CMSC423: Bioinformatic Algorithms, Databases and Tools

Phylogenetic trees

What is evolution?

Evolution is the change in the characteristics of a species over several generations and relies on the process of **natural selection**

What is a phylogenetic tree?

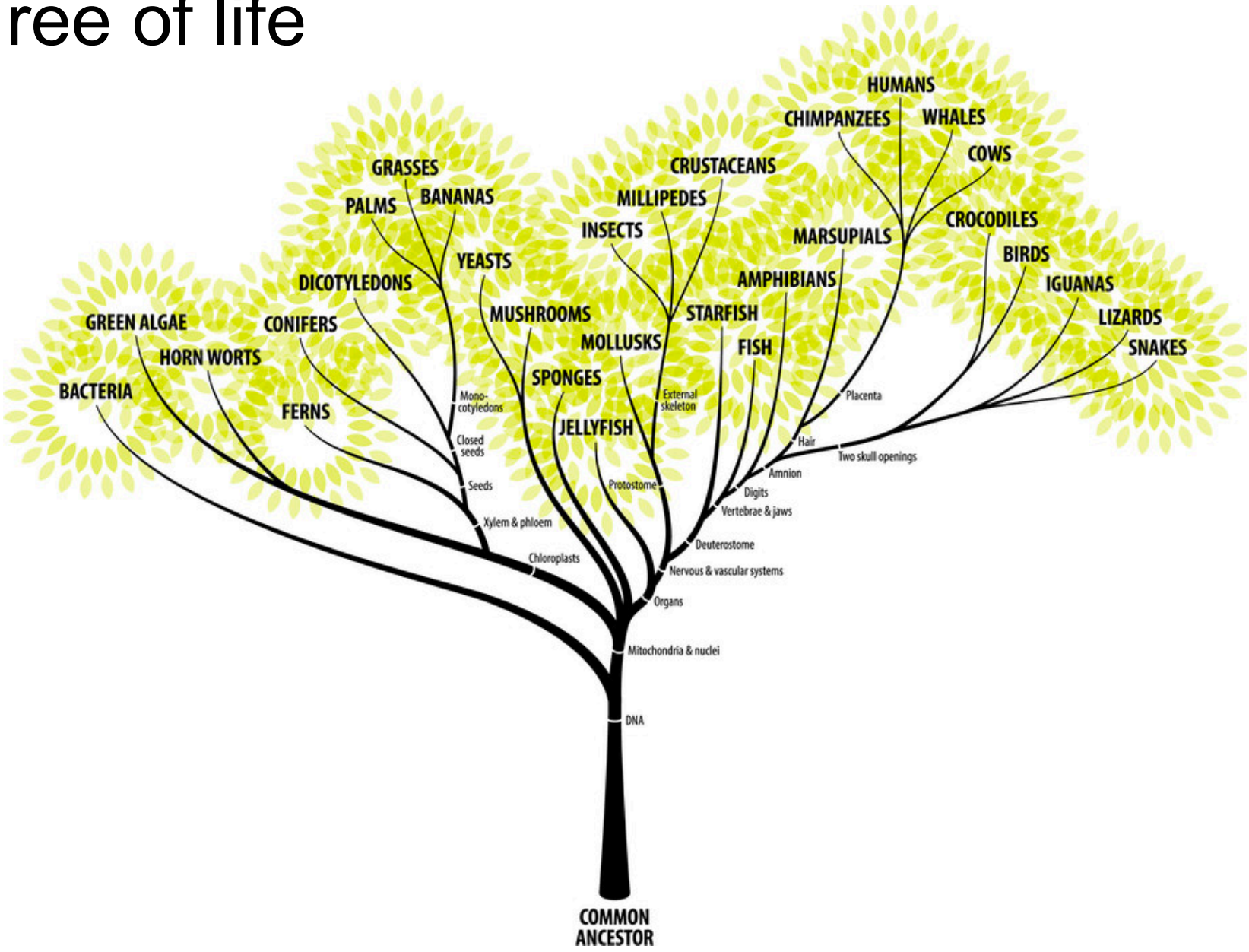
- A phylogenetic tree (evolutionary tree) is a branching diagram showing the inferred evolutionary relationships among various biological species or other entities

- It is based on similarity and differences in their physical or genetic characteristics

- The species joined together are implied to descend from a common ancestor

- A phylogenetic tree is used to help represent evolutionary relationships between organisms that are **believed to have some common ancestry**

Tree of life



http://tolweb.org/tree/phylogeny.html

Large scale efforts to understand evolution of different groups of interest

STUDY PROVIDES FRAMEWORK FOR 1 BILLION YEARS OF GREEN PLANT EVOLUTION

Oct 23, 2019 Alan Flurry (aflurry@uga.edu), U. Georgia; Katie Willis (kewillis@ualberta.ca), U. Alberta

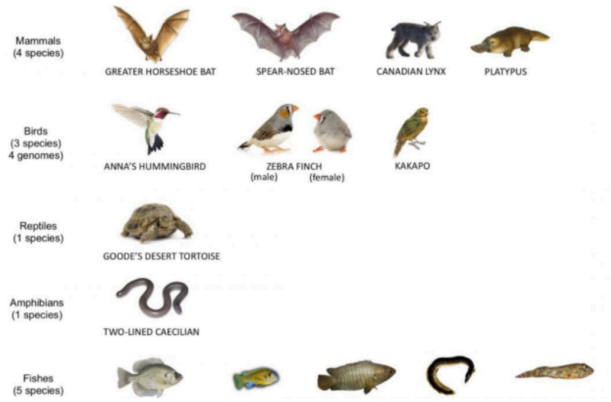
Gene sequences for more than 1100 plant species have been released by an international consortium of nearly 200 plant scientists, the culmination of a nine-year research project.

The One Thousand Plant Transcriptomes Initiative (1KP) is a global collaboration to examine the diversification of plant species, genes and genomes across the more than one-billion-year history of green plants dating back to the ancestors of flowering plants and green algae.

"In the tree of life, everything is interrelated," said Gane Ka-Shu Wong, lead investigator and professor in the University of

VERTEBRATE GENOME PROJECT IS BUILDING 66,000 GENOMES...WITH BIONANO

OCTOBER 15, 2018



CAPTION

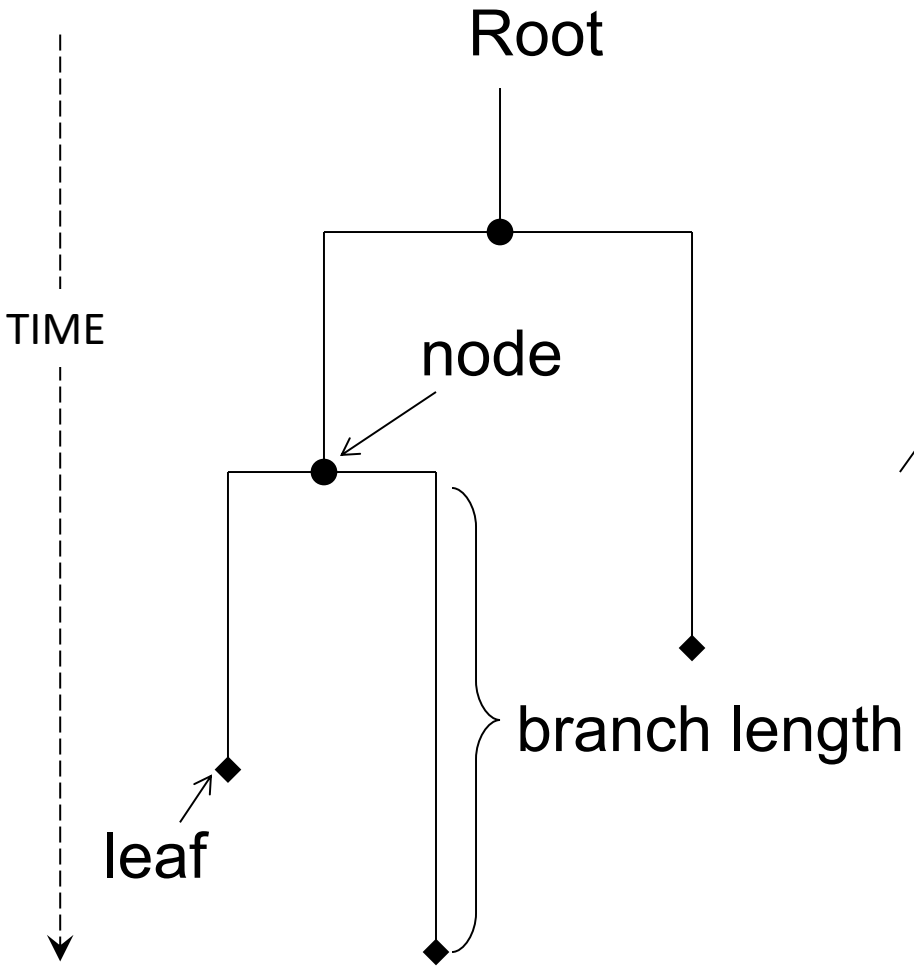
Researchers have worked out the evolutionary relationships of dozens of bird species. The findings add to the evidence that some traits -- such as vocal learning or foot-propelled underwater diving - evolved independently among different groups of birds.

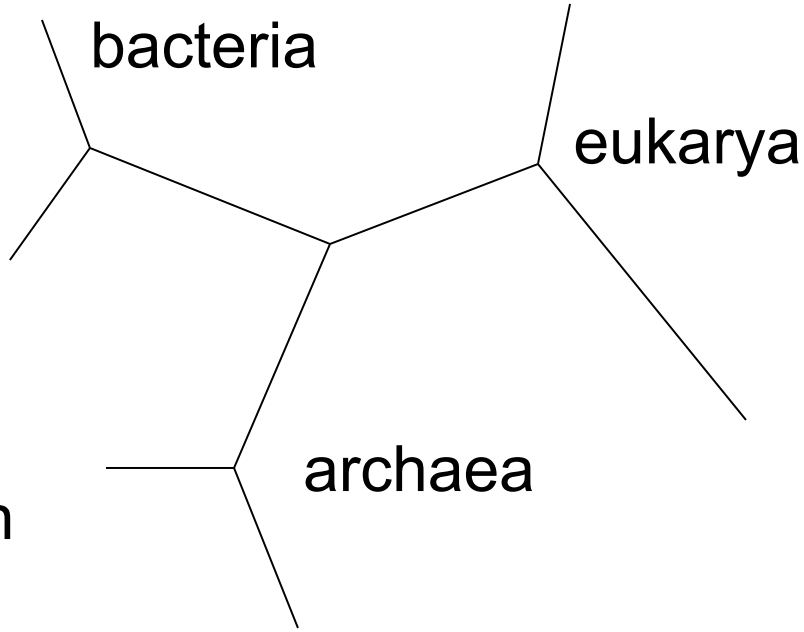
Applications

- Understanding human origin
- Understanding biodiversity
- Understanding origin of particular traits
- Understanding the process of molecular evolution
- Origin of disease
- Forensics
- Other use cases – evolution of languages, cancer tumor evolution, etc.

