

Phylogenetic methods

- Parsimony
- Maximum likelihood
- Distance

Weighted parsimony

- Transition/transversion ratio
- Codons
 - 3rd base position
- Stem - loops

Phylogenetic trees

- Efficiency
 - How fast is the method
- Power
 - How much data is needed
- Consistency
 - Will it converge to the correct answer given enough data
- Robustness
 - Will violations of the assumptions result in poor estimates
- Falsifiability
 - Will the method tell us when the assumptions are violated

How do we search such a large tree space?

<i>s</i>	# Unrooted	#Rooted
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425
20	2.2×10^{20}	8.2×10^{21}
50	2.8×10^{74}	2.7×10^{76}
100	1.7×10^{182}	3.4×10^{184}
1000	1.9×10^{2860}	3.8×10^{2863}

PAUP*

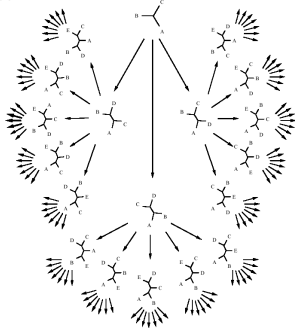
- Phylogenetic Analysis Using Parsimony
- Phylogenetic Analysis Using Parsimony* (*and other methods)
- "Phylogenetic Analysis Using PAUP"
- Available for:
 - Macintosh Classic (OS X/Carbon soon): full GUI interface
 - MS Windows: command-line interface with enhancements (editor, output scrollbar, some menus)
 - Unix/Linux/VMS: "Portable interface"

Capabilities

- Tree searching under parsimony, distance, and maximum-likelihood criteria
 - Exhaustive searching
 - Branch-and-bound
 - Criterion-based heuristics (stepwise addition, branch-swapping, star-decomposition)
 - NJ, UPGMA
- Bootstrapping/jackknifing for assessment of support
- Inference of ancestral states using parsimony and maximum likelihood
- Complete set of DNA substitution models including general time-reversible (GTR=REV) and all of its submodels (JC, K2P=K80, HKY, Tamura-Nei, etc.)
- Import/export of other formats (GCG/MSF, Phylip, Mega, NBRF, etc.)

Searching for trees

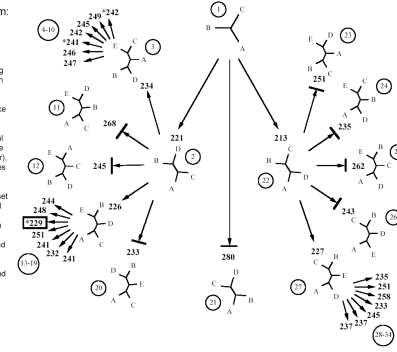
3. Full search tree:



Searching for trees

Branch and bound algorithm:

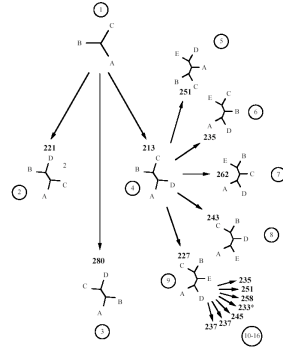
The branch-and-bound algorithm for exact solution of the problem of finding an optimal parsimony tree. The search tree is the same as for exhaustive search, with tree lengths for a hypothetical data set shown in boldface type. If a tree lying at a node of this search tree has a length that exceeds the current lower bound on the optimal tree length, this path of the search tree is terminated (indicated by a cross-bar), and the algorithm backtracks and takes the next available path. When a tip of the search tree is reached (i.e., when we arrive at a tree containing the full set of taxa), the tree is either optimal (and hence retained) or suboptimal (and rejected). When all paths leading from the initial 3-taxon tree have been explored, the algorithm terminates, and all most-parsimonious trees will have been identified. Asterisks indicate points at which the current lower bound is reduced. See text for additional explanation, and circled numbers represent the order in which phylogenetic trees are visited in the search tree.



Searching for trees

Heuristic search methods

A greedy stepwise-addition search applied to the example used for branch-and-bound. The best 4-taxon tree is determined by evaluating the lengths of the three trees obtained by joining taxon D to tree 1 containing only the first three taxa. Taxa E and F are then connected to the five and seven possible locations, respectively, on trees 4 and 9, with only the shortest trees found during each step being used for the next step. In this example, the 233-step tree obtained is not a global optimum. Circled numbers indicate the order in which phylogenetic trees are evaluated in the stepwise-addition search.



