

Spring 2020 CR/NR Grade Info

<https://comm.sjsu.edu/JK2EU00Z3Q0JgeU0MC00010>

CS123A
Bioinformatics
Module 3 – Week
12 – Presentation 1

Leonard Wesley
Computer Science Dept
San Jose State Univ

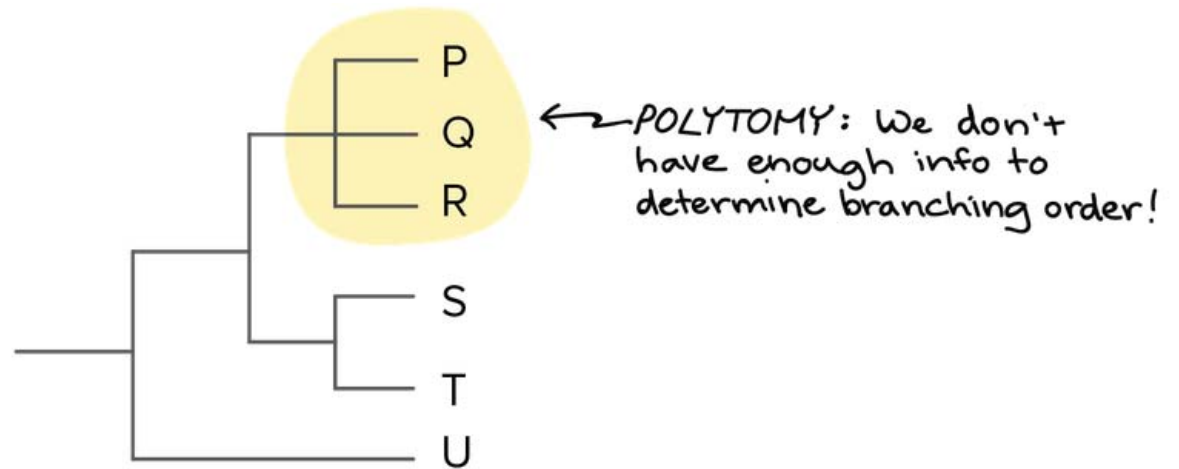


Image modified from *Taxonomy and phylogeny: Figure 2* by Robert Bear et al., CC BY 4.0

Agenda

- Quiz 3 on 4/16
 - Study guide posted to Canvas
- Neighbor Joining Tree Building
- Phylogenetic Tree Review Slides & Parsimony Slides

Introduction To Neighbor-Joining Tree Building Method

- Created by Naruya Saitou and Masatoshi Nei in 1987. Reformulated by Studier and Keppler in 1988. Criticized by many, but still used as benchmark.
- Neighbor –Joining does not assume a constant rate of evolution.
- The algorithm is based on the concept of minimum evolution; the true tree is the one for which the total branch length is minimum.
- The resulting tree is not rooted and is additive.

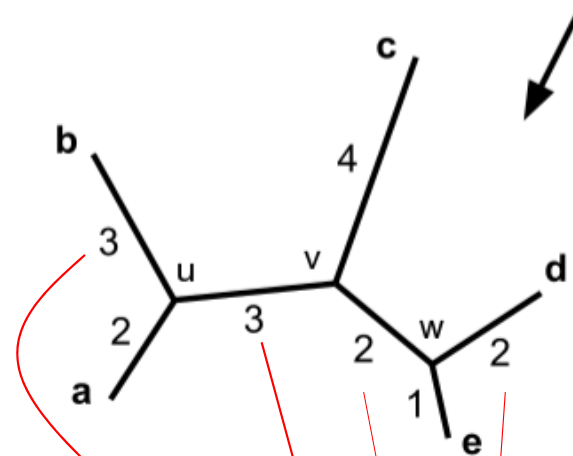
What Do We Mean By Not Rooted & Additive?

