Phylogenetic analysis

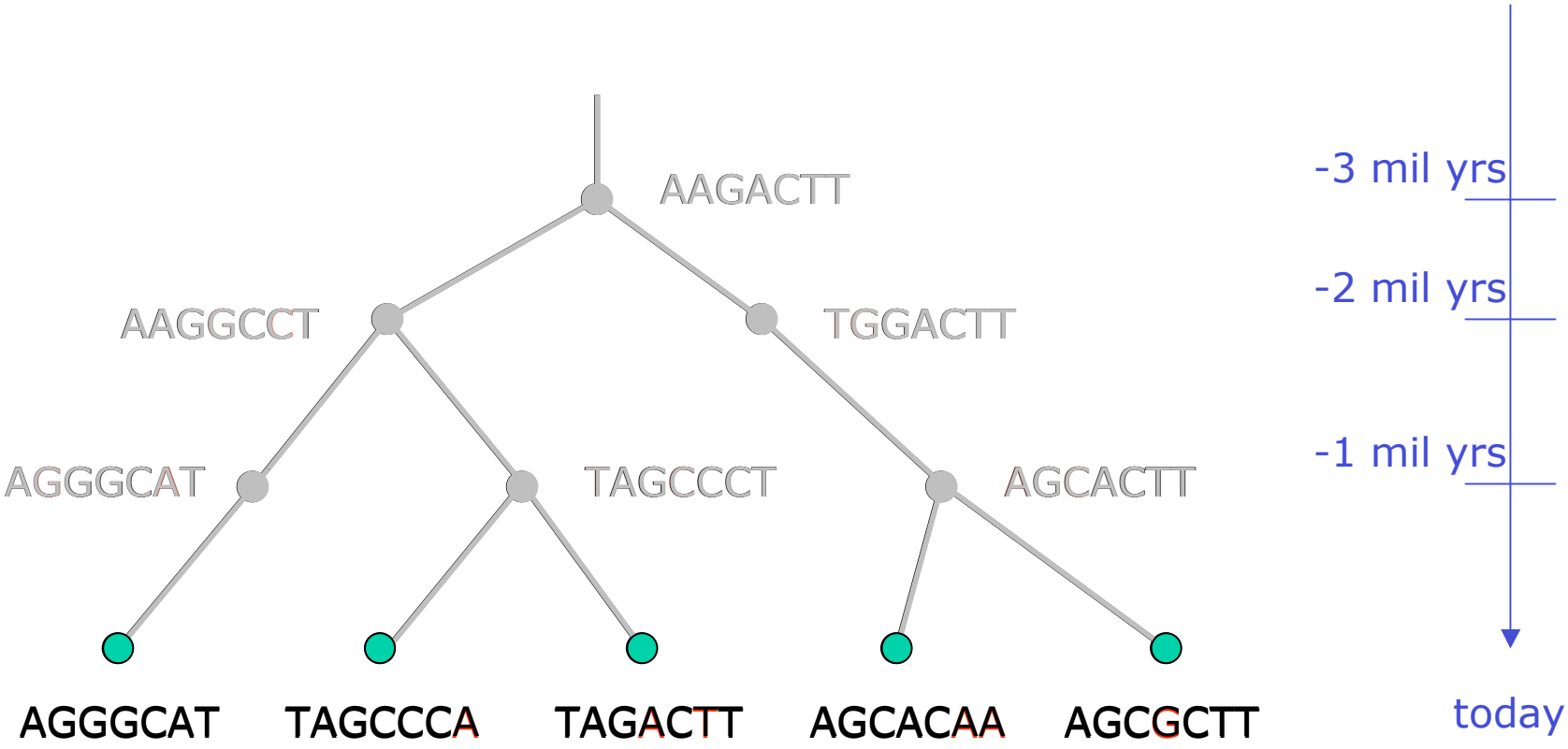
Shi Xiaoli

Jan. 13, 2007

Tree of the life

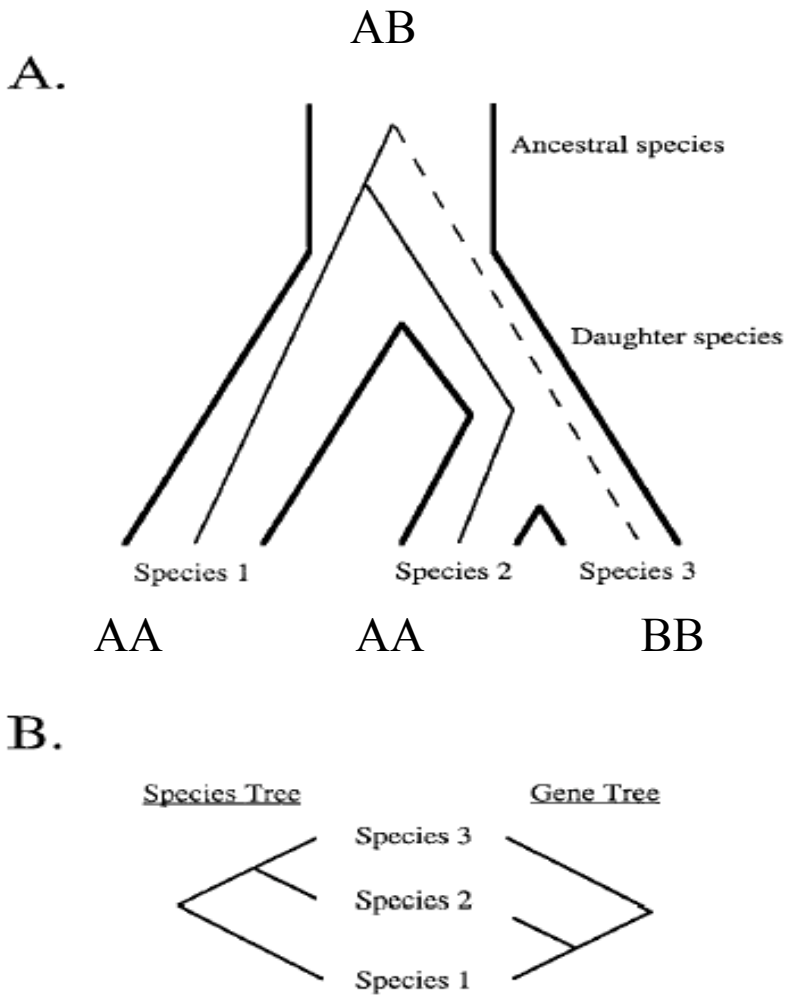
- From http://tolweb.org/tree/

DNA Sequence Evolution



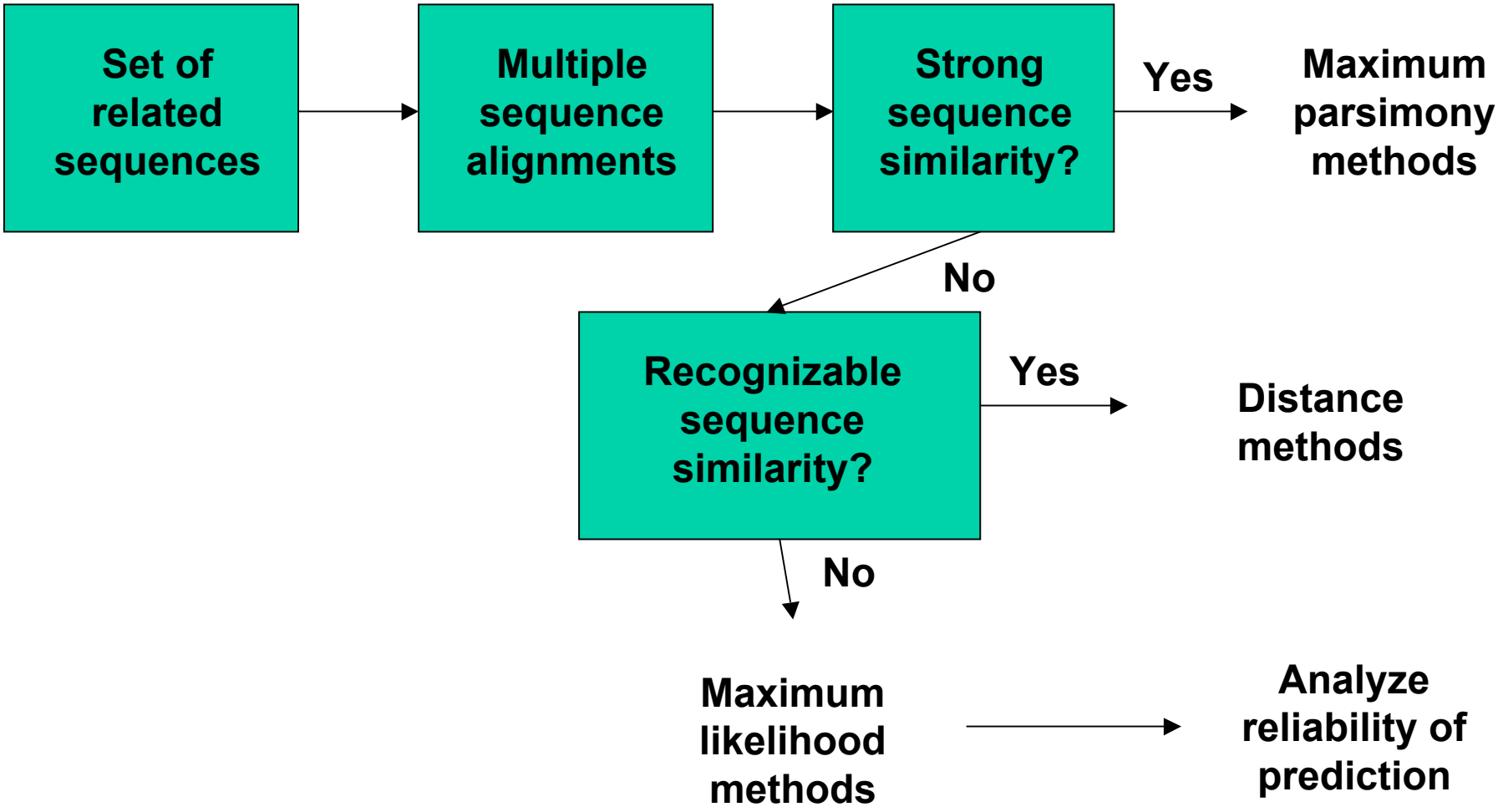
