CMSC423: Bioinformatic Algorithms, Databases and Tools

Phylogenetic trees

What is a phylogenetic tree?

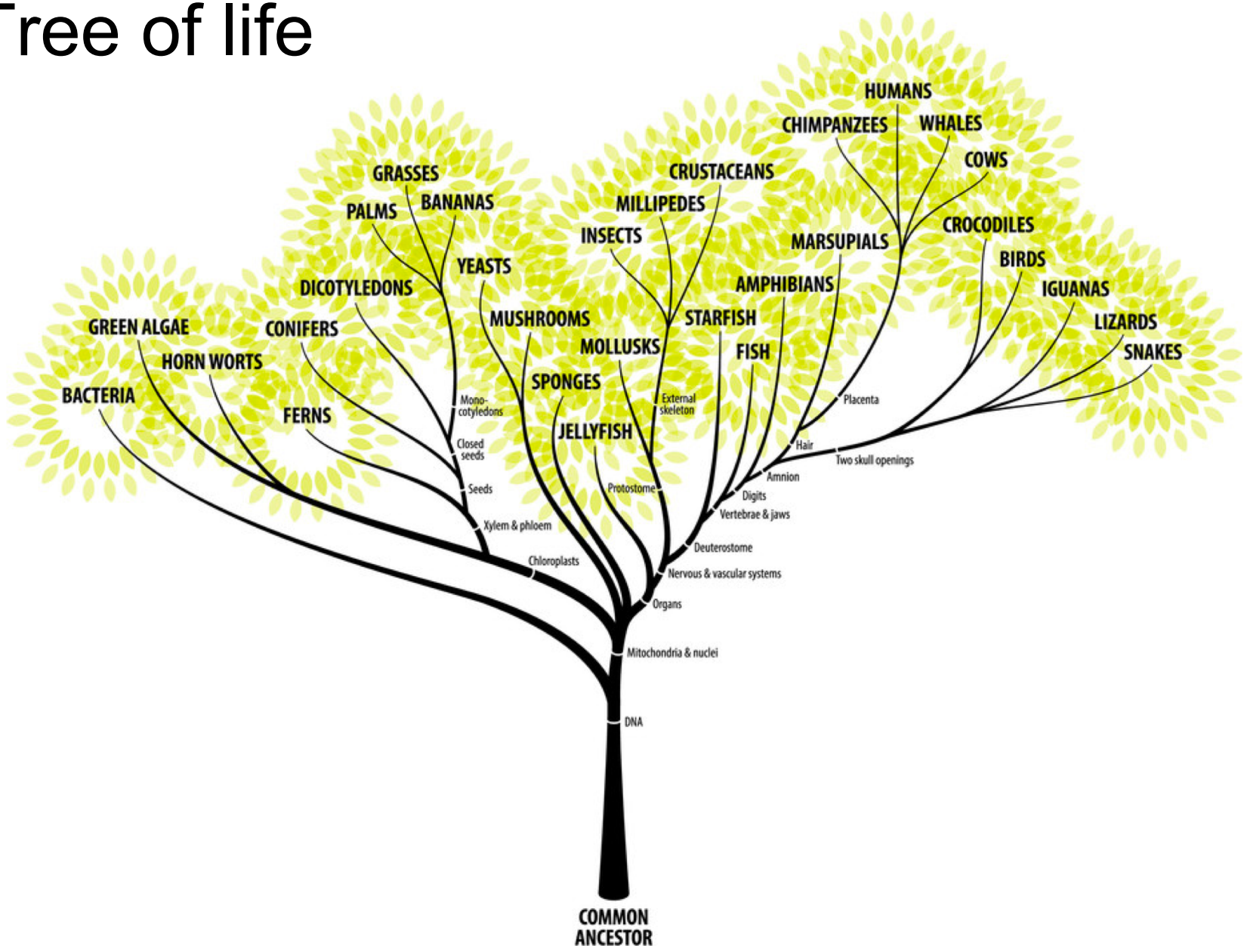
- A phylogenetic tree (evolutionary tree) is a branching diagram showing the inferred evolutionary relationships among various biological species or other entities

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

- The taxa 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

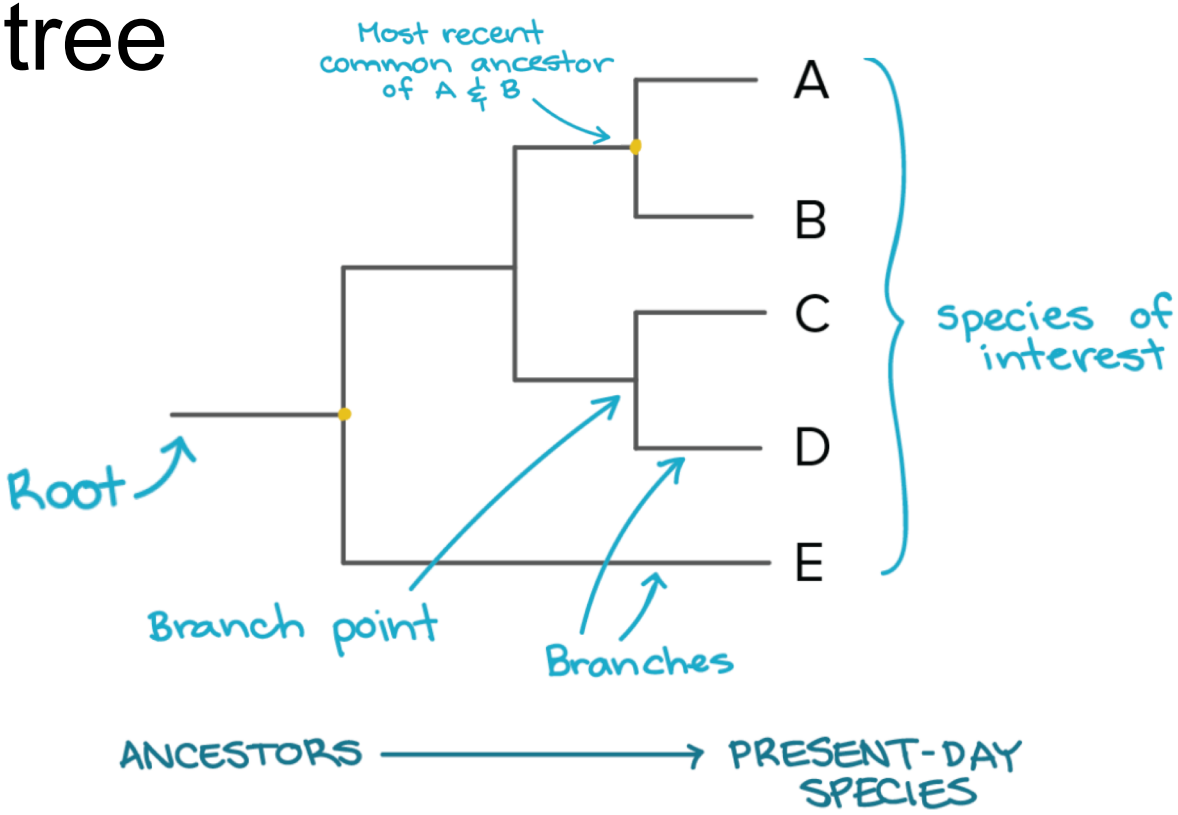

HUMANS
CHIMPANZEES
WHALES
COWS
CRUSTACEANS
GRASSES
MILLIPEDES
PALMS
BANANAS
INSECTS
CROCODILES
MARSUPIALS
BIRDS
YEASTS
AMPHIBIANS
IGUANAS
DICOTYLEDONS
MUSHROOMS
STARFISH
LIZARDS
GREEN ALGAE
CONIFERS
MOLLUSKS
FISH
SNAKES
HORN WORTS
SPONGES
BACTERIA
Mono-cotyledons
External skeleton
Placenta
FERNS
JELLYFISH
Hair
Closed seeds
Two skull openings
Amnion
Seeds
Protostome
Digits
Vertebrae & jaws
Xylem & phloem
Deuterostome
Chloroplasts
Nervous & vascular systems
Organs
Mitochondria & nuclei
DNA
COMMON ANCESTOR

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

Purpose of Phylogenetic tree

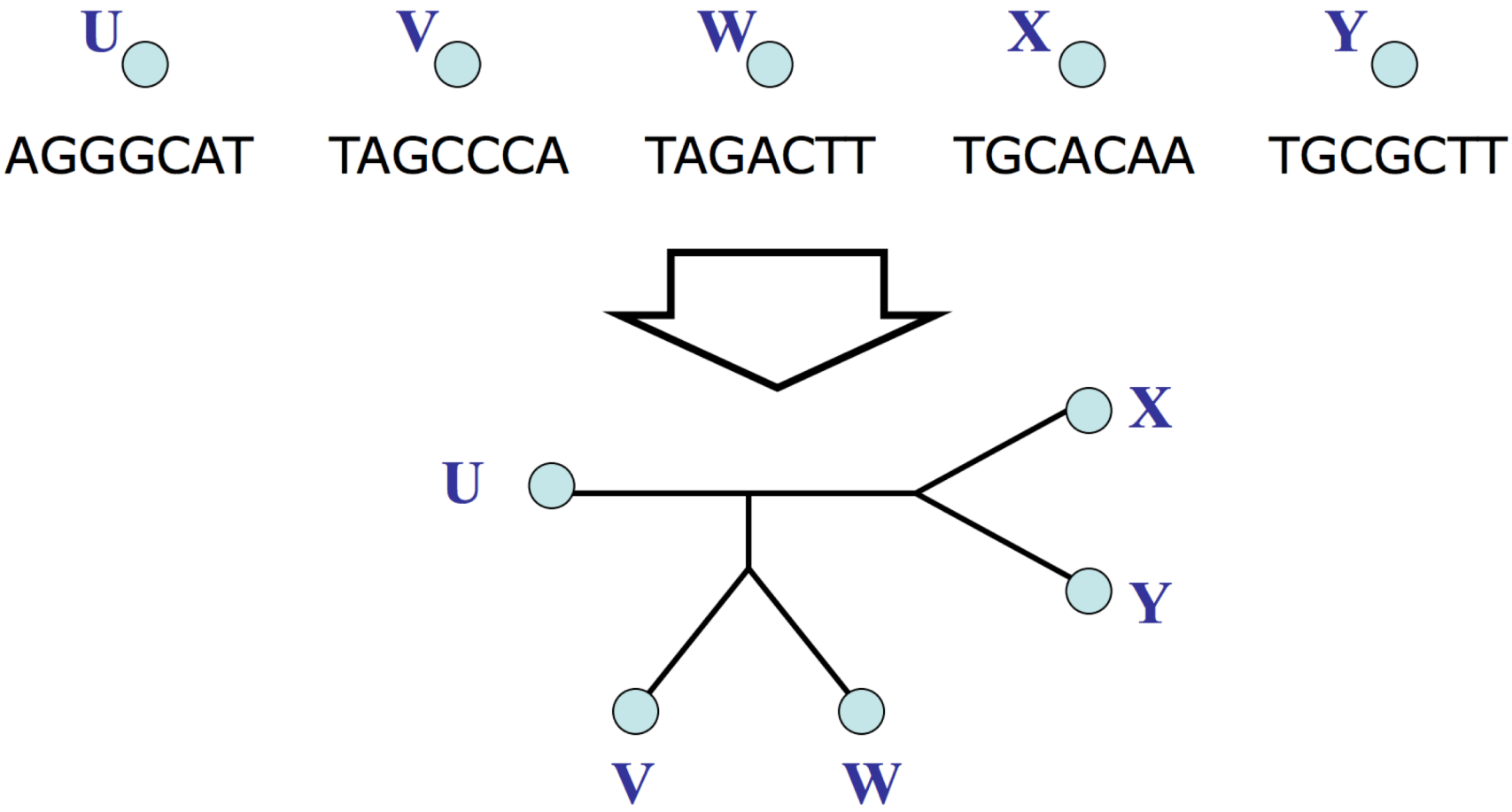
- Understanding human origin
- Understanding biogeography
- Understanding origin of particular traits
- Understanding the process of molecular evolution
- Origin of disease
- Forensics
- Other use cases – evolution of languages, cancer tumor evolution, etc.

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.

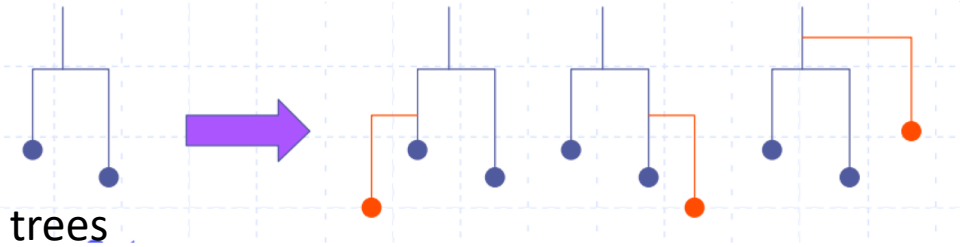
Phylogeny problem



Brute Force approach

- Brute Force
  - Enumerate all trees
  - Compute some measure of evolutionary likelihood
  - Select the best tree
- How many rooted trees are there with n leaves?
  - N=2 leaves → 1 tree
  - N=3 leaves

    attach 3rd leaf to 3 edges → 3 trees



- Let T(n) = # rooted trees with n leaves, E(n) = # of edges
  - T(2)=1, E(2)=3; T(3)=3, E(3)=5
  - Addition of a leaf creates two new edges => E(n)=E(n-1)+2=> E(n)=2n-1
  - T(n)=T(n-1)*E(n-1)=T(n-1)*(2n-3) => T(n)= 1*3*5*…(2n-3)
  - For n=20 leaves ~ 10^{21}

Distance based phylogeny (UPGMA and neighbor joining)

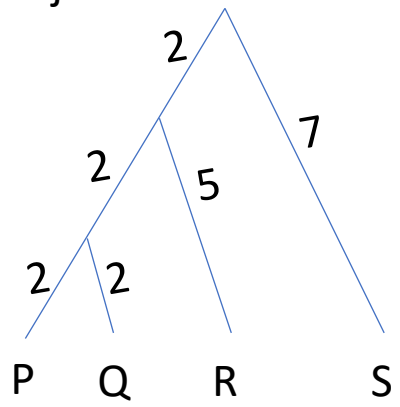
Distance based methods - Review

- UPGMA
- Neighbor joining

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

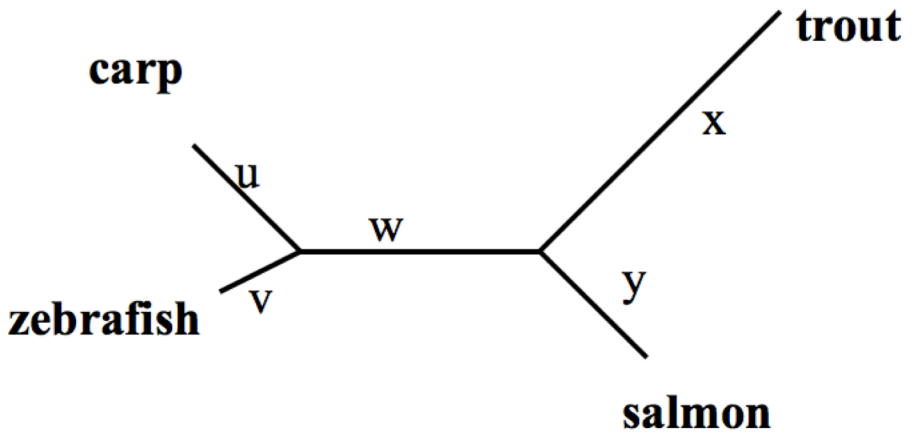


Distance matrix	P	Q	R	S
P	0	4	9	13
Q	4	0	9	11
R	9	9	0	14
S	13	11	14	0

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



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

$u + v = 3$

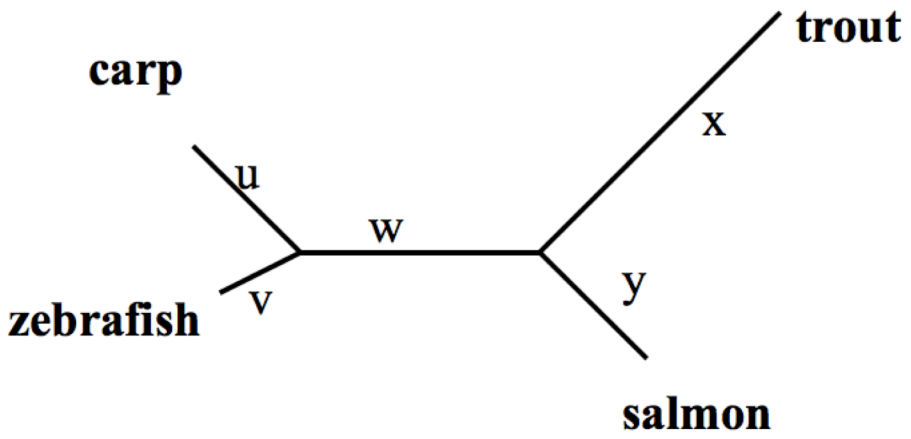
$u + w + y = 7$

$u + w + x = 9$

$v + w + y = 6$

$v + w + x = 8$

$x + y = 6$



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

$u + v = 3$

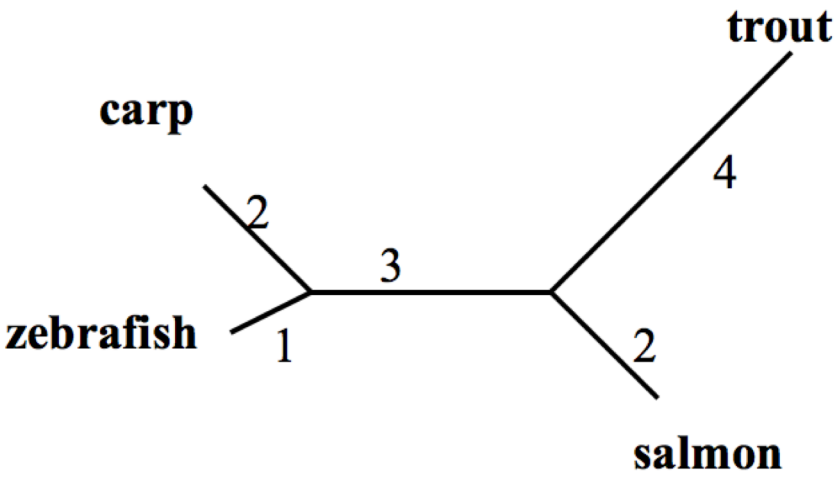
$u + w + y = 7$

$u + w + x = 9$

$v + w + y = 6$

$v + w + x = 8$

$x + y = 6$



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

$u + v = 3$

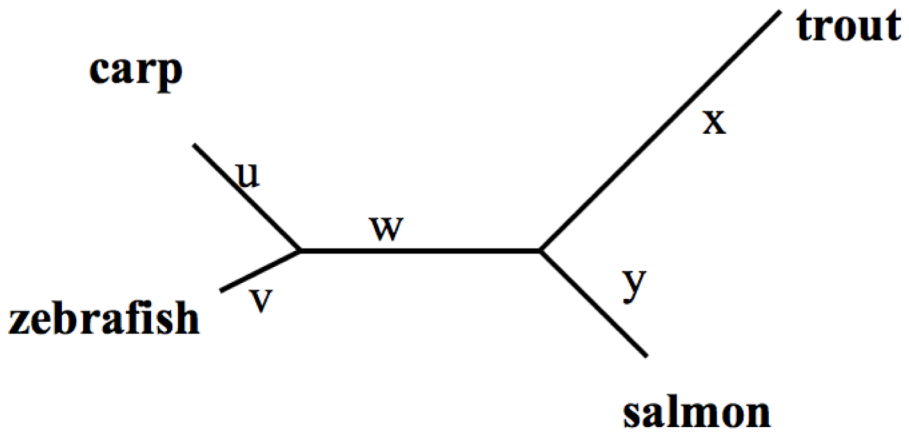
$u + w + y = 7$

$u + w + x = 9$

$v + w + y = 6$

$v + w + x = 8$

$x + y = 6$



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

A matrix can be fitted to a tree **if and only if** the equations have a solution.

$$u + v = 3$$

$$u + w + y = 7$$

$$u + w + x = 9$$

$$v + w + y = 6$$

$$v + w + x = 8$$

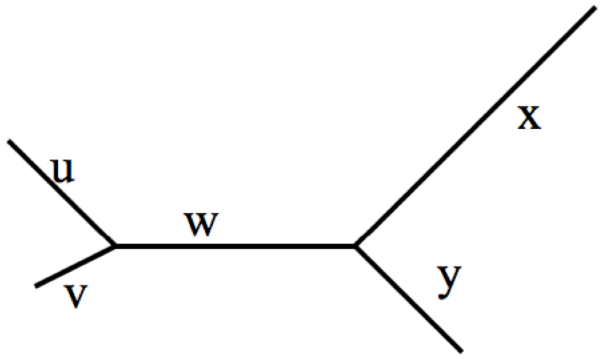
$$x + y = 6$$

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:

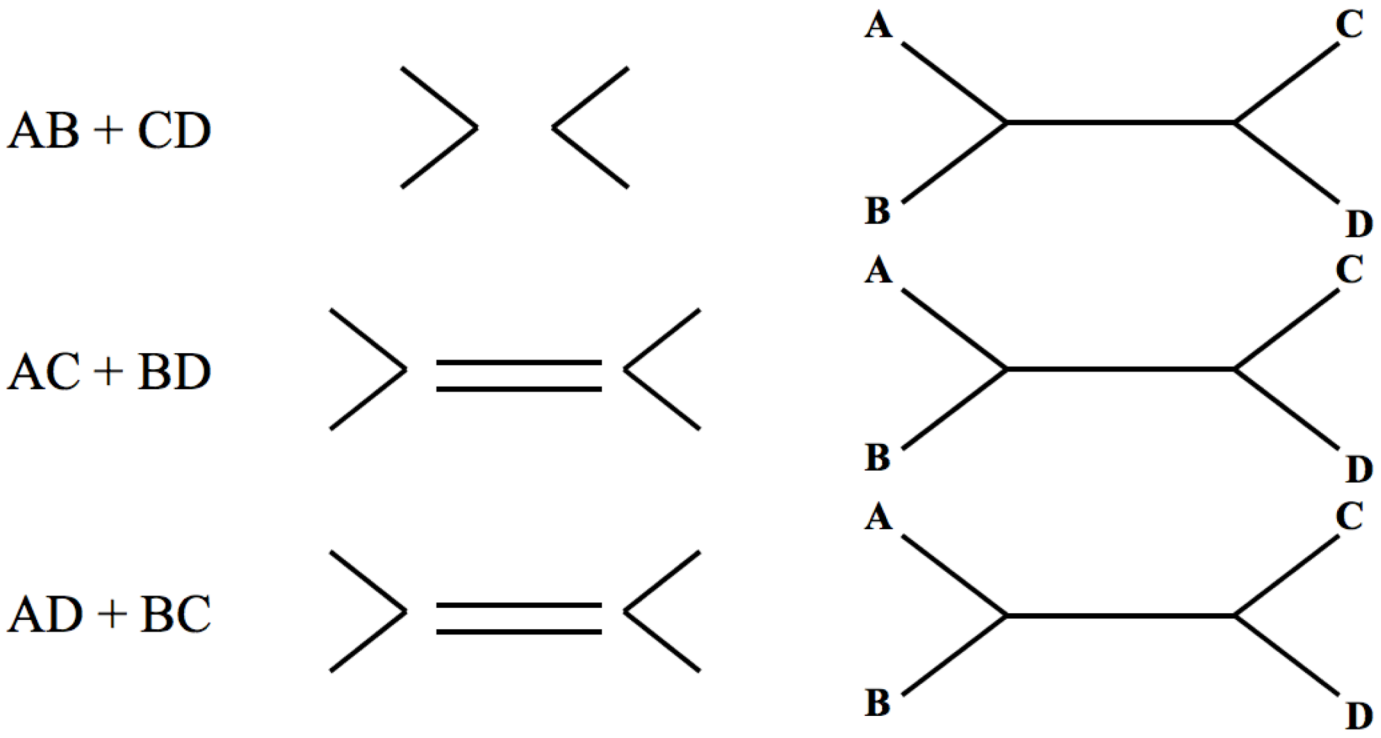
```
AB+CD <= max(AC+BD,AD+BC)
AC+BD <= max(AB+CD,AD+BC)
AD+BC <= max(AC+BD,AB+CD)
```



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

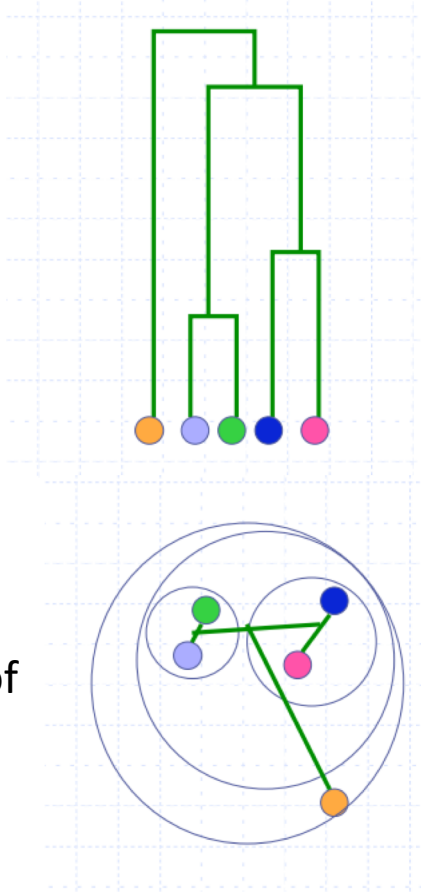
Greedy methods for distance-based phylogeny reconstruction

Taxa are points in a metric space with pairwise distances, D[i,j]. Tree building is equivalent to hierarchical clustering of these points.

These greedy algorithms maintain a forest of subtrees, beginning with the set of singleton trees (i.e., trees with one leaf and no edges). At each iteration, the algorithm merges two neighboring subtrees in the forest. The length(s) of edge(s) connecting the subtrees are calculated and the distance matrix is updated. This step is repeated until only one tree remains - the final result.

The algorithms differ in
– How neighbors to be merged are identified.
– How the branch lengths are computed.
– How the distance matrix is updated.

Unweighted Pair Group Method using Arithmetic averages (UPGMA)



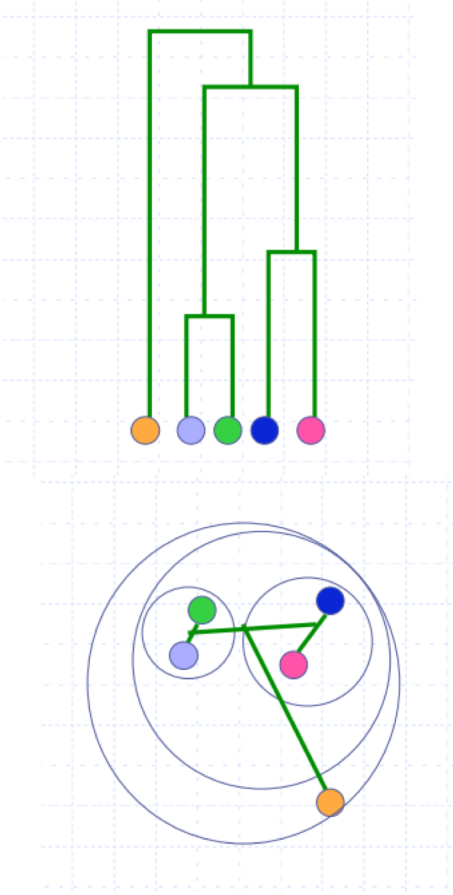
- The UPGMA algorithm is a variant of average linkage.
- UPGMA is based on the molecular clock assumption.

- 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

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

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.
- Key element – must be able to quickly compute distance between clusters (internal nodes) – weighted distance

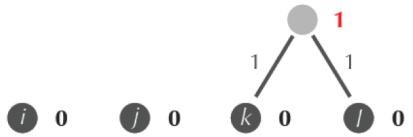


Read about Ultrametric matrix and three point condition
If a distance matrix, D, is ultrametric, then UPGMA will reconstruct the correct rooted tree in quadratic time.

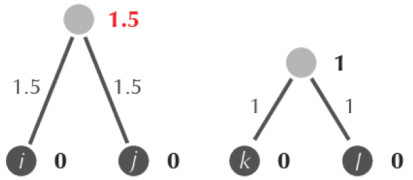
Last time

	i	*j*	*k*	*l*
i	0	3	4	3
j	3	0	4	5
k	4	4	0	2
l	3	5	2	0

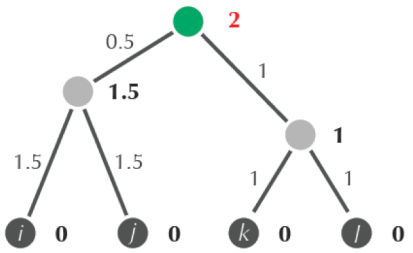
	i	*j*	*k*	*l*
i	0	3	4	3
j	3	0	4	5
k	4	4	0	**2**
l	3	5	**2**	0



	i	*j*	{*k*, *l*}
i	0	**3**	3.5
j	**3**	0	4.5
{*k*, *l*}	3.5	4.5	0



	{*i*, *j*}	{*k*, *l*}
{*i*, *j*}	0	**4**
{*k*, *l*}	**4**	0



UPGMA: An example

Observed distances:

	Q	R	S
P	9	9	4
Q	0	16	7
R		0	11

UPGMA: An example

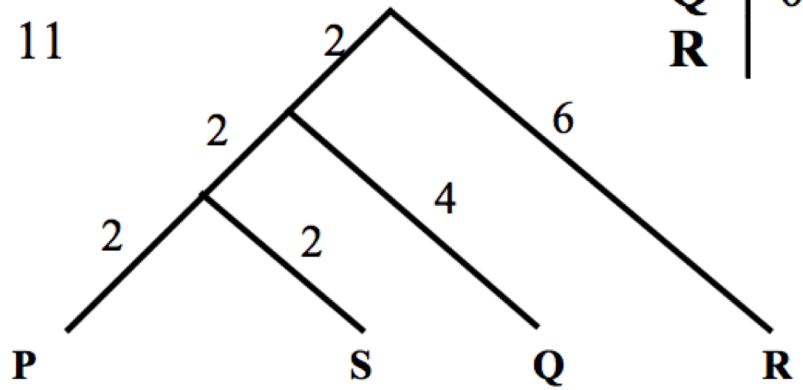
Note, however, that the tree distances are very different from the observed distances.

Observed distances:

O	Q	R	S
P	9	9	4
Q	0	16	7
R		0	11

Tree distances:

T	Q	R	S
P	8	12	4
Q	0	12	8
R		0	12



Problem? Gets incorrect tree for some additive matrices

UPGMA: Another example

As an exercise, verify that

1. this matrix is ultrametric and
2. when UPGMA is applied to this matrix, you obtain the correct tree

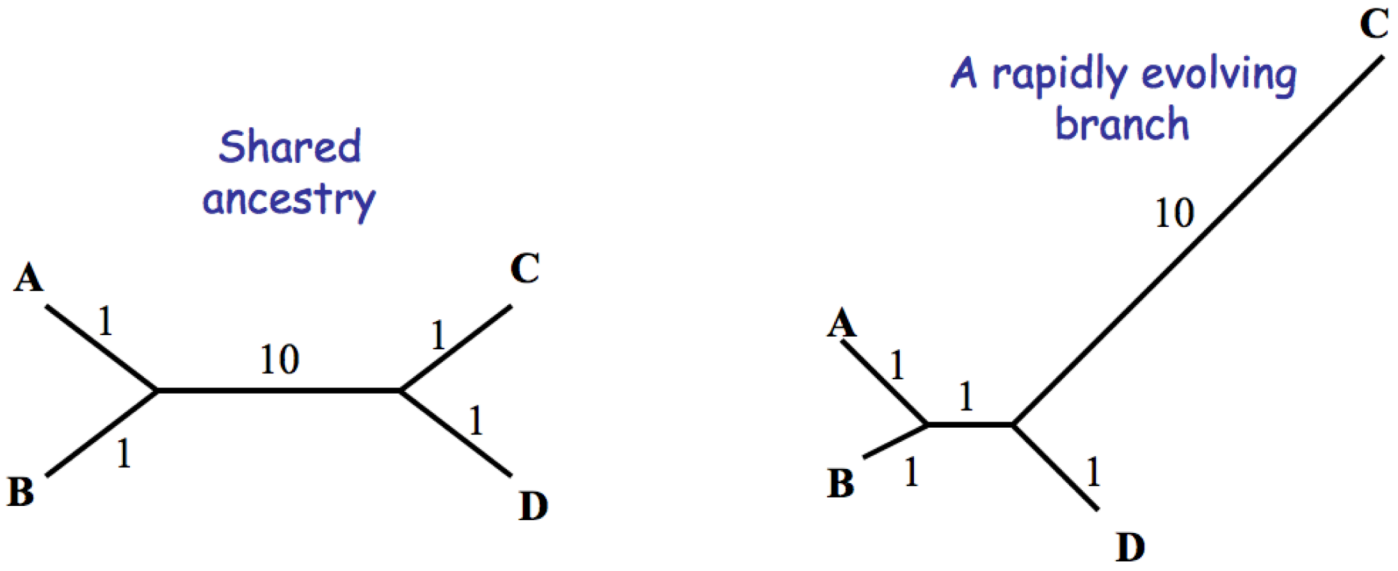
T	Q	R	S
P	8	12	4
Q	0	12	8
R		0	12

Neighbor joining algorithm

- Neighbor joining heuristics: join closest clusters that are far from the rest
- The NJ algorithm adjusts the distance matrix for variations in the rate of change. The “adjusted" distance between a pair of nodes is calculated by subtracting the average of the distances to all other leaves.

- Thm: – If D is additive, the pair of taxa that minimimize this "corrected" distance matrix are neighbors in the true tree.
- Proof: – Durbin et al., 7.8
- If D is additive, then NJ will reconstruct the correct unrooted tree in quadratic time.

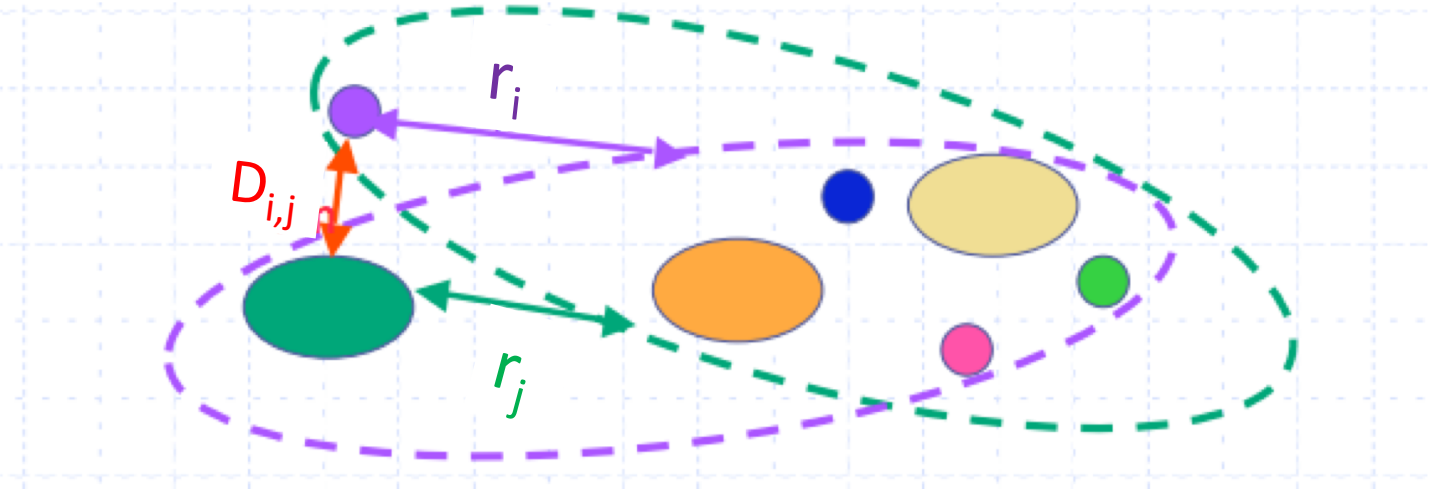
NJ intuition



Does a long branch indicate shared ancestry or a change in the substitution rate?

Neighbor joining algorithm

- Neighbor joining heuristics: join closest clusters that are far from the rest
- Define: $R_i=\Sigma_{i\neq k} D_{ik}$ the divergence of i
- Cluster nodes i and j that minimize $D^*_{i,j} = D_{i,j} – (R_i + R_j)/(n-2)$
- $(R_i + R_j)/(n-2)$ – computes the average distance of i and j from the rest of the leaves. (There are n leaves, so normalizing by n-2)
- Define $r_i = R_i /(n-2)$ for simplicity. $D^*_{i,j} = D_{i,j} – r_i - r_j$



Neighbor joining

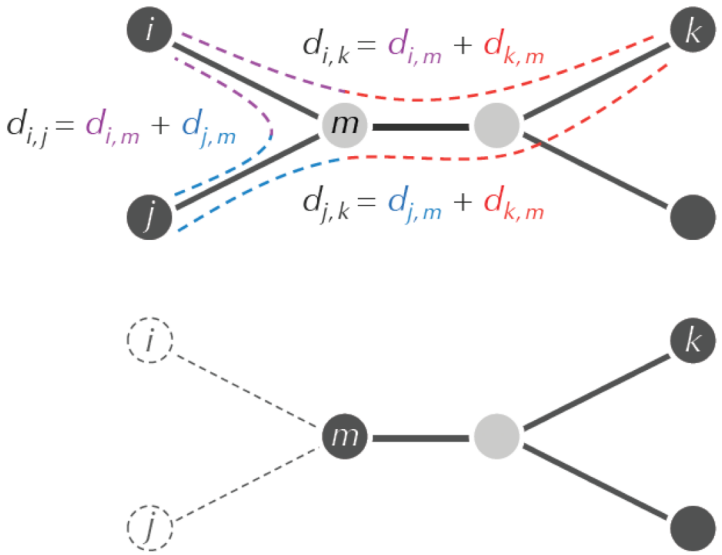
- Pick two nodes with NJdist(i,j) minimal

– Create parent m s.t.

– $D(m, k) = 0.5\ (D(i,k) + D(j,k) – D(i,j))$ for every other node k

– $D(i, m) = 0.5\ (D(i,j) + r_i – r_j)$ - length of branch between i & m

– $D(j, m) = 0.5\ (D(i,j) + r_j – r_i)$ – length of branch between j & m



Neighbor joining

- Complexity is $O(n^2)$
- Does not depend on molecular clock assumption
- Heavily used in practice [e.g., Clustal W]
- But can be sensitive to non-additivity

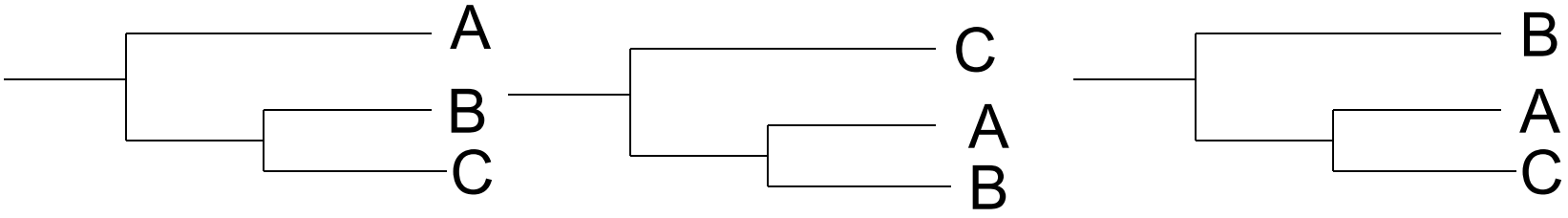
Maximum Parsimony (character based phylogeny)

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

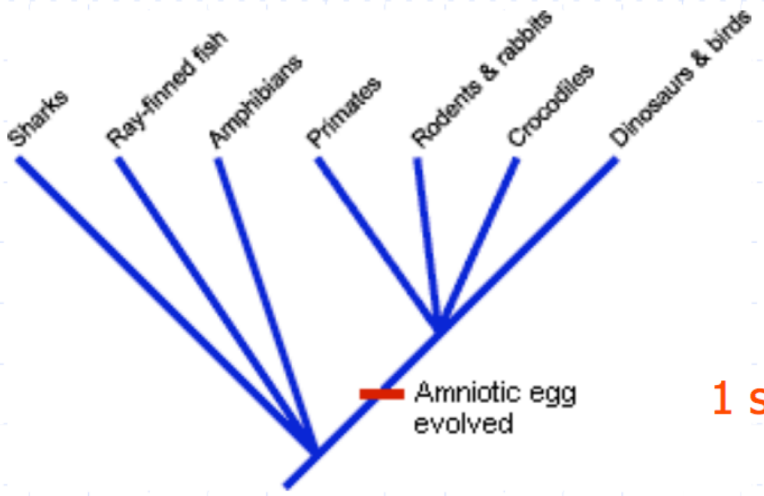
Example

	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

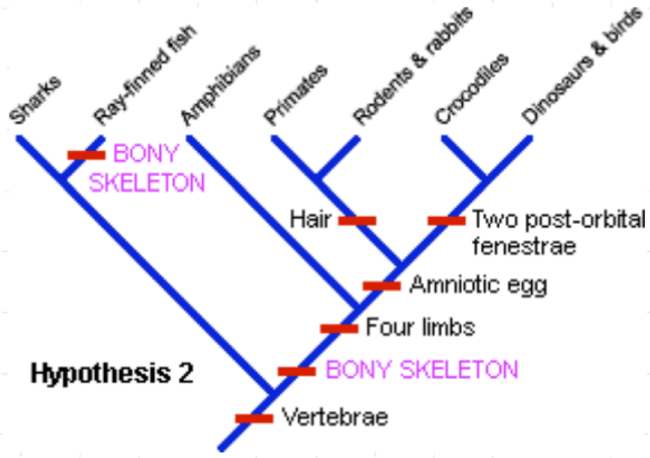


1 state change

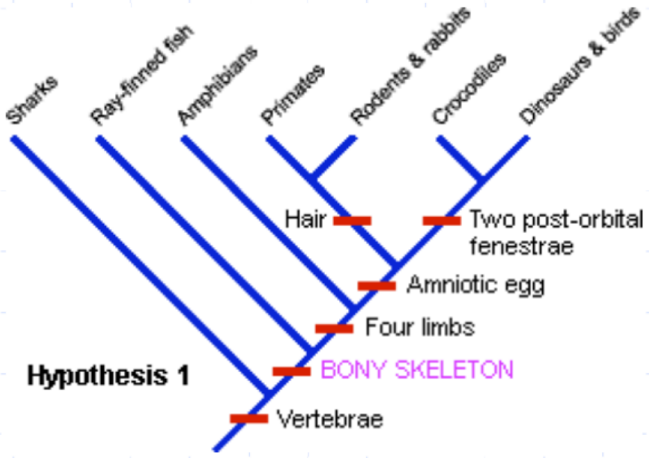
Example

	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

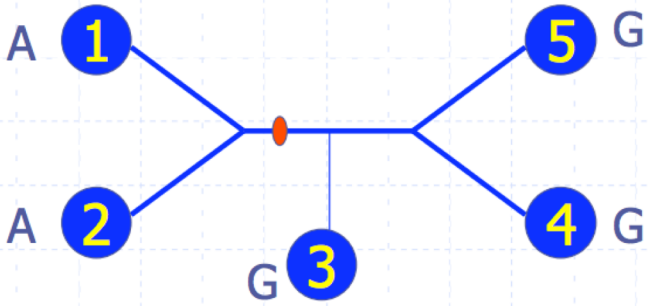


6 state changes

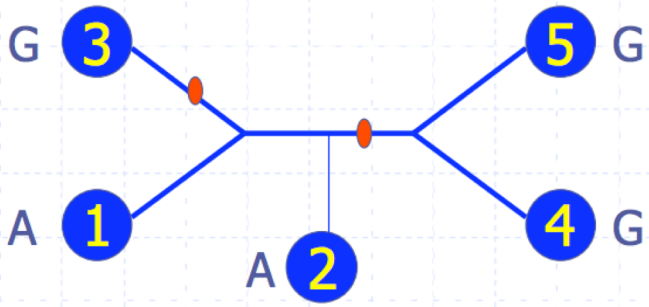


Example

- 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

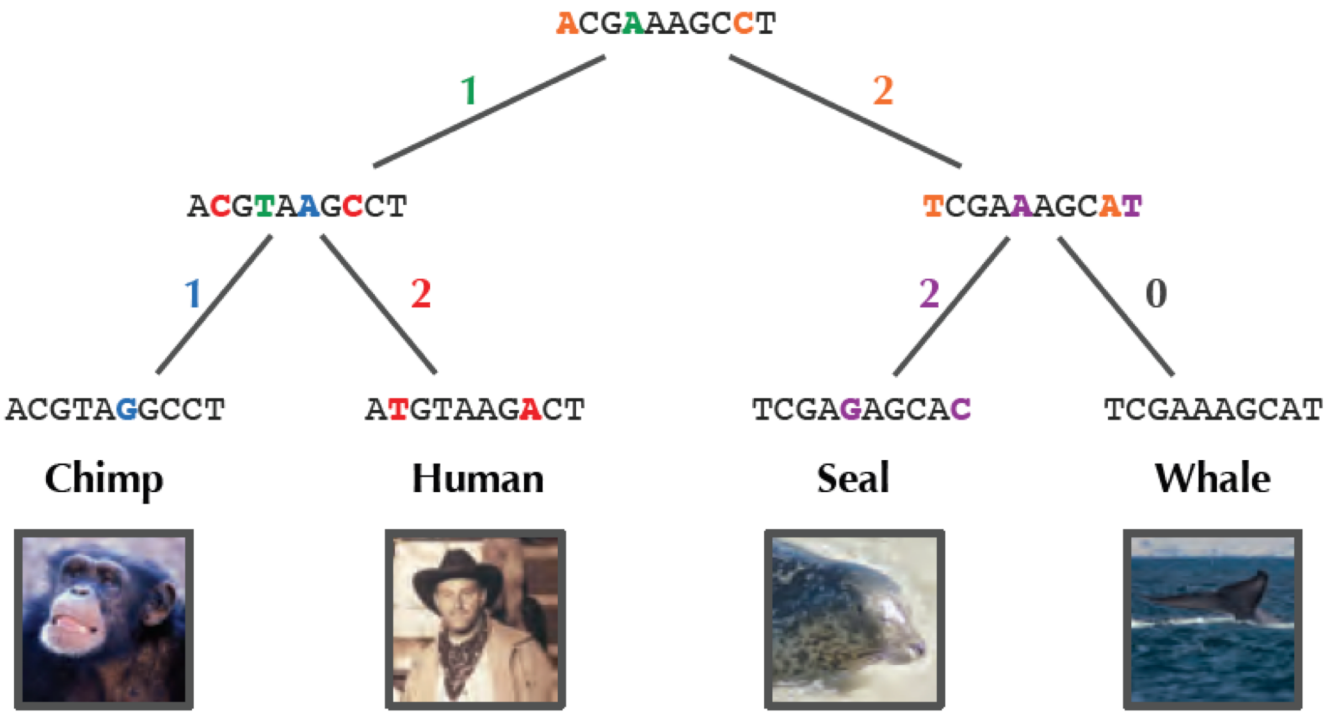
–find labels at internal nodes s.t. # of state transitions is minimzed

•Weighted small parsimony

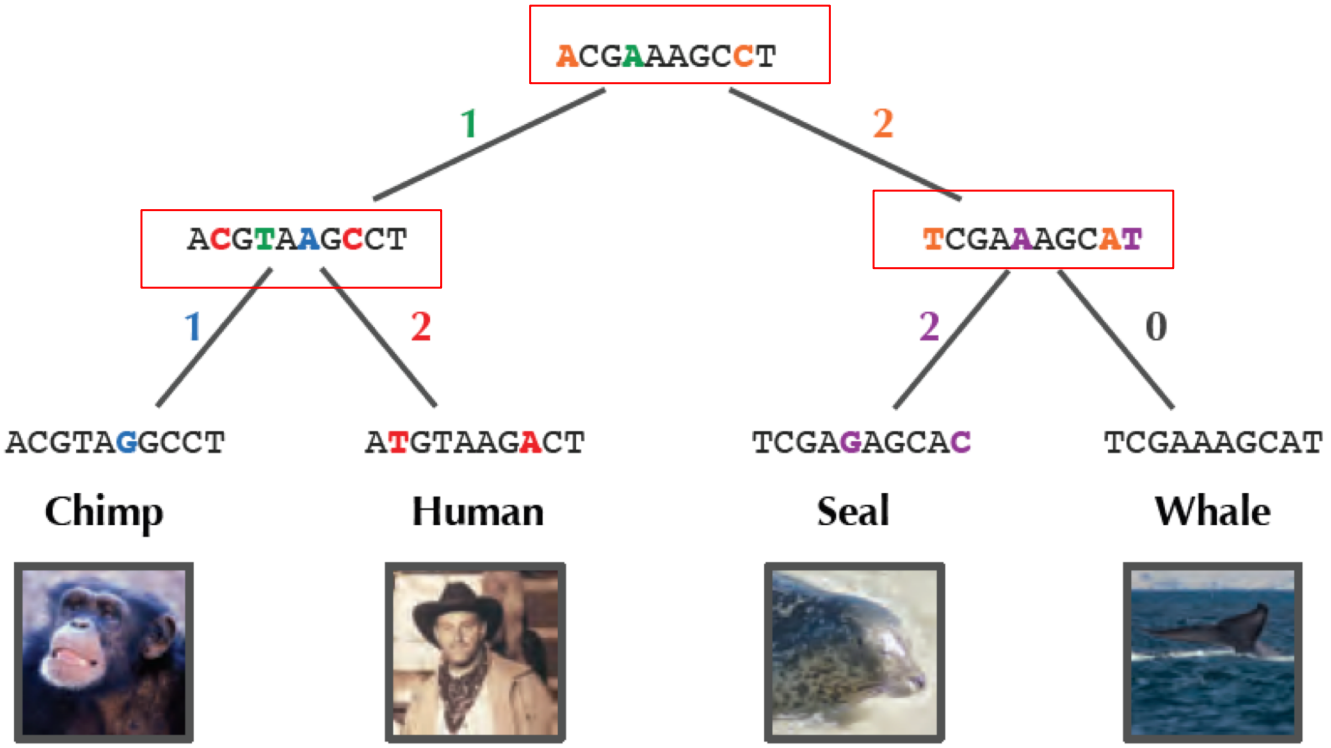
–same as parsimony except that state transitions are assigned weights

–minimize the overall weight of the tree

Scoring a tree – Sankoff's algorithm



Scoring a tree – Sankoff's algorithm

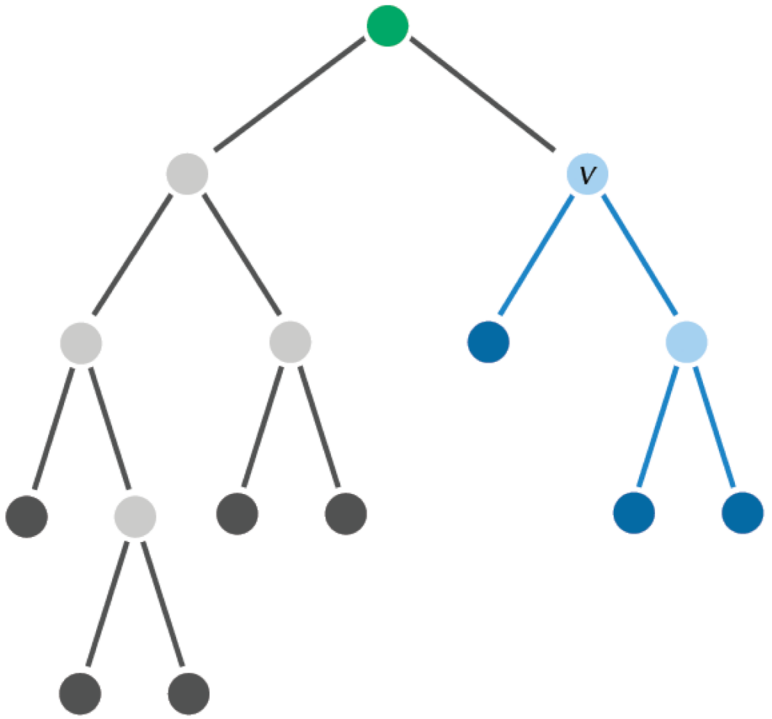


Make a simplifying assumption – all characters are independent in the sequence
i.e. Run separately for each character then merge results

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?

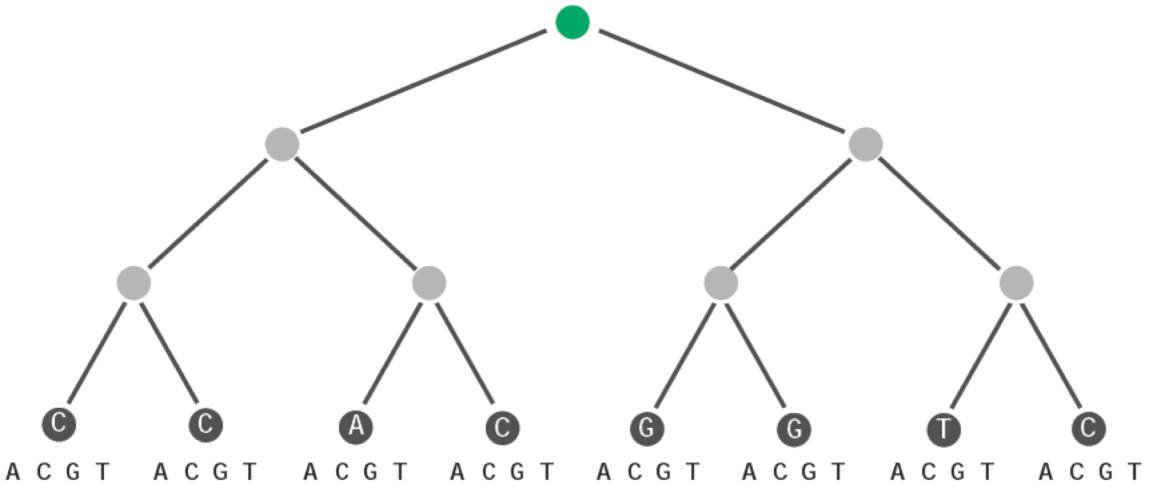


The (blue) subtree T_v of a node v within a larger rooted binary tree T.

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

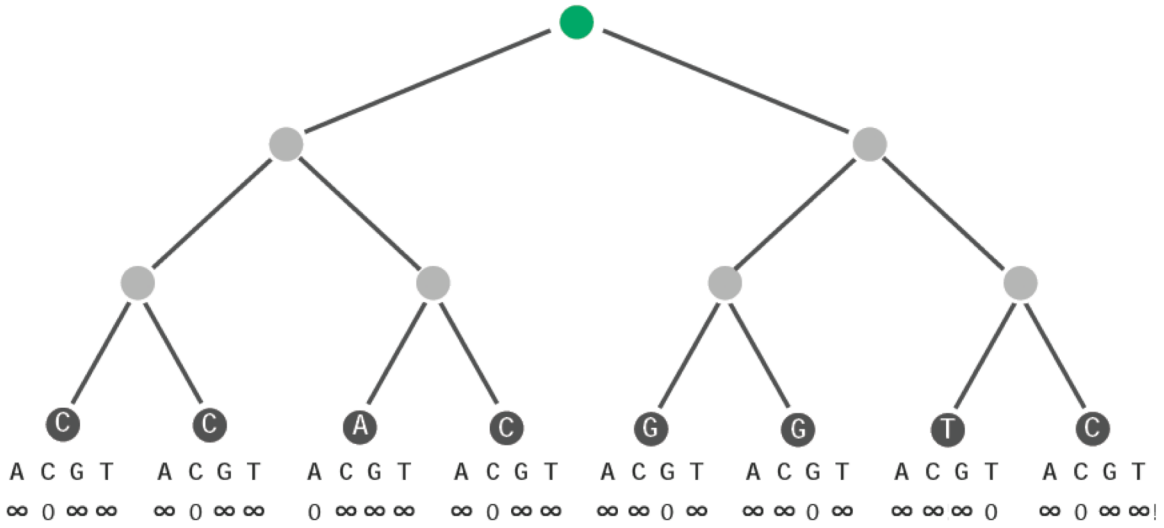
How to get a Base case?



Sankoff's algorithm – recurrence relation

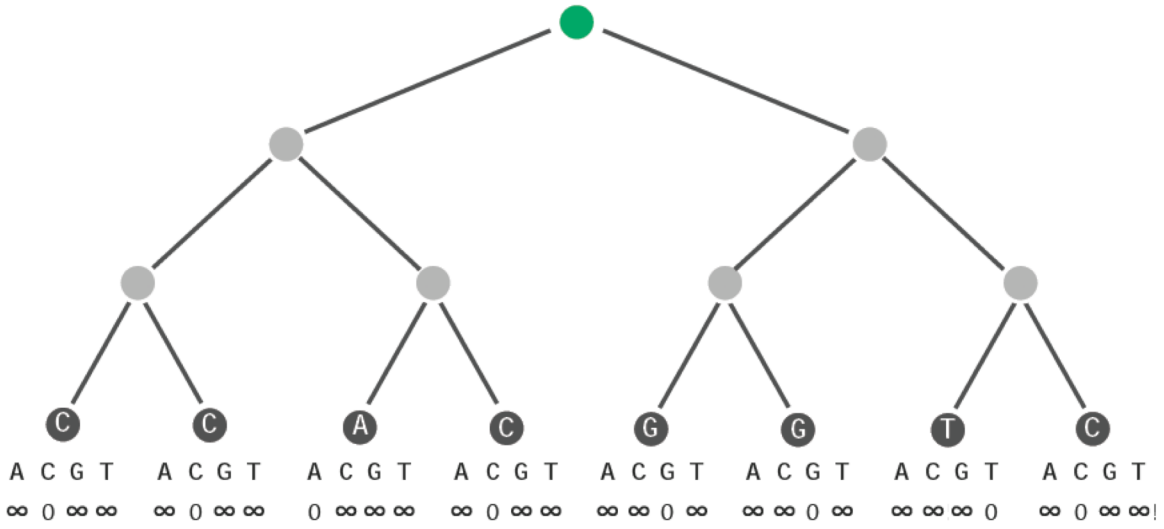
- At each node v in the tree store s(v,t) – best parsimony score for subtree rooted at v if character stored at v is t

How to get a Base case?



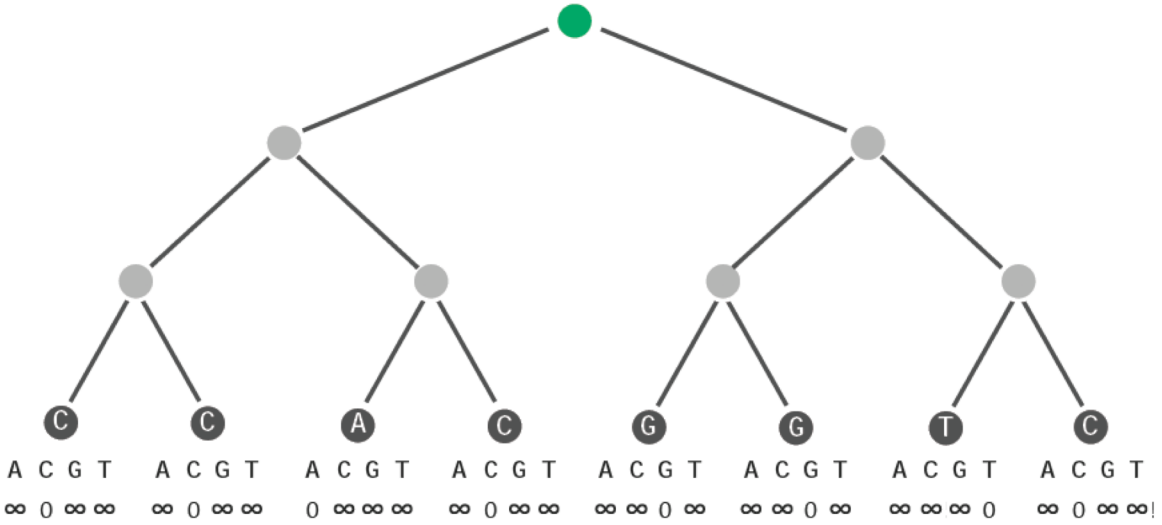
Sankoff's algorithm – recurrence relation

- 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

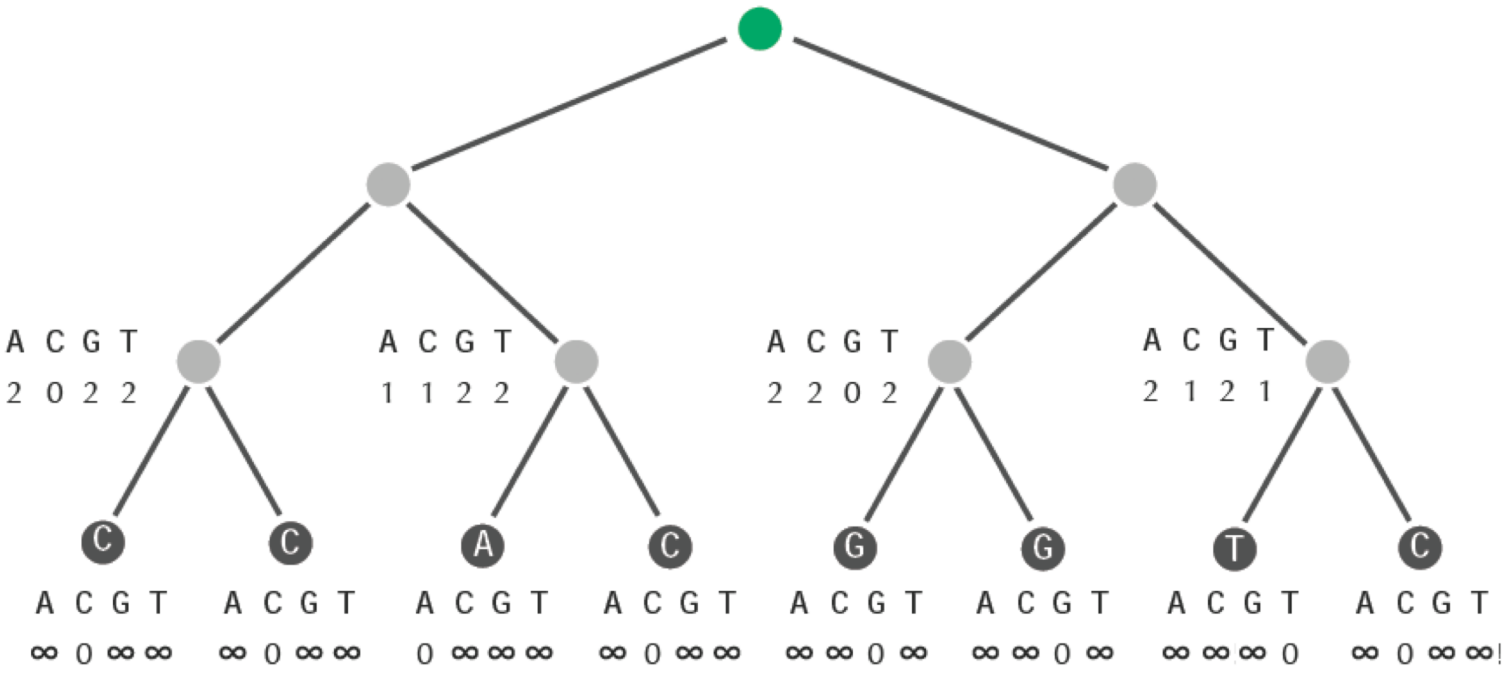
- At each node v in the tree store s(v,t) – best parsimony score for subtree rooted at v if character stored at v is t
- Traverse the tree in post-order and update s(v,t) as follows
  - assume node v has children u and w
  - $s(v,t) = \min_i \{s(u,i) + score(i,t)\} + \min_j \{s(w,j) + score(j,t)\}$



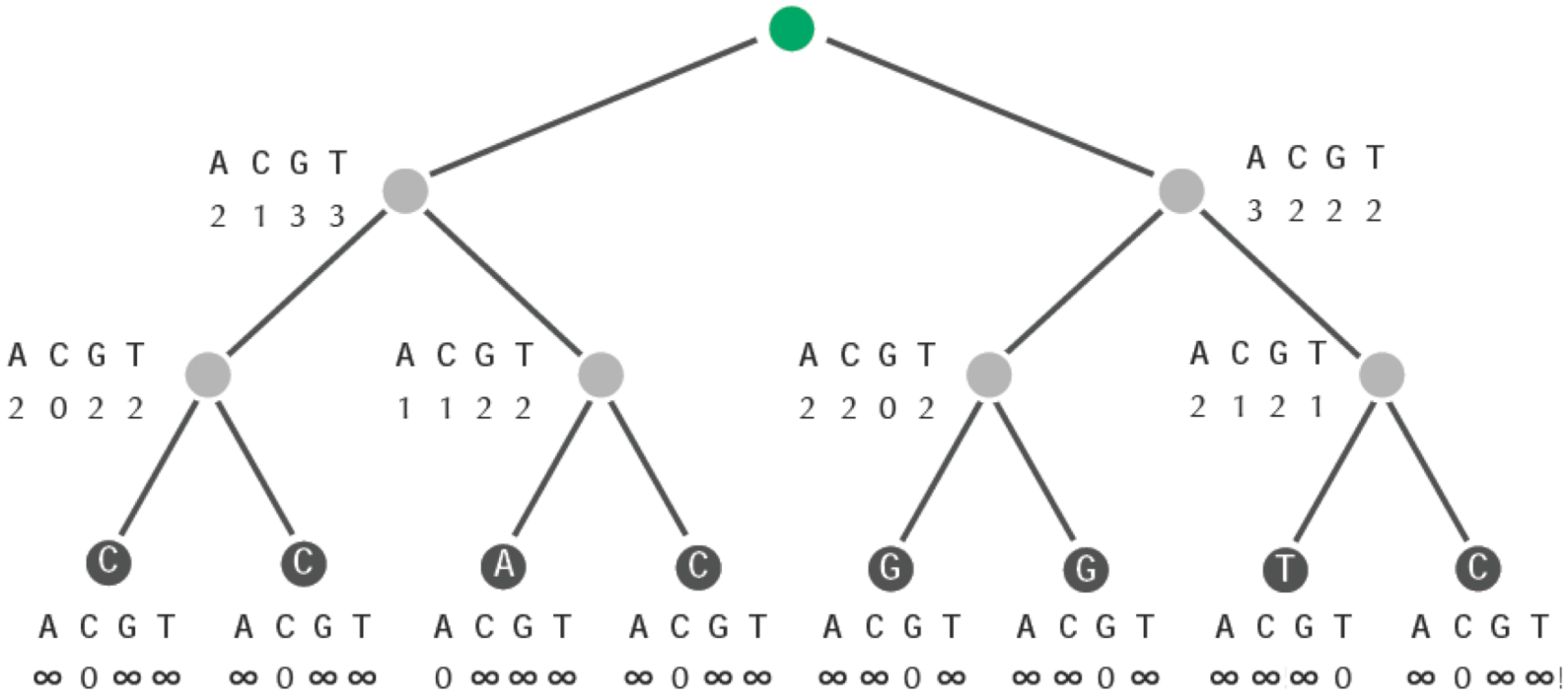
Sankoff's algorithm – recurrence relation

- 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

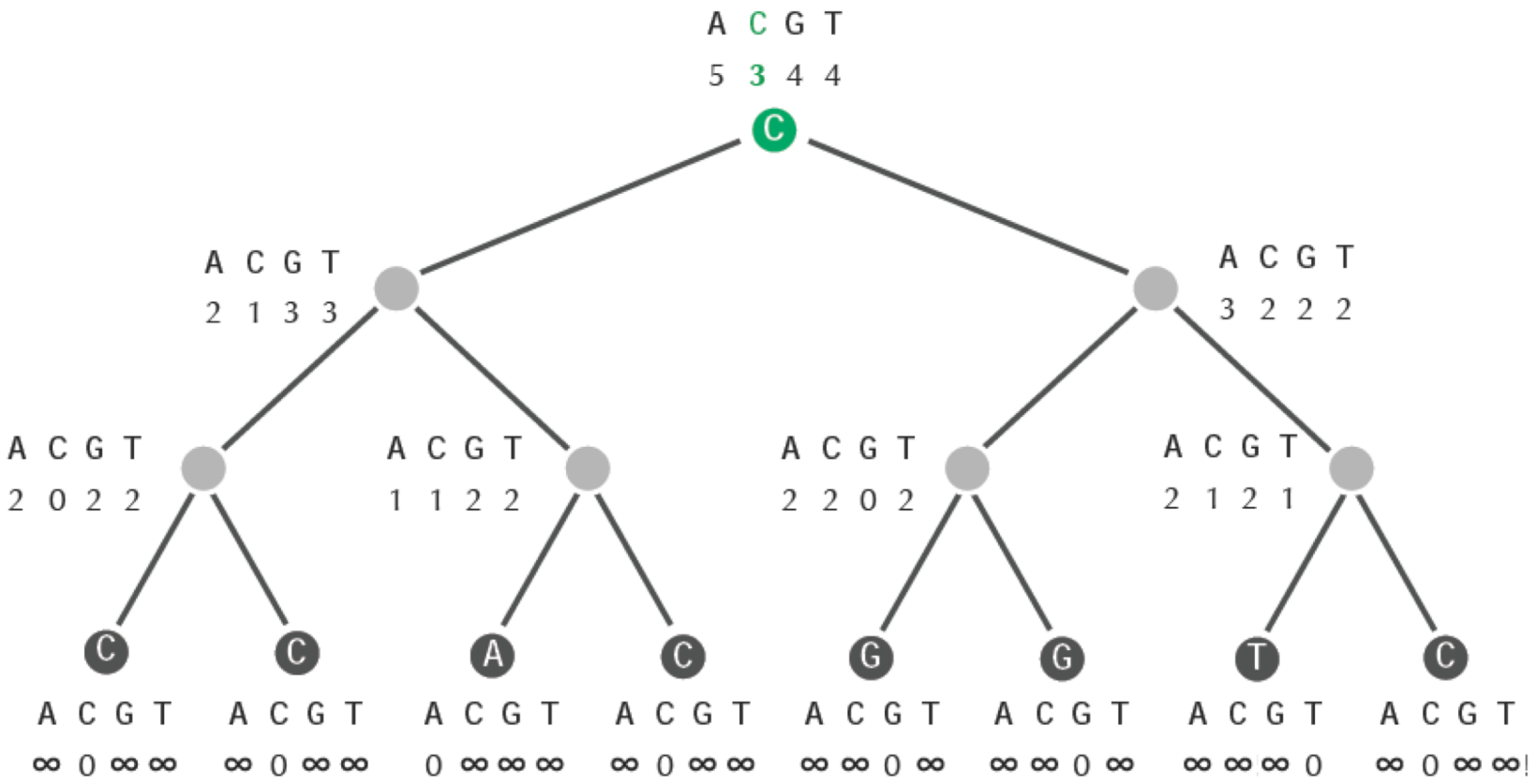
Sankoff's algorithm – example (continued)



Sankoff's algorithm – example (continued)



Sankoff's algorithm – example (continued)

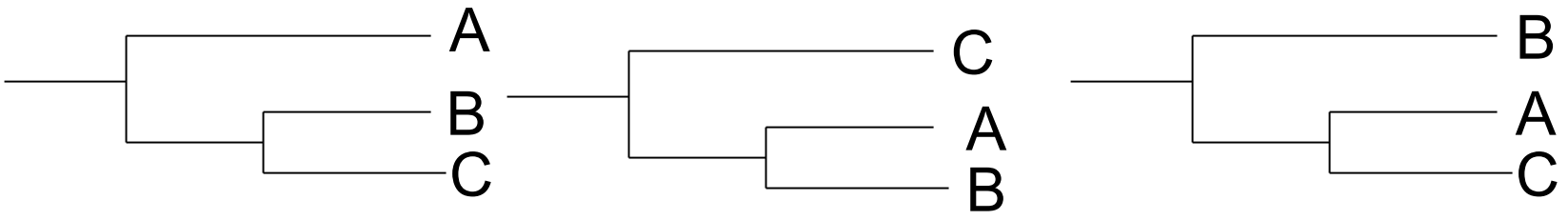


Optimal labeling can be computed in linear time O(nk) where n is number of leaves and k is number of character states

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

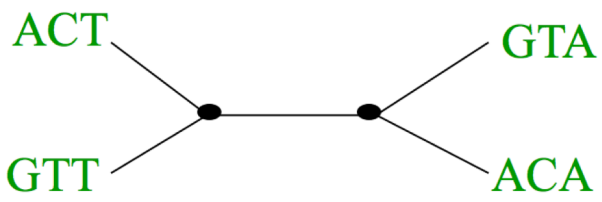


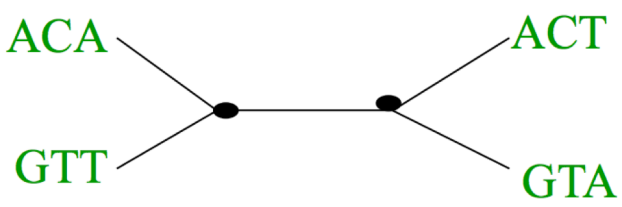
- Finding the optimal Maximum parsimonius tree is NP-hard
- Exponential number of trees
- Heuristics to bound the search space
  - Branch and bound
  - Nearest neighbor interchange (NNI) - switch subtrees
  - Prune scoring a tree if the score exceeds best score of a fully explored tree

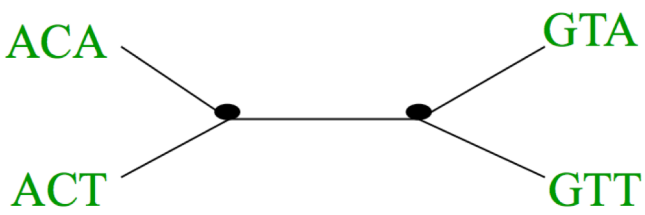
Maximum parsimony (example)

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

Maximum Parsimony







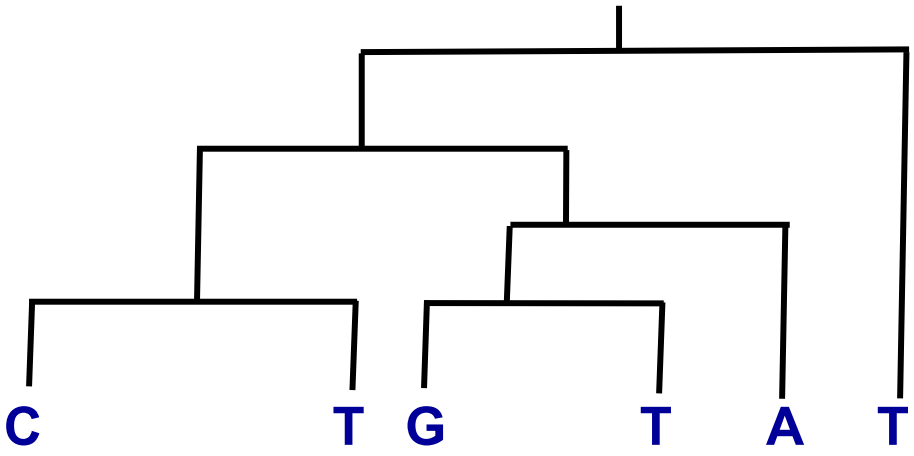
Maximum likelihood

- 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

Questions

- Why do you need a multiple alignment for phylogeny?
- What is the running time of the neighbor-joining algorithm, given k sequences of length L?
- What is the parsimony score of the following tree, and what are the labels at internal nodes?



Thanks!

Contact:
nidhi@cs.umd.edu