# Principles of Phylogenetics

## Reading and Inferring Trees

Finlay Maguire

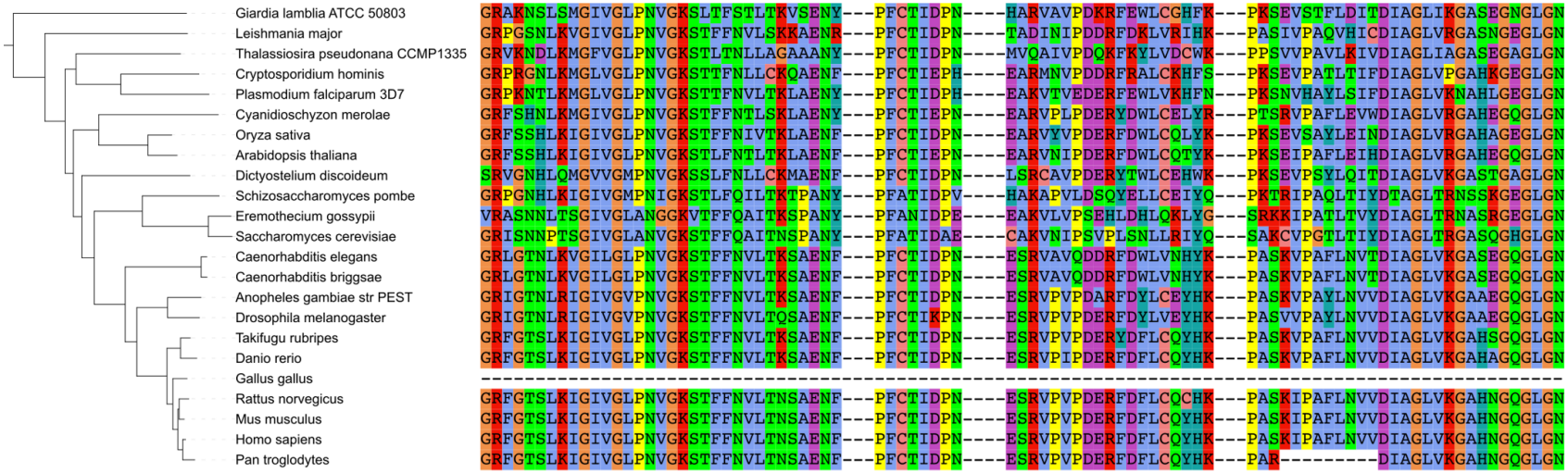
April 1, 2020

FCS, Dalhousie

# Table of contents

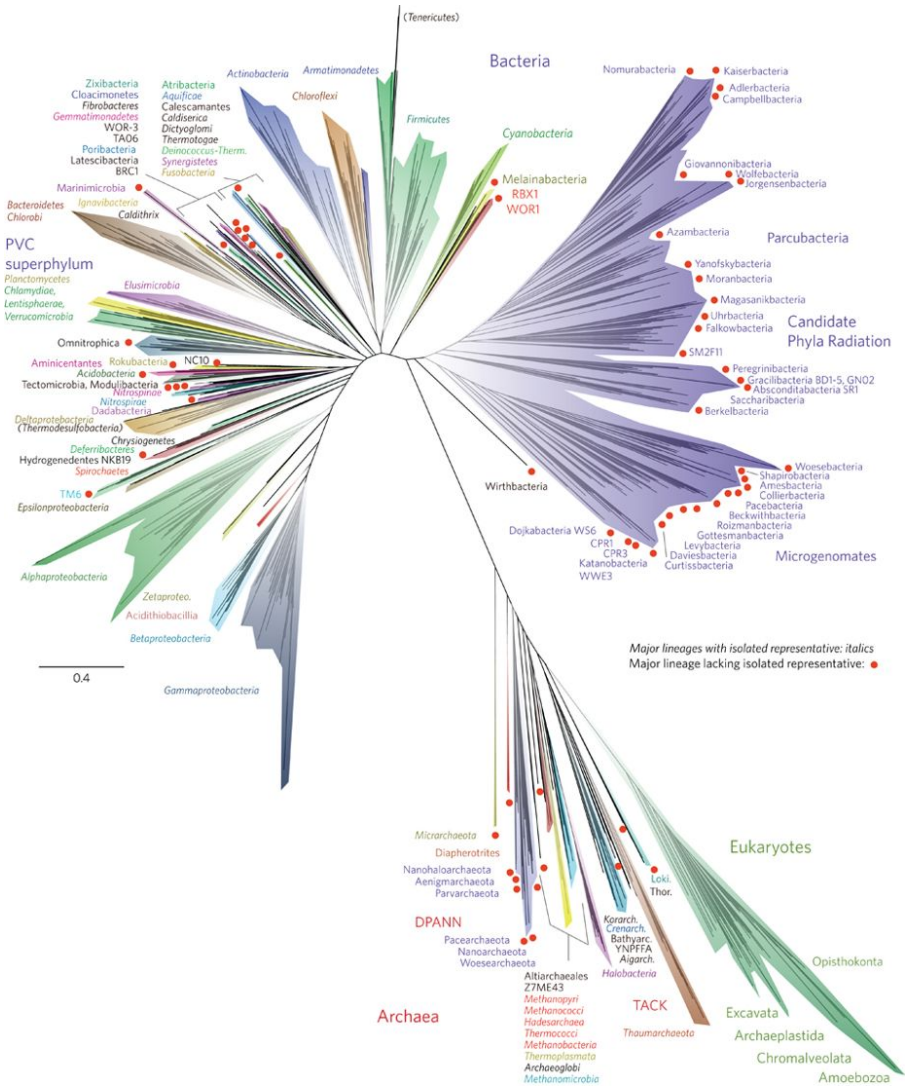
# What are phylogenies?

# Hypotheses for understanding alignments



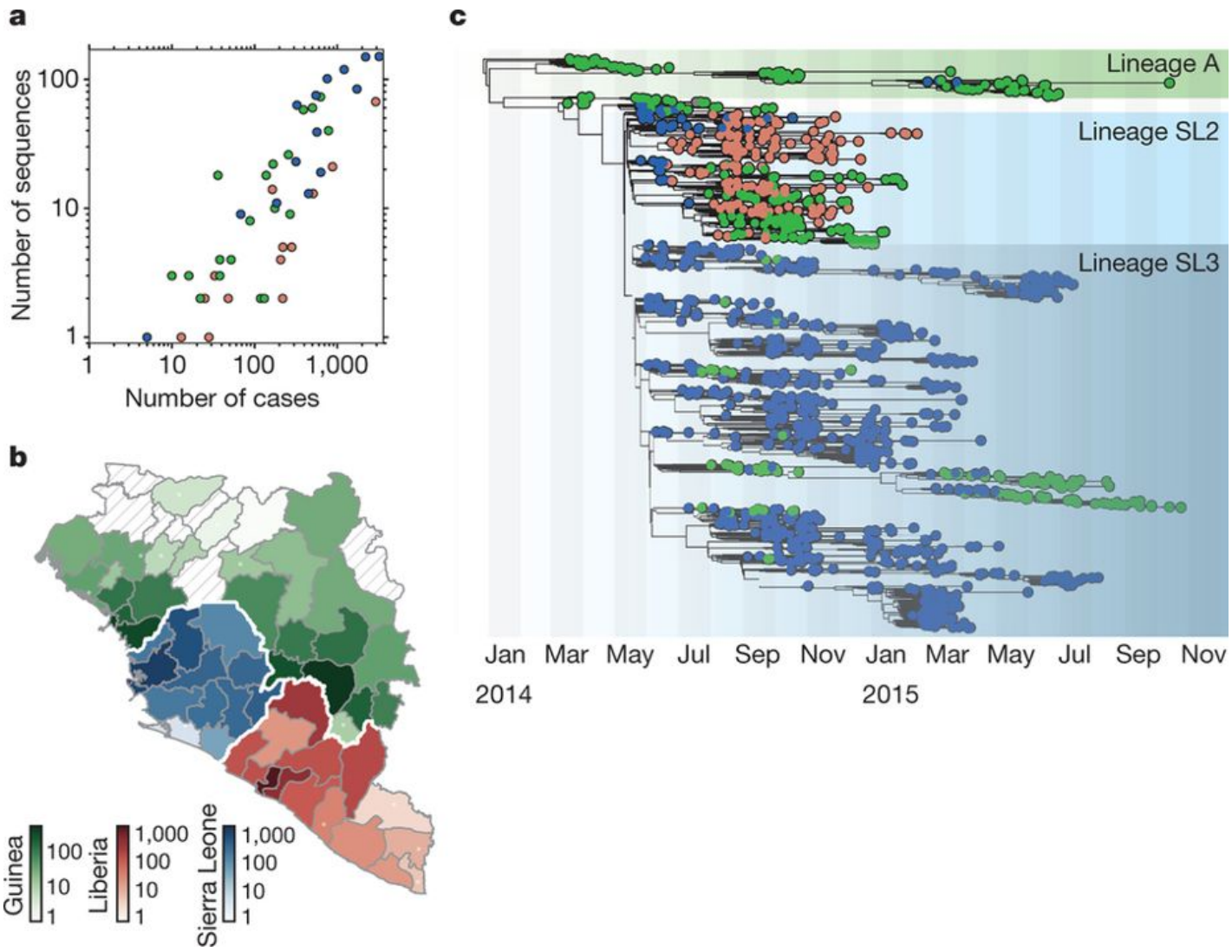

https://itol.embl.de/help.cgi

# Tree of Life



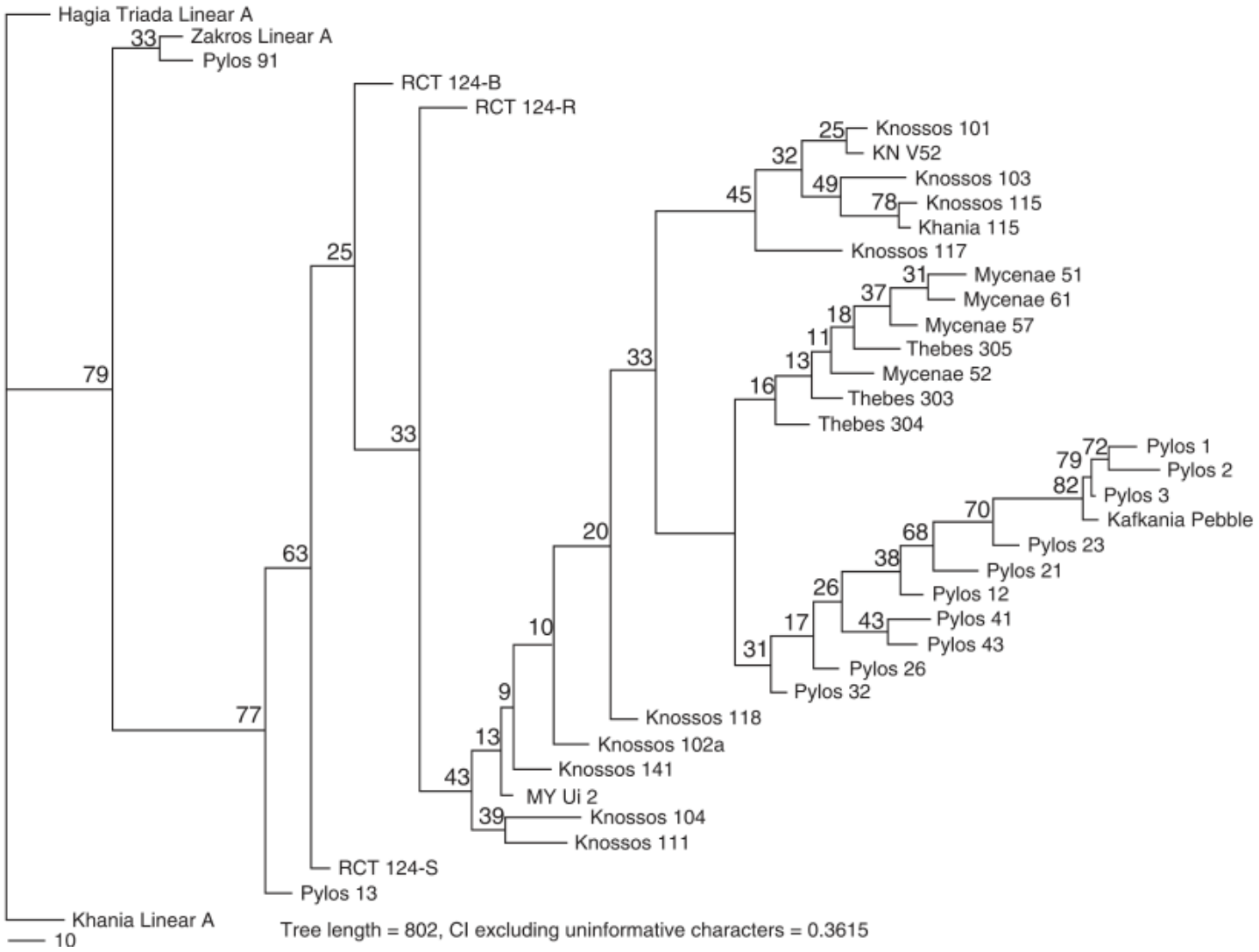

[Hug et al., 2016]

# 2013- Ebola Outbreak




[Holmes et al., 2016]

# Uses outside biology



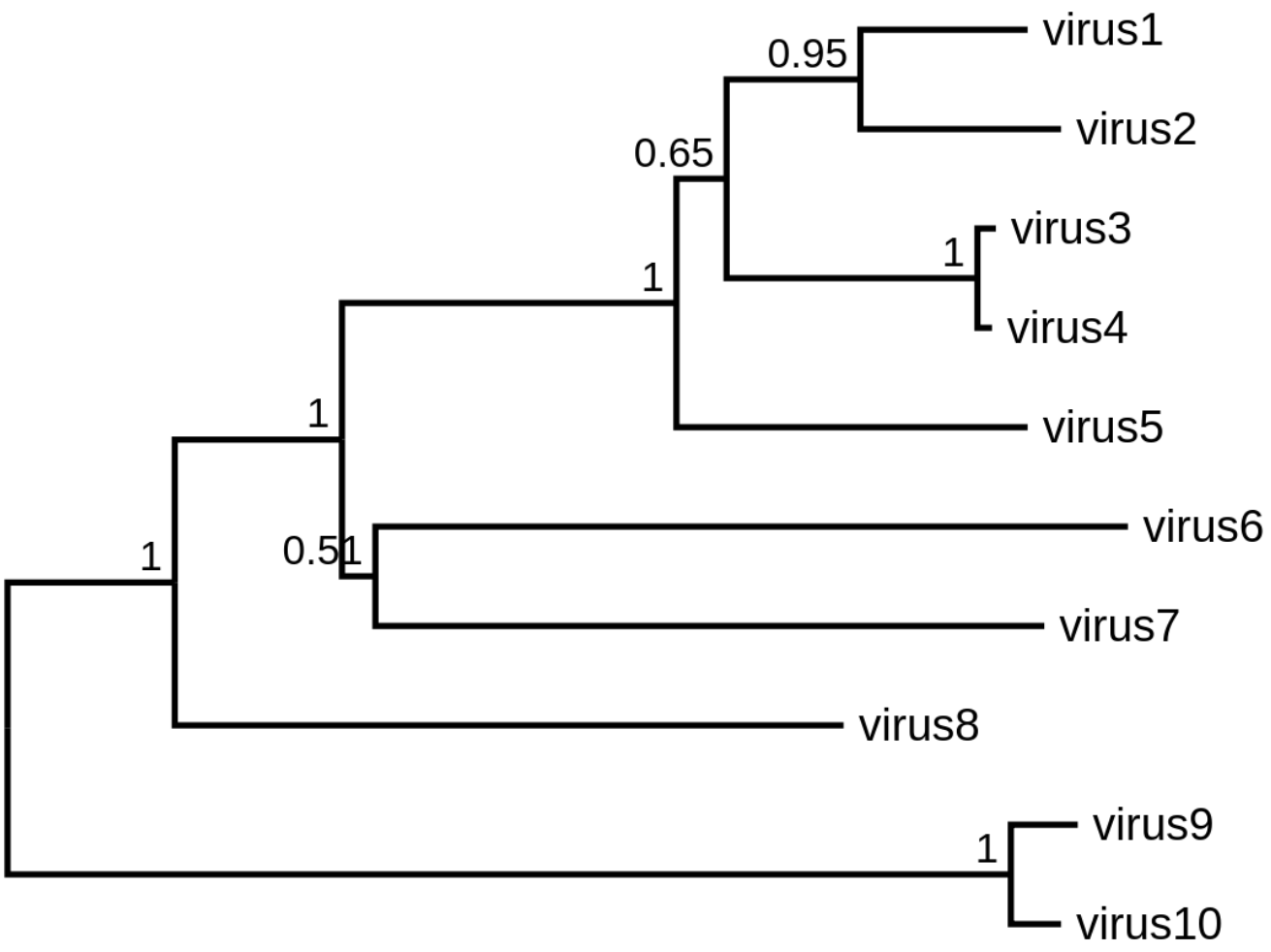

[Skelton, 2008]

# Uses outside biology

- Manuscript change [Barbrook et al., 1998]
- Social evolution (many examples, some questionable).
- Plagiarism [Ryu et al., 2008]
- Anything you can measure distances between.