Who’s more related? A whale and a manatee or a whale and a cow?

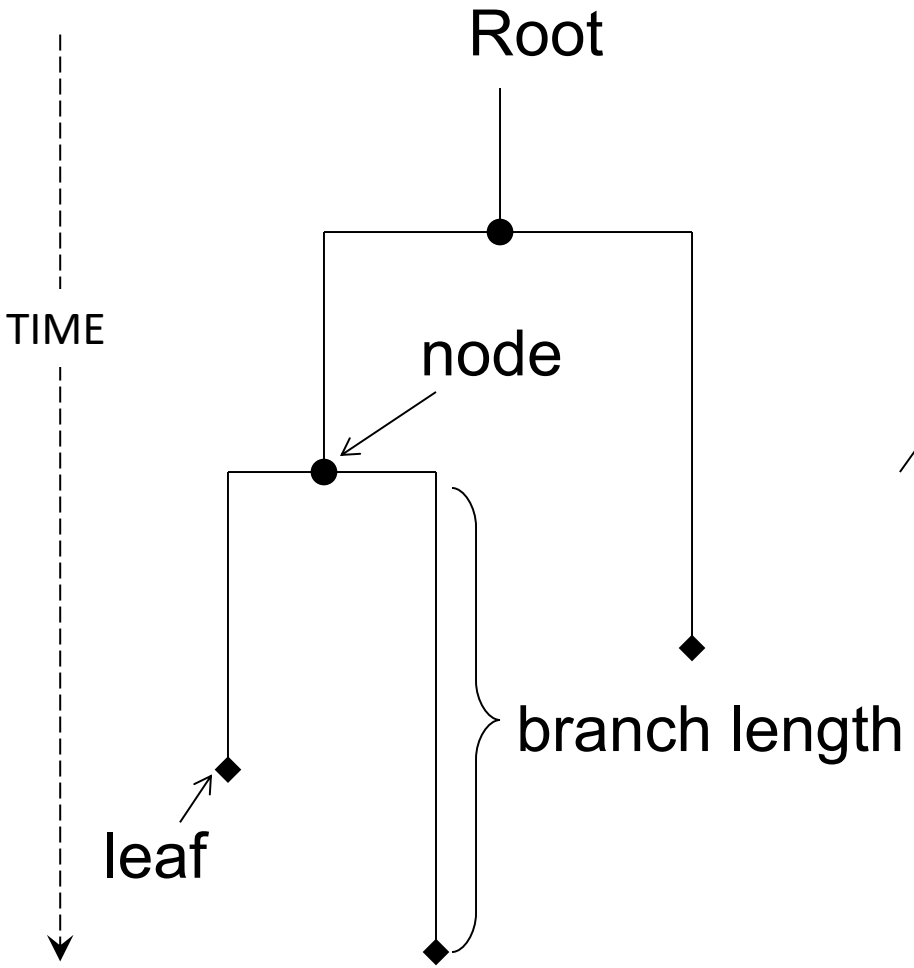
Anatomy of a tree

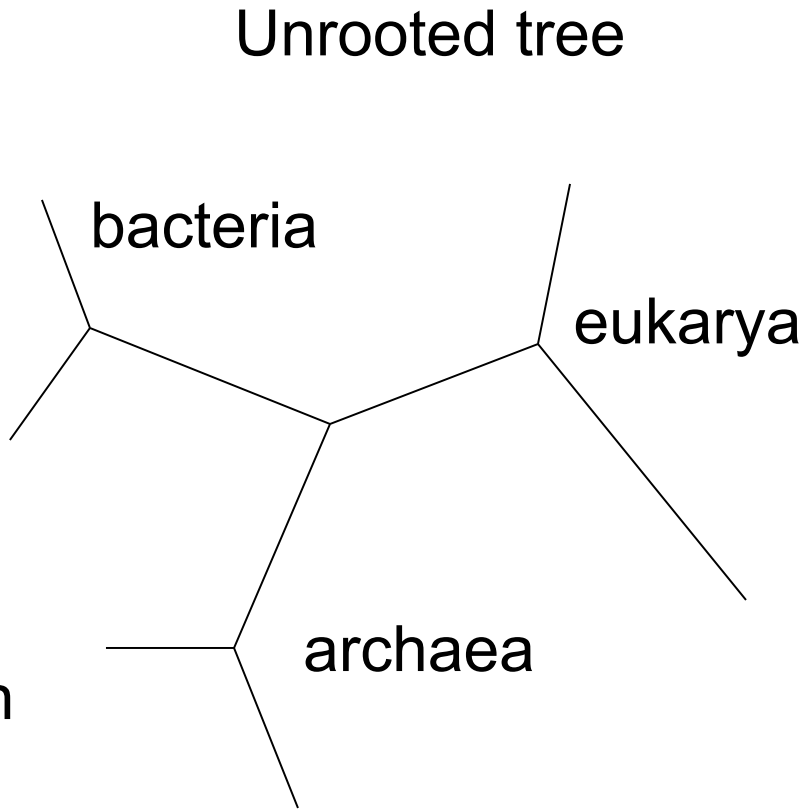




Connected and Acyclic

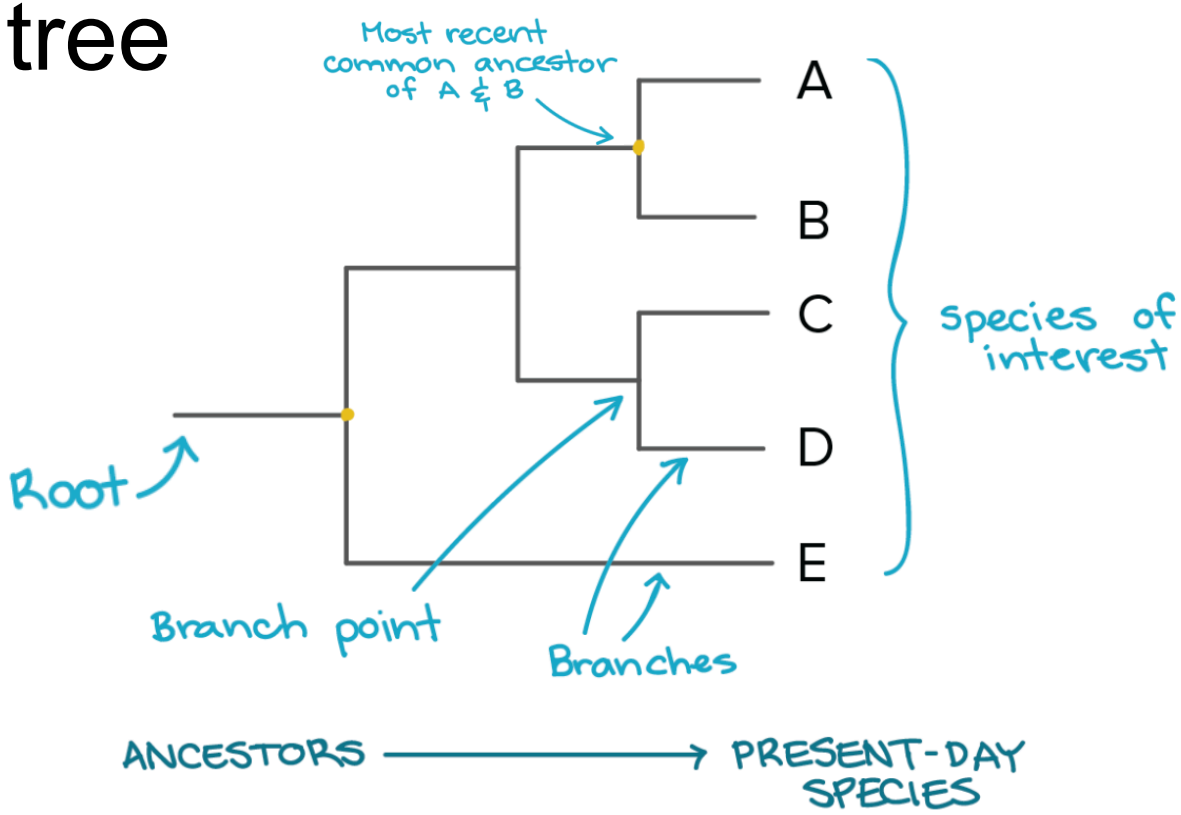
Anatomy of a tree





Phylogenetic trees are usually binary (though they don't have to)

Anatomy of tree



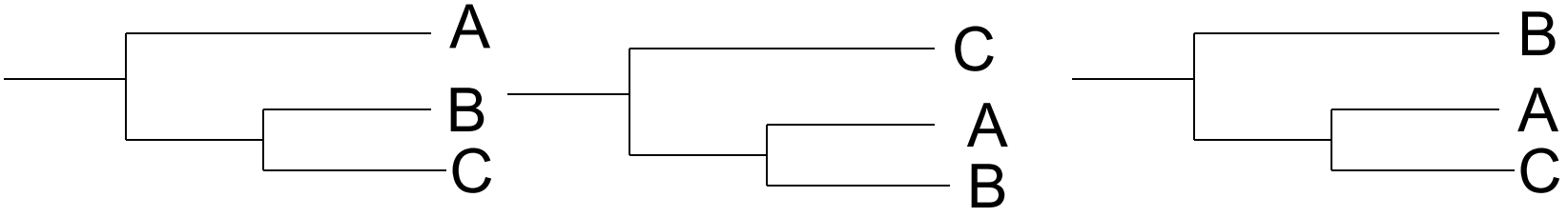
- Phylogenetic trees are usually binary (though they don't have to)
- Can be rooted or unrooted
- In trees, two species are **more related** if they have a more recent common ancestor and **less related** if they have a less recent common ancestor.

