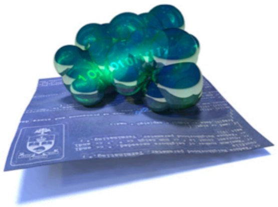
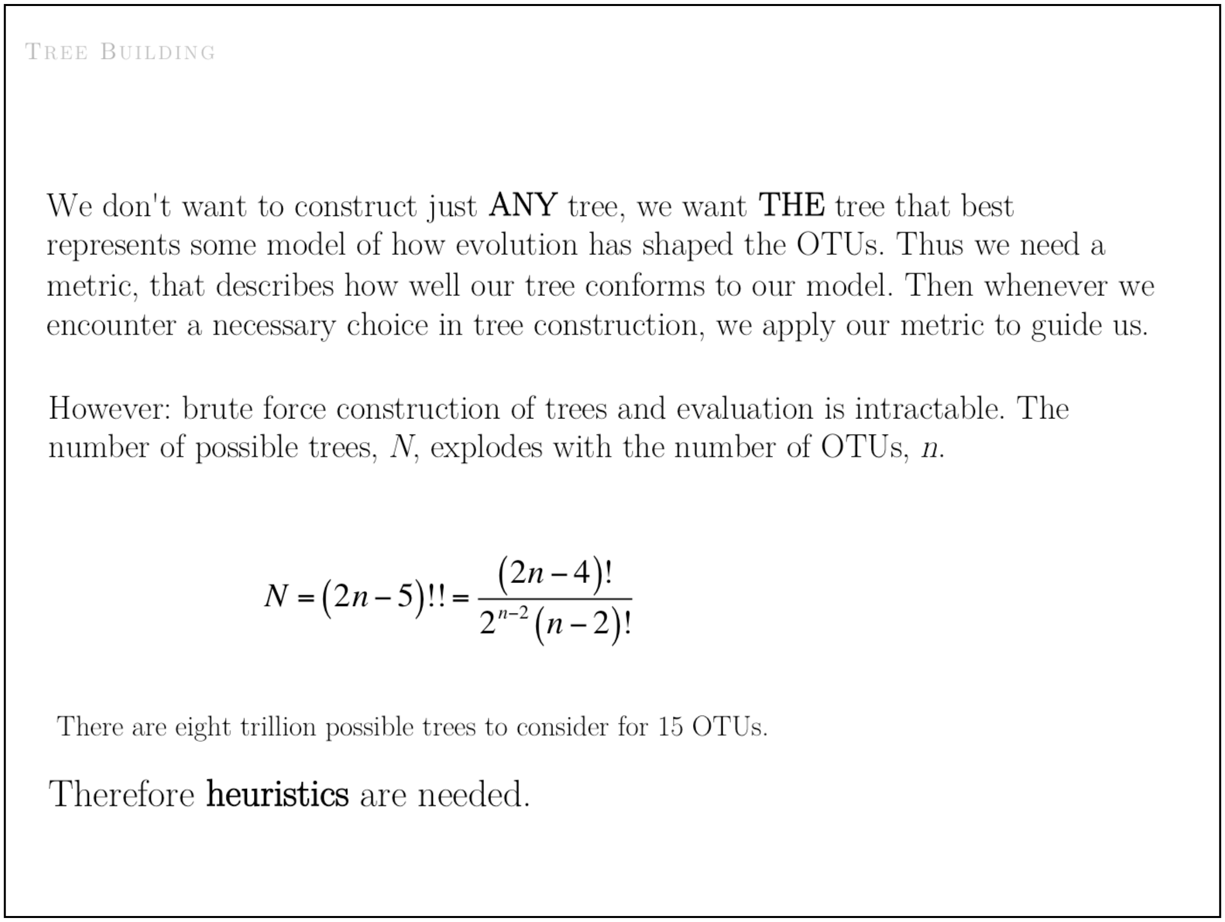
A
Bioinformatics
Course

Building Phylogenetic Trees



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



TREE BUILDING

We don't want to construct just **ANY** tree, we want **THE** tree that best represents some model of how evolution has shaped the OTUs. Thus we need a metric, that describes how well our tree conforms to our model. Then whenever we encounter a necessary choice in tree construction, we apply our metric to guide us.

