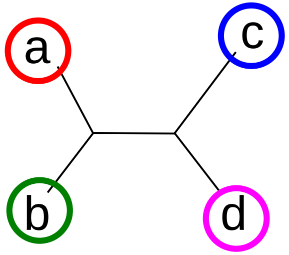
	1334567
a	AAATAAA
b	ATTATAA
c	TAAAATA
d	TTTAAAT

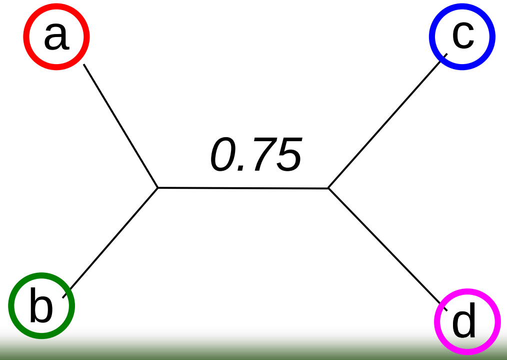


	1234567
a	ATATAAA
b	ATTATAA
c	TAAAATA
d	TATAAAT

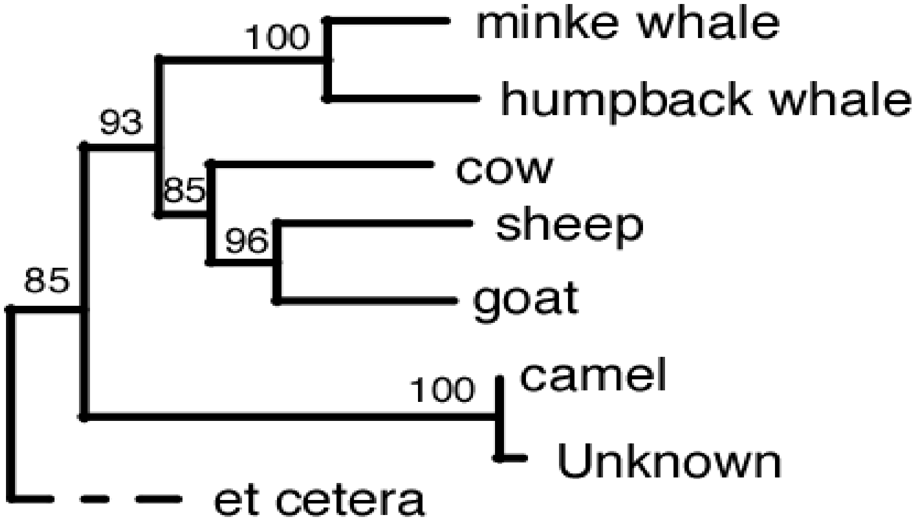


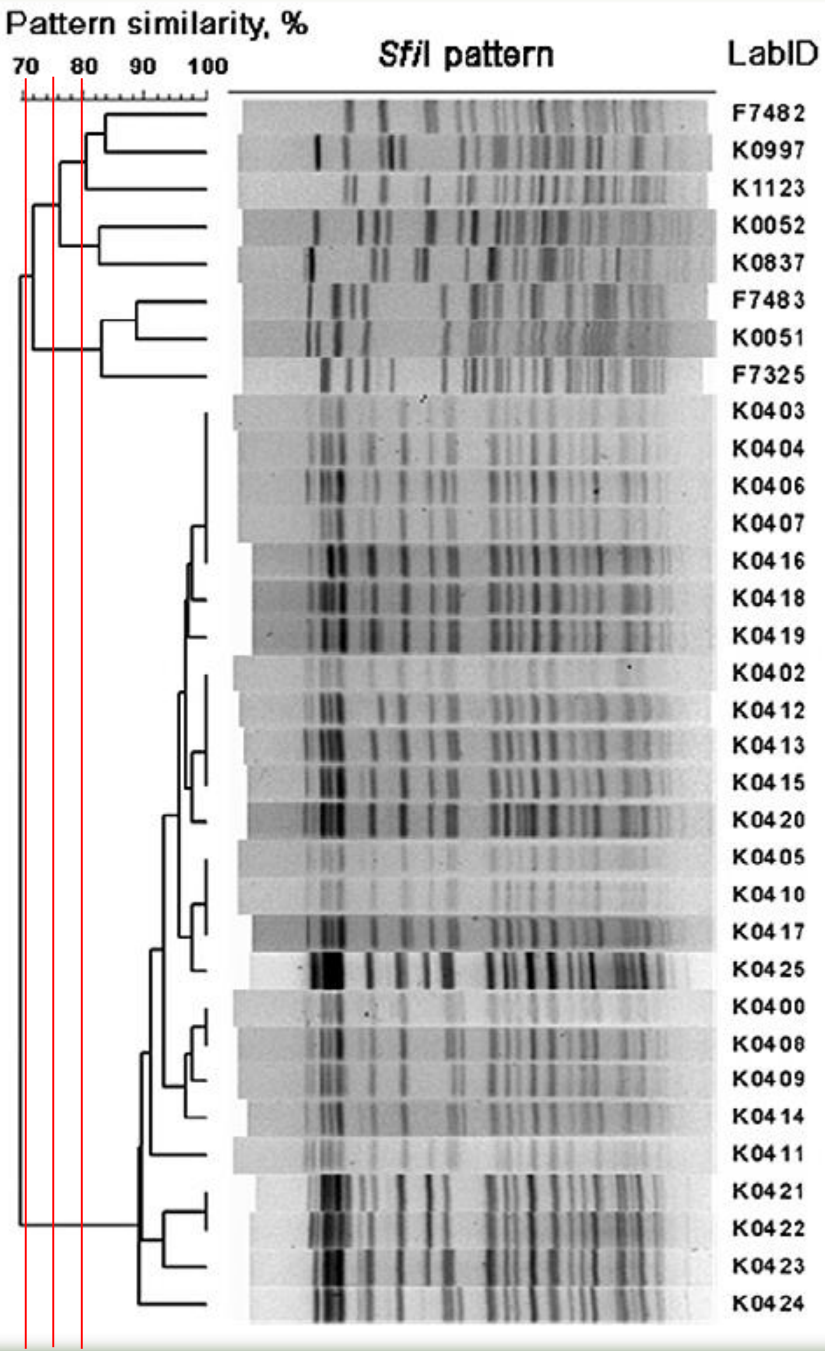
	1244567
a	ATTTAAA
b	ATAATAA
c	TAAAATA
d	TAAAAAT





- The number of bootstrap show the confidence of the tree topology (max = 100)
- In general: bootstrap value in particular branch > 95 = correct





Dendogram (RFLP)

“% similarity”

How many group at 80% similarity?

Methods for phylogenetic tree analysis

4 main (statistical) methods

1. Distance (NJ and UPGMA)
2. Parsimony
3. Maximum likelihood
4. Bayesian methods (not in detail!)

[Tree merging method e.g. consensus tree]

Different method may provide different tree
Which one I should select for my data?

1. Distance based method

- Construct trees by evolutionary distances
- Minimum Evolution = best tree is the shortest length

- Concept
  - Pairwise distances between taxa are calculated
  - Tree topology & branch lengths from distance matrix
  - not accurate but good for continuous data/ large data

- Most common methods
  - Neighbor Joining
  - UPGMA

% similarity

Seq A >>> AGAUUCGUCUGUAGGUUUCCAC C AA

Seq B >>> ACAUUCG UGUAUAGGUUU CCACU AA

Seq A: AGAUUCGUCUGUAGGUUUCCACCAA

| X | | | | | | | X | X | | | | | | | | | | | | | X | |

Seq B: ACAUUCGUGUAUAGGUUUCCACUAA

0100000010100000000000100

No. of different character = 4

the similarity between Seq A and Seq B
= 21/25 = **0.84**

There are **options** for calculation similarity

“Score of transversion > transition”

Seq A: AGAUUCGUCUGUAGGUUUCCACCAA

| X| | | | | | X|X| | | | | | | | | | | X| |

Seq B: ACAUUCGUGUAUAGGUUUCCACUAA

0200000020100000000000100

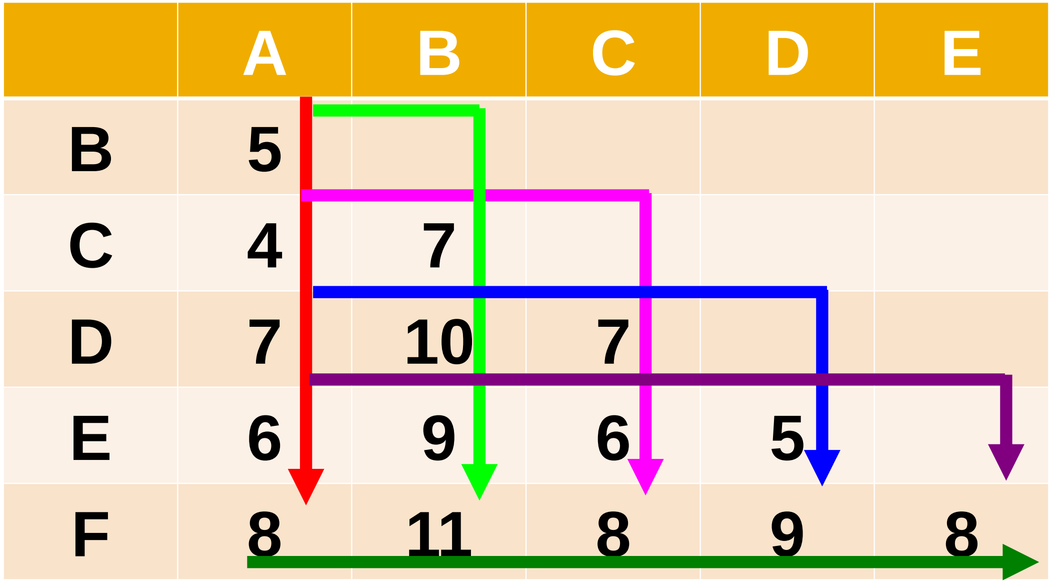
the similarity between Seq A and Seq B

= 19/25 = **0.76**

1.1 Neighbour joining

1. Calculate the distance for each taxon to others
2. Join the two nearest neighbours into a new node
3. Compute branch lengths from these two taxa to the new node
4. Compute the distance between the new node and all other taxa
5. Delete the joined taxa from the distance matrix and add the new node
6. Repeat until only 2 taxa remain, then join them