- From Tandy Warnow 2004

Gene Trees and Species Trees

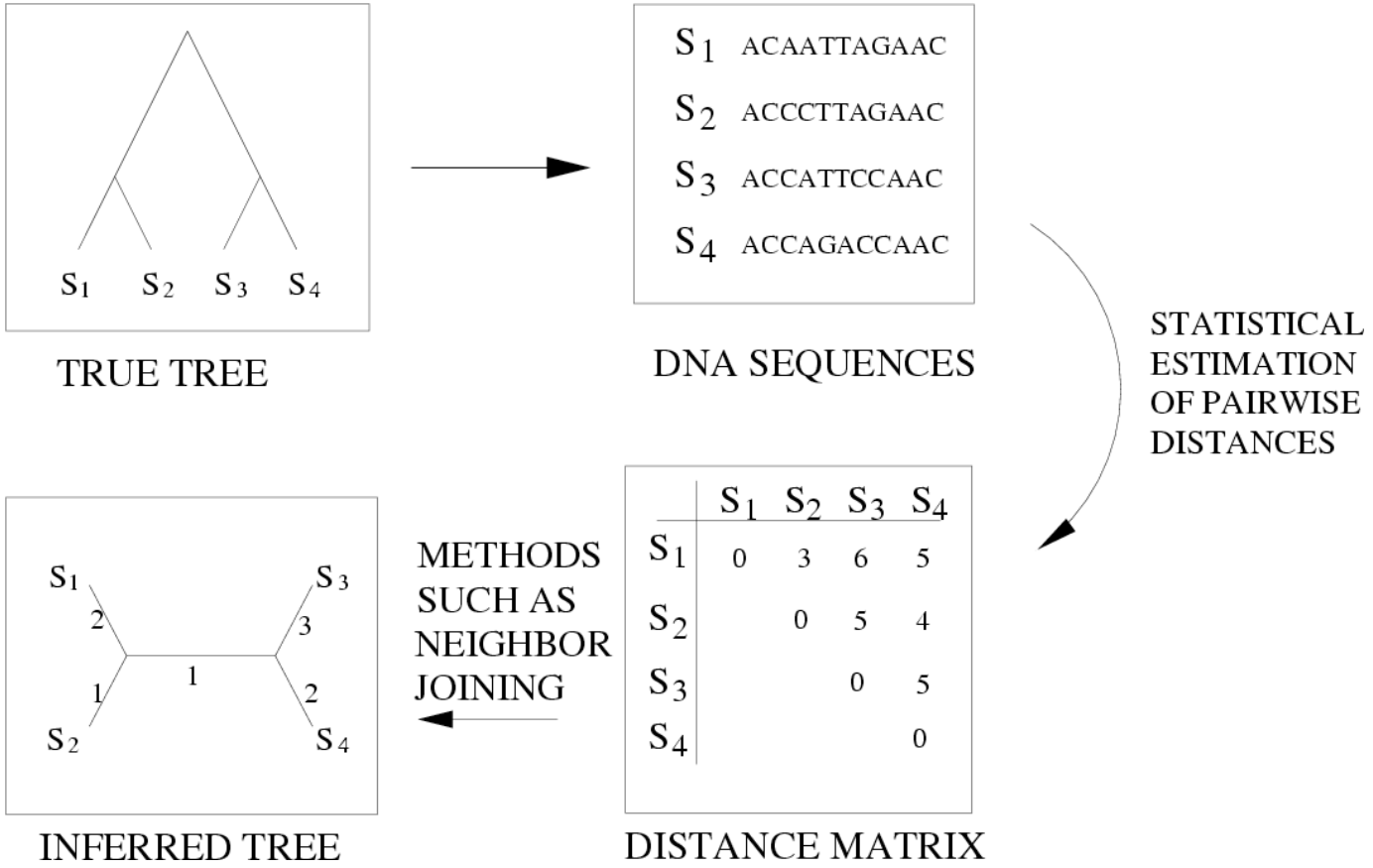


From Doyle and Gaut; Plant Molecular Biology 42:1-23, 2000

Methods for phylogenetic trees construction.



Distance Based Methods



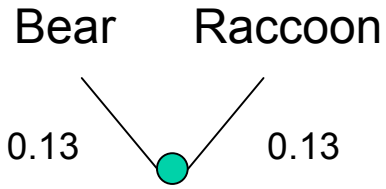
•From Tandy Warnow 2004

Distance Based Methods

- UPGMA algorithm
  (Michener & Sokal 1957)
- ME algorithm
  (Rzhetsky & Nei 1993)
- NJ algorithm
  (Saitou & Nei 1987)

UPGMA algorithm

1. AGGCCATGAATTAAGAATAA
2. AGCCCATGGATAAAGAGTAA
3. AGGACATGAATTAAGAATAA
4. AAGCCAAGAATTACGAATAA



D_{ij}	Bear	Raccoon	Weasel	Seal
Bear	-	**0.26**	0.34	0.29
Raccoon		-	0.42	0.44
Weasel			-	0.44
Seal				-

