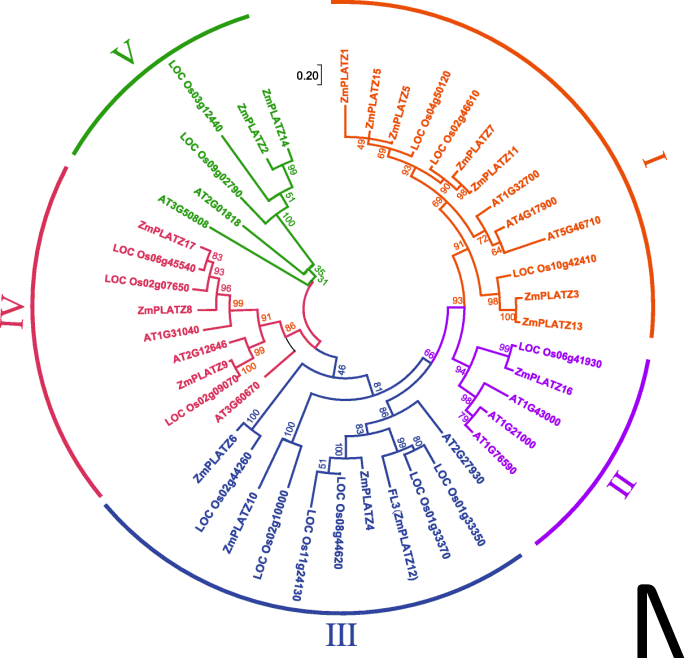
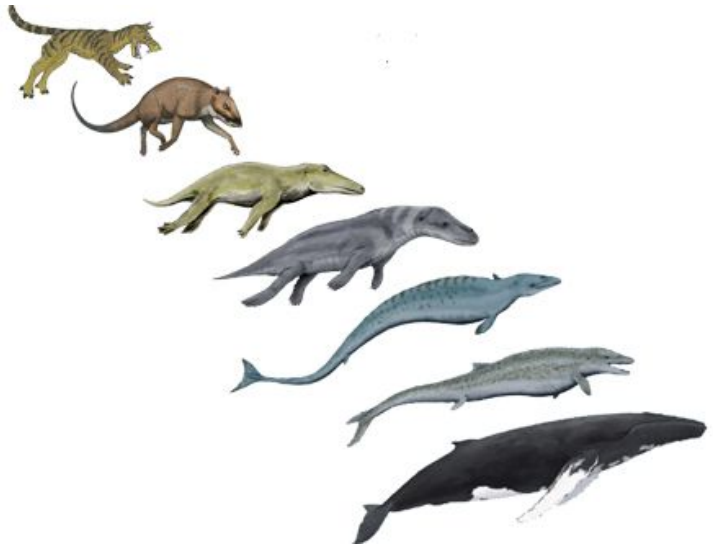
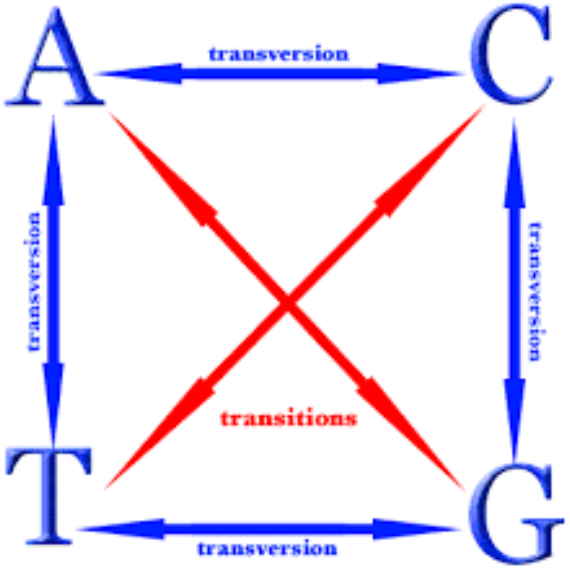
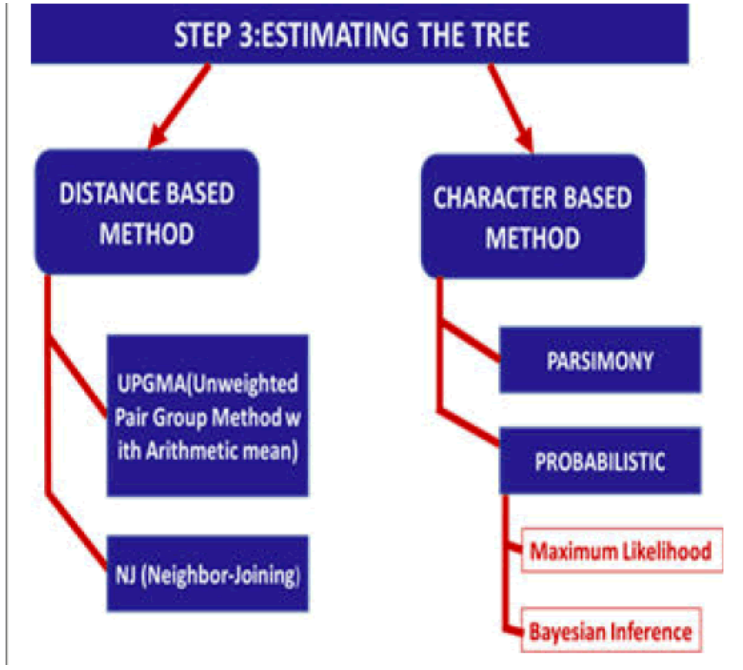




Maximum-likelihood

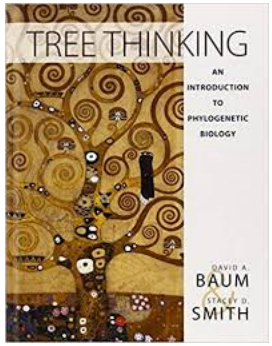
Nov 7th, 2019



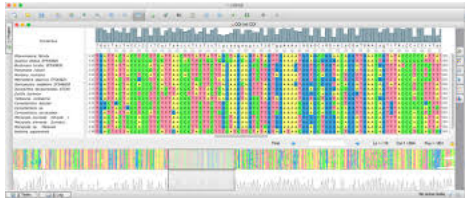


Covered on Tuesday

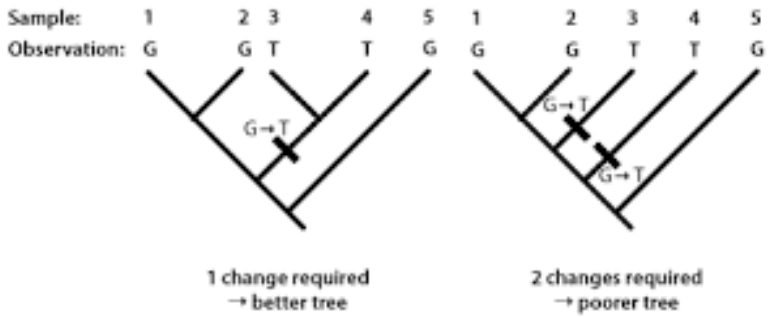
- Tree thinking



- Sequence alignment



- Parsimony for tree inference



Topics to cover today

- Specifying models of molecular evolution
- Maximum-likelihood inference
- Bootstrap support
- Great, you have a tree…now what? Comparative phylogenetic methods

Just like the Great White Shark these creatures never had to evolve…

Is this true?