Distance Matrix

| | a | b | c | d | e |
|---|---|----|----|----|---|
| a | 0 | 5 | 9 | 9 | 8 |
| b | 5 | 0 | 10 | 10 | 9 |
| c | 9 | 10 | 0 | 8 | 7 |
| d | 9 | 10 | 8 | 0 | 3 |
| e | 8 | 9 | 7 | 3 | 0 |

Additive means the distance in the final tree equals the distance in the distance between two taxa/OTU in the original distance matrix.

Unrooted NJ Tree



$$\text{b to d distance} = 3 + 3 + 2 + 2 = 10$$

Advantages Of NJ

- ADVANTAGE: It is fast, due in part to its being a polynomial-time algorithm.
 - $y=x$ (linear), $y=x^4$ (polynomial), $y=2^x$ (exponential)
- Practical for analyzing large data sets (hundreds or thousands of taxa) where max parsimony, maximum likelihood and bootstrapping may take too long.
- If the input distance matrix is correct, then the output tree will be correct. Furthermore, the correctness of the output tree topology is guaranteed as long as the distance matrix is 'nearly additive'.

Disadvantages Of NJ

- **DISADVANTAGE:** In practice the distance matrix rarely satisfies this condition (i.e., being additive), but neighbor joining often constructs the correct tree topology anyway.
- Neighbor joining has been largely superseded by other phylogenetic methods
- It often assigns negative lengths to some of the branches.


The NJ Idea

The raw data of the tree are represented by the following distance matrix:

| | 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 |

We have in total 6 OTUs (N=6).

Step 1: Calculate Dissimilarity/Divergence Matrix


$$\begin{aligned}r(A) &= 5 + 4 + 7 + 6 + 8 = 30 \\r(B) &= 42 \\r(C) &= 32 \\r(D) &= 38 \\r(E) &= 34 \\r(F) &= 44\end{aligned}$$

The NJ Idea (cont.)

Step 2: Now we calculate a new distance matrix using for each pair of OUTs the formula:

$M(ij) = d(ij) - [r(i) + r(j)] / (N - 2)$ or in the case of the pair A,B:

$M(AB) = d(AB) - [(r(A) + r(B))] / (N - 2) = -13$

$r(A) = 5 + 4 + 7 + 6 + 8 = 30$
 $r(B) = 42$
 $r(C) = 32$
 $r(D) = 38$
 $r(E) = 34$
 $r(F) = 44$

Original distance matrix

| | 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 |

N = number of taxa/OTUs

$$5 - (30 + 42) / (6 - 2) = 5 - (72 / 4) = 5 - 18 = -13$$

| | A | B | C | D | E |
|---|-----|---|---|---|---|
| B | -13 | | | | |
| C | | | | | |
| D | | | | | |
| E | | | | | |
| F | | | | | |

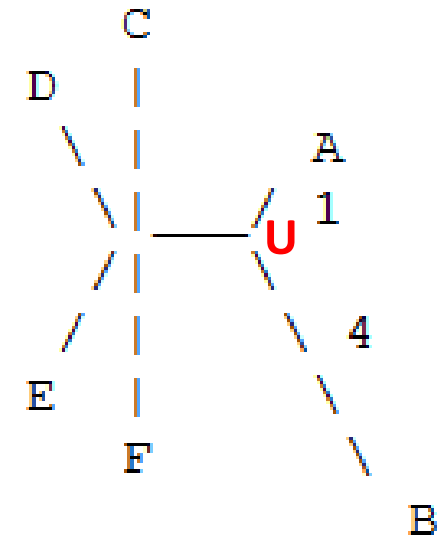
You Calculate The New Distance For (B, C)

| | 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 |

```
r (A)  =  5+4+7+6+8=30
r (B)  =  42
r (C)  =  32
r (D)  =  38
r (E)  =  34
r (F)  =  44
```