# Reading a Tree

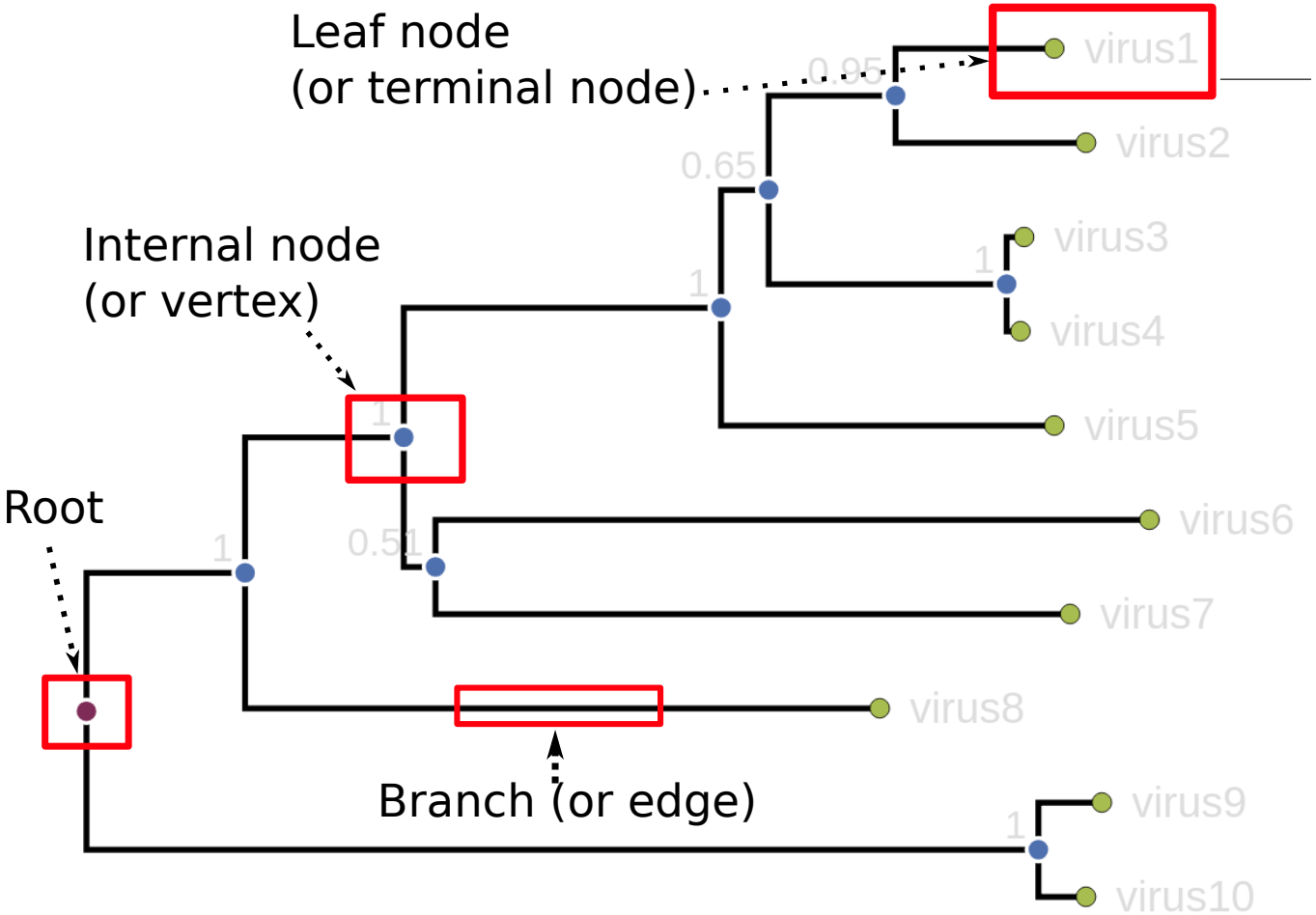
# Toy Tree



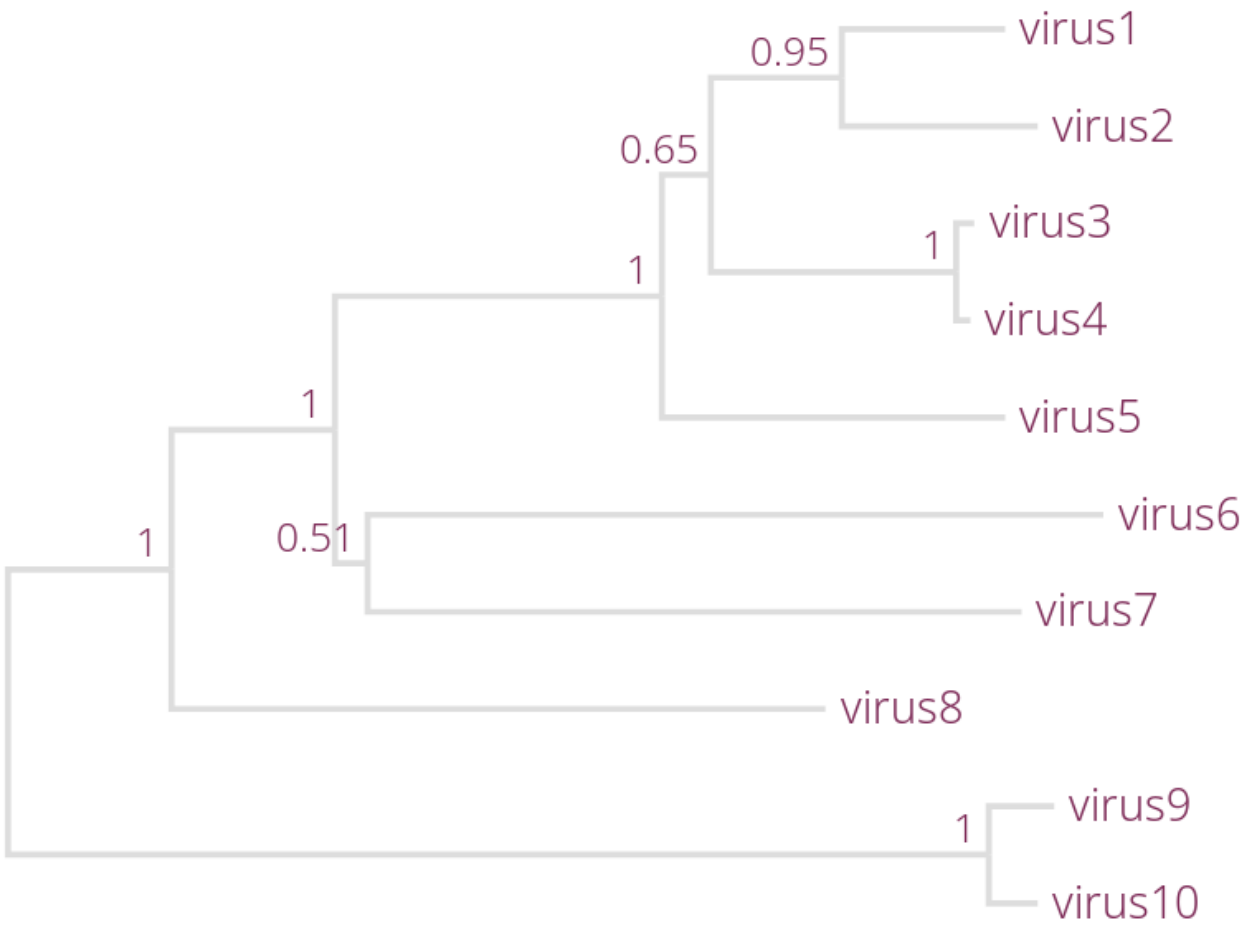

Andrew Rambaut's Tutorial

`http://artic.network/how-to-read-a-tree.html`

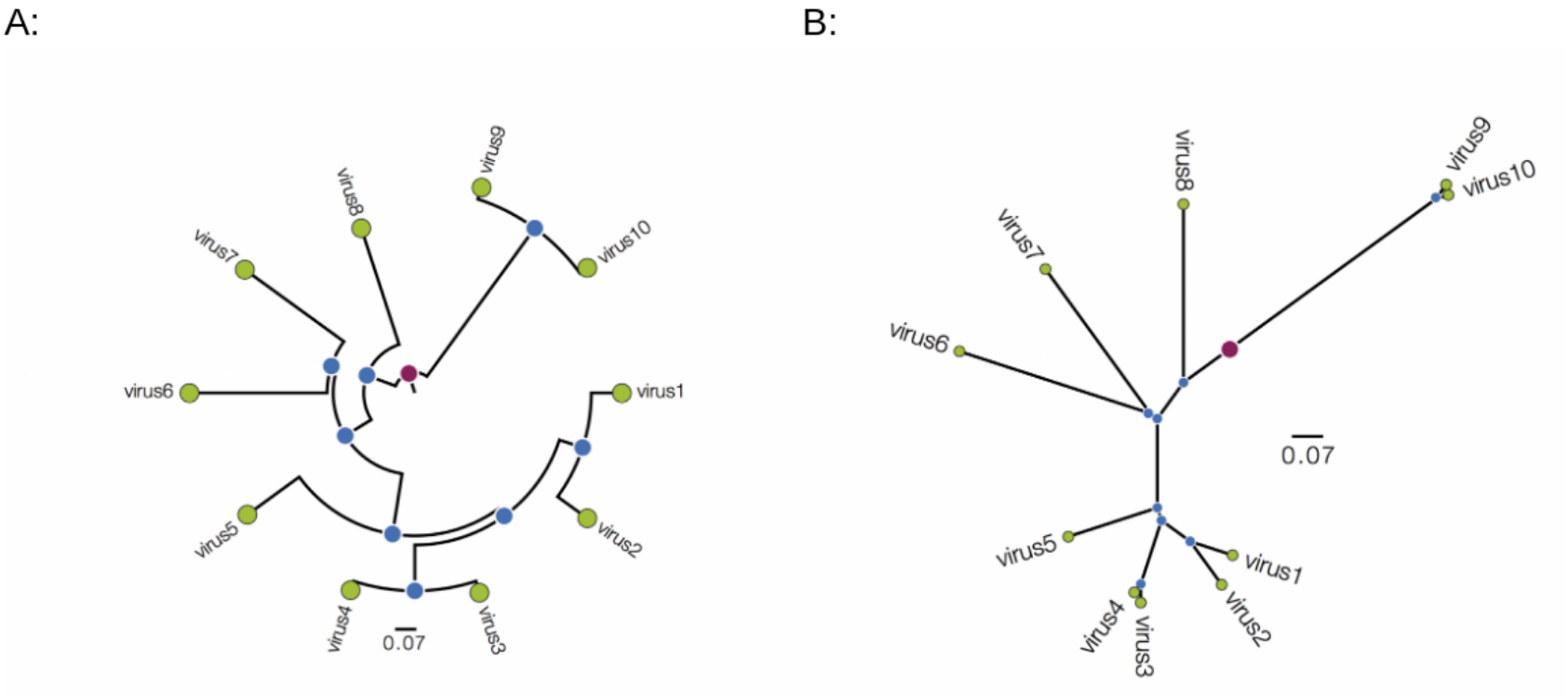
# Parts of the Tree



# Support Values

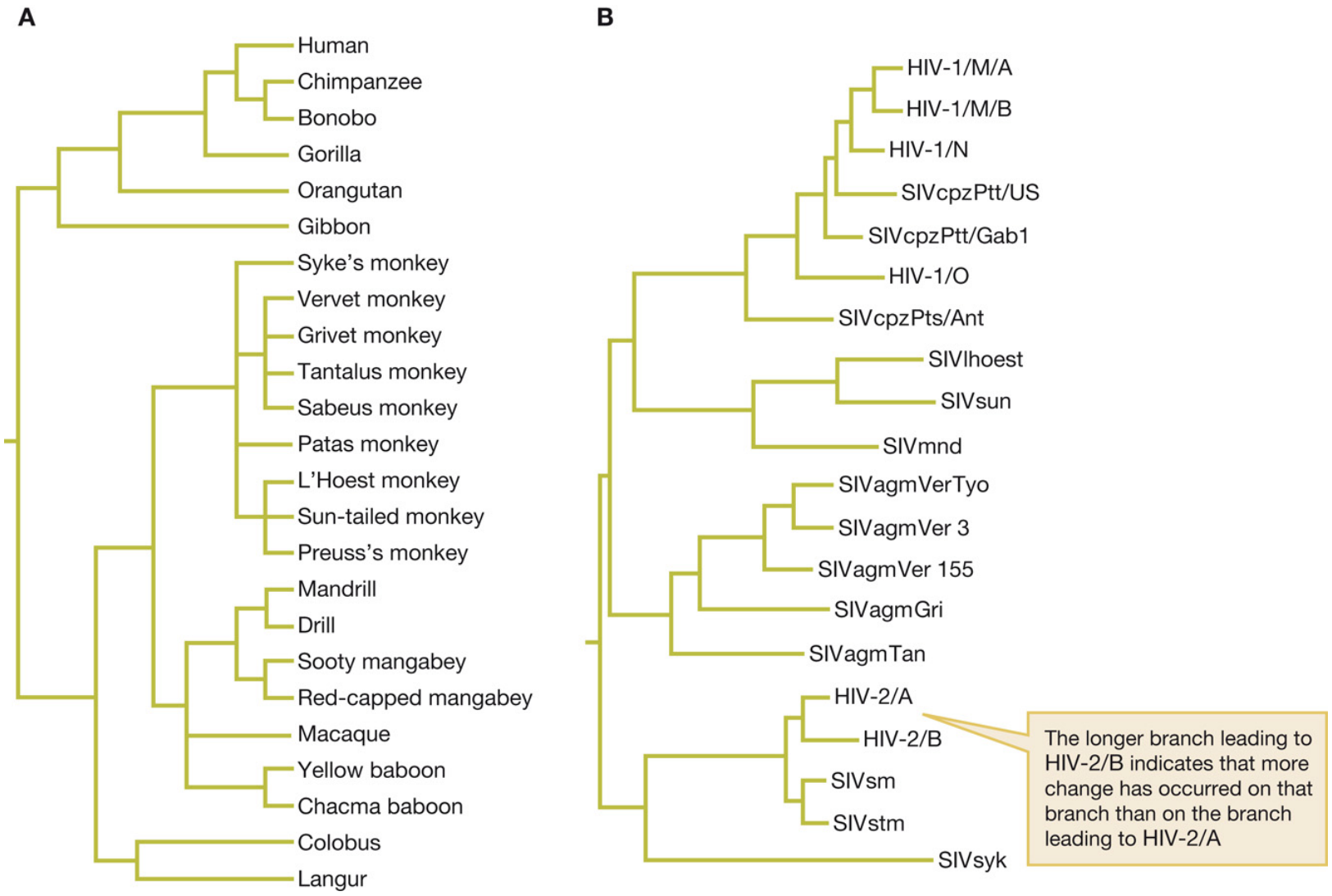


# Other formats

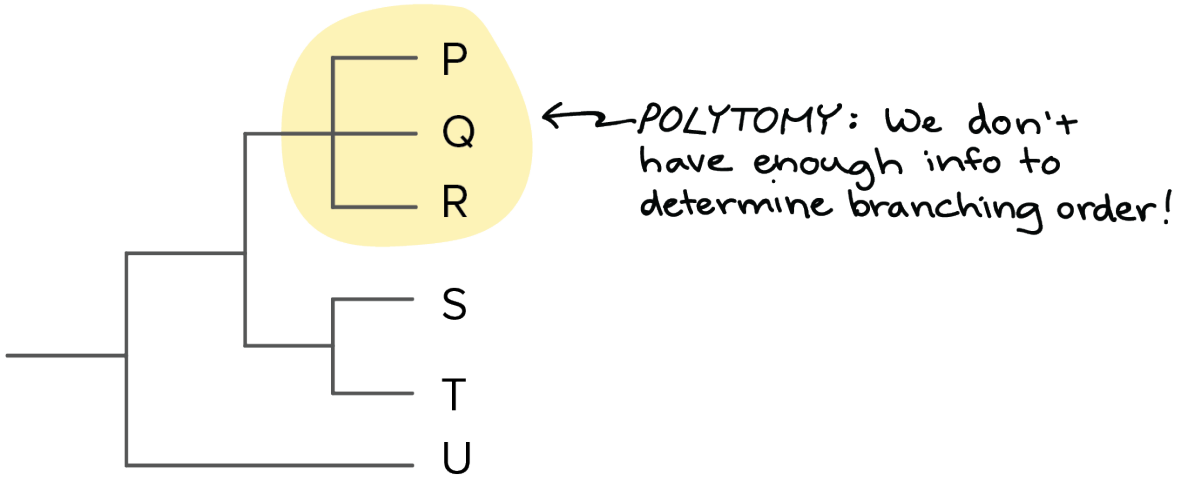



http://artic.network/how-to-read-a-tree.html

# Meaningful Branch Lengths

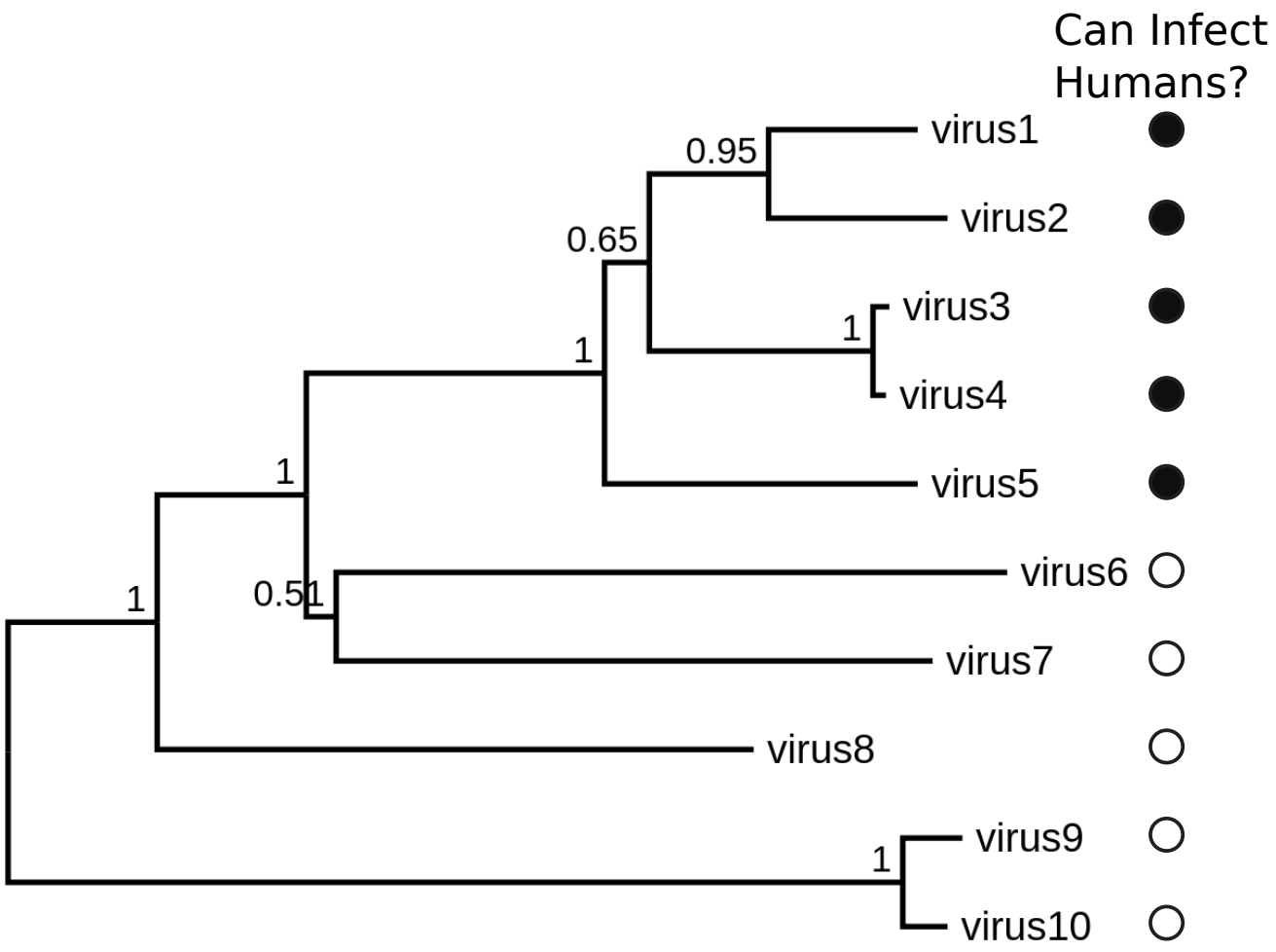



https://biology-forums.com/gallery/18099_27_04_12_2_16_20.jpeg

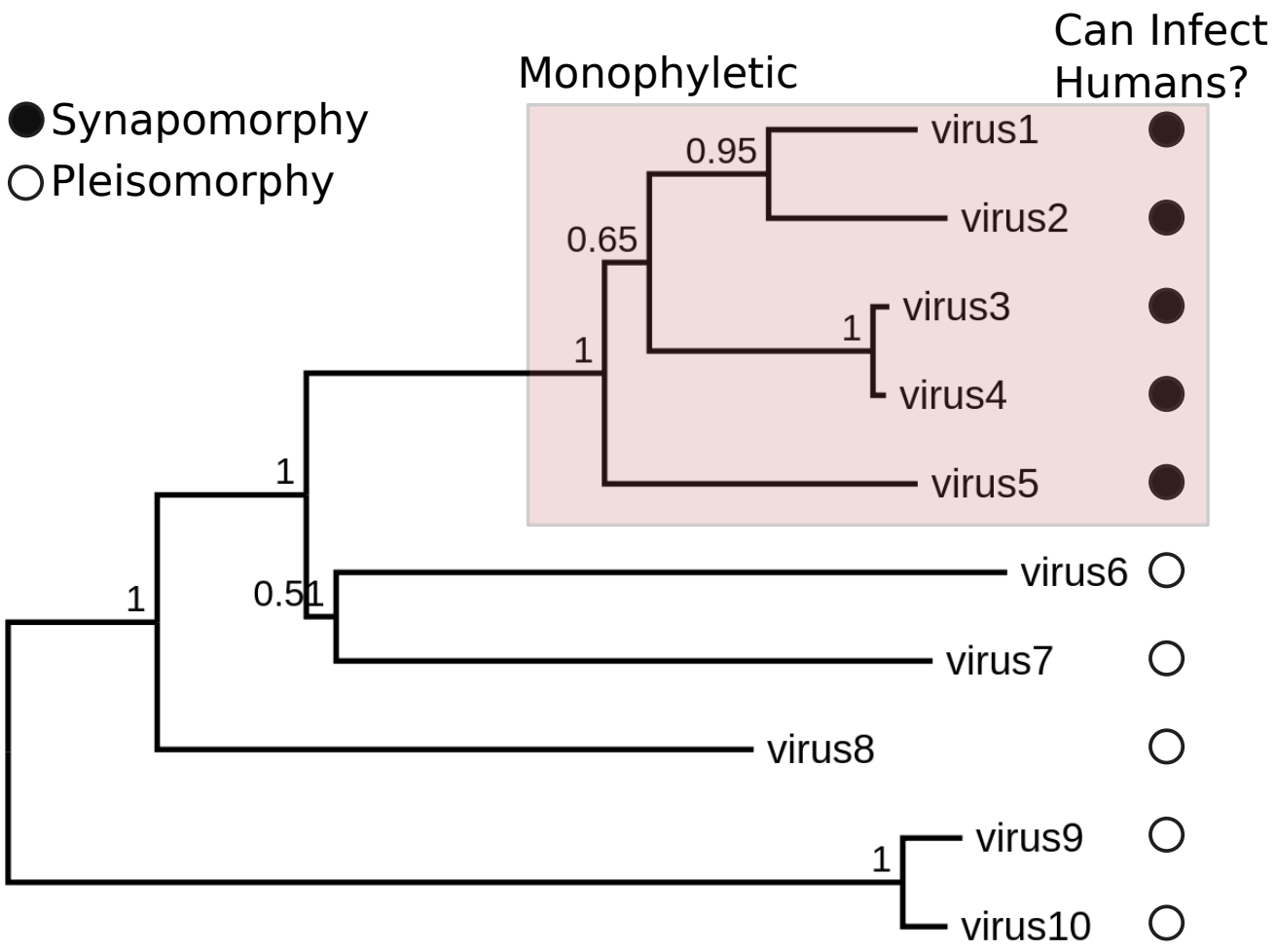



Khan Academy