Horseshoe crabs

Crocodiles

Equisetum

Ginkgo

Okapi

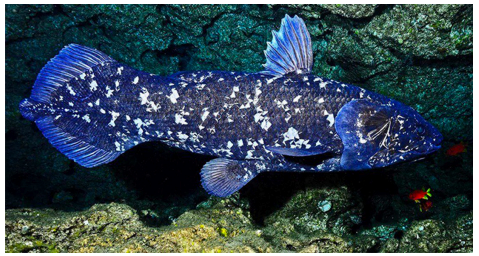
Coelacanth

Amborella

"Never had to evolve"…not really

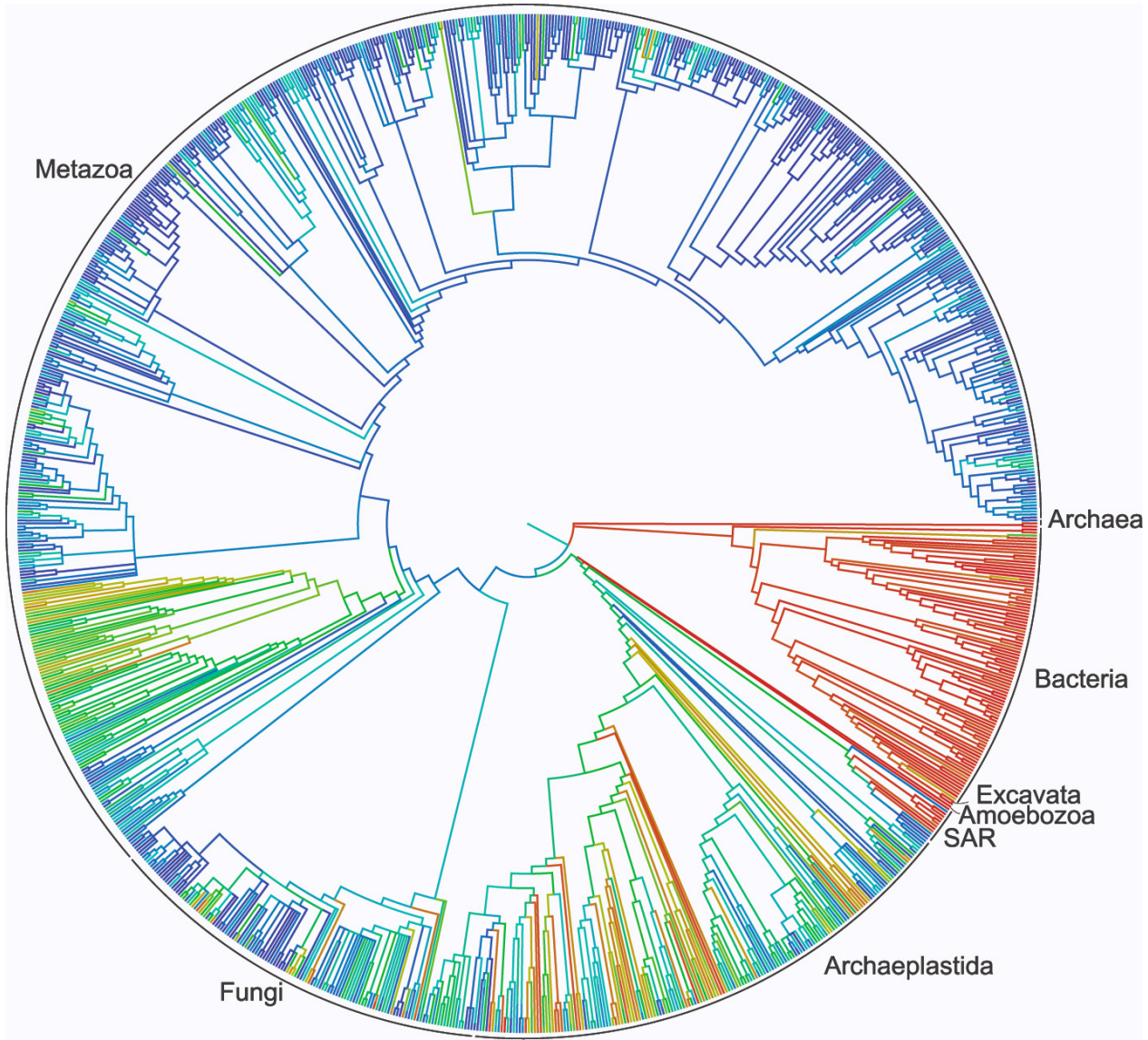
Horseshoe crabs

Ginkgo

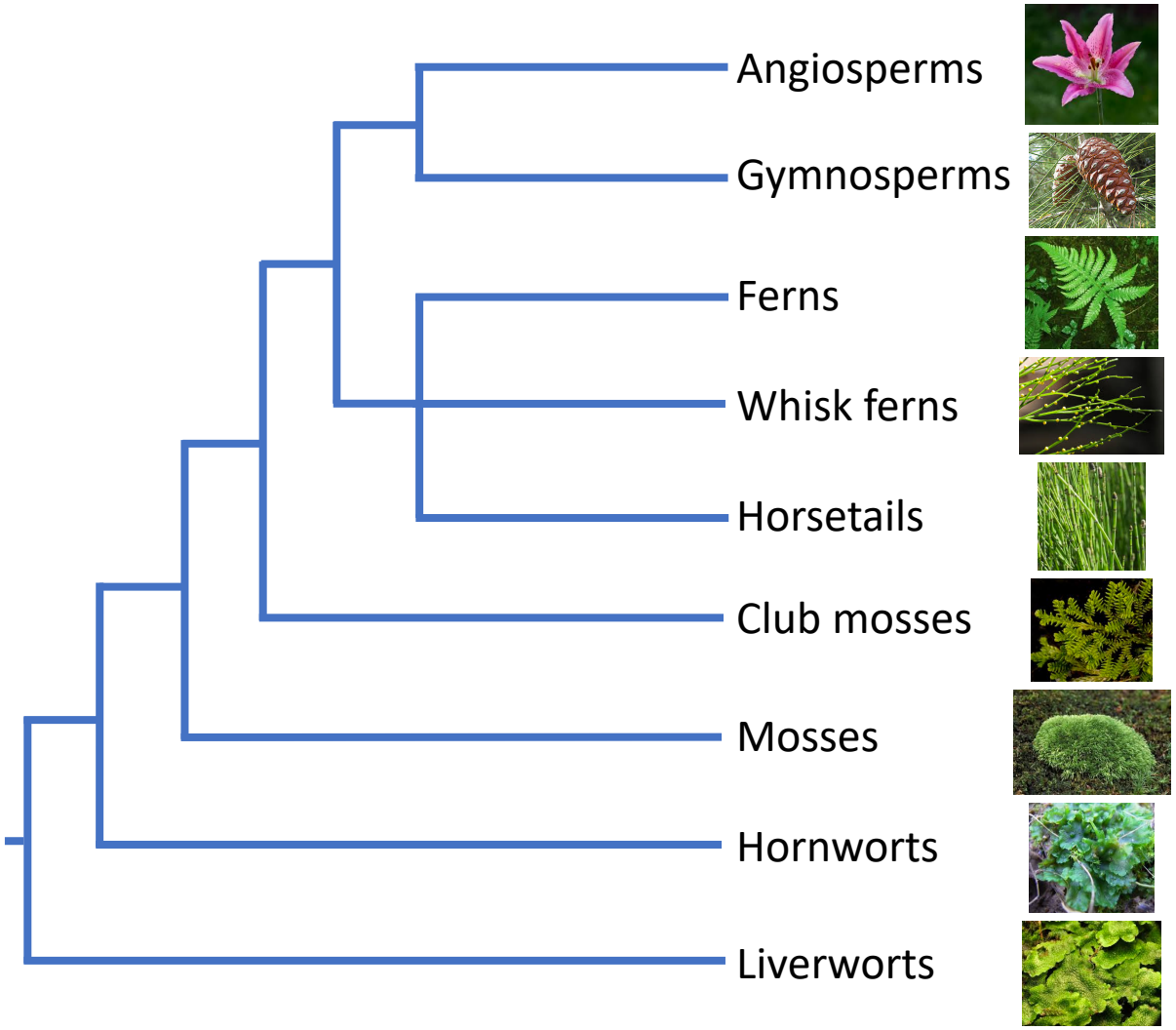


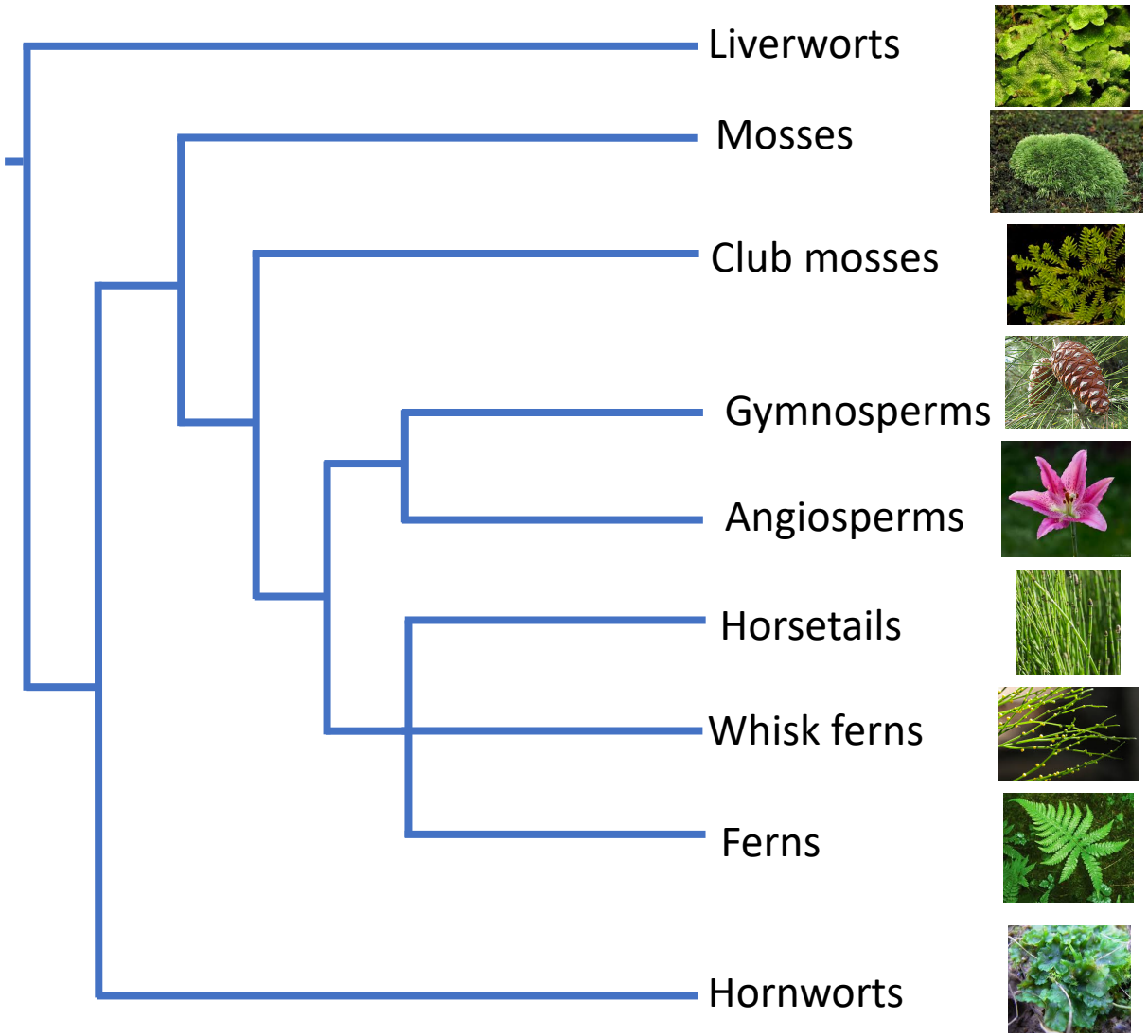
Coelacanth

Amborella



Which are “higher” plants?