Example of Neighbor-joining



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

Matrix 1

Step 1: calculation : Sx = (sum all Dx) / (leaves - 2)

- **S(A) = (5 + 4 + 7 + 6 + 8) / 4 = 7.5**
- **S(B) = (5 + 7 + 10 + 9 + 11) / 4 = 10.5**
- **S(C) = (4 + 7 + 7 + 6 + 8) / 4 = 8**
- **S(D) = (7+ 10 + 7 + 5 + 9) / 4 = 9.5**
- **S(E) = (6 + 9 + 6 + 5 + 8) / 4 = 8.5**
- **S(F) = (8 + 11 + 8 + 9 + 8) / 4 = 11**

Step 2: Calculate pair with smallest M

Mij = Distance ij – Si – Sj

- Smallest are
  - M(AB) = d(AB) – S(A) –S(B) = 5 – 7.5 – 10.5= -13
  - M(DE) = 5 – 9.5 – 8.5 = -13

	A	B	C	D	E
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

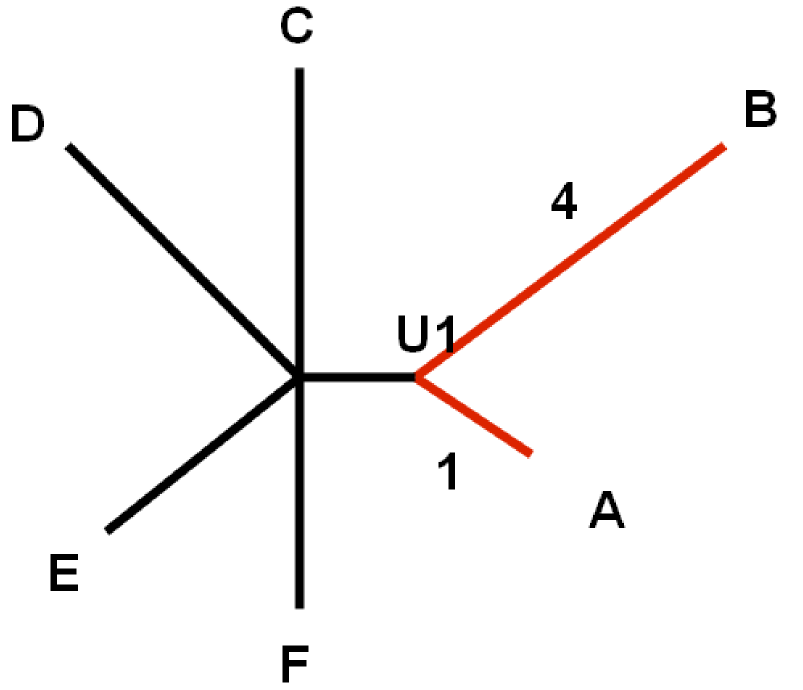
Matrix 2

Step 3: Create a node U

S1U = (Dij / 2) + (Si – Sj) / 2

- U1 joins A and B:
  - S(AU1) = d(AB) / 2 + (S(A) – S(B)) / 2
    = 5 / 2 + (7.5 - 10.5) / 2 = **1**
  - S(BU1) = d(AB) / 2 + (S(B) – S(A)) / 2
    = 5 / 2 + (10.5 – 7.5) / 2 = **4**

Step 4: Join A and B according to S, and make all other taxa in form of a star. Branches in black are unknown length and Branches in red are known length



Step5: Calculate new distance matrix

Dxu = (Dix + Djx – Dij) / 2

- d(CU) = (d(AC) + d(BC) - d(AB)) / 2
  = (4 + 7 - 5) / 2 = **3**
- d(DU) = d(AD) + d(BD) - d(AB) / 2 = **6**

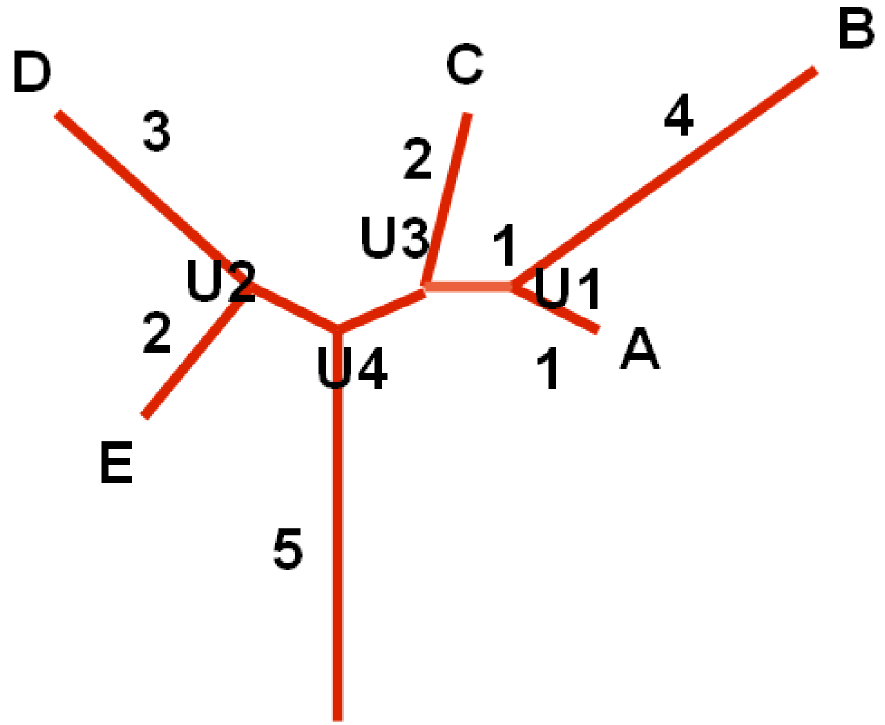
  Same as EU and FU

□ Then we get the new distance matrix

	U1	C	D	E
U1	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

Matrix 3

- □ Repeat 1 to 5 until all branches are done
- □ In this example, we will get this at the end



1.2 UPGMA

- **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean
- Assumes a constant rate of evolution over time or among lineages (molecular clock hypothesis)
- This assumption have to be tested and justified before analysis

New Mexicans for Science and Reason EXAMPLE CALCULATION OF PHYLOGENIES: THE UPGMA METHOD

http://www.nmsr.org/upgma.htm

UPGMA

1. Compare the differences among taxa & create distance matrix
2. Join and average the values of the closet match taxon (the smallest value have to be combined first)
3. The tree is build following the value of differences
4. Join the taxon until finish

Cytochrom C comparisons

(Fitch and Margoliash, Science Vol. 155, 20 Jan. 1967)

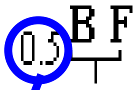
		Turtle	Man	Tuna	Chicken	Moth	Monkey	Dog
		A	B	C	D	E	F	G
Turtle	A	----	----	----	----	----	----	----
Man	B	19	----	----	----	----	----	----
Tuna	C	27	31	----	----	----	----	----
Chicken	D	8	18	26	----	----	----	----
Moth	E	33	36	41	31	----	----	----
Monkey	F	18	1	32	17	35	----	
Dog	G	13	13	29	14	28	12	----

19 difference in the amino acid sequences between man and turtle
1 difference in the amino acid sequences between man and monkey

Combine and average the closet match cells (same color code)

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

JOIN B, F →



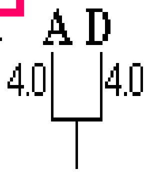
	A	BF	C	D	E	G
A						
BF	18.50					
C	27.00	31.50				
D	8.00	17.50	26.00			
E	33.00	35.50	41.00	31.00		
G	13.00	12.50	29.00	14.00	28.00	

The smallest number (boldface)
So, combine B and F

The smallest number/2 = distance
Distance between B and F = 0.5+0.5 = 1

	A	BF	C	D	E	G
A						
BF	18.50					
C	27.00	31.50				
D	**8.00**	17.50	26.00			
E	33.00	35.50	41.00	31.00		
G	13.00	12.50	29.00	14.00	28.00	