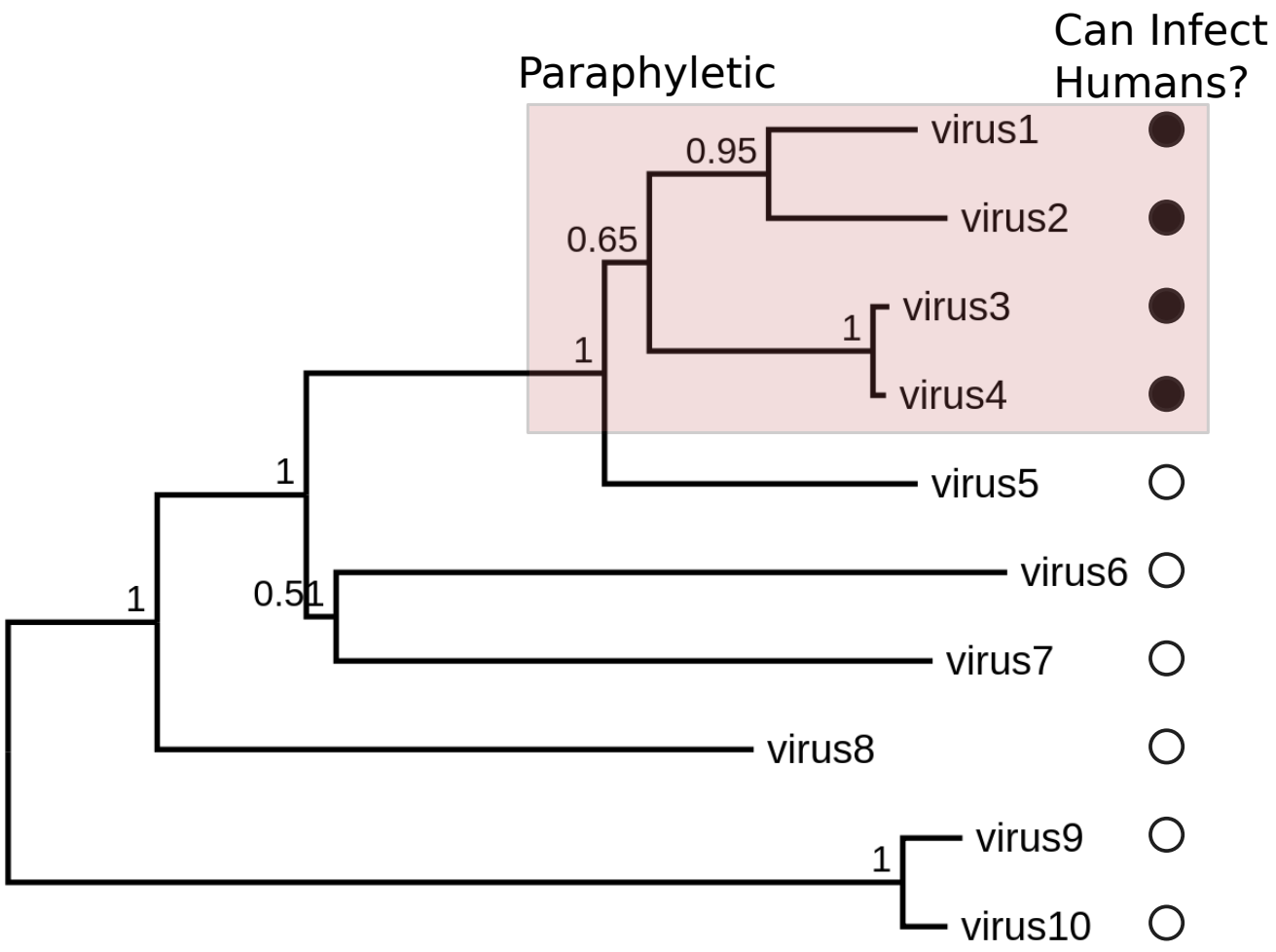
# Groupings on the Tree



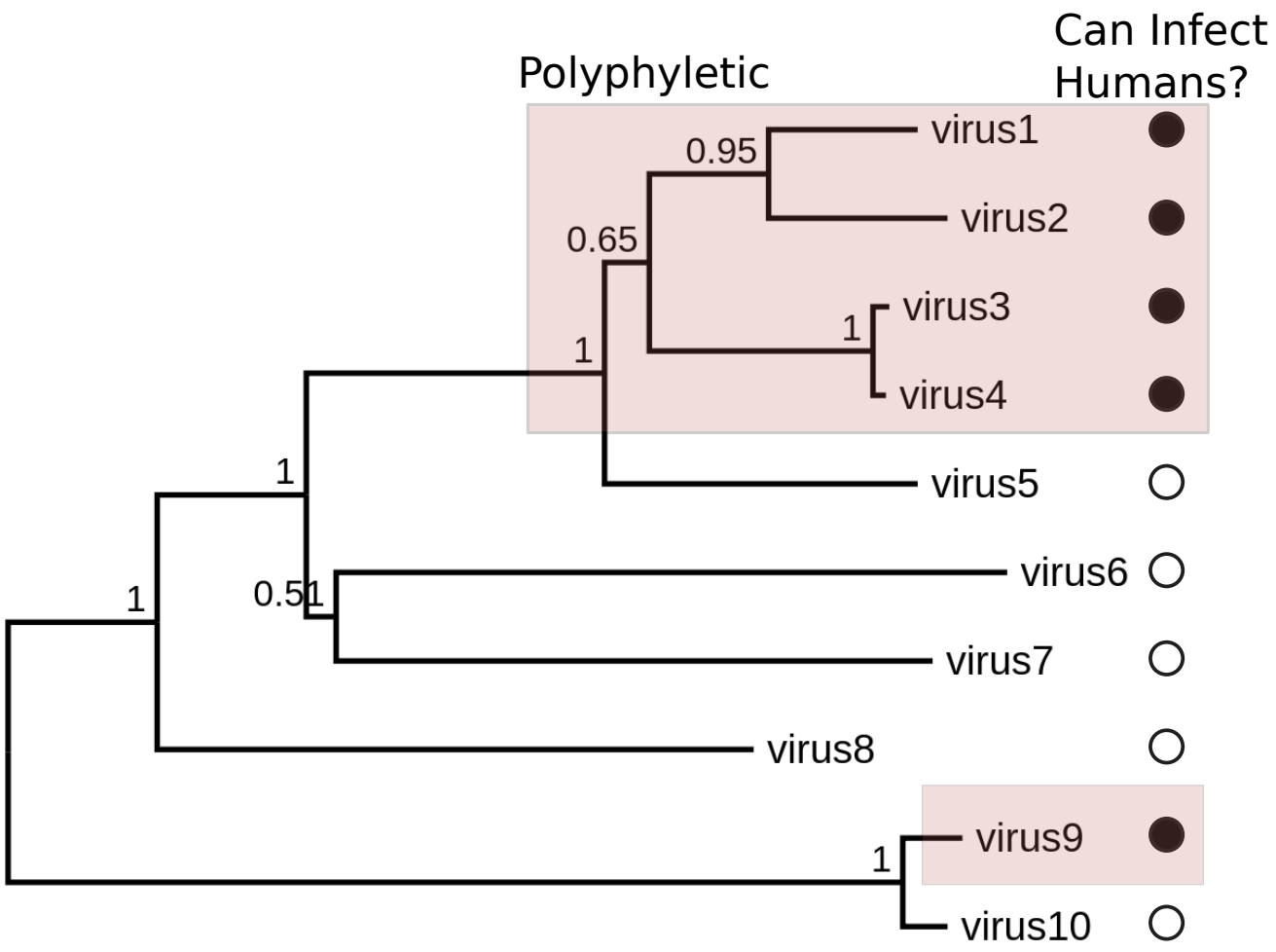
# Monophyletic



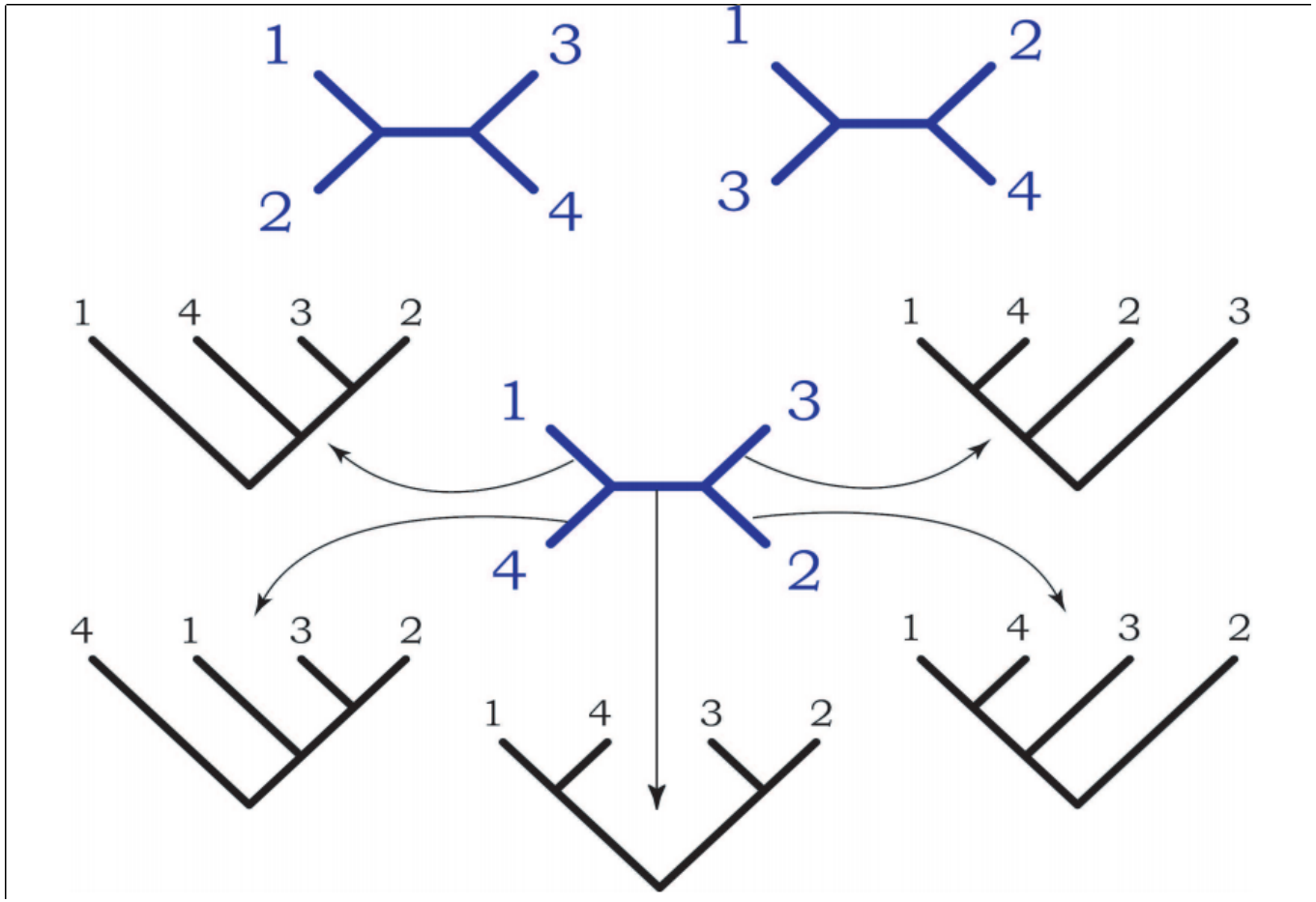
# Paraphyletic



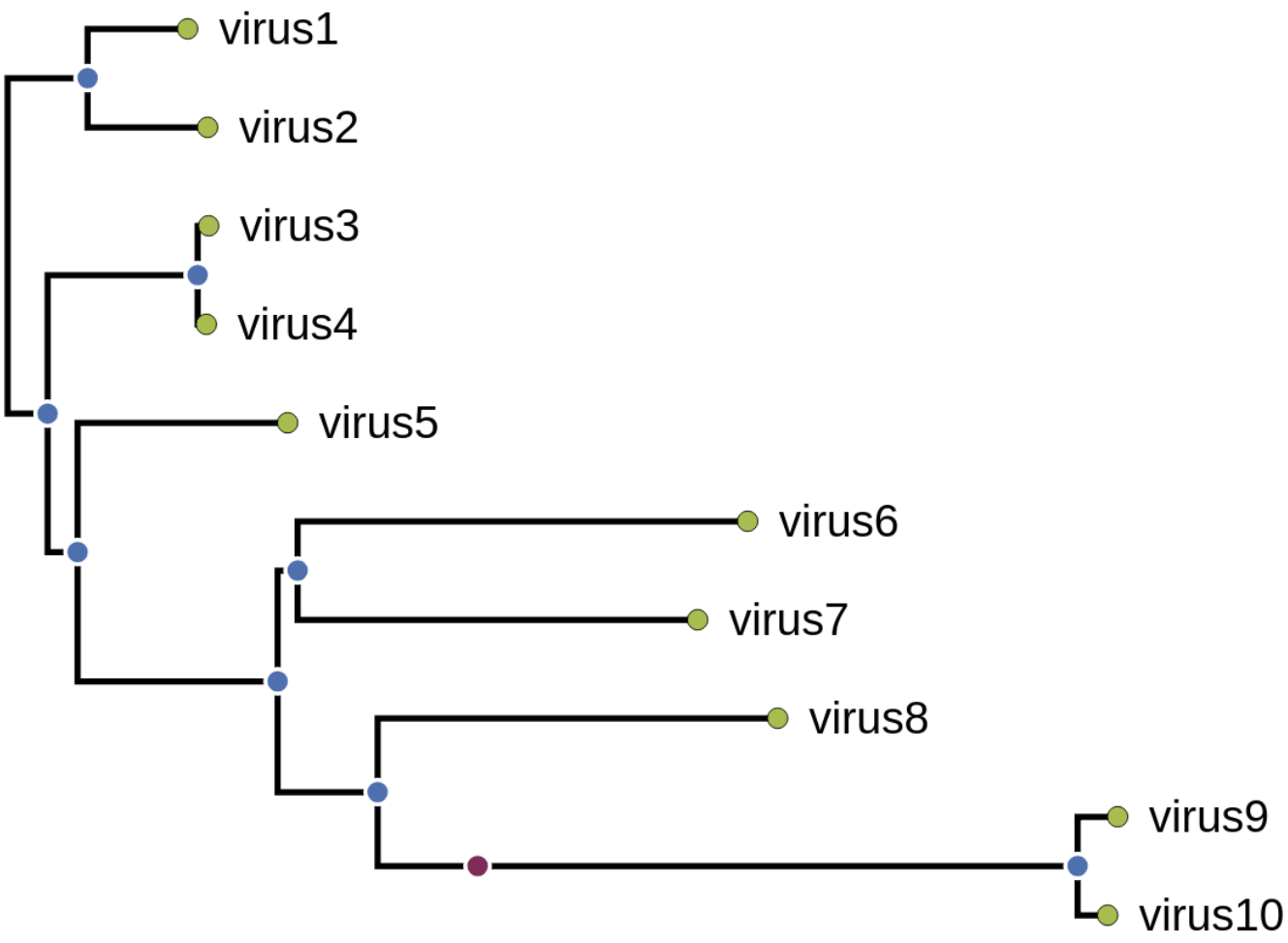
# Polyphyletic



# Rooting

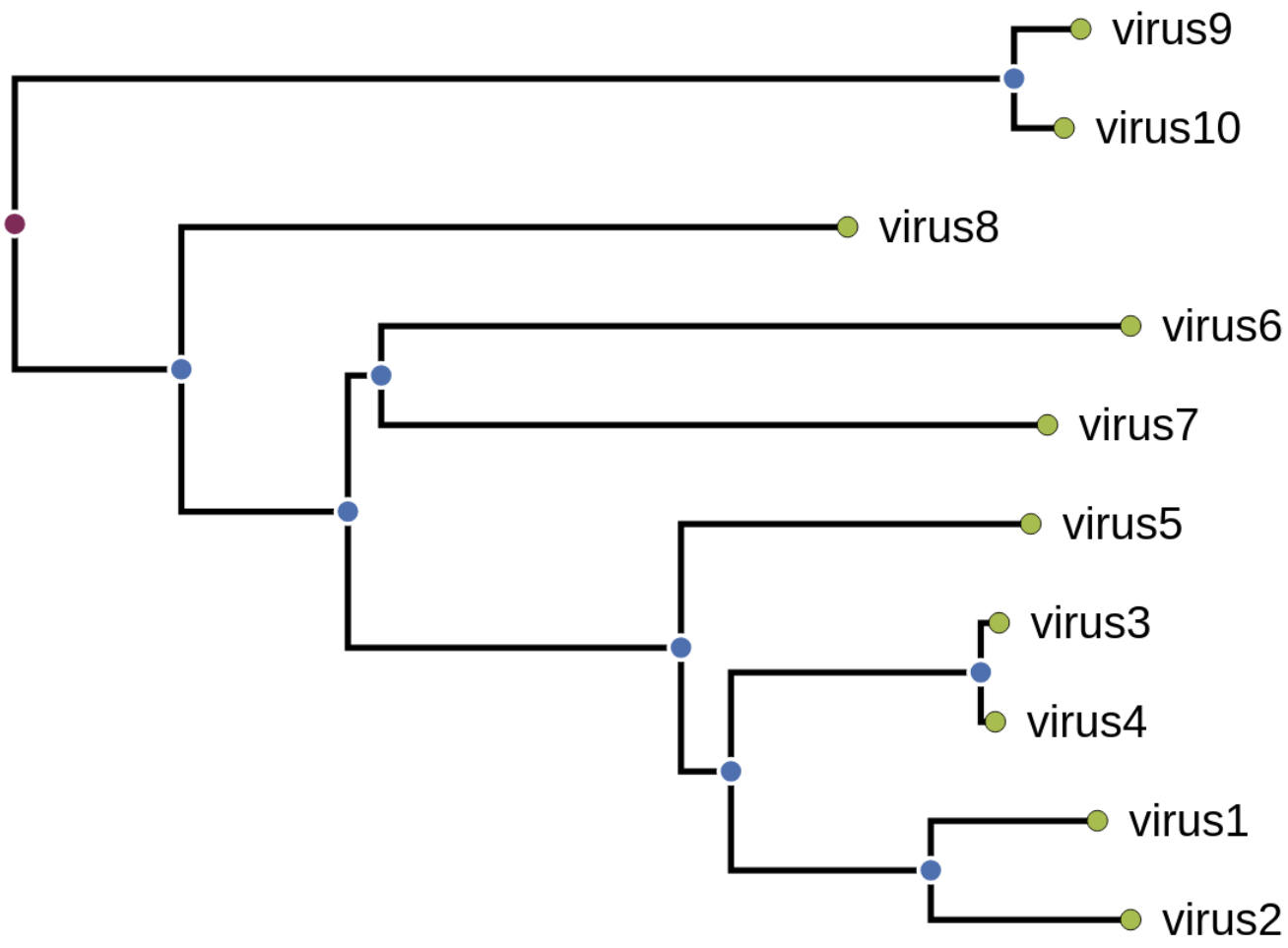


# Rooting



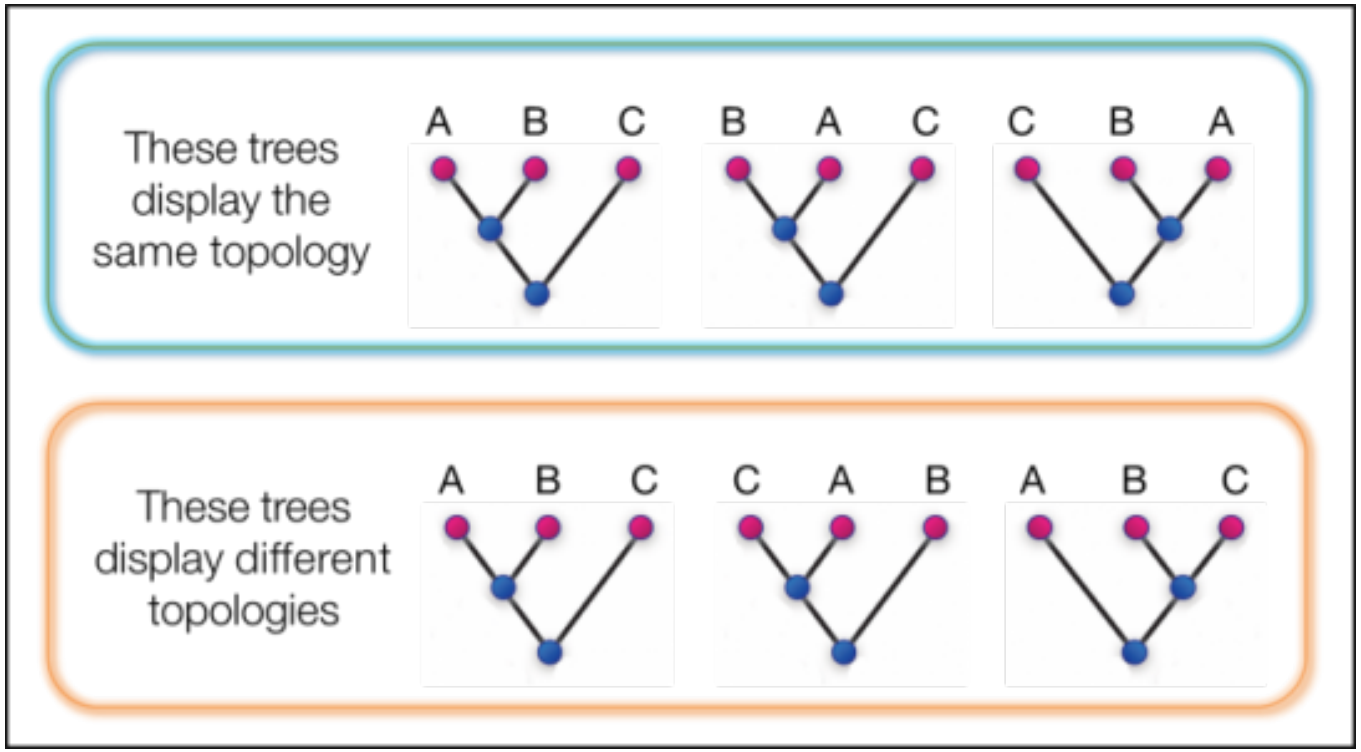

http://artic.network/how-to-read-a-tree.html

# Rooting

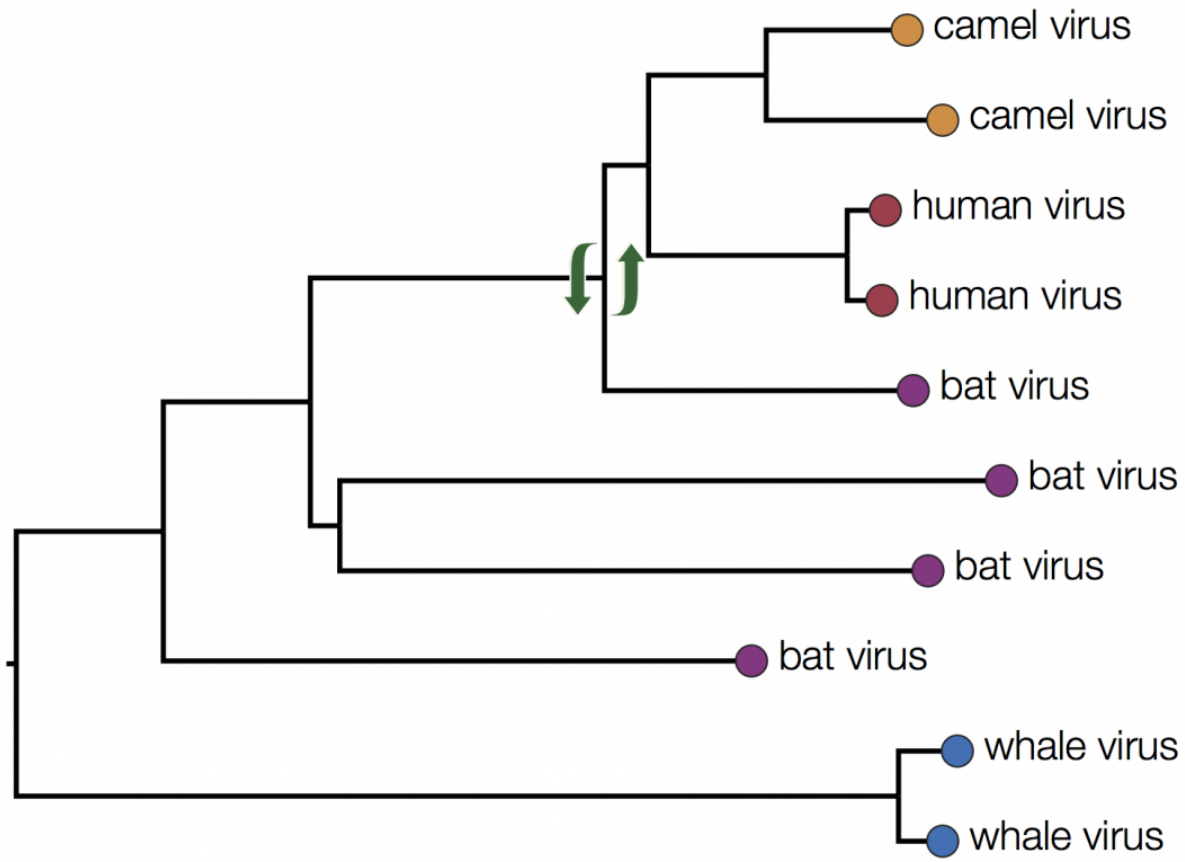



http://artic.network/how-to-read-a-tree.html

# Topology and rotation

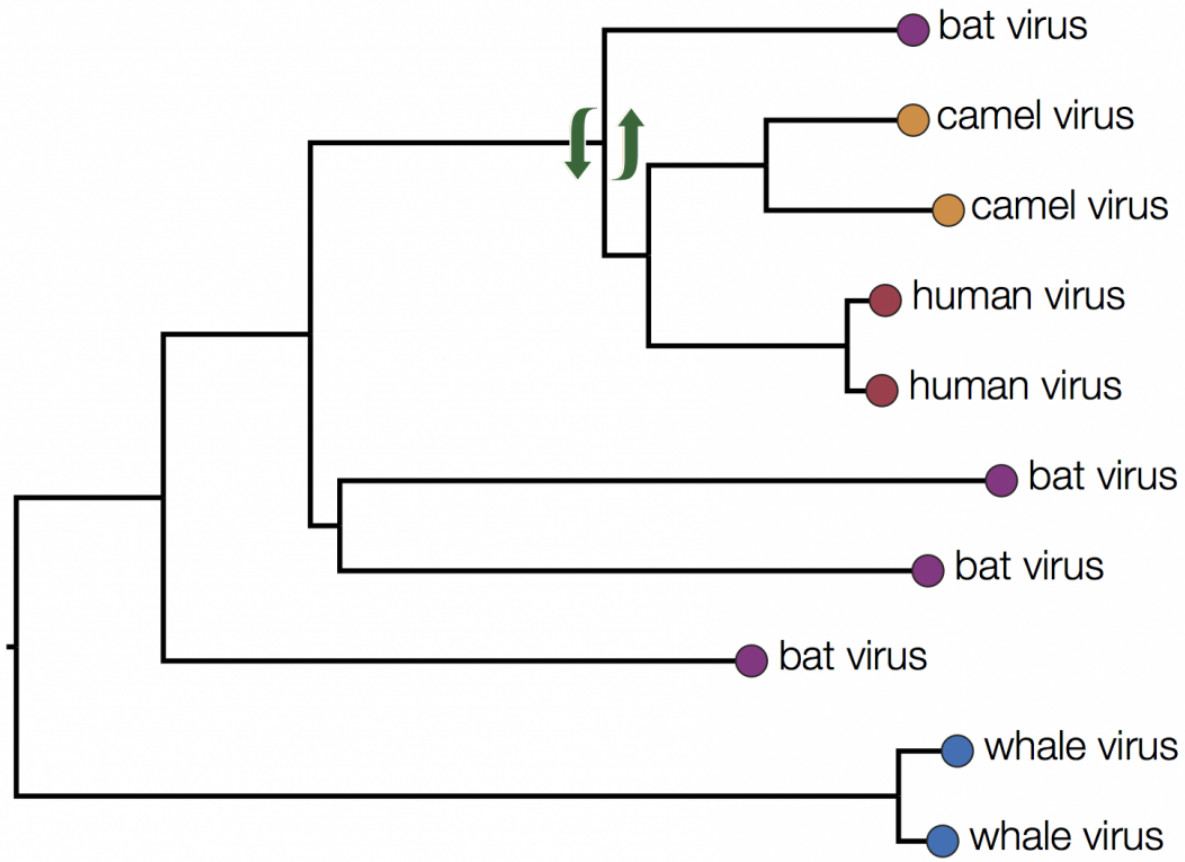


# Nodes can rotate



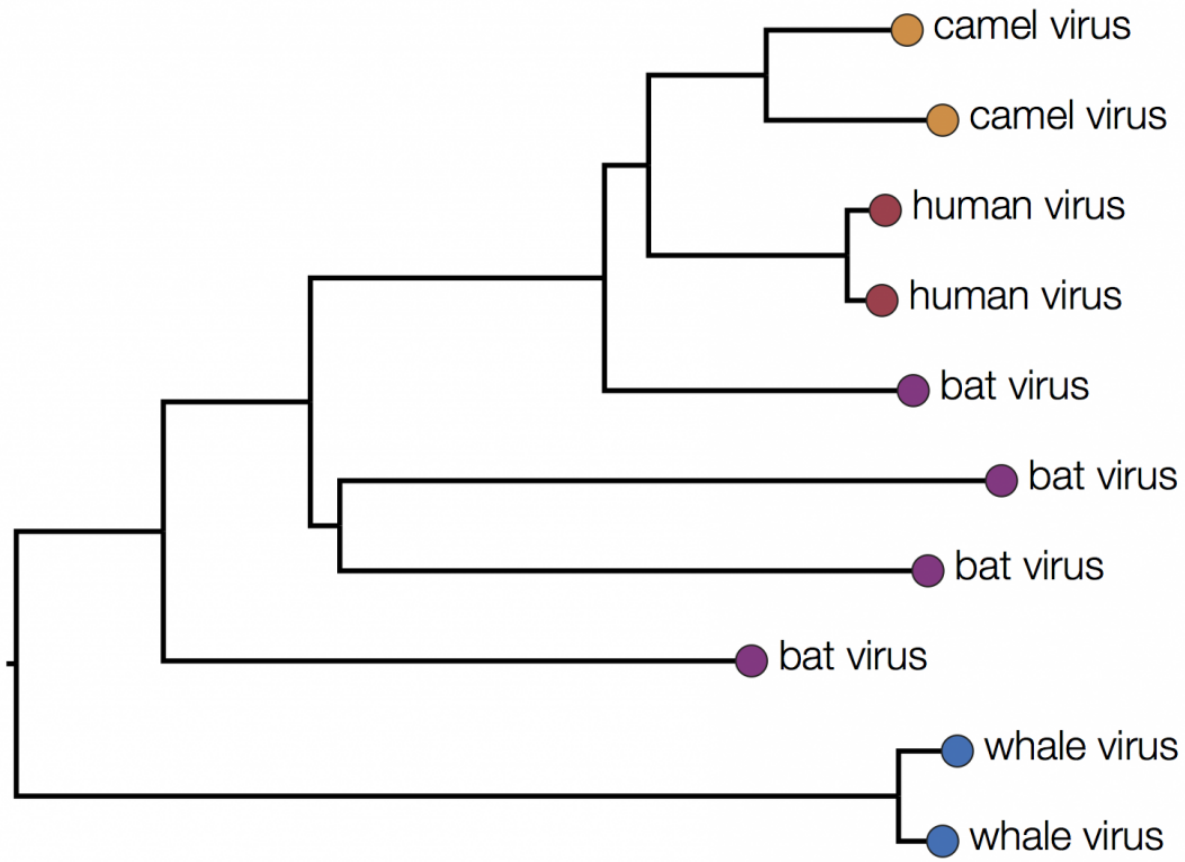