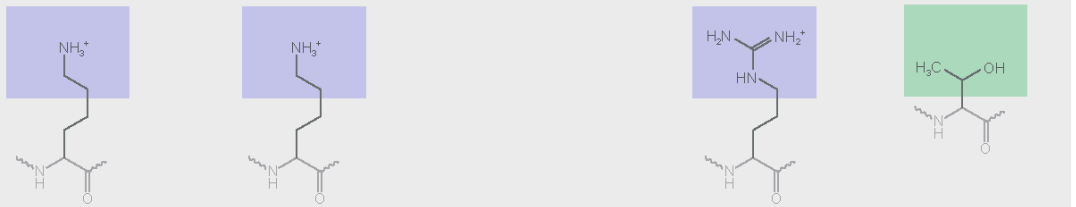




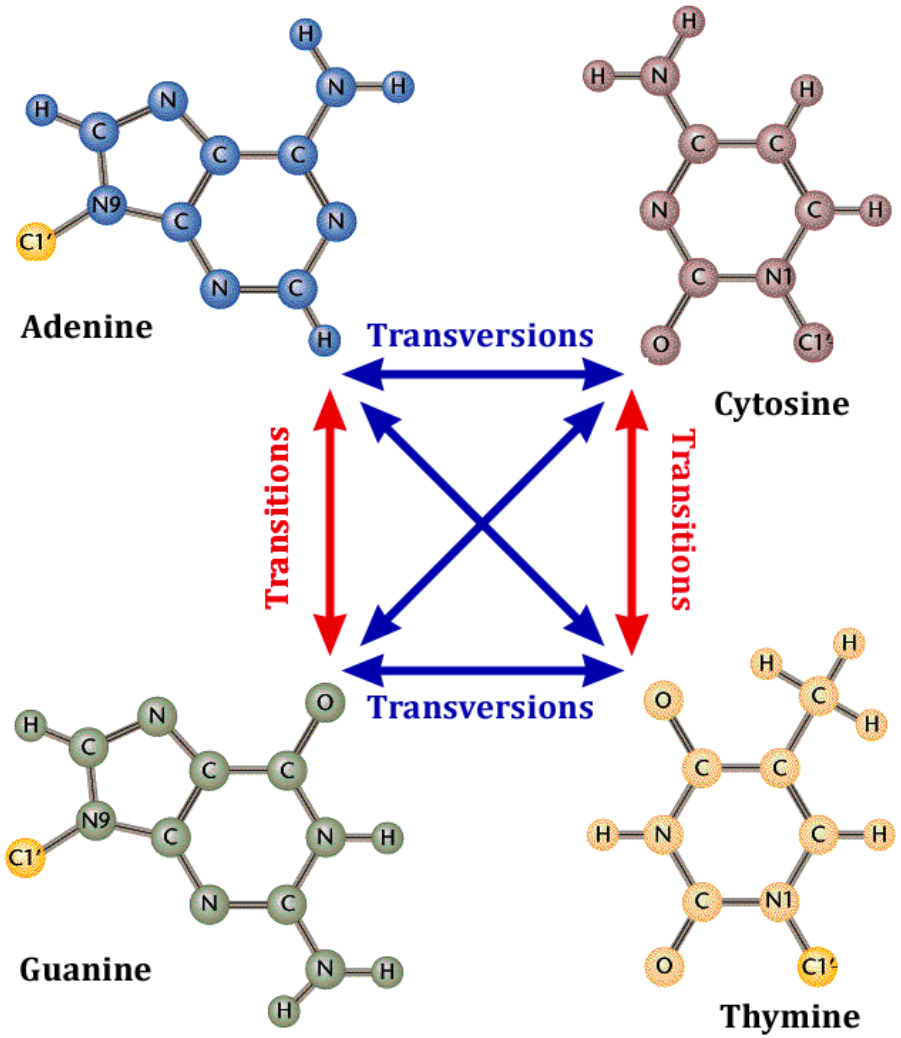
How do nucleotides change?

- Change at any given site is independent of the base in its prior iteration

	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr

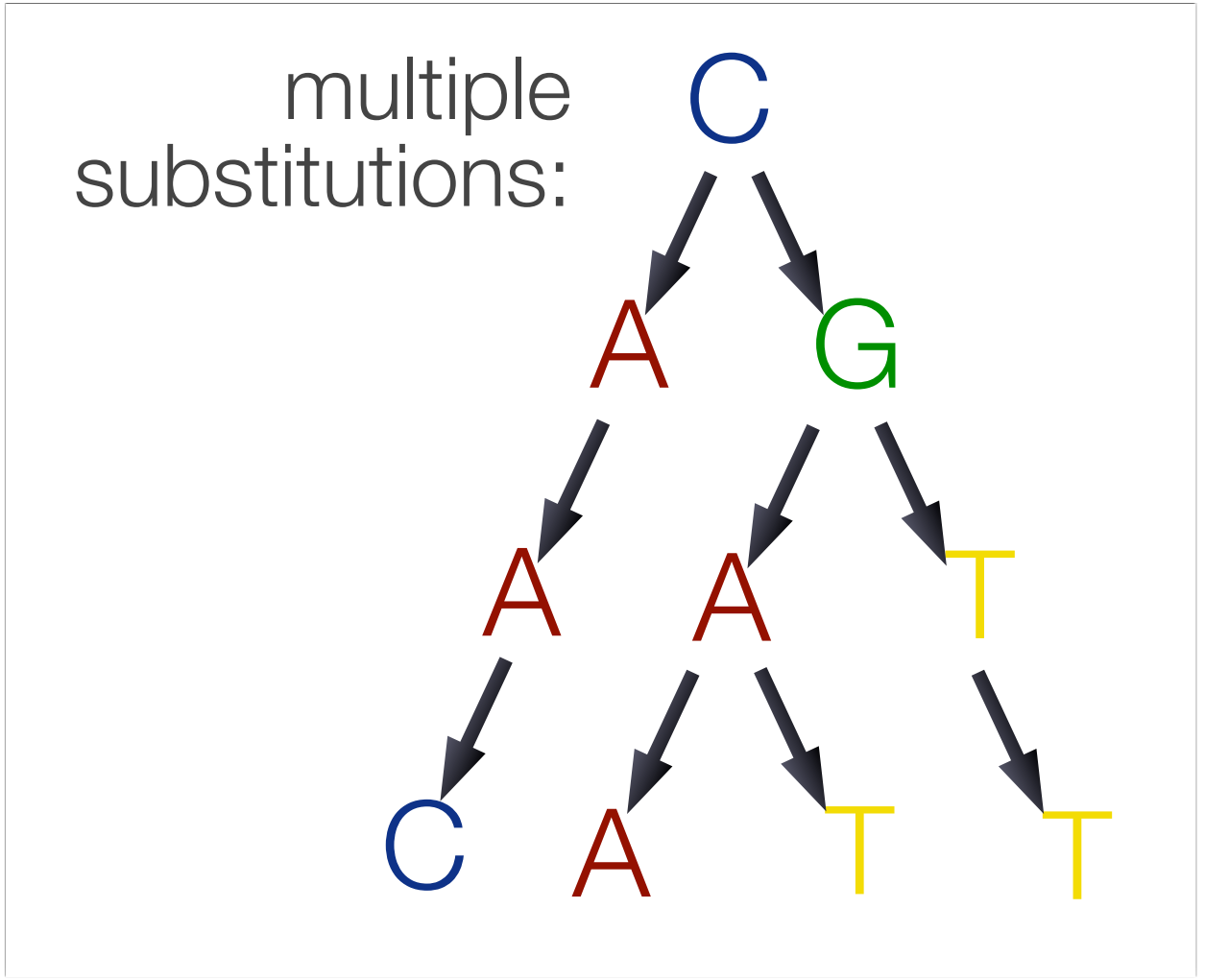


basic
polar



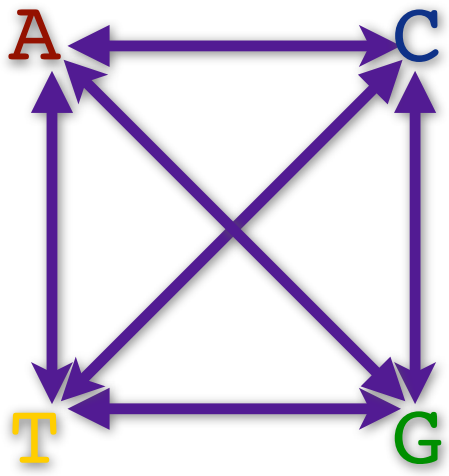
Multiple substitutions

- Given enough time, may be multiple changes
- 25% of nucleotide sites are expected to be identical by chance
- Models vary in complexity from very simple to extremely complex
- Model choice is important
- Model must suit the data



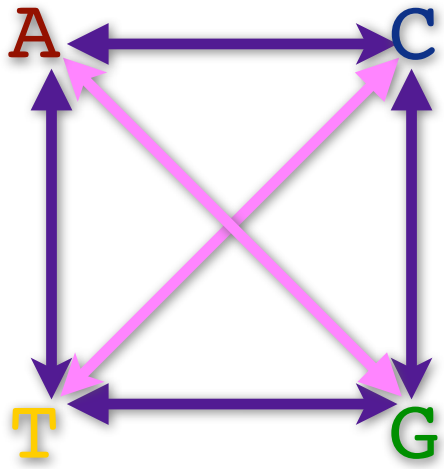
Underlying models

Jukes Cantor



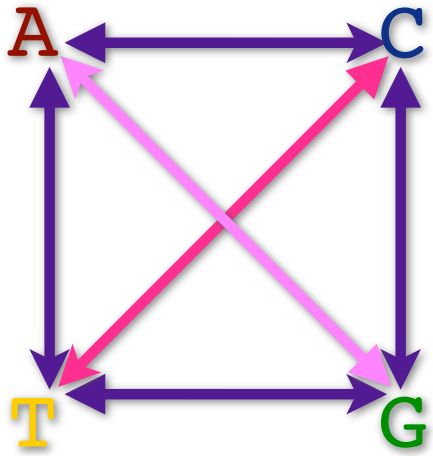
- All bases evolve independently
- All bases are equal frequency
- Each base can change with equal probability

Kimura



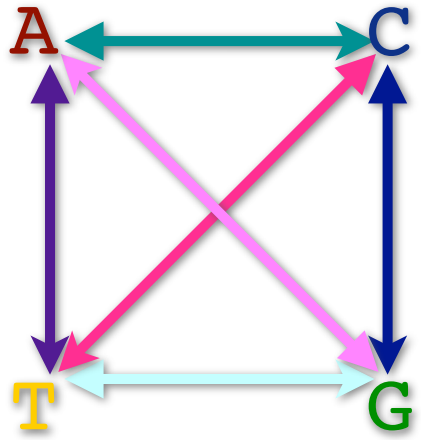
- All bases evolve independently
- All bases are equal frequency
- Transitions and transversions evolve at different rates

TrN



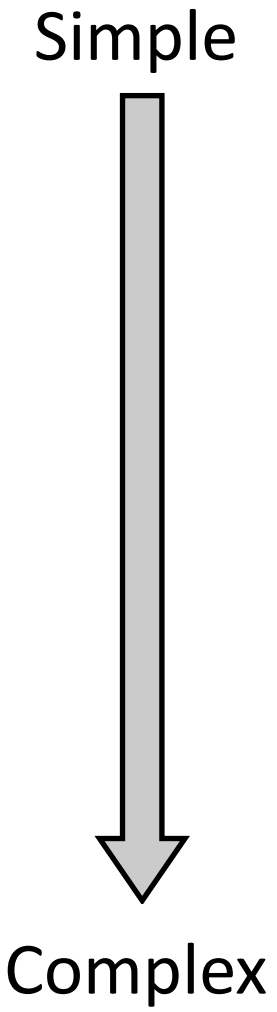
- All bases evolve independently
- All bases at unequal frequency
- Transitions and transversions evolve at two different rates

GTR



- All bases evolve independently
- All bases at unequal frequency
- All changes occur at different rates

Model hierarchy



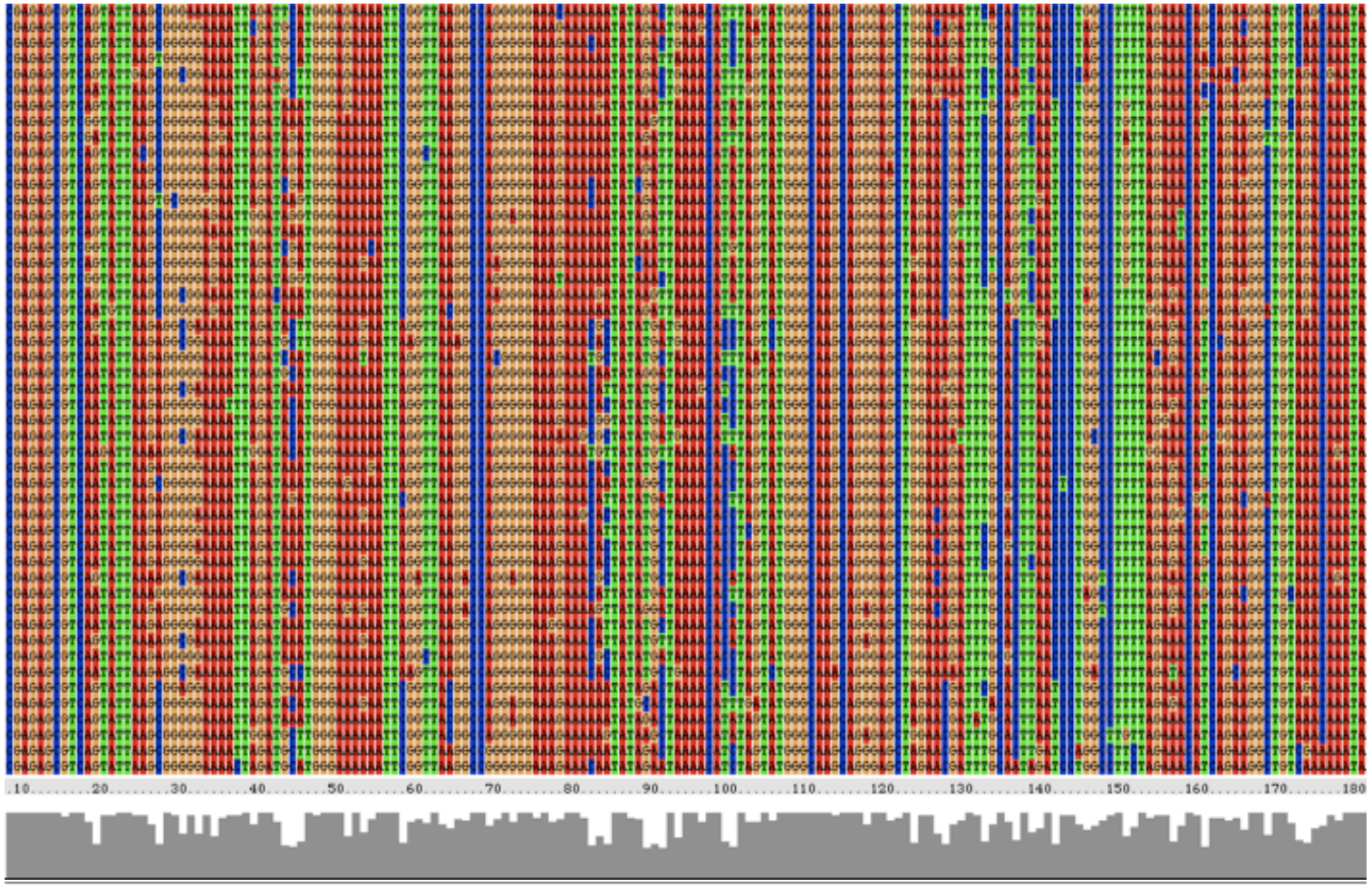
base frequencies are equal and
all substitutions are equally likely
(Jukes-Cantor)

base frequencies are equal but transitions and
transversions occur at different rates
(Kimura 2 parameter)

unequal base frequencies and transitions and
transversions occur at different rates
(Hasegawa-Kishino-Yano)

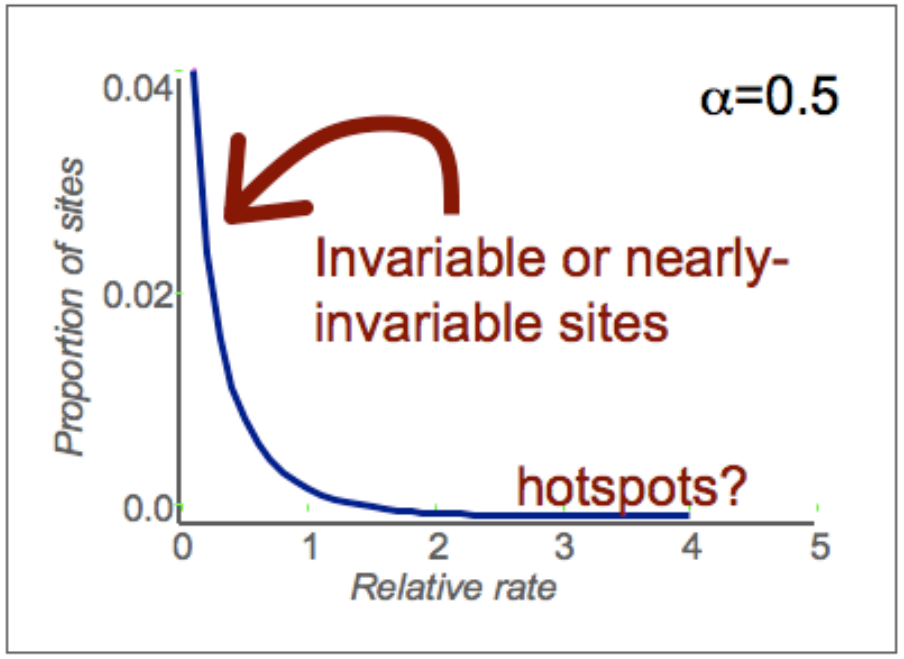
unequal base frequencies and all
substitution types occur at different rates
(General Reversible Model)

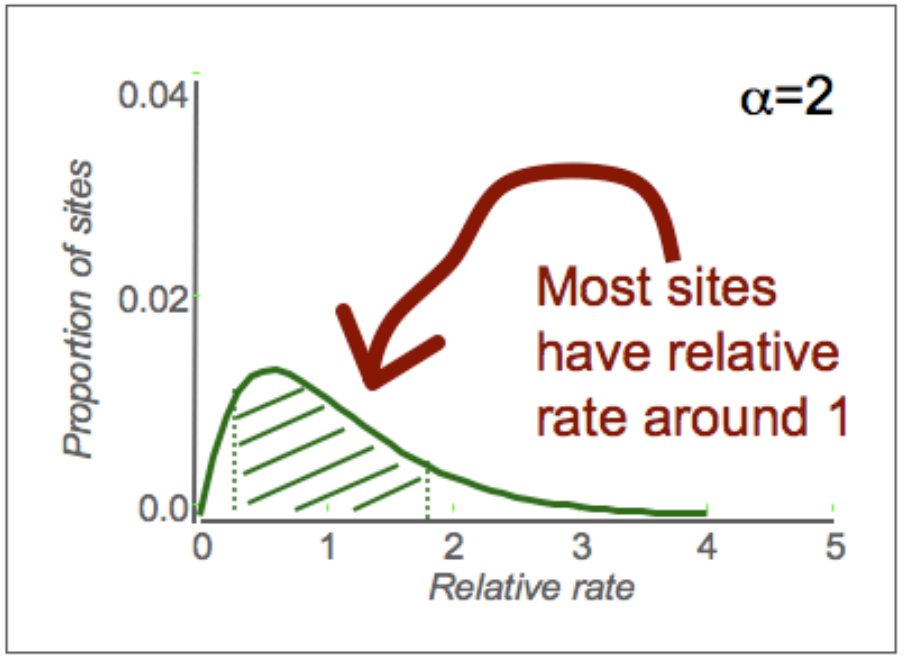
Rate Heterogeneity between sites



- Most models assume rate heterogeneity
  - 3rd position wobble
  - Hypervariable vs invariant sites
- Use a gamma distribution to model heterogeneity
  - Usually 4-8 discrete categories of rates

Rate heterogeneity





- Strong rate heterogeneity alpha < 1
- Weak rate heterogeneity alpha > 1

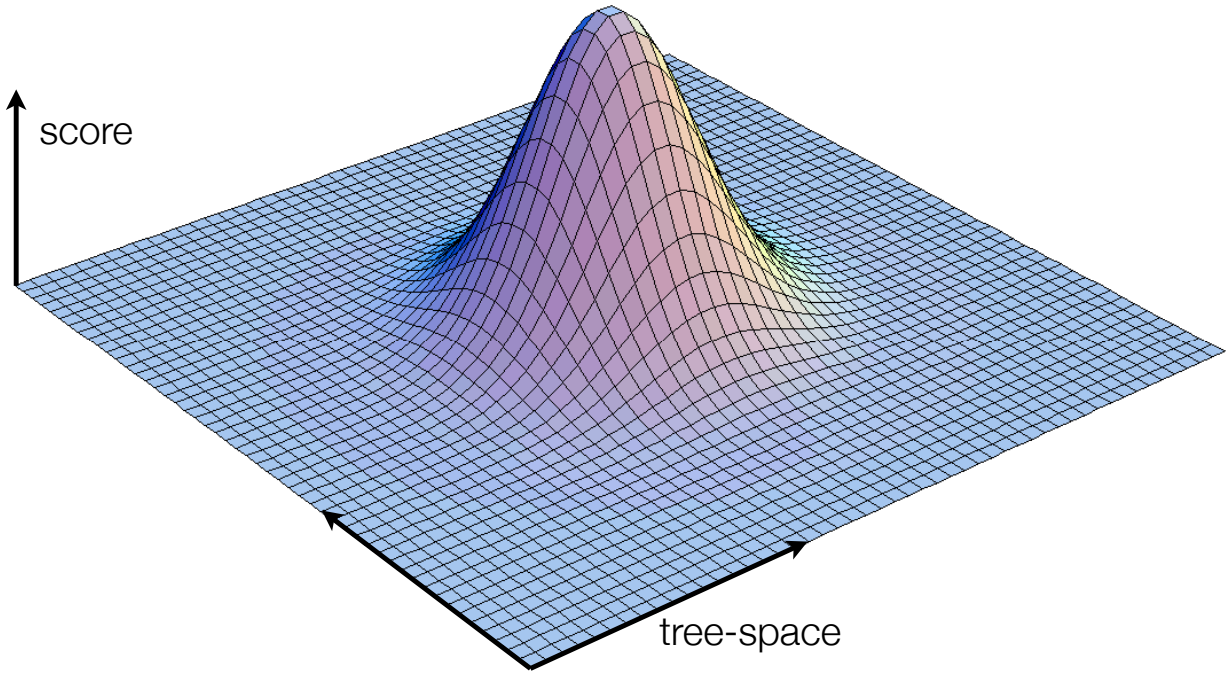
Maximum likelihood

- Closely related to the more common concept of probability
- Considered the most statistically valid approach for molecular phylogenetics
  - Along with Bayesian (next week)
- Incorporates detailed models of molecular evolution

Tree space

Starting Tree (T)

hill-climbing method



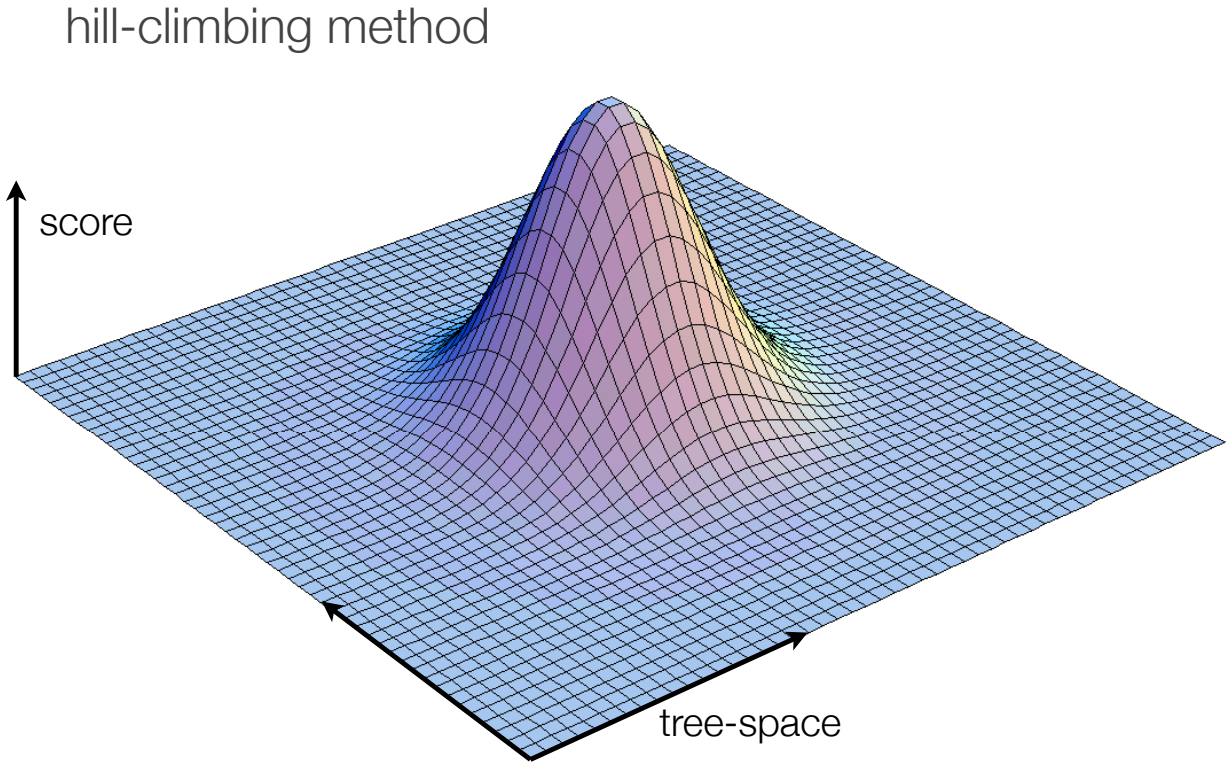
ML search

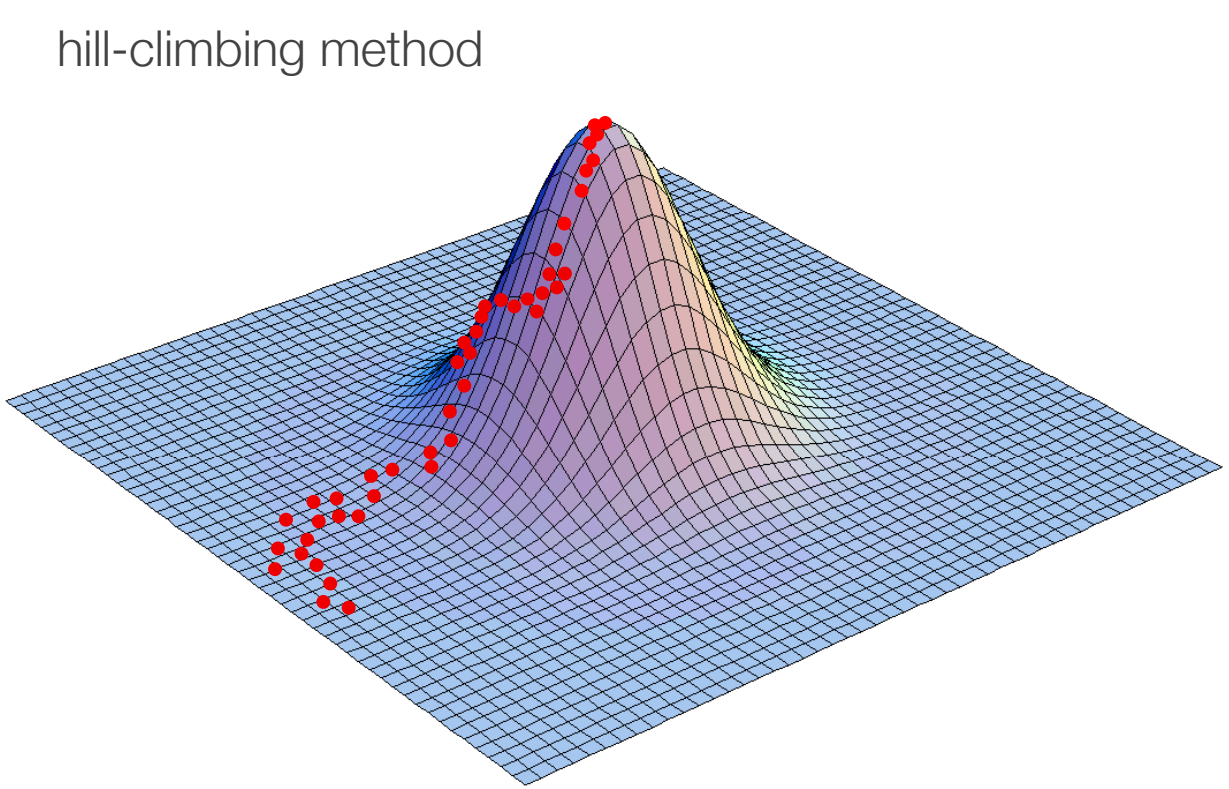
- Start with a tree (T)
- Perturb with branch-swapping (T’)
- Calculate likelihood of tree
- If T’ has better score than T, continue walk; if not, try different perturbations
- For each tree, need to optimize parameters of the model

Tree space

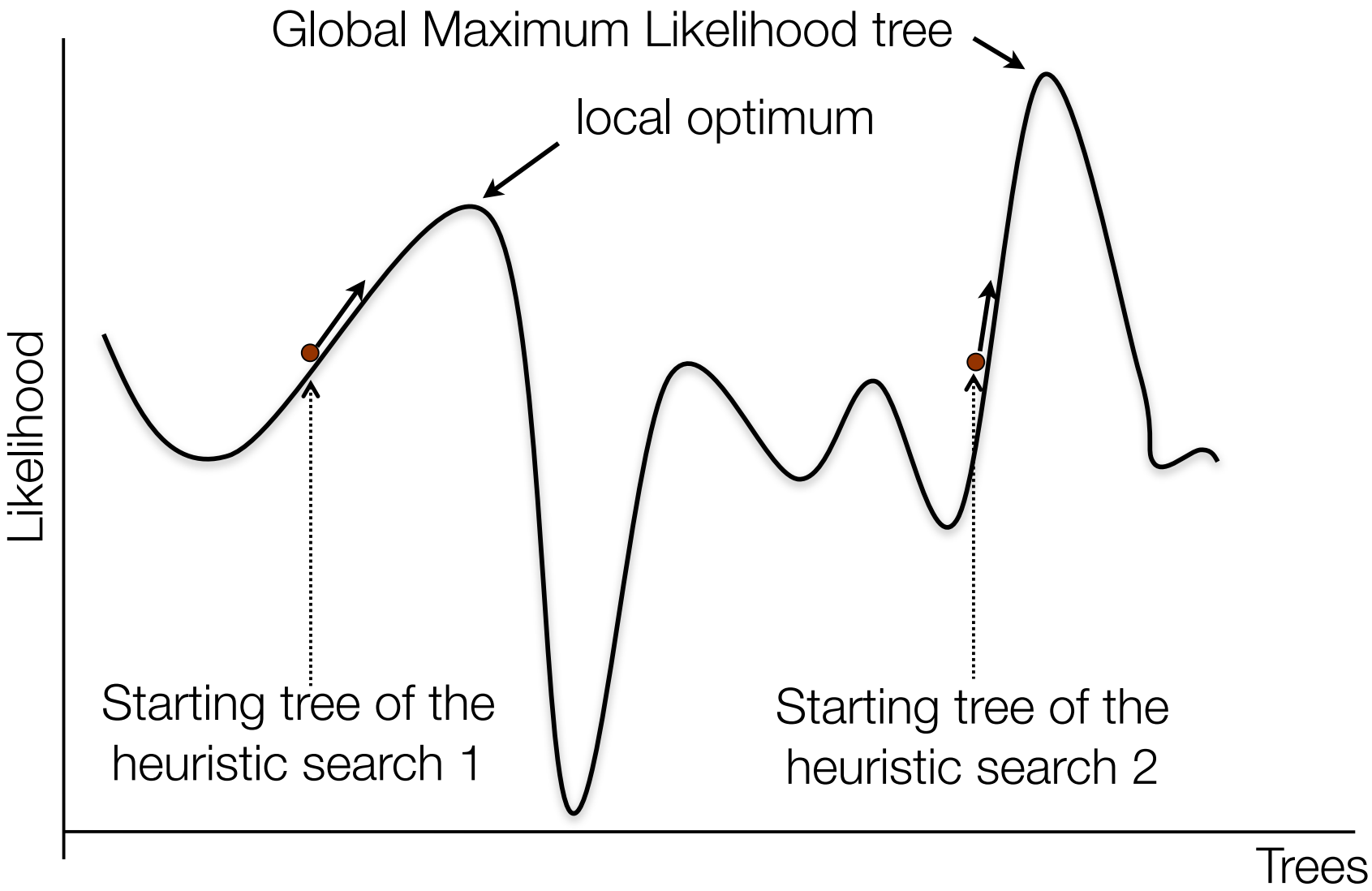
Starting Tree (T)

Random walk (T’)





Tree space


Global Maximum Likelihood tree
local optimum
Likelihood
Starting tree of the heuristic search 1
Starting tree of the heuristic search 2
Trees

Programs for inferring trees

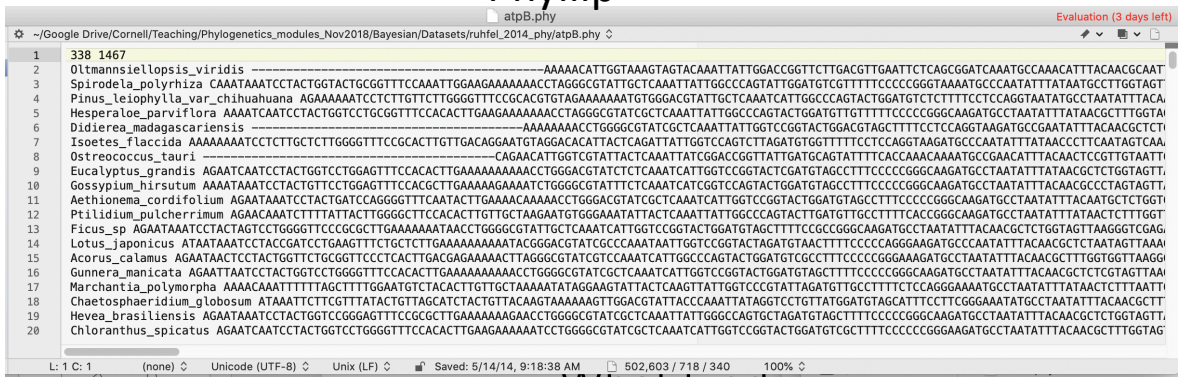
- Garli
  - Limited GUI but full functioning command line
  - Complete control of model selection
  - Great for small data sets <300 tips and one gene, but slow for many genes and tons of taxa
- RAxML
  - Great for large data sets; also have a RAXML NG for NGS data
  - Limitation in the number of models of molecular evolution
  - Command line, but there is a RAxML BlackBox
- Mega
  - GUI version, easy to learn
  - However not widely accepted in publications (not all reasons are supported)
  - May not work very well for large data sets, but fast on small data sets

What kind of data do we need?

- Alignment files in either fasta or phyllip format

Fasta

Phyllip

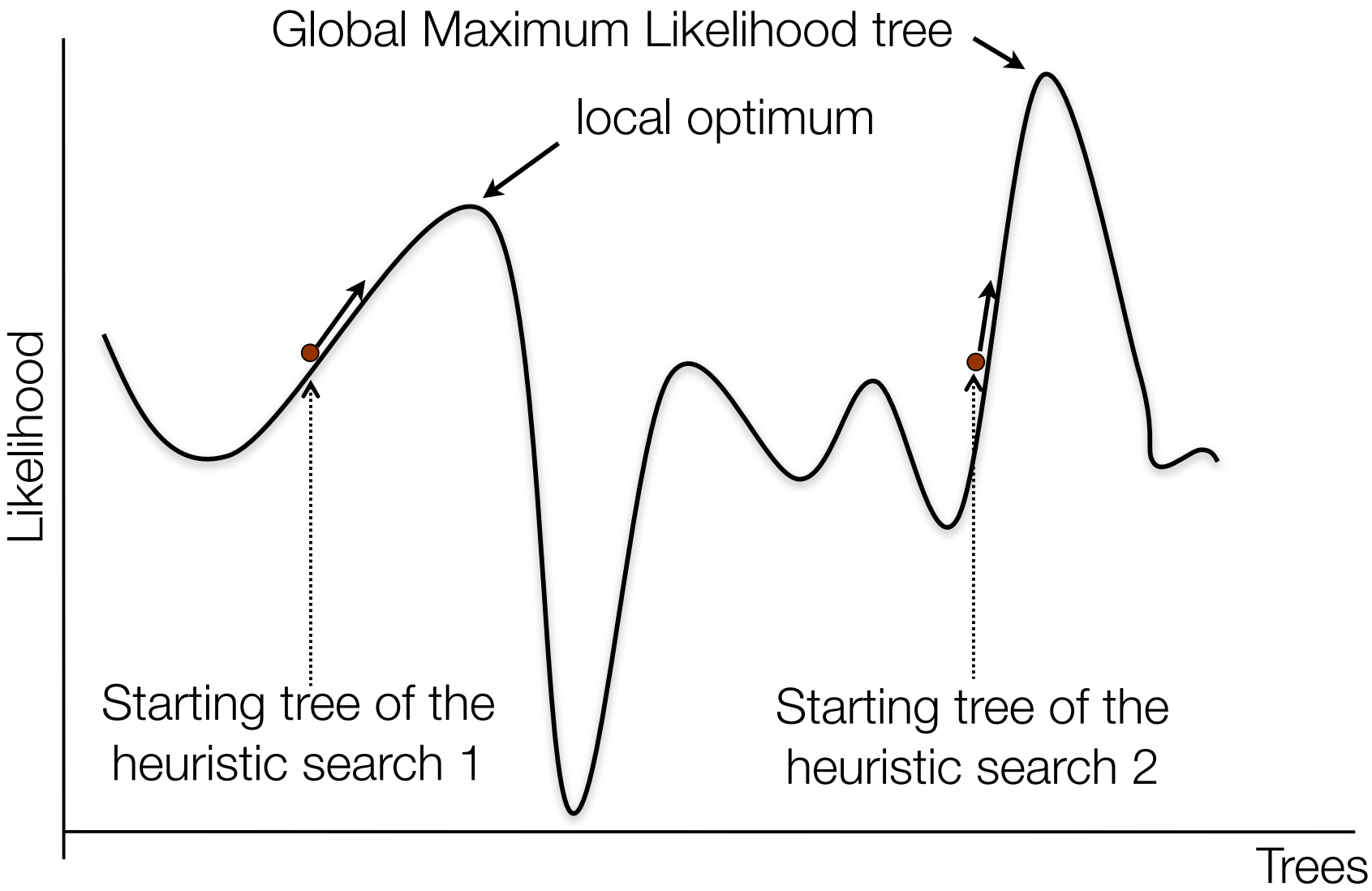


- Either nucleotide or amino acids; specify the appropriate model
- Missing data/gaps are treated as N's
- Should specify an outgroup, but will run without it
  - If you specify multiple, but they are not monophyletic, the first will be used

Once you hit go, what happens?

- RAxML generates starting trees using stepwise addition order parsimony
  - Taxa are inserted into the tree one at a time
- After all taxa added, subtree pruning re-grafting (SPR)
  - Similar to TBR as done in TNT
- Other programs us a Neighbor Joining distance tree as the starting tree
  - Parsimony subtrees seen as an advantage because each start is a distinct and different starting point, which helps search tree space
  - Why is this good?

Tree space

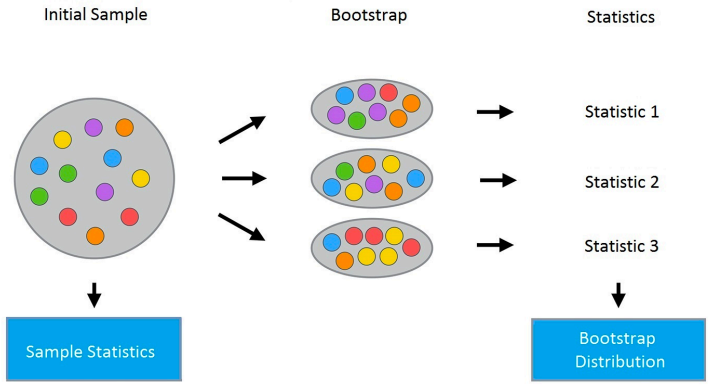

Global Maximum Likelihood tree
local optimum
Likelihood
Starting tree of the heuristic search 1
Starting tree of the heuristic search 2
Trees

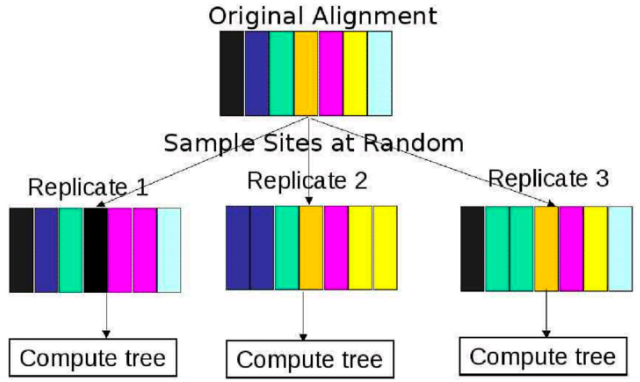
Scope of the question and data

- RAxML was designed for large data sets
  - Reason the models of molecular evolution are constrained to GTR
- For broad taxonomic questions
  - chloroplast genes (slower evolving the nuclear)
  - Amino acid/protein alignments – avoid the third position wobble
- Questions below the species level
  - Can use SNP data from RAD-Seq or Genome Resequencing
  - Models of molecular evolution are different, no invariant sites

How do we tell support for a topology?

- Most analyses use bootstrap support

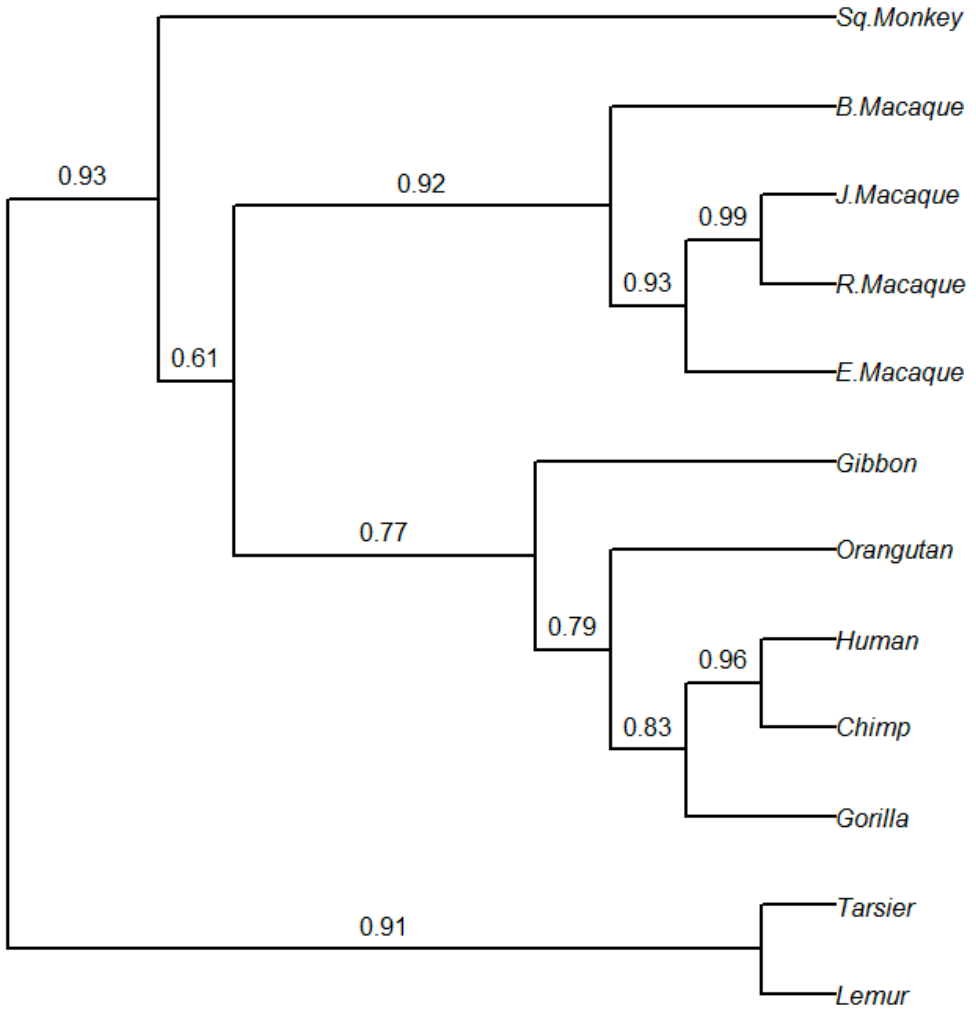




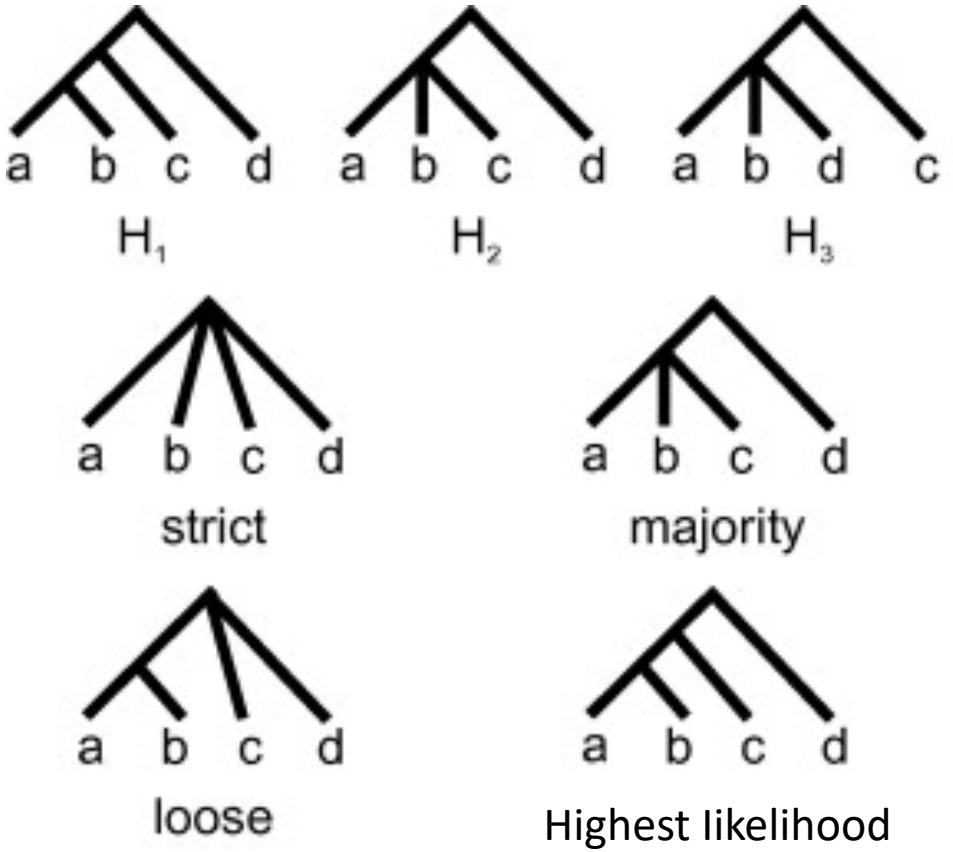
- In our case, we will sample with replacement the original alignment to generate 1000 alignments of the same length
- How often are the same taxa related to each other out of the 1000 replicates
- Usually a value of BS > 70 or more is considered strong support

What do we consider good support?