JOIN A, D →

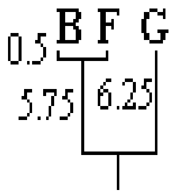


	AD	BF	C	E	G
AD					
BF	18.00				
C	26.50	31.50			
E	32.00	35.50	41.00		
G	13.50	12.50	29.00	28.00	

The smallest number (boldface)
So, combine A and D

	AD	BF	C	E	G
AD					
BF	18.00				
C	26.50	31.50			
E	32.00	35.50	41.00		
G	13.50	12.50	29.00	28.00	

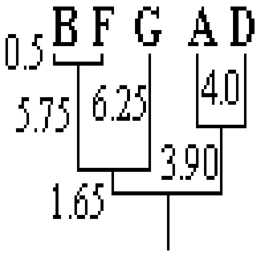
JOIN BF, G



	AD	BFG	C	E
AD				
BFG	15.80			
C	26.50	30.30		
E	32.00	31.80	41.00	

	AD	BFG	C	E
AD				
BFG	15.80			
C	26.50	30.30		
E	32.00	31.80	41.00	

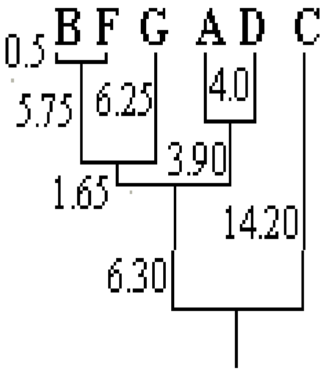
JOIN AD, BFG



	ADBFG	C	E
ADBFG			
C	28.40		
E	31.90	41.00	

	ADBFG	C	E
ADBFG			
C	28.40		
E	31.90	41.00	

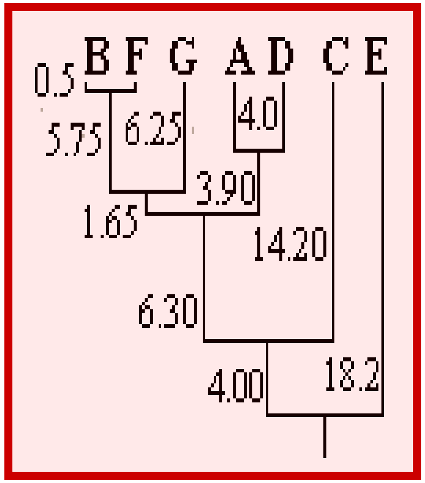
JOIN ADBFG, C



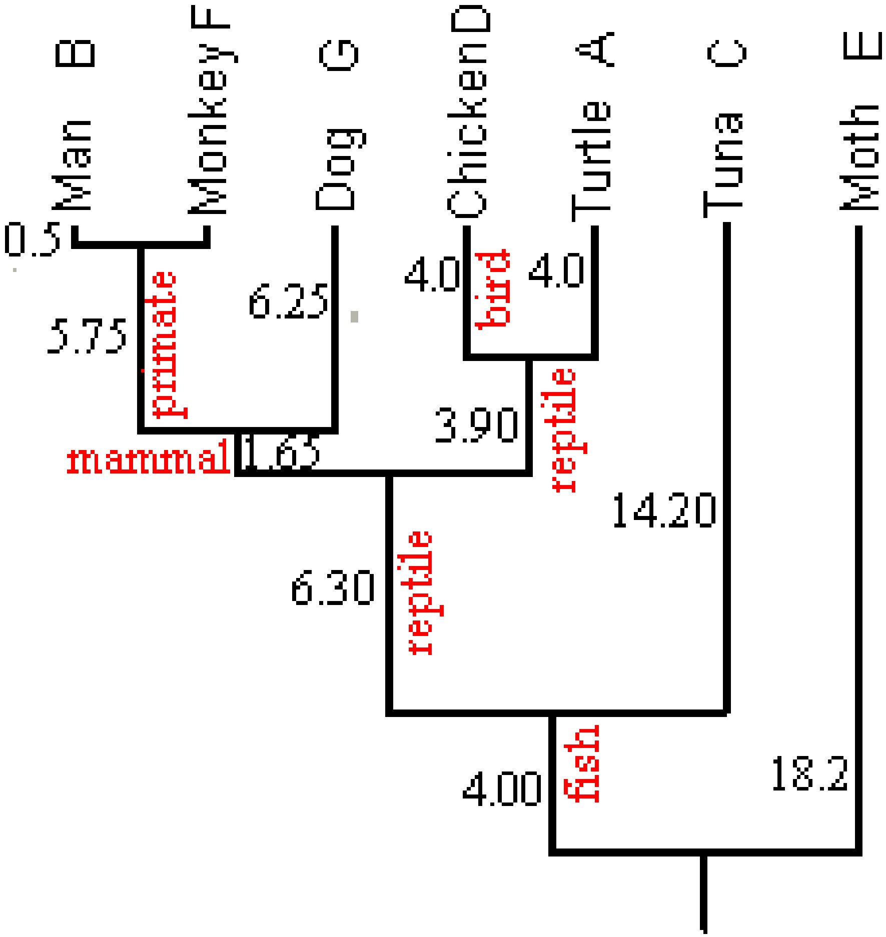
	ADBFGC	E
ADBFGC		
E	36.40	

	ADBFGC	E
ADBFGC		
E	36.40	

JOIN ADBFGC, E



UPGMA result



Interpretation

After the reptile/mammal split, birds splitting from reptiles…

It is in perfect match with the "fossil record"

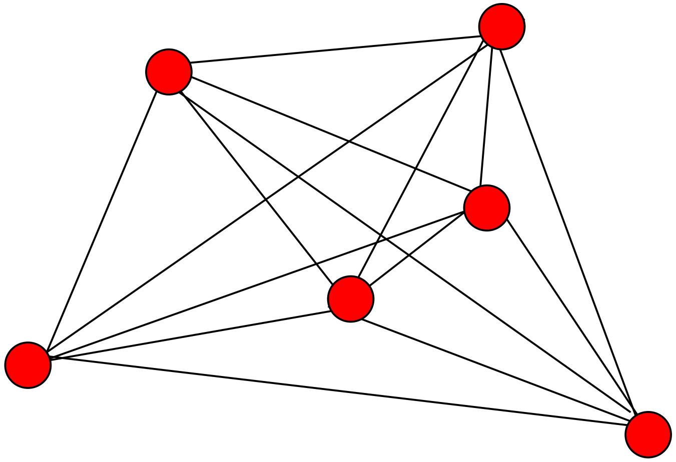
Distance based method

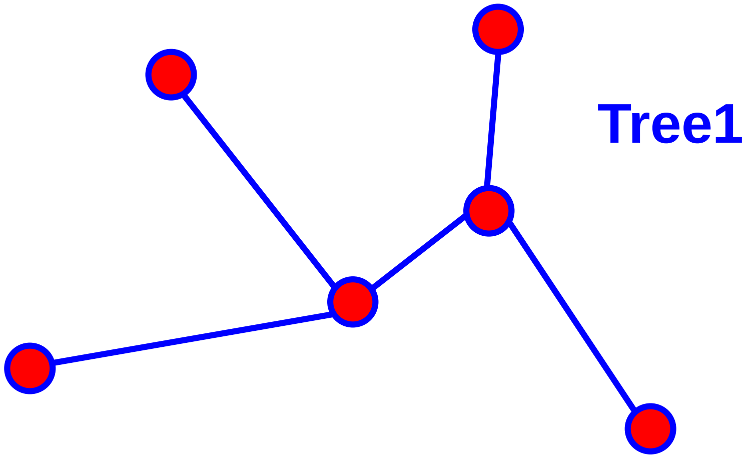
- **Advantages:**
  - **A single tree is estimated: easy**
  - **Fast with little computational expenditure**
  - **Easy to handle large numbers of sequences**

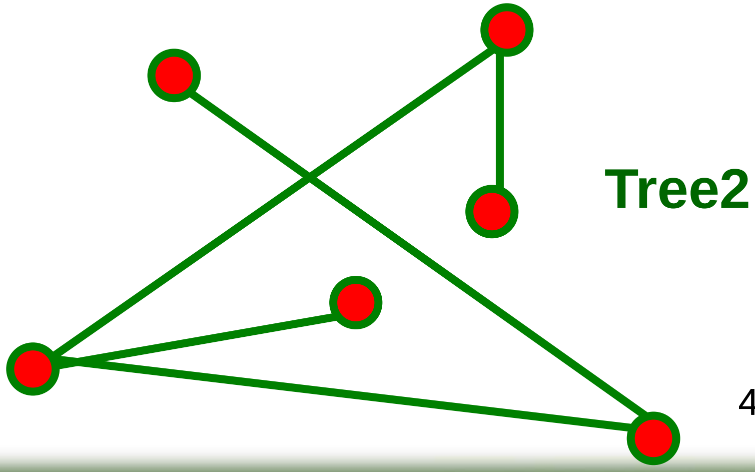
- **Disadvantages:**
  - **Lacks accuracy: no attempt to correct homoplasy**
  - **No optimizing criterion**
  - **Assume molecular clock (UPGMA)**
  - **A single tree is estimated: no confidence**

Minimum spanning tree

- A tree is a connected graph without cycles
- The MST = the shortest length (when reach out) that connect all points (vertices)