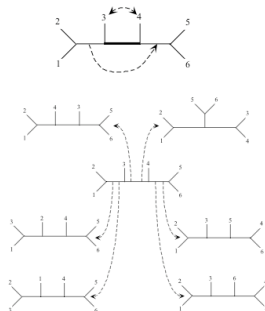
Once we obtain a tree, how do we check nearby trees for a better fit?

Searching for trees

Heuristic search methods continued

Nearest neighbor interchange:

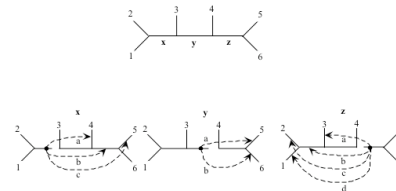
All possible NNIs on 6-taxon tree:



Searching for trees

Heuristic search methods continued

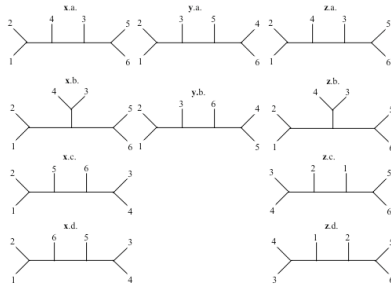
Subtree pruning regrafting:



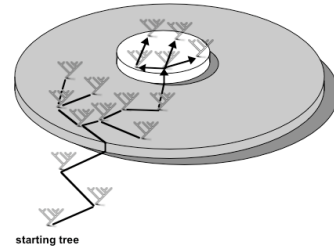
Searching for trees

Heuristic search methods continued

Trees resulting from SPR:

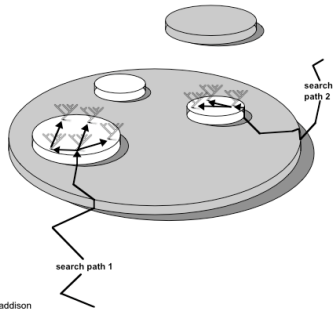


One search path



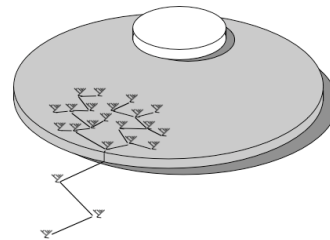
Drawing by David Maddison

Multiple islands



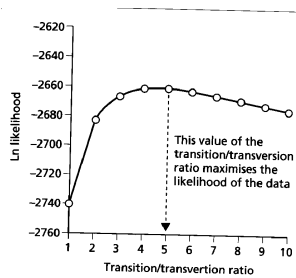
Drawing by David Maddison

Plateaus



Drawing by David Maddison

Maximum Likelihood



Likelihood calculations require...

- An explicit model of substitution that specifies change probabilities for a given branch length:

$$Q = \begin{pmatrix} \pi_A r_{AA} & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{CA} & \pi_C r_{CC} & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{GA} & \pi_C r_{GC} & \pi_G r_{GG} & \pi_T r_{GT} \\ \pi_A r_{TA} & \pi_C r_{TC} & \pi_G r_{TG} & \pi_T r_{TT} \end{pmatrix}$$

Jukes-Cantor
Kimura 2-parameter
Hasegawa-Kishino-Yano (HKY)
Felsenstein 1981, 1984
General time-reversible

- An estimate of optimal branch lengths in units of expected amount of change ($v = \text{rate} \times \text{time}$)

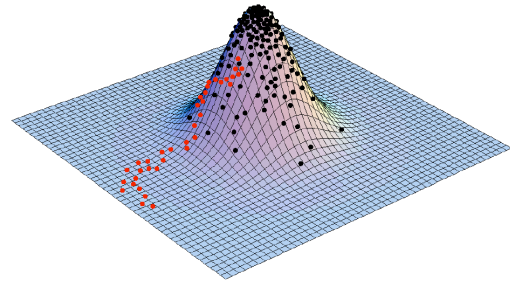
$$P(v) = e^{Qv}$$

Maximum Likelihood

One way to get the likelihood is to estimate them using Markov Chain Monte Carlo methods.

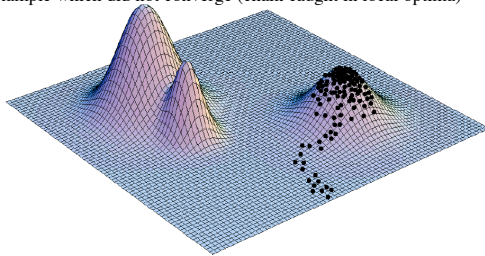
-analogy to walking up hill.

Red dots = "burn in" period.



Convergence = tested all of likelihood surface and found maximum

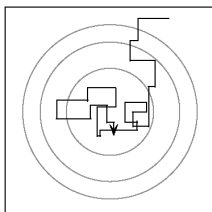
- example which did not converge (chain caught in local optima)



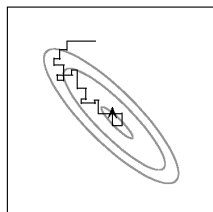
For an appropriately constructed and adequately run Markov chain, the proportion of the time any tree is visited is a valid approximation of the posterior probability of that tree (= Bayes)

MC-Robot demo:

A



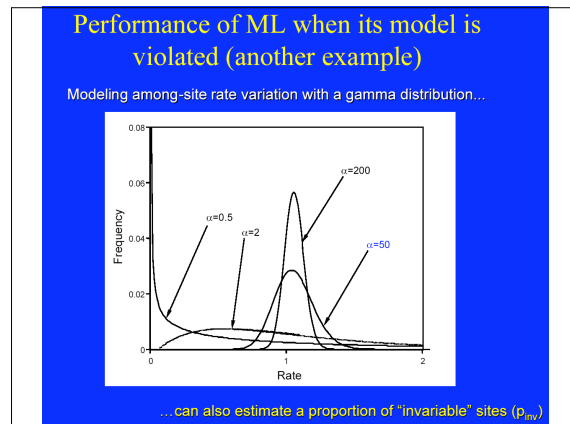
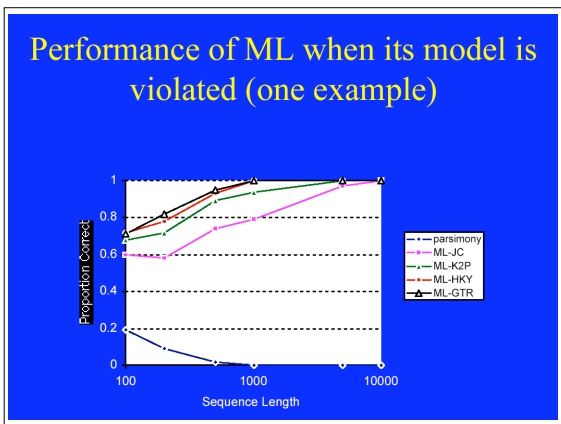
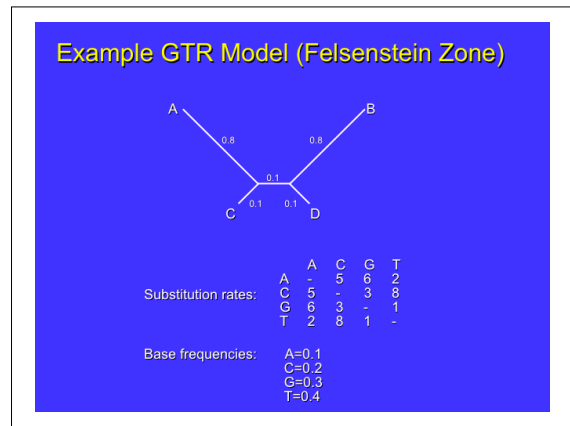
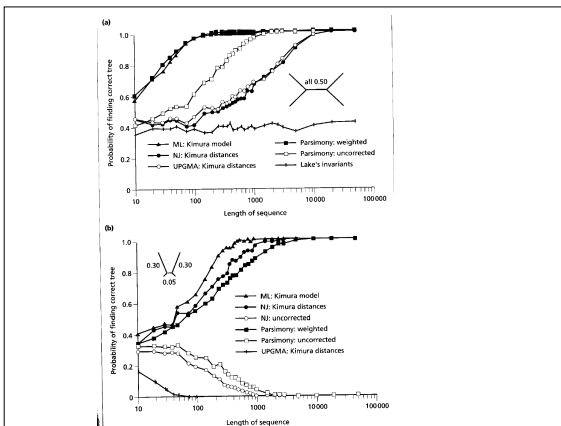
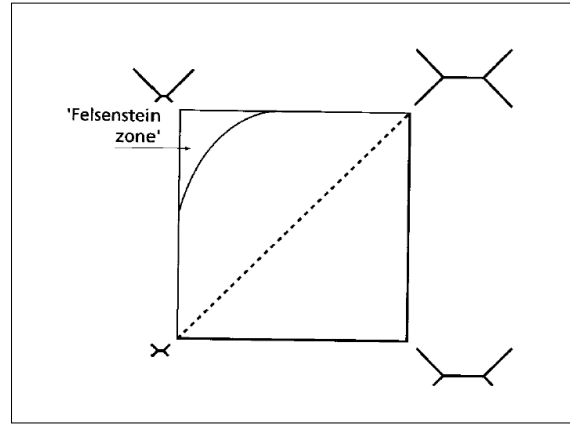
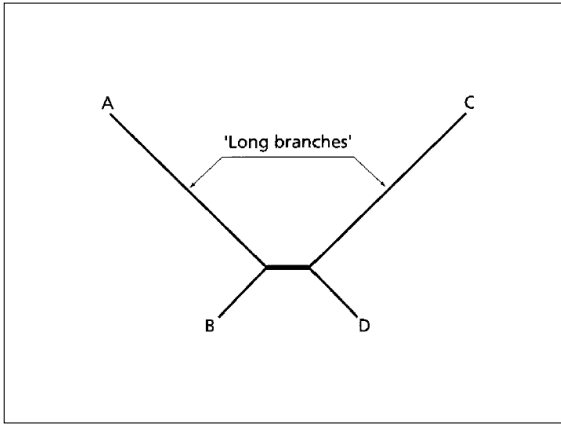
B