http://artic.network/how-to-read-a-tree.html

# Nodes can rotate



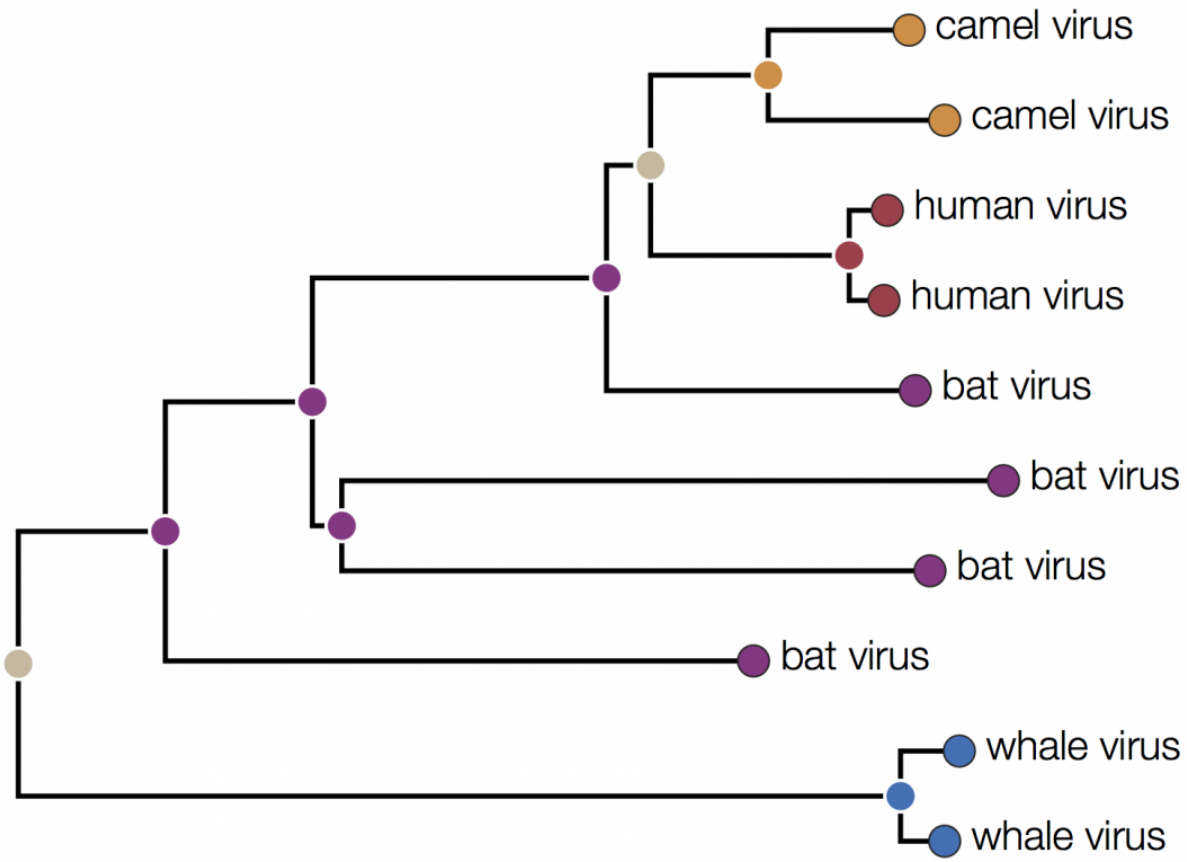

http://artic.network/how-to-read-a-tree.html

# Adding Metadata



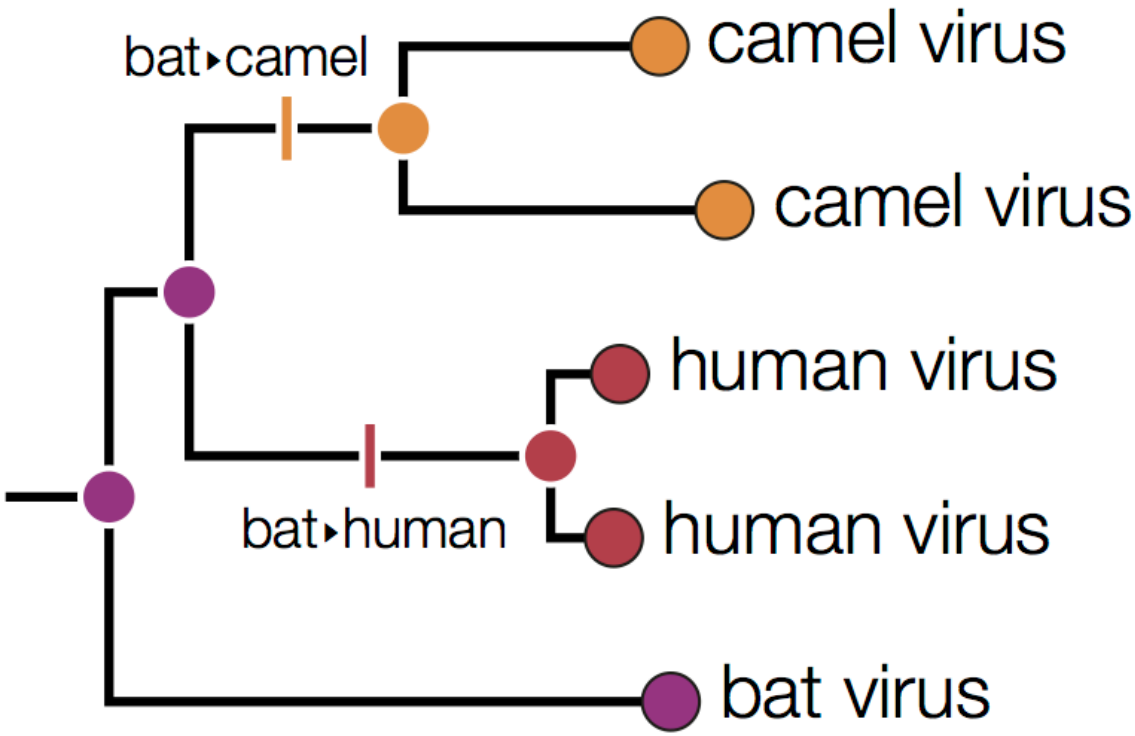

http://artic.network/how-to-read-a-tree.html

# Ancestral Node Reconstruction



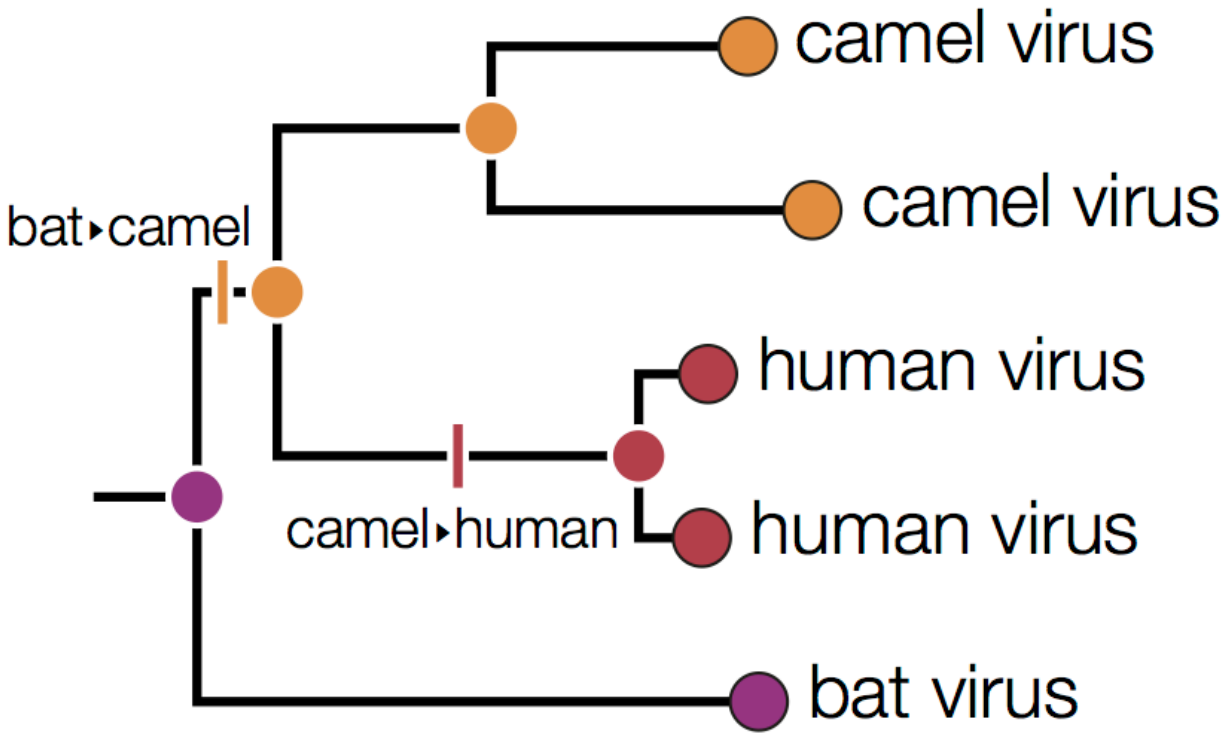

http://artic.network/how-to-read-a-tree.html

# Ancestral Node Reconstruction



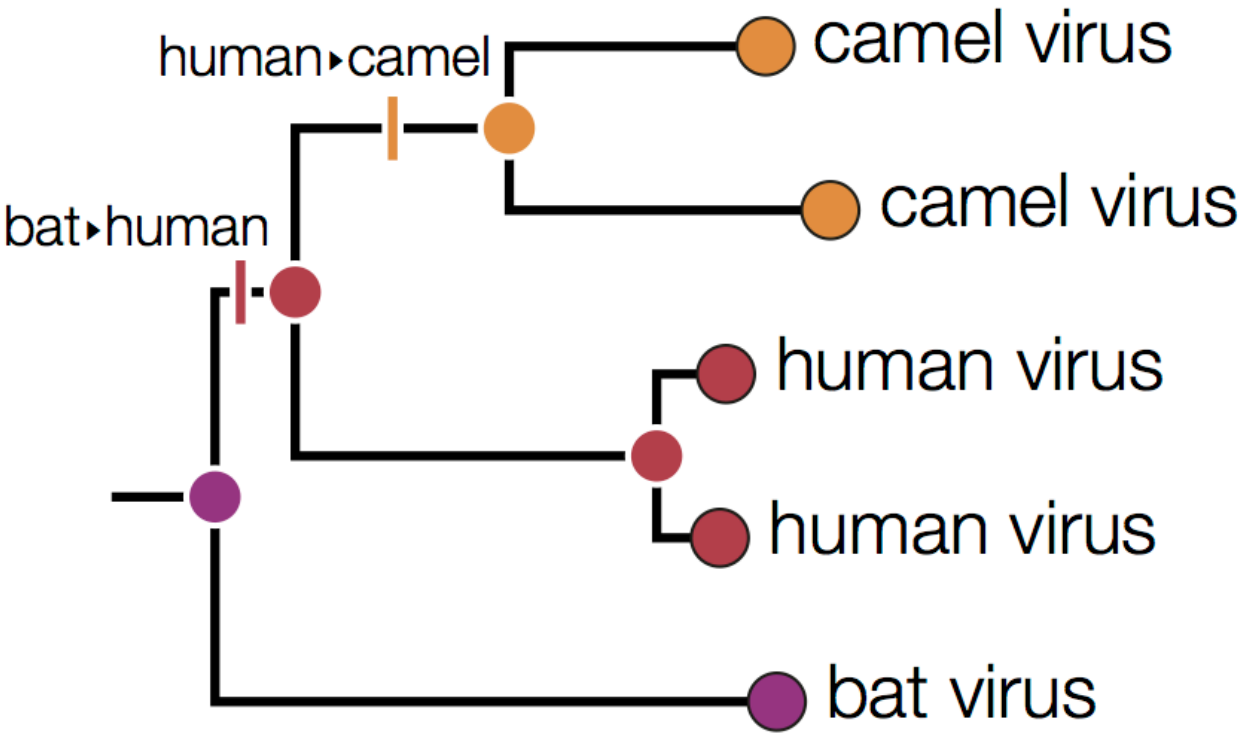

http://artic.network/how-to-read-a-tree.html

# Ancestral Node Reconstruction




http://artic.network/how-to-read-a-tree.html

# Ancestral Node Reconstruction




http://artic.network/how-to-read-a-tree.html

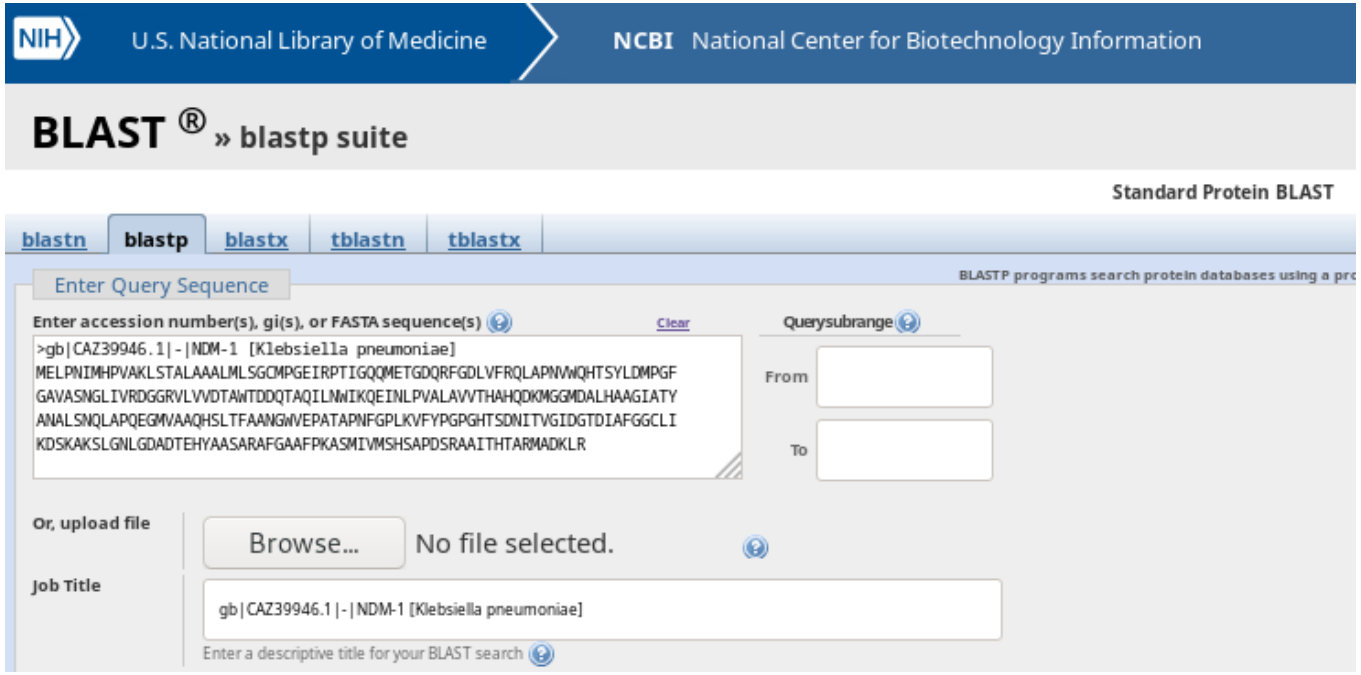
# Making a Tree

# Going from data to a tree

- Getting your data
- Aligning your data
- Tree-inference
  - Maximum Parsimony
  - Distance Methods
  - Maximum-Likelihood
  - Bayesian
- Sequence evolution models
- Exploring topology space
- Statistical support

# Getting and preparing your data

# Finding Similar Sequences



# BLAST

New Delhi metallo-beta-lactamase 1 [Acinetobacter baumannii]

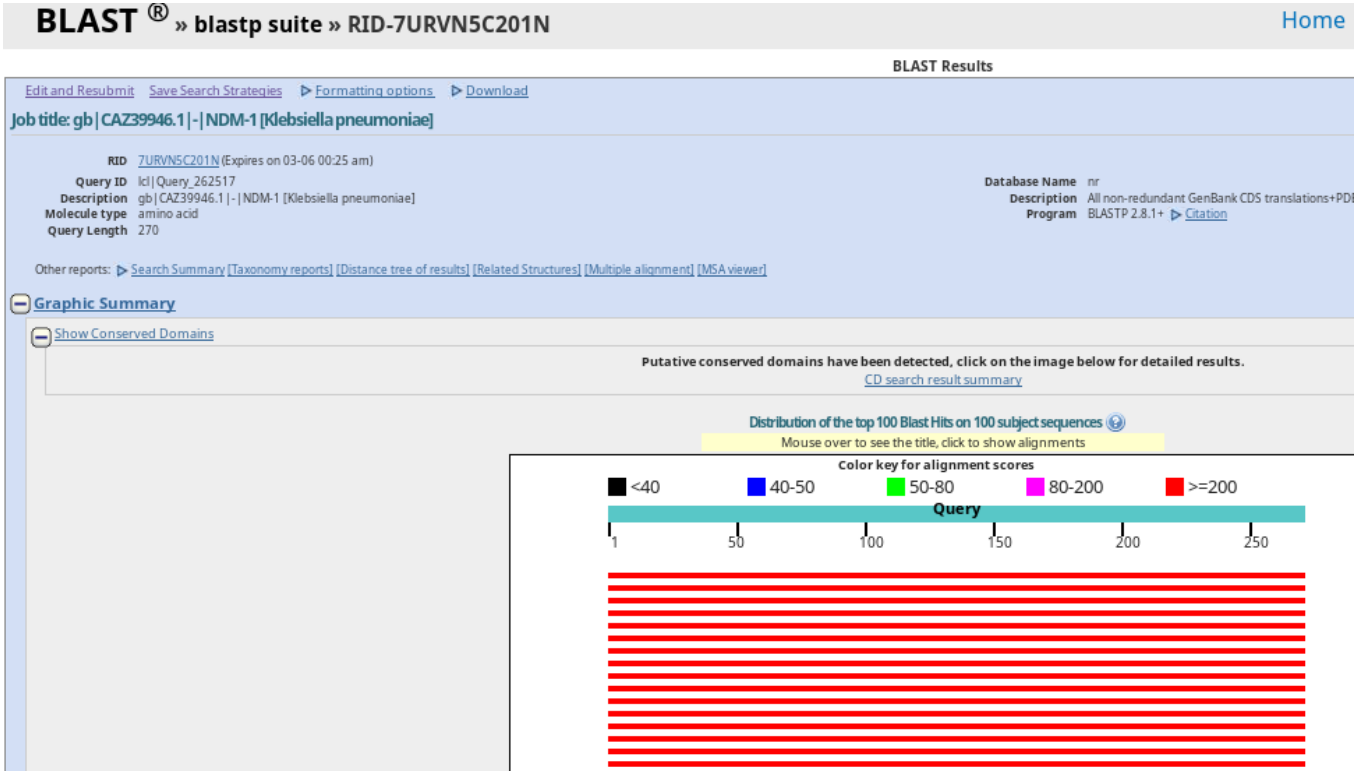
Sequence ID: BBA83870.1 Length: 261 Number of Matches: 1

Range 1: 3 to 260 GenPept Graphics ▼ Next Match ▲ Previous Match

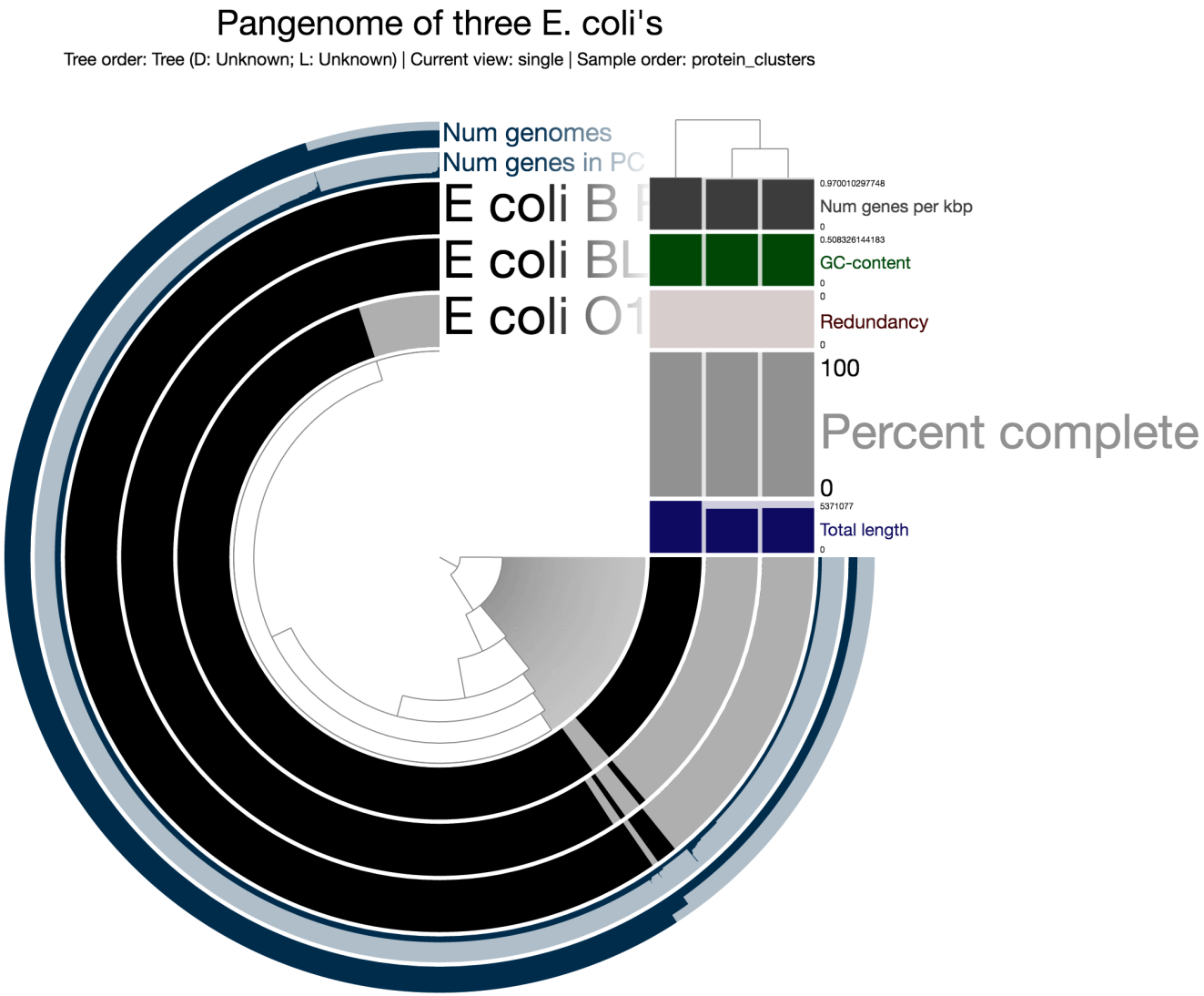
| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 498 bits(1282) | 2e-177 | Compositional matrix adjust. | 256/265(97%) | 257/265(96%) | 7/265(2%) |

```
Query  4    PNIMHPVAKLSTALAAALMLSGCMPGEIRPTIGQQMETGDQRFGDLVFRQLAPNVWQHTS  63
            PNIMHPVAKLSTALAAALMLSGCMPGEIRPTIGQQMETGDQRFGDLVFRQLAPNVWQHTS
Sbjct  3    PNIMHPVAKLSTALAAALMLSGCMPGEIRPTIGQQMETGDQRFGDLVFRQLAPNVWQHTS  62

Query  64   YLDMPGFGAVASNGLIVRDGGRVLVVDTAWTDDQTAQILNWIKQEINLPVALAVVTHAHQ  123
            YLDMPG GAVASNGLIVRDGGRV+V   AWTDDQTAQILNWIKQEINLPVALAVVTHAHQ
Sbjct  63   YLDMPG-GAVASNGLIVRDGGRVVV--DAWTDDQTAQILNWIKQEINLPVALAVVTHAHQ  119

Query  124  DKMGGMDALHAAGIATYANALSNQLAPQEGMVAAQHSLTFAANGWVEPATAPNFGPLKVF  183
            DKMGGMDALHAAGIATYANALSNQLAP EGMVAAQHSLTFAANGWVEPATAPNFGPLKVF
Sbjct  120  DKMGGMDALHAAGIATYANALSNQLAP-EGMVAAQHSLTFAANGWVEPATAPNFGPLKVF  178

Query  184  YPGPGHTSDNITVGIDGTDIAFGGCLIKDSKAKSLGNLGDADTEHYAASARAFGAAFPKA  243
            YPGPGHT DNITVGIDGTDIAFGGCLIKDSKAKSLGNLGDADTEHYAASARAFGAA PKA
Sbjct  179  YPGPGHT-DNITVGIDGTDIAFGGCLIKDSKAKSLGNLGDADTEHYAASARAFGAA-PKA  236

Query  244  SMIVMSHSAPDSRAAITHTARMADK  268
            SMIVMSH APDSRAAITHTARMADK
Sbjct  237  SMIVMSH-APDSRAAITHTARMADK  260
```

# BLAST

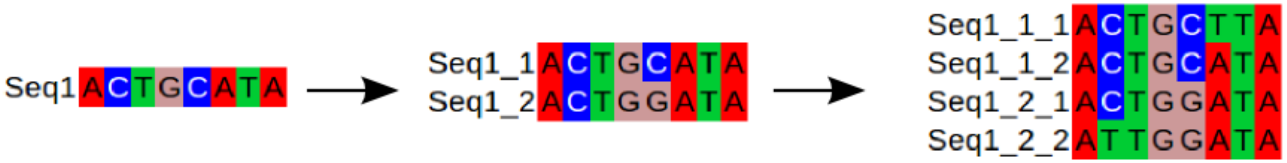

BLAST ® » blastp suite » RID-7URVN5C201N
Home
BLAST Results
Edit and Resubmit
Save Search Strategies
Formatting options
Download
Job title: gb|CAZ39946.1|-|NDM-1 [Klebsiella pneumoniae]
RID 7URVN5C201N (Expires on 03-06 00:25 am)
Query ID lcl|Query_262517
Description gb|CAZ39946.1|-|NDM-1 [Klebsiella pneumoniae]
Molecule type amino acid
Query Length 270
Database Name nr
Description All non-redundant GenBank CDS translations+PDB
Program BLASTP 2.8.1+ Citation
Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment] [MSA viewer]
Graphic Summary
Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
CD search result summary
Distribution of the top 100 Blast Hits on 100 subject sequences
Mouse over to see the title, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
50
100
150
200
250

# Core Genome Inference




anvi'o documentation

# Multiple Sequence Alignment

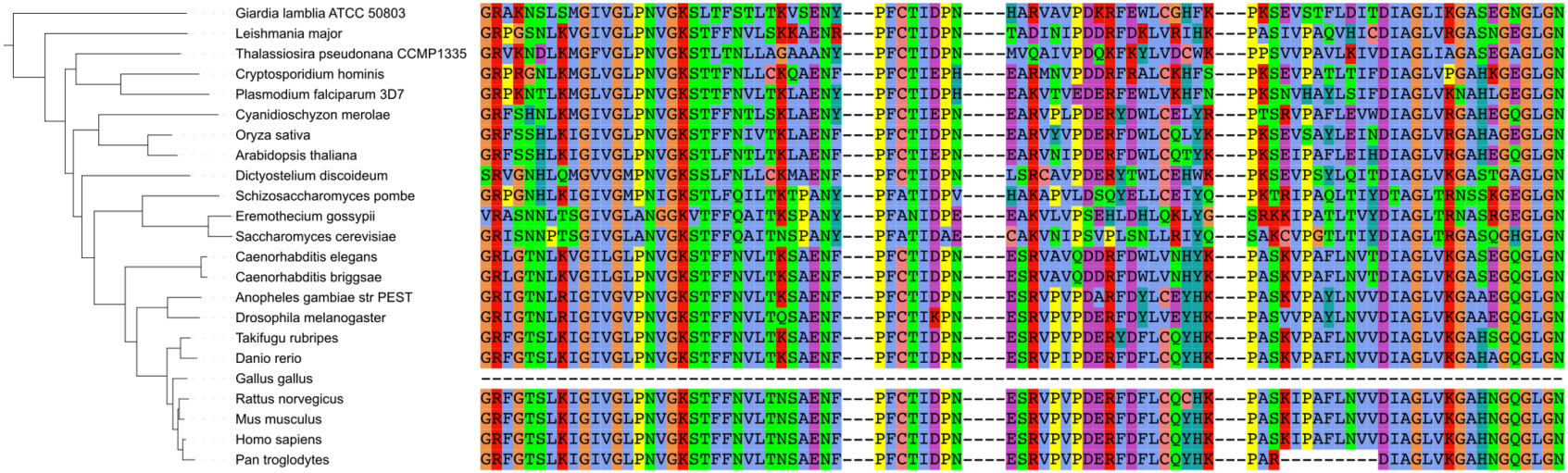



https://bioinf.comav.upv.es/courses/biotech3/theory/multiple.html

# Multiple Sequence Alignment


https://bioinf.comav.upv.es/courses/biotech3/theory/multiple.html

# Alignment Trimming




https://itol.embl.de/help.cgi

# Trimmed Alignment

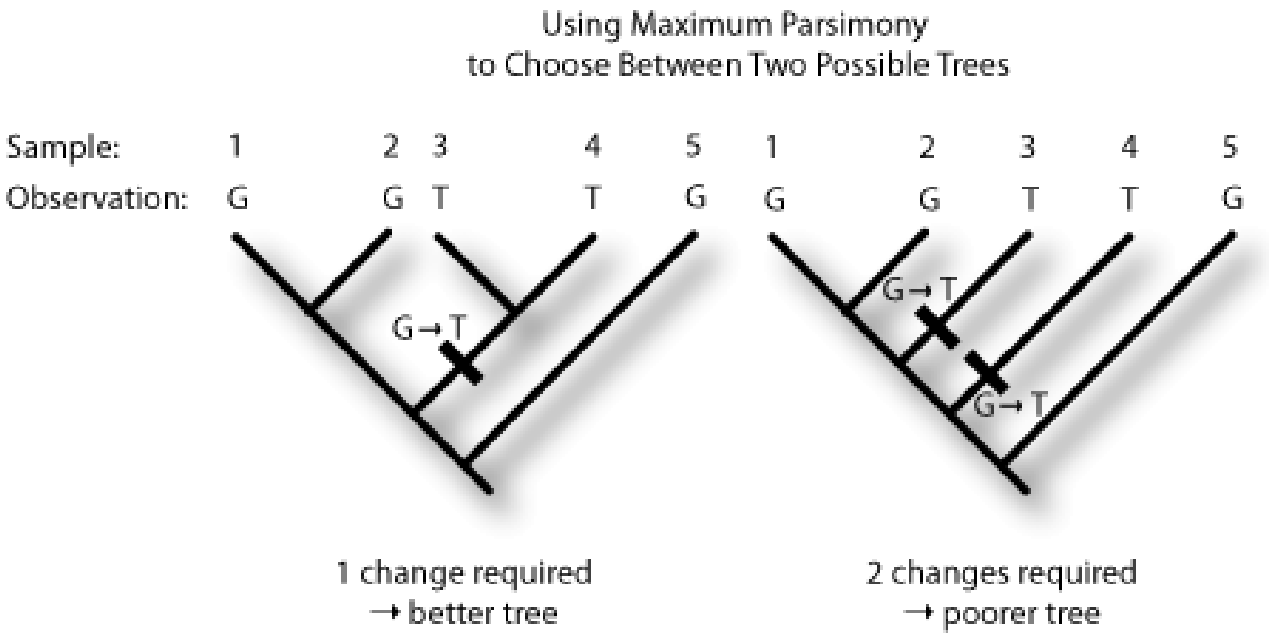

https://bioinf.comav.upv.es/courses/biotech3/theory/multiple.html

# Tree Inference methods

# Difficulties

- Huge number of possible trees
- unrooted = $\frac{2n-5!}{2^{n-3}(n-3)!}$
- rooted = $\frac{2n-3!}{2^{n-2}(n-2)!}$
- 10 taxa = 2,027,025 unrooted and 34,459,425 rooted topologies.
- 50 taxa = 2.84e74 unrooted and 2.75e76 rooted topologies.
- Topology space geometry is large and awkward to traverse.
- Large number of parameters to optimise
- How do you choose which tree is optimal? Which criterion?

# Parsimony



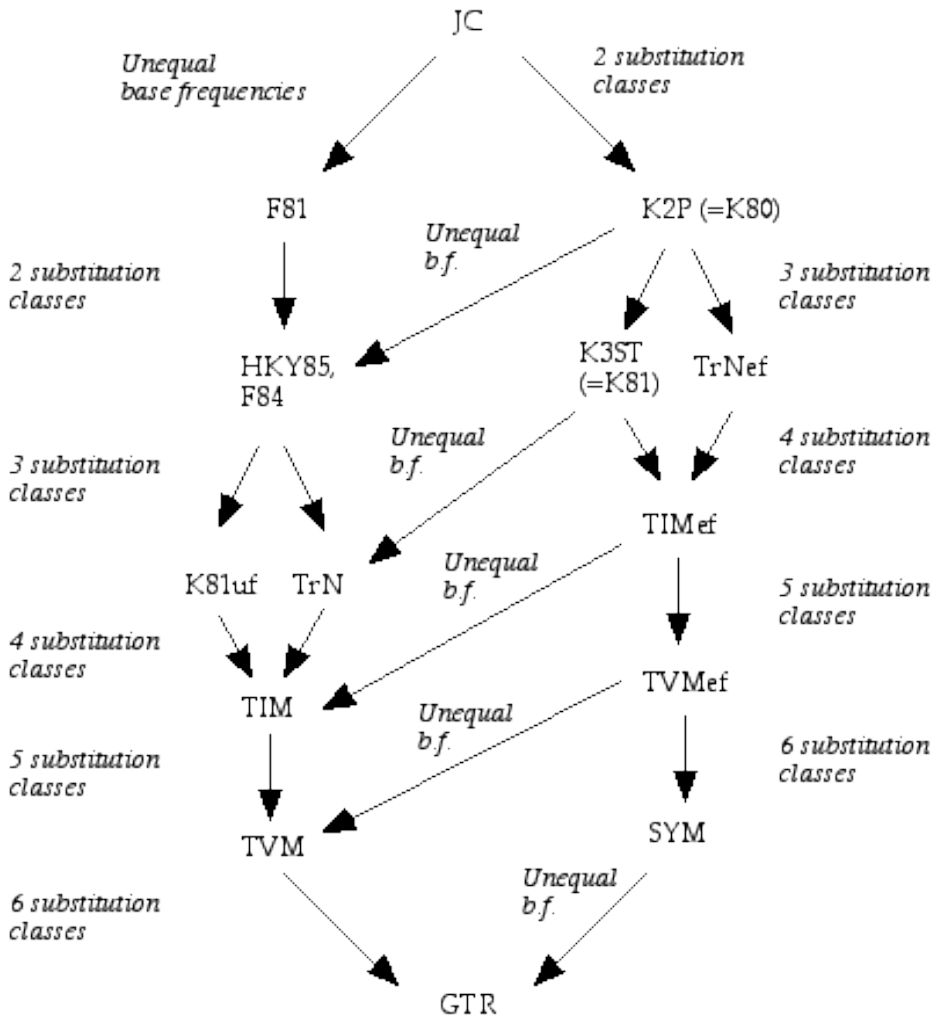

Intuitive: minimise the number of changes needed.

# Parsimony Pros/Cons

- Advantages:
  - Very simple
  - Works on any type of data (no explicit model).
- Disadvantages:
  - Very simple
  - Requires informative sites with consistent signal.
  - Poor handling of multiple substitutions.
  - Can’t incorporate extra information.
  - Not consistent for certain tree shapes (misleading support values).

# Evolutionary Models

# Sequence Evolution Models




http://carrot.mcb.uconn.edu/~olgazh/bioinf2010/class24.html

# Sequence Evolution Models

| | HIV-$W_m$ | HIV-$W_m$+F | HIV-$B_m$ | HIV-$B_m$+F | REV-1 step | JTT+F | JTT | WAG+F | MtMAM+F | rtREV | mtREV 24+F | WAG | Dayhoff+F | rtREV+F | Dayhoff | Equal Input | mtREV 24 | mtMAM | REV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-$W_m$ | 0 | 45 | 44 | 46 | 47 | 46 | 47 | 47 | 47 | 46 | 47 | 47 | 47 | 47 | 47 | 47 | 47 | 47 | 47 |
| HIV-$W_m$+F | 1 | 0 | 45 | 46 | 46 | 46 | 46 | 47 | 47 | 47 | 47 | 47 | 47 | 47 | 47 | 47 | 47 | 47 | 47 |
| HIV-$B_m$ | 0 | 1 | 0 | 15 | 43 | 30 | 39 | 43 | 46 | 46 | 46 | 46 | 46 | 47 | 47 | 47 | 47 | 47 | 47 |
| HIV-$B_m$+F | 0 | 0 | 15 | 0 | 43 | 37 | 40 | 44 | 47 | 46 | 47 | 46 | 47 | 47 | 47 | 47 | 47 | 47 | 47 |
| REV-1 step | 0 | 1 | 4 | 4 | 0 | 6 | 6 | 11 | 31 | 32 | 22 | 14 | 17 | 24 | 28 | 35 | 41 | 43 | 47 |
| JTT+F | 0 | 0 | 8 | 5 | 40 | 0 | 28 | 47 | 46 | 46 | 47 | 47 | 47 | 47 | 47 | 47 | 47 | 47 | 47 |
| JTT | 0 | 0 | 3 | 3 | 38 | 4 | 0 | 35 | 44 | 46 | 45 | 47 | 47 | 46 | 47 | 47 | 47 | 47 | 47 |
| WAG+F | 0 | 0 | 3 | 1 | 34 | 0 | 5 | 0 | 43 | 44 | 43 | 39 | 42 | 46 | 47 | 47 | 47 | 47 | 47 |
| MtMAM+F | 0 | 0 | 0 | 0 | 16 | 0 | 0 | 2 | 0 | 14 | 2 | 6 | 4 | 7 | 12 | 31 | 47 | 47 | 46 |
| rtREV | 0 | 0 | 0 | 0 | 12 | 0 | 1 | 2 | 29 | 0 | 8 | 1 | 3 | 3 | 4 | 39 | 47 | 47 | 47 |
| MtREV 24+F | 0 | 0 | 0 | 0 | 18 | 0 | 1 | 1 | 41 | 37 | 0 | 7 | 7 | 22 | 25 | 47 | 47 | 47 | 47 |
| WAG | 0 | 0 | 0 | 1 | 29 | 0 | 0 | 2 | 40 | 45 | 35 | 0 | 30 | 39 | 43 | 46 | 47 | 47 | 47 |
| Dayhoff+F | 0 | 0 | 0 | 0 | 26 | 0 | 0 | 0 | 39 | 43 | 29 | 8 | 0 | 36 | 43 | 46 | 47 | 47 | 47 |
| rtREV+F | 0 | 0 | 0 | 0 | 19 | 0 | 0 | 0 | 35 | 41 | 20 | 2 | 1 | 0 | 20 | 46 | 47 | 47 | 47 |
| Dayhoff | 0 | 0 | 0 | 0 | 18 | 0 | 0 | 0 | 32 | 39 | 17 | 0 | 1 | 17 | 0 | 44 | 47 | 47 | 47 |
| Equal Input | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 14 | 2 | 0 | 1 | 0 | 1 | 2 | 0 | 41 | 46 | 47 |
| mtREV 24 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 43 | 45 |
| mtMAM | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 44 |
| REV | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 0 |

[1]Models are arranged by decreasing rank performance (see Table 2)
doi:10.1371/journal.pone.0000503.t003

[Nickle et al., 2007]

# How do we select a model?

- Which model is most likely given the data?
- Information Criterion (regularisation to penalise overly complex models)
- Decision Theory: risk minimisation.

# What happens theoretically if the wrong model is specified?

- Increased Inaccuracy (wrong tree more often)
- Inconsistency (adding more data converges to wrong tree)
- Wrong branch lengths (important for certain analyses)
- Wrong tree support values

# What actually happens

[Abadi et al., 2019]

- Almost always use the most flexible model (GTR+I+G/LG)
- Criteria are inconsistent (BIC/AIC disagree in 62% of cases)
- Different models change the distance matrix trivially.
- ALL models lead to very similar topologies.
- Model only really important if branch length matters to you.

# Distance Matrix

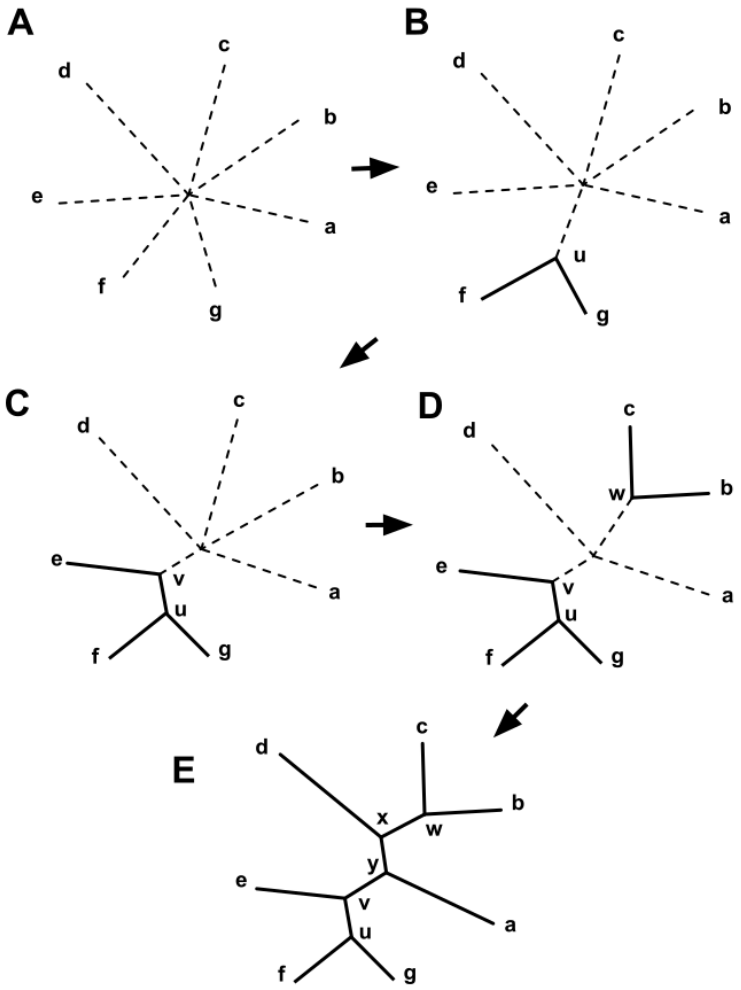
```
1   AGCCCGTCGC...
2   AGCTCGTCCC...
3   GGCTCGACCC...
⋮       ⋮
4   AGCCGGATCC...
```

$$\begin{array}{c|c} & \begin{array}{cccc} 1 & 2 & \cdots & n \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ n \end{array} & \begin{pmatrix} 0 & 0.3 & \cdots & 0.5 \\ 0.3 & 0 & \cdots & 0.3 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.4 & \cdots & 0 \end{pmatrix} \end{array}$$

https://slideplayer.com/slide/4422868/

# Neighbour-Joining

Iteratively pair off branches that minimise the total sum of branch lengths



# Distance Approaches Pros/Cons

- Advantages:
  - Very fast (often used as starting point)
  - Works well for clock-like and closely related sequences
- Disadvantages:
  - Requires a sequence evolution model
  - Pairwise distance isn't always error-free estimate of evolutionary distance (bigger problem with divergent sequences).
  - Doesn't use all available information
  - Cannot reconstruct character histories

# Aside: sources of error

# Sources of Error

- Bad data

# Sources of Error

- Bad data
- Sampling error

# Sources of Error

- Bad data
- Sampling error
- Misleading evolutionary events

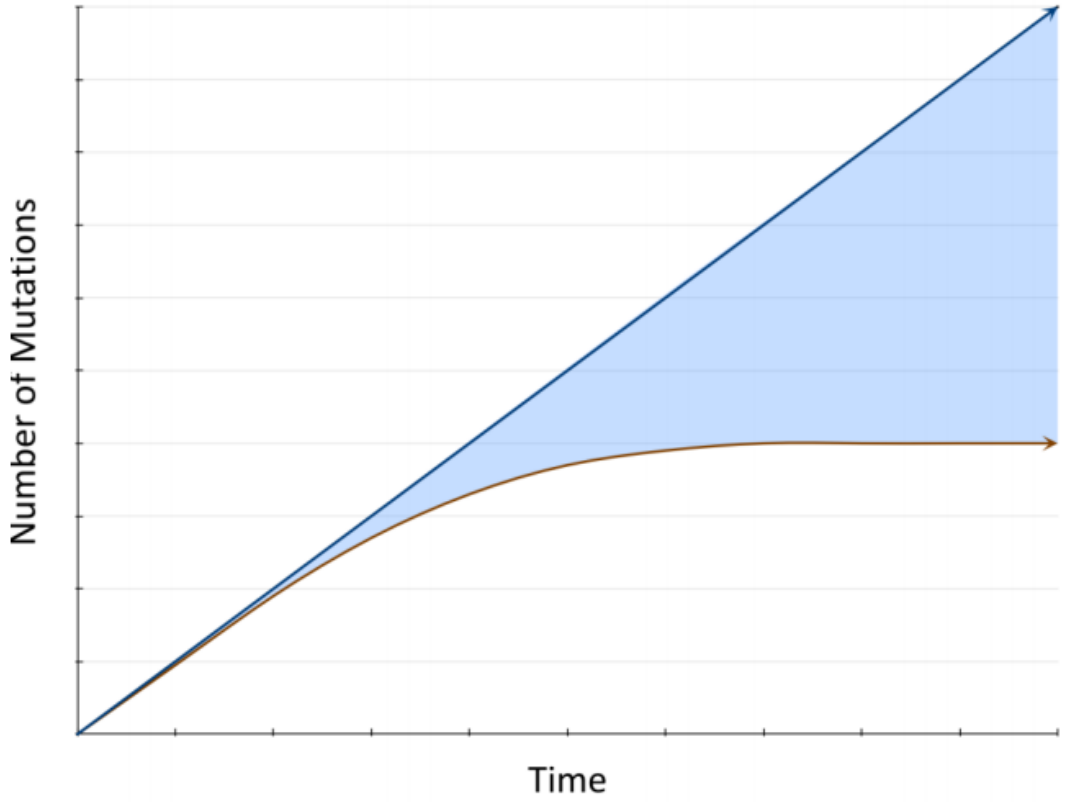
# Sources of Error

- Bad data
- Sampling error
- Misleading evolutionary events
- Misspecified models

# Sources of Error

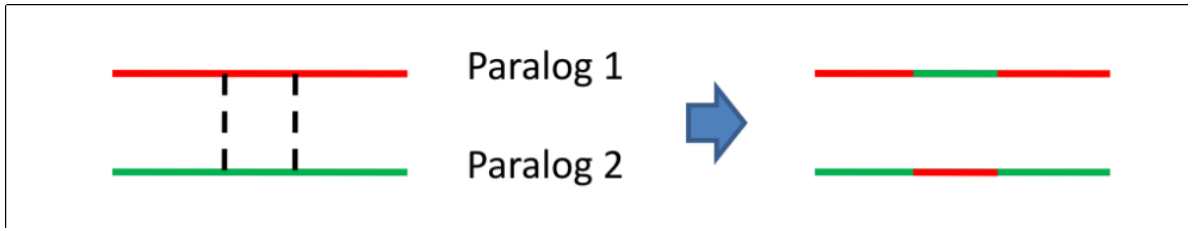
- Bad data
- Sampling error
- Misleading evolutionary events
- Misspecified models
- Inappropriate inference

# Saturation

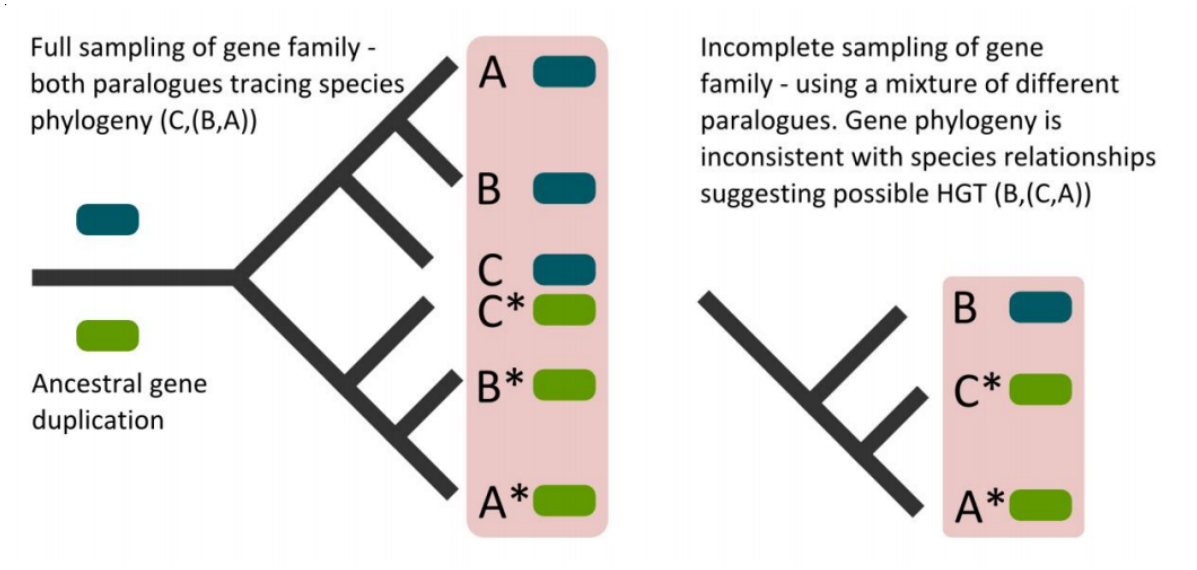



[Leonard, 2010]

# Misleading Signal: Recombination

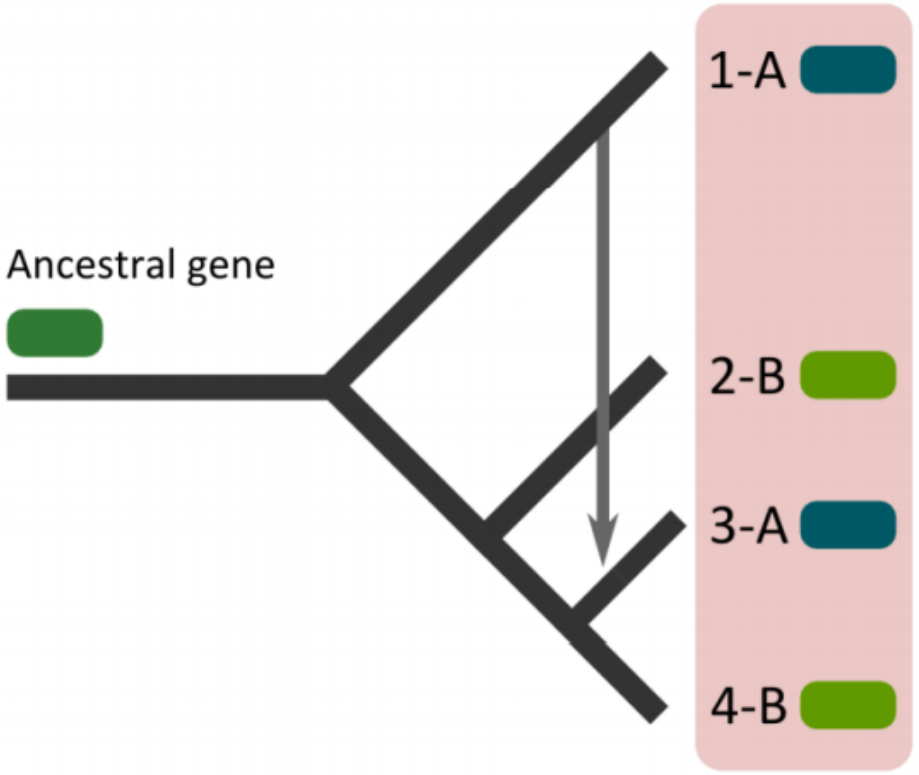


# Misleading Signal: Hidden Paralogy/Incomplete Sampling



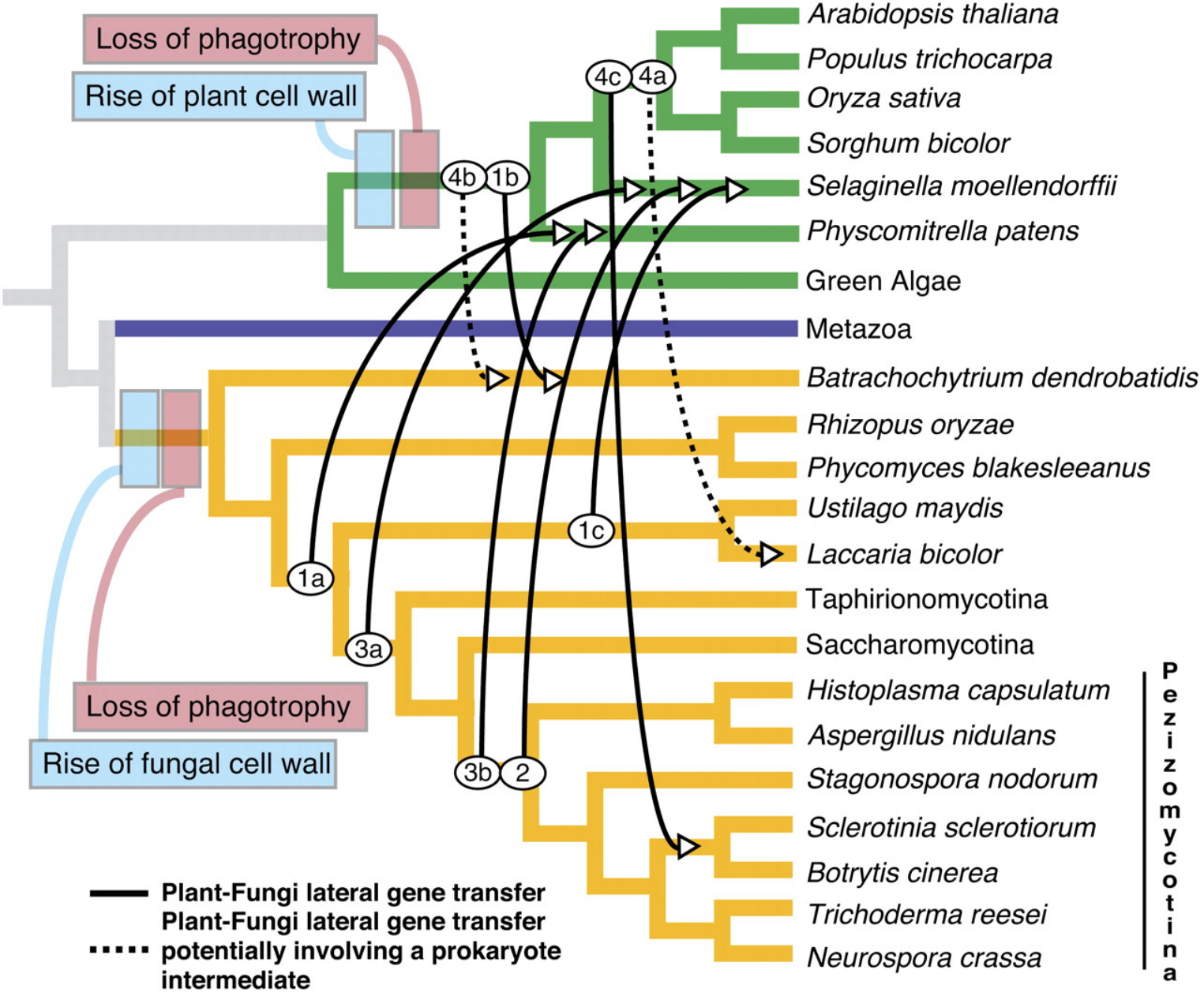

[Leonard, 2010]

# Misleading Signal: Horizontal Gene Transfer




[Leonard, 2010]

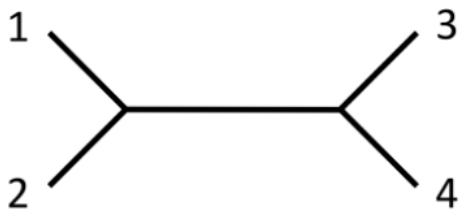
# Misleading Signal: Horizontal Gene Transfer



# Tree not always correct paradigm

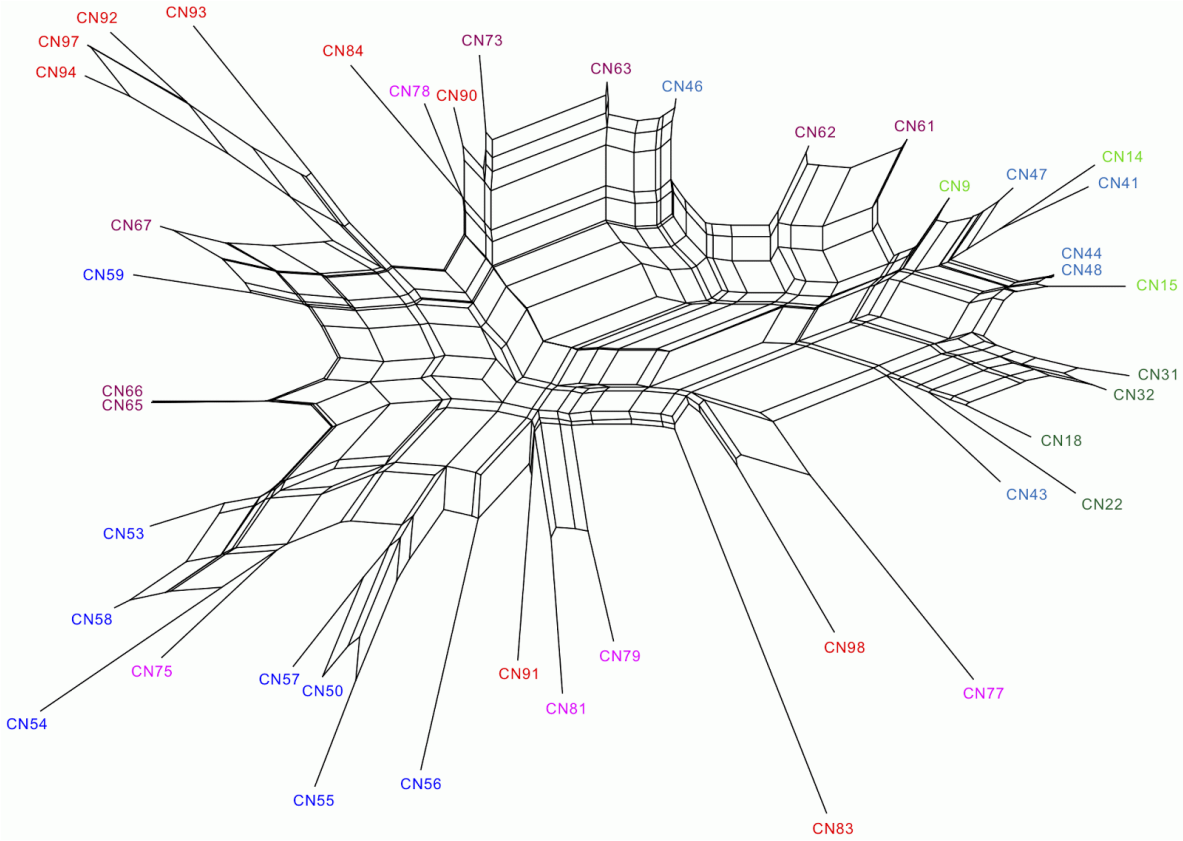
Ask for a tree get a tree.

```
1 ACCGAGCAA
2 ACCGAGCAA
3 ACCGAGCAA
4 ACCGAGCAA

1 ACCGAATGA
2 ACCGAGCAG
3 GTTAGGCAG
4 GTTAGATGA
```



# Tree not always correct paradigm

Ask for a tree get a tree.