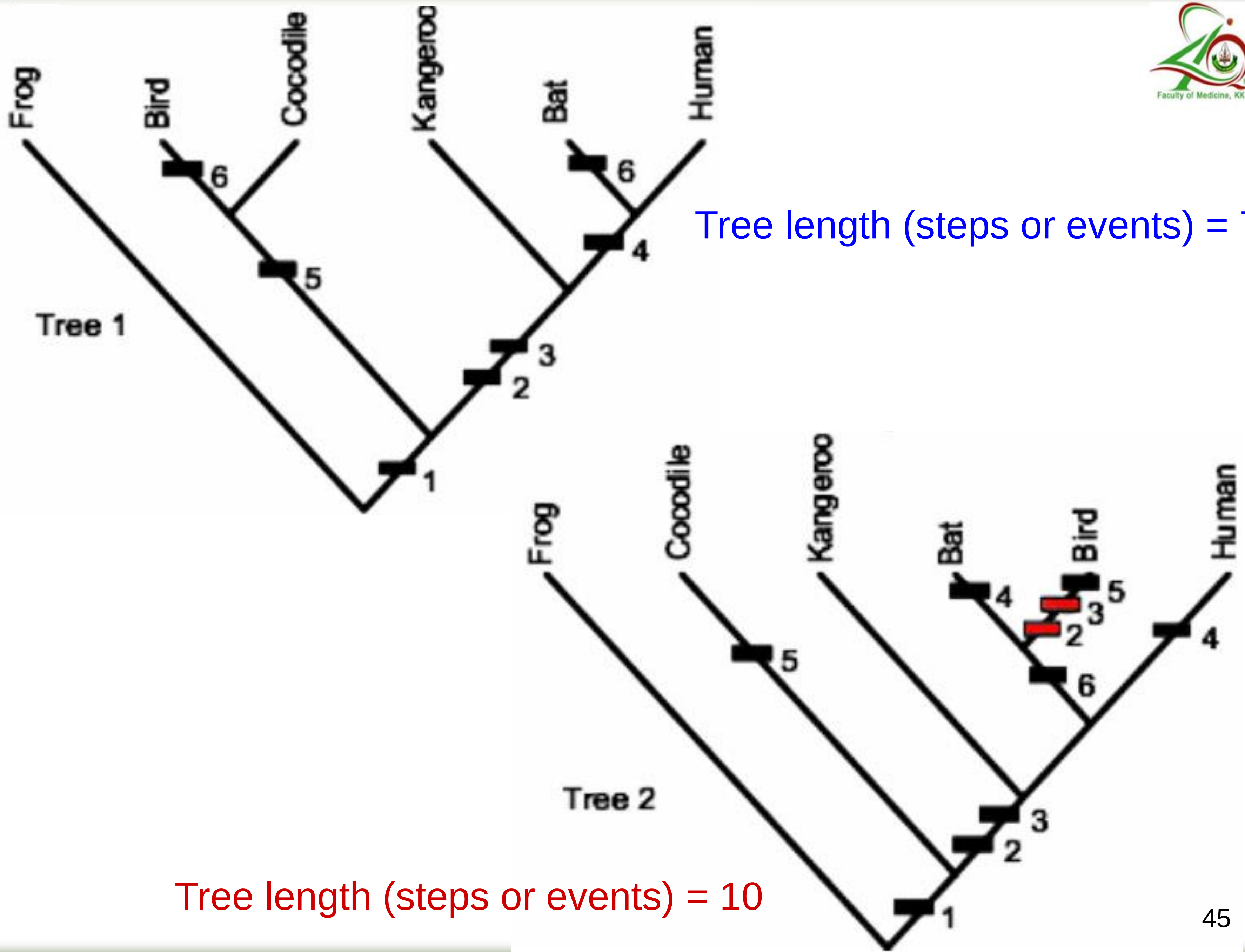
2. Parsimony based method

- Minimize the number of changes that are needed to explain the data

- Use a simple algorithm to determine how many "steps" are required to explain the distribution of each character
  (i.e., prefer the simpler relationship)

- The steps may be base or amino-acid substitutions for sequence data, or gain and loss events

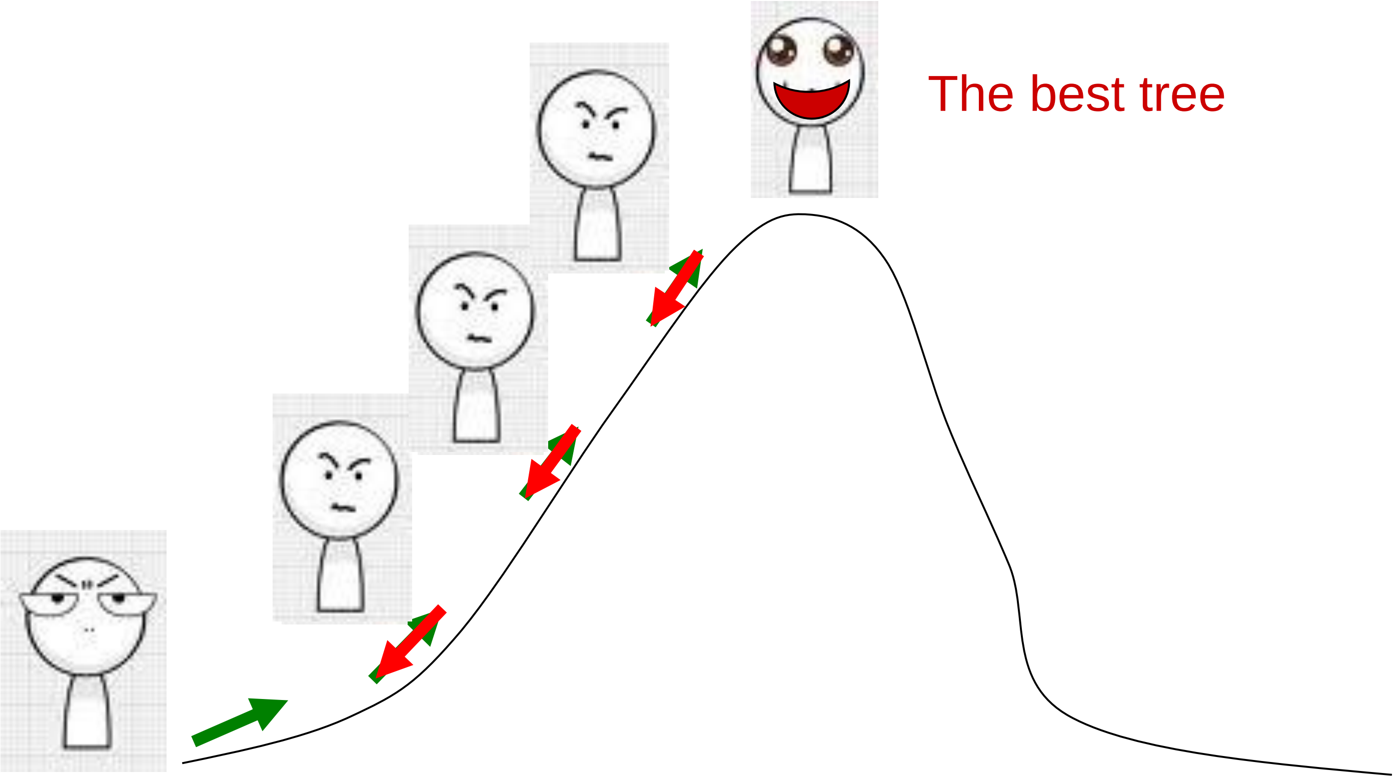
- Maximum parsimony tree: the most parsimonious distribution = preferred hypothesis of relationships among taxa