Phylogeny questions

•Given several organisms & a set of features (usually sequence, but also morphological: wing shape/color...)

•A. Given a phylogenetic tree – figure out what the ancestors looked like (what are the features of internal nodes)

•B. Find the phylogenetic tree that best describes the common evolutionary heritage of the organisms



Phylogeny questions

- A. Easy-ish – can be done with dynamic programming
- B. Hard – Many possible trees

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

rooted trees with n leaves

Phylogeny estimation methods

- Distance-based methods
- Maximum parsimony
- Maximum Likelihood

Distance-Based Phylogeny Problem

Reconstruct an evolutionary tree fitting a distance matrix

Input: A distance matrix

Output: A tree fitting this distance matrix

Note: A weighted unrooted tree T fits a distance matrix D if $d_{i,j}(T) = D_{i,j}$ for every pair of leaves i and j

Distance matrix

Distance matrix

- Symmetric (for all i,j $D_{i,j} = D_{j,i}$)
- Non-negative
- satisfy triangle's inequality (for all i, j, and k, $D_{i,j} + D_{j,k} >= D_{i,k}$)

	Carp	Zebrafish	Salmon	Trout
Carp	0	3	7	9
Zebrafish		0	6	8
Salmon			0	6
Trout				0

Construct a phylogenetic tree

	Carp	Zebrafish	Salmon	Trout
Carp	0	3	7	9
Zebrafish		0	6	8
Salmon			0	6
Trout				0

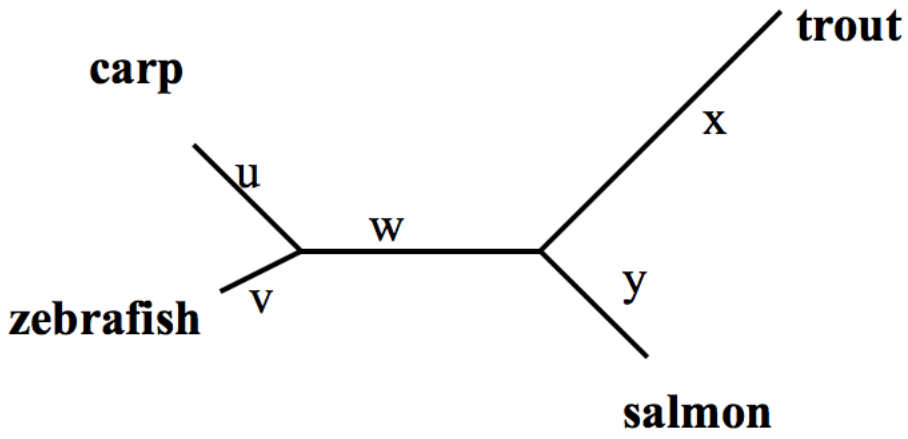
Observed distances

- Tree topology?
- Branch lengths?

Given a tree topology, estimating branch lengths is easy…

	Carp	Zebrafish	Salmon	Trout
Carp	0	3	7	9
Zebrafish		0	6	8
Salmon			0	6
Trout				0

Observed distances



Match distance matrix to branch lengths

Match distance matrix to branch lengths

	Carp	Zebrafish	Salmon	Trout
Carp	0	3	7	9
Zebrafish		0	6	8
Salmon			0	6
Trout				0

Observed distances

$u + v = 3$

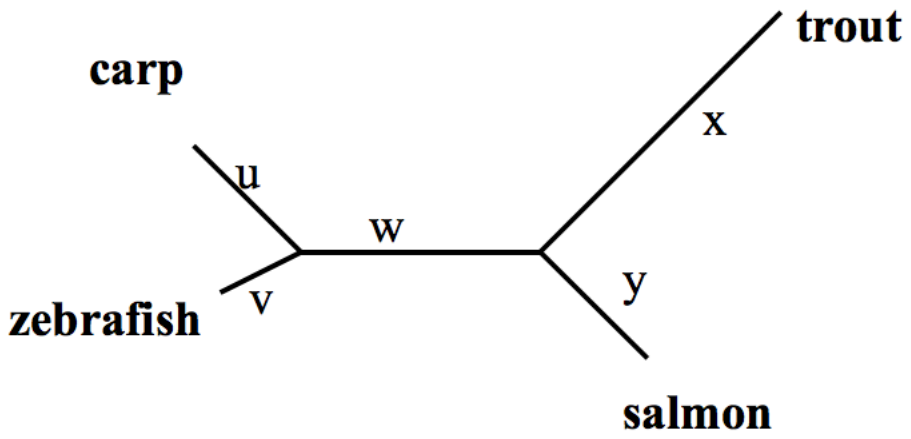
$u + w + y = 7$

$u + w + x = 9$

$v + w + y = 6$

$v + w + x = 8$

$x + y = 6$



Match distance matrix to branch lengths

	Carp	Zebrafish	Salmon	Trout
Carp	0	3	7	9
Zebrafish		0	6	8
Salmon			0	6
Trout				0

Observed distances

$u + v = 3$

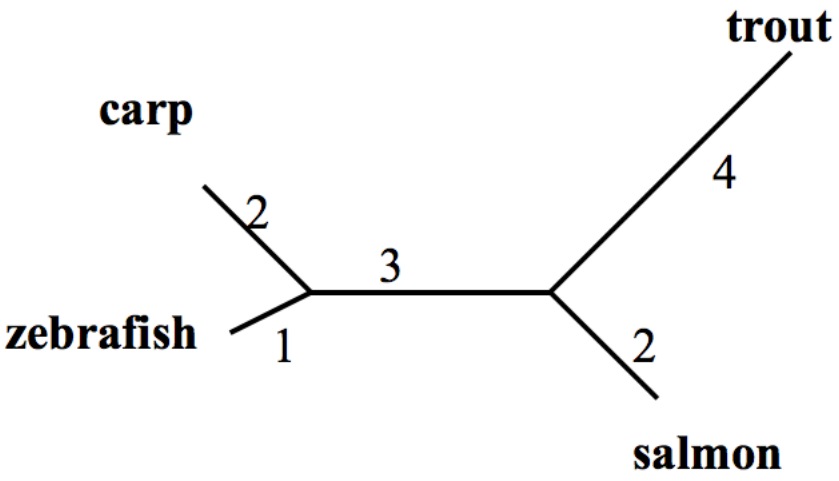
$u + w + y = 7$

$u + w + x = 9$

$v + w + y = 6$

$v + w + x = 8$

$x + y = 6$



Can every matrix be fitted to a tree?

	Carp	Zebrafish	Salmon	Trout
Carp	0	3	7	9
Zebrafish		0	6	8
Salmon			0	6
Trout				0

Observed distances

$u + v = 3$

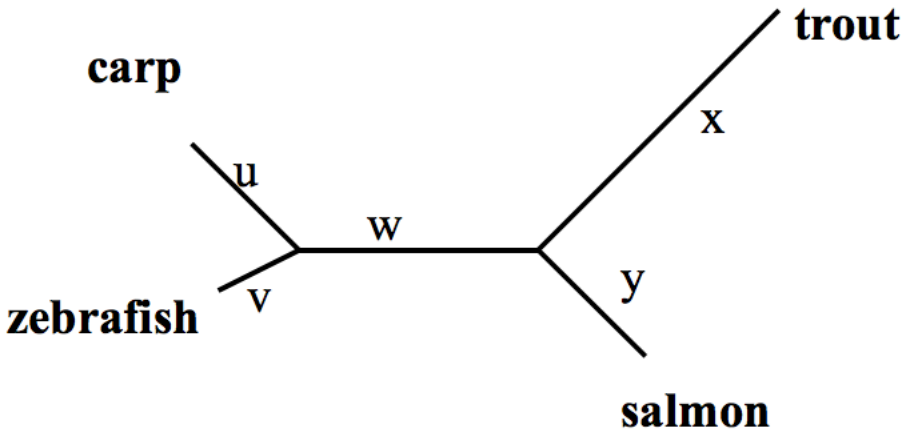
$u + w + y = 7$

$u + w + x = 9$

$v + w + y = 6$

$v + w + x = 8$

$x + y = 6$



Can every matrix be fitted to a tree? NO!!

	Carp	Zebrafish	Salmon	Trout
Carp	0	3	7	9
Zebrafish		0	6	8
Salmon			0	6
Trout				0

Observed distances

If I have a tree topology, I can verify that the distance matrix can be fitted to a tree **if and only if** the equations have a solution.

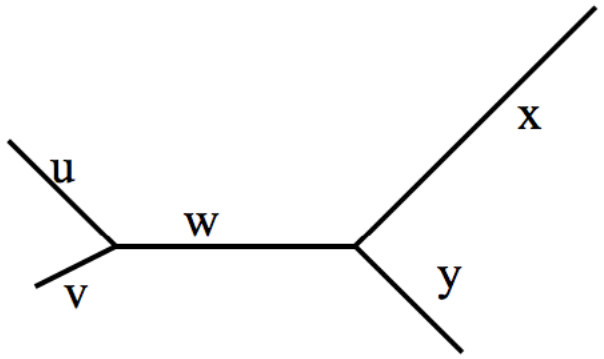
If I **don't** have a tree topology, how can I know I can verify that the distance matrix can be fitted to a tree? Distance matrix needs to be **additive.**

A matrix is additive if and only if it satisfies the four point condition.