However: brute force construction of trees and evaluation is intractable. The number of possible trees, N, explodes with the number of OTUs, n.

$$N = (2n-5)!! = \frac{(2n-4)!}{2^{n-2}(n-2)!}$$

There are eight trillion possible trees to consider for 15 OTUs.

Therefore **heuristics** are needed.

The formula refers to the number of possible unrooted binary trees with n labelled leaves. Such a tree with one, two or three leafs can only be formed in a single way. For $n \geq 3$, the number of edges is $(2n–5)$. Adding a node to a tree with n leaves can therefore be done in $2n–5$ places.

We don't want to construct just ANY tree, we want THE tree that best represents some model of how evolution has shaped the OTUs. Thus we need a metric, that describes how well our tree conforms to our model. Then whenever we encounter a necessary choice in tree construction, we apply our metric to guide us. However, we can phrase our objective in different ways.

The best tree ensures that the *most similar* OTUs share direct ancestors.

The best tree minimizes the *number of evolutionary events* in the tree.

The best tree maximizes the *likelihood* of the observed *alignment*.

The best tree maximizes the probability of the tree, given the alignment.

We don't want to construct just **ANY** tree, we want **THE** tree that best represents some model of how evolution has shaped the OTUs. Thus we need a metric, that describes how well our tree conforms to our model. Then whenever we encounter a necessary choice in tree construction, we apply our metric to guide us. However, we can phrase our objective in different ways.

The best tree ensures that the *most similar* OTUs share direct ancestors.

The best tree minim *nts* in the tree.

The best tree maxin *alignment.*

The best tree maxin en the alignment.

Distance based methods:

Consider aggregate properties of the individual sequences.

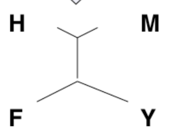
Distance Methods

Distance methods count the number of changes required between each pair of sequences.

1. Compute distance matrix
2. Closest species are "neighbours" and share ancestral node
3. Build tree (several options)

```
Human ... A G A G A T C C G ...
Mouse ... A G A T A T C C A ...
Fugu  ... A G C C G T G C G ...
Yeast ... A A G A G T G C A ...
```

	H	M	F	Y
H		2	4	6
M			5	5
F				4
Y				

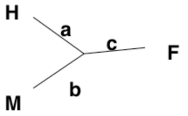


Distance Methods: Fitch/Margoliash

The Fitch/Margoliash method of tree building minimizes branch lengths:

1. Find closest neighbors A, B
2. Calculate distance from A and B to all other sequences combined.
3. Compute branch lengths
4. Combine A, B and repeat