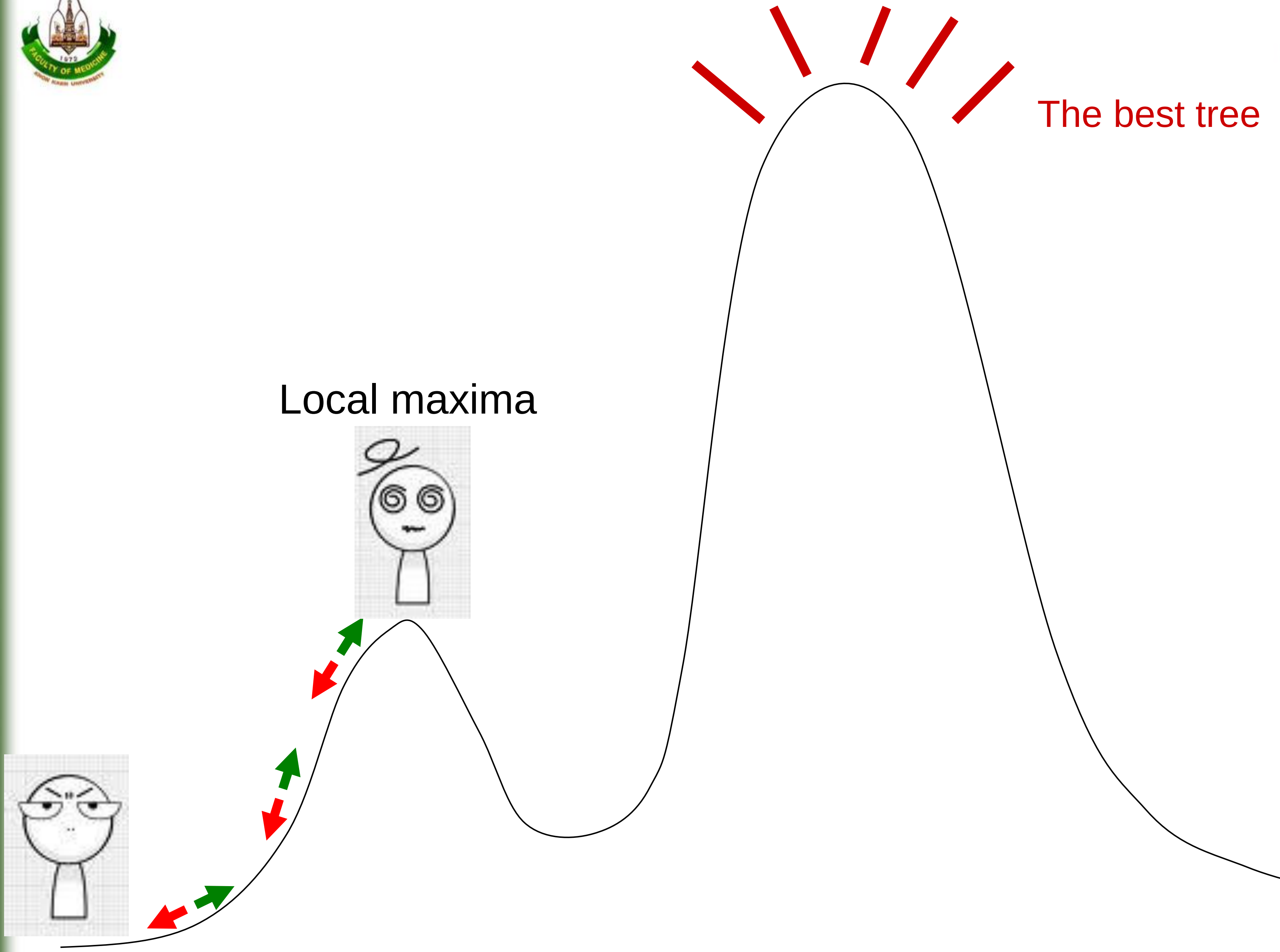
Tree length (steps or events) = 7

Tree length (steps or events) = 10

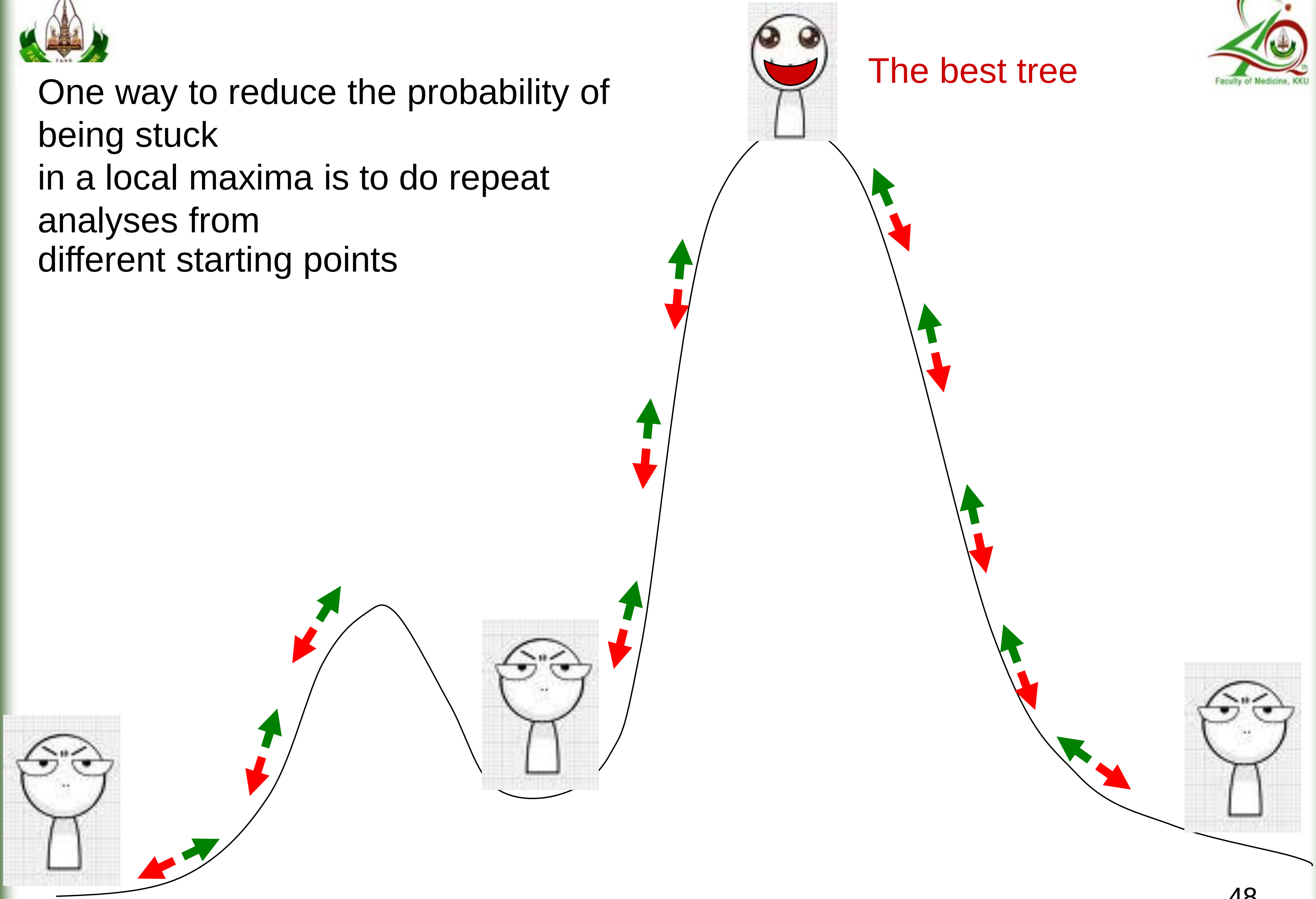
Searches through tree topologies in 'tree-space' (hill) using a 'hill-climbing' algorithm.



- **Accept uphill move**
- **Reject down hill move**


The best tree
Local maxima

One way to reduce the probability of being stuck
in a local maxima is to do repeat analyses from
different starting points



Parsimony

- **Advantages:**
  - When the data is simple = generally accurate method
  - Does not reduce sequence information to a single number (that found from distance methods)
  - Relatively fast and undemanding (faster than ML)

- **Disadvantages**
  - Several typical "shortest trees">> Potentially ambiguous consensus topology (if trees with the same score)
  - Prone to error under certain circumstances (homoplasy/ LBA)
  - Cladogram: not provide branch lengths

3. Maximum likelihood (ML) method

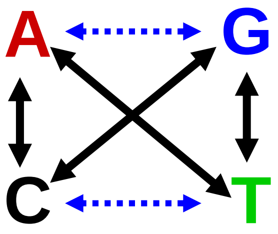
- Using a model for **sequence evolution**
- Create a tree that gives the **highest likelihood** of occurring with the given data
- The process of sequence evolution is not as simple as parsimony assumes

☀	🌧	🌨	FOG
0.25	0.25	0.25	0.25

- Probability to rain?
- Probability to rain in November?
- Probability to rain in November in Southern of Thailand?
- Probability to rain in 2 successive days, in November in Southern of Thailand?

A T C G

- Transition/ Transversion and likelihood of amino acid changes
- House keeping gene and Conservative region
- Back substitution
- Multiple substitution (in one site or one branch)



Maximum likelihood method

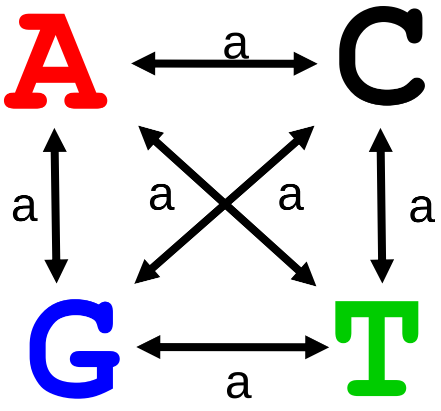
- Pick an evolutionary model (JC, K2P, GTR etc.)
- Generate all possible tree structures
- Calculate Likelihood of these trees and sum them to get the column likelihood for each OTU cluster.
- Calculate Tree Likelihood by multiplying the likelihood for each position
- Choose tree with greatest likelihood

Models of Sequence Evolution

Imagine tossing a coin and getting a head.
What is the likelihood of that result?

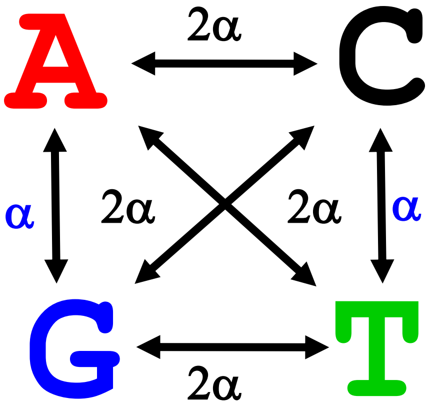
Pick an Evolutionary Model

Jukes Cantor (JC)



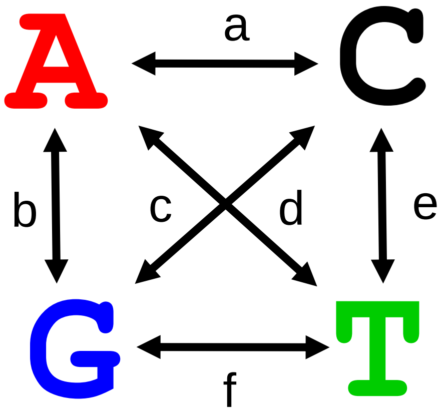
All equal

Kimura (K2P)



Transition/transversion

General (GTR)



All free

Evolutionary (nucleotide substitution) Model
: rates of change from one nucleotide to another

- JC
- K2P
- GTR (most general usable model)

Models to describe rate variation among sites in a sequence

- gamma distribution (G)
- proportion of invariable sites (I)

"GTR model of nucleotide substitution
with gamma model of rate of heterogeneity"

Sequence W: A C G C G T T G G G

Sequence X: A C G C G T T G G G

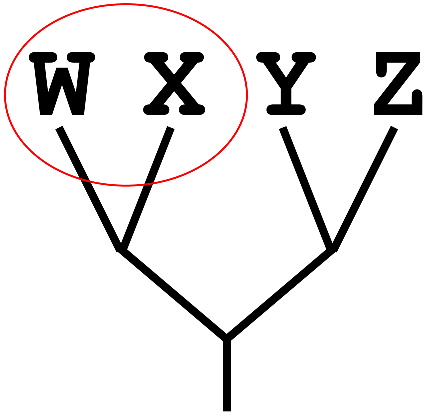
Sequence Y: A C G C A A T G A A

Sequence Z: A C A C A G G G A A

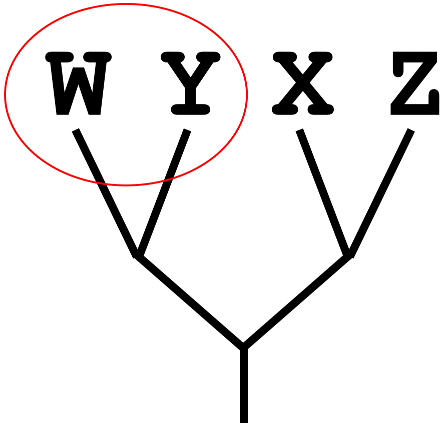
4 sequences
10 sites

Select No. 6th as example

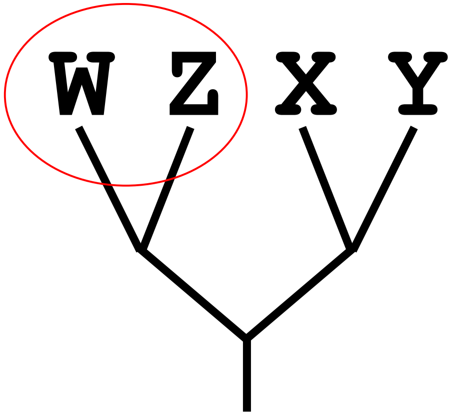
Tree 1



Tree 2



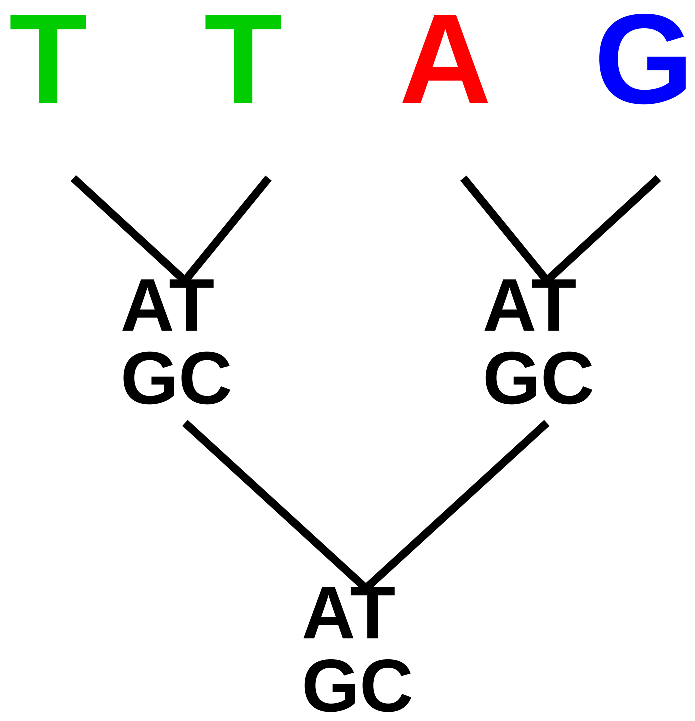
Tree 3



Possible Trees (unrooted)

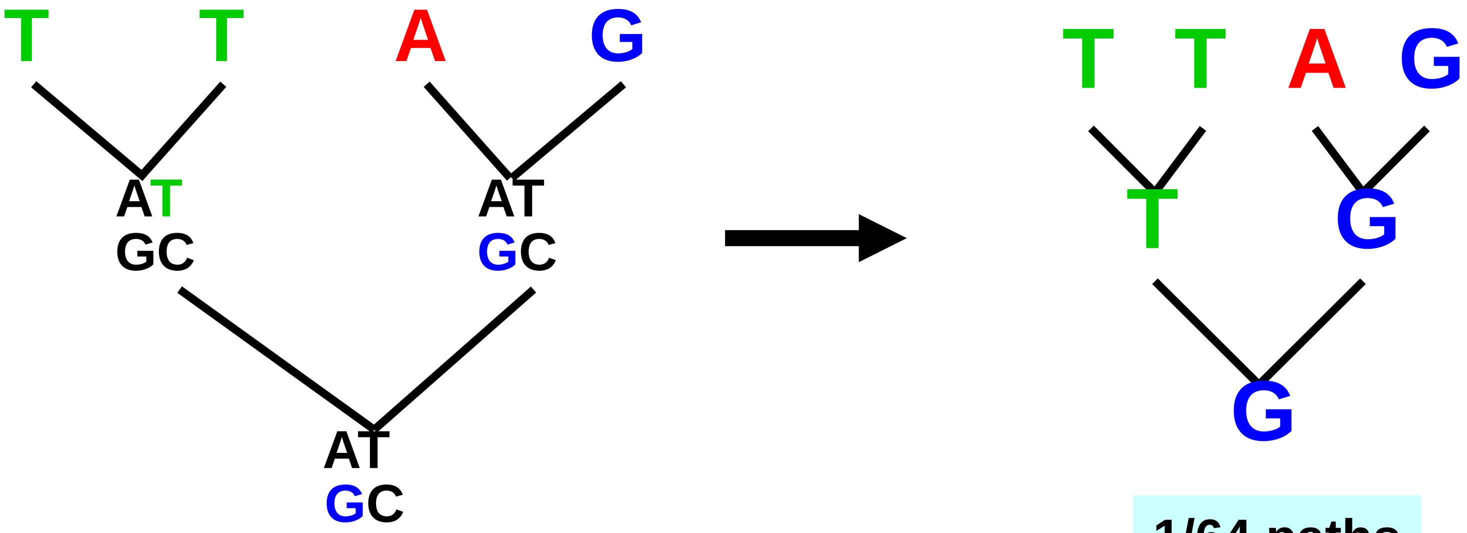
All possible evolutionary paths (position 6th)

Tree 1



Likelihood for one path of one position (6th)

Tree 1



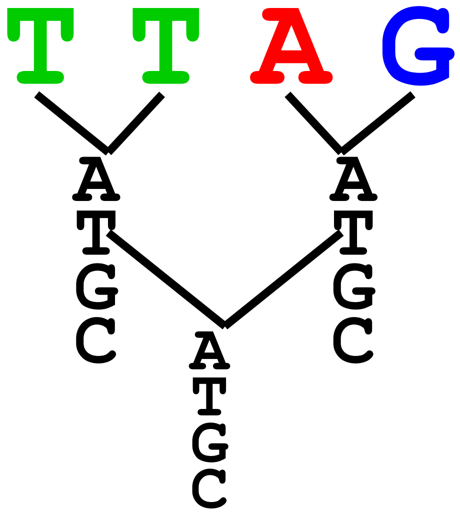
1/64 paths

$$L(path) = L(root) \times \prod L(branches)$$

$$= P(G \rightarrow T)P(G \rightarrow G)\ P(G \rightarrow A)P(G \rightarrow G)\ P(T \rightarrow T)P(T \rightarrow T)$$

Sum over all (64) paths of one position (6th)

Tree 1

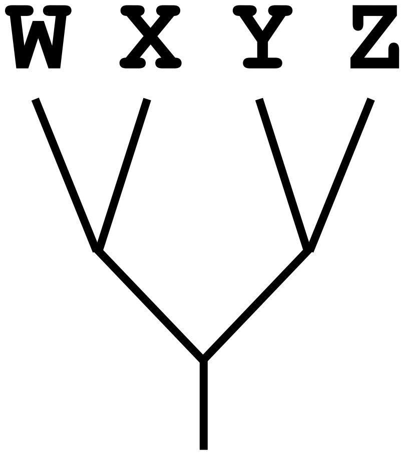


L(Column Cluster 1) = Σ L(all possible Evolutionary Paths)

= L(path1) + L(path2) + L(path3) + … + L(path64)

Whole Sequence Likelihood
For all (10) positions, all (640) paths

Tree 1



$$L(\text{Sequence}) = L(\text{root}) \times \Pi_i\, L\ (\text{each position } i)$$

Do the rest of the possible tree

↓

Choose the tree with the "Maximum Likelihood"

Maximum likelihood method

Advantages

- Highly accurate, allows various forms of homoplasy to be corrected
- Single tree is produced that is generally precise (choose the best likely tree)

Disadvantages

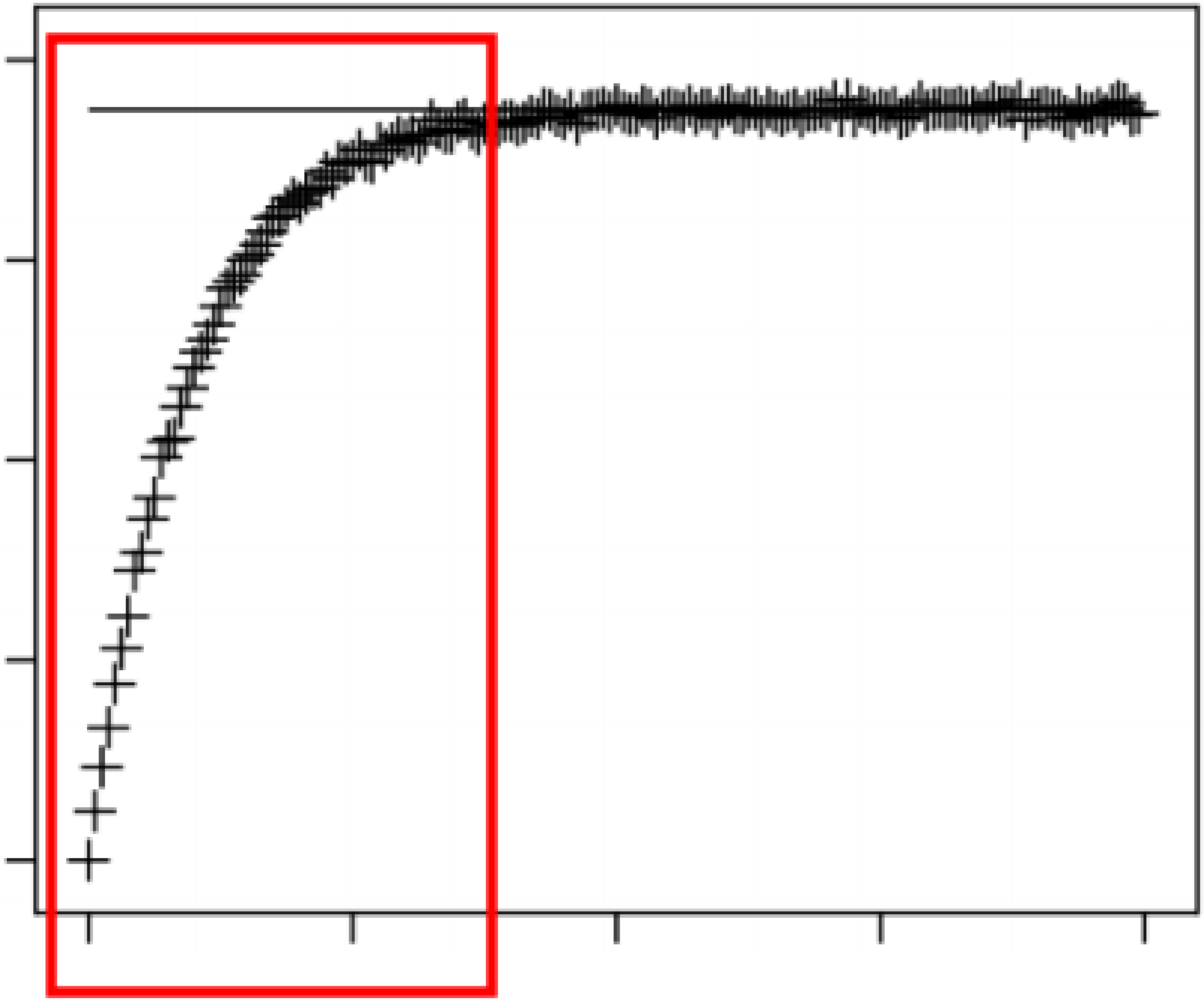
- Complexity process = slow and computationally demanding
- The hill-climbing algorithm is susceptible to local optima

4. Bayesian methods

- Maximum likelihood tries to find the best values (single tree with most likelhood) for the branch lengths and model parameters
- Bayesian inference allows these parameters to have some uncertainty (Distribution of trees)
- Maximum likelihood: the probability of the data given the model
- Bayesian inference: the probability of the model given the data (posterior probability)
- Bayesian inference requires a prior probability to be set for each parameter

Simon Harris, Wellcome Trust Sanger Institute

Bayesian methods



Simon Harris, Wellcome Trust Sanger Institute

Bayesian method

Advantages:

- Potential for any complex model (more complex than ML)
- Provides tree and support for the relationships in a single analysis
- Able to break out of local maxima

Disadvantages:

- Must be specified prior probabilities (require sufficient knowledge of these probabilities?)
- Must be run long enough (but never certain) for the result to smooth out

Simon Harris, Wellcome Trust Sanger Institute

Comparison of Methods

Distance	Maximum parsimony	Maximum likelihood
Uses only pairwise distances	Uses only shared derived characters	Uses all data
Minimizes distance between nearest neighbors	Minimizes total distance	Maximizes tree likelihood given specific parameter values
Very fast	Slow	*Very* slow
Easily trapped in local optima	Assumptions fail when evolution is rapid	Highly dependent on assumed evolution model
Good for generating tentative tree, or choosing among multiple trees	Best option when tractable (<30 taxa, homoplasy rare)	Good for very small data sets and for testing trees built using other methods

Pro and cons

- Character based (ML/ MP) better than distance based (not reduce character as single event), but demanding higher computer.
- So, with >1000 taxa, distance based is appropriate to cove with limitation.
- NJ is better than UPGMA (molecular clock assumptions)/ provide single tree but ignore other possible tree.
- ML is better than MP because correct homoplasy, still sensitive to local optima.
- Bayasien is the best (posterior probability/ break local optima).

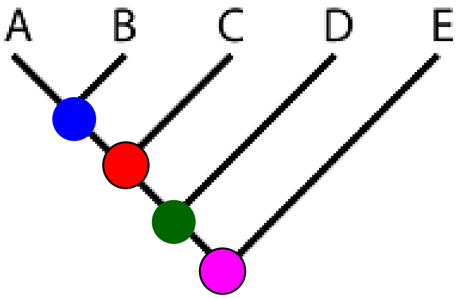
ML is generally the best, esp. for sequence data, but demanding

Most commonly used packages contain software for all three methods (Not Bayesian method)

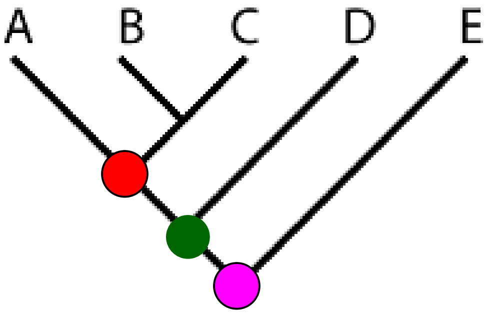
It would be more confident to use more than 1 method to built the tree

Consensus tree; When there are 3 trees from analysis?

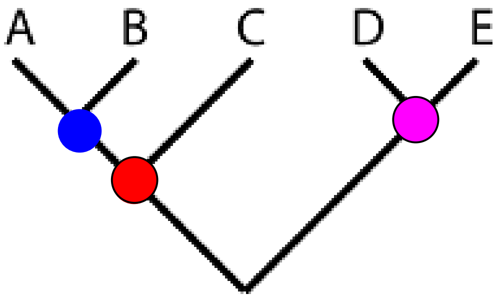
Tree 1



Tree 2

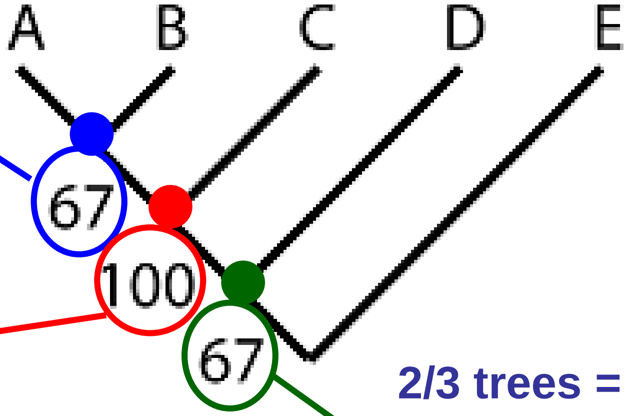


Tree 3



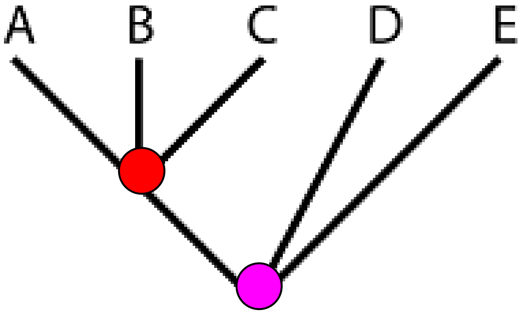
2/3 trees = 67%
A & B are from common ancestor
(except Tree2)

Majority rule consensus



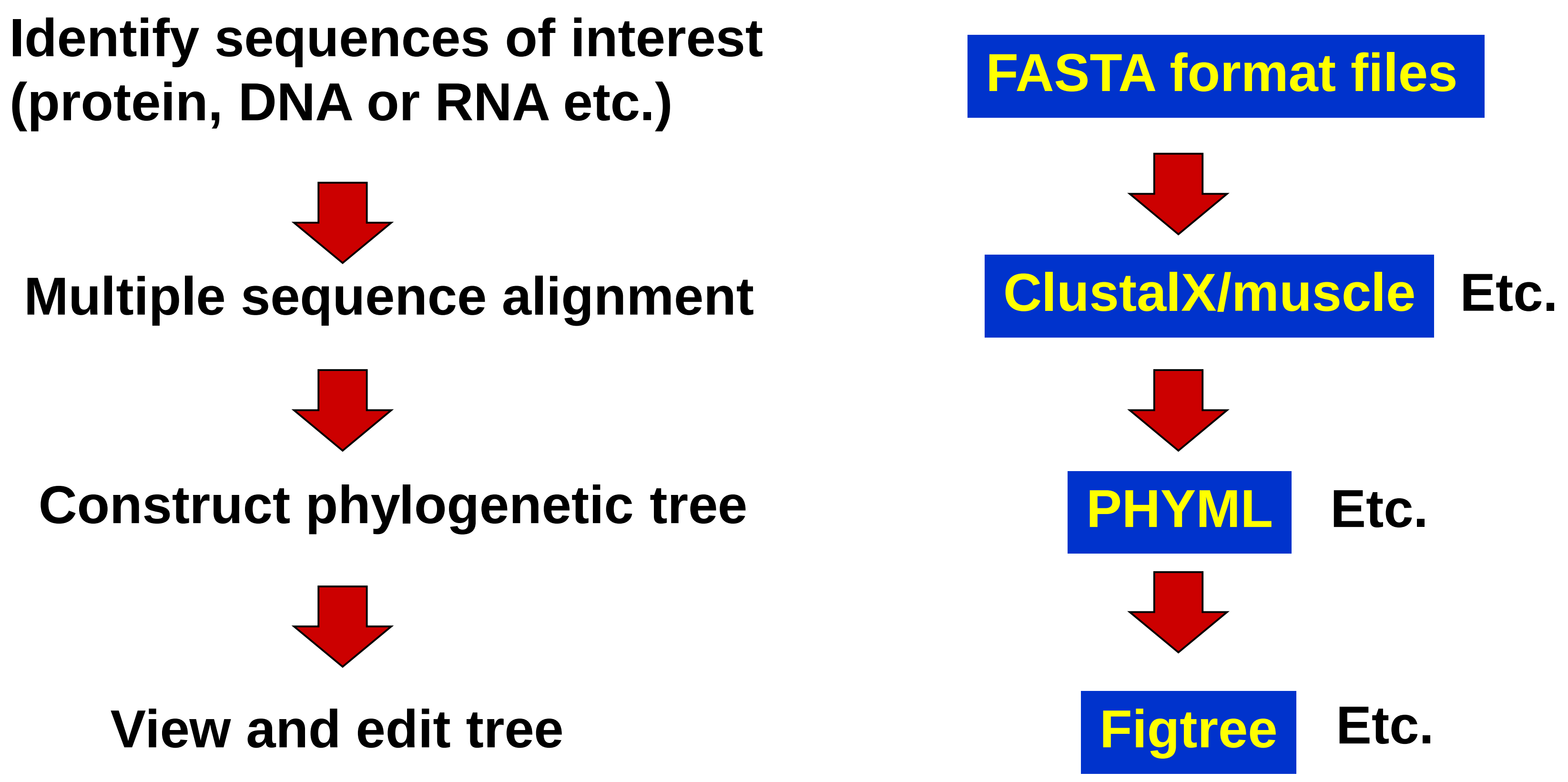
3/3 trees = 100%
A, B & C are from common ancestor

2/3 trees = 67%
A B C & D are from common ancestor
(except Tree3)



Strict Consensus

Steps for building the tree



Identify sequences of interest (protein, DNA or RNA etc.)