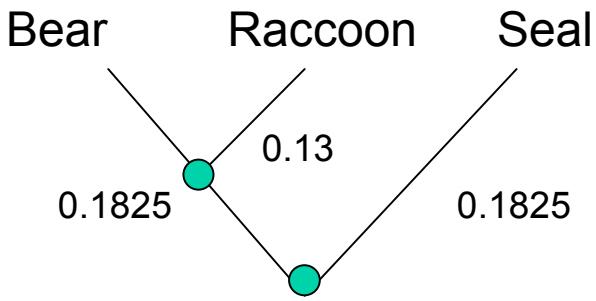
1. Pick smallest entry D_{ij}
2. Join the two intersecting species and assign branch lengths $D_{ij}/2$ to each of the nodes
3. Compute new distances to the other species using arithmetic means

$$D_{W(BR)} = \frac{D_{WB} + D_{WR}}{2} = \frac{0.34 + 0.42}{2} = 0.38$$

$$D_{S(BR)} = \frac{D_{SB} + D_{SR}}{2} = \frac{0.29 + 0.44}{2} = 0.365$$

UPGMA algorithm

D_{ij}	BR	Weasel	Seal
BR	-	0.38	0.365
Weasel		-	0.44
Seal			-

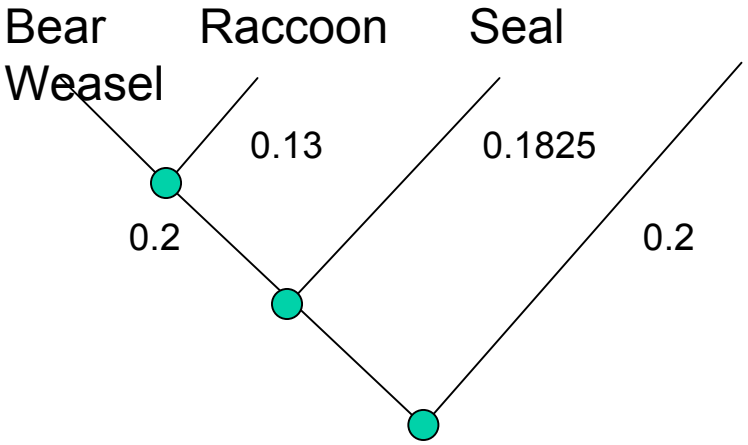


1. Pick smallest entry Dij
2. Join the two intersecting species and assign branch lengths Dij/2 to each of the nodes
3. Compute new distances to the other species using arithmetic means

$$D_{W(BRS)} = \frac{D_{WB} + D_{WR} + D_{WS}}{3} = \frac{0.34 + 0.42 + 0.44}{3} = 0.4$$

UPGMA algorithm

D_{ij}	BRS	Weasel
BRS	-	**0.4**
Weasel		-



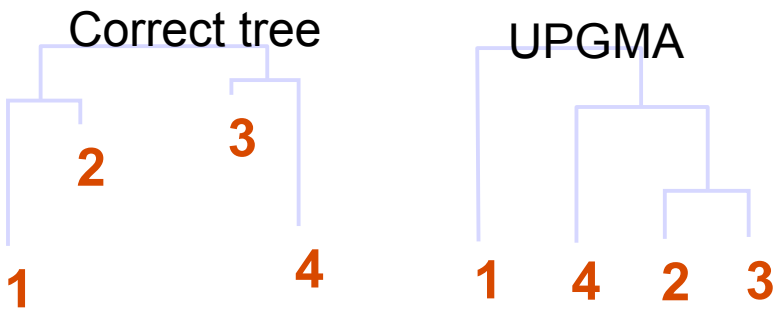
1. Pick smallest entry Dij.

