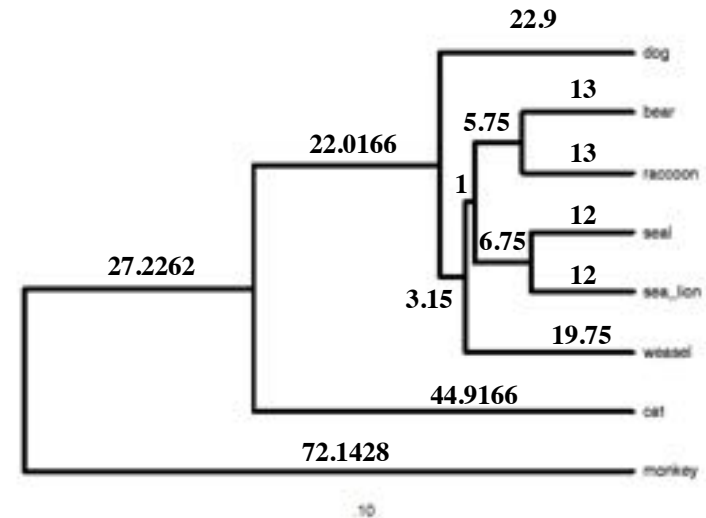


Introduction to Phylogenetics BCBio444

Dennis Lavrov (dlavrov@iastate.edu)

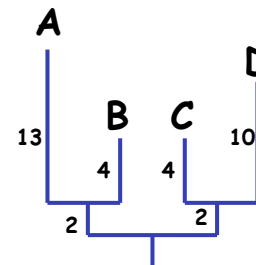
UPGMA: the tree



UPGMA

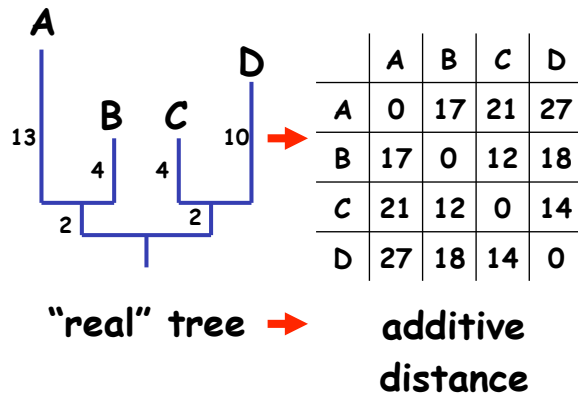
- "Unweighted Pair Group Method Using Arithmetic Averages"
- Works by clustering least distant groups
- Constructs an ultrametric (clocklike) tree
- It is not guaranteed to find the least square ultrametric phylogeny, but it often does quite well