Multiple sequence alignment

Construct phylogenetic tree

View and edit tree

FASTA format files

ClustalX/muscle **Etc.**

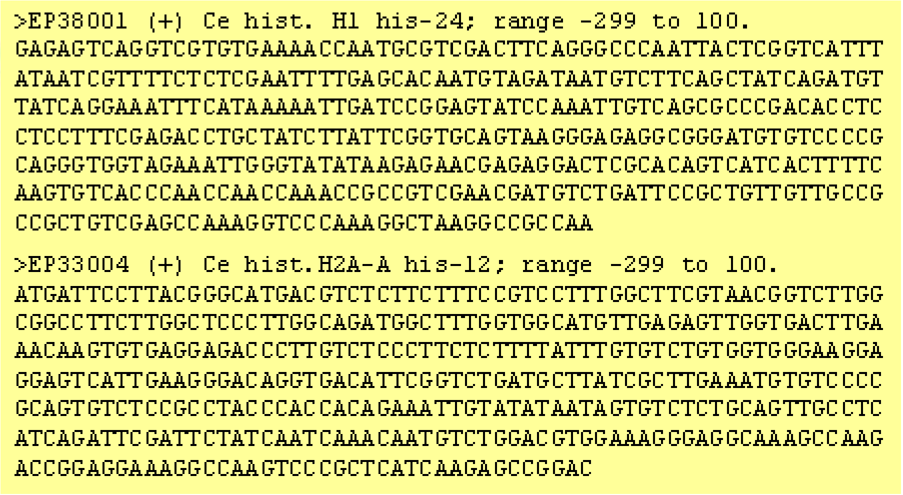
PHYML **Etc.**

Figtree **Etc.**

Alignment = hypothesis of **positional homology** between sequences

Phylogeny is meaningless unless it is based on a **well-made alignment**

FASTA format



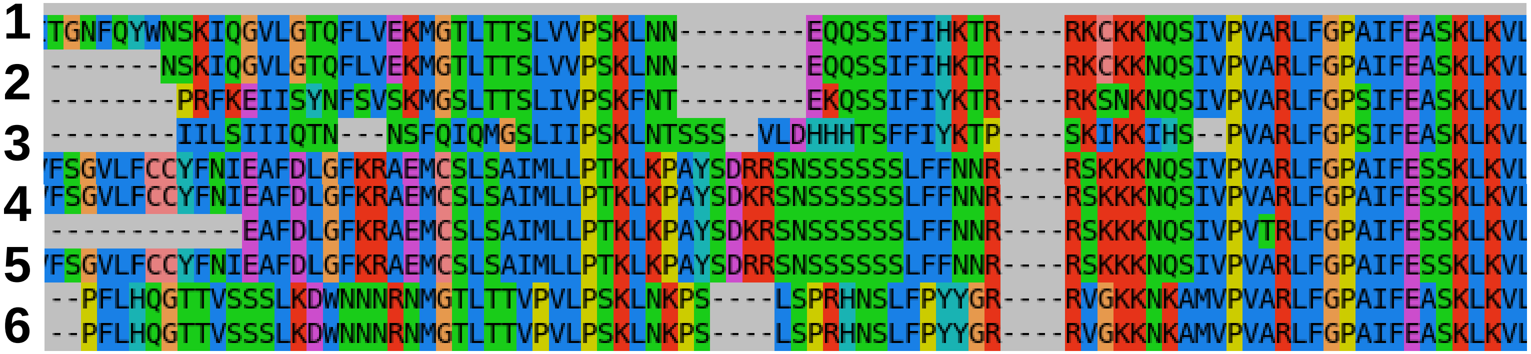
```
>EP38001 (+) Ce hist. H1 his-24; range -299 to 100.
GAGAGTCAGGTCGTGTGAAAACCAATGCGTCGACTTCAGGGCCCAATTACTCGGTCATTT
ATAATCGTTTTCTCTCGAATTTTGAGCACAATGTAGATAATGTCTTCAGCTATCAGATGT
TATCAGGAAATTTCATAAAAATTGATCCGGAGTATCCAAATTGTCAGCGCCCGACACCTC
CTCCTTTCGAGACCTGCTATCTTATTCGGTGCAGTAAGGGAGAGGCGGGATGTGTCCCCG
CAGGGTGGTAGAAATTGGGTATATAAGAGAACGAGAGGACTCGCACAGTCATCACTTTTC
AAGTGTCACCCAACCAACCAAACCGCCGTCGAACGATGTCTGATTCCGCTGTTGTTGCCG
CCGCTGTCGAGCCAAAGGTCCCAAAGGCTAAGGCCGCCAA

>EP33004 (+) Ce hist.H2A-A his-12; range -299 to 100.
ATGATTCCTTACGGGCATGACGTCTCTTCTTTCCGTCCTTTGGCTTCGTAACGGTCTTGG
CGGCCTTCTTGGCTCCCTTGGCAGATGGCTTTGGTGGCATGTTGAGAGTTGGTGACTTGA
AACAAGTGTGAGGAGACCCTTGTCTCCCTTCTCTTTTATTTGTGTCTGTGGTGGGAAGGA
GGAGTCATTGAAGGGACAGGTGACATTCGGTCTGATGCTTATCGCTTGAAATGTGTCCCC
GCAGTGTCTCCGCCTACCCACCACAGAAATTGTATATAATAGTGTCTCTGCAGTTGCCTC
ATCAGATTCGATTCTATCAATCAAACAATGTCTGGACGTGGAAAGGGAGGCAAAGCCAAG
ACCGGAGGAAAGGCCAAGTCCCGCTCATCAAGAGCCGGAC
```

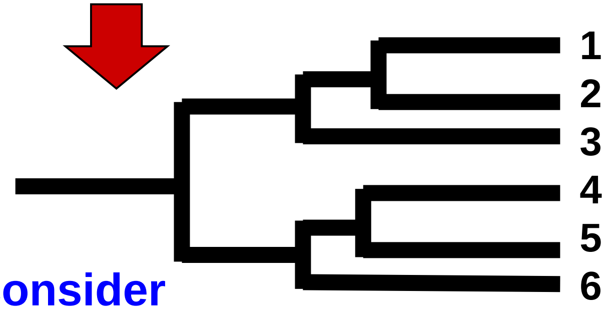
- You know the best what is your biology of interest!

-What are you doing!!!

Alignment is critical
Bad Alignment = bad tree



```
ITGNFQYWNSKIQGVLGTQFLVEKMGTLTTSLVVPSKLNN--------EQQSSIFIHKTR----RKCKKNQSIVPVARLFGPAIFEASKLKVL
-------NSKIQGVLGTQFLVEKMGTLTTSLVVPSKLNN--------EQQSSIFIHKTR----RKCKKNQSIVPVARLFGPAIFEASKLKVL
--------PRFKEIISYNFSVSKMGSLTTSLIVPSKFNT--------EKQSSIFIYKTR----RKSNKNQSIVPVARLFGPSIFEASKLKVL
--------IILSIIIQTN---NSFQIQMGSLIIPSKLNTSSS--VLDHHHTSFFIYKTP----SKIKKIHS--PVARLFGPSIFEASKLKVL
VFSGVLFCCYFNIEAFDLGFKRAEMCSLSAIMLLPTKLKPAYSDRRSNSSSSSSLFFNNR----RSKKKNQSIVPVARLFGPAIFESSKLKVL
VFSGVLFCCYFNIEAFDLGFKRAEMCSLSAIMLLPTKLKPAYSDKRSNSSSSSSLFFNNR----RSKKKNQSIVPVARLFGPAIFESSKLKVL
------------EAFDLGFKRAEMCSLSAIMLLPTKLKPAYSDKRSNSSSSSSLFFNNR----RSKKKNQSIVPVTRLFGPAIFESSKLKVL
VFSGVLFCCYFNIEAFDLGFKRAEMCSLSAIMLLPTKLKPAYSDRRSNSSSSSSLFFNNR----RSKKKNQSIVPVARLFGPAIFESSKLKVL
--PFLHQGTTVSSSLKDWNNNRNMGTLTTVPVLPSKLNKPS----LSPRHNSLFPYYGR----RVGKKNKAMVPVARLFGPAIFEASKLKVL
--PFLHQGTTVSSSLKDWNNNRNMGTLTTVPVLPSKLNKPS----LSPRHNSLFPYYGR----RVGKKNKAMVPVARLFGPAIFEASKLKVL
```



- Chose the most appropriate model to your data

- Maybe try many models, compare and consider

Free programs for phylogenetic analysis

There are hundreds of free available programs that involved in phylogenetic tree e.g.

- PhyML (ML): http://atgc.lirmm.fr/phyml/
- PAUP* (NJ, MP, ML): http://paup.csit.fdsu.edu
- **PHYLIP (NJ, MP, ML):** http://evolution.genetics.washington.edu/phylip.html
- MrBayes (Bayesian): http://mrbayes.csit.fdsu.edu
- Splitstree (Networks): http://www.splitstree.org
- FindModel (Model Test): http://www.hiv.lanl.gov/content/sequence/findmodel/findmodel.html
- SeaView (Contains Clustal, Muscle, PHYLIP and PhyML + a simple tree viewer: http://pbil.univ-lyon1.fr/software/seaview.html

- **Felsenstein's Phylogeny program page** (links to available software): **http://evolution.genetics.washington.edu/phylip/software.html**

Practices

- – Construct the tree from sequences
- – Using SeaView
  - Contains Clustal, Muscle + PHYLIP and PhyML + a simple tree viewer
- – Bootstrapping, branch length, Re-root
- – Interpret the tree

Recommended books, articles and links

- How to read a phylogenetic tree: http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny
- Hall Phylogenetic trees made easy. Sinauer Associates.
- Page & Holmes Molecular Evolution: A Phylogenetic Approach. Blackwell Science.
- Felsenstein Inferring Phylogenies. Sinauer Associates.

- **Natural selection and variation** http://www.blackwellpublishing.com/ridley/EVOC04.pdf
- **Evolutionary Developmental Biology** http://www.blackwellpublishing.com/ridley/EVOC20.pdf
- **Evolutionary and diversity** http://www.blackwellpublishing.com/ridley/EVOC13.pdf
- **Multiple alignment and phylogenetic analysis** http://cmgm.stanford.edu/classes/pdf/phylogenetic.pdf

- MultiPhyl (ML via email): http://distributed.cs.nuim.ie/multiphyl.php
- Phylogeny.fr (Robust Phylogenetic Analysis For The Non-Specialist): http://www.phylogeny.fr/

"Nothing in Biology Makes Sense Except in the Light of Evolution"