2. Join the two intersecting species and assign branch lengths Dij/2 to each of the nodes.

3. Done!

Neighbor Joining

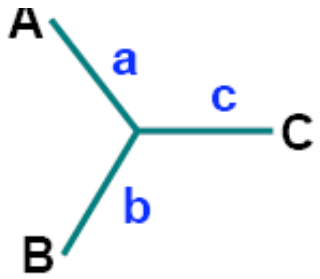
- Goal
  - Join closest ***neighbors*** (nodes with same parent) in tree
  - Avoids problem with UPGMA when rates of change differ
  - Closest leaves are not neighbors in correct tree but are joined first by UPGMA
- Assumptions
  - Rate of change can differ
  - Branch lengths for tree are ***additive***



Neighbor Joining

- Calculating branch length after join (additive tree)

	A	B	C
A	—	$d_{A,B}$	$d_{A,C}$
B		—	$d_{B,C}$
C			—

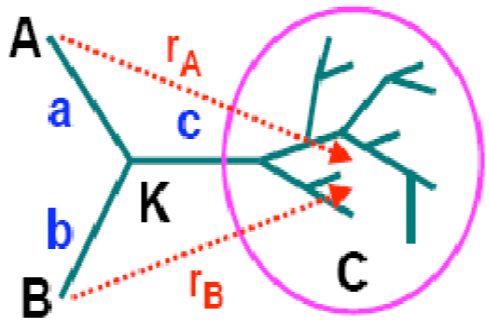


- Simple algebra shows
  - Given
    - $d_{A,B} = a + b$
    - $d_{A,C} = a + c$
    - $d_{B,C} = b + c$
  - We can calculate
    - $a = \frac{1}{2} (d_{A,B} + d_{A,C} - d_{B,C})$
    - $b = \frac{1}{2} (d_{A,B} + d_{B,C} - d_{A,C})$
    - $c = \frac{1}{2} (d_{B,C} + d_{A,C} - d_{A,B})$

Neighbor Joining

- Basic principle for neighbor joining algorithm
  - Simply treat all other nodes as C, and treat distance to C as r.

	A	B	C	
A	—	$d_{A,B}$	$d_{A,C}$	r_A
B		—	$d_{B,C}$	r_B
C			—	



- Replace distance to C
  - Use normalized divergence r_A, r_B(~avg. distance to nodes)
    - $a = \frac{1}{2}(d_{A,B} + d_{A,C} - d_{B,C}) \rightarrow \frac{1}{2}(d_{A,B} + r_A - r_B)$
    - $b = \frac{1}{2}(d_{A,B} + d_{B,C} - d_{A,C}) \rightarrow \frac{1}{2}(d_{A,B} + r_B - r_A)$
    - $c = \frac{1}{2}(d_{B,C} + d_{A,C} - d_{A,B}) \rightarrow \frac{1}{2}(d_{B,C} + d_{A,C} - d_{A,B})$

Neighbor Joining

- Approach
  - To find closest pair of neighbors

    Reduce branch length for a node by the average distance of the node from all other nodes

    Find smallest distance between nodes after reduction.
- Definitions

  For all pairs of nodes A & B in set of all nodes L, let

  $d_{A,B}$ = distance between A,B

  $R_X = \Sigma\, d_{X,N}$ where $N \in L$ (total distance from X to all N)

  $r_X = R_X / (|L| - 2)$, where $|L|$ = # of nodes

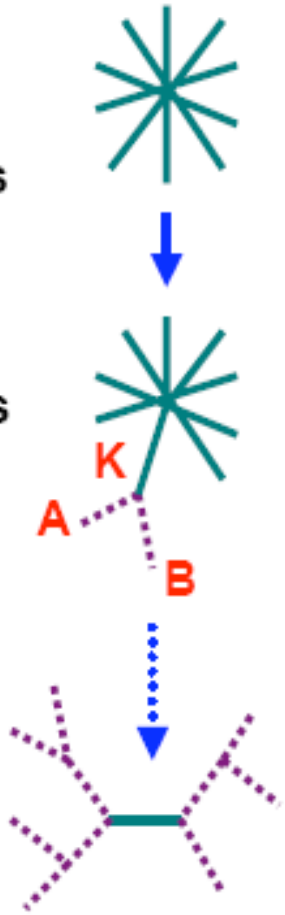
  (normalized divergence from X to all other nodes)