UPGMA: the problem

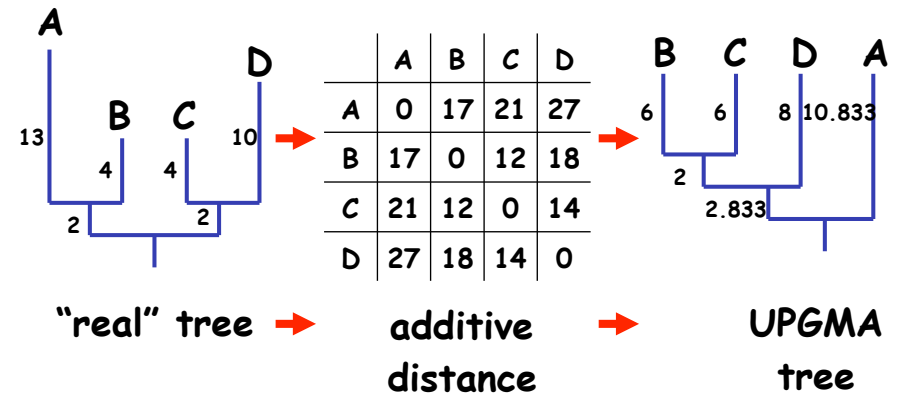


"real" tree

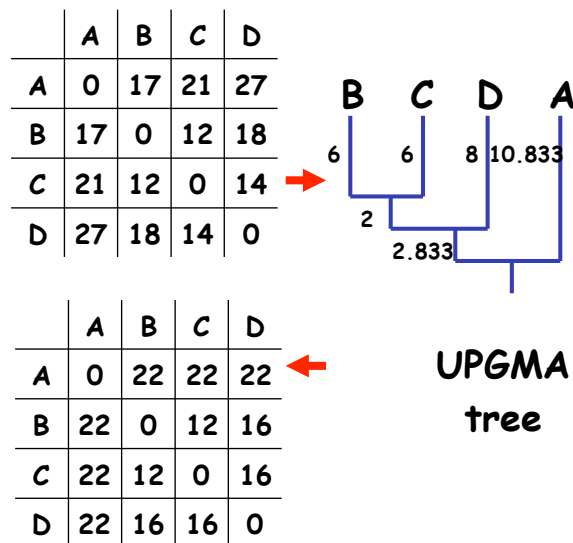
UPGMA: the problem



UPGMA: the problem



UPGMA: the problem



The least squares methods

$$Q = \sum_{i=1}^n \sum_{j=1}^n w_{ij} (D_{ij} - d_{ij})^2$$

Where w_{ij} are weights:

$w_{ij}=1$ (Cavalli-Sforza and Edwards, 1967)

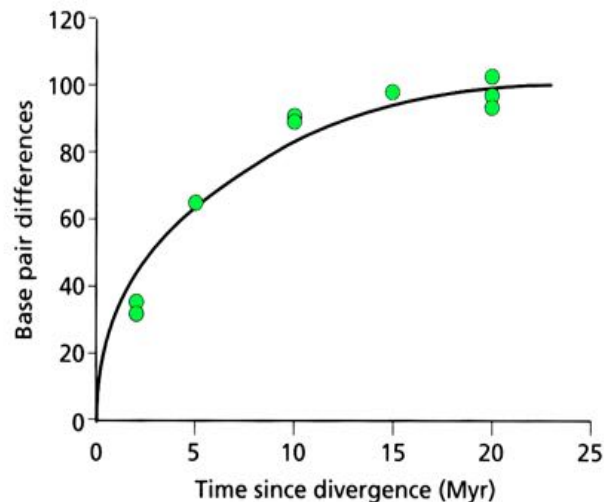
$w_{ij}=1/D_{ij}^2$ (Fitch and Margoliash, 1967)

$w_{ij}=1/D_{ij}$ (Beyer et al., 1974)

(Longer distances associated with larger errors)

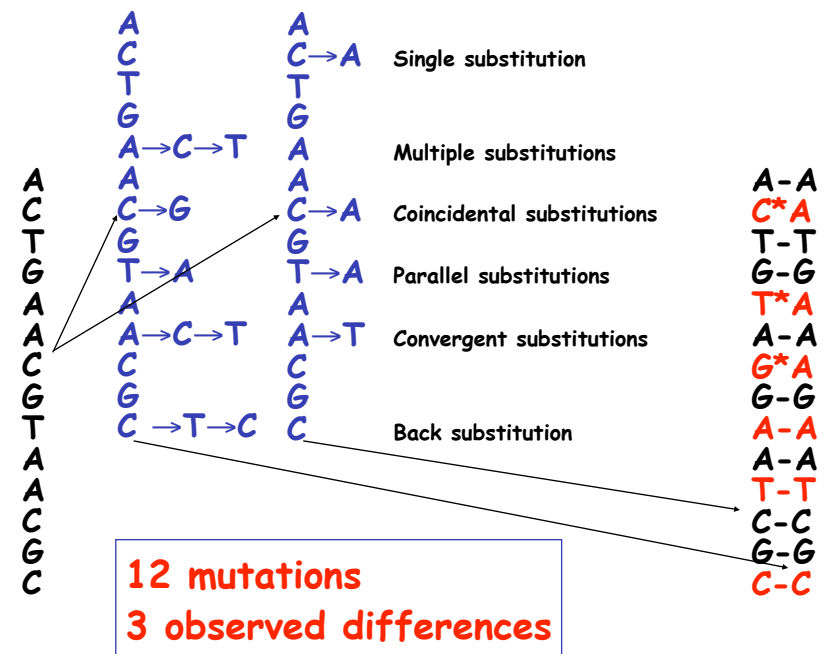
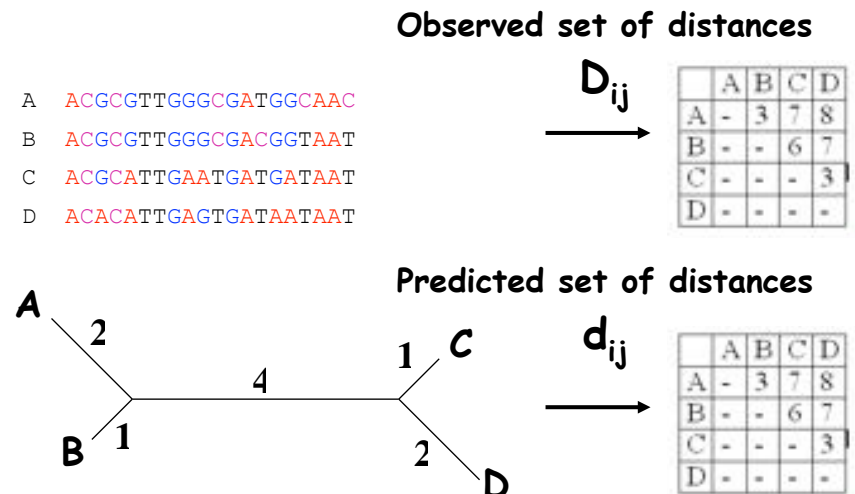
Neighbor-joining