	H	M	F	Y
H		2	4	6
M			5	5
F				4
Y				



(1) HM = a+b = 2

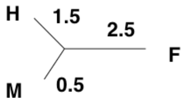
(2) HF = a+c = 4

(3) MF = b+c = 5

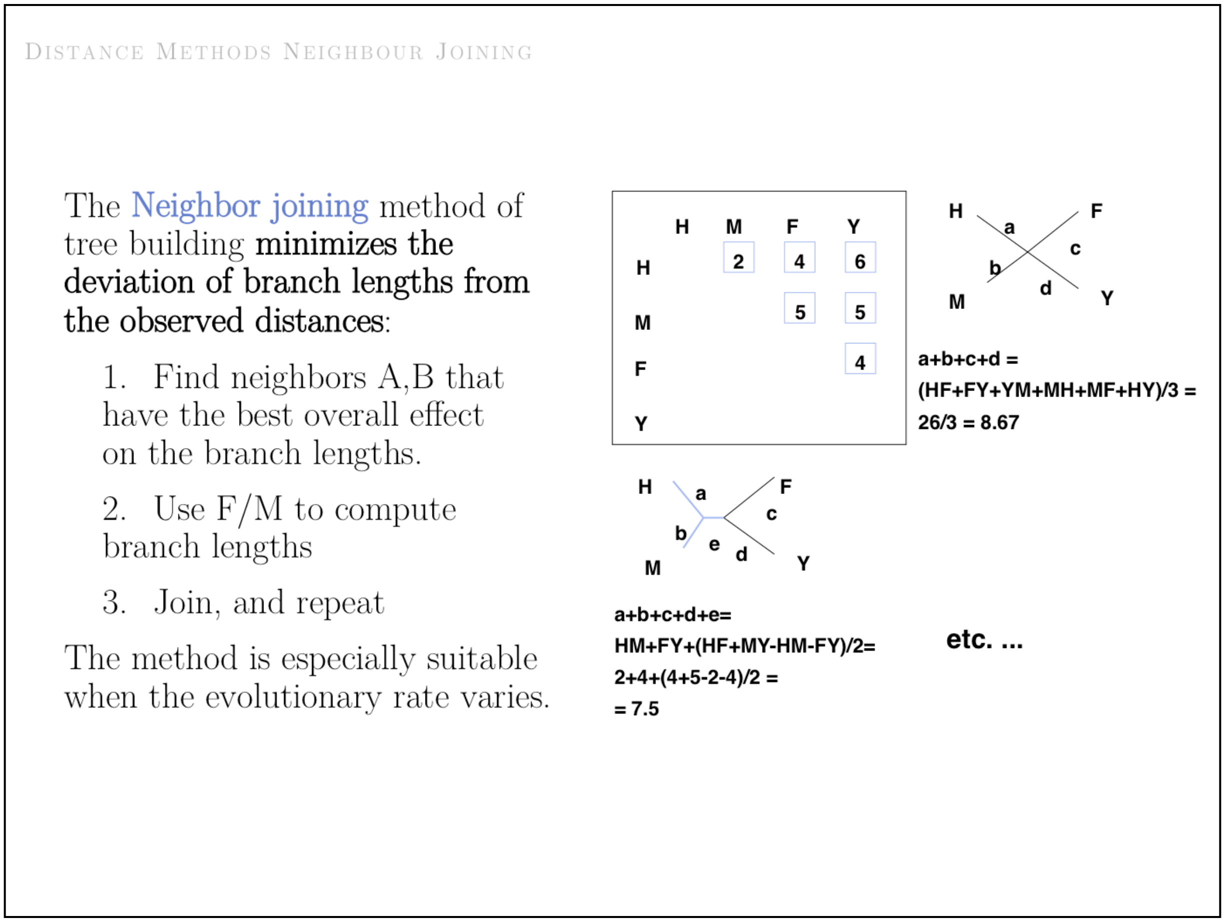
(4) (2)-(3) = a-b = 1

(5) (1)+(4) = 2a = 3; a = 1.5

Using (5) and (1) b = 0.5, using (5) and (2) c = 2.5



Phylip: DNADIST, PROTDIST



The Neighbor joining method of tree building **minimizes the deviation of branch lengths from the observed distances**:

1. Find neighbors A,B that have the best overall effect on the branch lengths.
2. Use F/M to compute branch lengths
3. Join, and repeat

The method is especially suitable when the evolutionary rate varies.

	H	M	F	Y
H		2	4	6
M			5	5
F				4
Y				

a+b+c+d =
(HF+FY+YM+MH+MF+HY)/3 =
26/3 = 8.67

a+b+c+d+e=
HM+FY+(HF+MY-HM-FY)/2=
2+4+(4+5-2-4)/2 =
= 7.5

etc. ...

Neigbour joining is also one of the standard methods for *hierarchical clustering*.

Tree Building

We don't want to construct just **ANY** tree, we want **THE** tree that best represents some model of how evolution has shaped the OTUs. Thus we need a metric, that describes how well our tree conforms to our model. Then whenever we encounter a necessary choice in tree construction, we apply our metric to guide us. However, we can phrase our objective in different ways.

The best tree ensures that the *most similar* OTUs share direct ancestors.