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

Four point condition:

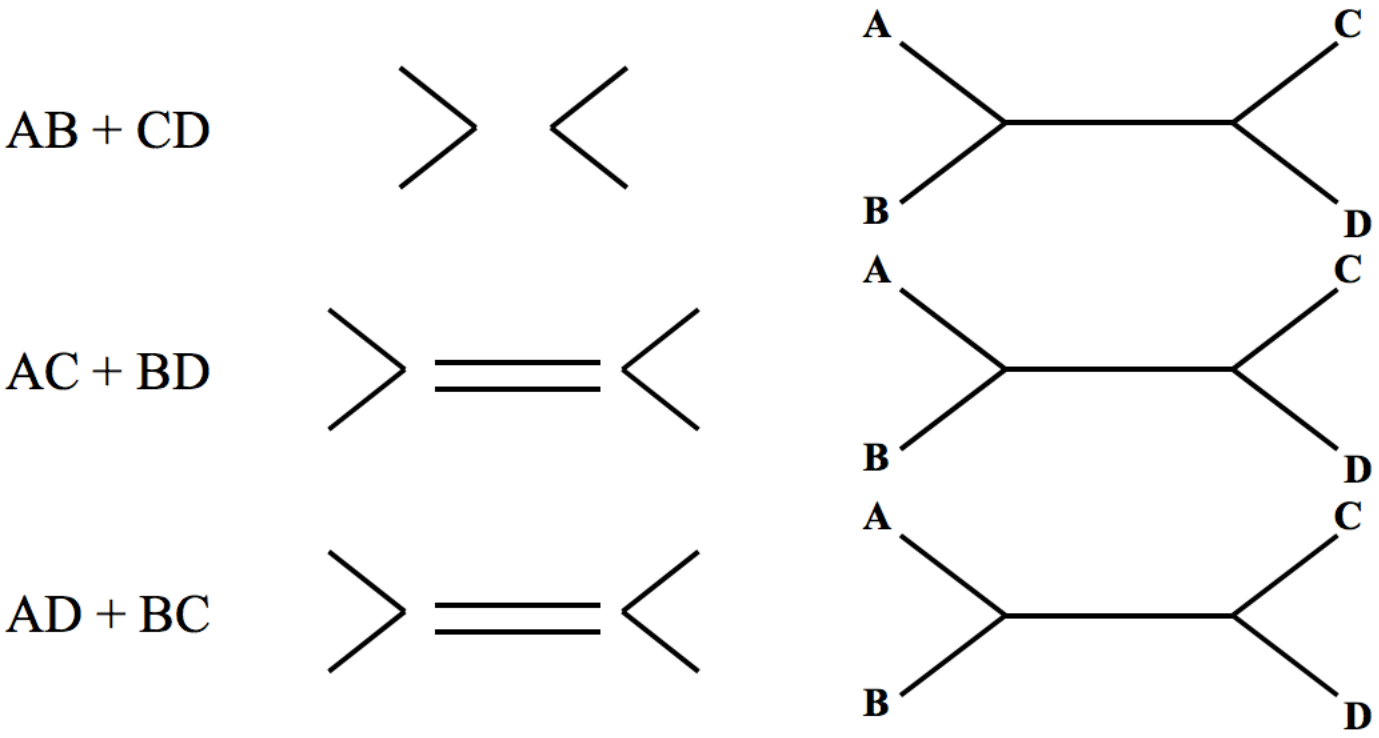
```
AB+CD <= max(AC+BD,AD+BC)
AC+BD <= max(AB+CD,AD+BC)
AD+BC <= max(AC+BD,AB+CD)
```



which generalizes the familiar triangle inequality (take C = D).

A matrix is additive if and only if it satisfies the four point condition.

```
AB+CD <= max(AC+BD,AD+BC)
AC+BD <= max(AB+CD,AD+BC)
AD+BC <= max(AC+BD,AB+CD)
```

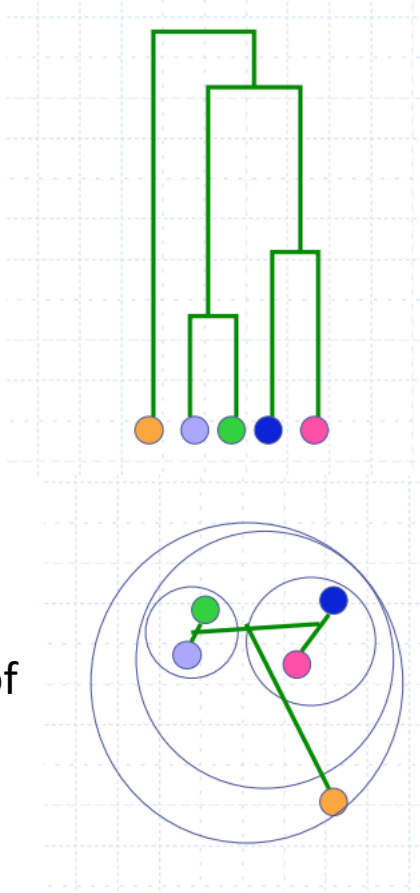


Does this matrix satisfy four point condition?

	Carp	Zebrafish	Salmon	Trout
Carp	0	3	7	9
Zebrafish		0	6	8
Salmon			0	6
Trout				0

	i	*j*	*k*	*l*
i	0	13	21	22
j	13	0	12	13
k	21	12	0	13
l	22	13	13	0

Unweighted Pair Group Method using Arithmetic averages (UPGMA)



- The UPGMA algorithm is a variant of average linkage
- UPGMA is based on the molecular clock assumption

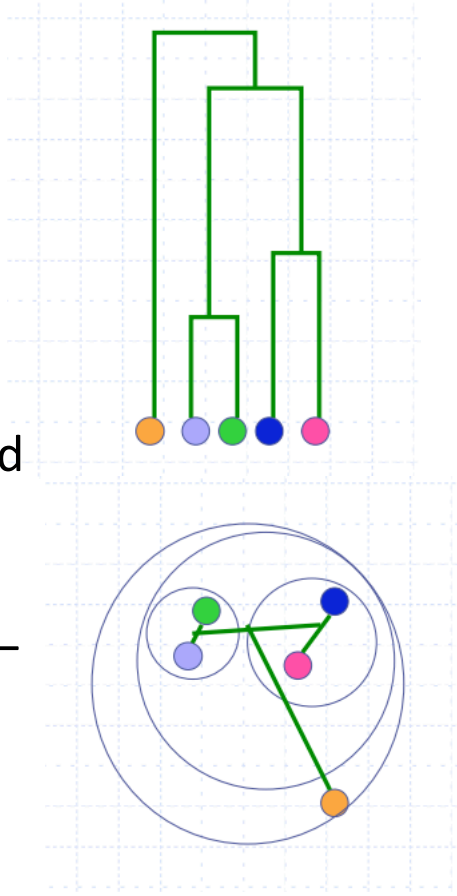
- The consequences of this assumption are that
  - At each step, the two closest taxa are selected as neighbors
  - The height of the least common ancestor of any pair of leaves is half the distance between the leaves
  - It assumes an ultrametric tree in which the distances from the root to every branch tip are equal

$$D(cl_1, cl_2) = \frac{1}{|cl_1||cl_2|} \sum_{p \in cl_1, q \in cl_2} D(p, q)$$

Construct a UPGMA tree

	i	*j*	*k*	*l*
i	0	13	21	22
j	13	0	12	13
k	21	12	0	13
l	22	13	13	0

Unweighted Pair Group Method using Arithmetic averages (UPGMA)



- UPGMA is based on the molecular clock assumption
- Key element – must be able to quickly compute distance between clusters (internal nodes) – weighted distance
- Note that UPGMA does not estimate branch lengths – they are all assumed equal

Read about Ultrametric matrix and three point condition
If a distance matrix, D, is ultrametric, then UPGMA will reconstruct the correct rooted tree in quadratic time.

Neighbor joining (NJ) algorithm

- Neighbor joining heuristics: join closest clusters that are far from the rest
- distance between two sequences is not sufficient – must also know how each sequence compares to every other sequence
- Statistically consistent
- Does not depend on molecular clock assumption
- Heavily used in practice [e.g., Clustal W]
- If D is additive, then NJ will reconstruct the correct unrooted tree in quadratic time.
- But can be sensitive to non-additivity

Limitations of Distance based phylogeny estimation approach

- Calculating distance matrix can be challenging
- Loss of information contained in the alignment
- Cannot obtain ancestral sequences (sequences at the internal nodes of the tree)

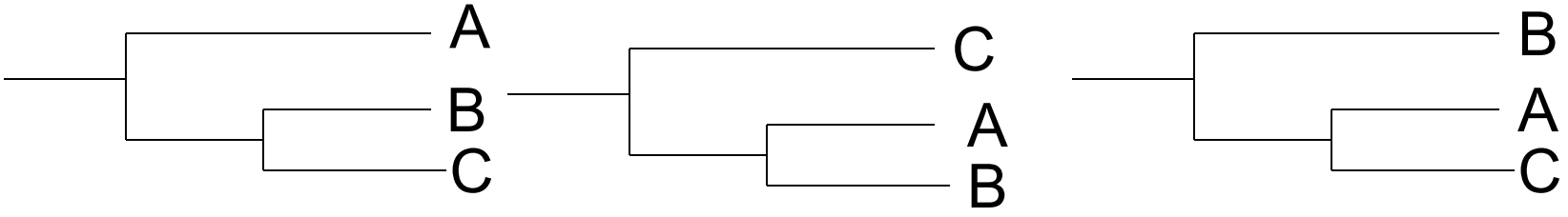
Maximum Parsimony (character based phylogeny)

Phylogeny questions

•Given several organisms & a set of features (usually sequence, but also morphological: wing shape/color...)

•A. Given a phylogenetic tree – figure out what the ancestors looked like (what are the features of internal nodes)

•B. Find the phylogenetic tree that best describes the common evolutionary heritage of the organisms



Phylogeny questions

•Given several organisms & a set of features (usually sequence, but also morphological: wing shape/color...)