- Developed by Saitou and Nei in 1987
- Does not assume a clock
- Approximates the minimum evolution (at each stage of taxon clustering a ME principle is used)
- Is guaranteed to recover the true tree if the distance matrix happens to be an exact reflection of a tree.



Relationship between seq. difference and the time elapsed since divergence is not linear

Another problem



Jukes and Cantor

$$P_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix} \quad \mathbf{f} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

substitution probability matrix

base composition vector

$$d = -3/4 \ln(1 - 4/3 p)$$

The natural logarithm \ln is used to correct for superimposed changes at the same site

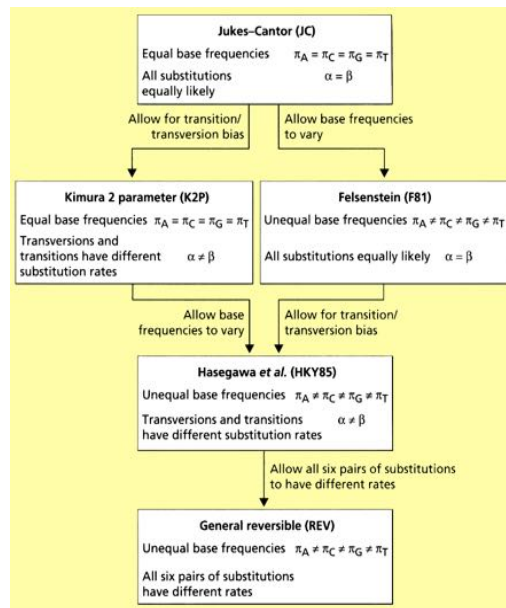
- If two sequences are 95% identical they are different at 5% or 0.05 (p) of sites thus:

$$d_{xy} = -3/4 \ln(1 - 4/3 \cdot 0.05) = 0.0517$$

- However, if two sequences are only 50% identical they are different at 50% or 0.50 (p) of sites thus:

$$d_{xy} = -3/4 \ln(1 - 4/3 \cdot 0.5) = 0.824$$

Summary Comparison of Models



Maximum Parsimony

Pluralitas non est ponenda sine necessitas (plurality shouldn't be posited without necessity)

- William of Ockham, English monk (c. 1285-1349)

- Ockham's razor – best hypothesis to explain a phenomenon is the one that requires the smallest number of assumptions (i.e. is the most parsimonious)
- First mentioned by Edwards and Cavalli-Sforza (1963)
- Later adopted for sequence data (Eck & Dayhoff 1966)
- residues treated as character states

Maximum Parsimony

- Depends on the idea of the fit of a character to a tree
- Initially, we can define the fit of a character to a tree as the minimum number of steps required to explain the observed distribution of character states among taxa
- The sum of steps over all characters is called **Tree Length**
- **Most parsimonious trees (MPTs)** have the minimum tree length needed to explain the observed distributions of all the characters

Two questions:

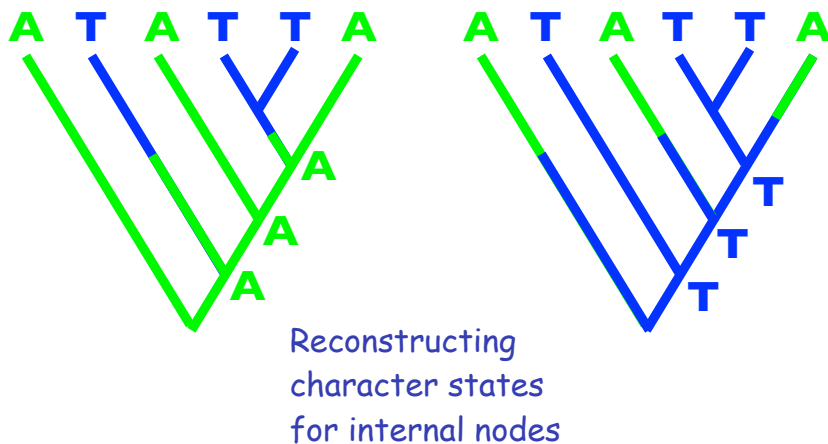
How to count (optimize) steps on a given tree?

Need an algorithm

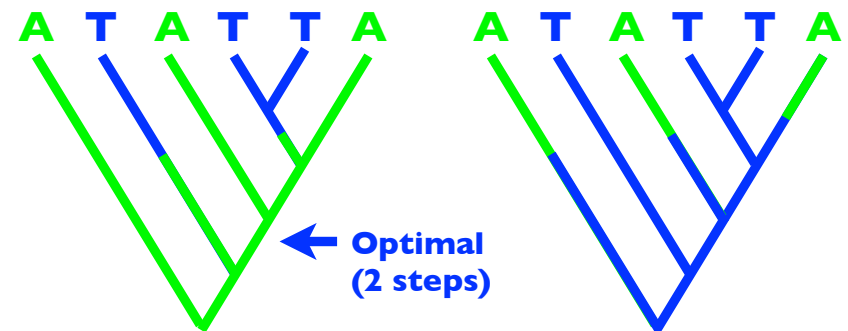
How to evaluate different trees?

Too many trees - which ones to evaluate?

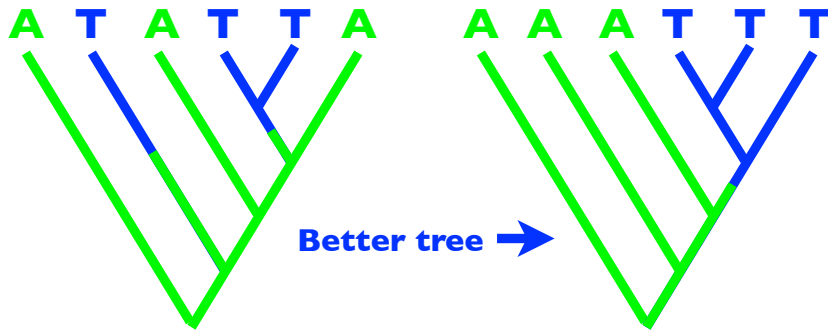
Character Optimization



Character Optimization



Tree Search



Parsimony types

- Characters may differ in the costs (contribution to tree length) made by different kinds of changes
- **Wagner** - ordered, additive
0 — 1 — 2 (morphology, unequal costs)
- **Fitch** - unordered, non-additive
 (equal costs for all changes)
- **Sankoff** - generalized parsimony (variable costs for different kinds of change)

Character State Optimization

(The Sankoff algorithm)

A dynamic programming algorithm for counting the smallest number of possible (weighted) state changes needed on a given tree.

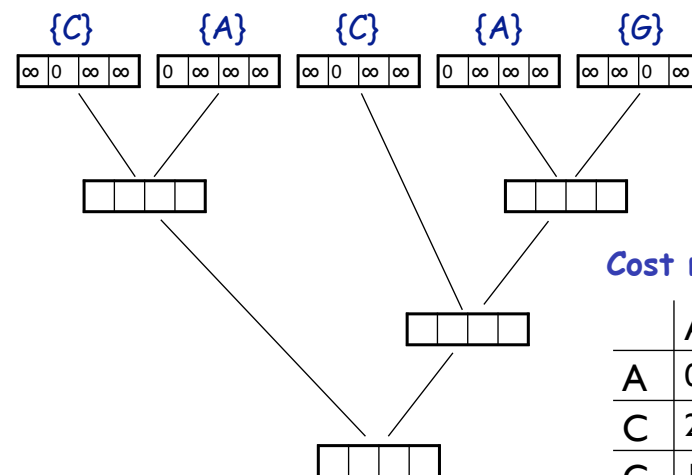
Let $S_j(i)$ be the smallest (weighted) number of steps needed to evolve the subtree at or above node j , given that node j is in state i .

Suppose that c_{ij} is the cost of going from state i to state j . Initially, at tip (say) j

$$S_j(i) = \begin{cases} 0 & \text{if node } j \text{ has (or could have) state } i \\ \infty & \text{if node } j \text{ has any other state} \end{cases}$$

Character State Optimization

(The Sankoff algorithm)



Cost matrix:

	A	C	G	T
A	0	2.5	1	2.5
C	2.5	0	2.5	1
G	1	2.5	0	2.5
T	2.5	1	2.5	0

Character State Optimization

(The Sankoff algorithm)

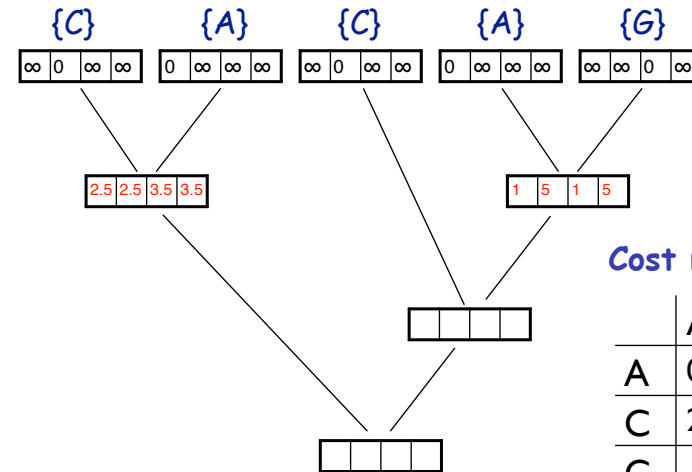
Then proceeding down the tree (postorder tree traversal) for node a whose immediate descendants are b and c

$$S_a(i) = \min_j [c_{ij} + S_b(j)] + \min_k [c_{ik} + S_c(k)]$$

The minimum number of (weighted) steps for the tree is found by computing at the bottom node (0) the $S_0(i)$ and taking the smallest of these.

Character State Optimization

(The Sankoff algorithm)

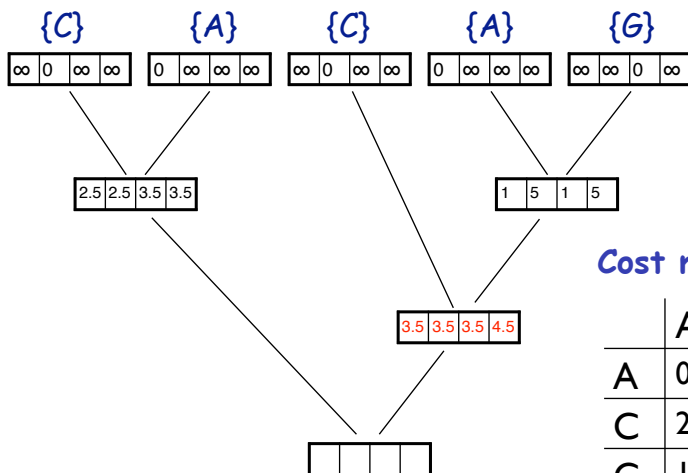


Cost matrix:

	A	C	G	T
A	0	2.5	1	2.5
C	2.5	0	2.5	1
G	1	2.5	0	2.5
T	2.5	1	2.5	0

Character State Optimization

(The Sankoff algorithm)

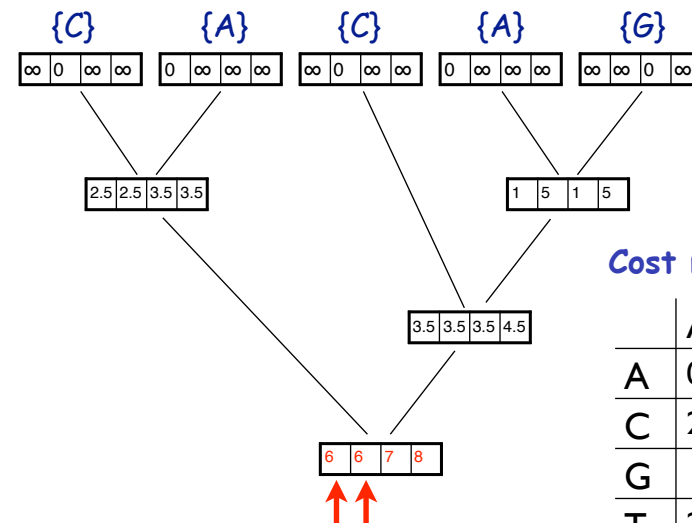


Cost matrix:

	A	C	G	T
A	0	2.5	1	2.5
C	2.5	0	2.5	1
G	1	2.5	0	2.5
T	2.5	1	2.5	0

Character State Optimization

(The Sankoff algorithm)



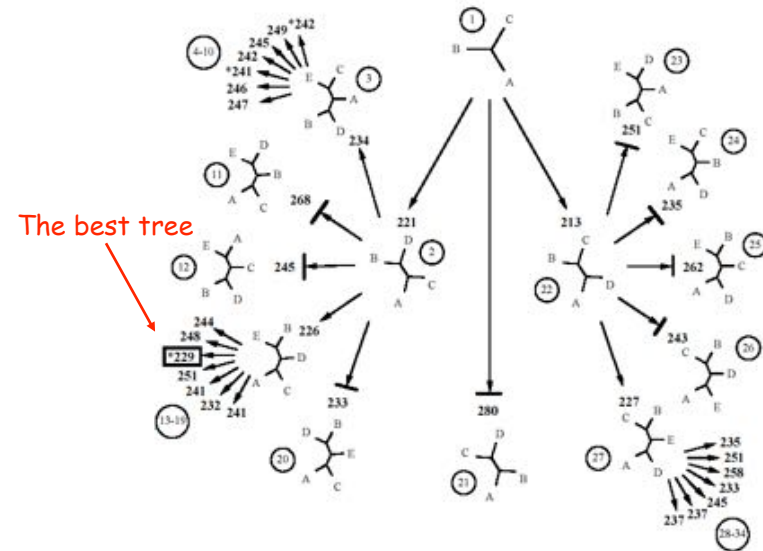
Cost matrix:

	A	C	G	T
A	0	2.5	1	2.5
C	2.5	0	2.5	1
G	1	2.5	0	2.5
T	2.5	1	2.5	0

Branch and bound search

Tree Search Options:

- Exhaustive Search
- Branch and Bound Search
- Heuristic Searches



Heuristic search

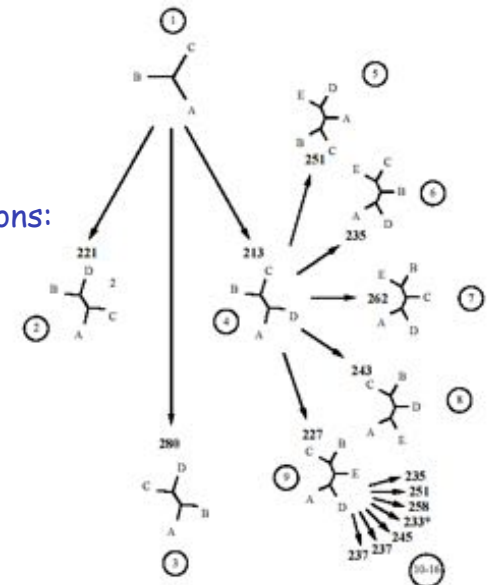
- Obtain a starting tree
 - stepwise addition
 - star decomposition
 - random
- Rearrange this tree
 - don't bother
 - NNI
 - SPR
 - TBR

Heuristic search

Stepwise addition only
No branch swapping

Stepwise Addition Options:

Simple
Closest
As Is
Random

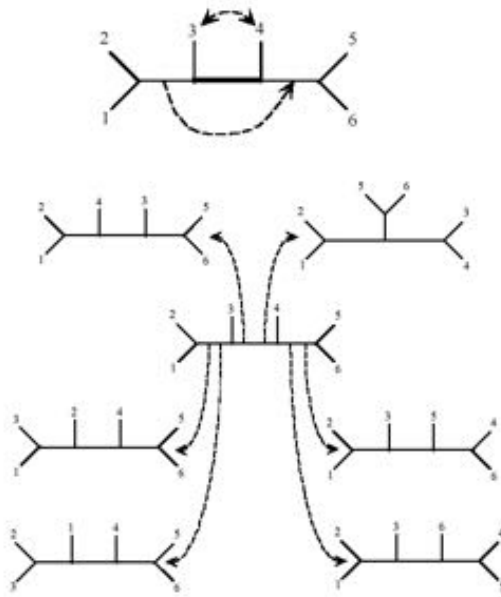


Heuristic search

Nearest-neighbor interchange (NNI)

• erases an interior branch on the tree, and the two branches connected to it

2(n-3) possibilities
How greedy to be?



Maximum Likelihood

The likelihood supplies a natural order of preferences among the possibilities under consideration.