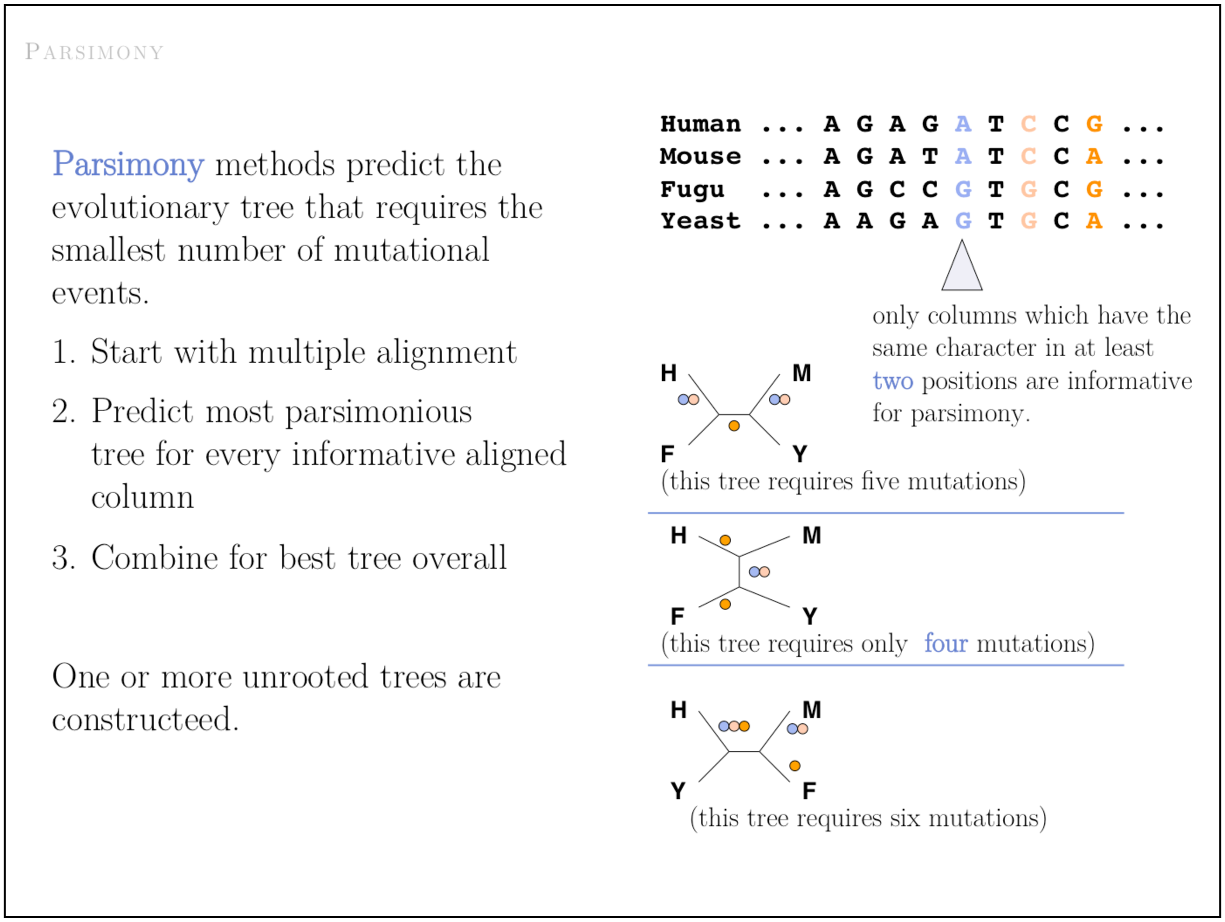
The best tree minimizes the *number of evolutionary events* in the tree.

The best tree maximizes the *likelih*

The best tree maximizes the proba alignment.

Parsimony and distance based methods:

Consider explicit change.



Parsimony

Parsimony methods predict the evolutionary tree that requires the smallest number of mutational events.

1. Start with multiple alignment
2. Predict most parsimonious tree for every informative aligned column
3. Combine for best tree overall

One or more unrooted trees are constructeed.

```
Human ... A G A G A T C C G ...
Mouse ... A G A T A T C C A ...
Fugu  ... A G C C G T G C G ...
Yeast ... A A G A G T G C A ...
```

only columns which have the same character in at least two positions are informative for parsimony.

(this tree requires five mutations)

(this tree requires only four mutations)

(this tree requires six mutations)

Note that potentially many ancestral states could give rise to the trees that are being considered.

