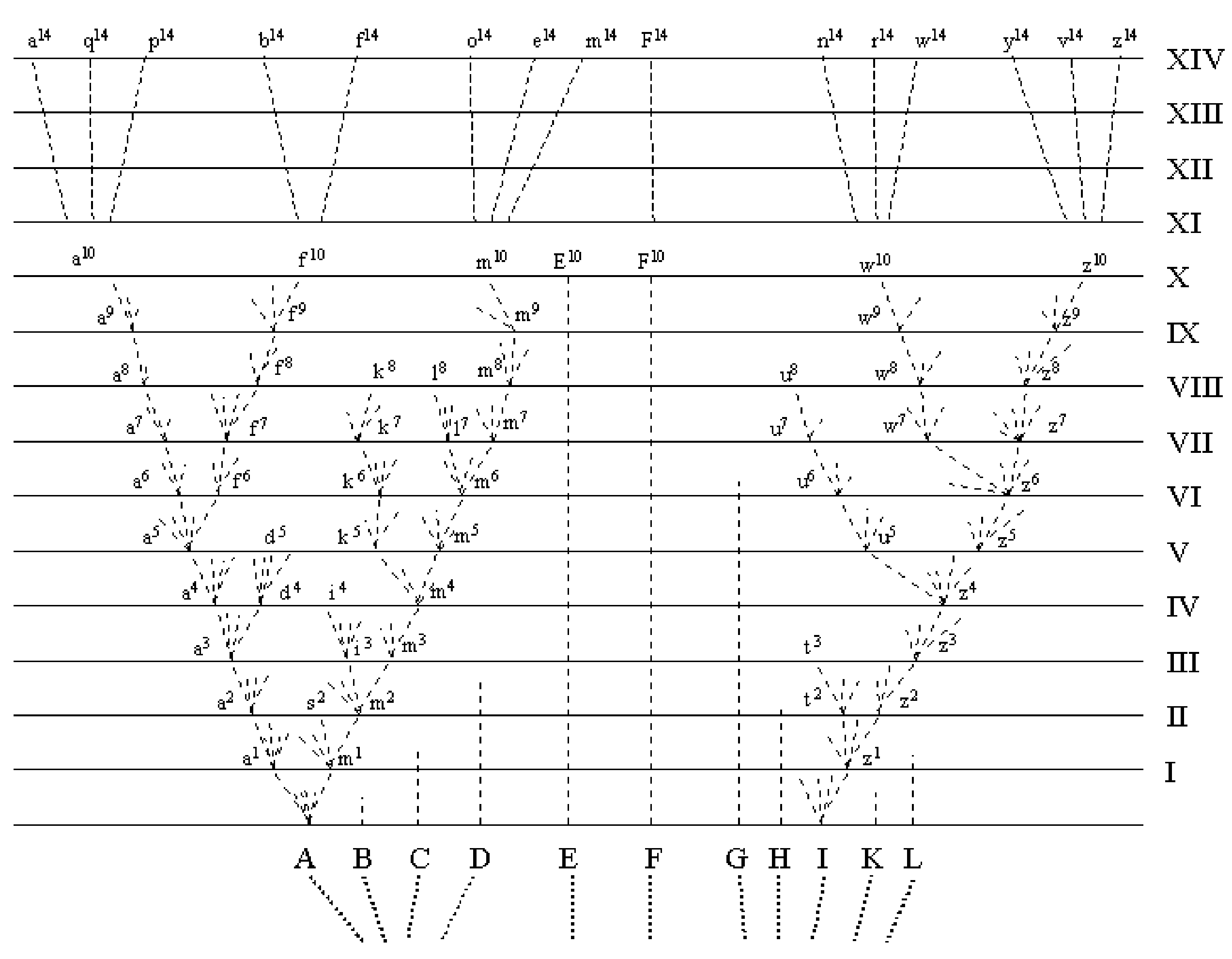
Inference of phylogenies, with some thoughts on statistics and geometry

Joe Felsenstein

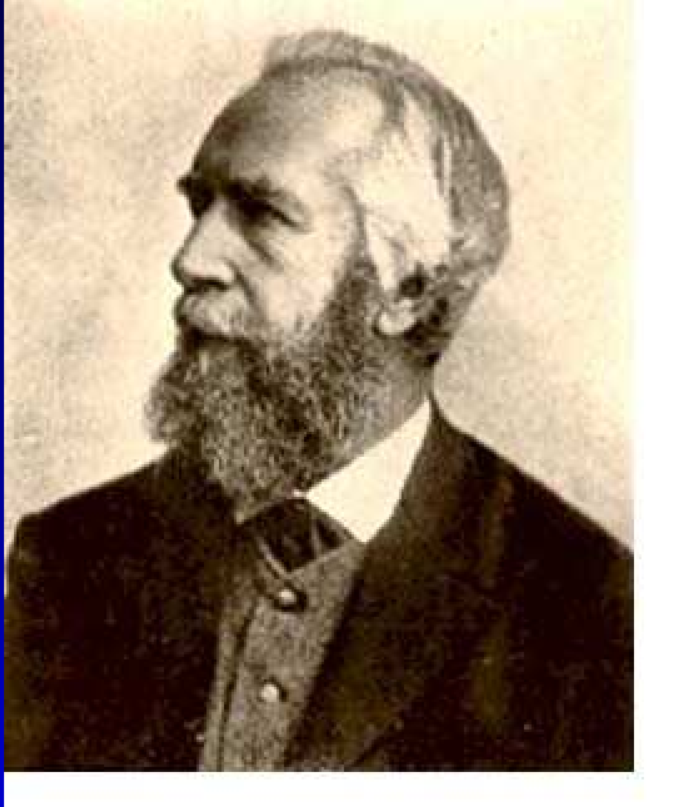
University of Washington

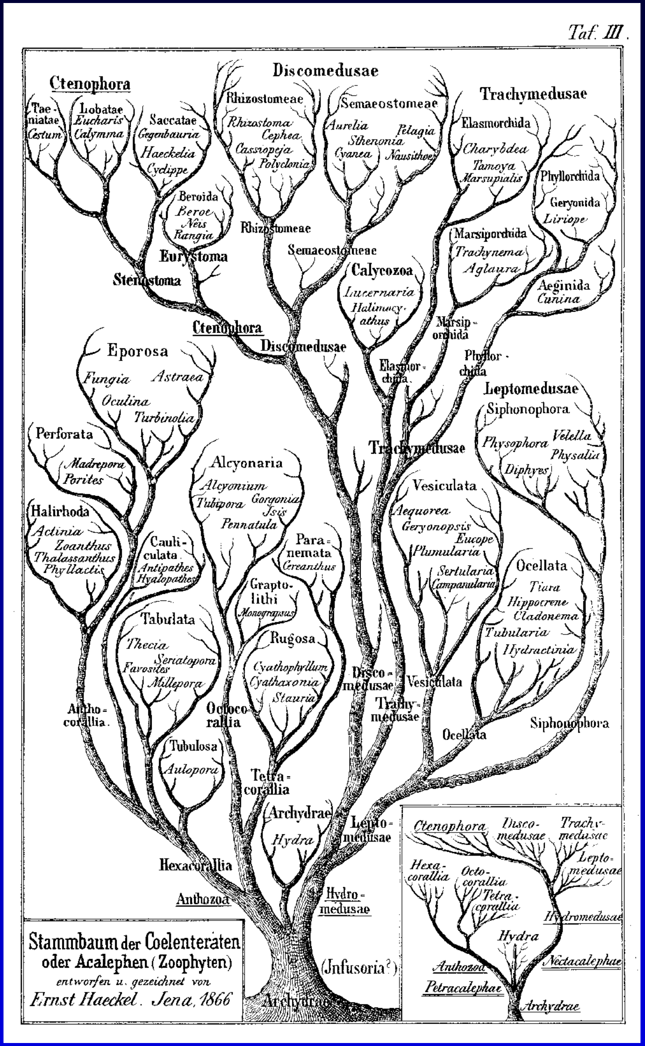
Darwin’s tree


a^{14} q^{14} p^{14} b^{14} f^{14} o^{14} e^{14} m^{14} F^{14} n^{14} r^{14} w^{14} y^{14} v^{14} z^{14}
XIV
XIII
XII
XI
a^{10} f^{10} m^{10} E^{10} F^{10} w^{10} z^{10}
X
a^9 f^9 m^9 w^9 z^9
IX
a^8 f^8 k^8 l^8 m^8 u^8 w^8 z^8
VIII
a^7 f^7 k^7 l^7 m^7 u^7 w^7 z^7
VII
a^6 f^6 k^6 m^6 u^6 z^6
VI
a^5 d^5 k^5 m^5 u^5 z^5
V
a^4 d^4 i^4 m^4 z^4
IV
a^3 i^3 m^3 t^3 z^3
III
a^2 s^2 m^2 t^2 z^2
II
a^1 m^1 z^1
I
A B C D E F G H I K L