-R.A. Fisher, 1956

An ideal parsimony method (after J. Felsenstein)

Ideally, we'd like to have a parsimony method that:

- Takes into account less parsimonious as well as most parsimonious state reconstructions
- Weights changes differently if they occur in a branch of different length
- Weights different kinds of events (e.g. transitions, transversions) differently

There is such a method! It is maximum likelihood.

Given two explanations for a particular outcome which should we choose?

- The explanation that makes the observed outcome more likely...
- More formally if given some data D and a hypothesis H the likelihood of that data is given by $L_H = \Pr(D | H)$
- Which is the probability of D given H .



"How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth."

Sherlock Holmes to Dr. Watson in
The Sign of Four, by A. Conan Doyle.

Three dice example



6 face die



8 face die



12 face die

A person rolls two dice and obtains the score "14"

Which pair of dice are the most likely
to have yielded this result?

Equivalent to: which tree is most likely to have yielded these sequences?

How many ways of obtaining the score "14" are
there for each pair?



$$6 + 8$$

1



$$\begin{aligned} 2 + 12 \\ 3 + 11 \\ 4 + 10 \\ 5 + 9 \\ 6 + 8 \end{aligned}$$

5



$$\begin{aligned} 2 + 12 \\ 3 + 11 \\ 4 + 10 \\ 5 + 9 \\ 6 + 8 \\ 7 + 7 \\ 8 + 6 \end{aligned}$$

7



$$1/6 \times 1/8$$

$$= 1/48$$



$$1/6 \times 1/12$$







$$= 1/72$$



$$1/8 \times 1/12$$

$$= 1/96$$

Now multiply ways of obtaining the score "14" by the probability of any single outcome to get the likelihood.

 + 	 + 	 + 
$1/48 \times 1$	$1/72 \times 5$	$1/96 \times 7$
↓	↓	↓
0.0208	0.0694	0.0729
		maximum likelihood

Notice that none of the likelihoods are very likely, but (8+12) is more likely than the other two

Important to distinguish between likelihood and probability

- Probabilities sum to 1, likelihoods do not.
- Given a tree and a model, could work out probability of obtaining all possible data sets. The sum of these probabilities would = 1
- But we are interested in just one data set - the one we observed
- note: the likelihood is not the probability that the tree is the true tree, merely the probability that the tree gave rise to the DATA.

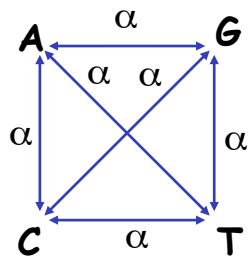
In the context of molecular phylogenetics...

- **D** is the set of sequences being compared
- **H** is a phylogenetic tree + the model of sequence evolution
- We want to find the likelihood of obtaining the observed data given the tree.
- The tree that makes the data the most probable evolutionary outcome is the Maximum Likelihood estimate of the phylogeny.

Applying a maximum likelihood approach in phylogenetic analysis

- **Requires 3 elements:** Tree, model and observed data
- **TWO challenges:** (A) For a given topology, which branch lengths make the data most likely (B) which of all of the possible topologies is the most likely.

Jukes and Cantor



$$P_{ii(t)} = 1/4 + (3/4)e^{-4\alpha t}$$

$$P_{ij(t)} = 1/4 - (1/4)e^{-4\alpha t}$$

Likelihood of the simplest tree

our sequences are only 2 nucleotides long

CT ————— CC

$$L = L_{(1)} \cdot L_{(2)} = [\text{Pr}(C)\text{Pr}(C \rightarrow C)] [\text{Pr}(T)\text{Pr}(T \rightarrow C)] =$$

$$[1/4] [1/4 + 3/4 e^{-4\alpha t}] [1/4] [1/4 - 1/4 e^{-4\alpha t}]$$

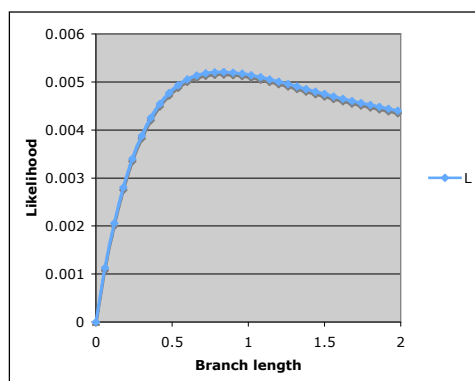
αt - is proportional to the branch length (1/3)

Likelihood of the simplest tree

CT —————^{3 αt} CC

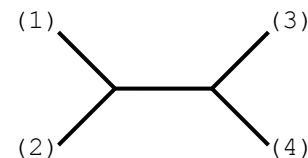
$$L = L_{(1)} \cdot L_{(2)} = [\text{Pr}(C)\text{Pr}(C \rightarrow C)] [\text{Pr}(T)\text{Pr}(T \rightarrow C)] =$$

$$[1/4] [1/4 + 3/4 e^{-4\alpha t}] [1/4] [1/4 - 1/4 e^{-4\alpha t}]$$



The likelihood of a tree

1 j N
(1) C...GGACACGTTTA...C
(2) C...AGACACCTCTA...C
(3) C...GGATAAGTTAA...C
(4) C...GGATAGCCTAG...C



$$L_{(j)} = \text{Prob} \left(\begin{array}{c} C \ C \ A \ G \\ \diagdown \ \diagup \\ A \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \ C \ A \ G \\ \diagup \ \diagdown \\ A \end{array} \right) + \dots + \text{Prob} \left(\begin{array}{c} C \ C \ A \ G \\ \diagdown \ \diagup \\ T \end{array} \right)$$

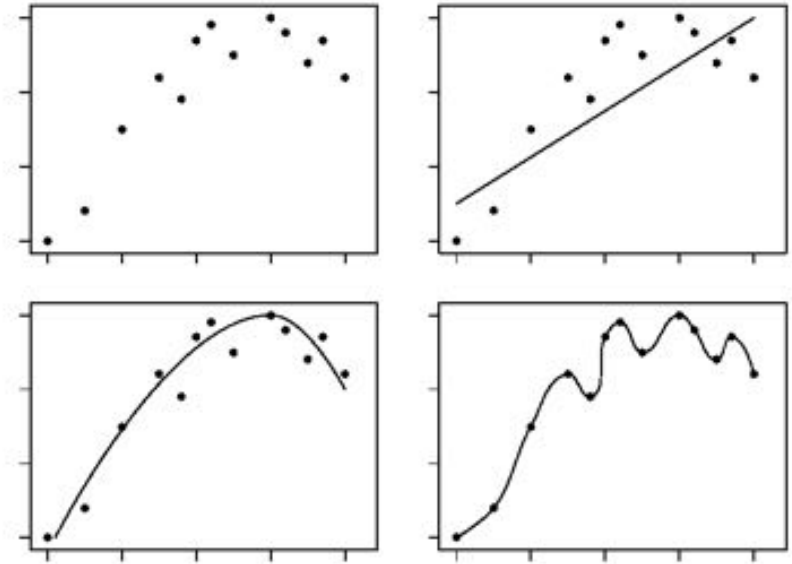
$$L = L_{(1)} \cdot L_{(2)} \cdot \dots \cdot L_{(N)} = \prod_{j=1}^N L_{(j)}$$

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \dots + \ln L_{(N)} = \sum_{j=1}^N \ln L_{(j)}$$

Parameters

- Models differ in their free, i.e. adjustable, parameters
- More parameters are often necessary to better approximate the reality of evolution
- The more free parameters, the better the fit (higher the likelihood) of the model to the data. (Good!)
- The more free parameters, the higher the variance, and the less power to discriminate among competing hypotheses. (Bad!)
- We do not want to “over-fit” the model to the data

What is the best model for these data?



Probability of the hypothesis

- In ML, we choose the hypothesis that gives the highest (maximized) likelihood to the data. The likelihood is the probability of the data given the hypothesis.
- But what we may really like to know is the probability of the hypothesis given the data
- An equation derived by and named after the English mathematician and Presbyterian minister Thomas Bayes shows us how opinions about a hypothesis held before the experiment $\Pr(\theta)$ should be modified by the evidence of the experiment:

$$\Pr(\theta | D) = \frac{\Pr(\theta)\Pr(D | \theta)}{\Pr(D)}$$

What the difference?

- Suppose that we have a rare genetic disease in the human population, which the occurrence rate 1 in 1000.
- Suppose that you bought one of the fancy genomic screening kits, which includes a test for this disease and the test comes out positive (although you have no symptoms yet!)
- You frantically read the instructions and find out that the test has a 99% hit rate, meaning that if a person has the disease, then the test result is positive 99% of the time.
- You also read that the test has a false alarm rate of 5%
- What is the probability that you have a disease?