  $D_{A,B} = d_{A,B} - (r_A + r_B)$ (rate-corrected distance)
- Key points- 2 nodes with minimum D are always neighbors

Neighbor Joining

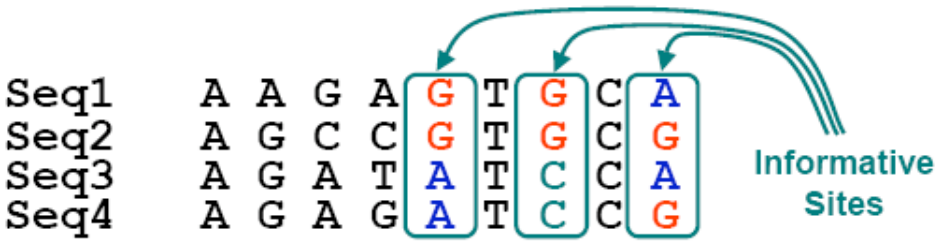
- Algorithm

1. Begin with star tree & all sequences as nodes in L
2. Find pair of nodes A & B $\in$ L with minimum $D_{A,B}$
3. Create & insert new join (node K) w/ branch lengths
   - $d_{A,K} = \frac{1}{2} (d_{A,B} + r_A - r_B)$
   - $d_{B,K} = \frac{1}{2} (d_{A,B} + r_B - r_A)$
4. For remaining nodes C $\in$ L, update distance to K as
   - $d_{K,C} = \frac{1}{2} (d_{A,C} + d_{B,C} - d_{A,B})$
5. Insert K and remove A, B from L
6. Repeat steps 2–5 until only two nodes left



Maximum parsimony

- Maximum parsimony
  - Minimize number of sequence change in tree
  - Assume fewest changes is most likely
- Informative site
  - Must have more than 2 different bases and each of them at least appears in two sequences.



Maximum parsimony

- Most parsimony tree
  - Tree with fewest total number of changes at informative sites.

Maximum parsimony

Weakness:

- Misleading if rates of changes vary among branches
- long branch attraction which is random similarity due to long periods of divergence among ***some numbers*** of clades.

Maximum Likelihood

- Goal
  - Given the probability P(x|y,t) for a sequence y to evolve (mutant) to sequence x along an edge of length t (time)
  - Find tree that has highest probability of taking place
- Mutation probabilities
  - Bases: Jukes-Cantor model; Kimura two parameters model
  - Amino acids: PAM; BLOSUM.
- Algorithm
  - Search over all tree topologies & sequences alignments
  - For each topology & assignment, search all branch lengths.
- Characteristics
  - Very computationally expensive

Robustness and reliability

- Bootstrapping is a statistical technique that can use random resampling of data to determine sampling error for tree topologies
- Bootstrapp proportions are not the same as confidence intervals. There is no simple mapping between bootstrap values and confidence intervals.

Computer Software for Phylogenetics

- **PHYLIP** is a free package that includes 30 programs that compute various phylogenetic algorithms on different kinds of data.
- The **GCG** package (available at most research institutions) contains a full set of programs for phylogenetic analysis including simple distance-based clustering and the complex **cladistic** analysis program **PAUP** (**P**hylogenetic **A**nalysis **U**sing **P**arsimony)
- **CLUSTALW** is a multiple alignment program that includes the ability to create tress based on **Neighbor Joining.**
- **MrBayes** is a program for **Bayesian inference** of phylogeny using Markov Chain Monte Carlo methods.
- **MEGA** is an integrated tool for automatic and manual sequence alignment, inferring **phylogenetic** trees, mining web-based databases, estimating rates of molecular evolution, and testing evolutionary hypotheses.

Thanks