This is the only illustration in *The Origin of Species*

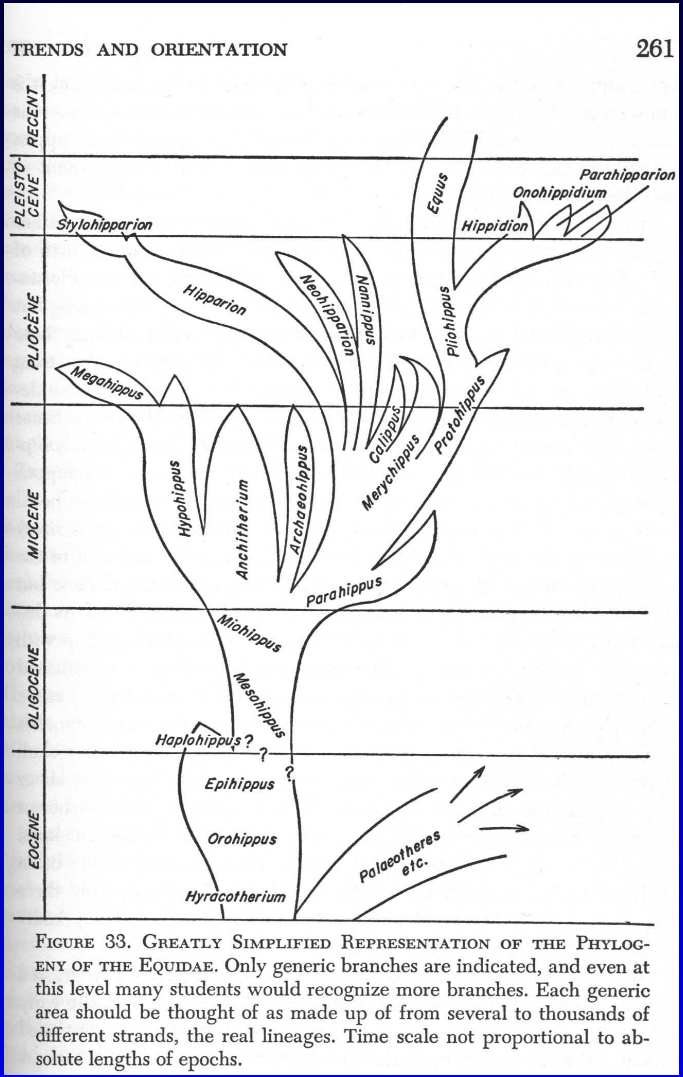
One of Ernst Haeckel’s trees (1866)





(Bark! on some of his trees, even leaves!) He coined the word “phylogeny".

George Gaylord Simpson, 1940’s



TRENDS AND ORIENTATION 261

FIGURE 33. GREATLY SIMPLIFIED REPRESENTATION OF THE PHYLOGENY OF THE EQUIDAE. Only generic branches are indicated, and even at this level many students would recognize more branches. Each generic area should be thought of as made up of from several to thousands of different strands, the real lineages. Time scale not proportional to absolute lengths of epochs.

Structure starts to get imprecise. Far from algorithms.

Algorithmic methods for inferring phylogenies

Starting in the 1960's, biologists invented a number of methods to infer phylogenies:

- Parsimony methods. Find that tree that minimizes the number of changes of state needed to evolve the observed data. (Related: compatibility methods)
- Distance matrix methods. Measure inferred distances between all pairs of species. Find the tree, with branch lengths, that predicts this table best.
- Maximum likelihood methods. Use a probability model of evolution (also used in distance matrix methods) and find that tree that makes the observed data most probable. (Related: Bayesian inference, which assumes also prior probabilities for the possible trees).

Parsimony methods

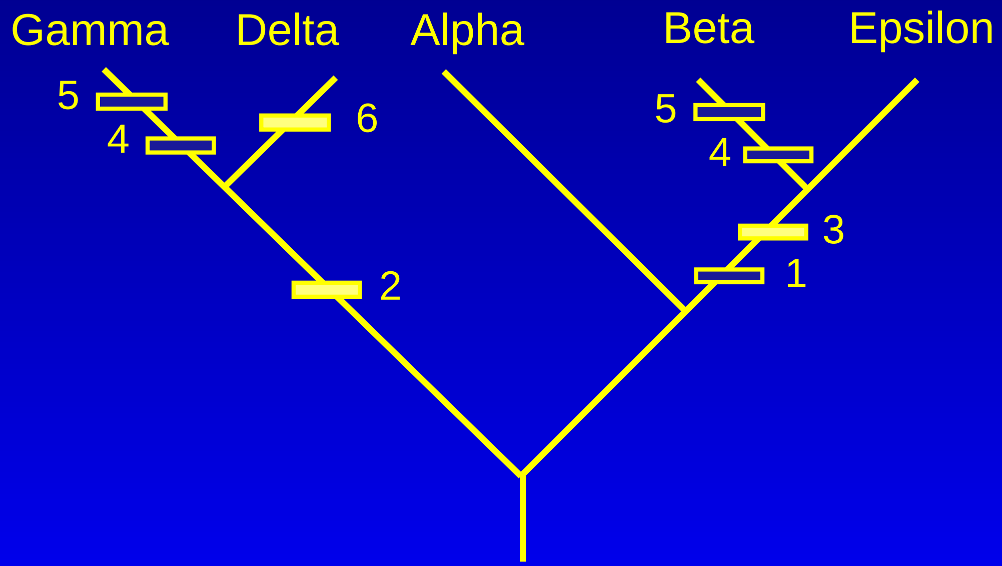
Two ways of looking at the parsimony method:

- Species are points in a sequence space. Find the Steiner tree for these points in this space.
- Trees live in a tree space; for each of them we can evaluate how many changes of state are needed to evolve the data set.

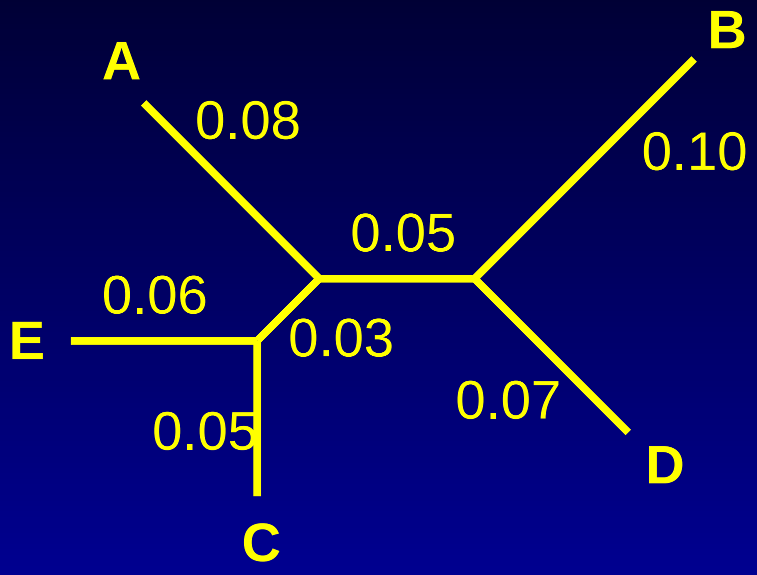
Parsimony on a simple data set

	site					
species	1	2	3	4	5	6
Alpha	A	T	G	A	G	C
Beta	C	T	C	T	A	C
Gamma	A	G	G	T	A	C
Delta	A	G	G	A	G	T
Epsilon	C	T	C	A	G	C

Here is the most parsimonious tree (drawn as if rooted, though root location doesn't matter for the parsimony score in this case)



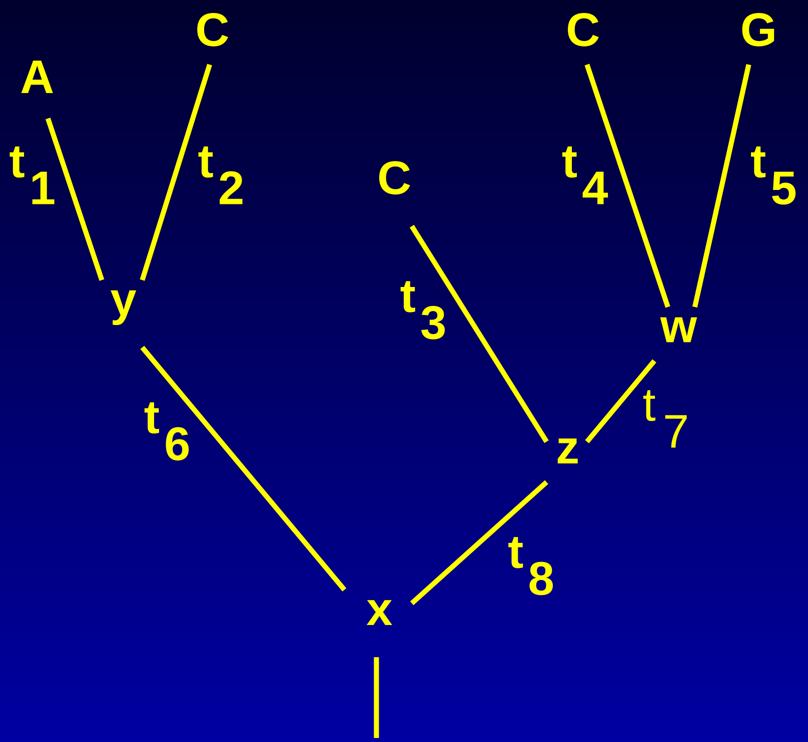
Distance matrix methods



	A	B	C	D	E
A	0	0.23	0.16	0.20	0.17
B	0.23	0	0.23	0.17	0.24
C	0.16	0.23	0	0.15	0.11
D	0.20	0.17	0.15	0	0.21
E	0.17	0.24	0.11	0.21	0

This table of expected distances is computed from the tree and compared to pairwise distances (branch lengths of a 2-species unrooted tree) computed from the data.

Likelihood methods



Likelihood is the probability of the data given the tree, where the tree has branch lengths, and a model of evolution (occurring independently in each site and independently in different lineages)

Likelihood computed separately in each site:

The likelihood is the product over sites (as they are independent given the tree):

$$\mathsf{L} \;=\; \mathrm{Prob}\,(\mathsf{D}|\mathsf{T}) \;=\; \prod_{\mathsf{i}=1}^{\mathsf{m}} \mathrm{Prob}\,\left(\mathsf{D}^{(\mathsf{i})}|\mathsf{T}\right)$$

For site i, summing over all possible states at interior nodes of the tree:

$$\mathrm{Prob}\,\left(\mathsf{D}^{(\mathsf{i})}|\mathsf{T}\right) \;=\; \sum_{\mathsf{x}}\sum_{\mathsf{y}}\sum_{\mathsf{z}}\sum_{\mathsf{w}} \mathrm{Prob}\,(\mathsf{A},\mathsf{C},\mathsf{C},\mathsf{C},\mathsf{G},\mathsf{x},\mathsf{y},\mathsf{z},\mathsf{w}|\mathsf{T})$$

Using independence assumptions to simplify

By independence of events in different branches (conditional only on their starting states):

$$\begin{aligned}
\text{Prob}\,(\mathsf{A}, \mathsf{C}, \mathsf{C}, \mathsf{C}, \mathsf{G}, \mathsf{x}, \mathsf{y}, \mathsf{z}, \mathsf{w} | \mathsf{T}) = \\
\text{Prob}\,(\mathsf{x})\ \text{Prob}\,(\mathsf{y}|\mathsf{x}, \mathsf{t}_6)\ \text{Prob}\,(\mathsf{A}|\mathsf{y}, \mathsf{t}_1)\ \text{Prob}\,(\mathsf{C}|\mathsf{y}, \mathsf{t}_2) \\
\text{Prob}\,(\mathsf{z}|\mathsf{x}, \mathsf{t}_8)\ \text{Prob}\,(\mathsf{C}|\mathsf{z}, \mathsf{t}_3) \\
\text{Prob}\,(\mathsf{w}|\mathsf{z}, \mathsf{t}_7)\ \text{Prob}\,(\mathsf{C}|\mathsf{w}, \mathsf{t}_4)\ \text{Prob}\,(\mathsf{G}|\mathsf{w}, \mathsf{t}_5)
\end{aligned}$$

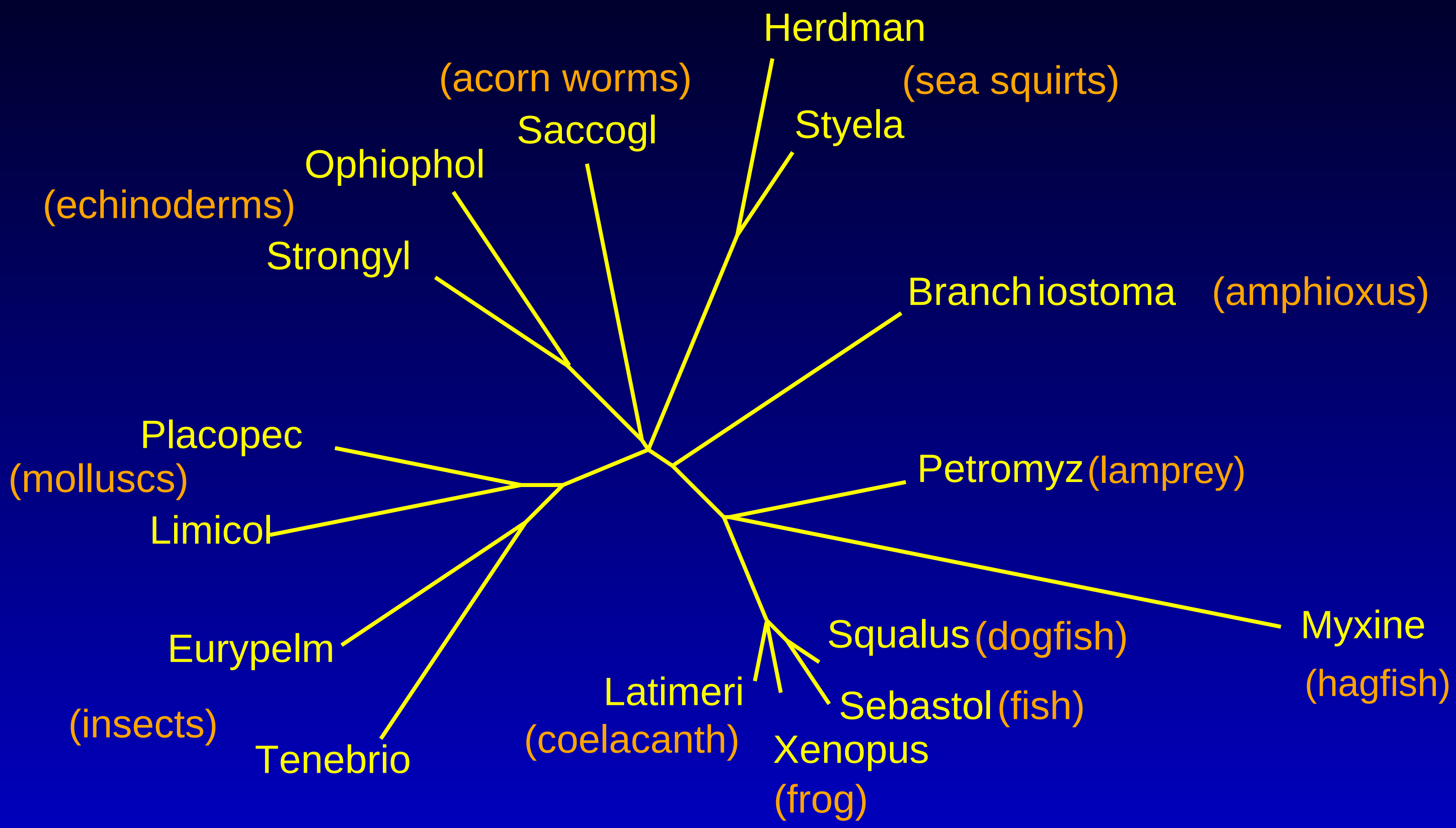
There are, of course, numerical optimization issues in finding the best branch lengths for each topology examined.

An example: Turbeville et al., 28s RNA, chordates

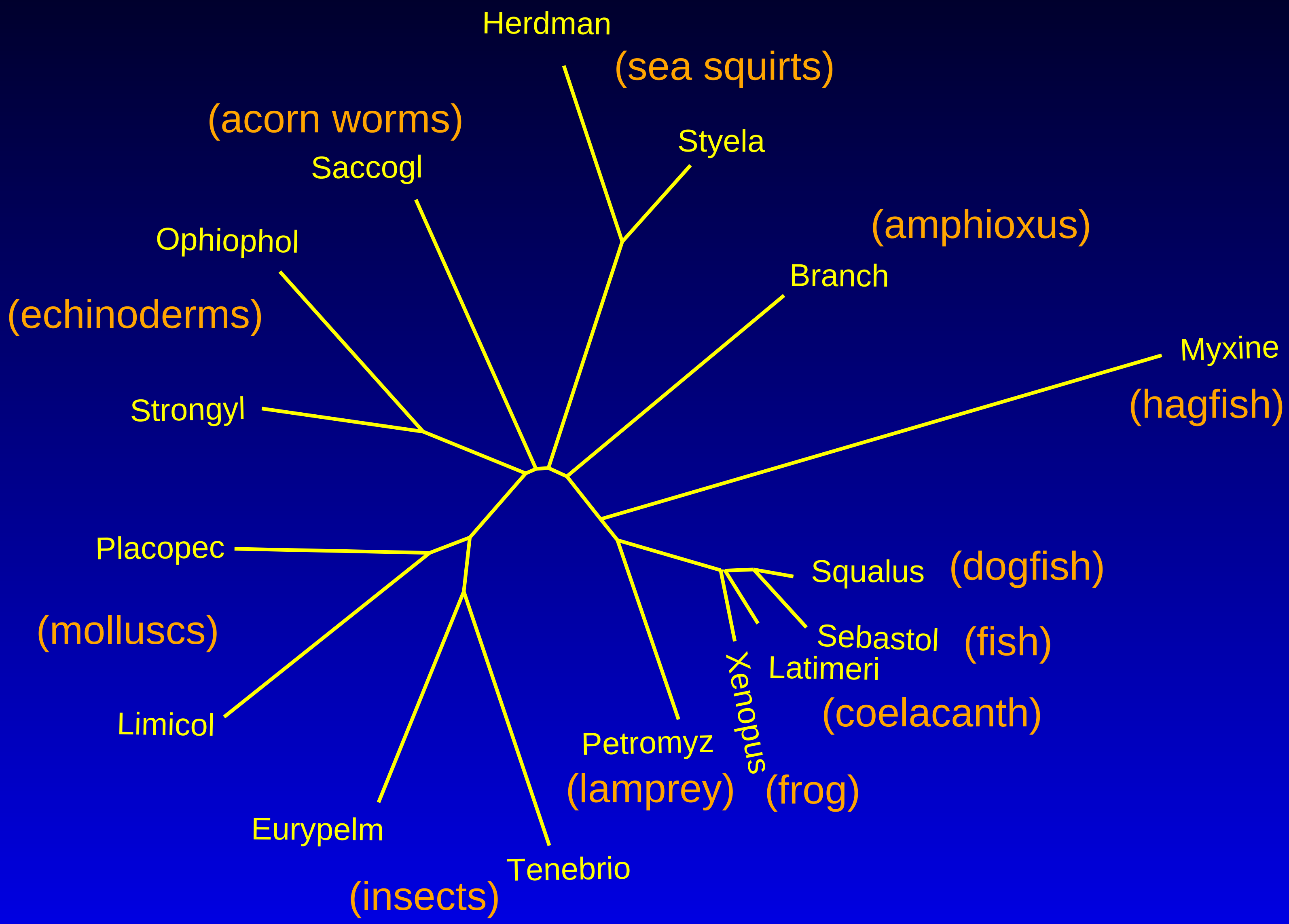
```
Xenopus      ?TACCTGGTTGATCCTGCCAGTAG-CATATG...
Sebastol     ??????????????????????AG-CATATG...
Latimeri     ?TACCTGGTTGATCCTGCCAGTAG-CATATG...
Squalus      ??????????????????????AG-CATATG...
Myxine       ??CCCTGGTTGATCCTGCCAGCCG-CATATG...
Petromyz     ???CCTGGTTGATCCTGCCAGTAG-CATATG...
Branch       ???CCTGGTTGATCCTGCCAGTAGTCATATG...
Styela       ??ATCTGGTTGATCCTGCCAGTAGTGATATG...
Herdman      ?TATCTGGTTGATCCTGCCAGTAGTGATATG...
Saccogl      ??ACCTGGTTGATCCTGCCAGTAGTCATATG...
Ophiophol    ??ACCTGGTTGATCCTGCCAGTAGTCATATG...
Strongyl     ??ACCTGGTTGATCCTGCCAGTAGTCATATG...
Placopec     CAACCTGGTTGATCCTGCCAGTAGTCATATG...
Limicol      ?TATCTGGTTGATCCTGCCAGTAGTCATATG...
Eurypelm     ?TACCTGGTTGATCCTGCCAGTAGTCATATG...
Tenebrio     ?TCCCTGGTTGATCCTGCCAGTAGTCATATG...
```

```
(and so on for 11 more pages)
```

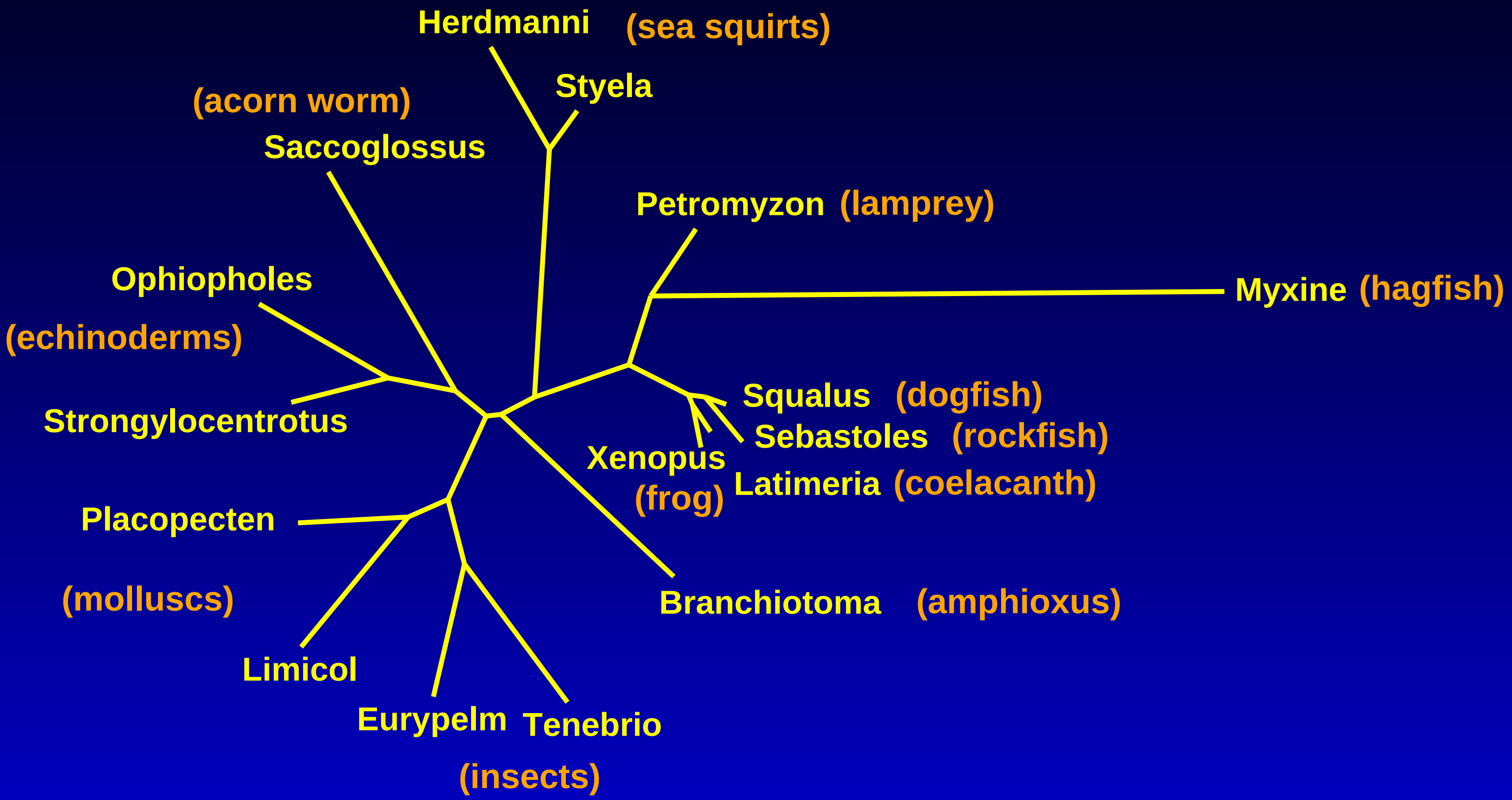
The tree with parsimony



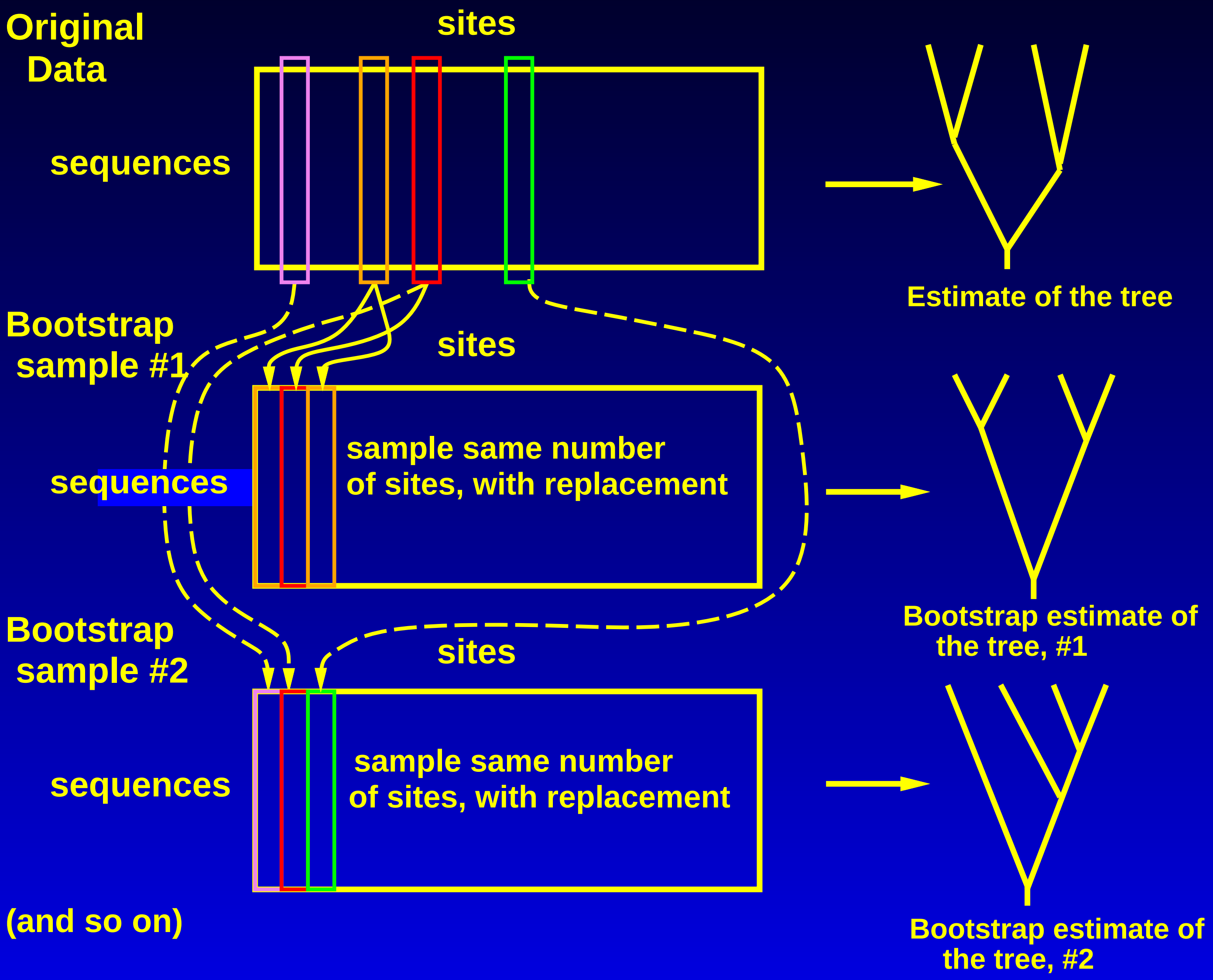
The tree with the Fitch-Margoliash distance matrix method



The tree with maximum likelihood



The bootstrap shows the uncertainty in our trees



Original
Data

sites

sequences

Estimate of the tree

Bootstrap
sample #1

sites

sequences

sample same number
of sites, with replacement

Bootstrap estimate of
the tree, #1

Bootstrap
sample #2

sites

sequences

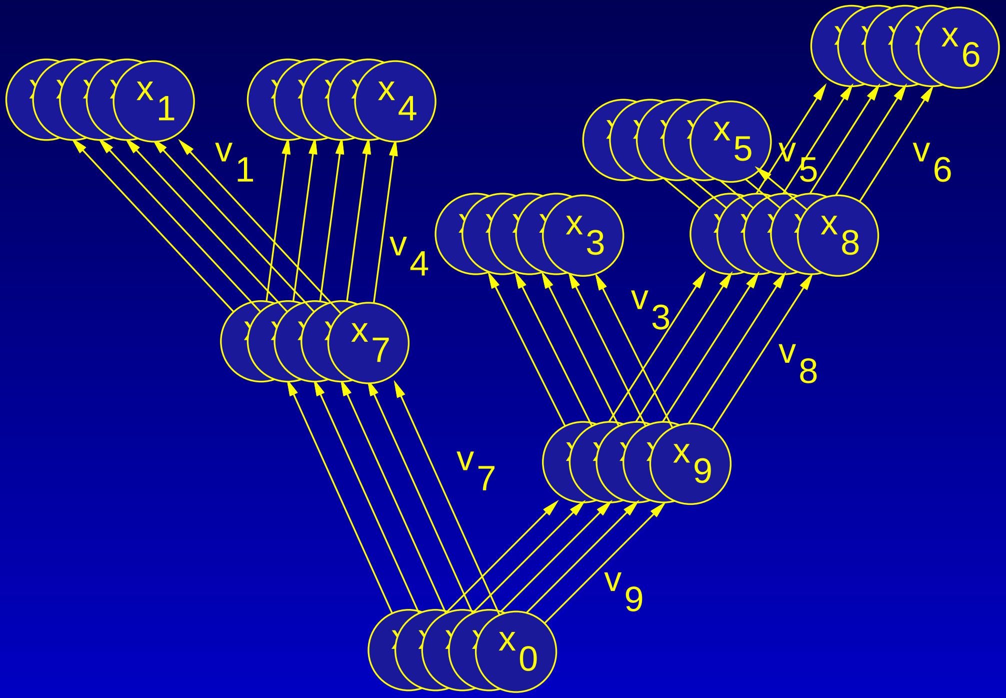
sample same number
of sites, with replacement

Bootstrap estimate of
the tree, #2

(and so on)

Phylogenies can be seen as graphical models

We have observations at a number of characters (wuch as sites in DNA molecules). Each site has a graphical model and is independently evolving on the same tree (by assumption). The graphical model looks like this:



Imagine a version of this 500 or so sites deep.

Problem: given observations on the x's at the tips, to reconstruct the tree topology and the v's.

Invariants (a method with a future?)

- With DNA data and 4 species there are 256 data patterns possible: AAAA through TTTT.
- With the simplest symmetric model (the Jukes-Cantor model) of nucleotide change, the tree has 5 parameters (the branch lengths
- These 5 parameters define a 5-dimensional space that defines where in the space of expected pattern frequency the trees are.
- This leaves 250 degrees of freedom to account for: there are 250 equations the expected pattern frequencies must satisfy.
- Many are simple symmetry requirements such as $f_{AAAA} = f_{CCCC}$
- a few are phylogenetic invariants that depend on the tree topology.

Mathematical approaches to trees that have been useful

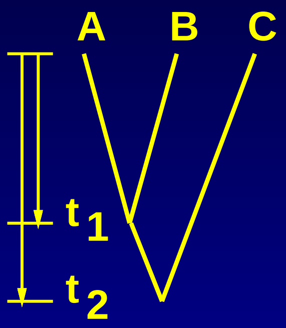
- The analogy between trees and recursion. This was discovered early by many mathematically-oriented biologists.
- Invariants (J. Cavender and J. Lake). Holds promise for the future but so far not of great practical use. Lots of geometry here?
- The Hadamard conjugation (M. Hendy / D. Penny). Remarkable results, but so far confined to certain symmetric models of base change.

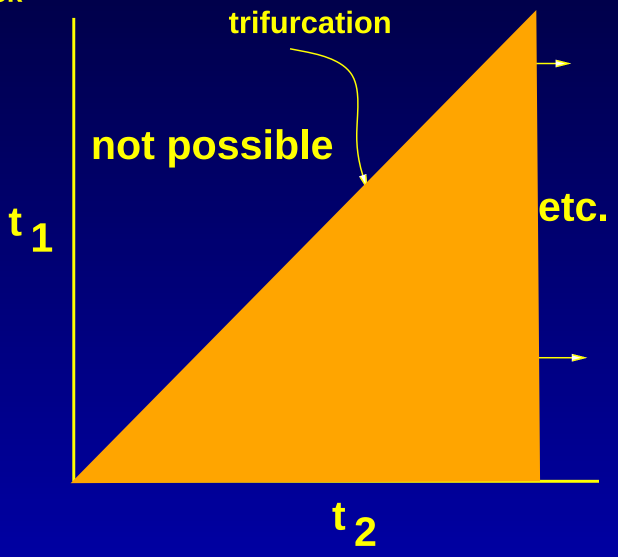
But on the whole, this is a short list. We are waiting for a mathematical result that makes a real impact, rather than restating the obvious or just rigorizing something we already use.

Tree space as explained by me a few years ago

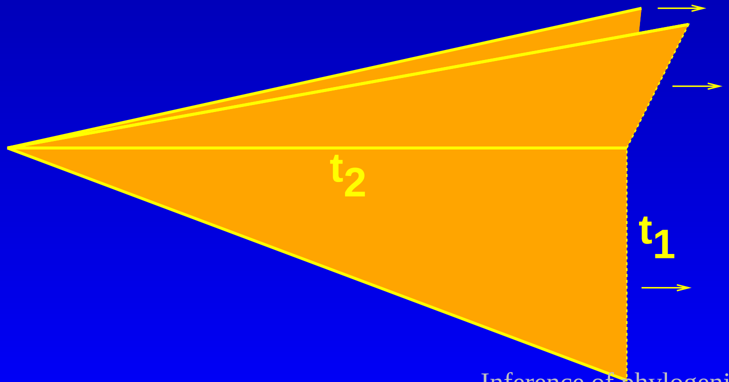
The space of trees with branch lengths

an example: three species with a clock

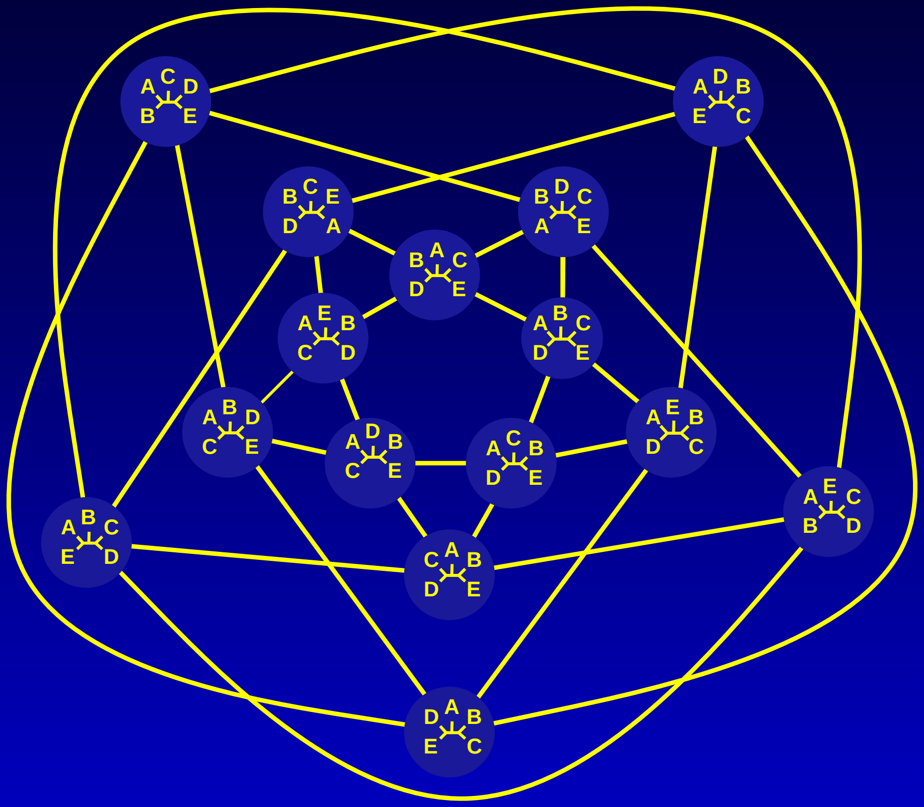




when we consider all three possible topologies, the space looks like:



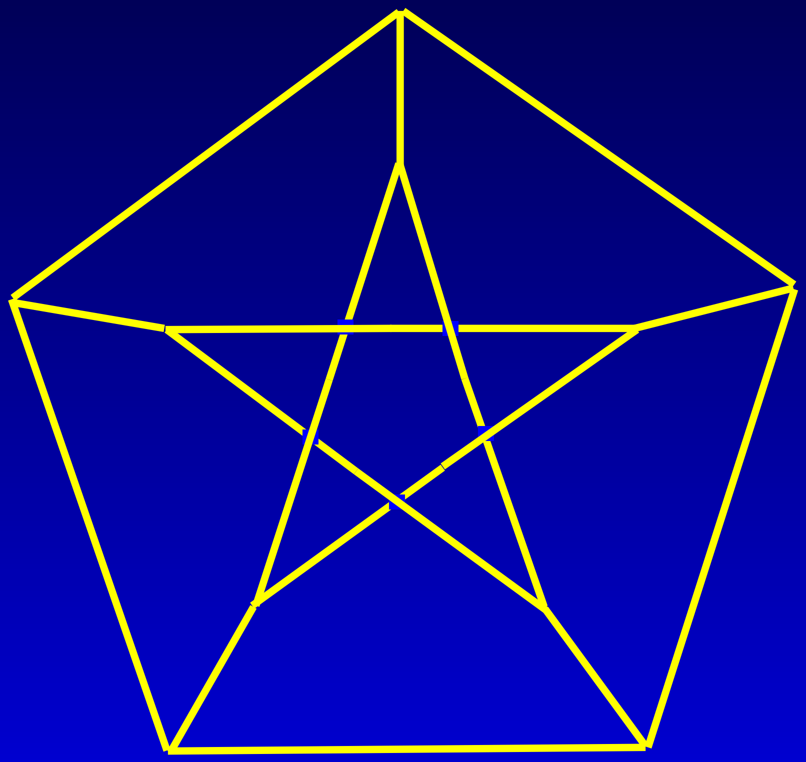
Tree space, for 5 species



This graph appeared on the 1999 T-shirt of the Molecular Evolution Workshop, Marine Biological Laboratory, Woods Hole, Massachusetts.

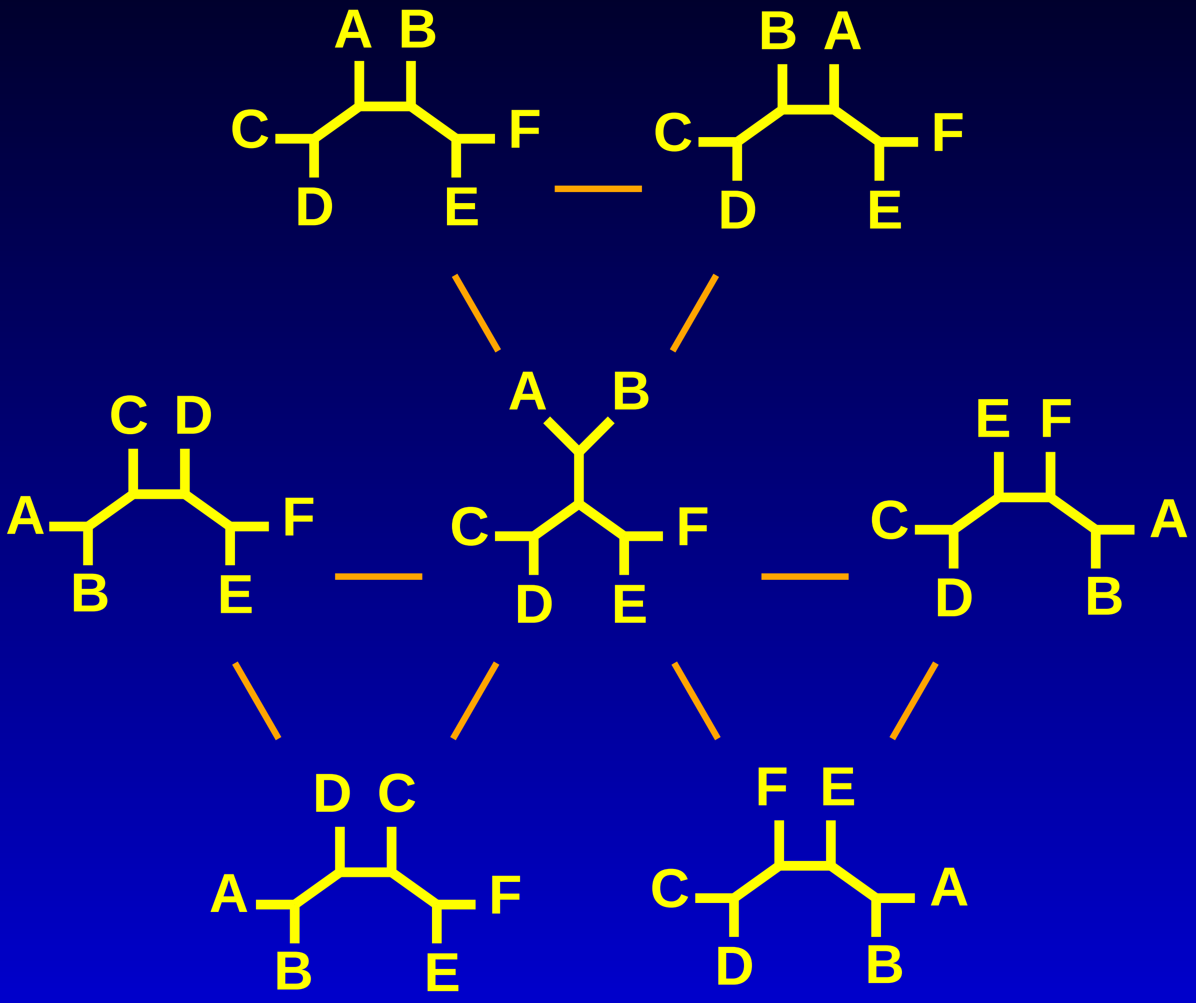
A dual

If we take the cliques (3 trees each) of the NNI graph, and make a dual in which the points are cliques, and the lines connect them when they share a tree, we get (ta-da!)



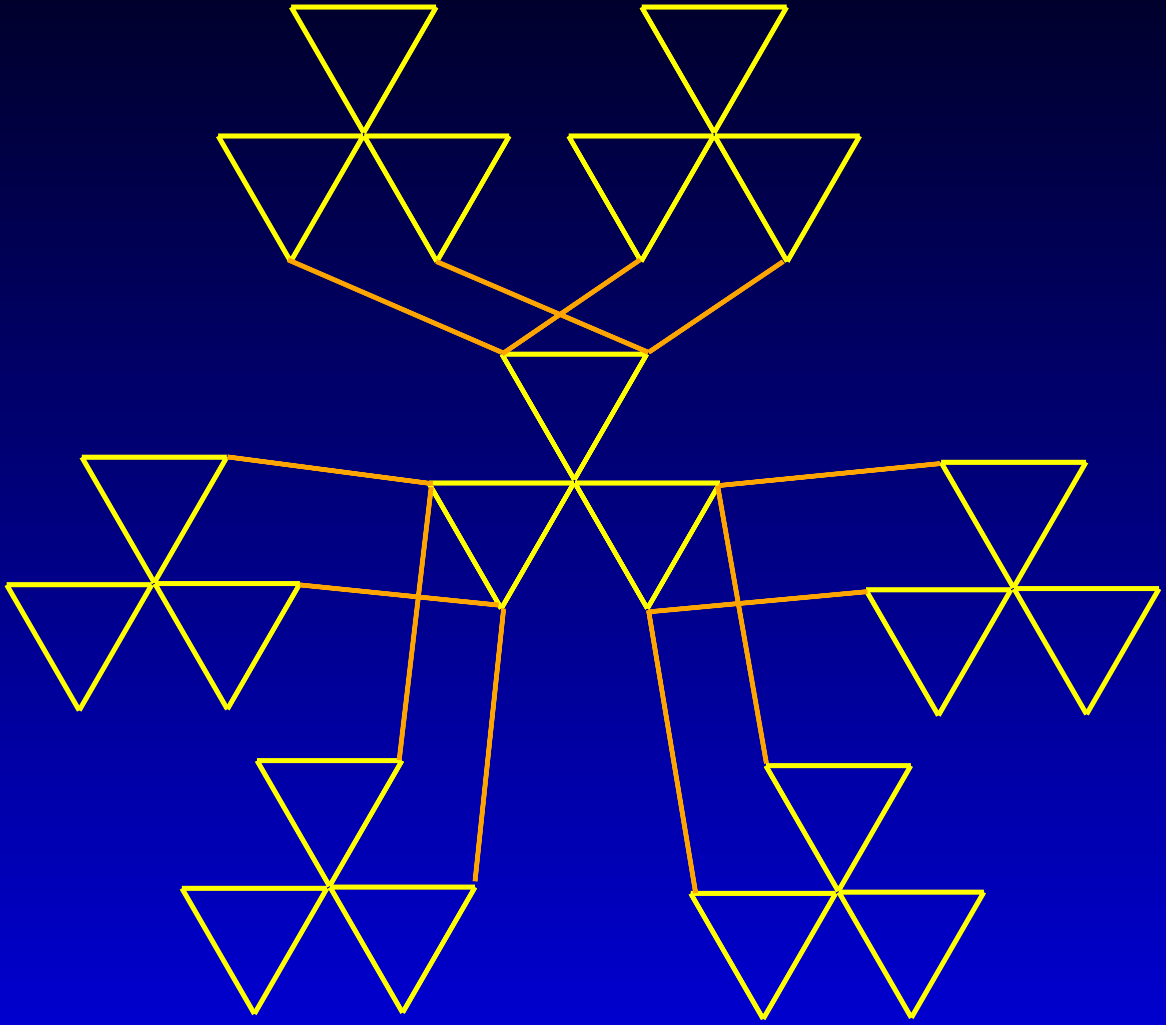
The infamous Petersen Graph.

A “shamrock", part of the NNI graph for 6 species



Each tree of shape ((x,x),(x,x),(x,x)) is surrounded by 6 trees of shape (((x,x),x),(x,(x,x))). There are 15 of these shamrocks, 105 trees in all.

More about the 6-species NNI graph



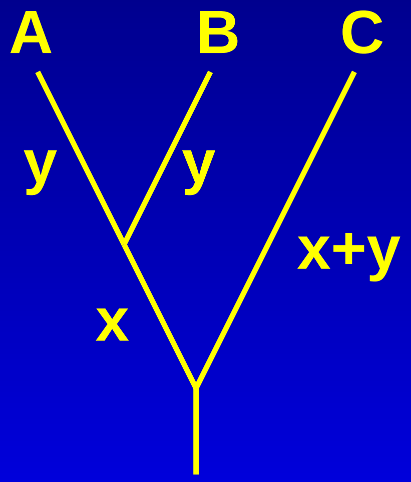
Each shamrock connects from each leaf to two others. The 6 shamrocks that each shamrock connects to are the ones that share at least one of its two-species sets.

A test case: 3 species with clock, Jukes-Cantor model

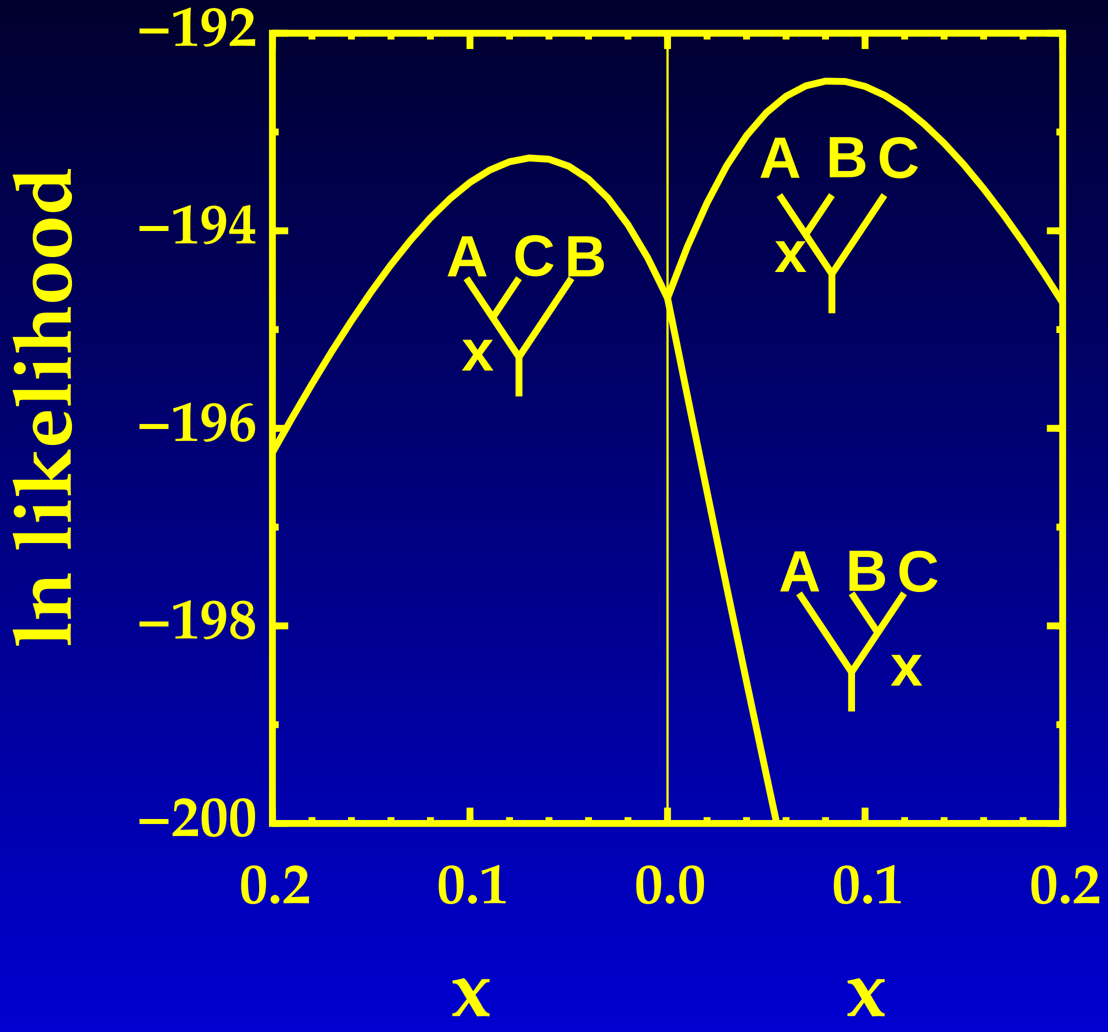
The data set (variant nucleotides lower-cased):

```
     3       64
Alpha            gtcaACGTACGTACGTACGTACGTACGTACGT
Beta             ACGTgtcaACGTgtcagtcaACGTACGTACGT
Gamma            ACGTACGTgtcaACGTACGTgtcagtcagtca

                 ACGTACGTACGTACGTACGTACGTACGTACGT
                 ACGTACGTACGTACGTACGTACGTACGTACGT
                 ACGTACGTACGTACGTACGTACGTACGTACGT
```

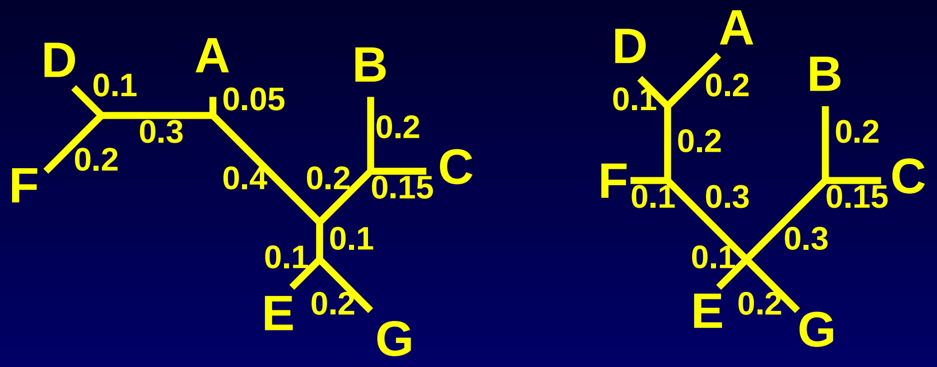


This leads to these profile likelihoods



(Profile likelihoods are $\max\limits_{y} \ln L(x, y)$)

The branch score (Kuhner and Felsenstein, 1994)



partitions	branch lengths	
{AD | BCEFG}	none	0.2
{ADF | BCEG}	0.4	0.3
{BC | ADEFG}	0.2	0.2
{DF | ABCEG}	0.3	none
{EG | ABCDF}	0.1	none
{A | BCDEFG}	0.05	0.2
{B | ACDEFG}	0.2	0.2
{C | ABDEFG}	0.15	0.15
{D | ABCEFG}	0.1	0.1
{E | ABCDFG}	0.1	0.1
{F | ABCDEG}	0.2	0.1
{G | ABCDEF}	0.2	0.2

The perfect tree distance

Can we make a tree distance with perfect statistical meaning? What would that look like if we did?

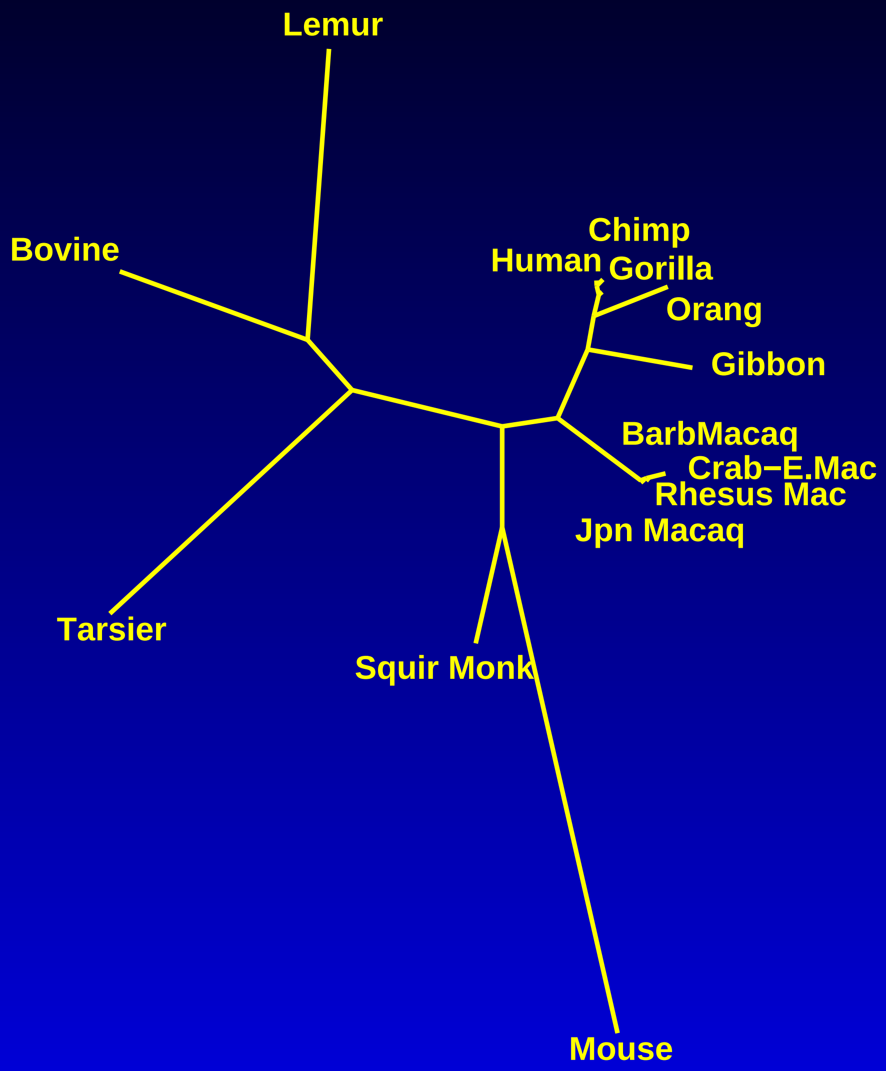
The (squared?) distance would reflect the difference of log likelihood between the two trees:

$$D^2 = \left| \sum_{i}^{\text{patterns}} \frac{n_i}{n} \ln \left(\frac{p_i}{q_i} \right) \right|$$

where the p_i and the q_i are the expected frequencies of the possible data patterns in the two trees, and the n_i are the numbers of each pattern seen.

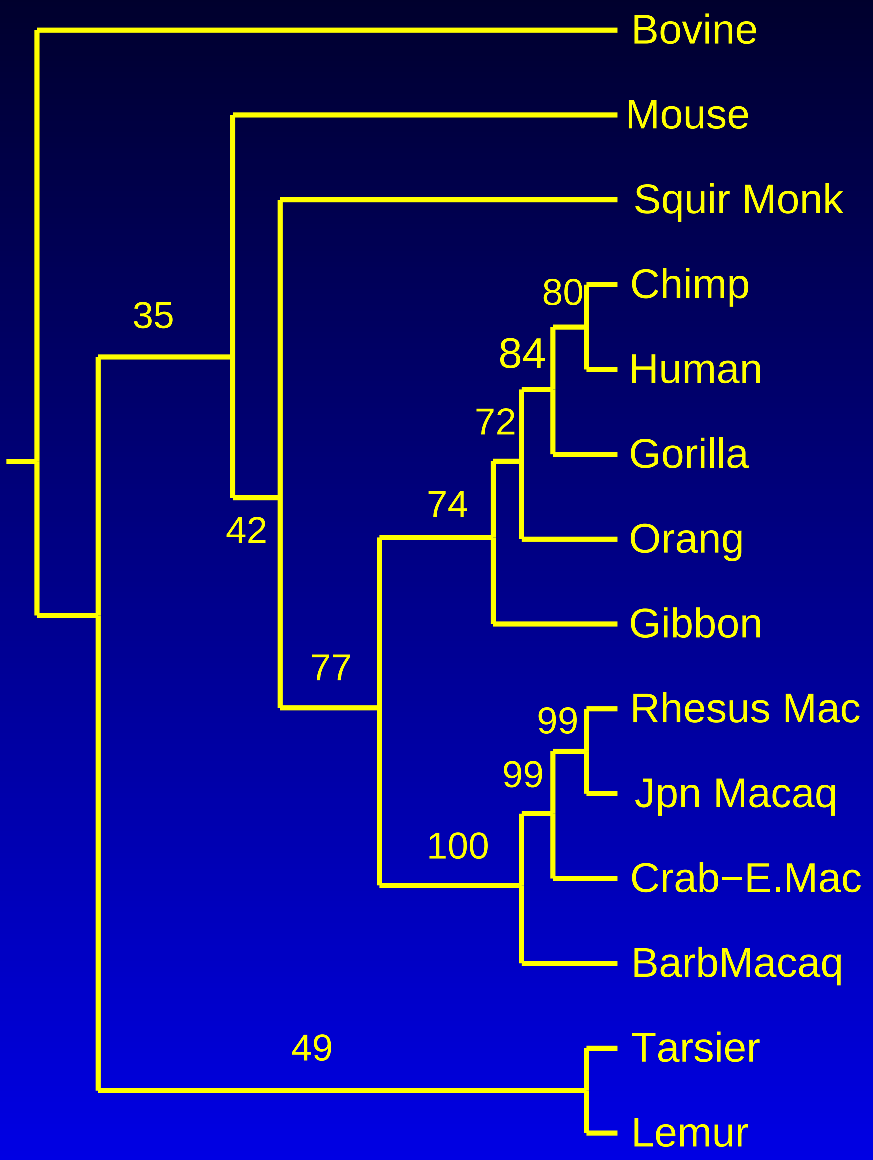
(This formula is related to the Kullback-Leibler separator but is slightly different because it has the data in it, and is not just a function of the two trees.)
I don't see any way to make a distance that magically reflects the data, without using the data.

An example of the problem: primates data



Tree from 232 nucleotides of mitochondrial DNA data

Bootstrap sampling finds these partition frequencies:



(drawing the tree as if rooted)

How to get distance-ish methods to use data?

A modest proposal

Use the cloud of bootstrap trees, jackknife trees, or posterior distribution of trees in a Bayesian analysis. The variation in these accurately reflects the uncertainty in the data. Can we somehow use this variation to evaluate which directions in tree space our data allows us to go?

References

Turbeville. J. McC., Schulz, J .R. and R. A. Raff. 1994. Deuterostome phylogeny and the sister group of the chordates: evidence from molecules and morphology. *Molecular Biology and Evolution*

How it was done

This projection produced as a PDF, not a PowerPoint file, and viewed using the Full Screen mode (in the View menu of Adobe Acrobat Reader):

- using the `prosper` style in LaTeX,
- using Latex to make a `.dvi` file,
- using `dvips` to turn this into a Postscript file,
- using `ps2pdf` to mill it into a PDF file, and
- displaying the slides in Adobe Acrobat Reader.

Result: nice slides using freeware.