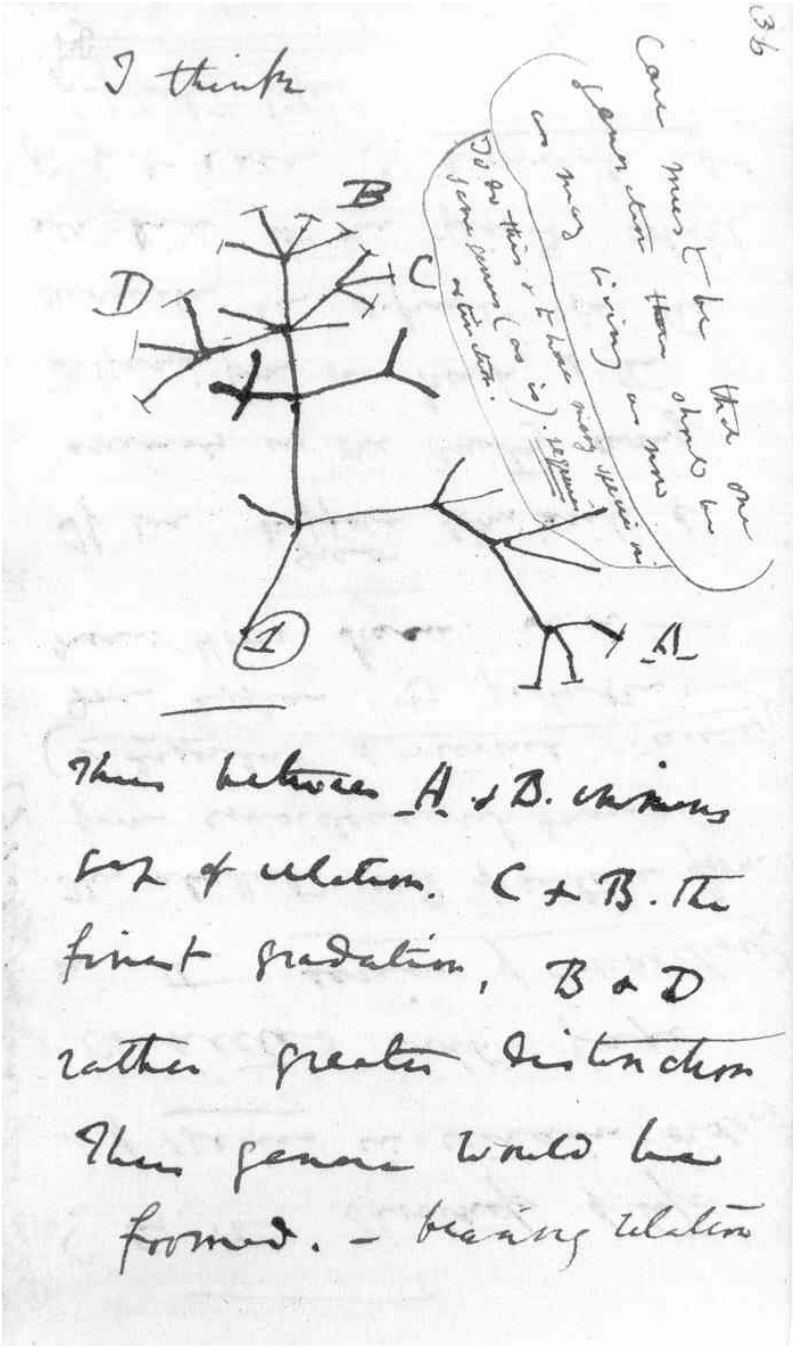
Phylogenetics

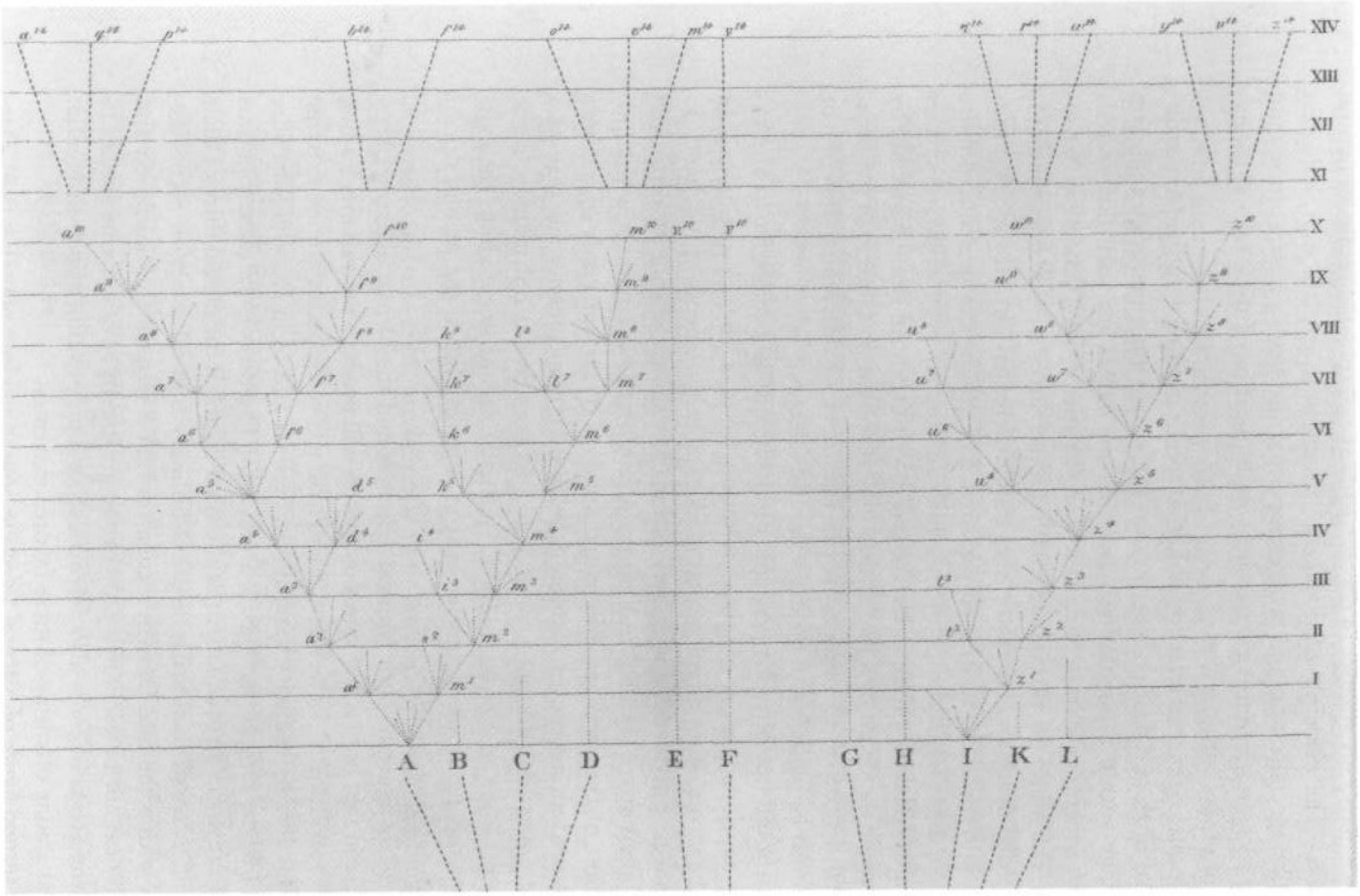
April 2023

Gustav Klimt 1905 “Tree of Life”

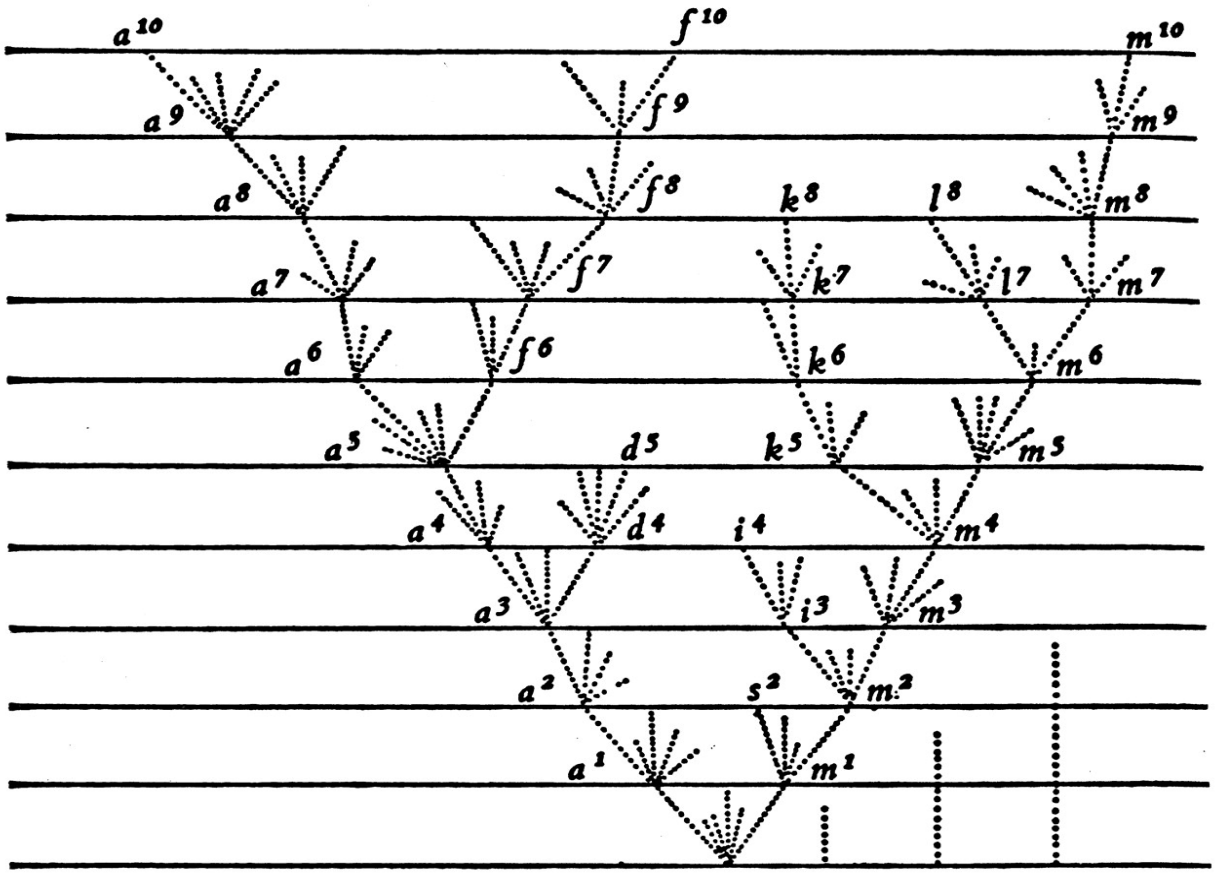


Notebook around July 1837

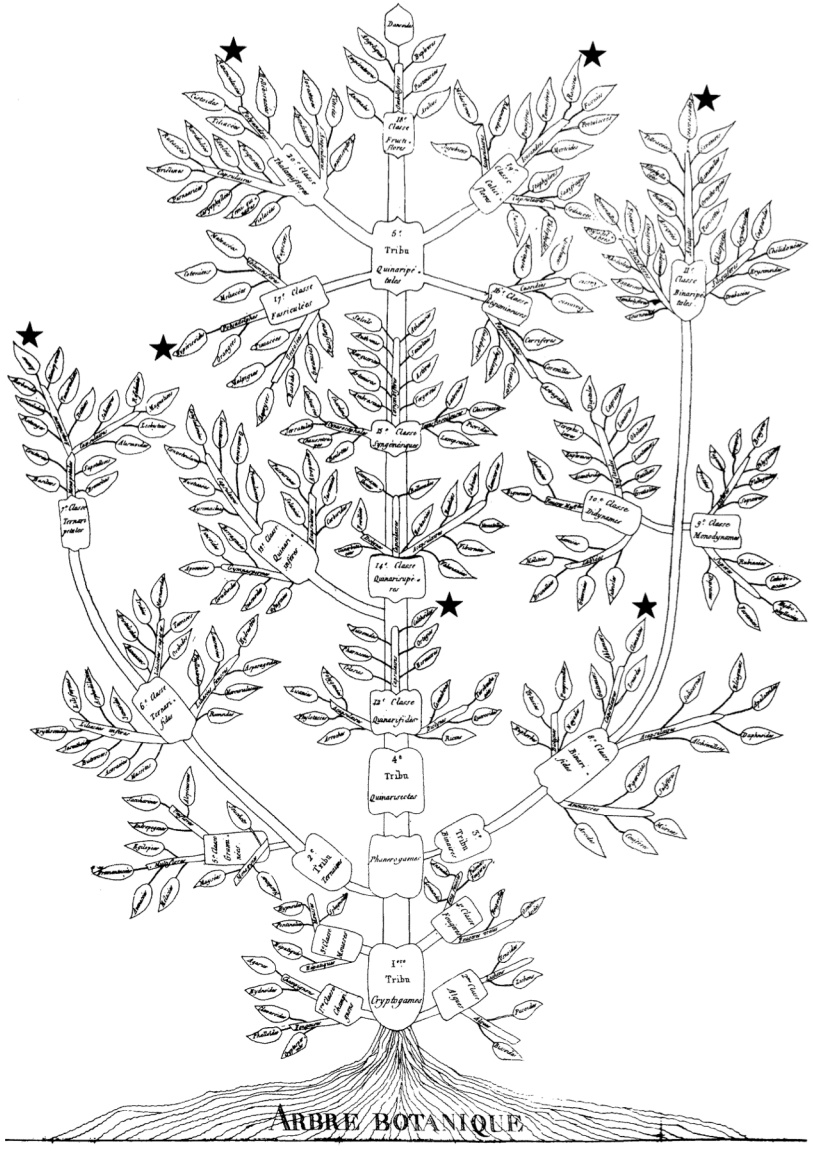

B
C
D
A
C. Darwin 1837


XIV
XIII
XII
XI
X
IX
VIII
VII
VI
V
IV
III
II
I
A
B
C
D
E
F
G
H
I
K
L

Detail from only figure in origin of species



Darwin 1859



Augustin Augier 1801

TABLEAU

Servant à montrer l'origine des différens animaux.

Vers.

Infusoires.
Polypes.
Radiaires.

Insectes.
Arachnides.
Crustacés.

Annelides.
Cirrhipèdes.
Mollusques.

Poissons.
Reptiles.

Oiseaux.

Monotrèmes.

M. Amphibies.

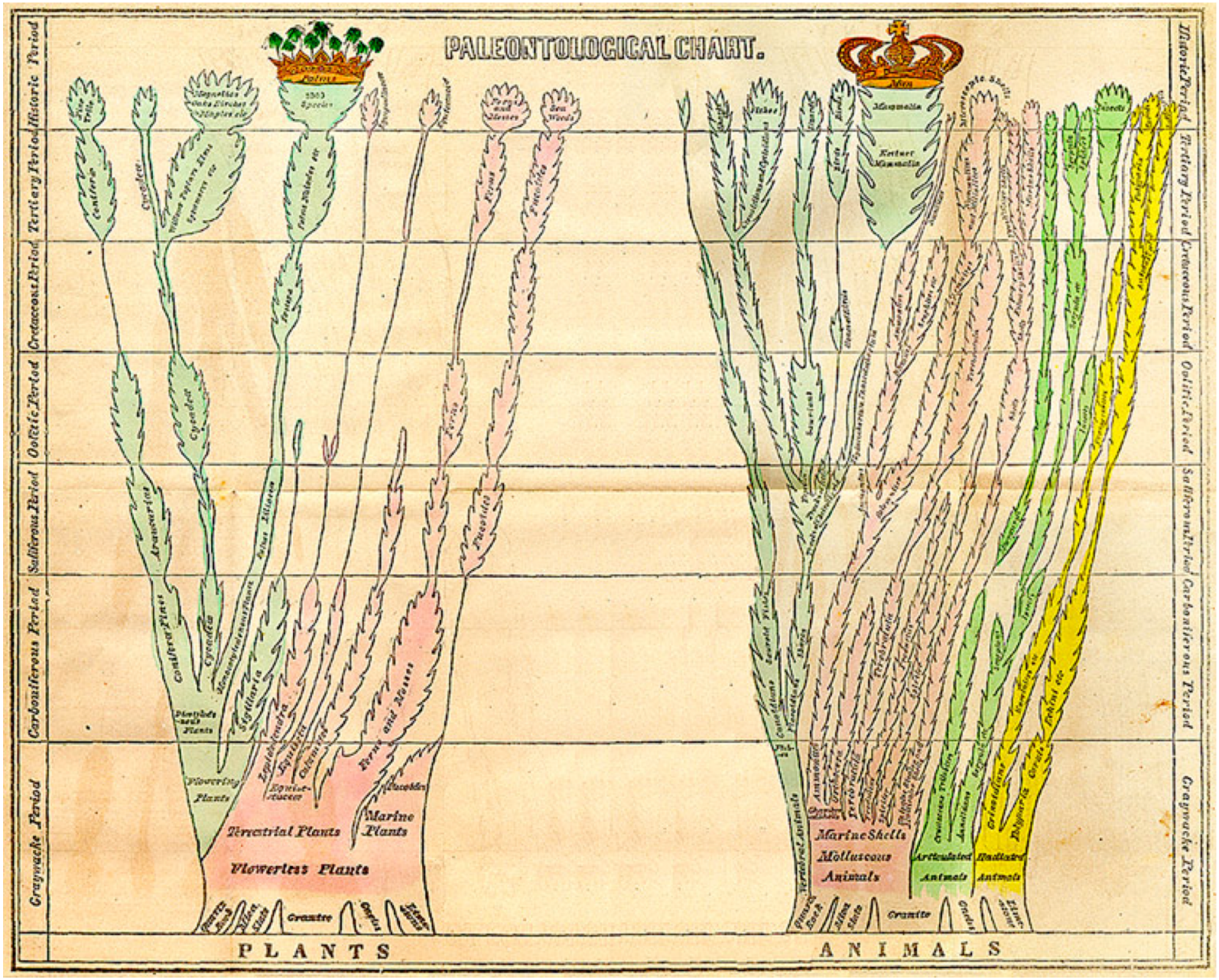
M. Cétacés.

M. Ongulés.

M. Onguiculés.

Cette série d'animaux commençant par deux

Jean-Baptiste Lamark 1809



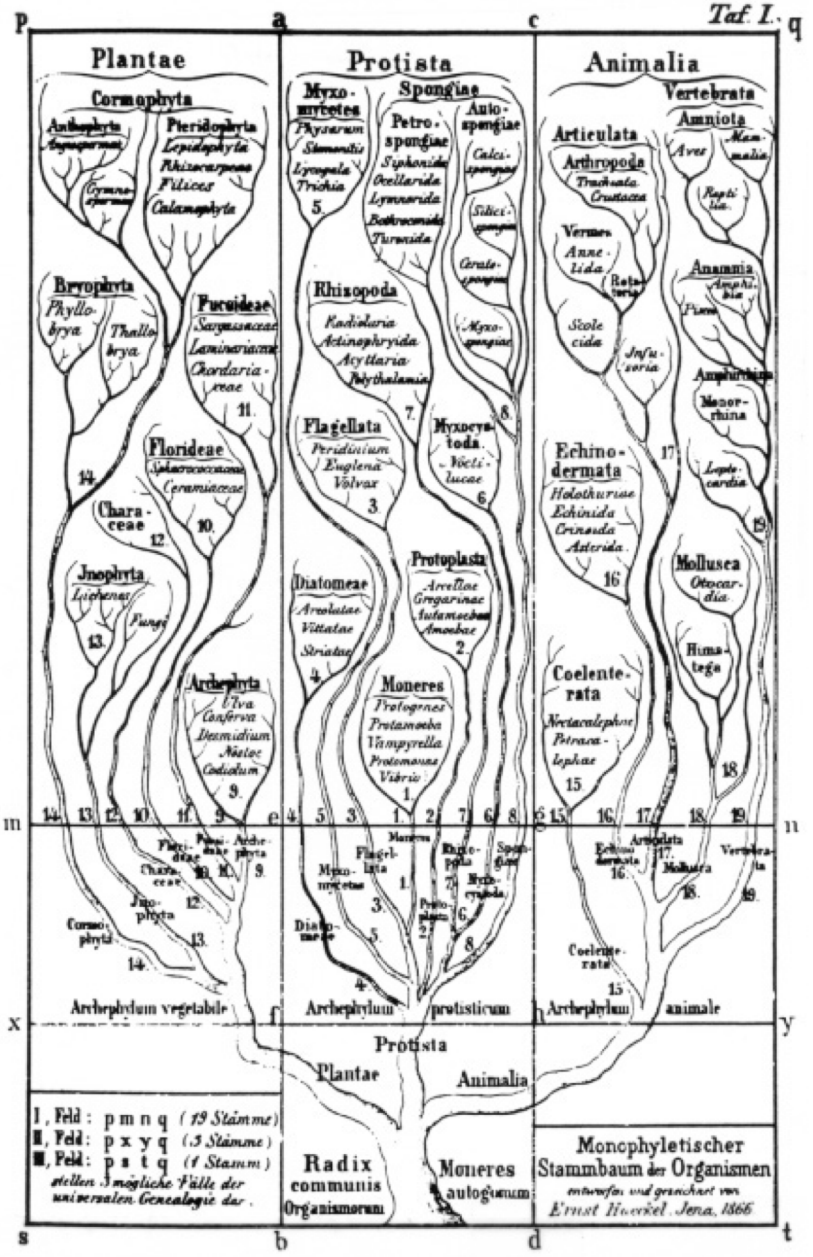
Elementary Geology Edward Hitchcock (1840)

Ernst Haeckel MD.PhD. (1834-1919)

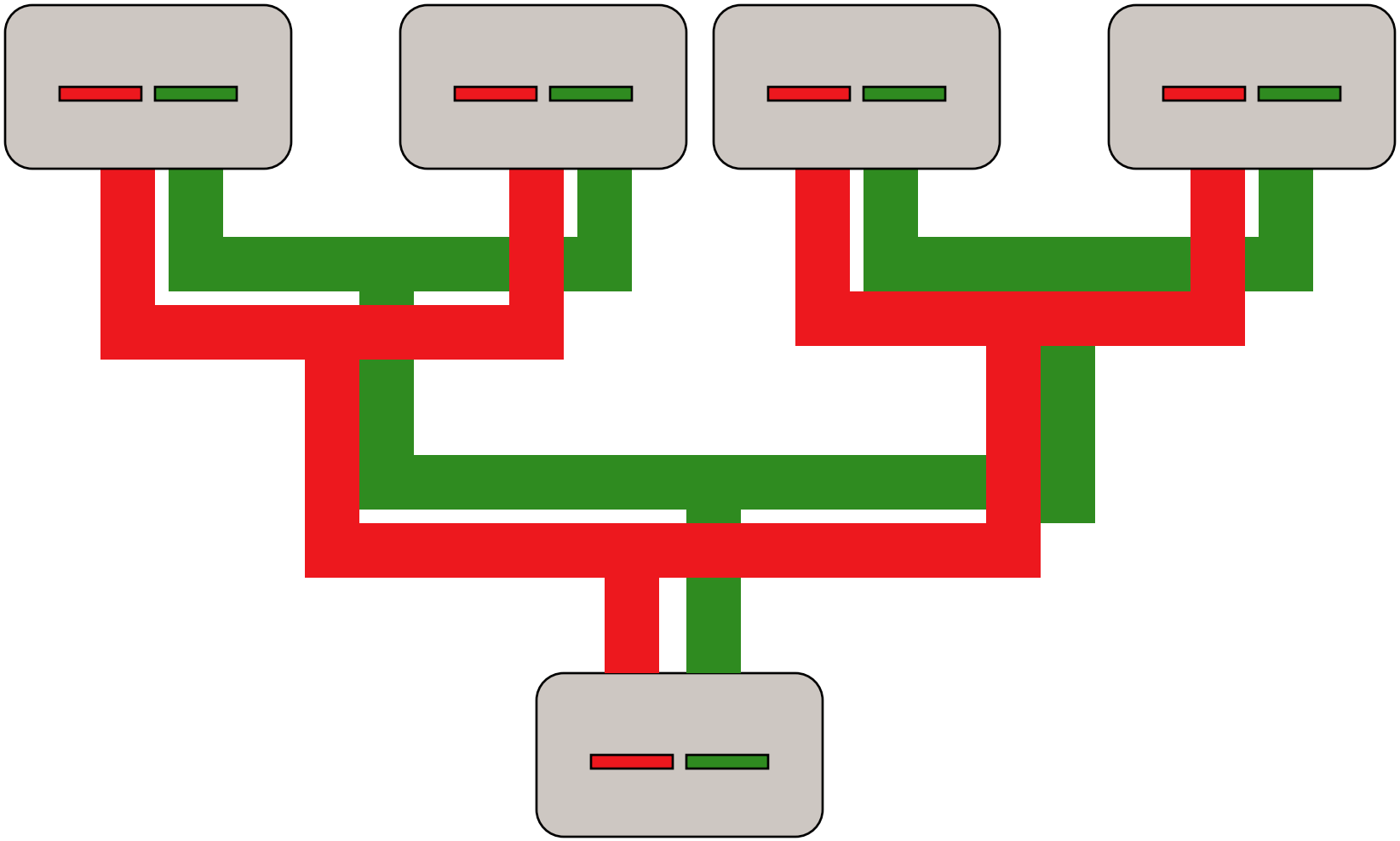
- “ontogeny recapitulates phylogeny”

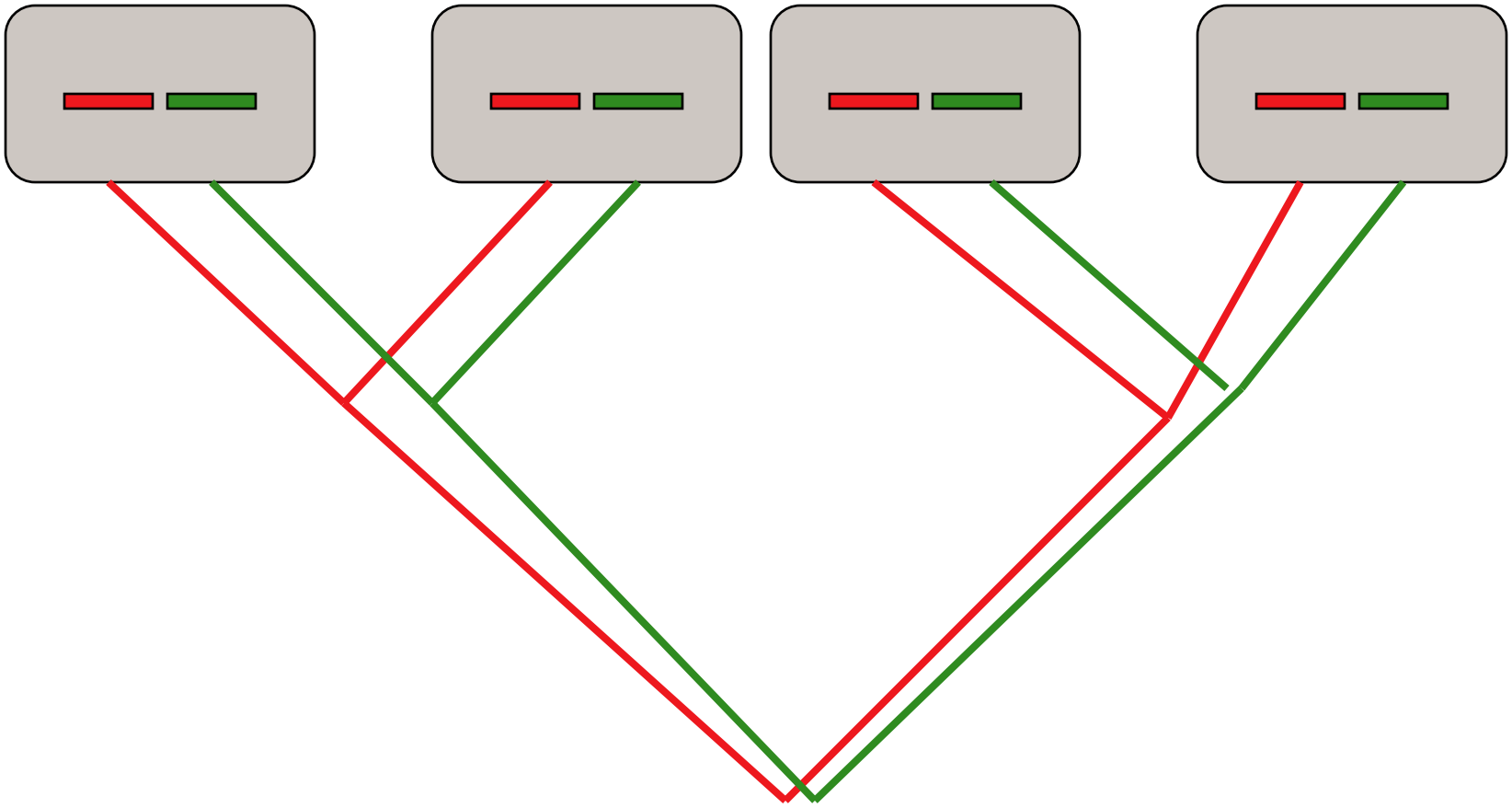
-coined words “protista”, “phylum”, “ecology”, “phylogeny”

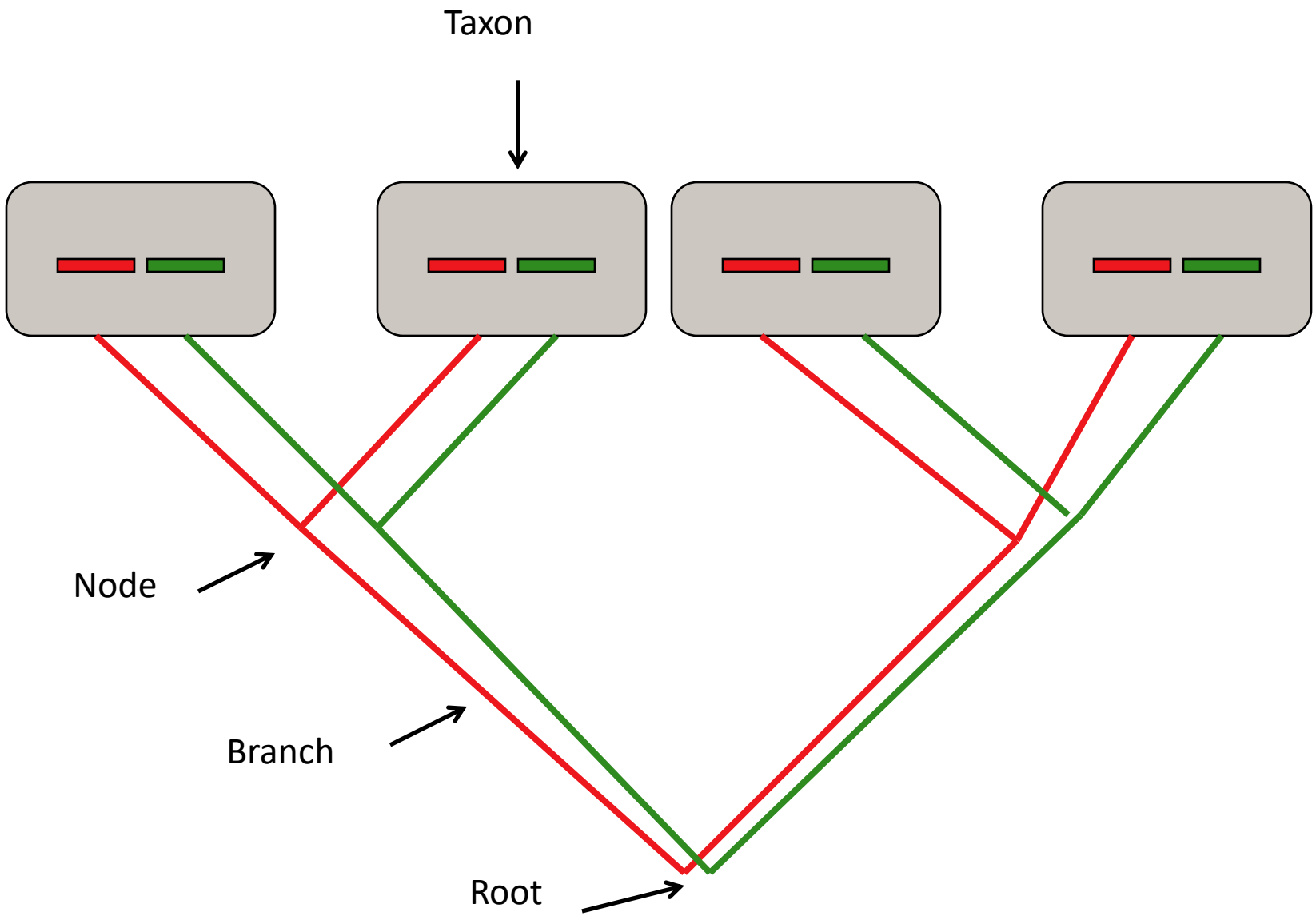
-TREE OF LIFE

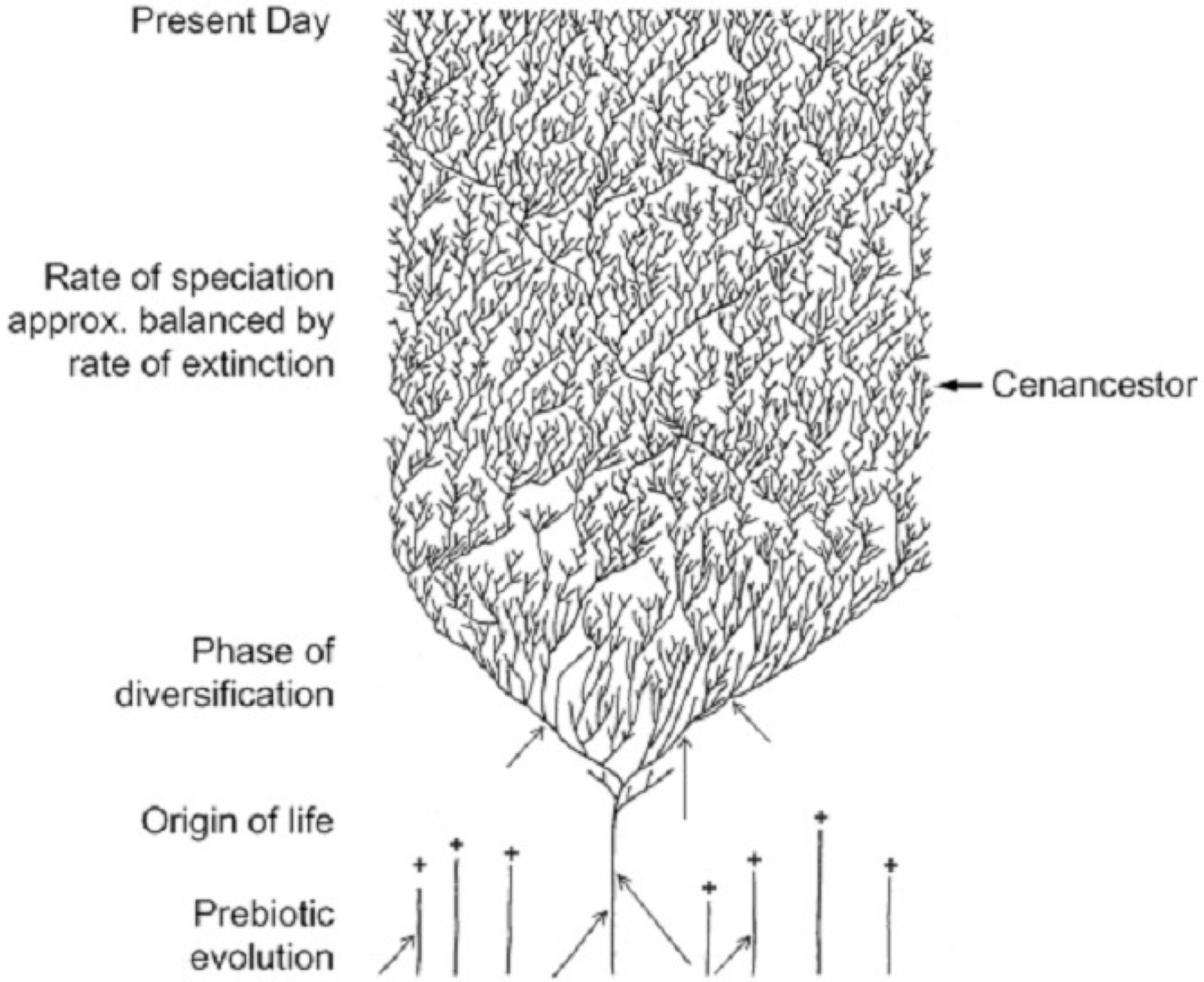


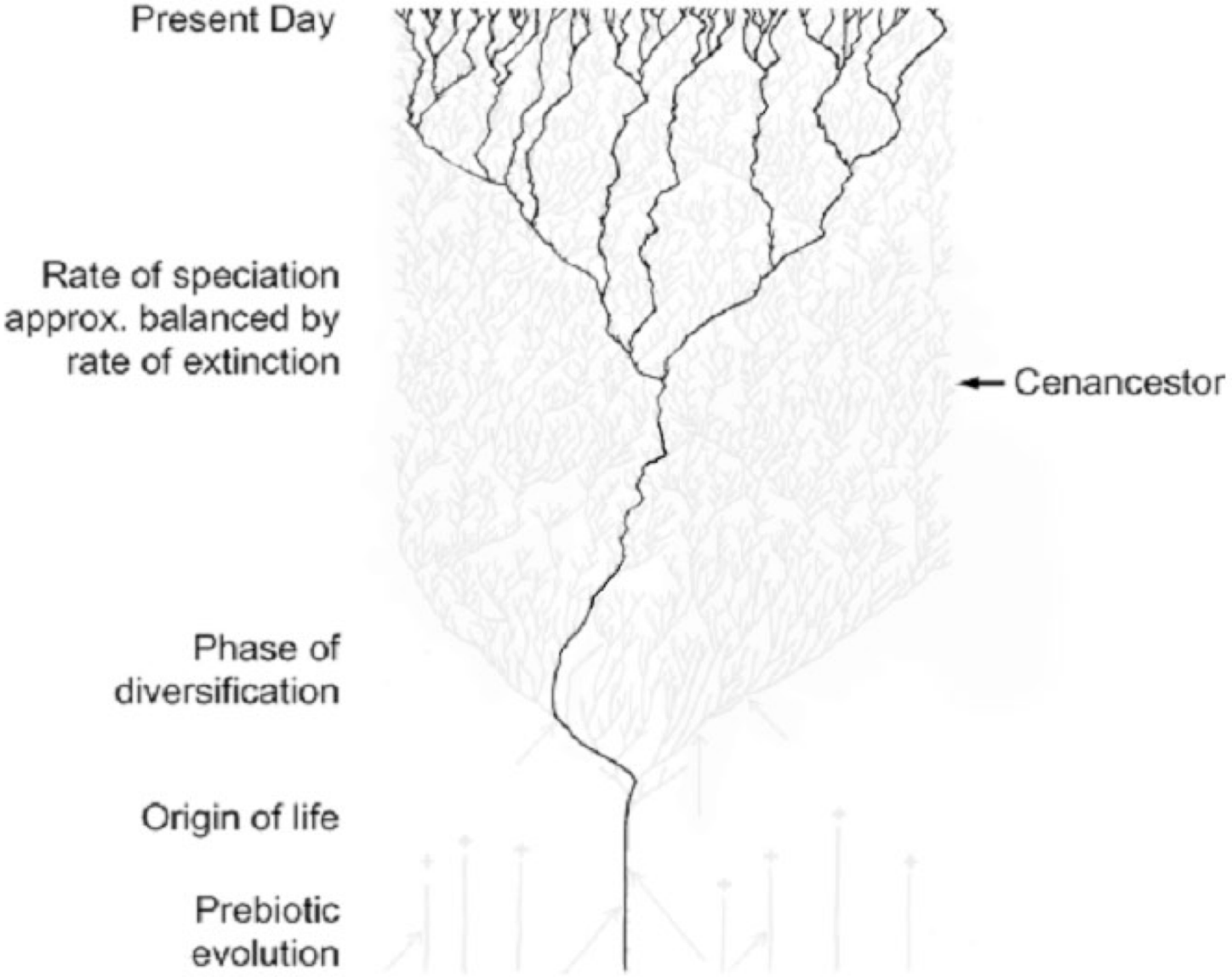
The Evolution of Man (1879)



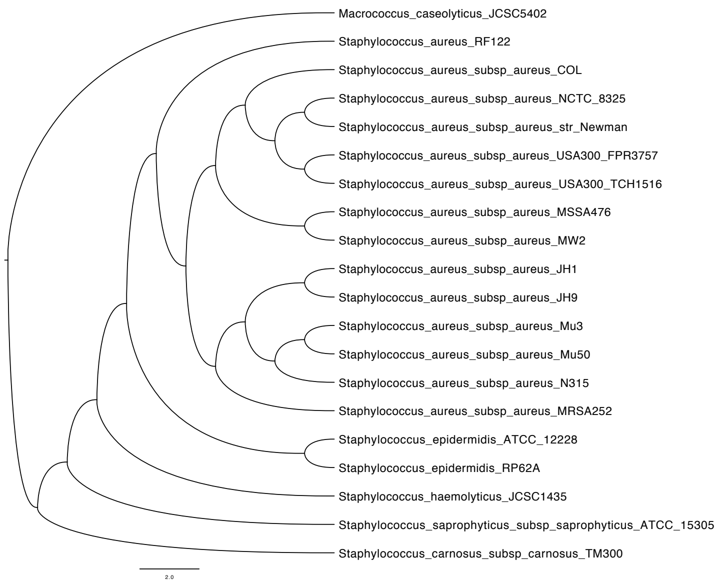


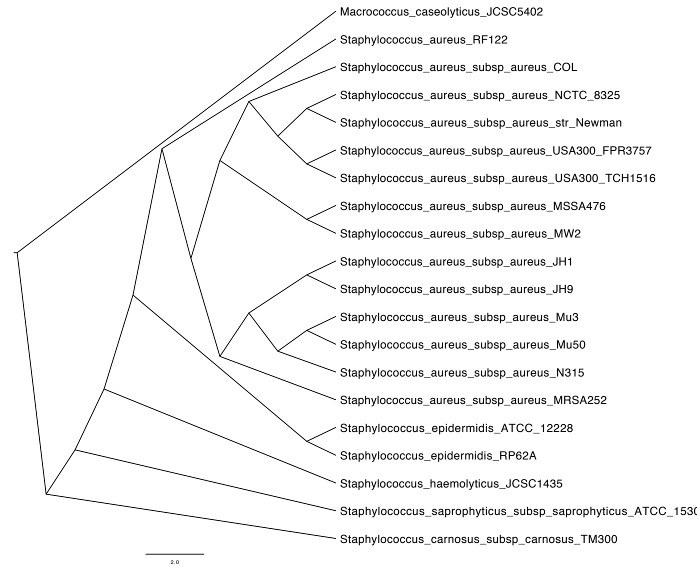

Taxon
Node
Branch
Root

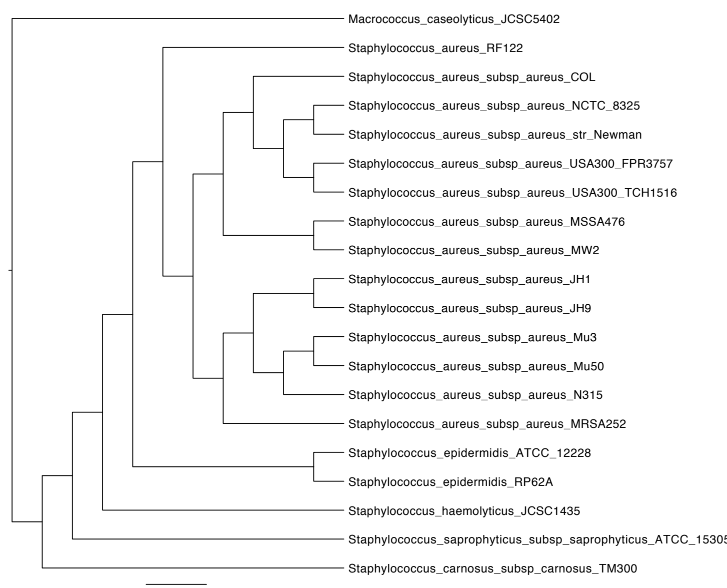

Present Day
Rate of speciation approx. balanced by rate of extinction
Cenancestor
Phase of diversification
Origin of life
Prebiotic evolution


Present Day
Rate of speciation approx. balanced by rate of extinction
Cenancestor
Phase of diversification
Origin of life
Prebiotic evolution

Look Different but Same Topologies



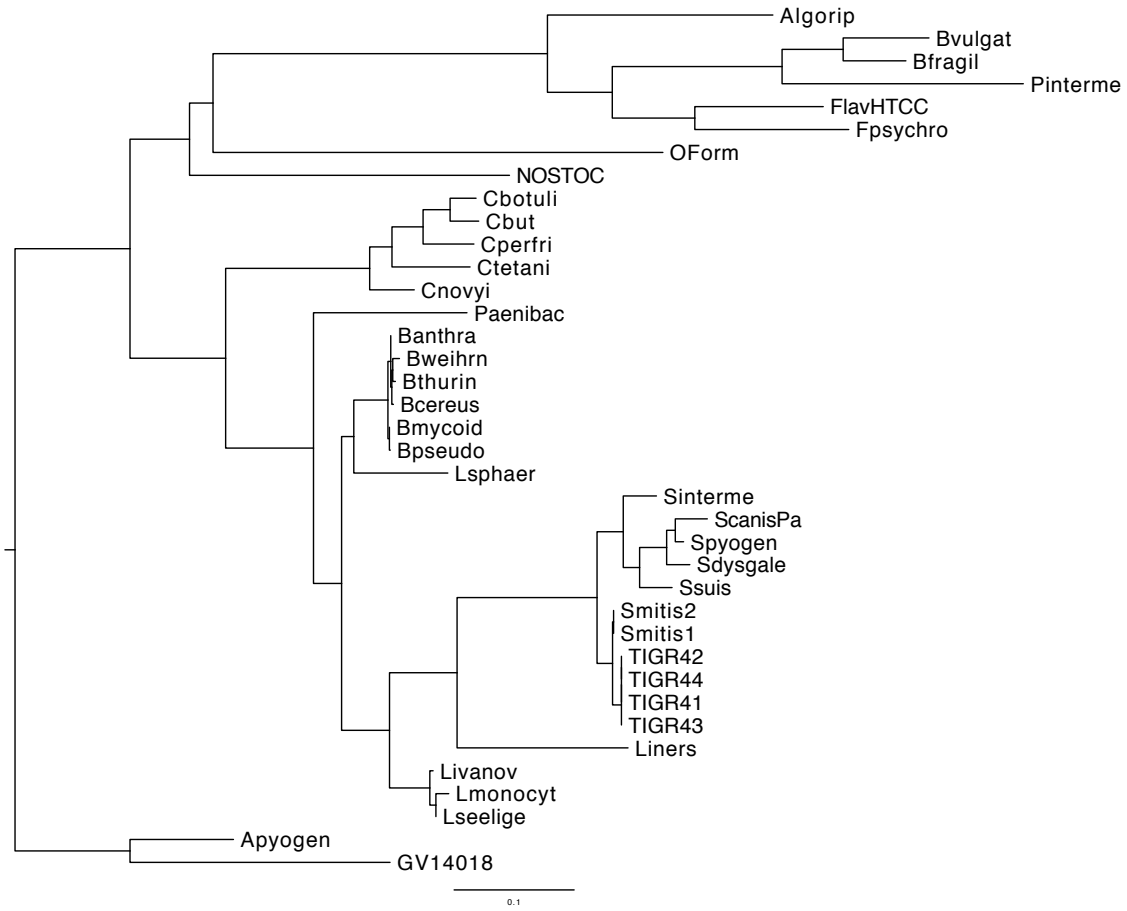


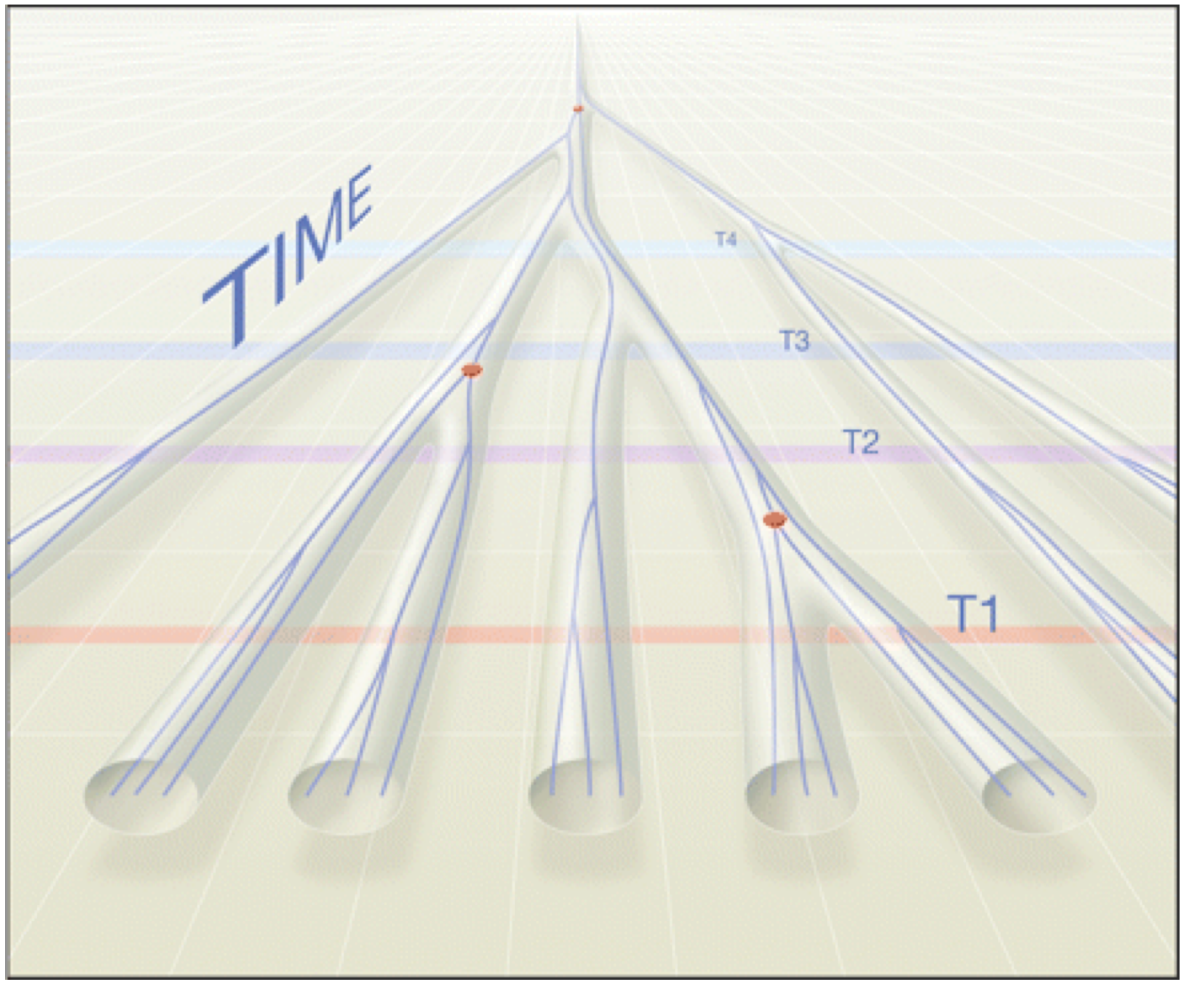


Time

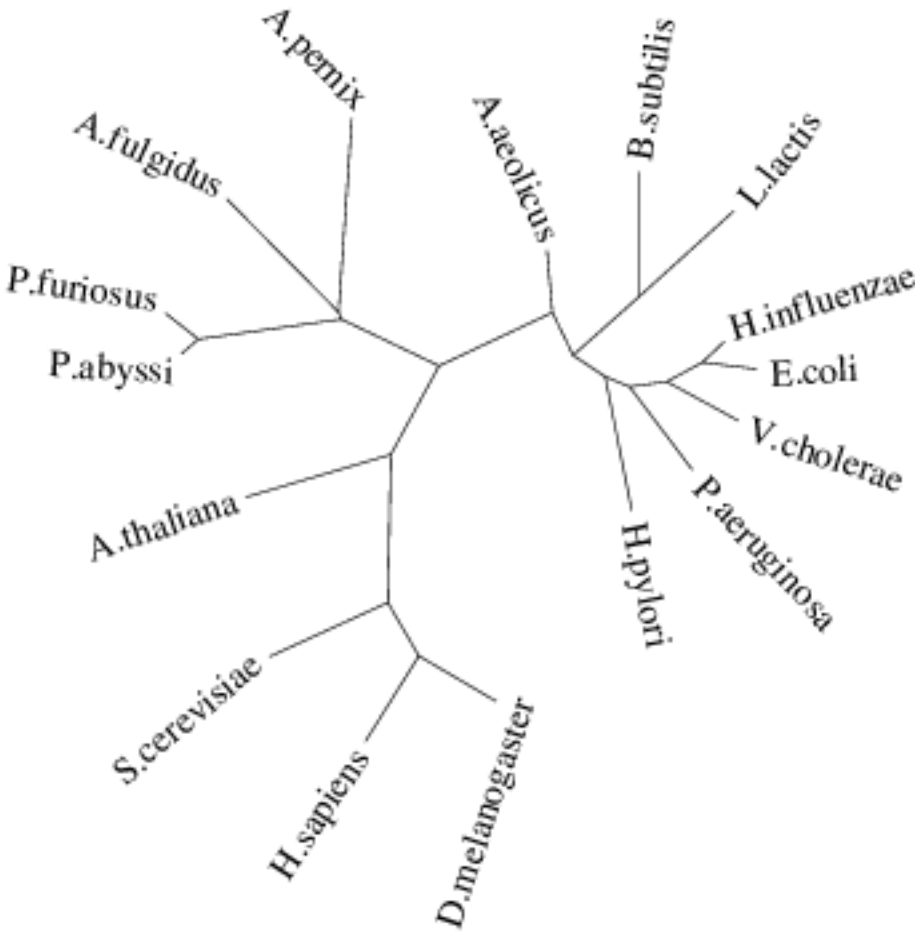
What about branch length:

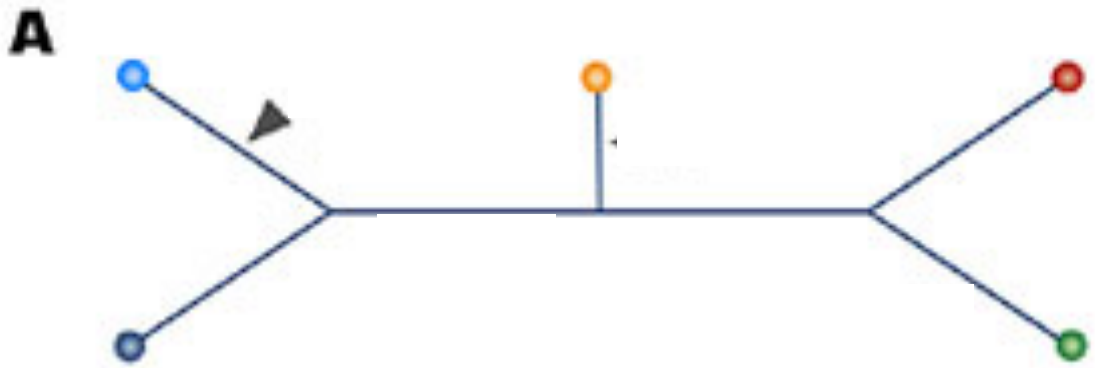
1. Branch length can signify many different things (time, difference, inferred changes)
2. General rule: longer branch length indicates **more evolutionary change**

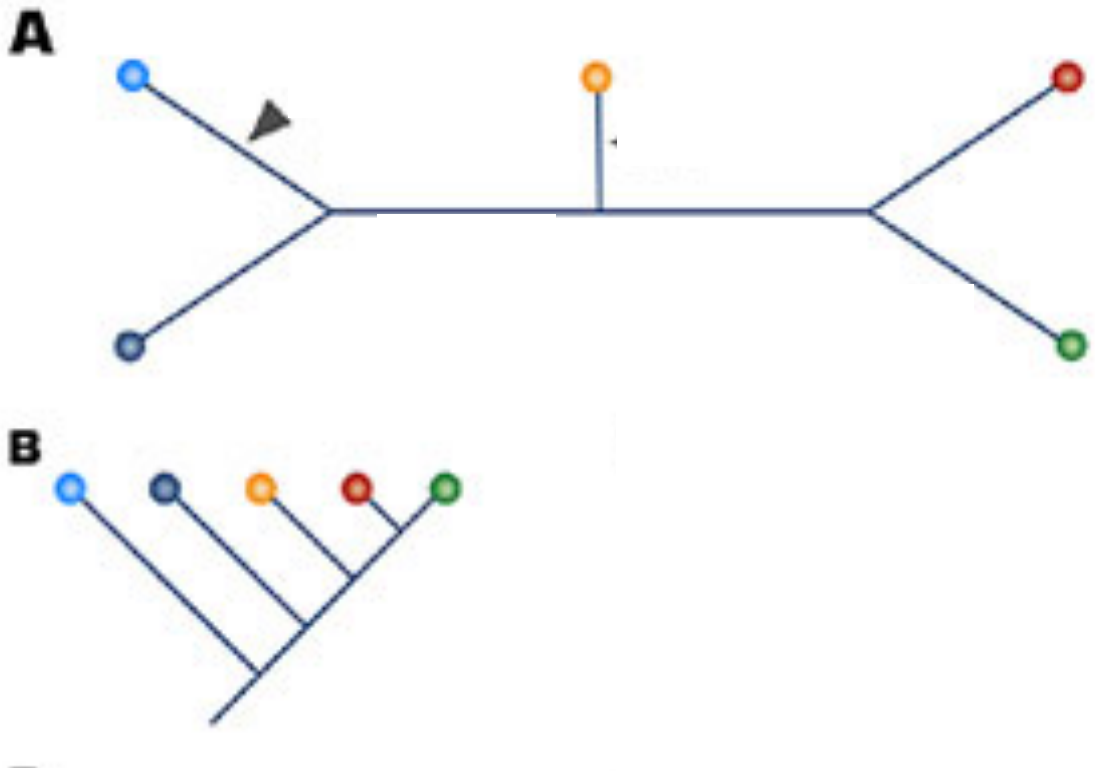


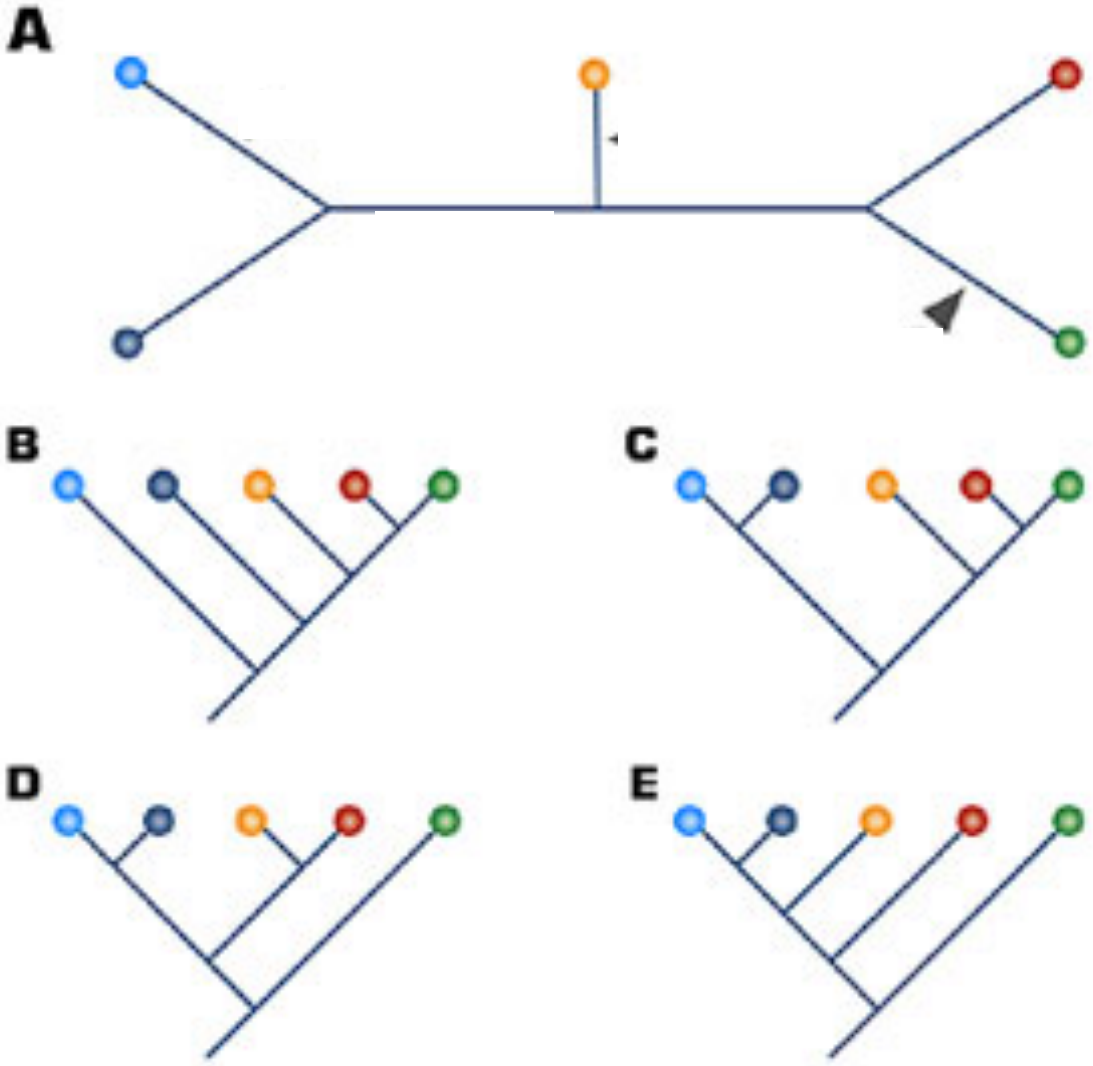

TIME
T4
T3
T2
T1

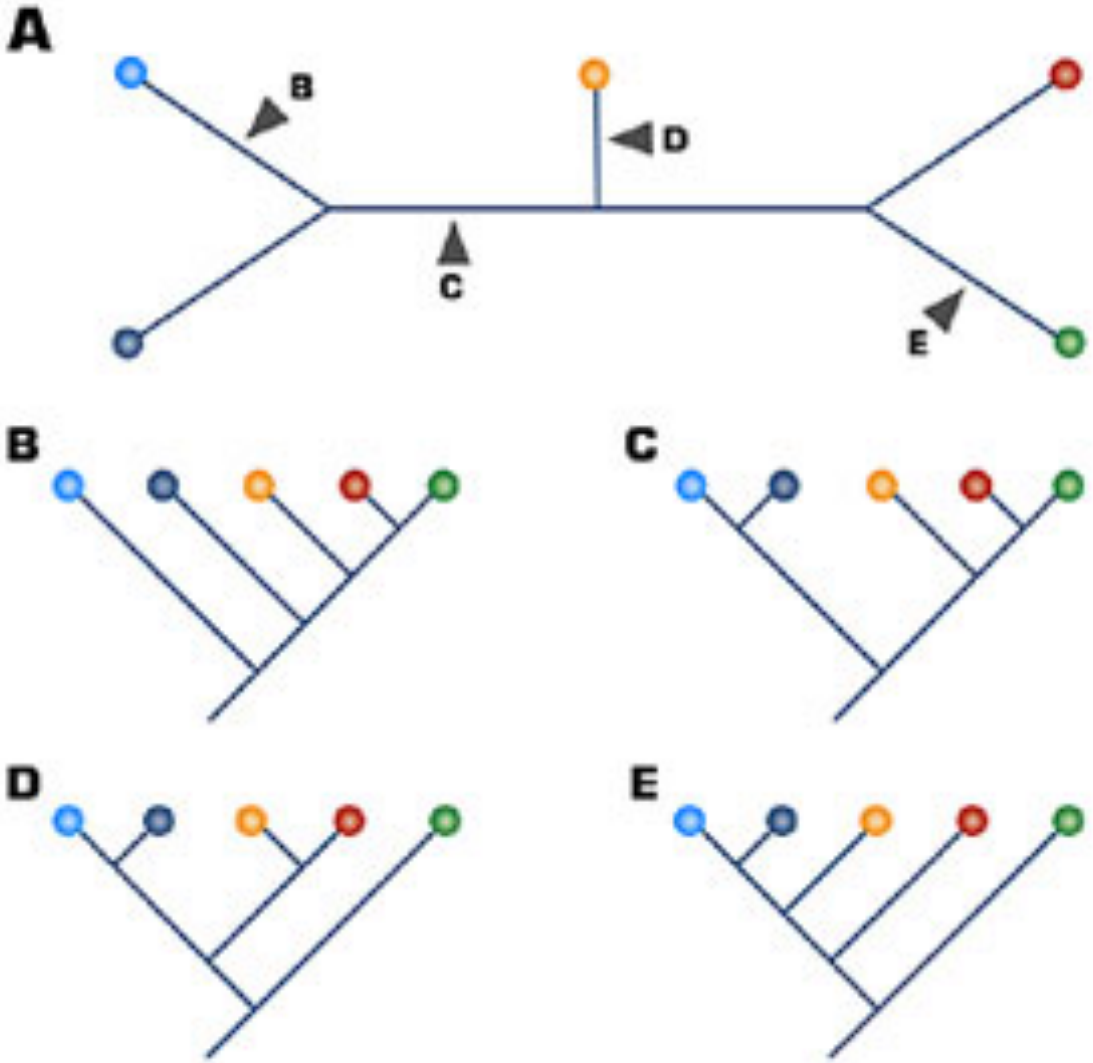
UNROOTED TREES




A


A
B

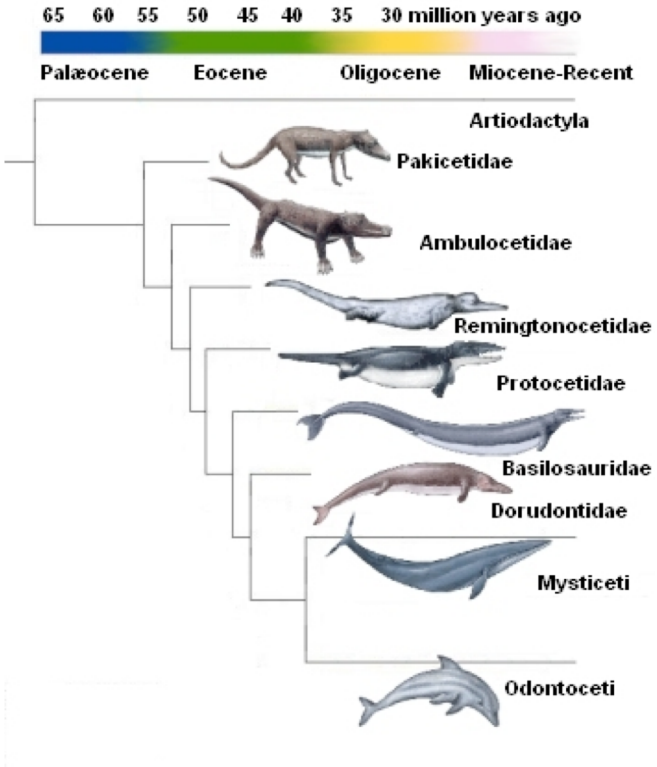

A
B
C
D
E


A
B
C
D
E
B
C
D
E

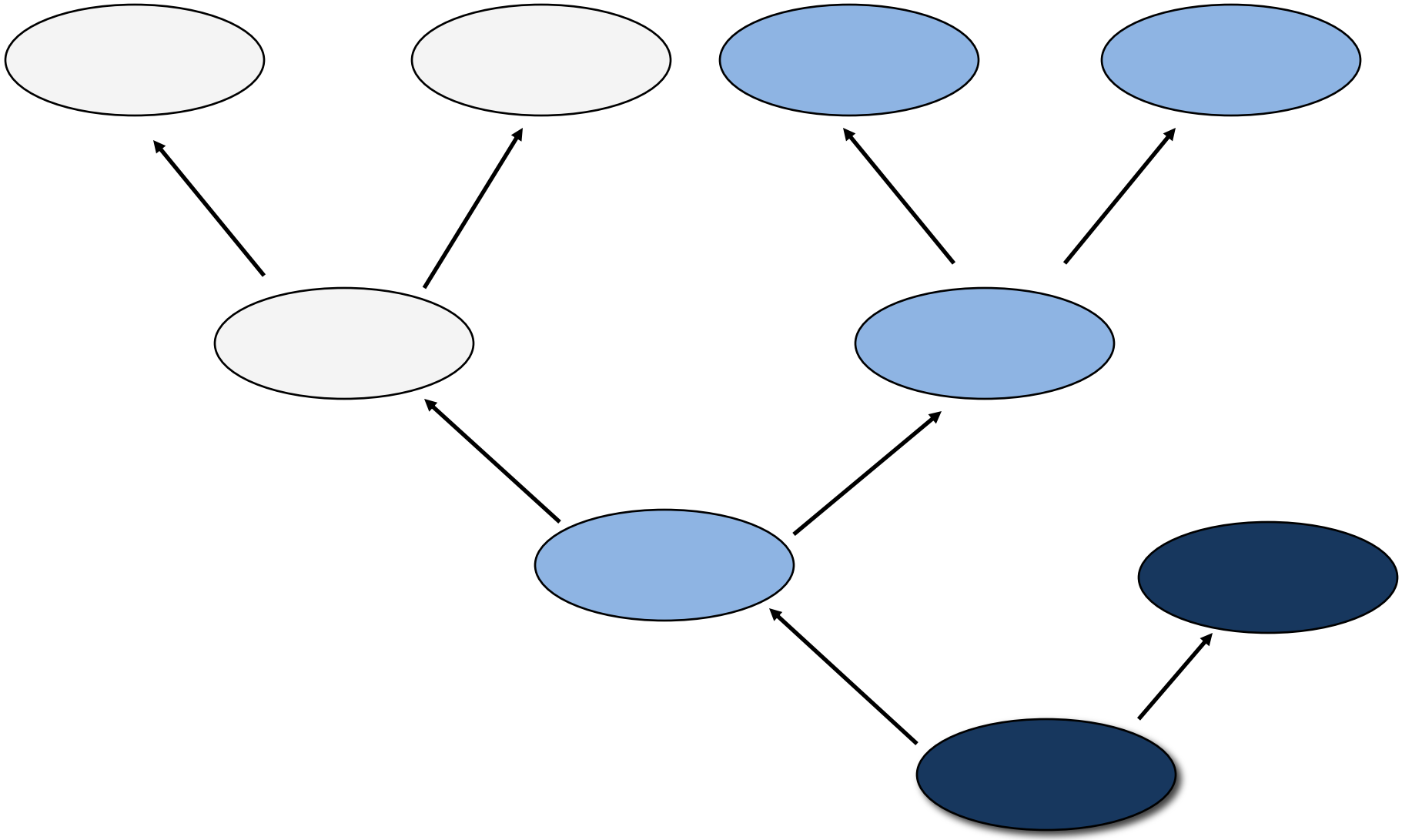
How do you decide how to root a phylogeny?

1. Use an **outgroup**:
    a. The best outgroup is a taxon that is a close relative that you are pretty sure is not in the ingroup.
    b. The more outgroups the better.
    c. The branch of the outgroup becomes the root.

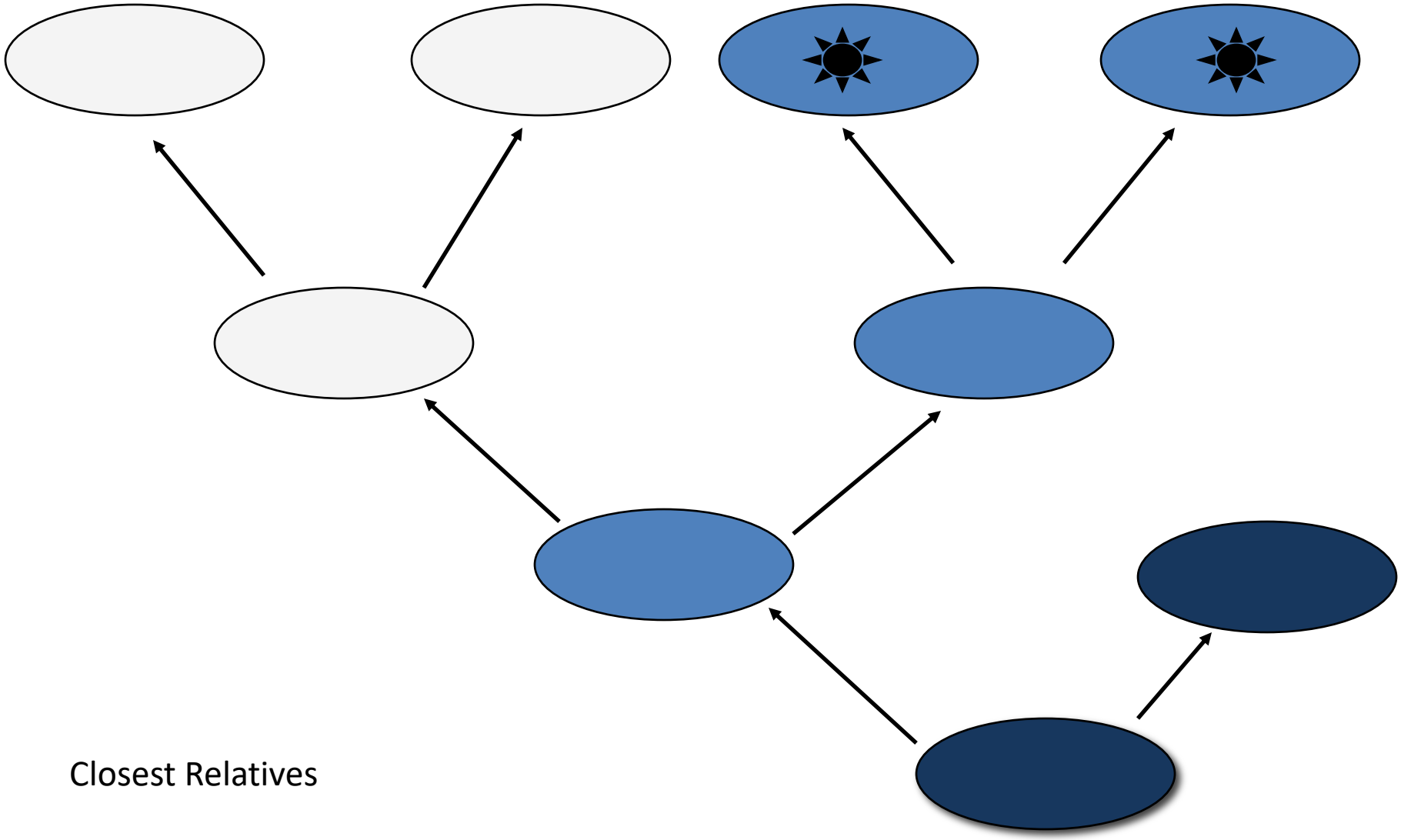
2. **Character polarity**
    a. Polarity is the direction of evolutionary change (ie., forelimb evolves into wing)
    b. Make sure that the topology has all of the assumed polarities going in the right direction.



What info is in a tree?

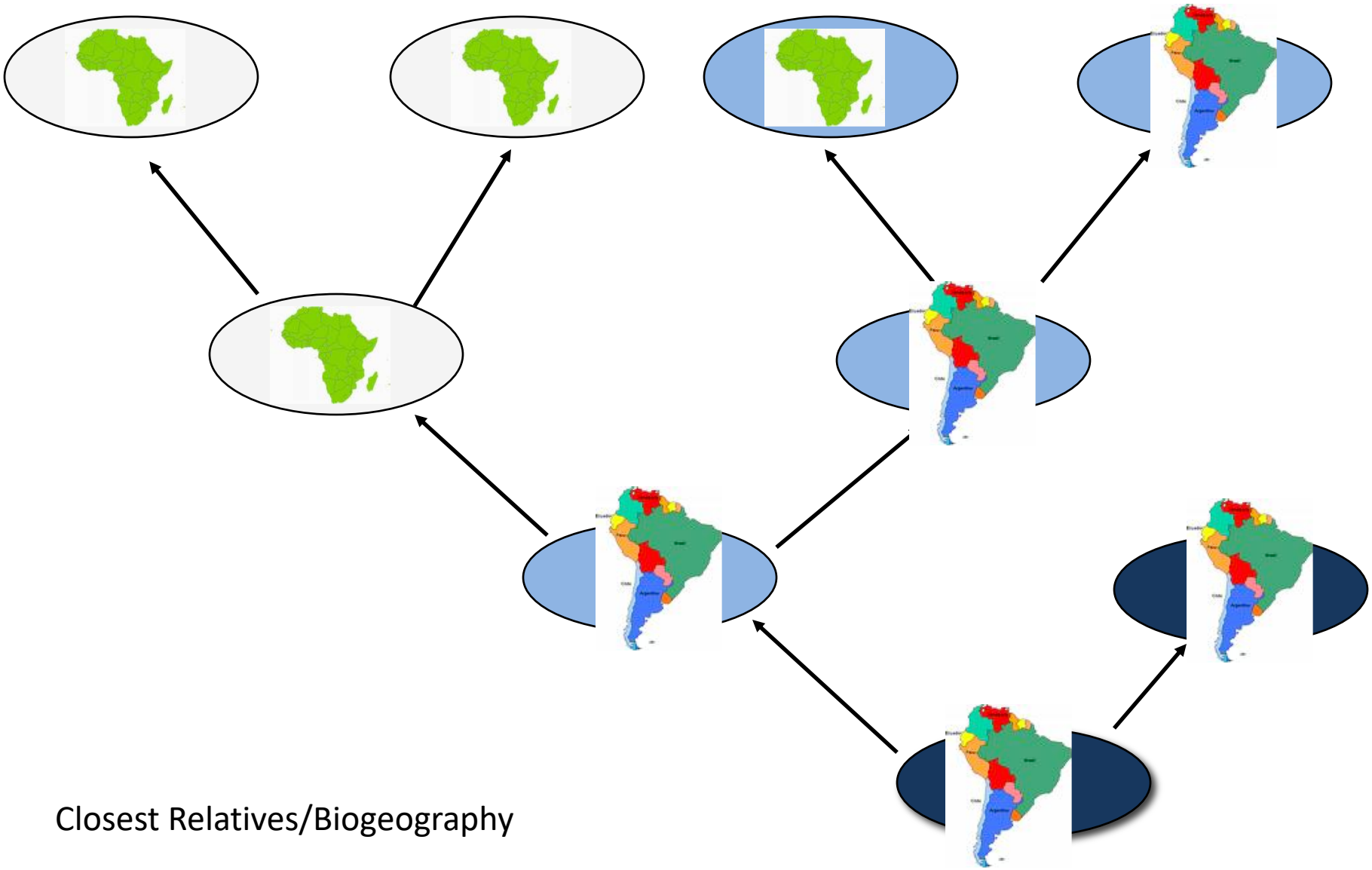


What info is in a tree?



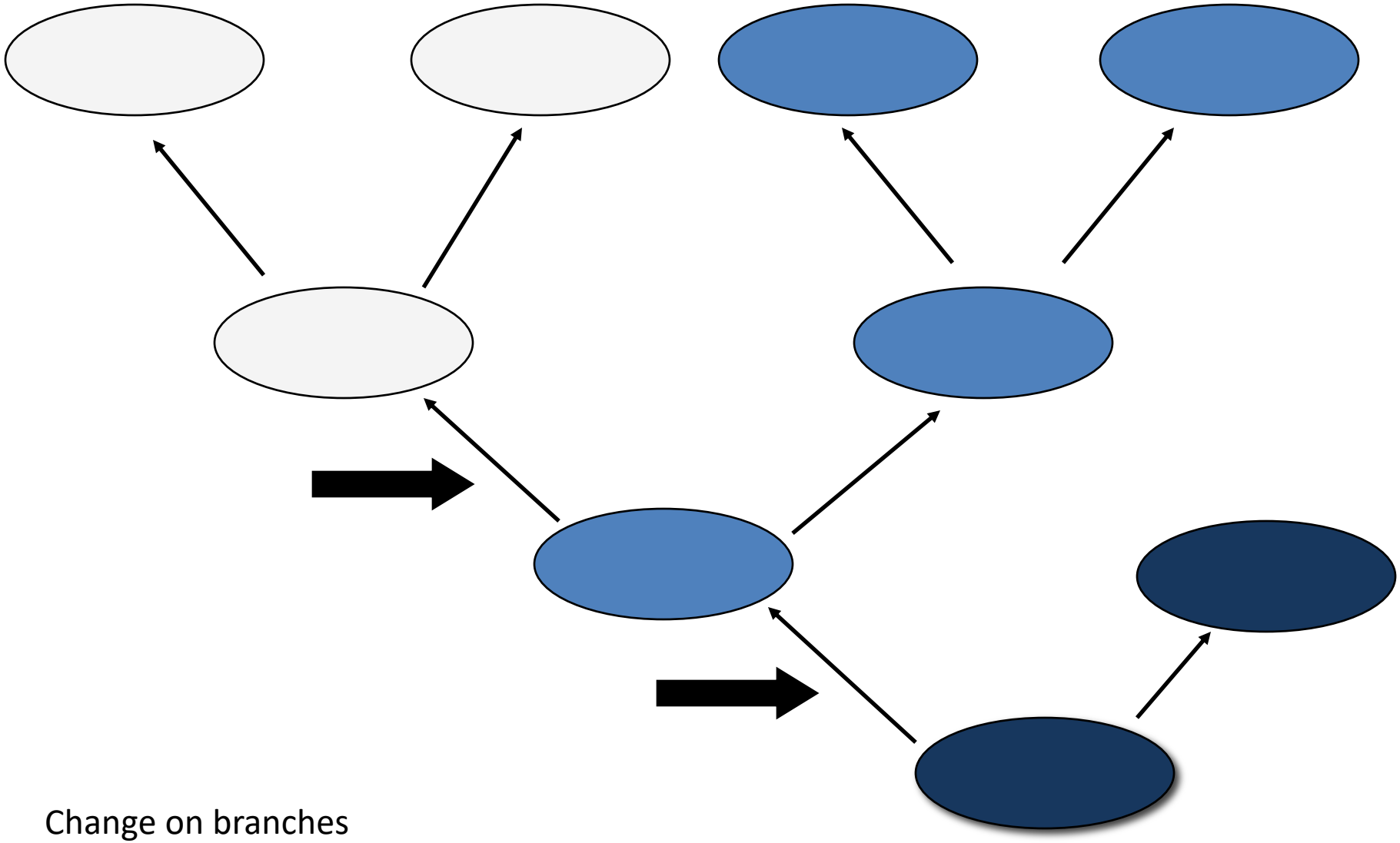
Closest Relatives

What info is in a tree?

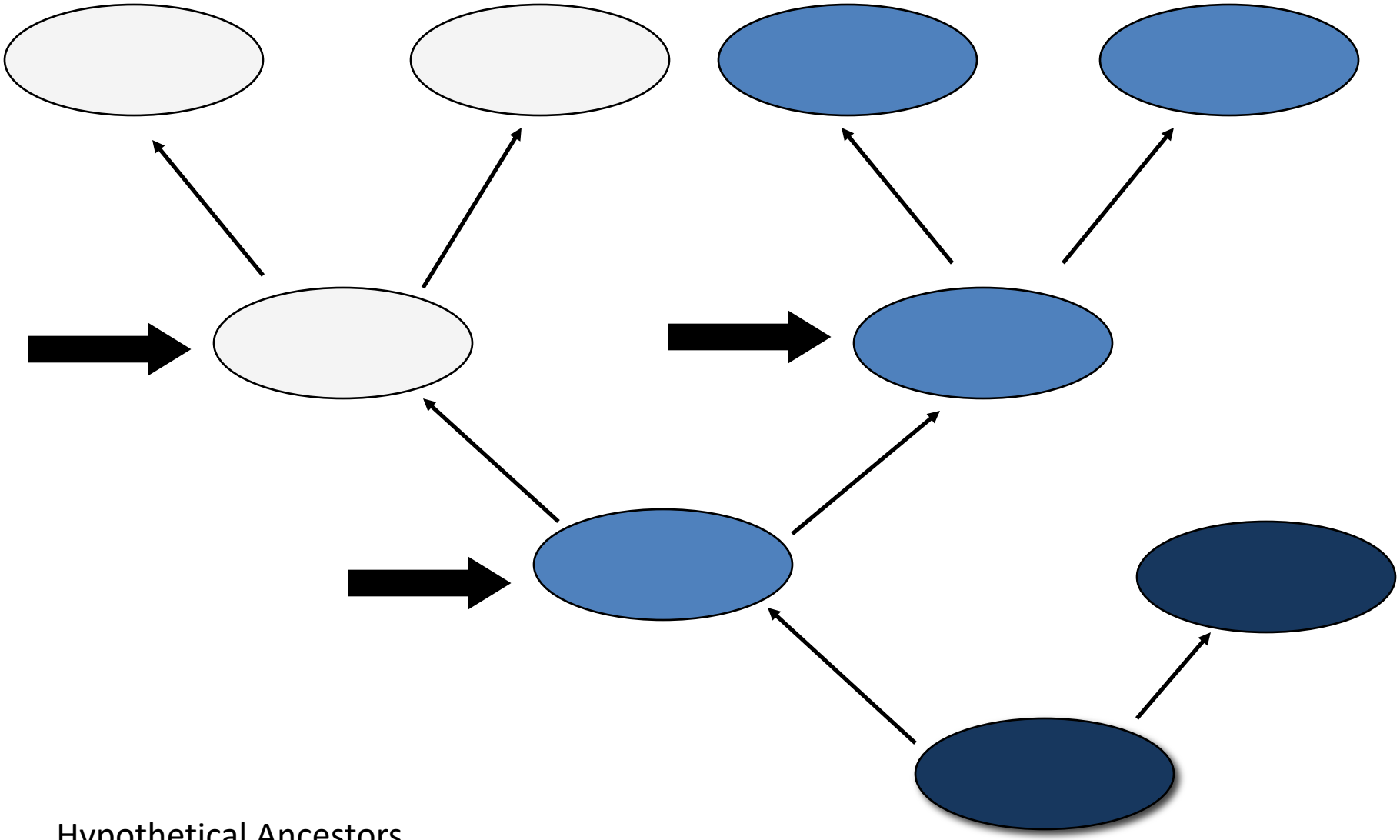


Closest Relatives/Biogeography

What info is in a tree?

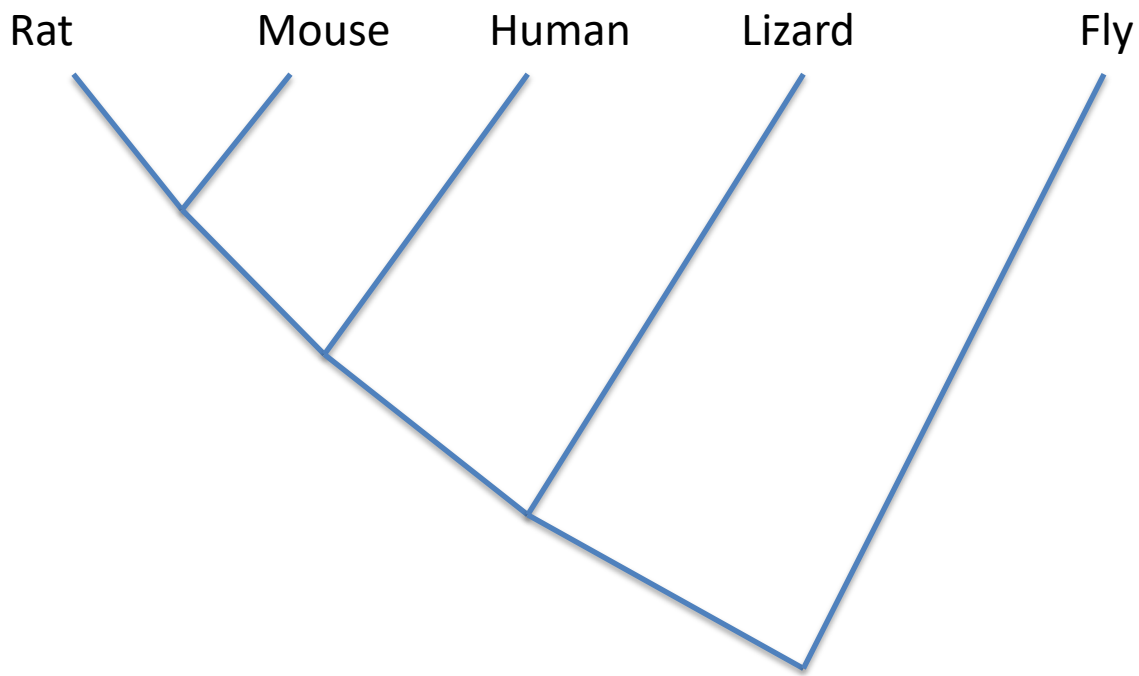


Change on branches
Evolutionary Direction/Polarity

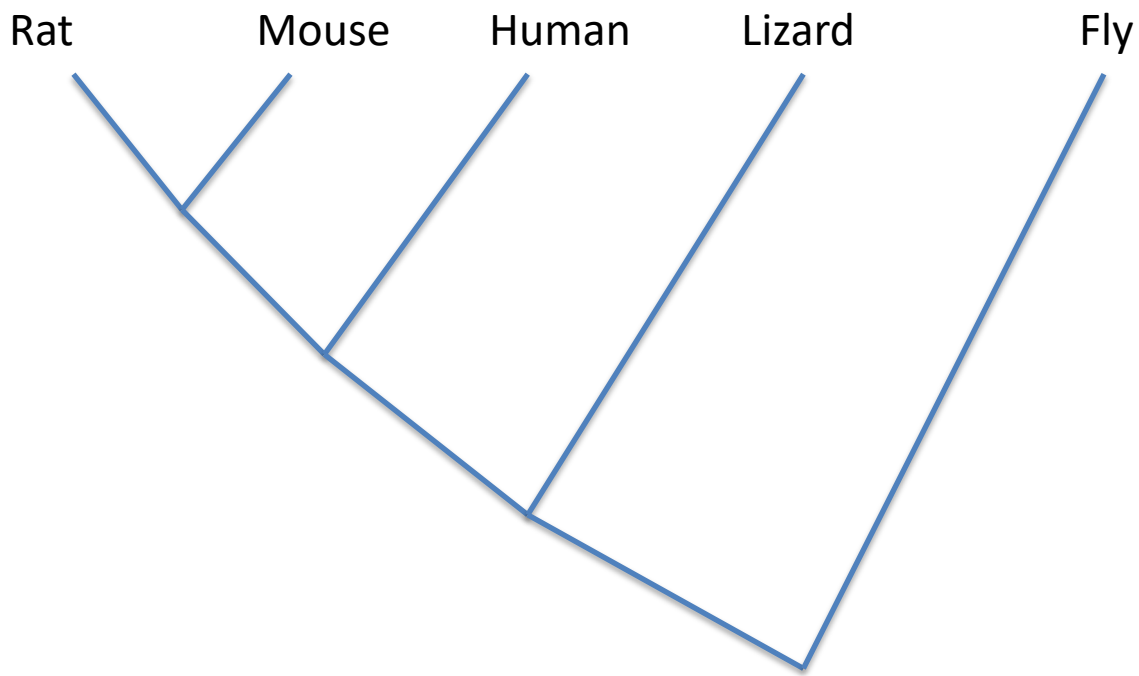


Hypothetical Ancestors

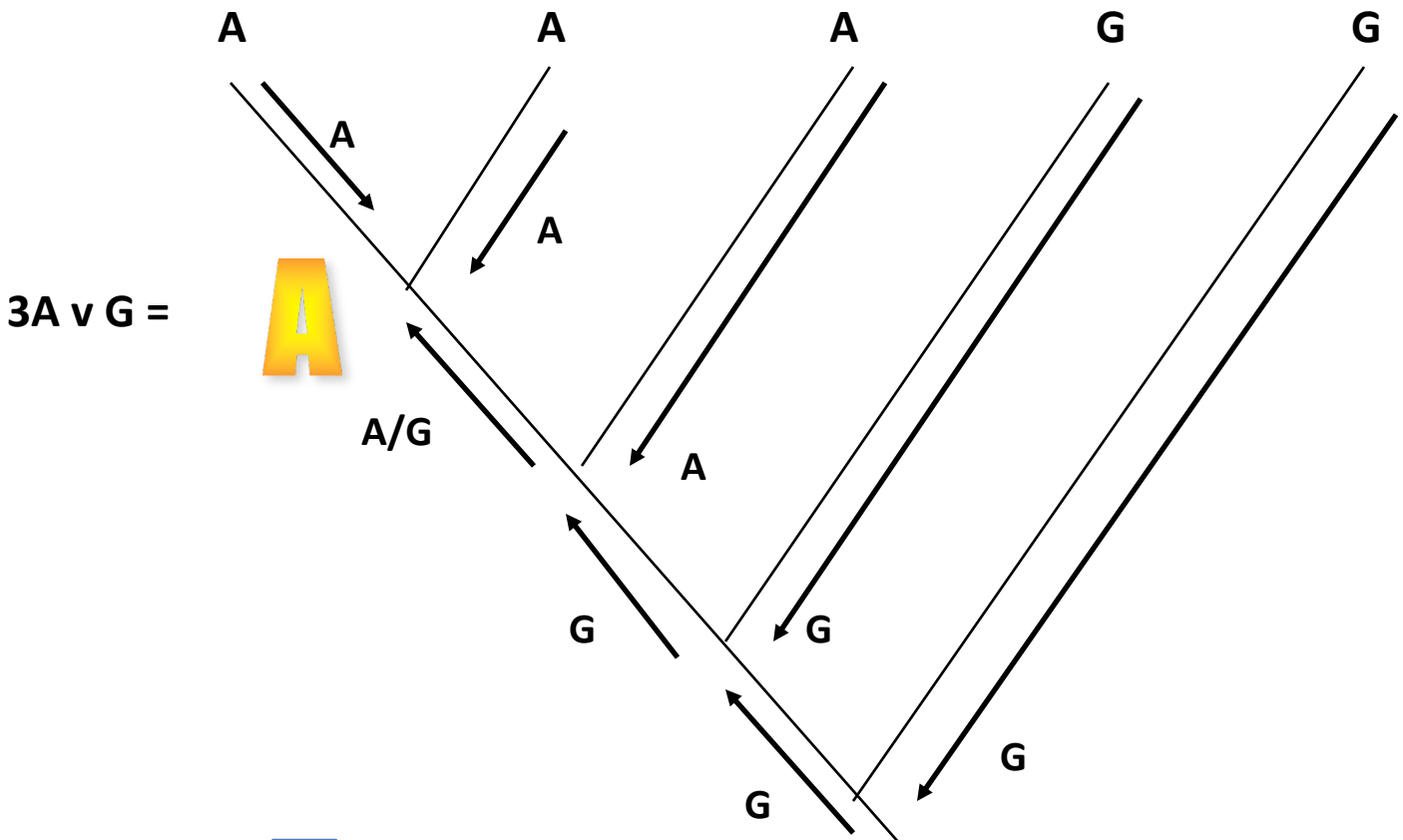
Rat	ATGGATCGTCTATTGTA
Mouse	ATGGATCGTCTATTGTA
Human	ATGGATCGTCTATTGTA
Lizard	ATGGGTCGTCTATTGTA
Fly	ATGGGTCGTCTATTGTA



```
Rat      ATGGATCGTCTATTGTA
Mouse    ATGGATCGTCTATTGTA
Human    ATGGATCGTCTATTGTA
Lizard   ATGGGTCGTCTATTGTA
Fly      ATGGGTCGTCTATTGTA
```

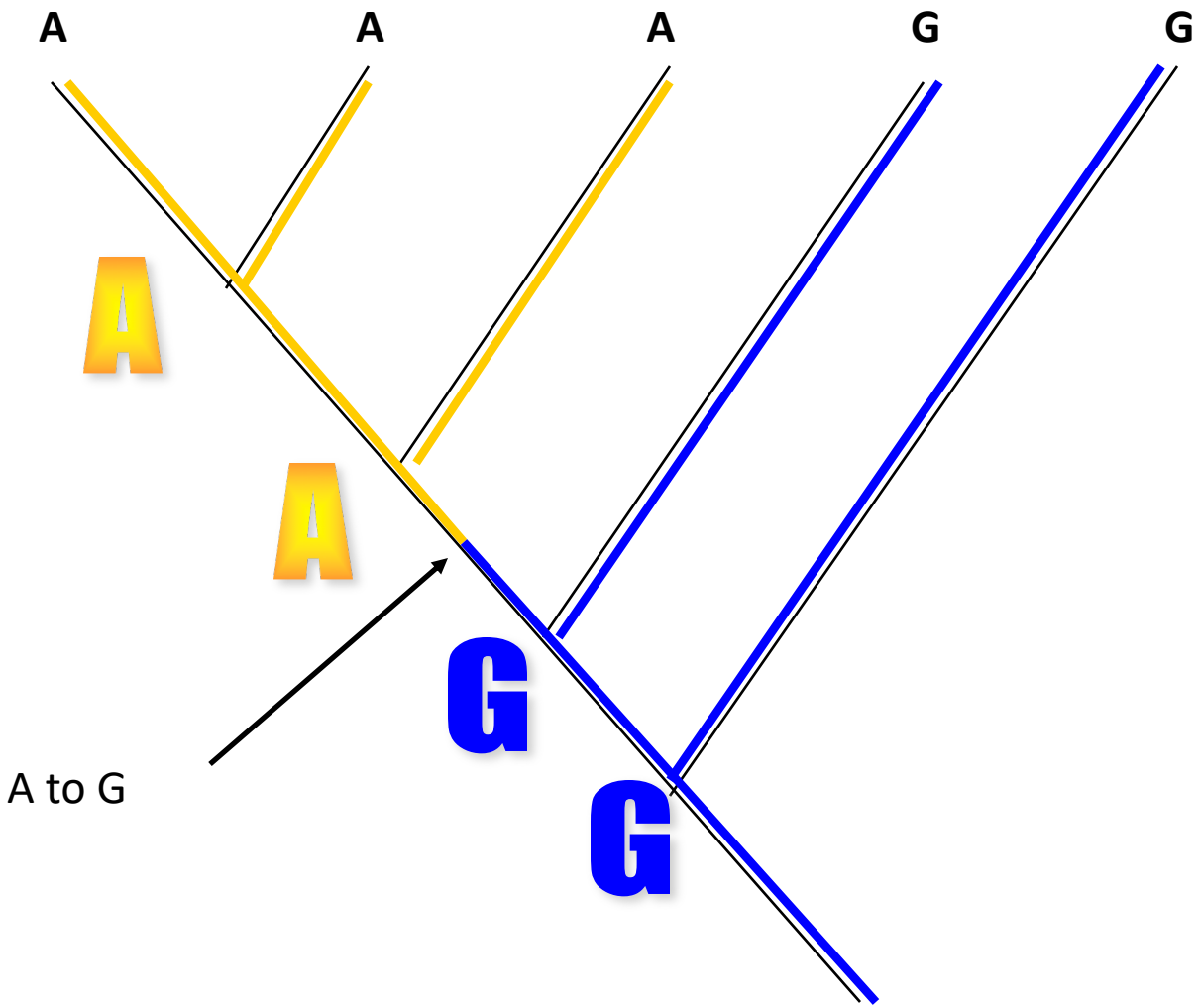


FITCH OPTIMIZATION




```
Rat      ATGGATCGTCTATTGTA
Mouse    ATGGATCGTCTATTGTA
Human    ATGGATCGTCTATTGTA
Lizard   ATGGGTCGTCTATTGTA
Fly      ATGGGTCGTCTATTGTA
```

FITCH OPTIMIZATION



Disclaimer: There are other more complicated statistical methods for this using maximum likelihood and Bayesian analysis but I did not think it was necessary to make you integrate!

How to make a tree from scratch starting with this sequence:

Human ACTCCAATGATAAATTTATAGTGGTGGAGCGCAATAAGAAATCACTGACAACTTCAC

Find Homologs and Align

```
Human    ACTCCAATGATAAATTTATAGTGGTGGAGCGCAATAAGAAATCACTGACAACTTCAC
Chimp    ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACGTCAC
Chimp2   AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG
Gorilla  ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACTTCAC
Gorilla2 AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG
Gibbon   AAACTCCAAATAAATATGTTGTAGTAAAACGAGAAAAGAAAAGTGTTTCAGATTCTA
Siamang ATATTGAGGATAAATTCATTGTTGTAGAACGTACAAAAAAATCTATAAATACAACTC
Human2  AAAATCCAGATAAATATATAGTAATAAAGCGTGAAAAGAAAAGCATATCAGATTCAA
Lucy     AGACTGGTAATAAATTTATAGTTGTAGAACGTCAAAAAAGATCCCTTACAACATCAC
Macaca   AAAATCCAGATAAGTTCATTGTGGTAAAACGTGAGAAGAAGAGTATTTCAGATTCCA
```

Find Homologs and Align

```
Human    ACTCCAATGATAAATTTATAGTGGTGGAGCGCAATAAGAAATCACTGACAACTTCAC
Chimp    ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACGTCAC
Chimp2   AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG
Gorilla  ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACTTCAC
Gorilla2 AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG
Gibbon   AAACTCCAAATAAATATGTTGTAGTAAAACGAGAAAAGAAAAGTGTTTCAGATTCTA
Siamang ATATTGAGGATAAATTCATTGTTGTAGAACGTACAAAAAAATCTATAAATACAACTC
Human2  AAAATCCAGATAAATATATAGTAATAAAGCGTGAAAAGAAAAGCATATCAGATTCAA
Lucy     AGACTGGTAATAAATTTATAGTTGTAGAACGTCAAAAAAGATCCCTTACAACATCAC
Macaca   AAAATCCAGATAAGTTCATTGTGGTAAAACGTGAGAAGAAGAGTATTTCAGATTCCA
```

Align Homologs

```
Human    ACTCCAATGATAAATTTATAGTGGTGGAGCGCAATAAGAAATCACTGACAACTTCAC
Chimp    ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACGTCAC
Chimp2   AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG
Gorilla  ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACTTCAC
Gorilla2 AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG
Gibbon   AAACTCCAAATAAATATGTTGTAGTAAAACGAGAAAAGAAAAGTGTTTCAGATTCTA
Siamang ATATTGAGGATAAATTCATTGTTGTAGAACGTACAAAAAAATCTATAAATACAACTC
Human2  AAAATCCAGATAAATATATAGTAATAAAGCGTGAAAAGAAAAGCATATCAGATTCAA
Lucy     AGACTGGTAATAAATTTATAGTTGTAGAACGTCAAAAAAGATCCCTTACAACATCAC
Macaca   AAAATCCAGATAAGTTCATTGTGGTAAAACGTGAGAAGAAGAGTATTTCAGATTCCA
```

Find the best tree: that optimizes “some function”

Align Homologs

```
Human    ACTCCAATGATAAATTTATAGTGGTGGAGCGCAATAAGAAATCACTGACAACTTCAC
Chimp    ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACGTCAC
Chimp2   AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG
Gorilla  ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACTTCAC
Gorilla2 AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG
Gibbon   AAACTCCAAATAAATATGTTGTAGTAAAACGAGAAAAGAAAAGTGTTTCAGATTCTA
Siamang ATATTGAGGATAAATTCATTGTTGTAGAACGTACAAAAAAATCTATAAATACAACTC
Human2  AAAATCCAGATAAATATATAGTAATAAAGCGTGAAAAGAAAAGCATATCAGATTCAA
Lucy     AGACTGGTAATAAATTTATAGTTGTAGAACGTCAAAAAAGATCCCTTACAACATCAC
Macaca   AAAATCCAGATAAGTTCATTGTGGTAAAACGTGAGAAGAAGAGTATTTCAGATTCCA
```

Find the best tree: that optimizes "some function"

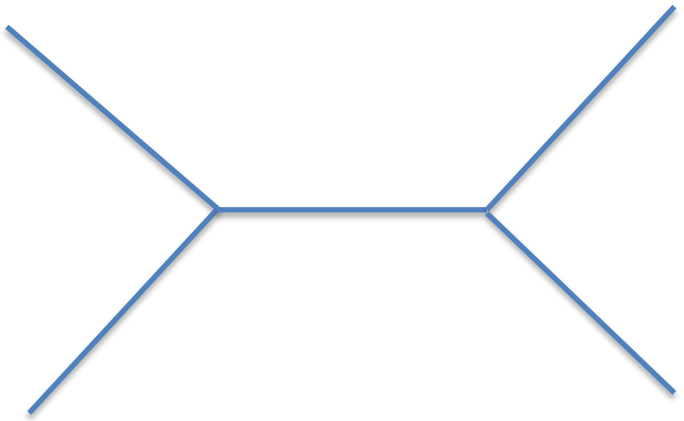
How many trees do I need to look through?

FOR UNROOTED TREES:
(2n-5)!/[2n-3(n-3)!]

FOR UNROOTED TREES:
(2n-3)!/[2n-2(n-2)!]

How tree building actually works.

A AAAAT
B AATTA
C TTTTA
D AATAT

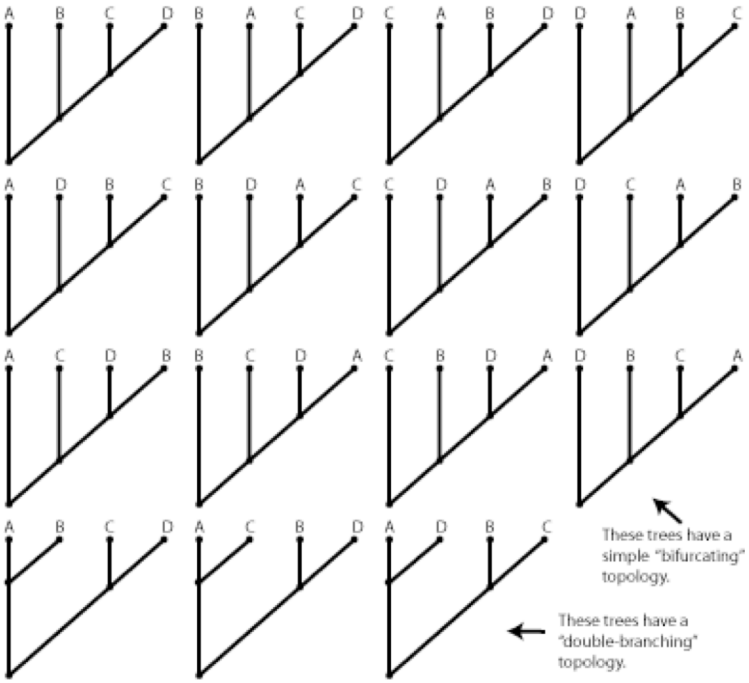


Number of Taxa	Number of Rooted Trees
3	3
4	15
5	105
6	945
7	10395
8	135135
9	2027025
10	34459425
11	654729075
12	13749310575
13	316234143225
14	7905853580625
15	213458046676875
16	6190283353629375
17	191898783962510625
18	6332659870762850625
19	221643095476699771875
20	8200794532637891559375
21	319830986772877770815625
22	13113070457687988603440625
23	563862029680583509947946875
24	25373791335626257947657609375
25	1192568192774434123539907640625

Align Homologs

```
Human   ACTCCAATGATAAATTTATAGTGGTGGAGCGCAATAAGAAATCACTGACAACTTCAC
Chimp   ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACGTCAC
Gorilla ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACTTCAC
Orang   AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG
```

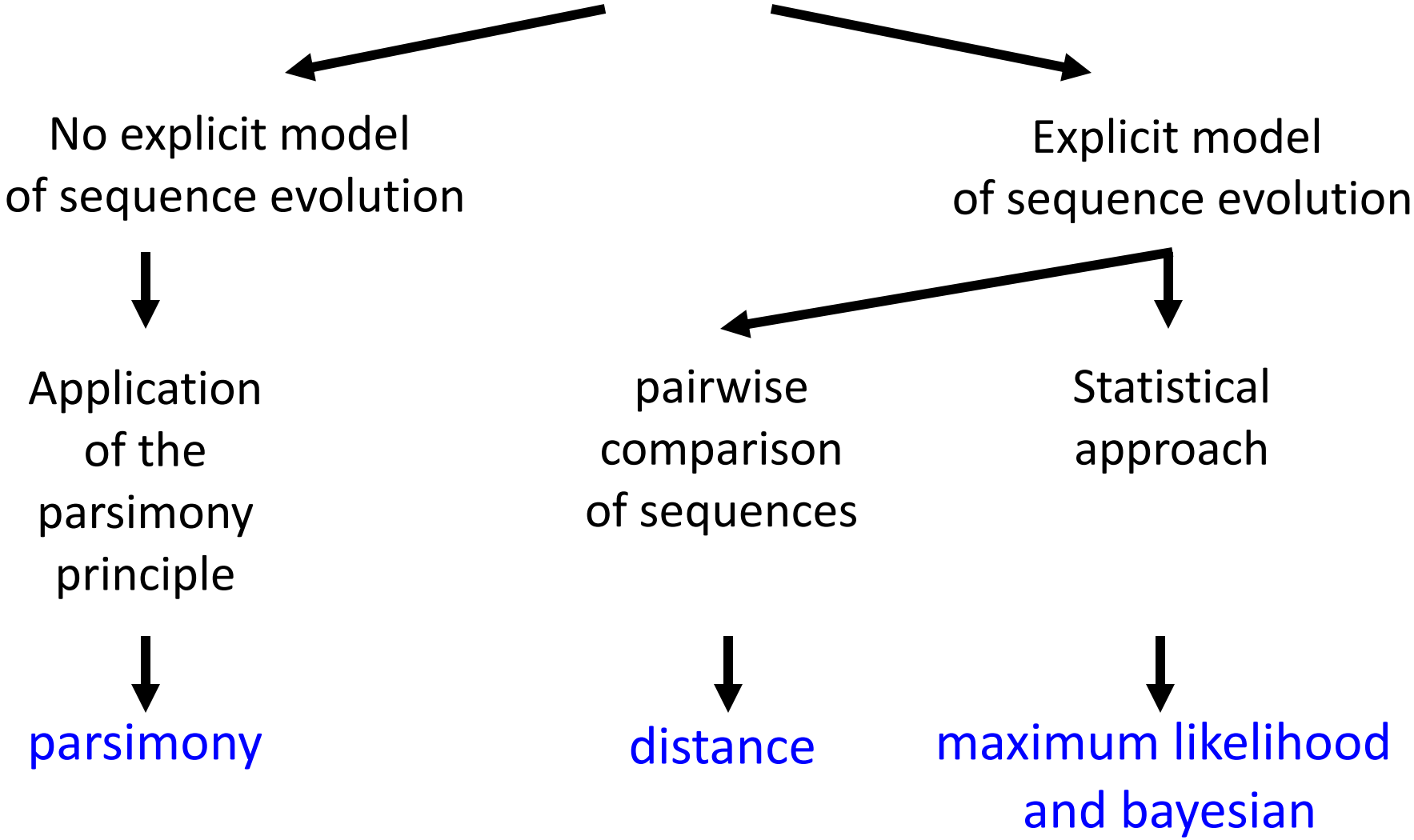
4 taxa = 15 trees



Deciding which tree to pick: Optimality Criteria

- Maximum Parsimony (what is the simplest explanation that accounts for all the data)
- Maximum Likelihood (what is the highest likelihood of observing the data given a tree)
- Bayesian (what is the highest probability that the tree is correct given the data)
- Distance/Similarity (minimize the distance or group the most similar things together)

Tree-Building Methods

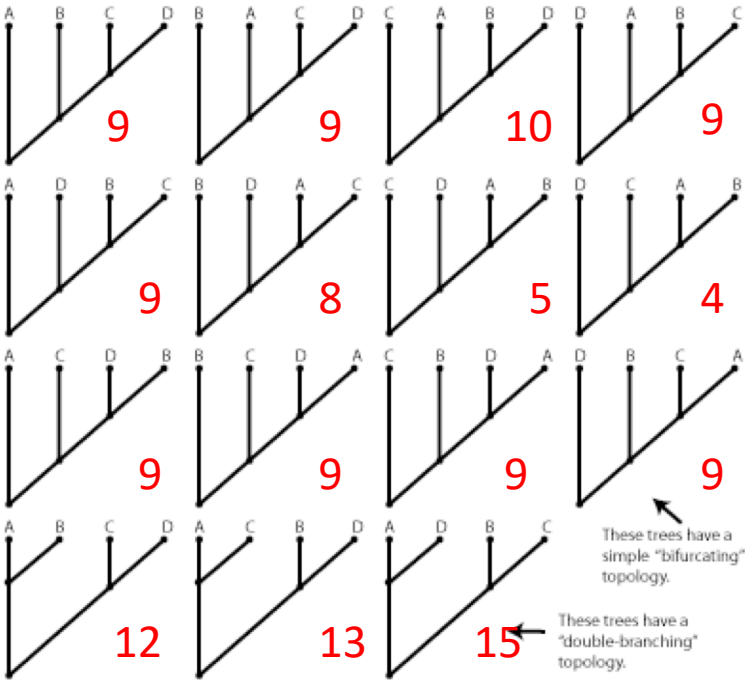


Align Homologs

A. Human ACTCCAATGATAAATTTATAGTGGTGGAGCGCAATAAGAAATCACTGACAACTTCAC
B. Chimp ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACGTCAC
C. Gorilla ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACTTCAC
D. Orang AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG

4 taxa = 15 trees

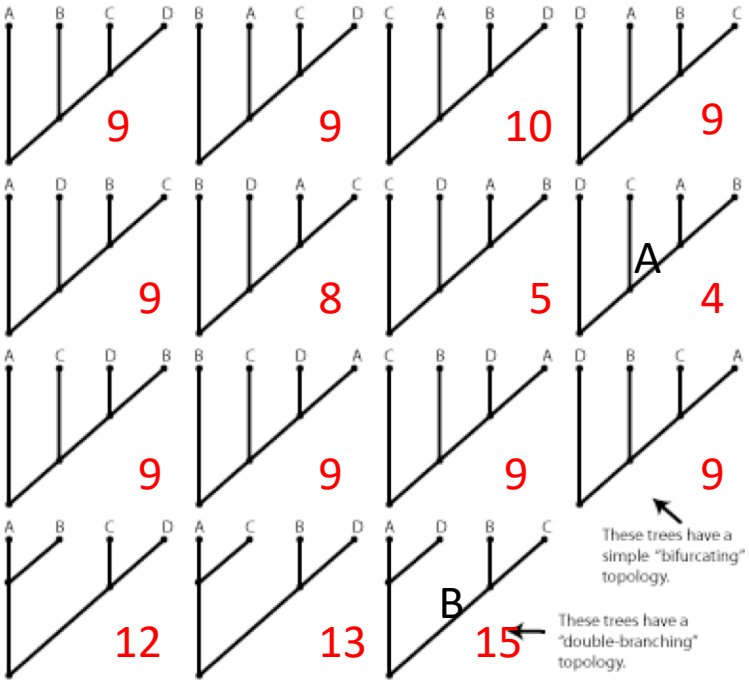
Steps in red



Align Homologs

A. Human ACTCCAATGATAAATTTATAGTGGTGGAGCGCAATAAGAAATCACTGACAACTTCAC
B. Chimp ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACGTCAC
C. Gorilla ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACTTCAC
D. Orang AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG

4 taxa = 15 trees

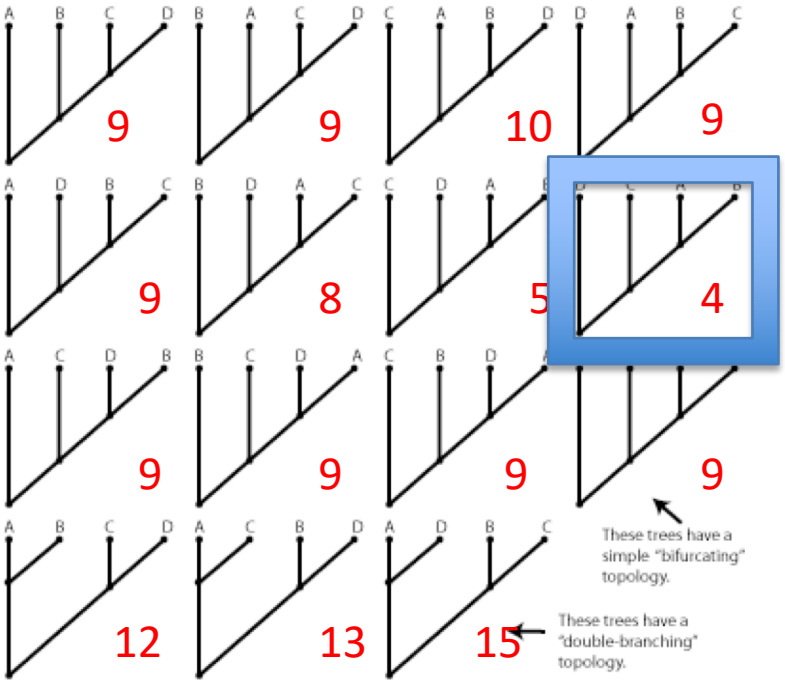


Steps in red

Align Homologs

A. Human ACTCCAATGATAAATTTATAGTGGTGGAGCGCAATAAGAAATCACTGACAACTTCAC
B. Chimp ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACGTCAC
C. Gorilla ATTCAAATGGTAAATTTGTAGTAGTGGAACGCGAGAAAAAATCACTTACAACTTCAC
D. Orang AAATCATTTCTGAGACATTATGGGATAATGGGTTTGGTAAAACAATTCATGCACAGG

4 taxa = 15 trees



Steps in red

Estimating Genetic Distance

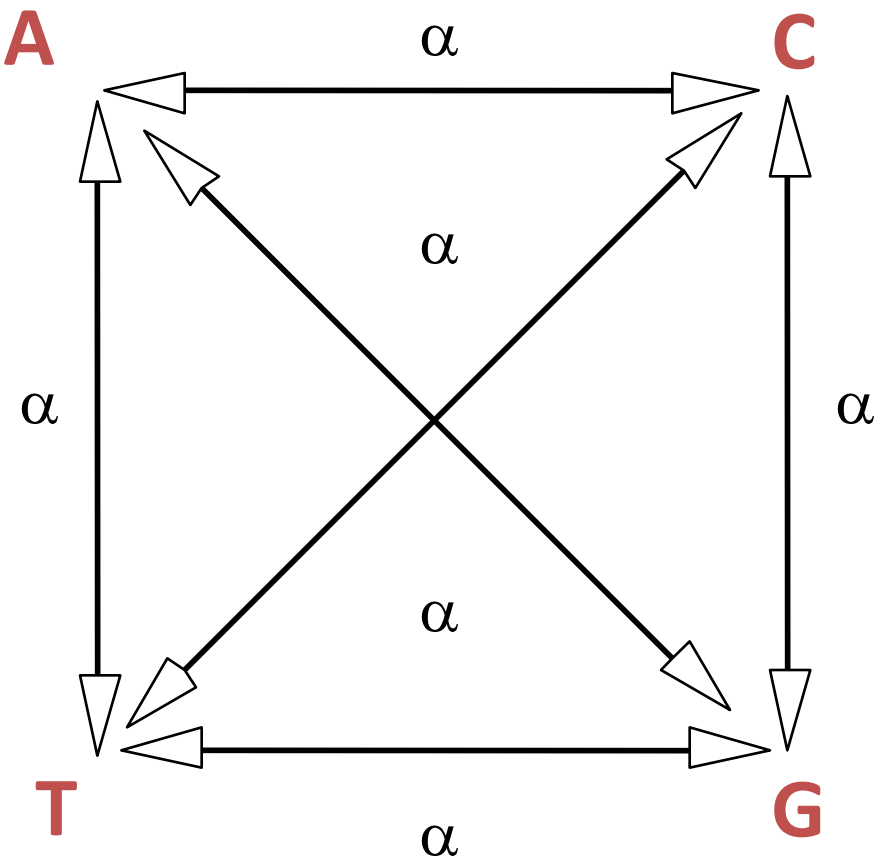
```
SIVcpz    ATGGGTGCGA GAGCGTCAGT TCTAACAGGG GGAAAATTAG ATCGCTGGGA
HIV-1     ATGGGTGCGA GAGCGTCAGT ATTAAGCGGG GGAGAATTAG ATCGATGGGA

SIVcpz    AAAAGTTCGG CTTAGGCCCG GGGGAAGAAA AAGATATATG ATGAAACATT
HIV-1     AAAAATTCGG TTAAGGCCAG GGGGAAAGAA AAAATATAAA TTAAAACATA

SIVcpz    TAGTATGGGC AAGCAGGGAG CTGGAAAGAT TCGCATGTGA CCCCGGGCTA
HIV-1     TAGTATGGGC AAGCAGGGAG CTAGAACGAT TCGCAGTTAA TCCTGGCCTG

SIVcpz    ATGGAAAGTA AGGAAGGATG TACTAAATTG TTACAACAAT TAGAGCCAGC
HIV-1     TTAGAAACAT CAGAAGGCTG TAGACAAATA CTGGGACAGC TACAACCATC

SIVcpz    TCTCAAAACA GGCTCAGAAG GACTGCGGTC CTTGTTTAAC ACTCTGGCAG
HIV-1     CCTTCAGACA GGATCAGAAG AACTTAGATC ATTATATAAT ACAGTAGCAA

SIVcpz    TACTGTGGTG CATACATAGT GACATCACTG TAGAAGACAC ACAGAAAGCT
HIV-1     CCCTCTATTG TGTGCATCAA AGGATAGAGA TAAAAGACAC CAAGGAAGCT

SIVcpz    CTAGAACAGC TAAAGCGGCA TCATGGAGAA CAACAGAGCA AAACTGAAAG
HIV-1     TTAGACAAGA TAGAG--GAA -----GAGCA AAACAAAAGT AA---GAAAA

SIVcpz    TAACTCAGGA AGCCGTGAAG GGGGAGCCAG TCAAGGCGCT AGTGCCTCTG
HIV-1     AAGCACAGCA AGC-----AG CAGCTGACA- -CAGGACAC- AG--CAGC--

SIVcpz    CTGGCATTAG TGGAAATTAC
HIV-1     CAGG--TCAG CCAAAATTAC
```

Models of DNA Substitution

• Models of DNA sequence evolution are required to recover the missing information through correcting for multiple substitutions.

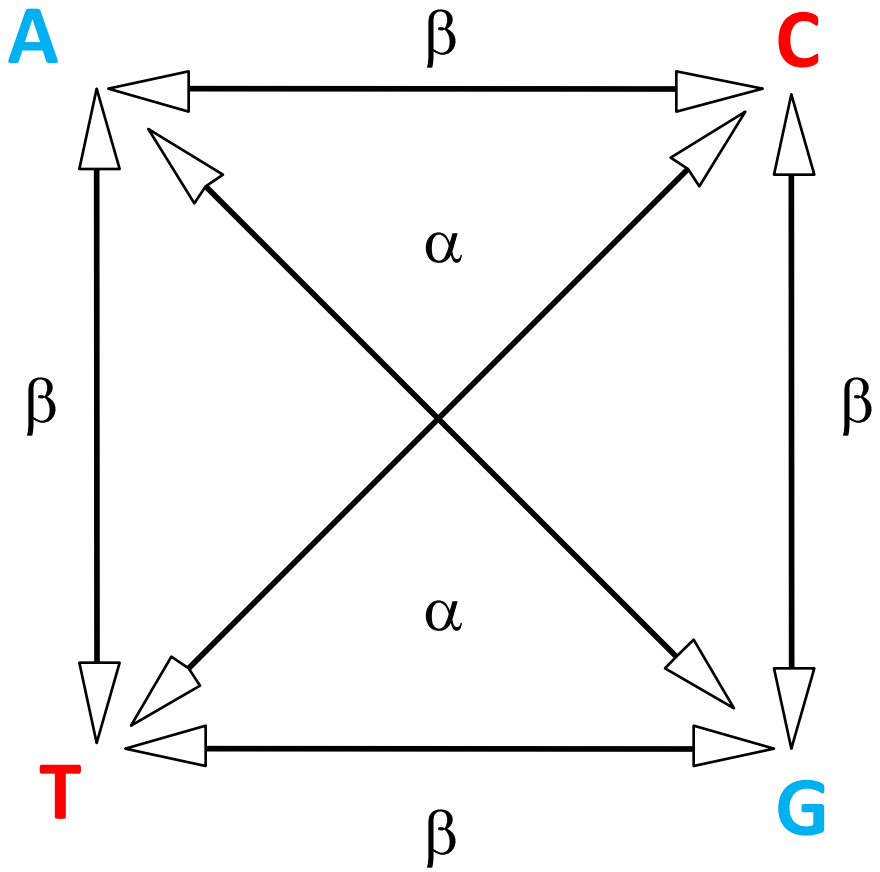
i. The probability of substitution between bases (e.g. A to C, C to T…)

ii. The probability of substitution along a sequence (different sites/regions evolve at different rates)



All substitutions occur at the same rate (α)

Is this model too simple for real data?



Transitions (α) and transversions (β) occur at a different rate

Models of DNA Substitution

Simplest
(few parameters)

1. Base frequencies are equal and all substitutions are equally likely (Jukes-Cantor)

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

3. Unequal base frequencies and transitions and transversions occur at different rates (Hasegawa-Kishino-Yano)

4. Unequal base frequencies and all substitution types occur at different rates (General Reversible Model)

Most complex
(many parameters)

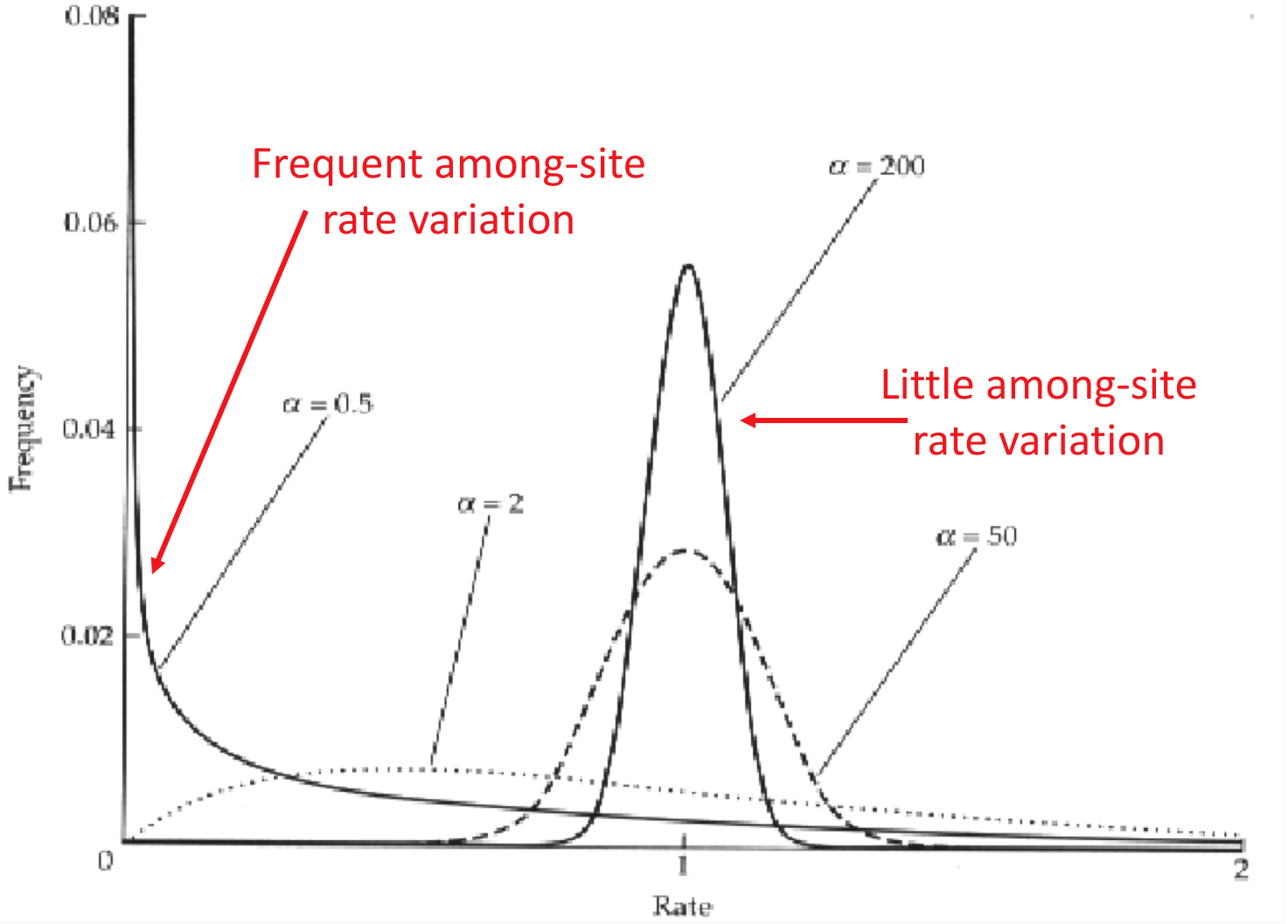
All these models can be tested using the program jMODELTEST (darwin.uvigo.es/software/jmodeltest.html)

Models of DNA Substitution

i. The probability of substitution between bases (e.g. A to C, C to T…)

ii. The probability of substitution along a sequence (different sites/regions evolve at different rates)

A Gamma Distribution Can be Used to Model Among-Site Rate Heterogeneity

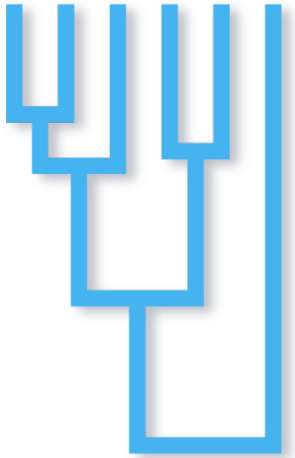


Non-Parametric Bootstrapping

```
1 A C C T G G
2 A C C T G G
3 A C C T A C
4 A T C T A T
5 A T C T A T
6 A T G G A A
```

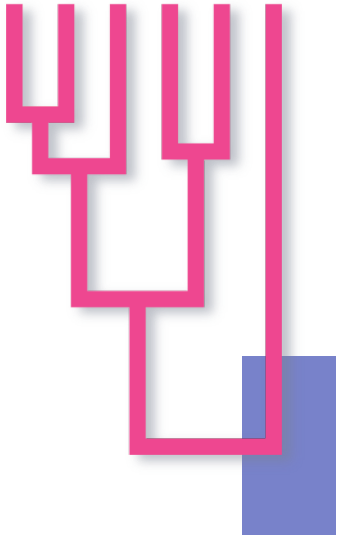
1

```
1 G G G A T C
2 G G G A T C
3 A A C A T C
4 A A T A T T
5 A A T A T T
6 A A A A G T
```



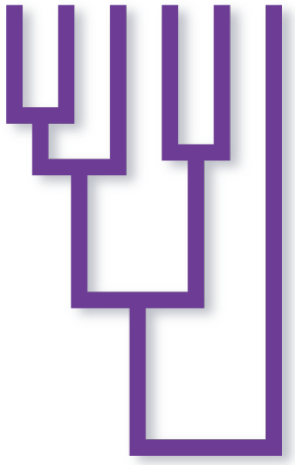
Resample with replacement multiple times

1 2 3 4 5 6



1000

```
1 T G C C A G
2 T G C C A G
3 T C C C A A
4 T T C C A A
5 T T C C A A
6 G A G G A A
```



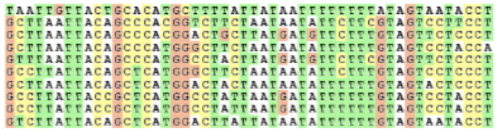
LARGE-SCALE PHYLOGENETIC ANALYSIS

1. Genome Concatenation or Whole Genome alignments

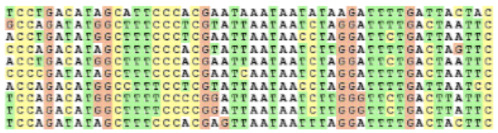
2. SNP calling with reference to a matrix

LARGE-SCALE STANDARD PHYLOGENETIC ANALYSIS

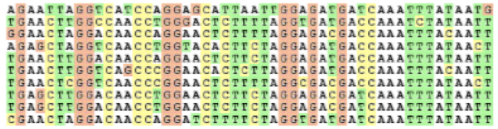
Concatenation



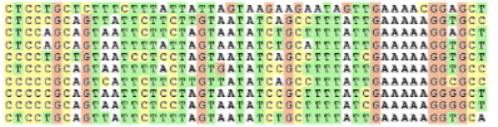
Gene 1



Gene 2



Gene 3



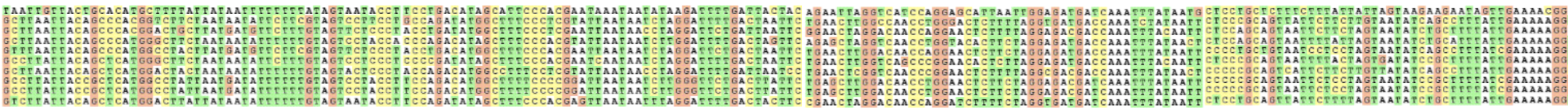
Gene 4

LARGE-SCALE PHYLOGENETIC ANALYSIS

A CONCATENATED MATRIX

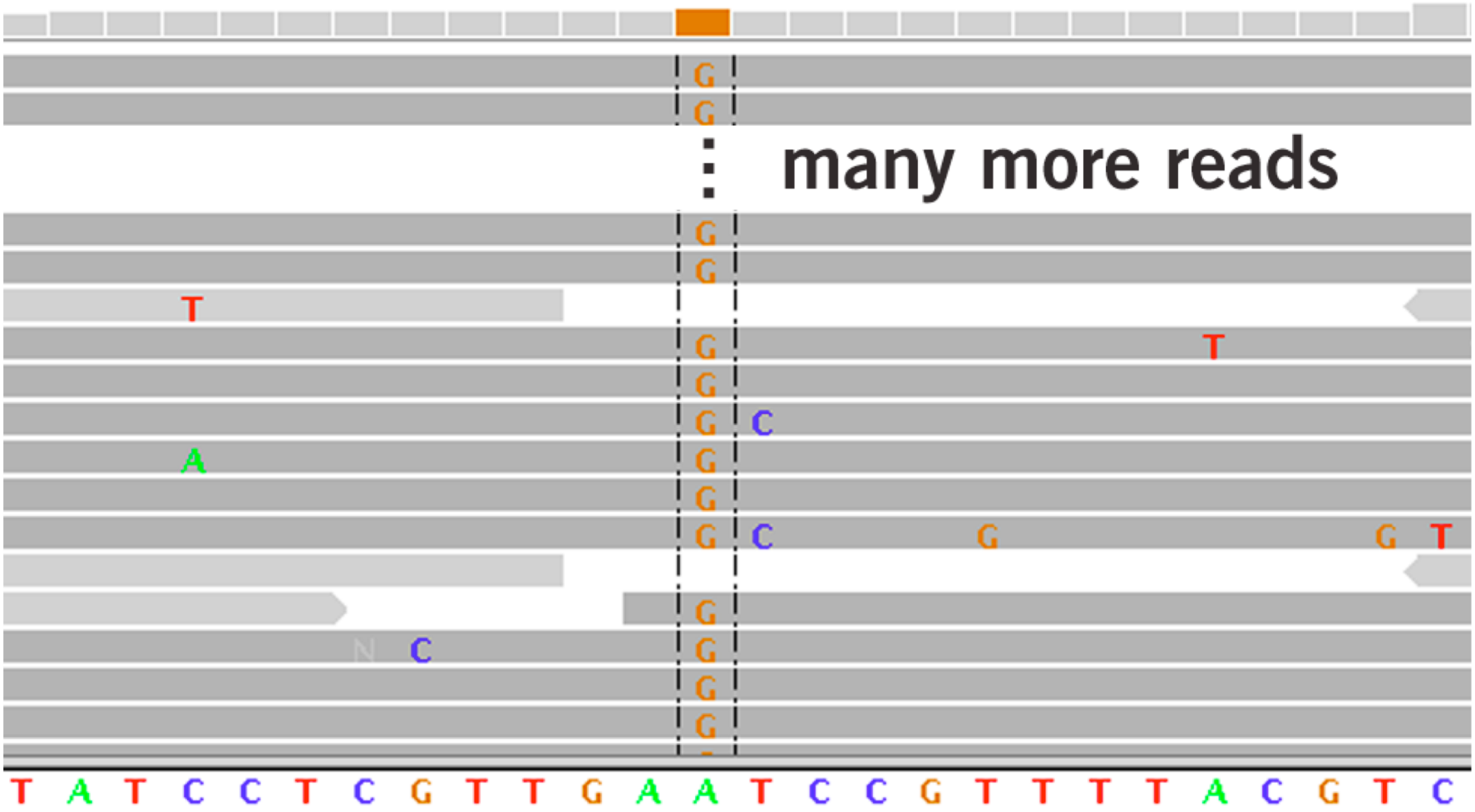
Final matrix has both variant and invariant positions.

Gene 1 Gene 2 Gene 3 … Gene n



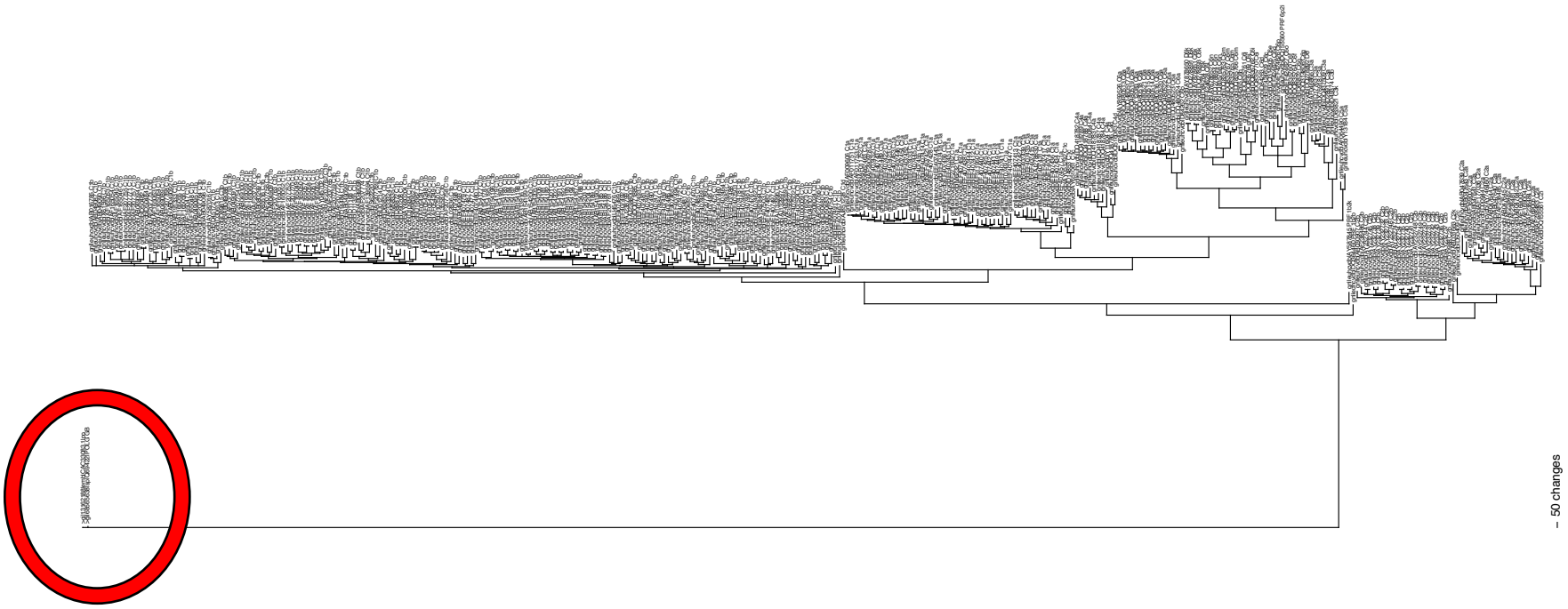
LARGE-SCALE PHYLOGENETIC ANALYSIS (Reference Based)

SNP Calling (final matrix has only invariant positions=danger ascertainment bias)



Things to think about when you look at a tree

1. How is it rooted?
2. What do the branch lengths mean?
3. Do the branch lengths make sense?



RAxML Settings:
Model of Binary: GTRCAT (GTR + Optimization of substitution rates + Optimization of site-specific evolutionary rates which are categorized into numberOfCategories distinct rate categories for greater computational efficiency.
Bootstrapping Criteria: autoMRE (for the majority-rule tree based criteria)
BFGS: Disabled automatic usage of BFGS method to optimize GTR rates on unpartitioned DNA datasets
Ascertainment Bias Correction: Paul Lewis Method
Rapid Bootstrap Number Seed: 123
Parsimony Random Seed: 256