- If there is strong support for relationships in the original data, it will come through in the bootstrap support



Support for topology



Bootstrap trees

- If topology is different between analyses, need to look at how well-supported relationships are
- The ”best” tree may not be well-supported at
- For small data sets this is easy to test by eye, for larger ones we will use R

So you have a tree, are you done?

……

So you have a tree, are you done?

……

I would say you are just getting started

Comparative phylogenetic biology

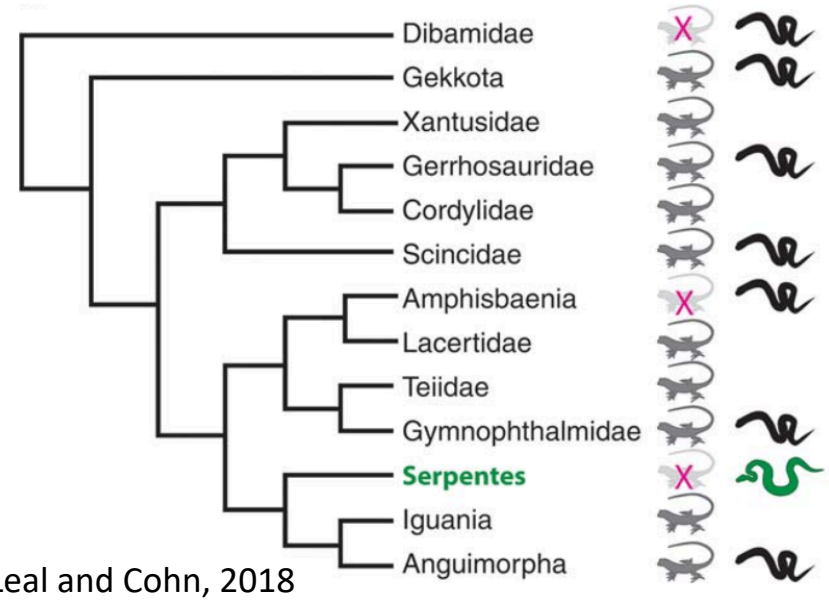
- “Nothing in biology makes sense except in the light of evolution” – Theodosius Dobzhansky, 1973

- We use comparative data of species and a phylogeny to make inferences about evolutionary process and history

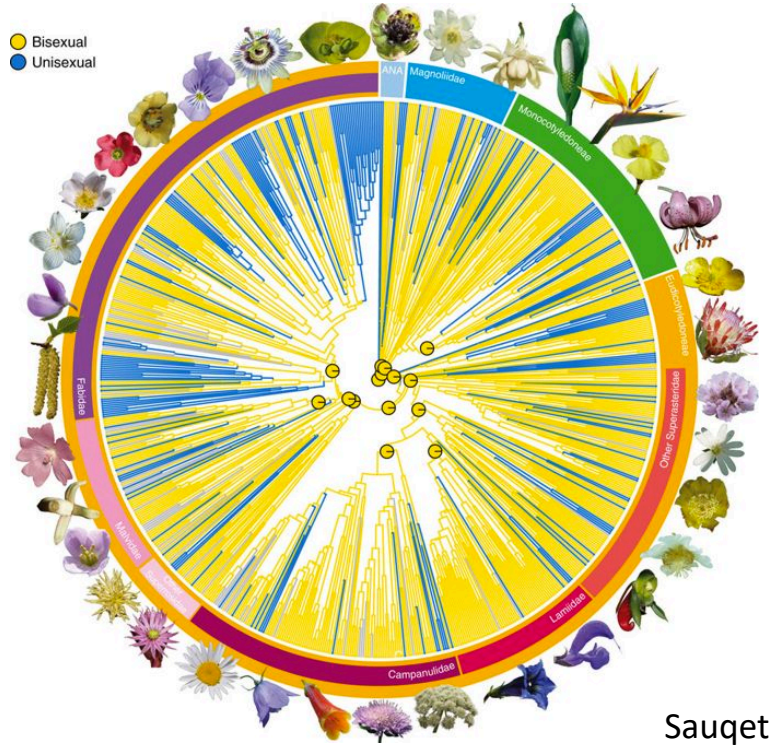
- Reconstructing the ancestral phenotypes of extinct hypothetical ancestral species is a major goal of phylogenetic comparative analyses

Why would do we care about ancestral states?

- If we study reptiles, may be we are interested in how many times limbs evolved.
- Or probably more along this audience, what the ancestral flower looked like.



Leal and Cohn, 2018



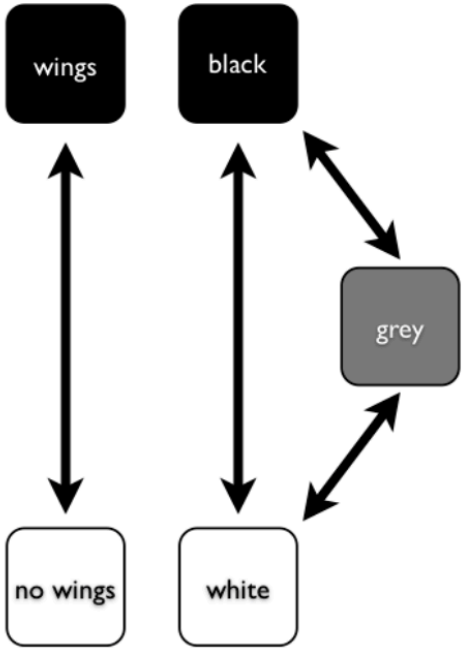
Sauqet et al., 2017

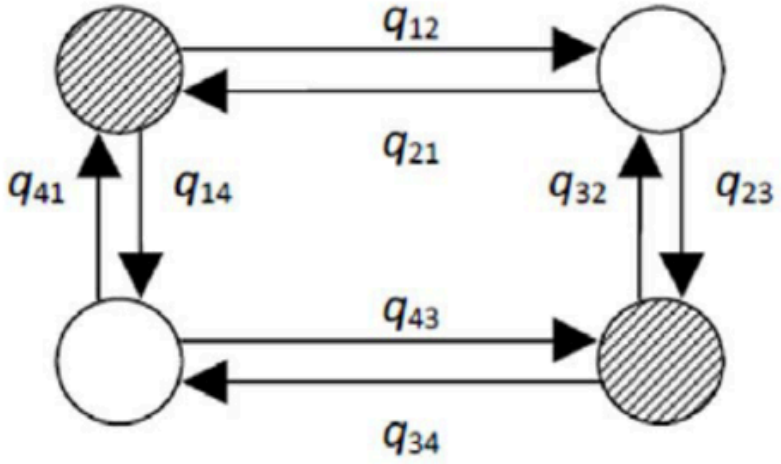
Ancestral state reconstructions

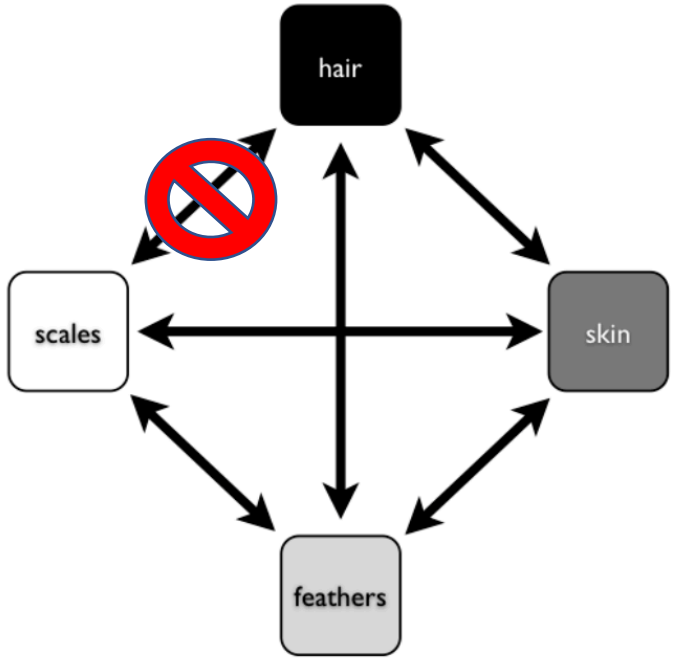
- The first step is to identify the type of data we are analyzing
- Do we have continuous or discrete characters?
- The distinction between this is not always straightforward
  - For instance: days to flowering, bristle number on *Drosophila*, number of scales along the midline, etc.
- What may be the best model of evolution?

Models of discrete change

- Equal Rates
- Symmetrical Rates
- All Rates Different
- Custom







Discrete ancestral state reconstructions

- Mk model most often used
  - M stands for Markov – modeling process is a continuous-time Markov chain
  - k number of states

- Central attribute is the Q matrix, or transition matrix

$$\mathbf{Q} = \begin{bmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{bmatrix}$$

R

- We will be doing some of the final steps in R
- For new grad students, I personally think it will be hard to graduate without knowing at least the very basics in R
- R for biologists
  - https://www.rforbiologists.org/

Specifically what are we doing today?

- Test for the best model of molecular evolution
- Run RAxML on the small data set together
- Use the R script to test which clades are shared
- Use the R script to simulate trait data and perform ancestral state reconstruction

- Then you will use your gene from last time to redo the analyses
- Next Tuesday we will go into Bayesian phylogenetics

If you want to know more about any of this

- Feel free to come talk to me, I'm in 510 Mann Library; e-mail is jbl256@cornell.edu

- Luke Harmon's Phylogenetic Comparative Methods: Learning from Trees is a great resource