Tree Building

We don't want to construct just **ANY** tree, we want **THE** tree that best represents some model of how evolution has shaped the OTUs. Thus we need a metric, that describes how well our tree conforms to our model. Then whenever we encounter a necessary choice in tree construction, we apply our metric to guide us. However, we can phrase our objective in different ways.

The best tree ensures that the *most s* ancestors.

The best tree minimizes the *number o*

The best tree maximizes the *likelihood* of the observed *alignment*.

The best tree maximizes the probability of the tree, given the alignment.

Likelihood based methods:

Consider explicit models of evolution.

Maximum Likelihood

The Maximum Likelihood (ML) method uses probability methods to find the tree that best accounts for the data. It is similar to parsimony methods, but allows inclusion of different substitution frequency models and evolutionary rates. :

1. Define model of evolution
2. For all possible trees:

   calculate probability that the observed sequence alignment (data)
   would have been generated by the tree (model).

Use e.g. branch-and-bound or heuristics to keep calculation tractable. This is compute intensive but flexible and gives good results. ML is one of the the state-of-the-art approaches.

Given sufficient computational resources,
ML or Bayesian methods are considered the method of choice!

Tree Building

We don't want to construct just **ANY** tree, we want **THE** tree that best represents some model of how evolution has shaped the OTUs. Thus we need a metric, that describes how well our tree conforms to our model. Then whenever we encounter a necessary choice in tree construction, we apply our metric to guide us. However, we can phrase our objective in different ways.

The best tree ensures that the *most similar* OTUs share direct ancestors.

The best tree minimizes the *number*

The best tree maximizes the *likeliho*

The best tree maximizes the probability of the tree, given the alignment.

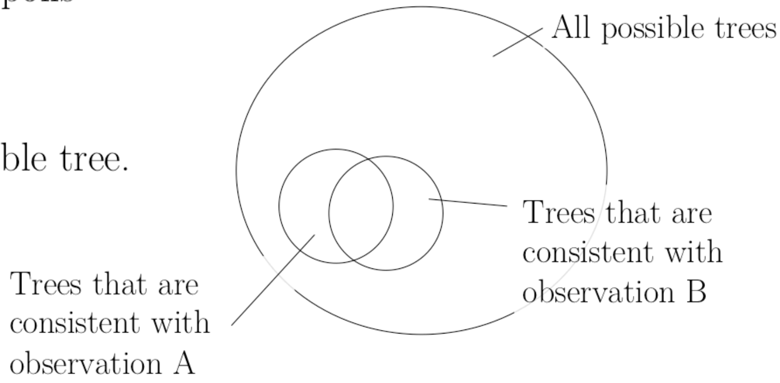
Bayesian methods:

Consider the predictions of explicit trees.

Bayes' Theorem For Tree Construction

Bayesian methods in phylogenetic analysis apply Bayes' Theorem to estimate the probability of each possible tree, given the observed data. The big advantage is that this is a consistent method to unify the effect of different parameters, such as branch lengths, topology, differing evolutionary models and differing rates for sites.

- Define trees in terms of parameters and topology.
- Use MCMC/Metropolis-Hastings to explore parameter space.
- Return most probable tree.