•A. Given a phylogenetic tree – figure out what the ancestors looked like (what are the features of internal nodes)

- Taxa are considered as sets of attributes: characters
- “character” = DNA position, genes order, morphological feature…
- “character state” = a value assumed by a character
- Characters evolve through state changes
- Evolutionary tree represents changes in character states
- MP-tree seeks to minimize state changes

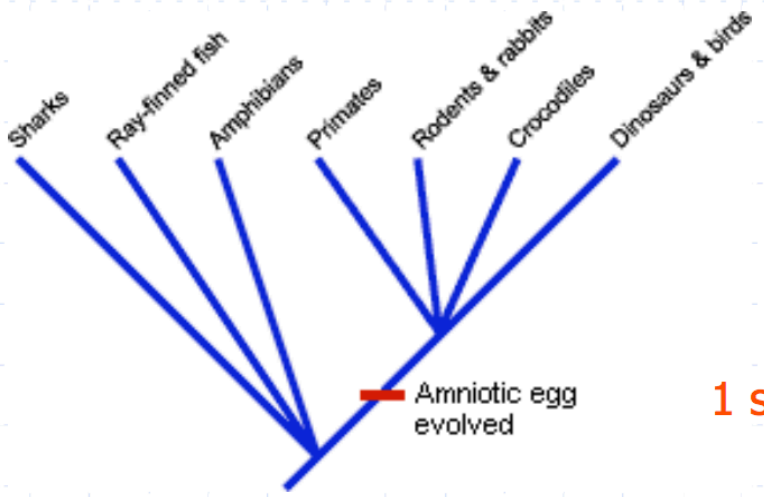
Example

	Vertebrae?	Bony skeleton?	Four limbs?	Amniotic egg?	Hair?	Two post-orbital fenestrae?
Sharks and relatives	YES	no	no	no	no	no
Ray-finned fishes	YES	YES	no	no	no	no
Amphibians	YES	YES	YES	no	no	no
Primates	YES	YES	YES	YES	YES	no
Rodents and rabbits	YES	YES	YES	YES	YES	no
Crocodiles and relatives	YES	YES	YES	YES	no	YES
Dinosaurs and birds	YES	YES	YES	YES	no	YES

Characters

Binary states

Taxa

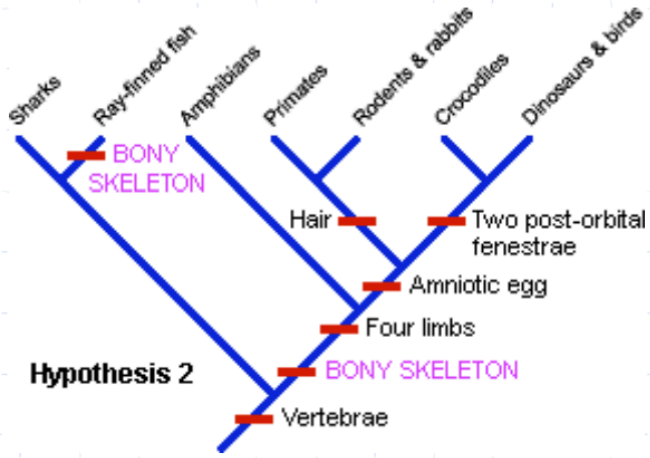


1 state change

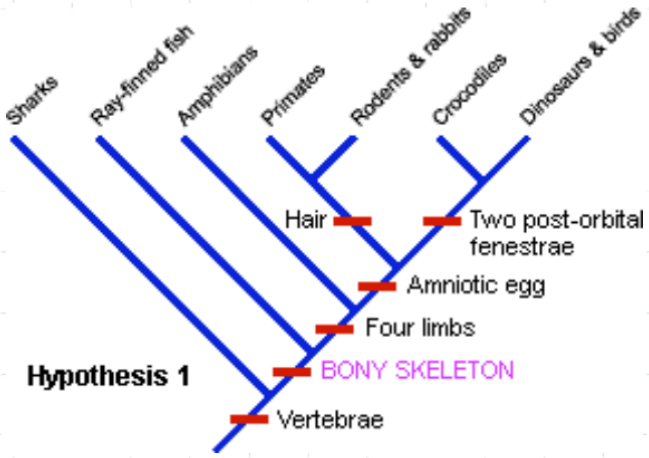
Example

	Vertebrae?	Bony skeleton?	Four limbs?	Amniotic egg?	Hair?	Two post-orbital fenestrae?
Sharks and relatives	YES	no	no	no	no	no
Ray-finned fishes	YES	YES	no	no	no	no
Amphibians	YES	YES	YES	no	no	no
Primates	YES	YES	YES	YES	YES	no
Rodents and rabbits	YES	YES	YES	YES	YES	no
Crocodiles and relatives	YES	YES	YES	YES	no	YES
Dinosaurs and birds	YES	YES	YES	YES	no	YES

7 state changes

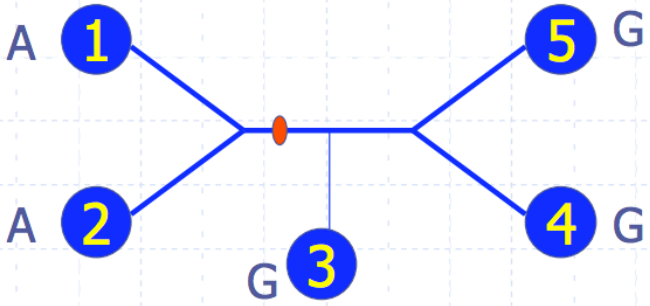


6 state changes

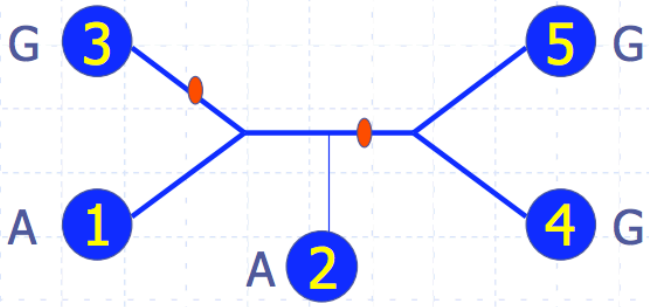


Example

- Consider {ATA,ATT, GTT, GTA, GGT}
  - First column admits 2 arrangements & identifies likely mutation



MP (1 mutation)



2 mutations

ATA
ATT
GTT
GTA
GGT

Scoring a tree – Sankoff's algorithm

•Assumption – we try to minimize # of state changes from root to leaves – Parsimony approach

•Small parsimony

–given a tree where leaves are labeled with m-character strings

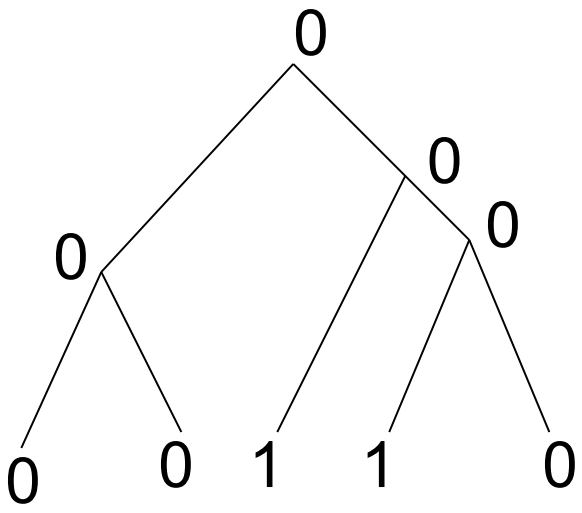
–find labels at internal nodes s.t. # of state transitions is minimzed

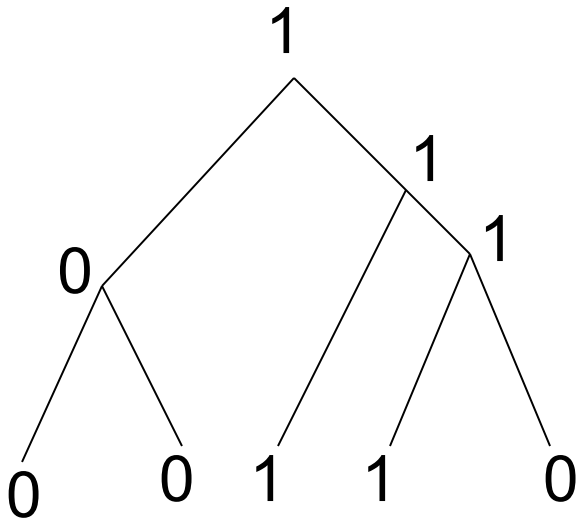
•Weighted small parsimony

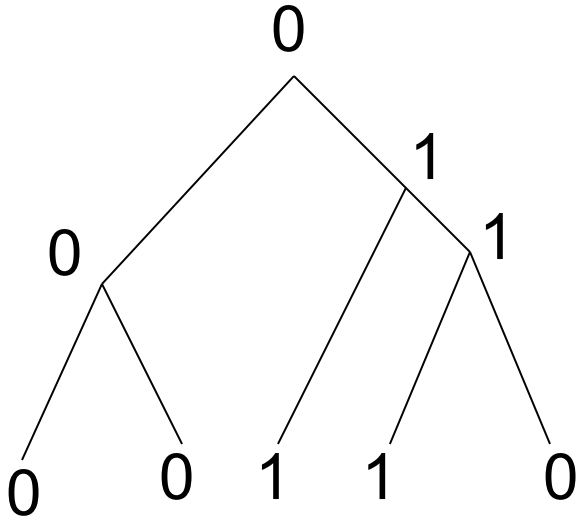
–same as parsimony except that state transitions are assigned weights

–minimize the overall weight of the tree

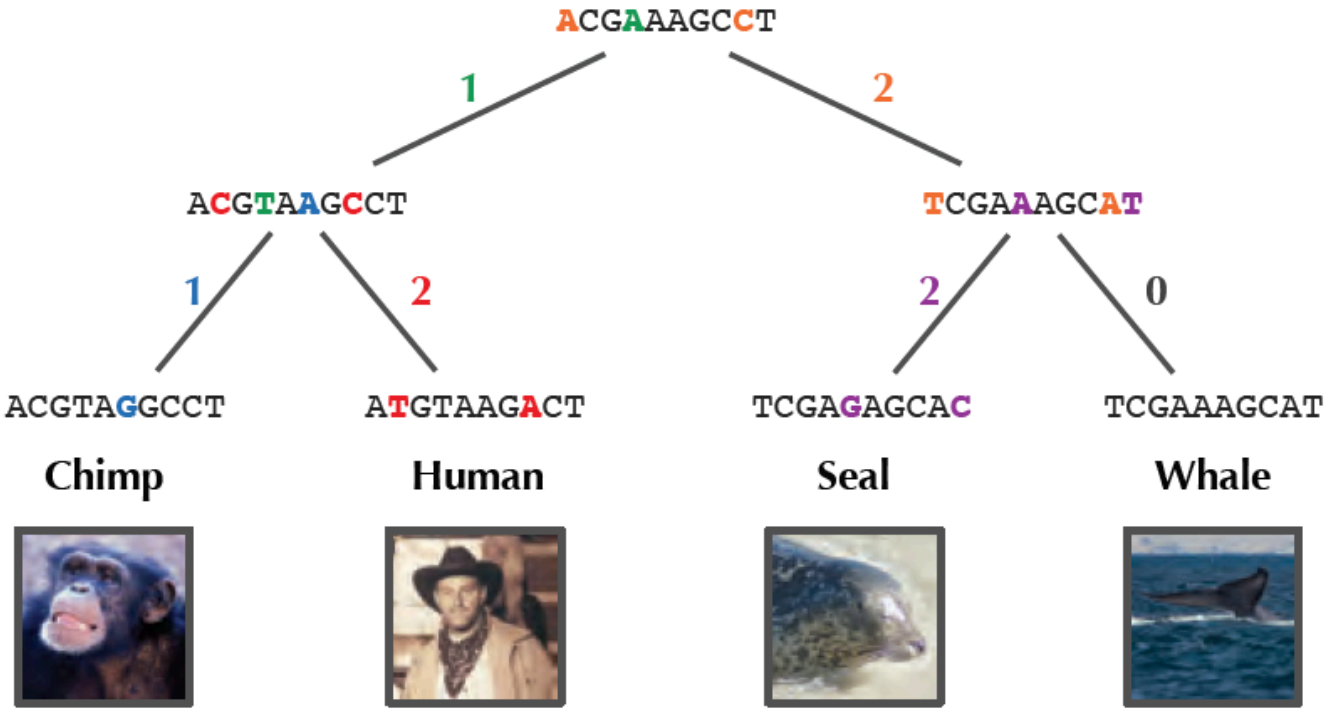
Example



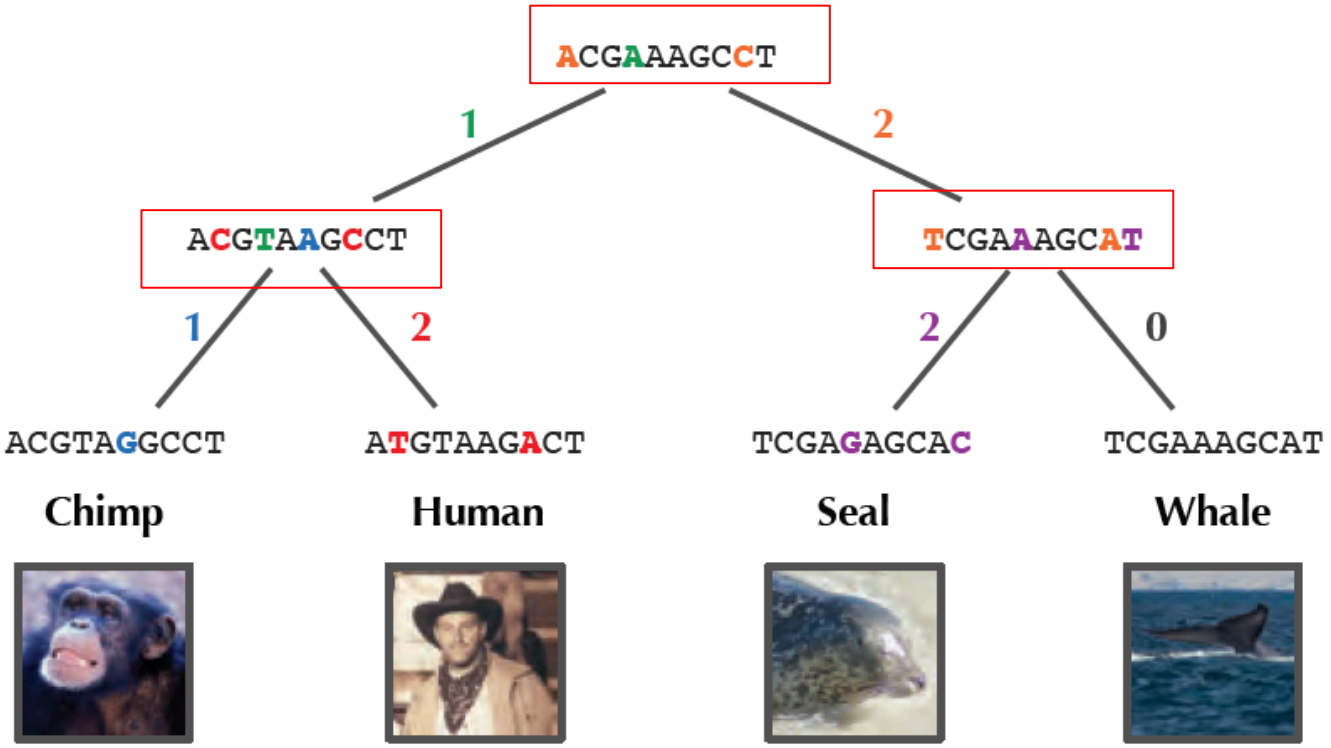




Scoring a tree – Sankoff's algorithm



Scoring a tree – Sankoff's algorithm

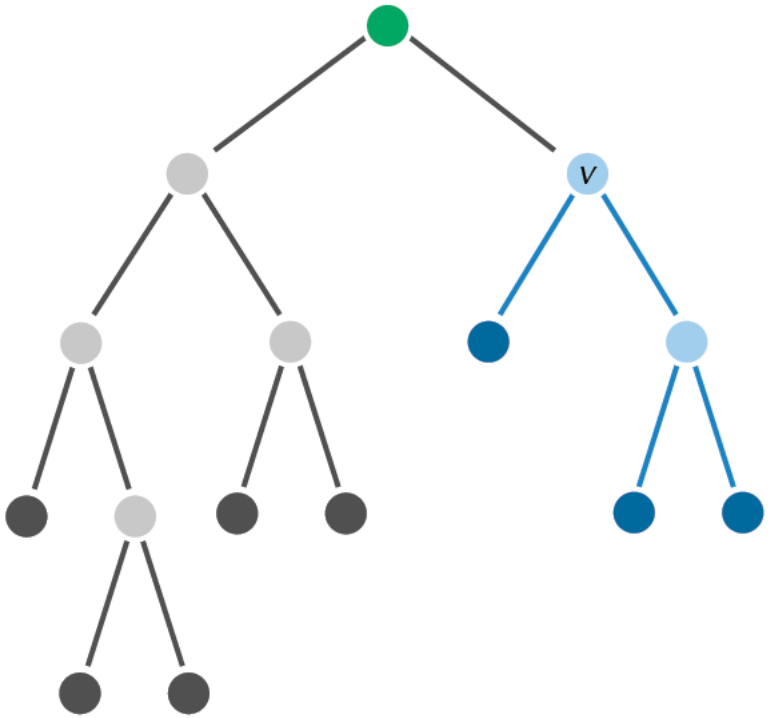


Make a simplifying assumption – all characters are independent in the sequence
i.e. Run separately for each character then merge results

Sankoff's algorithm – recurrence relation

- At each node v in the tree store s(v,t) – best parsimony score for subtree rooted at v if character stored at v is t

How to get a Base case?

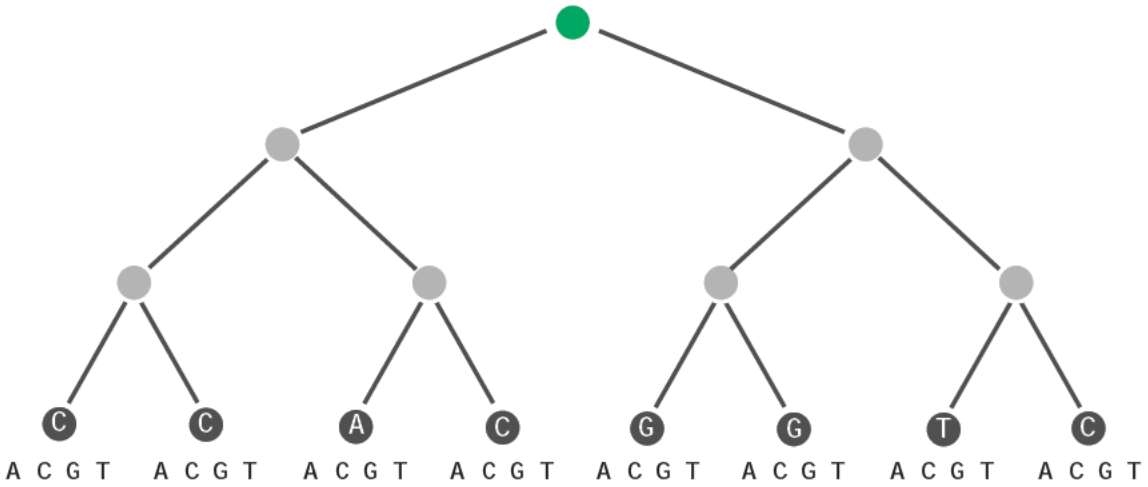


The (blue) subtree T_v of a node v within a larger rooted binary tree T.

Sankoff's algorithm – recurrence relation

•At each node v in the tree store s(v,t) – best parsimony score for subtree rooted at v if character stored at v is t

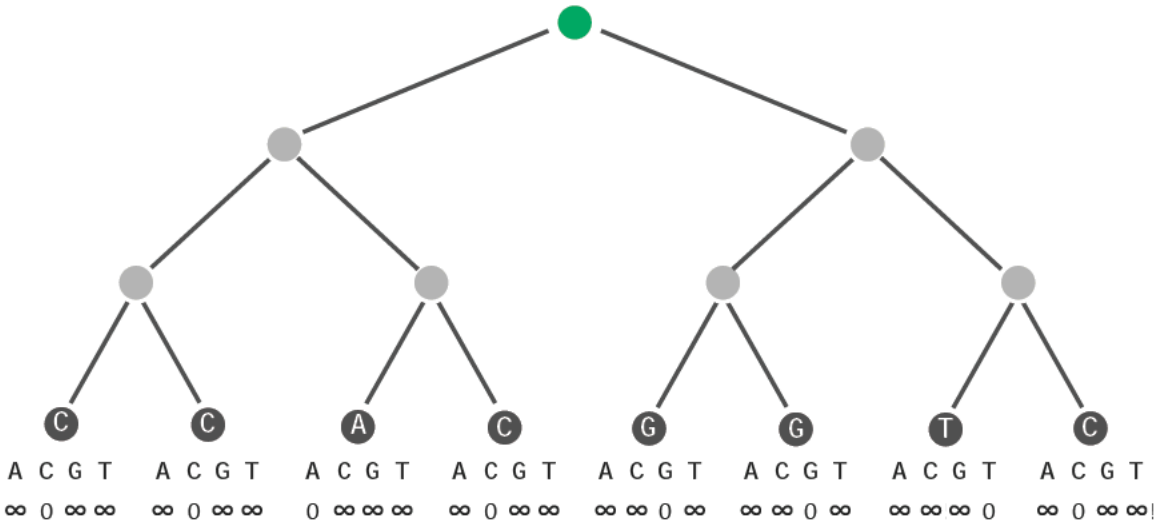
How to get a Base case?



Sankoff's algorithm – recurrence relation

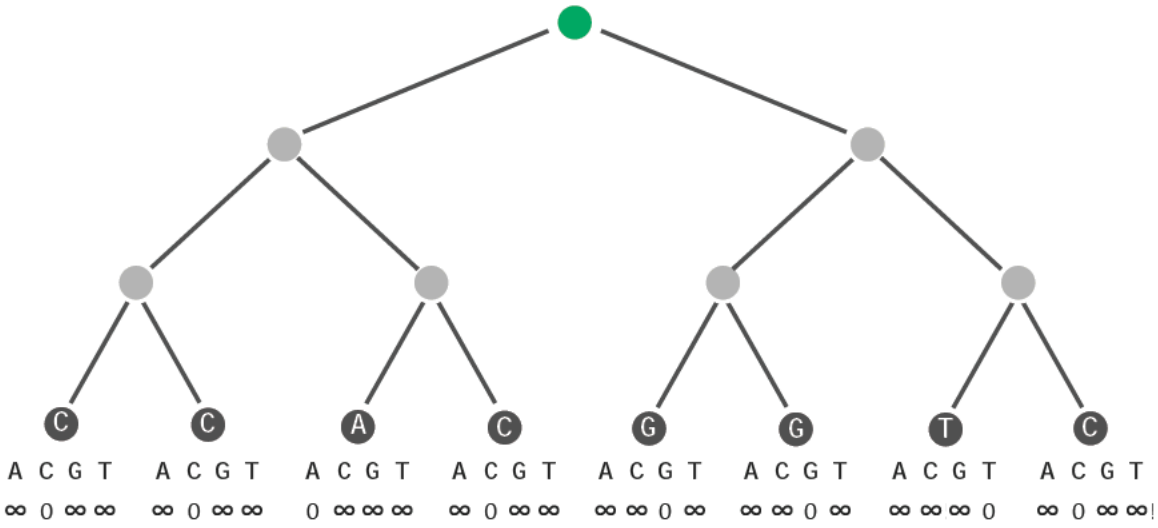
- At each node v in the tree store s(v,t) – best parsimony score for subtree rooted at v if character stored at v is t

How to get a Base case?



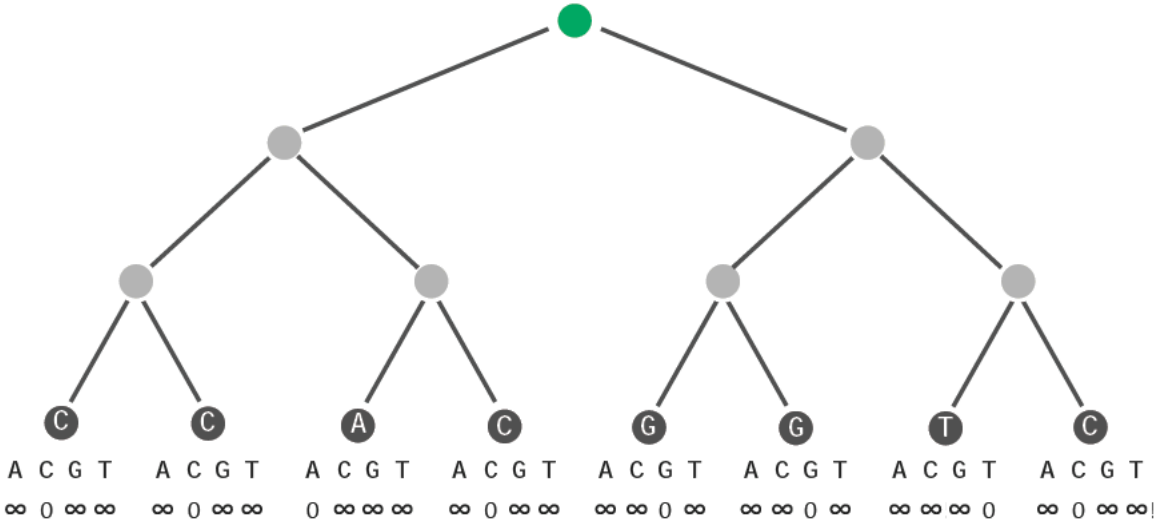
Sankoff's algorithm – recurrence relation

- At each node v in the tree store s(v,t) – best parsimony score for subtree rooted at v if character stored at v is t
- Traverse the tree in post-order and update s(v,t) as follows



Sankoff's algorithm – recurrence relation

- At each node v in the tree store s(v,t) – best parsimony score for subtree rooted at v if character stored at v is t
- Traverse the tree in post-order and update s(v,t) as follows
  - assume node v has children u and w
  - $s(v,t) = \min_i \{s(u,i) + score(i,t)\} + \min_j \{s(w,j) + score(j,t)\}$



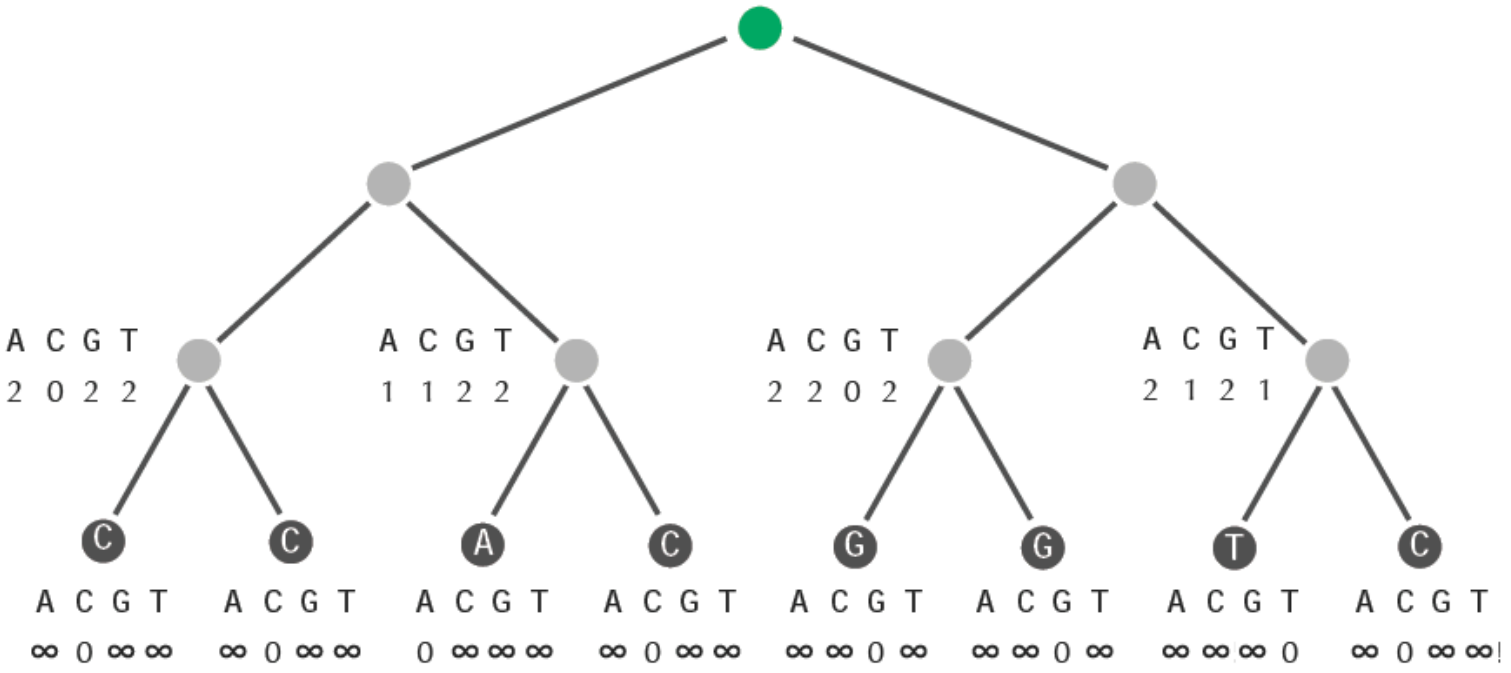
Sankoff's algorithm – recurrence relation

- At each node v in the tree store s(v,t) – best parsimony score for subtree rooted at v if character stored at v is t
- Traverse the tree in post-order and update s(v,t) as follows
  - assume node v has children u and w
  - $s(v,t) = \min_i \{s(u,i) + score(i,t)\} + \min_j \{s(w,j) + score(j,t)\}$

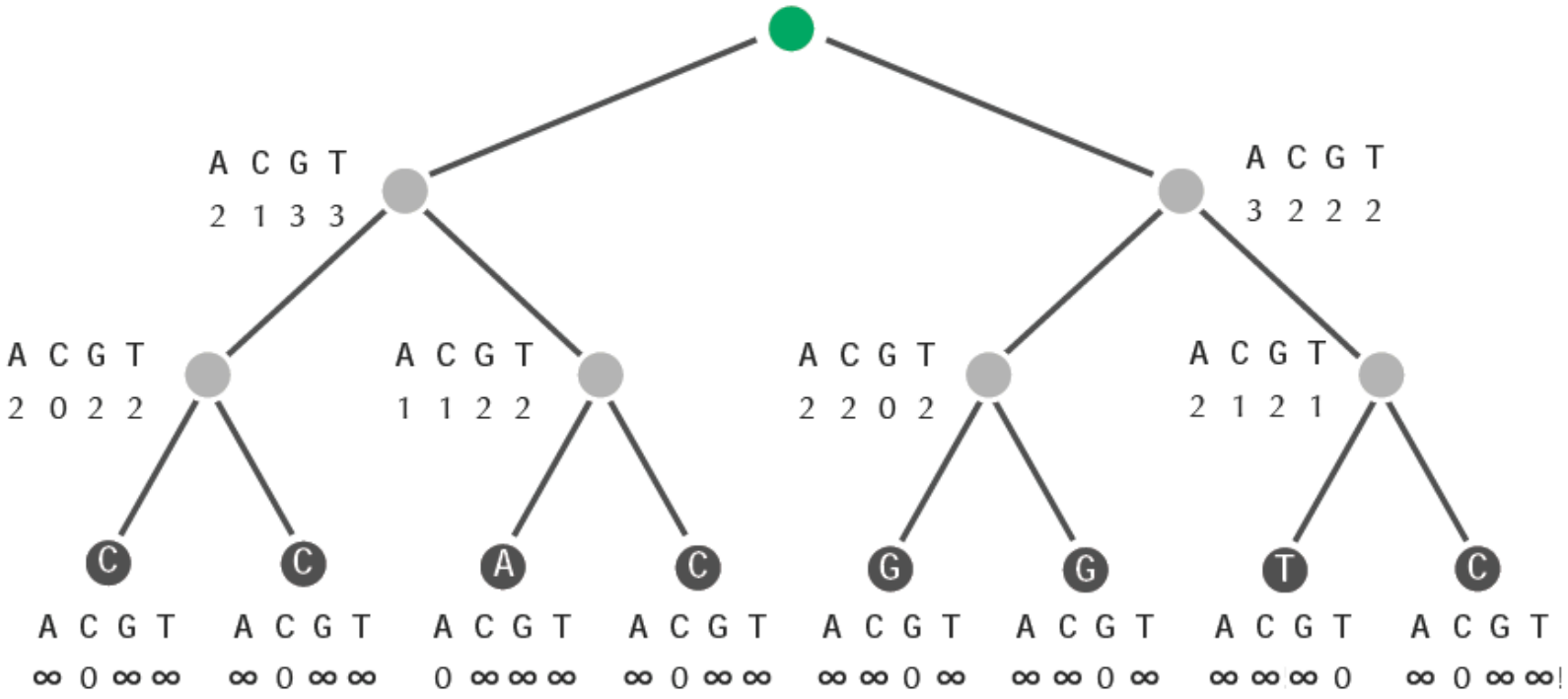
- the minimum parsimony score is given by the smallest score s(root,t) over all symbols t

- Note – this solves the weighted version. For unweighted set score (i,i) = 0, score(i,j) = 1 for any i,j

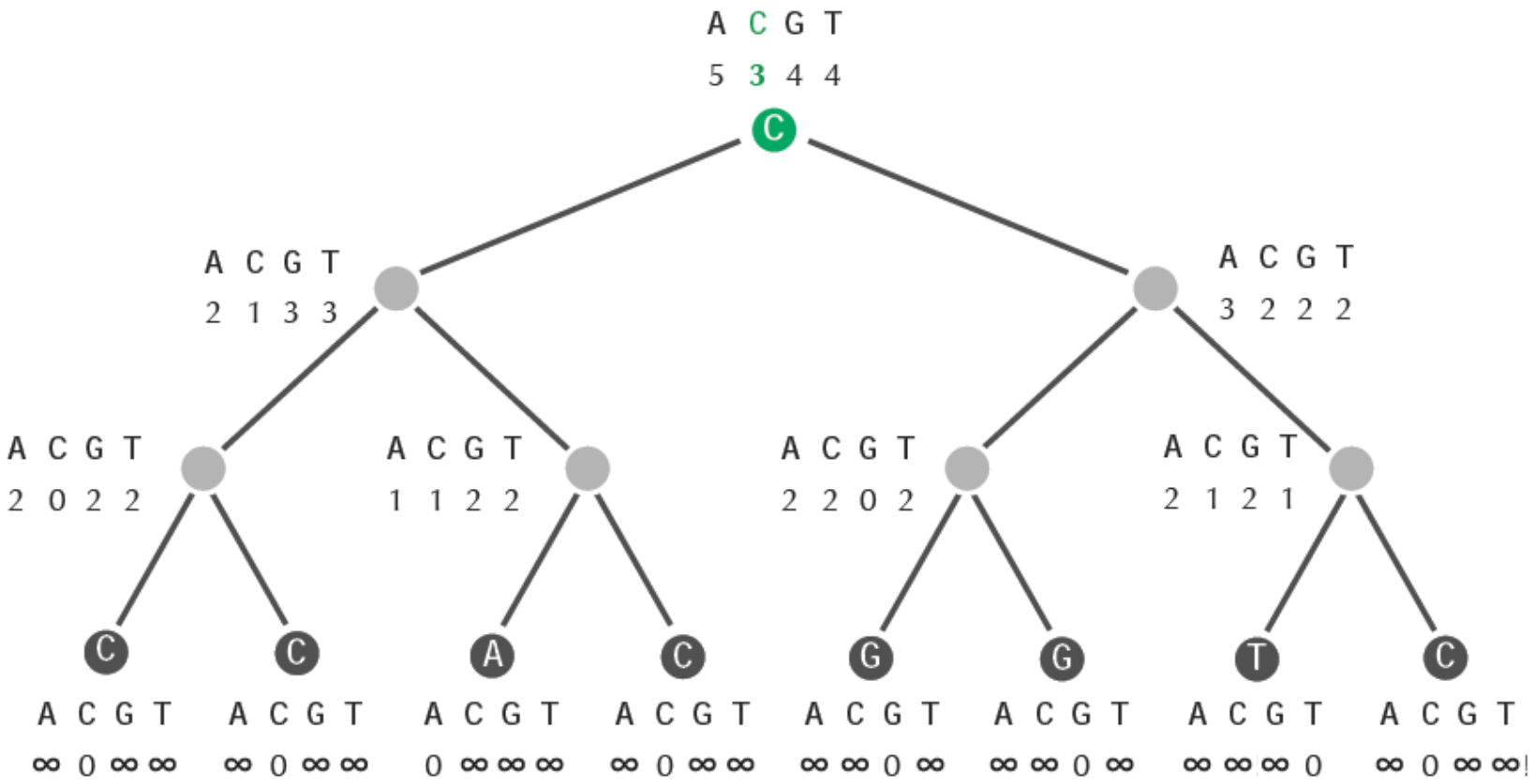
Sankoff's algorithm – example (continued)



Sankoff's algorithm – example (continued)



Sankoff's algorithm – example (continued)

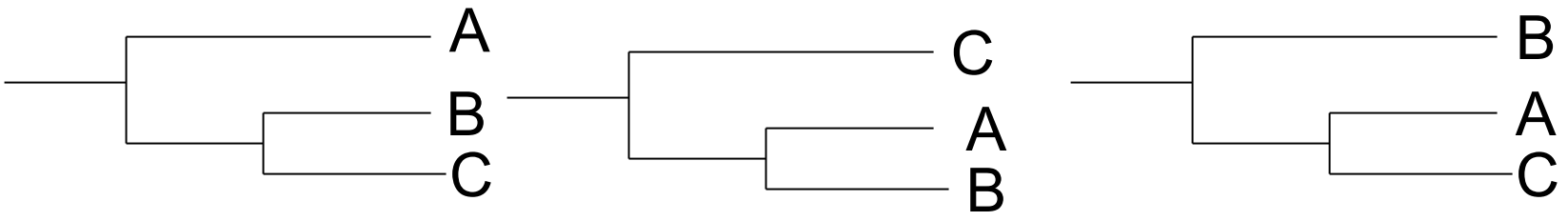


Optimal labeling can be computed in linear time O(nk) where n is number of leaves and k is number of character states

Phylogeny questions

•Given several organisms & a set of features (usually sequence, but also morphological: wing shape/color...)

B. Find the phylogenetic tree that best describes the common evolutionary heritage of the organisms

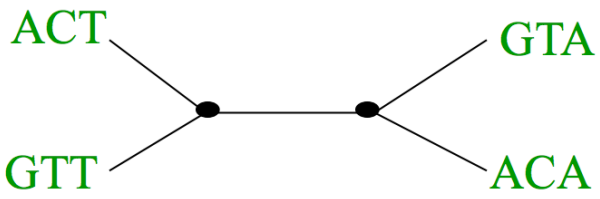


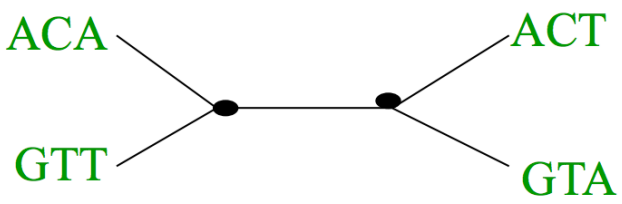
- Finding the optimal Maximum parsimonius tree is NP-hard
- Exponential number of trees
- Heuristics to bound the search space
  - Branch and bound
  - Nearest neighbor interchange (NNI) - switch subtrees
  - Prune scoring a tree if the score exceeds best score of a fully explored tree

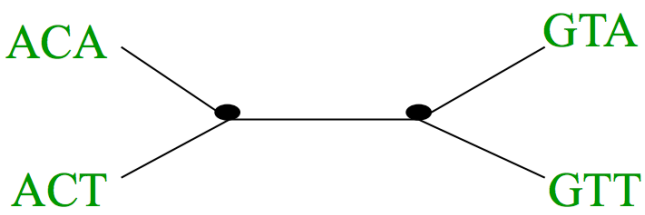
Maximum parsimony (example)

- **Input**: Four sequences
  - ACT
  - ACA
  - GTT
  - GTA
- **Question**: which of the three trees has the best MP scores?

Maximum Parsimony







Maximum likelihood

- For every branch S->T of length t, compute P(T|S,t) – likelihood that sequence S could have evolved in time t into sequence T
- Find tree that maximizes the likelihood

- Note that likelihood of a tree can be computed with an algorithm similar to Sankoffs
- However, no simple way to find a tree given the sequences – most approaches use heuristic search techniques
- Often, start with NJ tree – then "tweak" it to improve likelihood

Thanks!

Contact:
nidhi@cs.umd.edu

References

- http://www.cs.columbia.edu/4761/notes07/chapter3.2-phylo.pdf
- http://people.cs.uchicago.edu/~ridg/digbio08/talkaddree.pdf
- https://phys.org/news/2016-12-evolutionary-tree.html
- https://www.khanacademy.org/science/high-school-biology/hs-evolution/hs-phylogeny/a/phylogenetic-trees
- http://www.cs.cmu.edu/~durand/03-711/2011/Lectures/Trees11-3.pdf
- http://www.cs.cmu.edu/~durand/03-711/2011/Lectures/Trees11-4.pdf
- http://tandy.cs.illinois.edu/MaximumParsimony-598.pdf
- http://www.bioinf.uni-leipzig.de/~steigele/phylogeny.pdf