Next: Start With A Star Tree

- **Step 3:** Now we choose as neighbors those two OTUs for which M_{ij} is the smallest.
- These are A and B and D and E.
- Let's take A and B as neighbors and we form a new internal node called U.
- Now we calculate the branch length from the internal node U to the external OTUs A and B.



$$S(AU) = d(AB) / 2 + [r(A) - r(B)] / 2(N-2) = 1$$

$$5/2 + (30-42)/8 = 2.5 + (-12/8) = 2.5-1.5 = 1$$

$$S(BU) = d(AB) - S(AU) = 4$$

$$5 - 1 = 4$$

| | 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 |

Update Distance Matrix

- **Step 4:** Now we define new distances from U to each other terminal node:

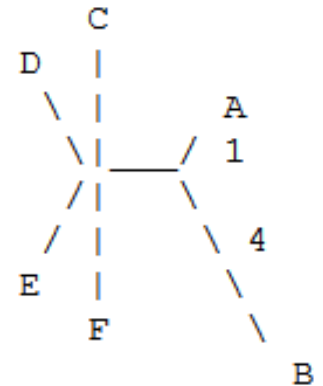
| | A | B | C | D | E |
|---|---|----|---|---|---|
| B | 5 | | | | |
| C | 4 | 7 | | | |
| D | 7 | 10 | 7 | | |
| E | 6 | 9 | 6 | 5 | |
| F | 8 | 11 | 8 | 9 | 8 |

$$(4 + 7 - 5) / 2 = 6 / 2 = 3$$

- $d(CU) = (d(AC) + d(BC) - d(AB)) / 2 = 3$
- $d(DU) = (d(AD) + d(BD) - d(AB)) / 2 = 6$
- $d(EU) = (d(AE) + d(BE) - d(AB)) / 2 = 5$
- $d(FU) = (d(AF) + d(BF) - d(AB)) / 2 = 7$

| | U | C | D | E |
|---|---|---|---|---|
| C | 3 | | | |
| D | 6 | 7 | | |
| E | 5 | 6 | 5 | |
| F | 7 | 8 | 9 | 8 |

NOW N = 5



Repeat Steps 1 to 4 Until
All Branches Have Distances

A NJ Example

Primate mitochondrial DNA sequences, HindIII

Hayasaka, K., T. Gojobori, and S. Horai. MBE (1988) 5:626-644.

| | Gorilla | Orangutan | Human | Chimp | Gibbon |
|-----------|---------|-----------|--------|--------|--------|
| Gorilla | 0 | 0.1890 | 0.1100 | 0.1130 | 0.2150 |
| Orangutan | 0.1890 | 0 | 0.1790 | 0.1920 | 0.2110 |
| Human | 0.1100 | 0.1790 | 0 | 0.0940 | 0.2050 |
| Chimp | 0.1130 | 0.1920 | 0.0940 | 0 | 0.2140 |
| Gibbon | 0.2150 | 0.2110 | 0.2050 | 0.2140 | 0 |