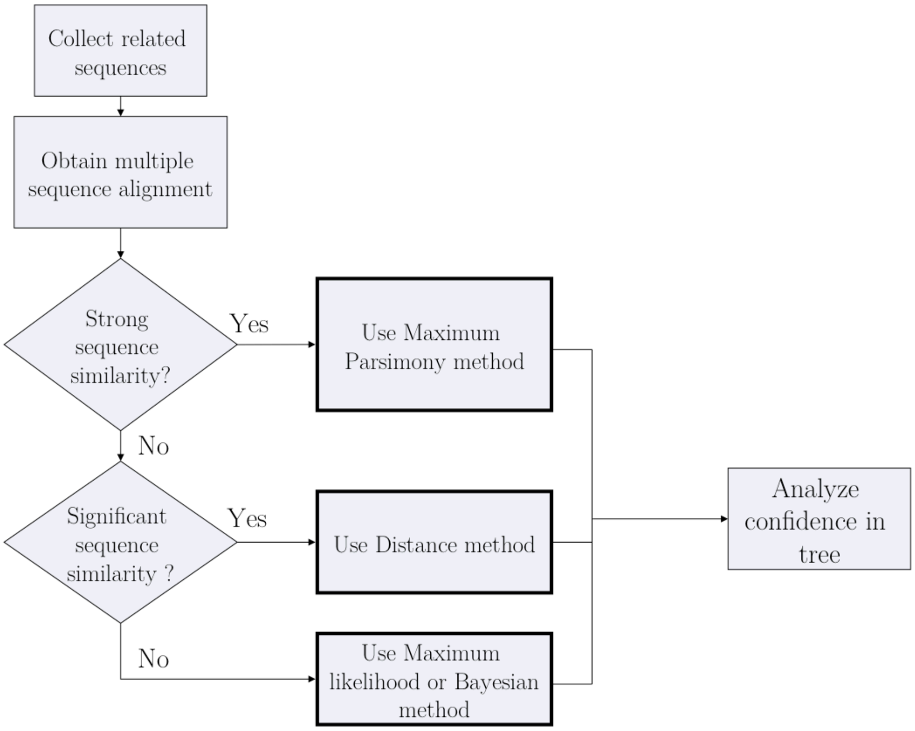
All possible trees

Trees that are consistent with observation B

Trees that are consistent with observation A

Example: Mr. Bayes

TREE BUILDING



Collect related sequences

Obtain multiple sequence alignment

Strong sequence similarity?

Yes

Use Maximum Parsimony method

No

Significant sequence similarity ?

Yes

Use Distance method

No

Use Maximum likelihood or Bayesian method

Analyze confidence in tree

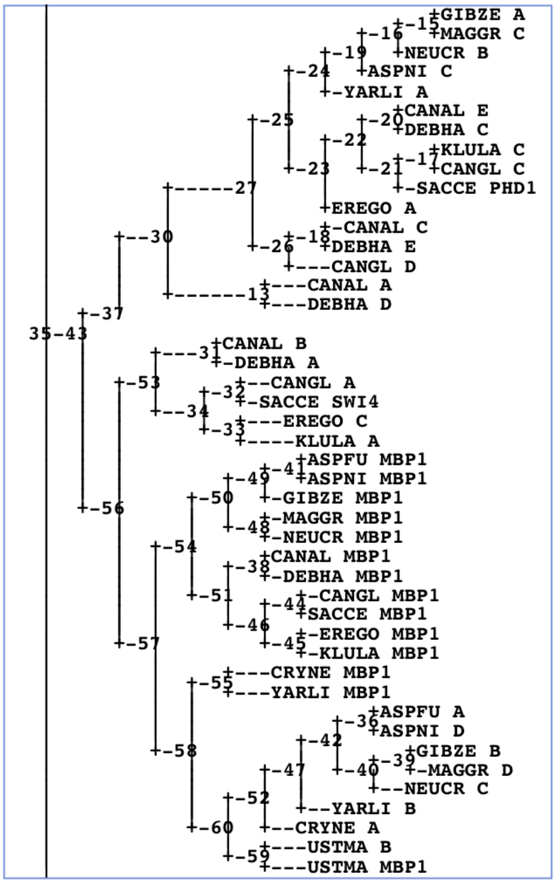
Bootstrapping

Bootstrapping gives an estimate how robust a tree is against small variations of the input data:

1. Randomly resample columns (with replacement) for an alignment of the same length. This produces trees that are based on only part of the data.
2. Compute tree
3. Repeat many (1000) times
4. Count number of times a specific bifurcation appears in the tree.
5. Report bifurcation frequency together with branching point in final tree.

Significant branching points should have $p > 0.7$
If a branching point is not well supported,
report this relationship as a *multifurcation*.

Mixed Gene Trees



Text

However! A real mixed gene tree is likely to deviate significantly from the evolutionary truth. In order to interpret such a tree, we must be **absolutely clear on what patterns of branching we would expect, given the possible speciation and duplication events, and the underlying relationship of species.**

http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA