Reconstructing phylogenies: how? how well? why?

Joe Felsenstein

Department of Genome Sciences and Department of Biology
University of Washington, Seattle

A review that asks these questions

- What are some of the strengths and weaknesses of different ways of reconstructing evolutionary trees (phylogenies)?

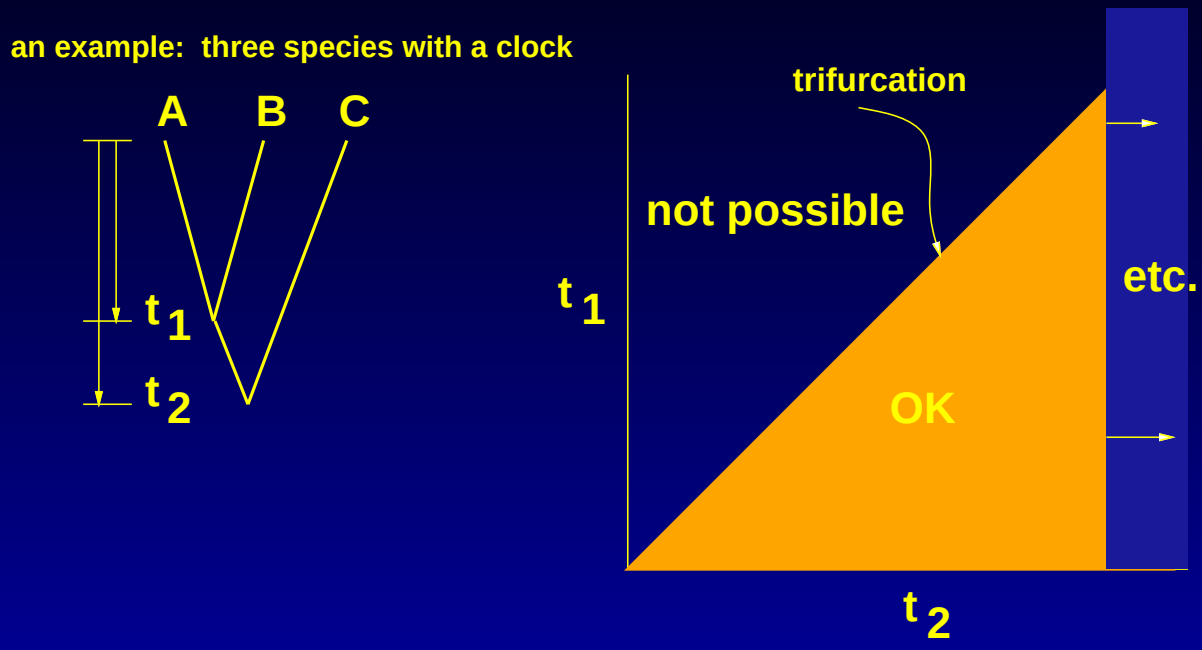
A review that asks these questions

- What are some of the strengths and weaknesses of different ways of reconstructing evolutionary trees (phylogenies)?
- How can we find out how accurate we may have been in reconstructing the phylogeny?

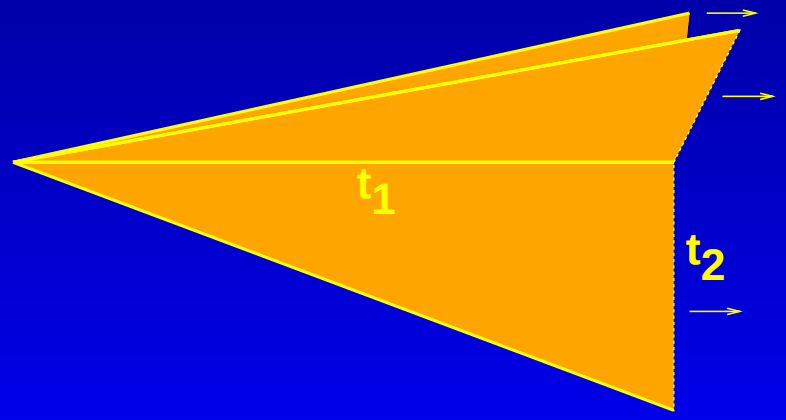
A review that asks these questions

- What are some of the strengths and weaknesses of different ways of reconstructing evolutionary trees (phylogenies)?
- How can we find out how accurate we may have been in reconstructing the phylogeny?
- Why do we want to reconstruct it? What are phylogenies used for?

What does "tree space" (with branch lengths) look like?

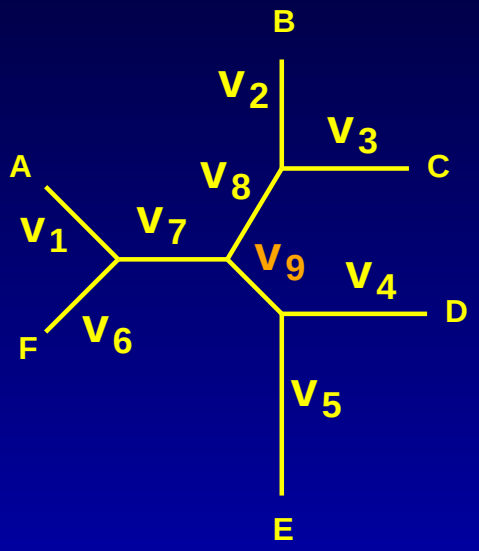


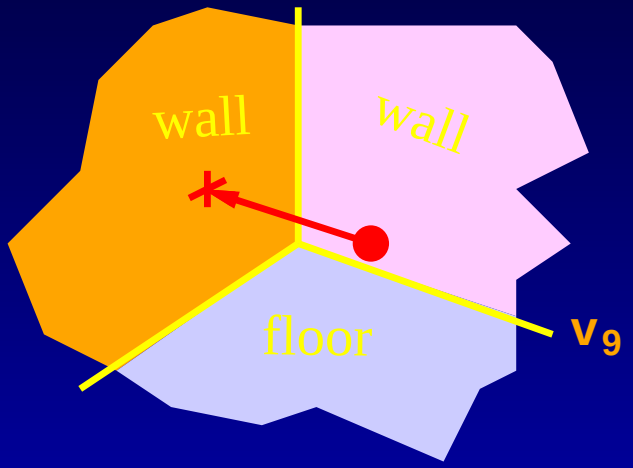
when we consider all three possible topologies, the space looks like:



For one tree topology

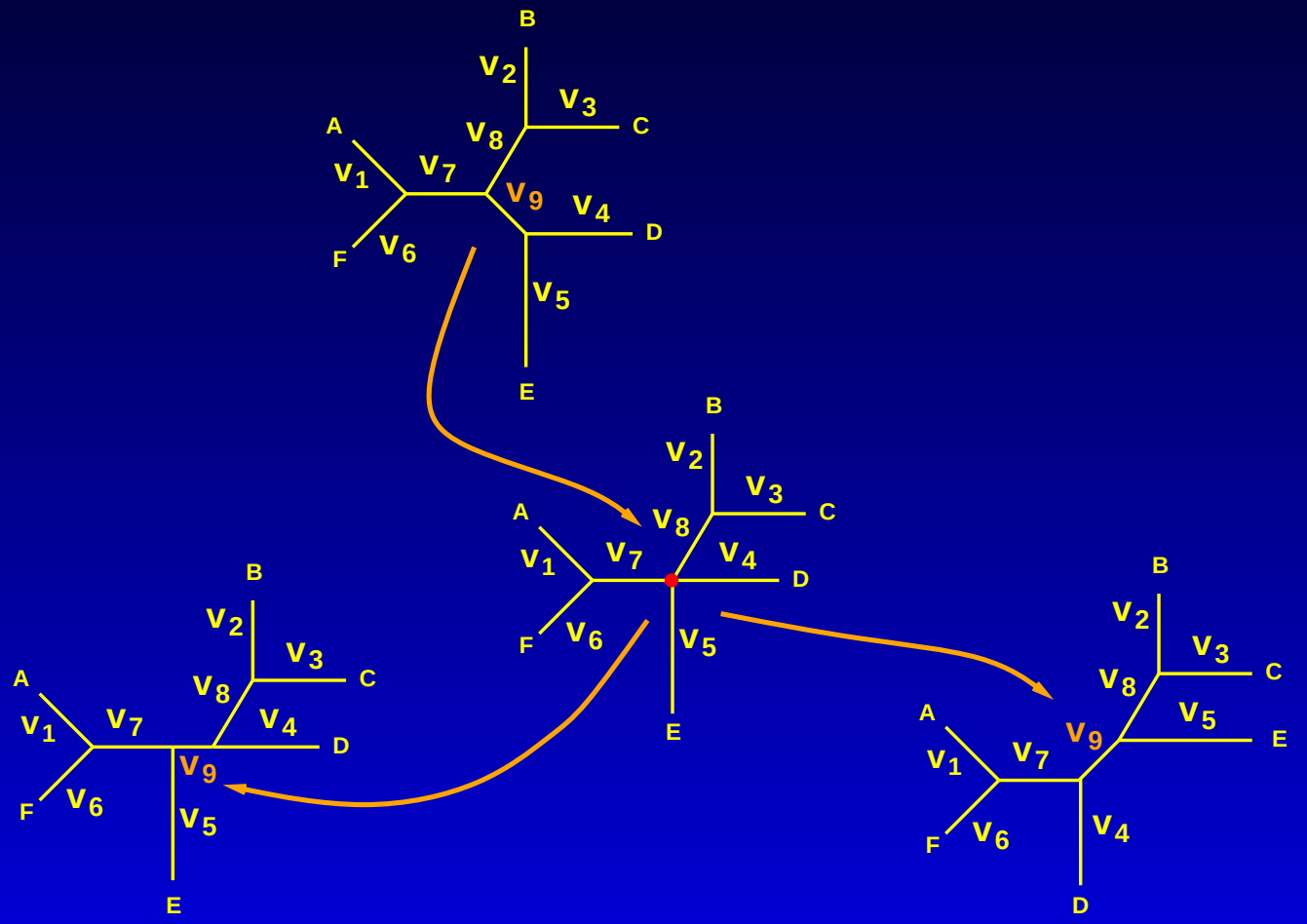
The space of trees varying all $2n-3$ branch lengths, each a nonegative number, defines an “orthant" (open corner) of a $2n-3$-dimensional real space:





Through the looking-glass

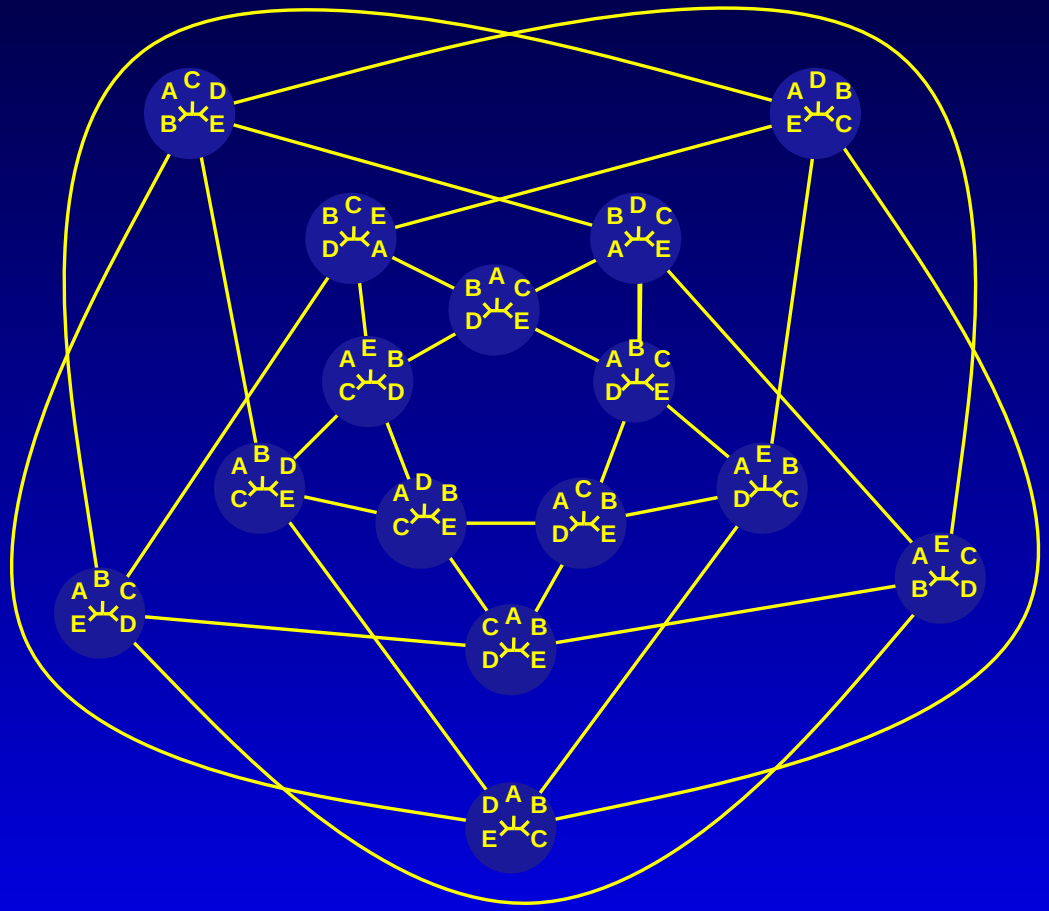
Shrinking one of the $n-1$ interior branches to 0, we arrive at a trifurcation:



Here, as we pass “through the looking glass" we are also touch the space for two other tree topologies, and we could decide to enter either.

The graph of all trees of 5 species

The space of all these orthants, one for each topology, connecting ones that share faces (looking glasses):



The Schoenberg graph (all 15 trees of size 5 connected by NNI's)

There are very large numbers of trees

For 21 species, the number of possible unrooted tree topologies exceeds Avogadro's Number: it is

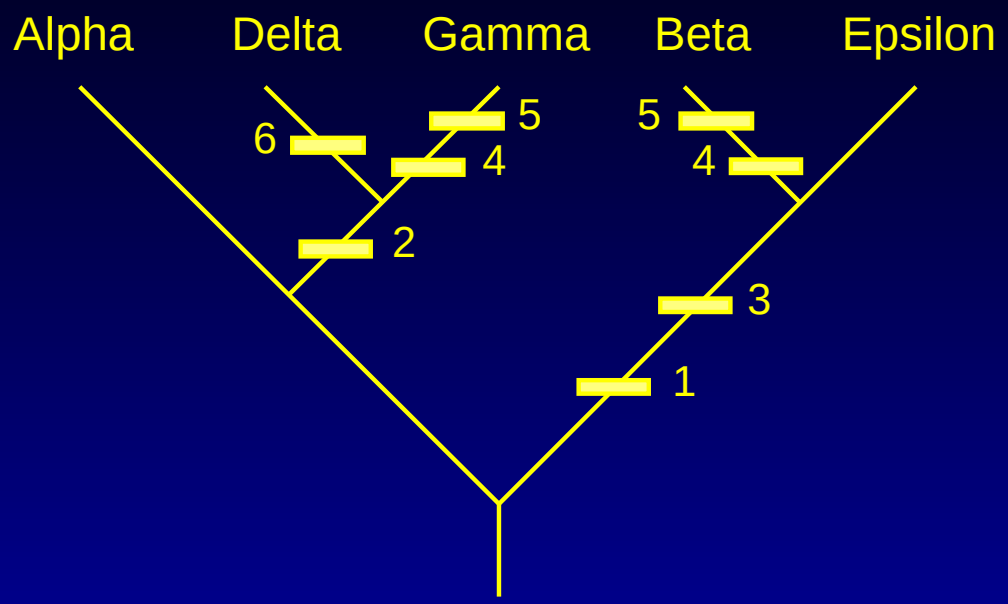
$$3 \times 5 \times 7 \times 9 \times 11 \times 13 \times 15 \times 17 \times 19$$
$$\times 21 \times 23 \times 25 \times 27 \times 29 \times 31 \times 33 \times 35 \times 37$$

$$= 8,200,794,532,637,891,559,375$$

... and that's not even asking about how hard it is to optimize the 39 branch lengths for each of these trees.

What this goes with is that most methods of finding the best tree are NP-hard, and not easy to approximate either.

Parsimony methods



Species	Sites 1	2	3	4	5	6
Alpha	T	A	G	C	A	T
Beta	C	A	A	G	C	T
Gamma	T	C	G	G	C	T
Delta	T	C	G	C	A	A
Epsilon	C	A	A	C	A	T

Advantages and disadvantages of parsimony methods

- Disadvantage: not model-based so people think it makes no assumptions.

Advantages and disadvantages of parsimony methods

- Disadvantage: not model-based so people think it makes no assumptions.
- Advantage: reasonably fast, no search of branch lengths needed and quick to compute the criterion.

Advantages and disadvantages of parsimony methods

- Disadvantage: not model-based so people think it makes no assumptions.
- Advantage: reasonably fast, no search of branch lengths needed and quick to compute the criterion.
- Advantage: good statistical properties when amounts of change are small.

Advantages and disadvantages of parsimony methods

- Disadvantage: not model-based so people think it makes no assumptions.
- Advantage: reasonably fast, no search of branch lengths needed and quick to compute the criterion.
- Advantage: good statistical properties when amounts of change are small.
- Disadvantage: statistical misbehavior (inconsistency) when some nearby branches on the tree are long (Long Branch Attraction).

Advantages and disadvantages of parsimony methods

- Disadvantage: not model-based so people think it makes no assumptions.
- Advantage: reasonably fast, no search of branch lengths needed and quick to compute the criterion.
- Advantage: good statistical properties when amounts of change are small.
- Disadvantage: statistical misbehavior (inconsistency) when some nearby branches on the tree are long (Long Branch Attraction).
- Disadvantage: likely to make you think you have William of Ockham's endorsement.

Advantages and disadvantages of parsimony methods

- Disadvantage: not model-based so people think it makes no assumptions.
- Advantage: reasonably fast, no search of branch lengths needed and quick to compute the criterion.
- Advantage: good statistical properties when amounts of change are small.
- Disadvantage: statistical misbehavior (inconsistency) when some nearby branches on the tree are long (Long Branch Attraction).
- Disadvantage: likely to make you think you have William of Ockham's endorsement.
- Disadvantage: may lead to the delusion that you know exactly what happened in evolution, in detail.

Distance matrix methods

The sequences:

A	CCTAACCTCTGACCC ...
B	CGTAACCTCCGGCCC ...
C	CGTAACCTCTGGCCC ...
D	CGCAACCTCTGGCTC ...
E	CCTAACCTCTGGCCC ...

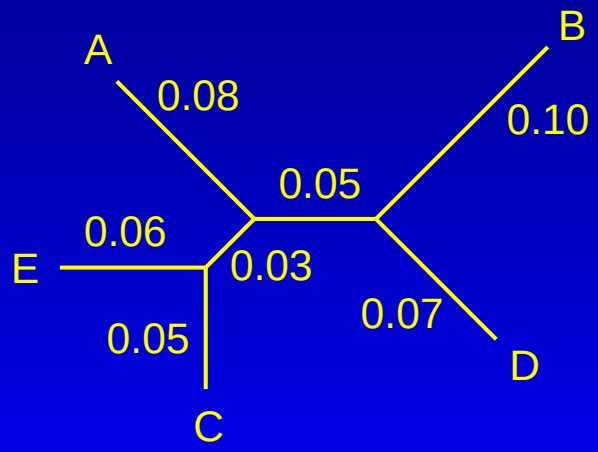
yield distances:

	A	B	C	D	E
A	0	0.19	0.17	0.20	0.16
B	0.19	0	0.24	0.15	0.24
C	0.17	0.24	0	0.25	0.10
D	0.20	0.15	0.25	0	0.24
E	0.16	0.24	0.10	0.24	0

alter tree until predictions match observed distances as closely as possible

compare:

A suggested tree:



predicts:

	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	0.20	0.11
D	0.20	0.17	0.20	0	0.21
E	0.17	0.24	0.11	0.21	0

Advantages and disadvantages of distance methods

- Advantage: model-based so assumptions are clearer.

Advantages and disadvantages of distance methods

- Advantage: model-based so assumptions are clearer.
- Advantage: it's geometry so mathematical scientists love it.

Advantages and disadvantages of distance methods

- Advantage: model-based so assumptions are clearer.
- Advantage: it's geometry so mathematical scientists love it.
- Advantage: often fast (especially Neighbor-Joining method), can handle large numbers of sequences.

Advantages and disadvantages of distance methods

- Advantage: model-based so assumptions are clearer.
- Advantage: it's geometry so mathematical scientists love it.
- Advantage: often fast (especially Neighbor-Joining method), can handle large numbers of sequences.
- Disadvantage: not using data fully statistically efficiently.

Advantages and disadvantages of distance methods

- Advantage: model-based so assumptions are clearer.
- Advantage: it's geometry so mathematical scientists love it.
- Advantage: often fast (especially Neighbor-Joining method), can handle large numbers of sequences.
- Disadvantage: not using data fully statistically efficiently.
- Advantage: when tested by simulation, found to be surprisingly efficient anyway.

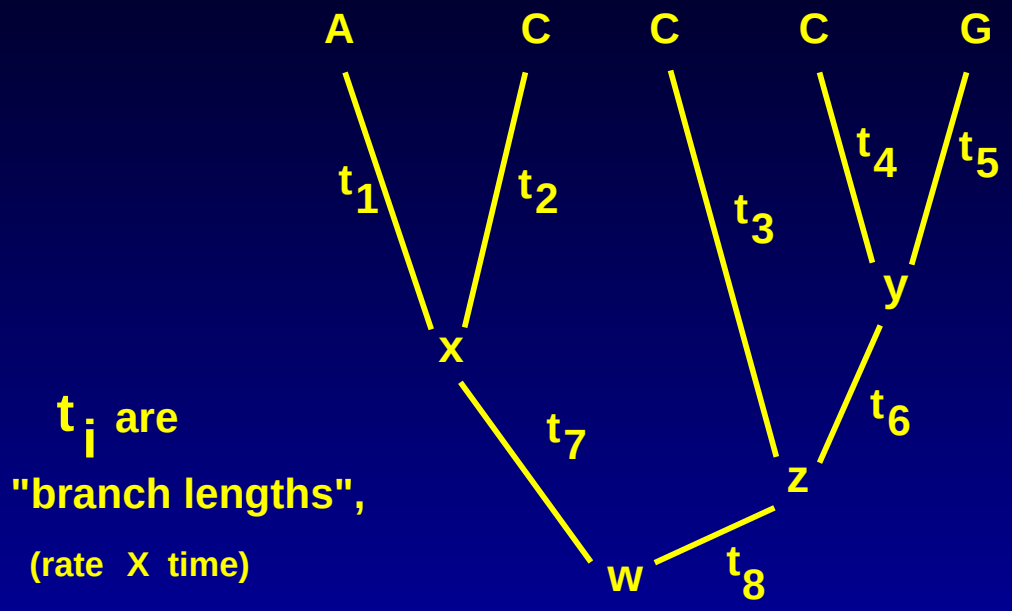
Advantages and disadvantages of distance methods

- Advantage: model-based so assumptions are clearer.
- Advantage: it's geometry so mathematical scientists love it.
- Advantage: often fast (especially Neighbor-Joining method), can handle large numbers of sequences.
- Disadvantage: not using data fully statistically efficiently.
- Advantage: when tested by simulation, found to be surprisingly efficient anyway.
- Disadvantage: cannot easily propagate some information about local features in the sequences from one distance calculation to another.

Advantages and disadvantages of distance methods

- Advantage: model-based so assumptions are clearer.
- Advantage: it's geometry so mathematical scientists love it.
- Advantage: often fast (especially Neighbor-Joining method), can handle large numbers of sequences.
- Disadvantage: not using data fully statistically efficiently.
- Advantage: when tested by simulation, found to be surprisingly efficient anyway.
- Disadvantage: cannot easily propagate some information about local features in the sequences from one distance calculation to another.
- Disadvantage: it's geometry so mathematical scientists hang onto it beyond the point of reason.

Maximum likelihood



To compute the likelihood for one site, sum over all possible states (bases) at interior nodes:

$$
\begin{aligned}
L^{(i)} = & \sum_x \sum_y \sum_z \sum_w \text{Prob}\,(w)\ \text{Prob}\,(x \mid w, t_7) \\
& \times \text{Prob}\,(A \mid x, t_1)\ \text{Prob}\,(C \mid x, t_2)\ \text{Prob}\,(z \mid w, t_8) \\
& \times \text{Prob}\,(C \mid z, t_3)\ \text{Prob}\,(y \mid z, t_6)\ \text{Prob}\,(C \mid y, t_4)\ \text{Prob}\,(G \mid y, t_5)
\end{aligned}
$$

Advantages and disadvantages of likelihood

- Advantage: uses a model, so assumptions are clear.

Advantages and disadvantages of likelihood

- Advantage: uses a model, so assumptions are clear.
- Advantage: fully statistically efficient.

Advantages and disadvantages of likelihood

- Advantage: uses a model, so assumptions are clear.
- Advantage: fully statistically efficient.
- Disadvantage: computationally slower.

Advantages and disadvantages of likelihood

- Advantage: uses a model, so assumptions are clear.
- Advantage: fully statistically efficient.
- Disadvantage: computationally slower.
- Advantage: statistical testing by likelihood ratio tests available

Advantages and disadvantages of likelihood

- Advantage: uses a model, so assumptions are clear.
- Advantage: fully statistically efficient.
- Disadvantage: computationally slower.
- Advantage: statistical testing by likelihood ratio tests available
- Disadvantage: can't use the LRT test to test tree topologies.

Bayesian inference methods

Basically uses the likelihood machinery, and adds priors on parameters and on trees.

Implemented by Markov chain Monte Carlo methods to sample from the posterior on trees (or parameters, or both).

Very popular right now.

- Advantage: interpretation is straightforward, once the assumptions are met.

Bayesian inference methods

Basically uses the likelihood machinery, and adds priors on parameters and on trees.

Implemented by Markov chain Monte Carlo methods to sample from the posterior on trees (or parameters, or both).

Very popular right now.

- Advantage: interpretation is straightforward, once the assumptions are met.
- Advantage: gives you what you want, the probability of the result.

Bayesian inference methods

Basically uses the likelihood machinery, and adds priors on parameters and on trees.

Implemented by Markov chain Monte Carlo methods to sample from the posterior on trees (or parameters, or both).

Very popular right now.

- Advantage: interpretation is straightforward, once the assumptions are met.
- Advantage: gives you what you want, the probability of the result.
- Disadvantage: how long is long enough to run the MCMC?

Bayesian inference methods

Basically uses the likelihood machinery, and adds priors on parameters and on trees.

Implemented by Markov chain Monte Carlo methods to sample from the posterior on trees (or parameters, or both).

Very popular right now.

- Advantage: interpretation is straightforward, once the assumptions are met.
- Advantage: gives you what you want, the probability of the result.
- Disadvantage: how long is long enough to run the MCMC?
- Disadvantage: where do we get priors from, what effect do they have?

Bayesian inference methods

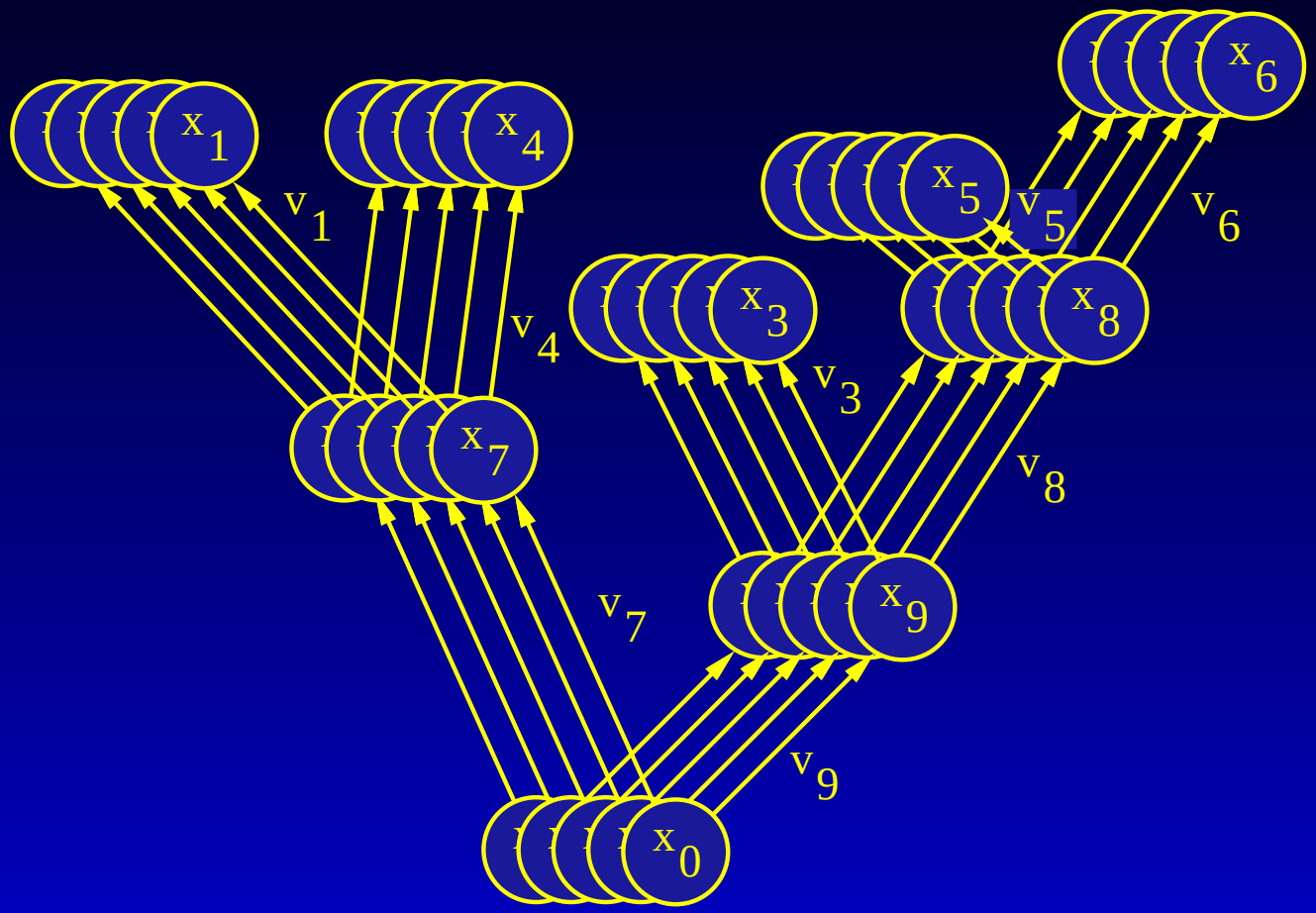
Basically uses the likelihood machinery, and adds priors on parameters and on trees.

Implemented by Markov chain Monte Carlo methods to sample from the posterior on trees (or parameters, or both).

Very popular right now.

- Advantage: interpretation is straightforward, once the assumptions are met.
- Advantage: gives you what you want, the probability of the result.
- Disadvantage: how long is long enough to run the MCMC?
- Disadvantage: where do we get priors from, what effect do they have?
- Disadvantage: they keep chanting in unison "We are the statisticians of Bayes – you will be assimilated."

Aren't these graphical models?



(You have to imaging it going back 500 layers or so). The problem is to use the data, which is at the tips but not available for the interior nodes, to infer the topology and branch lengths of the tree that is shared by all sites.

Could we use graphical model machinery here?

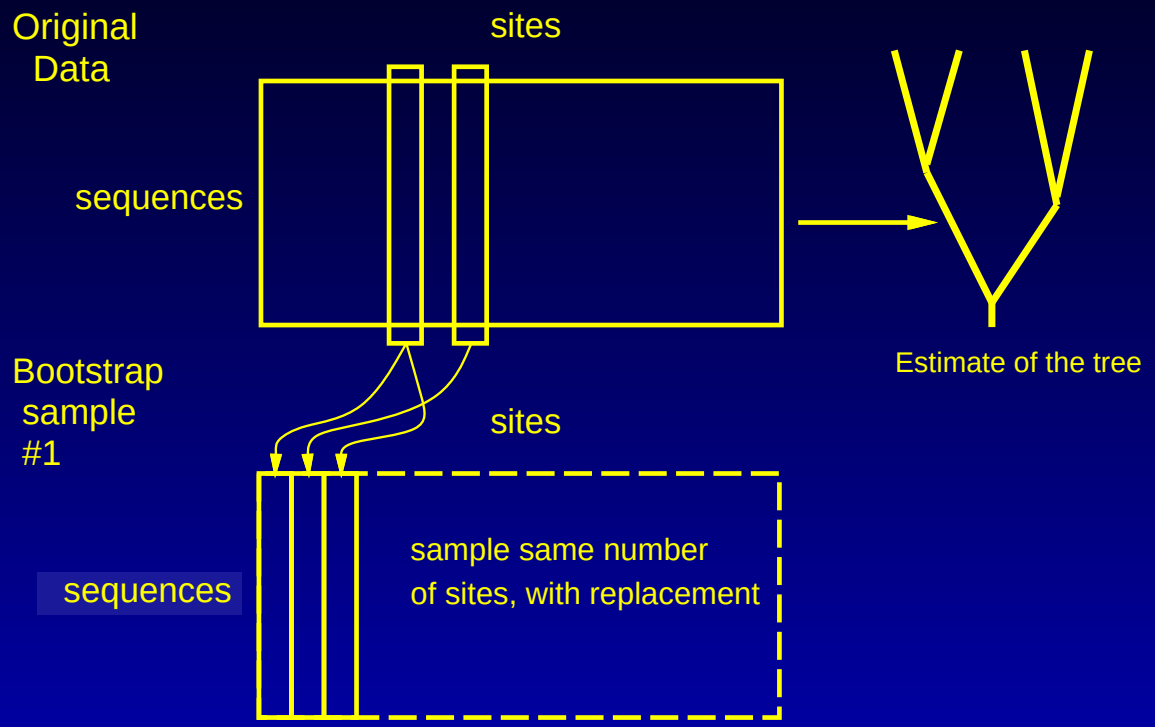
Like Moliére's character who is delighted to discover that he's been speaking prose all his life, we found we had already been using the relevant Graphical Model machinery since about 1973.

So alas there was nothing to gain.

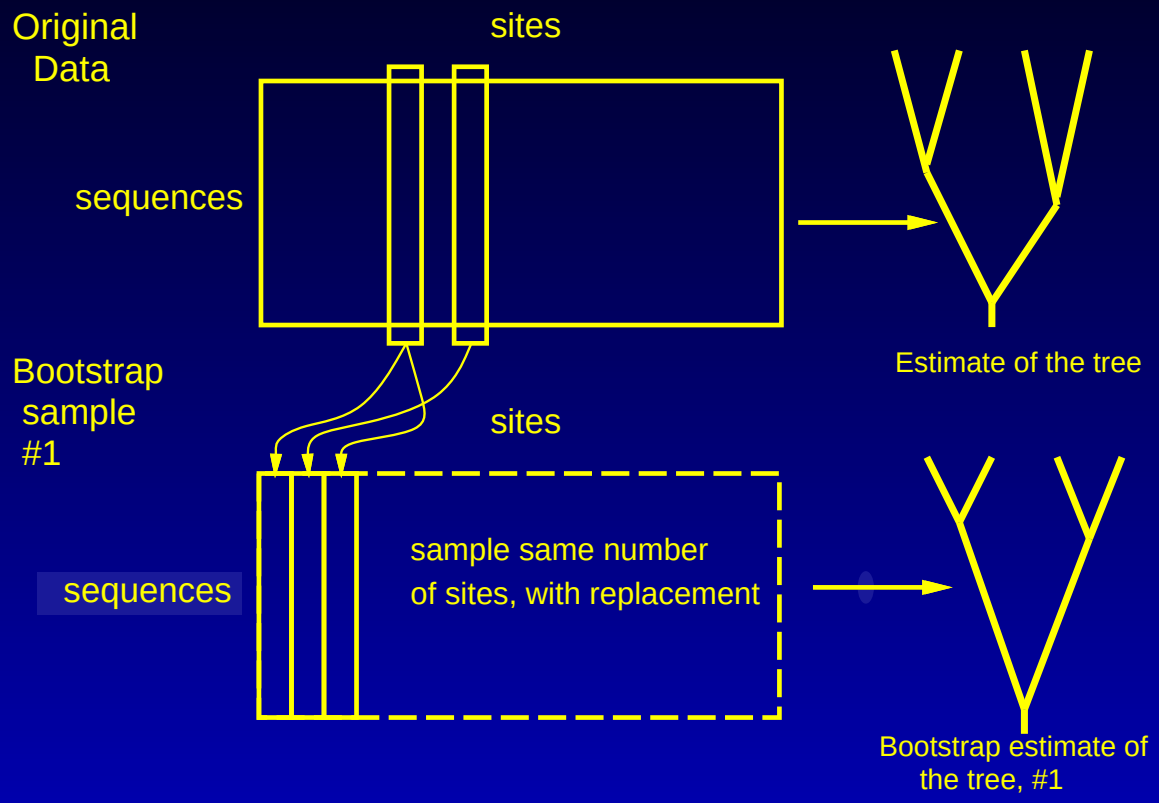
The same thing is true for statistical genetics, where the graphical model machinery reinvents the standard "peeling" algorithms for computing likelihoods on pedigrees, in use since 1970.

Bootstrap sampling of phylogenies

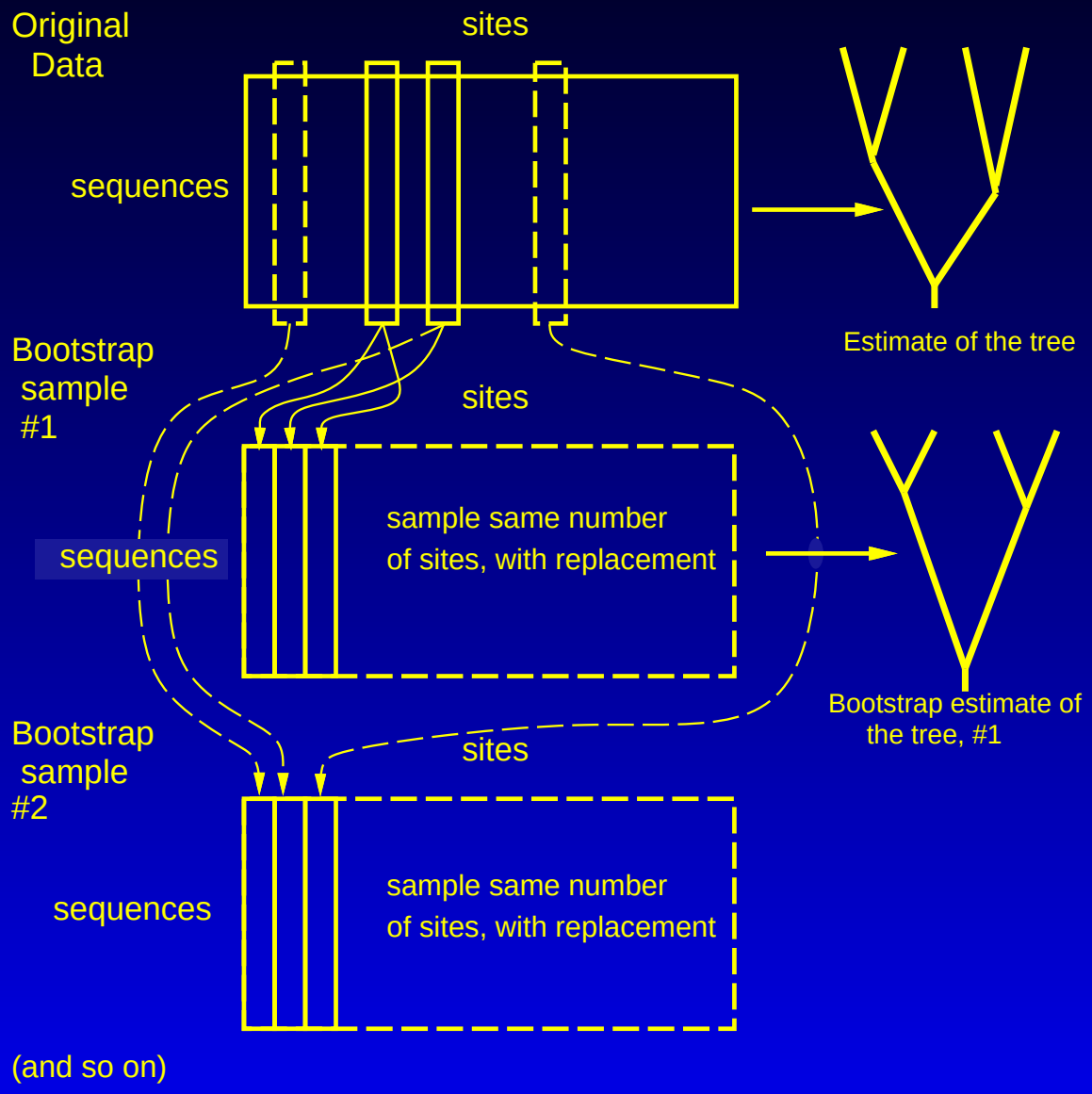
Draw columns randomly with replacement



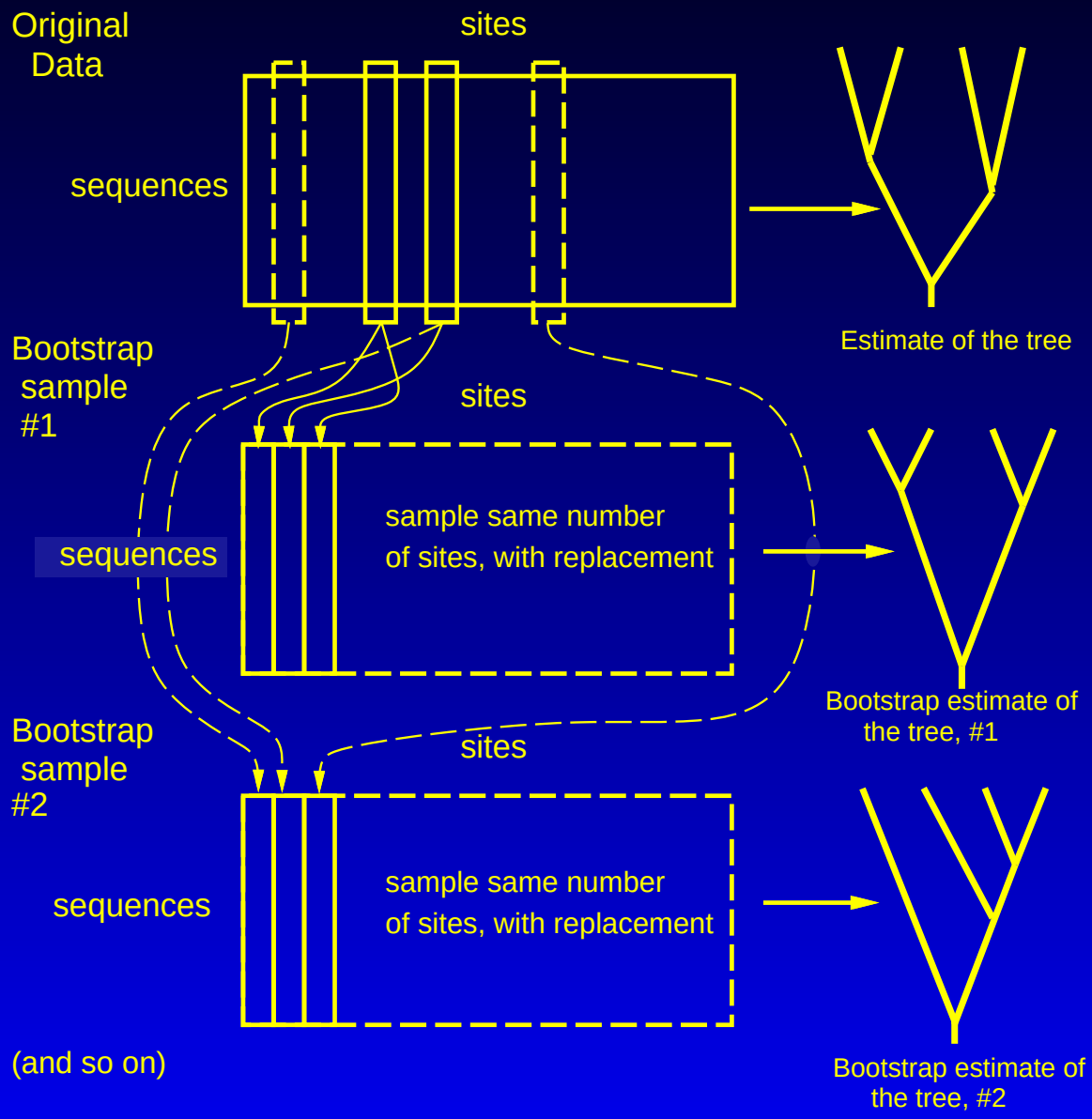
Make a tree from that resampled data set



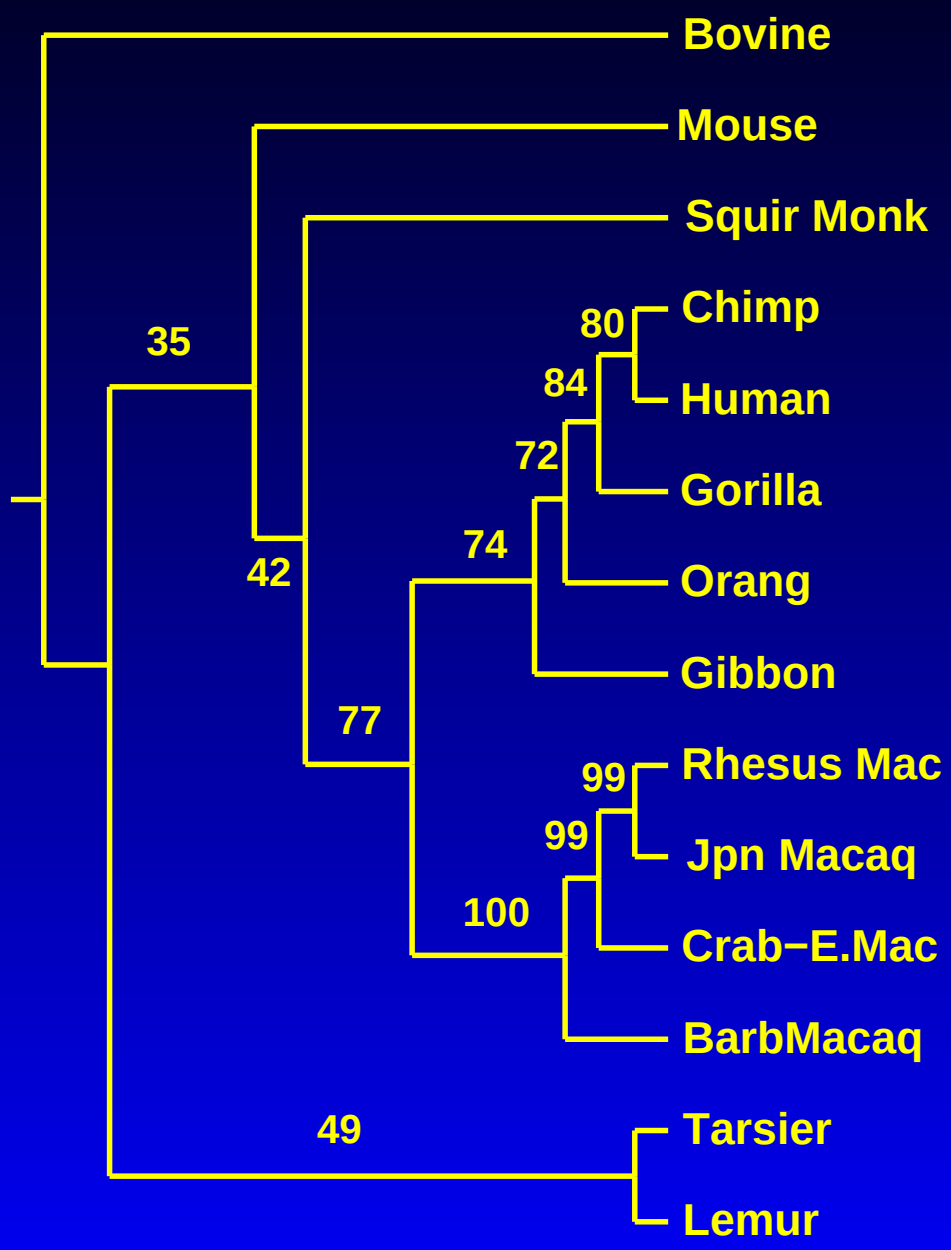
Draw another bootstrap sample



... and get a tree for it too. And so on.



Summarizing the cloud of trees by support for branches



Some alternatives to bootstrapping

- Parametric bootstrapping – same, but simulate data sets from our best estimate of the tree instead of sampling sites.

Some alternatives to bootstrapping

- Parametric bootstrapping – same, but simulate data sets from our best estimate of the tree instead of sampling sites.
- Bayesian inference of course gets statistical support information from the posterior.

Some alternatives to bootstrapping

- Parametric bootstrapping – same, but simulate data sets from our best estimate of the tree instead of sampling sites.
- Bayesian inference of course gets statistical support information from the posterior.
- The Kishino-Hasegawa-Templeton test (KHT test) which compares prespecified trees to each other by paired sites tests.

Why want to know the tree?

It affects all parts of the genomes – it is the essential part of propagating information about the evolution of one part of the genome to inquiries about another part.

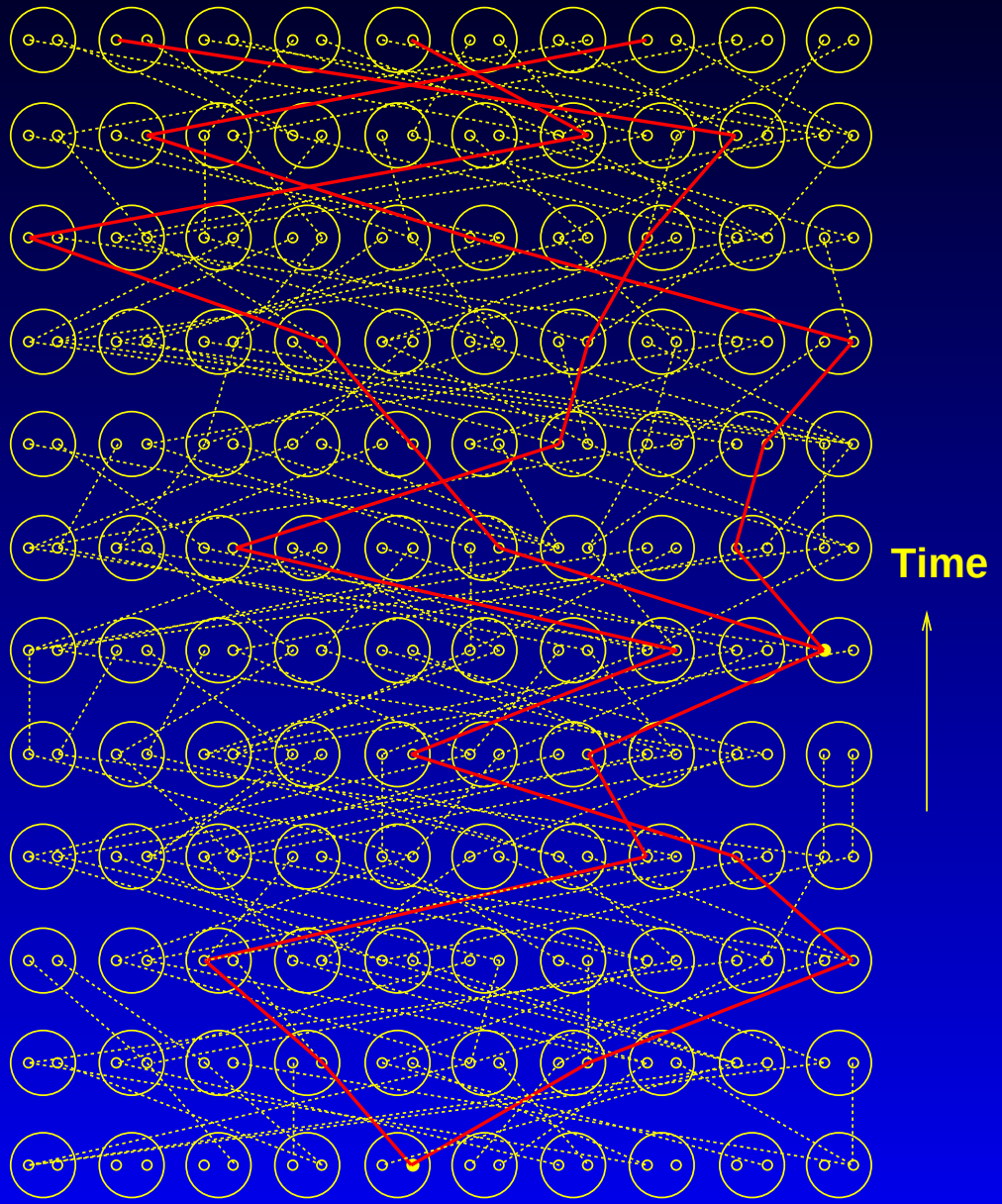
The standard method for finding functional regions of the genome is now using “PhyloHMMs” which use Hidden Markov Model machinery together with phylogenies to find regions that have unusually low rates of evolution.

Another kind of tree: the coalescent

Coalescent trees are trees of ancestry of copies of a single gene locus within a species. They are weakly inferrable as most have only a few sites (SNPs) varying among individuals.

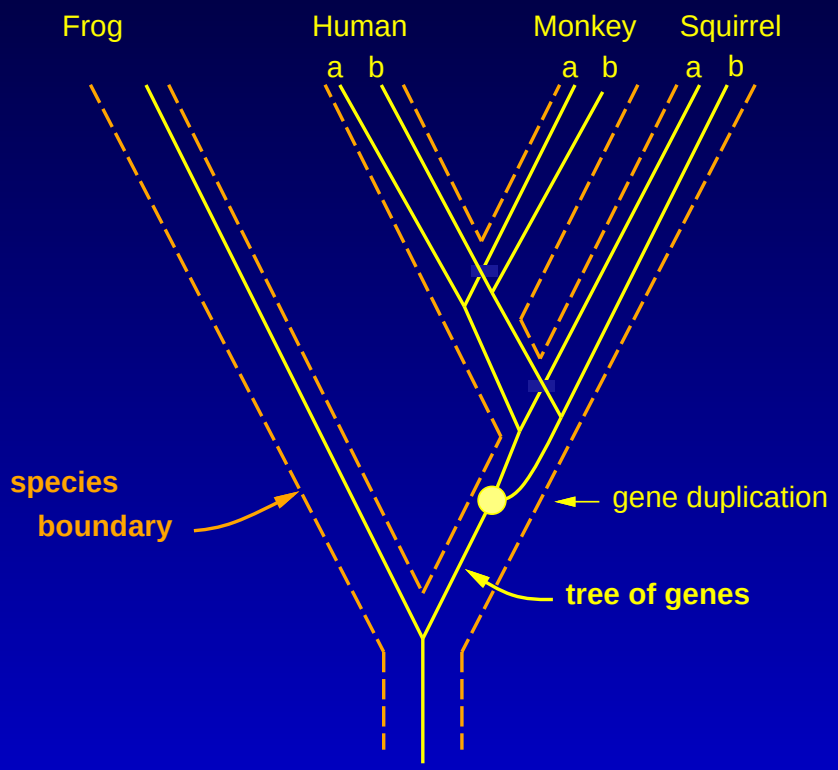
- Since each coalescent tree applies to a very short region of genome, maybe as little as one gene, there is less interest in the tree.
- But they do illuminate the values of parameters such as population size, migration rates, recombination rates etc. This allows us to accumulate information across different loci (genes).
- To don this we have to sum over our uncertainty about the tree by using MCMC methods, accumulating the information (as log likelihood or using Bayesian machinery) to make inferences about the parameters.
- This is the interface between within-species population genetics and between-species work on phylogenies.
- It is also the statistical foundation of inferences from mitochondrial genealogies ("mitochondrial Eve") and Y chromosome genealogies, and of the samples from the rest of the genome that are now being added to this.

A coalescent



Yet another kind of tree: trees of gene families

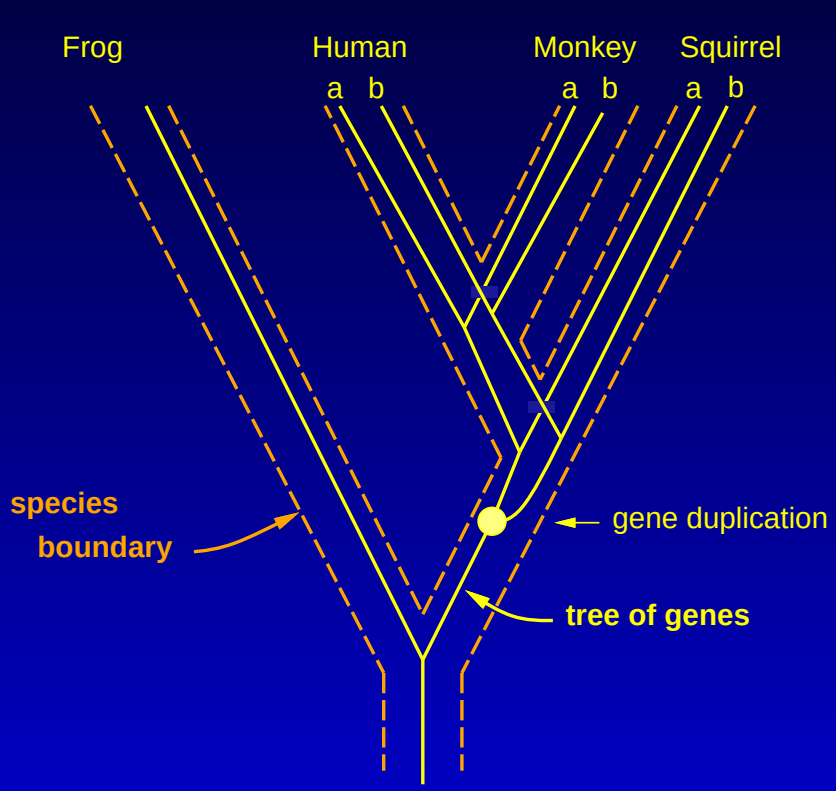
Gene duplications in evolution create new genes. Both the new gene and the original one then evolve.

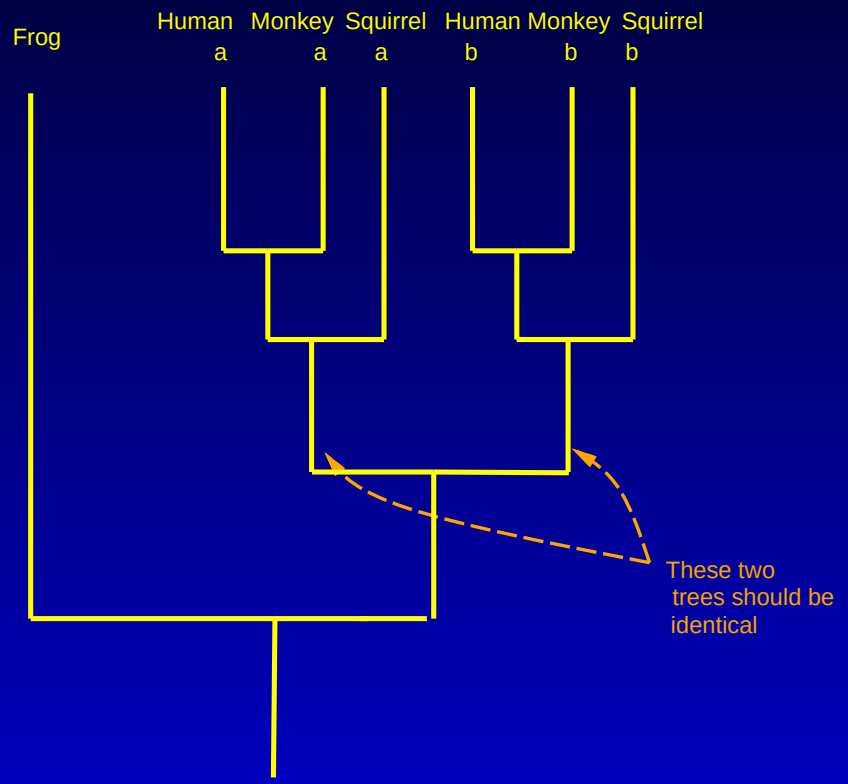


Some forks are gene duplications, leading to subtrees that are all supposed to have the same phylogeny as they are in the same set of species. Example: Hemoglobin proteins.

Yet another kind of tree: trees of gene families

Gene duplications in evolution create new genes. Both the new gene and the original one then evolve.





Some forks are gene duplications, leading to subtrees that are all supposed to have the same phylogeny as they are in the same set of species. Example: Hemoglobin proteins.

References

Felsenstein, J. 2004. *Inferring Phylogenies.* Sinauer Associates, Sunderland, Massachusetts. [Book you and all your friends must rush out and buy]

Semple, C. and M. Steel. 2003. *Phylogenetics.* Oxford Lecture Series in Mathematics and Its Applications, 24. Oxford University Press. [More rigorous mathematical treatment]

Yang, Z. 2007. *Computational Molecular Evolution.* Oxford Series in Ecology and Evolution. Oxford University Press, Oxford. [Careful survey of molecular phylogeny methods, from a leader]

For a list of 348 phylogeny programs, many available free, see

`http://evolution.gs.washington.edu/phylip/software.html`