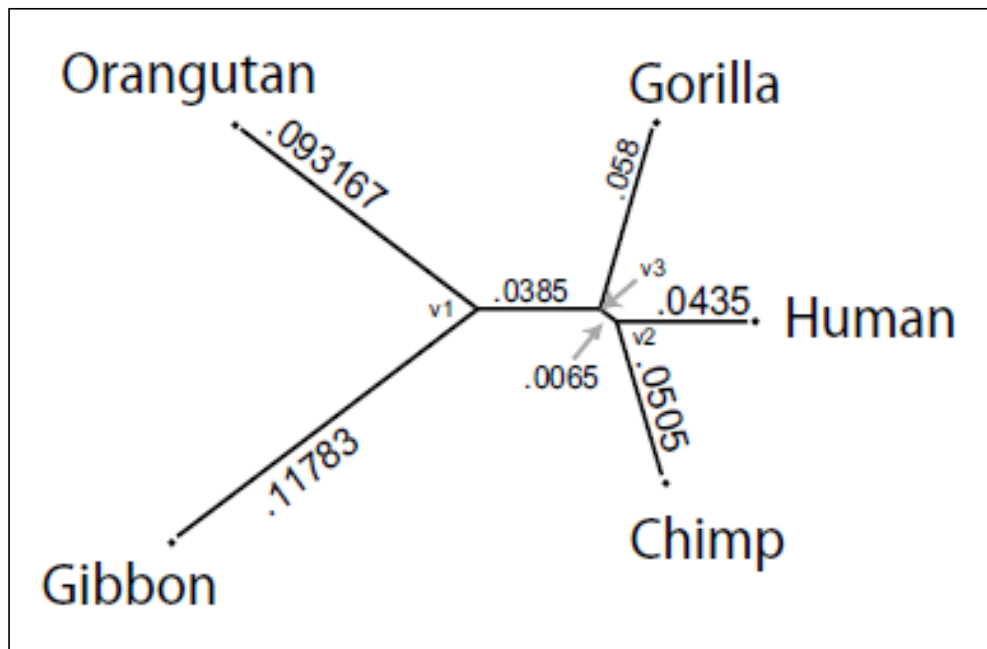
$R_i = \sum_{j=1}^N d_{ij}$, total dissimilarity between taxon S_i and all other taxa

$$R_1 = R_{\text{Gorilla}} = .189 + .11 + .113 + .215 = .627$$

YOU: Complete One Iteration Of NJ
Algorithm For Hayasaka Distance Matrix On
Previous Slide

i.e., Iterate through First 4 Steps

NJ: Final Result



Final Result

If your result does not
Have identical values, they
Should be relative.

Phylogenetic Tree Review Summary

Phylogenetic Distance Algorithms

- UPGMA: Un-weighted pair group method with arithmetic mean. A clustering method, joins branches based on distance between pairs and average of joined pairs.
- NJ: Neighbor Joining inserts branches between pairs of closest neighbors and terminals in tree.
- FM: Fitch Margoliash maximizes fit of observed pair wise distances to a tree by minimizing the squared deviation of all possible observed distances.
- ME: Minimum Evolution tries to find shortest tree that is consistent with path lengths measured in a manner similar to FM.

Two Main Types Of Tree Building Methods

Clustering Methods

- Follow a set of steps (an algorithm) and arrive at a tree.
- Use distance data.
- Produce a single tree.
- Do not use objective functions to compare the current tree to other trees.

Optimality Criterion

- Use objective functions to compare different trees.
- First define an optimality criterion, i.e. minimum branch length, and then find the tree with the best value for the objective function.

Strength Of Clustering

- The strength of clustering algorithms is:
 - Their speed
 - Their robustness
 - Their ability to reconstruct trees for very large numbers (thousands) of sequences.
 - Most clustering methods reconstruct phylogenetic trees for a set of sequences on the basis of their pairwise evolutionary distances.

Strength Of Optimality-Based Methods

- Can be more accurate if you have a good objective function and substitution data.
- Can be used to compare trees

Classification Of Tree Building Methods

| | | Tree Building Methods | |
|--------------|-----------------|---------------------------|---|
| | | Clustering Algorithm | Optimality Criterion |
| Type of Data | Distance-Based | UPGMA Neighbor Joining | Fitch-Margoliash |
| | Character-Based | | Maximum Parsimony Maximum Likelihood |

PARSIMONY
Not Covered In Lecture
Will Not Be On Exams
The Following Slides Here Just FYI

At Least Two Ways To Measure Distance In Trees

- Distance-based: uses measures of gene mutation, time,...
 - First calculate the overall distance between all pairs of sequences, then construct a tree based on the distances
- Character-based: morphological features (e.g., number of legs), DNA/protein sequences.
 - Uses the individual substitutions among sequences to determine the most likely ancestral relationships. The tree is constructed based on the gain or loss of traits (i.e., “character”).

Models Of Substitution Rates Between Bases

- Substitutions between bases or amino acids that are more chemically similar.
- DNA:
 - $A \rightarrow G$, $G \rightarrow A$, $C \rightarrow T$, $T \rightarrow C$ are usually more frequent than
 - $A \rightarrow C$, $A \rightarrow T$, $C \rightarrow G$, $G \rightarrow T$, and the reverse
- Amino Acids: Similar idea... too many to list hear.

Relative Rates Of Substitutions In Square Matrices

- DNA (4 square matrix), Amino Acids (20 square matrix), Codons (61 square matrix)
- Off-diagonal elements = cost of changing a base/amino acid/codon ...

| | A | C | G | T |
|---|---|---|---|---|
| A | - | 2 | 1 | 2 |
| C | 2 | - | 2 | 1 |
| G | 1 | 2 | - | 2 |
| T | 2 | 1 | 2 | - |

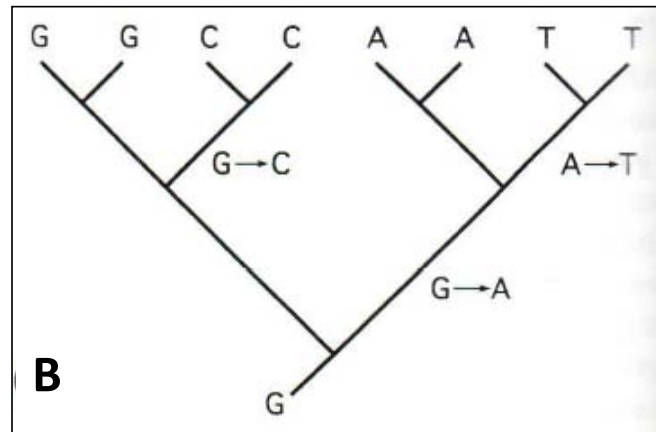
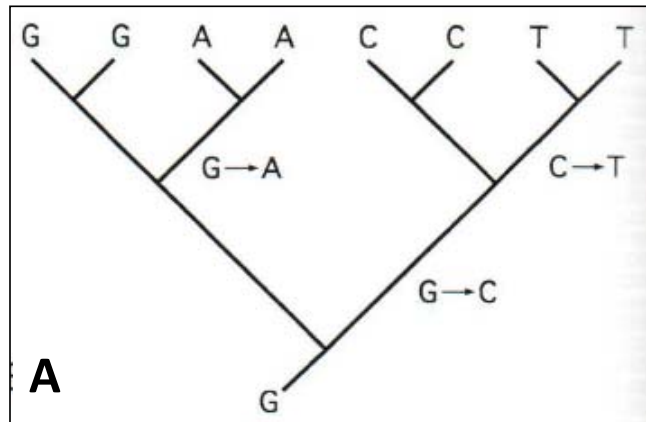
Weighted Parsimony Method

(MP = Maximum Parsimony)

A Character-Based Method

- Parsimony: Find or use easiest, most economical, least complex,...etc. explanation or method.
- Cost schedule/matrix is fixed, tree building can calculate an exact cost of a mutated sequence.