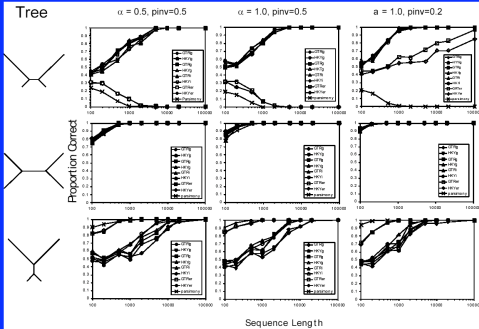
θ_1

How well do these methods perform

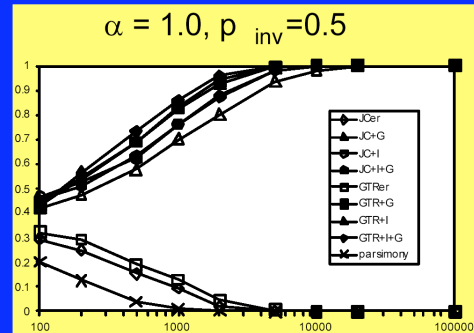
- Simulate datasets.
- Estimate phylogeny using variety of methods and models.
- Compare estimate to known answer.



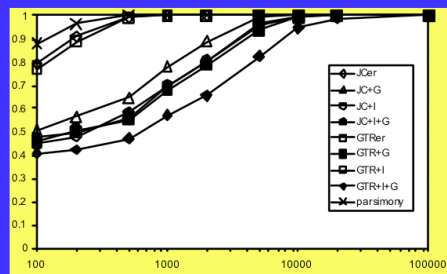
Performance of ML when its model is violated (another example)



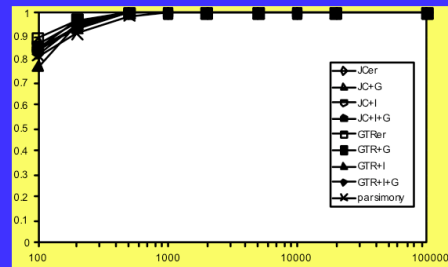
“MODERATE”–Felsenstein zone



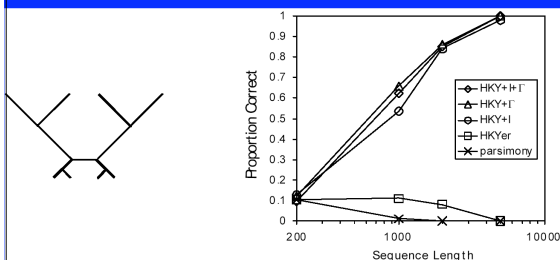
“MODERATE”–Inverse-Felsenstein zone



“MODERATE”–Equal branch lengths



Extension to more taxa...



Estimating Parameters of ML Models

Goal of ML Estimation:

Find the combination of branch lengths, model parameters, and topology that optimizes the likelihood score.

Estimating Parameters of ML Models

Goal of ML Estimation:

Find the combination of branch lengths, model parameters, and topology that optimizes the likelihood score.

Ideal Strategy:

Optimize all these variables on every tree we examine during a tree search.

Successive-approximations Approach

1: Conduct initial analysis (i.e. parsimony) to get initial tree(s).

Successive-approximations Approach

1: Conduct initial analysis (i.e. parsimony) to get initial tree(s).

2: Estimate model parameters on current tree(s).

Successive-approximations Approach

1: Conduct initial analysis (i.e. parsimony) to get initial tree(s).

2: Estimate model parameters on current tree(s).

3: Conduct ML analysis with model parameters fixed to values obtained on previous tree.

Successive-approximations Approach

1: Conduct initial analysis (i.e. parsimony) to get initial tree(s).

2: Estimate model parameters on current tree(s).

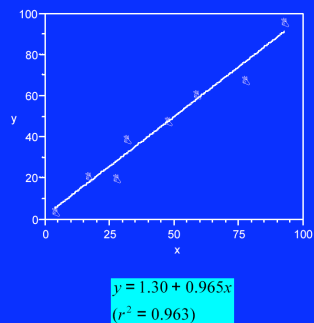
3: Conduct ML analysis with model parameters fixed to values obtained on previous tree.

4: Is topology of new tree identical to that of previous tree?
Yes - Done.
No - Go back to step 2.

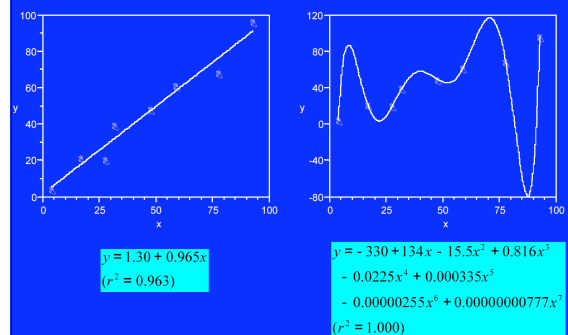
Model Selection

- Desirable to choose models that represent a good compromise between realism and utility
 - Models should be complex enough to incorporate important features of evolutionary process but not so complex that extra parameters absorb random errors (under-fitting versus over-fitting)

Overparameterizing a model

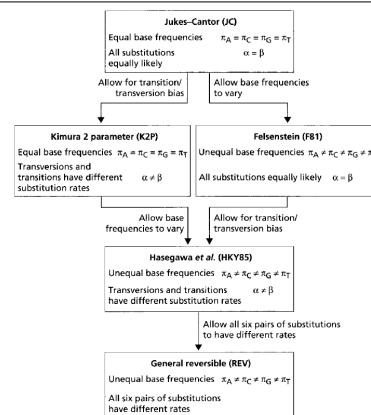


Overparameterizing a model



Model Selection

- Desirable to choose models that represent a good compromise between realism and utility
 - Models should be complex enough to incorporate important features of evolutionary process but not so complex that extra parameters absorb random errors (under-fitting versus over-fitting)
- Formal model selection strategies
 - Likelihood ratio tests
 - Akaike information criterion (AIC)
 - $AIC = -2 \ln L + 2k$ (k = "number of free parameters")



Is the branching order significant?

- Bootstrap test for significance.
- Resample data with replacement
- Reconstruct phylogeneny
- Redo 100s of times
- Build consensus tree, bootstrap values are percent on times that branch was recovered.

ORIGINAL DATA

AAATCTGCTAGCATGCA
AAATCTGCTACCATGCA

BOOTSTRAPPED DATA SETS

AAAAACCCGGCTCAGCA
AAAAACCCGGCTCAGCA

AATTTTCCCATTTGCC
AATTTTCCCATTTGCC

AATTTCACCAACCCTCA
AATTTCACCAACCCTCA

AATTGCCCCCCCCAAA
AATTGCCCCCCCCAAA

AAATCCCGGTACTTGCA
AAATCCCGGTACTTGCA

AATTGGAAACCAAGGCA
AATTGGAAACCAAGGCA

AAAATCCCATGGGAA
AAAATCCCATGGGAA

AATTGGCACCCAAAAC
AATTGGCACCCAAAAC

AAATTCCTCCTGGCAA
AAATTCCTCCTGGCAA

AAAATTCCACCATTCGA
AAAATTCCACCATTCGA

PAUP* demo