Non-Weighted Parsimony Example

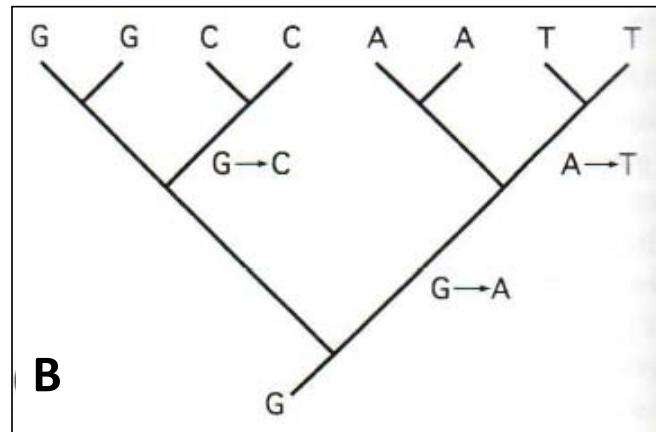
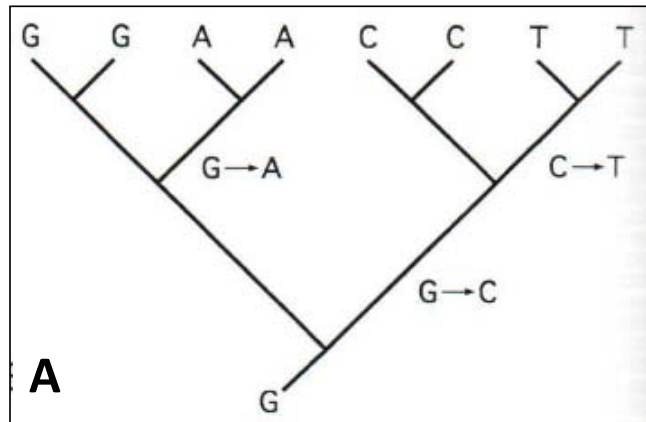


Consider a character with eight sequences

| | | |
|------------|---|---------------------|
| | T | T |
| | T | T |
| A → | C | A ← B |
| | C | A |
| | A | C |
| | A | C |
| | G | G |
| | G | G |

Both possible trees are equally plausible as explanations of substitutions.

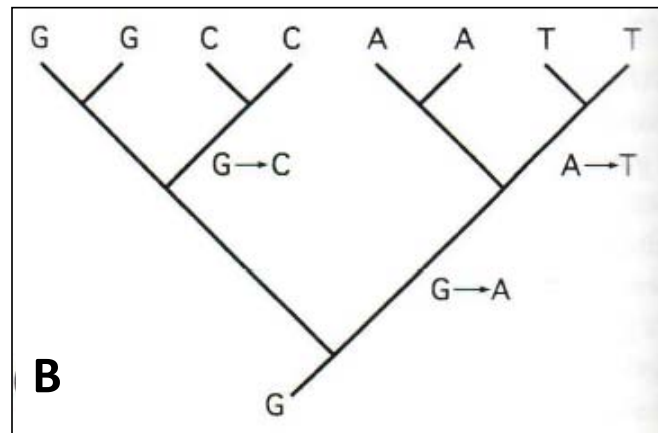
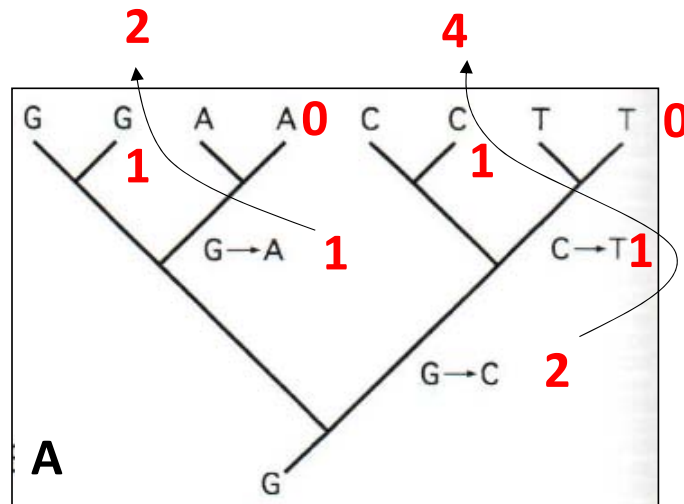
Weighted Parsimony Example



| | A | C | G | T |
|---|---|---|---|---|
| A | - | 2 | 1 | 2 |
| C | 2 | - | 2 | 1 |
| G | 1 | 2 | - | 2 |
| T | 2 | 1 | 2 | - |

Presumed that a **transversion** costs twice as much as a **transition**.

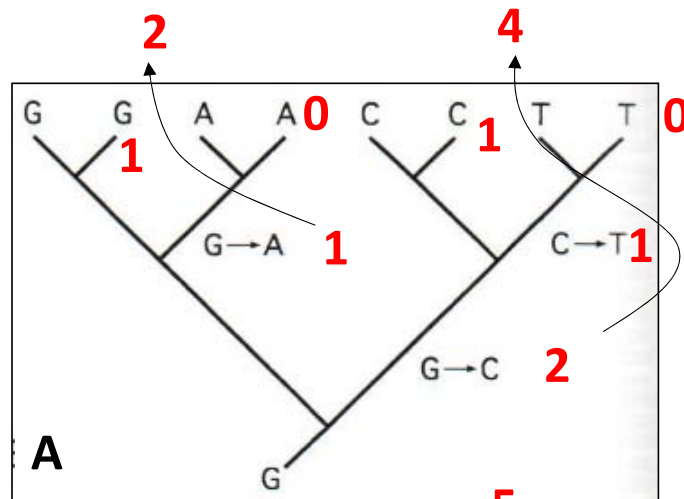
Weighted Parsimony Example



| | A | C | G | T |
|---|---|---|---|---|
| A | - | 2 | 1 | 2 |
| C | 2 | - | 2 | 1 |
| G | 1 | 2 | - | 2 |
| T | 2 | 1 | 2 | - |

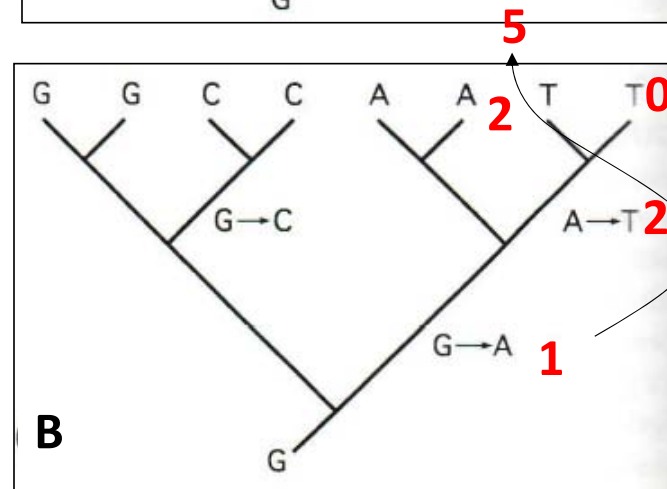
Presumed that a **transversion** costs twice as much as a **transition**.

Weighted Parsimony Example



*Parsimony says A is
The better explanation
(i.e., tree) because it has the
shortest branch lengths*

| | A | C | G | T |
|---|---|---|---|---|
| A | - | 2 | 1 | 2 |
| C | 2 | - | 2 | 1 |
| G | 1 | 2 | - | 2 |
| T | 2 | 1 | 2 | - |



Presumed that a **transversion**
costs twice as much as a
transition.

In-Lecture Exercise

- Which Tree is Better?

| | A | C | G | T |
|---|---|---|---|---|
| A | - | 2 | 1 | 2 |
| C | 2 | - | 2 | 1 |
| G | 1 | 2 | - | 2 |
| T | 2 | 1 | 2 | - |

Consider a site with eight sequences

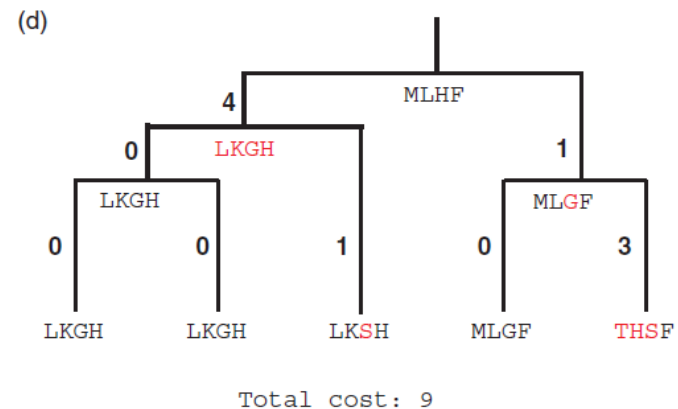
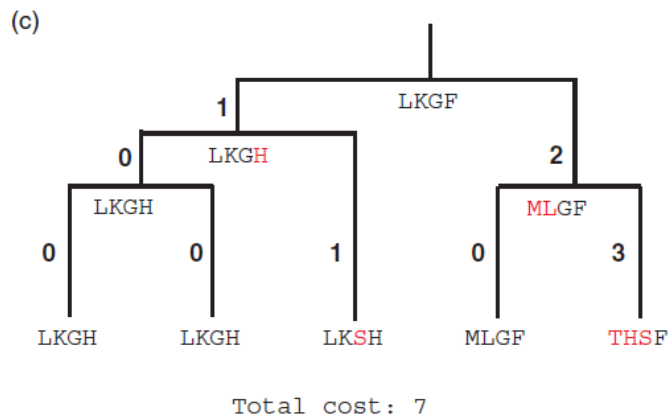
| | | |
|------------|---|--------------|
| | T | T |
| | T | T |
| | C | A |
| A → | C | A ← B |
| | T | A |
| | T | A |
| | G | G |
| | G | G |

Another Parsimony Method: Count Changes (Fig 7.26 in textbook)

(b)

| | |
|-------------------|------|
| kangaroo | LKGH |
| porpoise | LKGH |
| gray seal | LKSH |
| horse α | MLGF |
| kangaroo α | THSF |

Proteins



This Week's (i.e., week 12) HW Assignment

- Read the Section “*Phylogenetic Inference: Maximum Parsimony*” in *textbook*. (pages 287-289)
- Which is the better tree for

| | |
|------------|------|
| kangaroo | LKGH |
| porpoise | LKGH |
| gray seal | LKSH |
| horse α | MLGF |
| kangaroo α | THSF |

- If we start with
 - A: HKFL
 - B: SFKT

Maximum Likelihood (ML) Parsimony

$$\begin{array}{c}
 \begin{array}{c}
 A \\
 C \\
 G \\
 T
 \end{array}
 \begin{bmatrix}
 A & C & G & T \\
 \begin{array}{c}
 -(a_1+a_2+a_3) \quad a_1 \quad a_2 \quad a_3 \\
 a_4 \quad -(a_4+a_5+a_6) \quad a_5 \quad a_6 \\
 a_7 \quad a_8 \quad -(a_7+a_8+a_9) \quad a_9 \\
 a_{10} \quad a_{11} \quad a_{12} \quad -(a_{10}+a_{11}+a_{12})
 \end{array}
 \end{bmatrix}
 \end{array}$$

The off-diagonal values represent a product of an instantaneous rate of change, a relative rate between the different substitutions, and the frequency of the target base. Forward rates (upper triangular values) are presumed to equal the reverse rates corresponding lower triangular values). The diagonal elements are nonzero, which accounts for the possibility that more divergent sequences are more likely to share the same base by chance.

Models of Substitution Rates Between Amino Acids

- PAM Matrices
- BLOSUM matrices
- PROTDIST: A computer program that produces substitution models for proteins.

ML Estimators Most Popular

- **PAUP:** Phylogenetic **A**nalysis **U**sing ML **P**arsimony
(paup.csit.fsu.edu/) ← Currently under development
- Others
 - **Mega2:** (megasoftware.net)
 - **PHYLIP:** (evolution.genetics.washington.edu/phylip.html)
 - **Treeview:** (taxonomy.zoology.gla.ac.uk/rod/treeview.html)
 - **List:** (evolution.genetics.washington.edu/phylip/software.html)