Reanalysis of [Marwick, 2012] from
`http://phylonetworks.blogspot.ca/2013/02/`

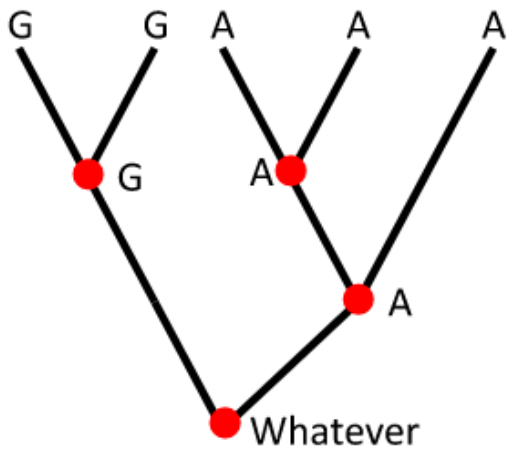
# Back to inference

# Maximum-Likelihood

- Likelihood = p(data | topology, branch, evolutionary model) = $p(D|\tau, \theta)$
- Maximum likelihood is the topology, branch lengths and model parameters with the highest likelihood.
- Performed site by site, search topology space then finding optimal tree parameters.
- Too expensive to exhaustively search likelihood surface so heuristics.
- Most methods start with distance-based starting tree and greedily traverse model space.

# Maximum-Likelihood

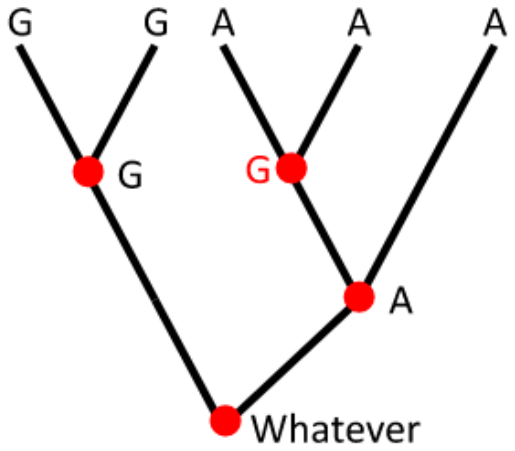
| | |
|---|---|
| Seq1 | A |
| Seq2 | A |
| Seq3 | A |
| Seq4 | G |
| Seq5 | G |




Parsimony's answer for internal states

# Maximum-Likelihood

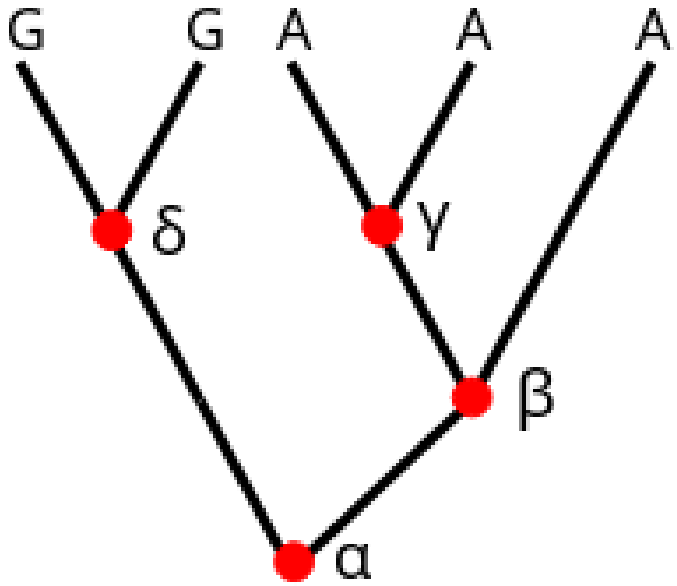
| | |
|---|---|
| Seq1 | A |
| Seq2 | A |
| Seq3 | A |
| Seq4 | G |
| Seq5 | G |




But likelihood will consider this too
(and all other possibilities)

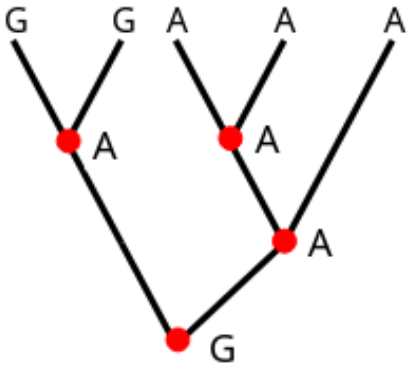
# The "intuition" version:

| | |
|---|---|
| Seq1 | A |
| Seq2 | A |
| Seq3 | A |
| Seq4 | G |
| Seq5 | G |




- $p(D|\tau,\theta) = \sum_\alpha \sum_\beta \sum_\gamma \sum_\delta = p(A,A,A,G,G,\alpha,\beta,\gamma,\delta|\tau,\theta)$

# Maximum-Likelihood




Likelihood of a given tree
(branching order AND branch lengths)

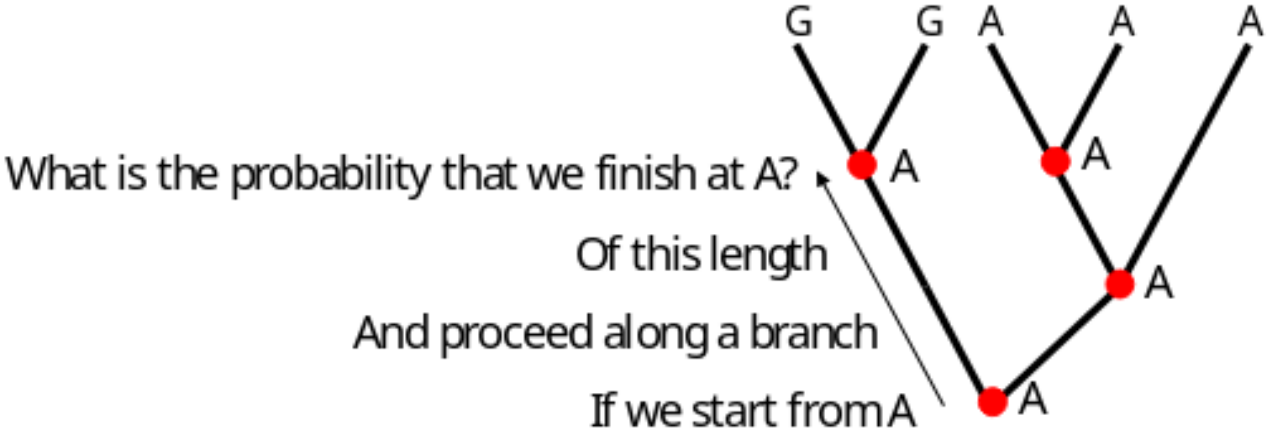
Probability of the states, given the tree

$$P(Data \mid T) = \sum_{\alpha} \sum_{\beta} \sum_{\gamma} \sum_{\delta} P(A, A, A, G, G, \alpha, \beta, \gamma, \delta \mid T)$$

Sum over ALL possible internal states

# Maximum-Likelihood

$$P(A, A, A, G, G, \alpha, \beta, \gamma, \delta \mid T)$$




Depends on SUBSTITUTION MATRIX
and BRANCH LENGTH

# Maximum-Likelihood Pros/Cons

- Advantages:
  - Maximum use of information in data
  - Explicit Model
  - Can handle complex models
  - Robust and consistent (for correct model)
  - Allows comparison of trees (which is ‘best’ and by how much)
- Disadvantages:
  - Default treatment sites as independent.
  - Very slow for exhaustive search
  - Model mispecification issues
  - Difficult to extend.
  - Question formulation can be unintuitive

# Bayesian

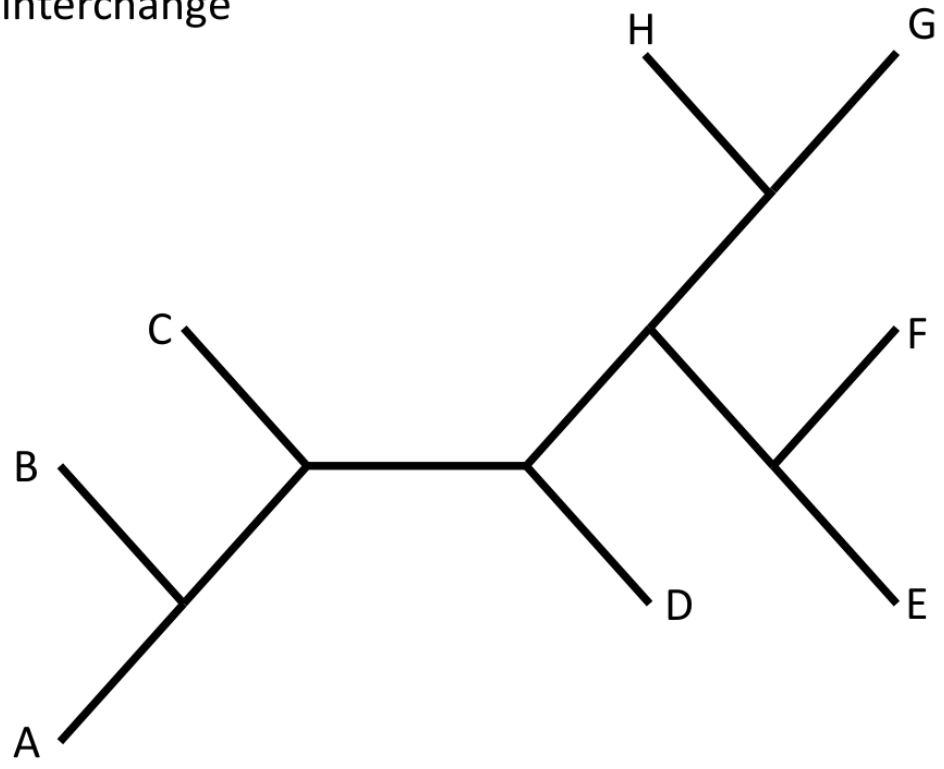
- Bayes Rule: $p(\theta|X) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$
- For trees: $p(\theta, \tau|D) = \frac{p(D|\theta,\tau)p(\theta)p(\tau)}{\int^{\theta}\int^{\tau} p(D|\theta,\tau)p(\theta)p(\tau)d\tau d\theta}$
- Approximate marginal probability using Markov-Chain Monte-Carlo
- Run multiple chains to estimate convergence

# Bayesian Pros/Cons

- Advantages:
  - Fast (relatively)
  - Can infer many different parameters
  - More flexible framework
  - More intuitive formulation
- Disadvantages:
  - Choice of priors
  - Difficulty determining convergence
  - Model mispecification issues.

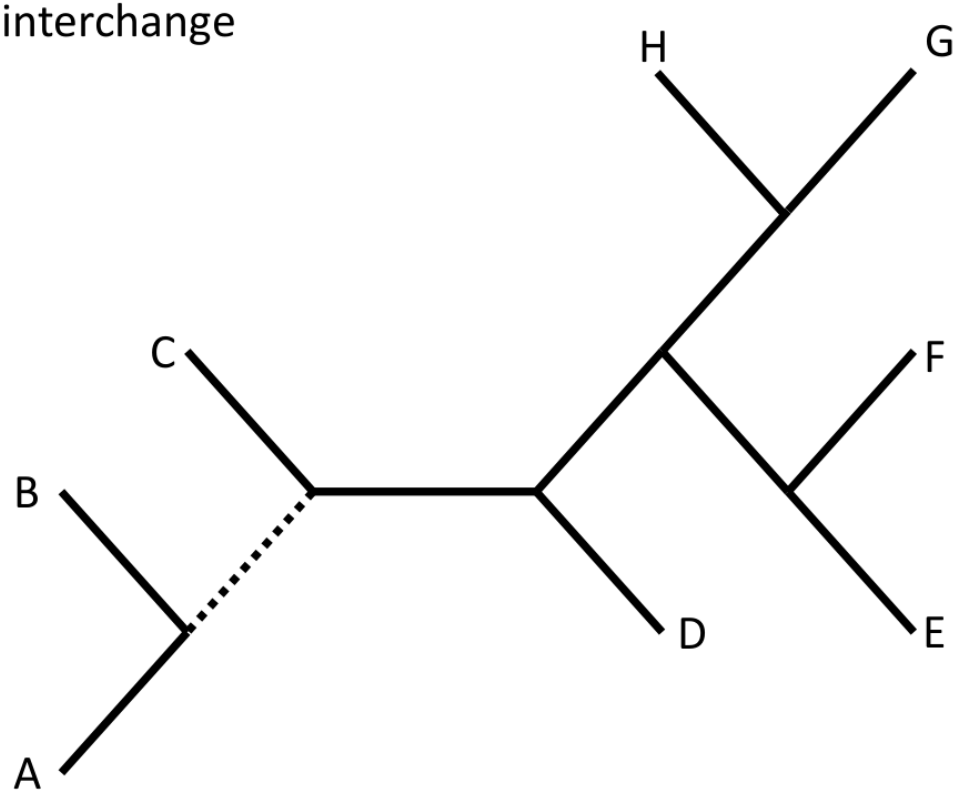
# Searching Tree-Space: NNI

Nearest-neighbor interchange



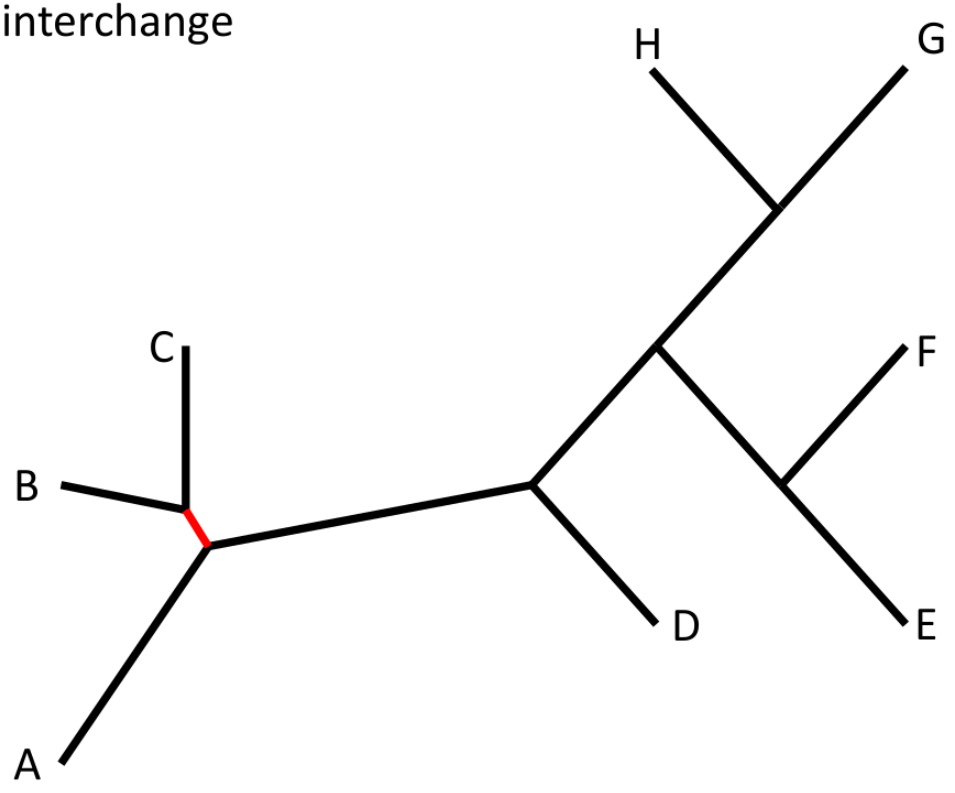
# Searching Tree-Space

Nearest-neighbor interchange



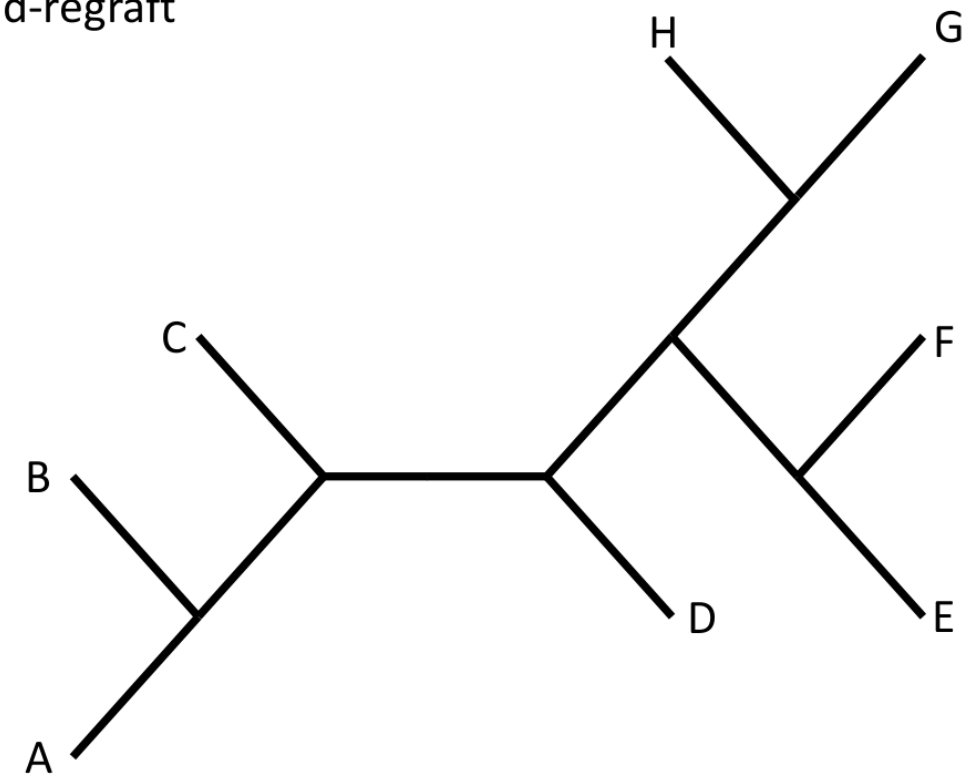
# Searching Tree-Space

Nearest-neighbor interchange



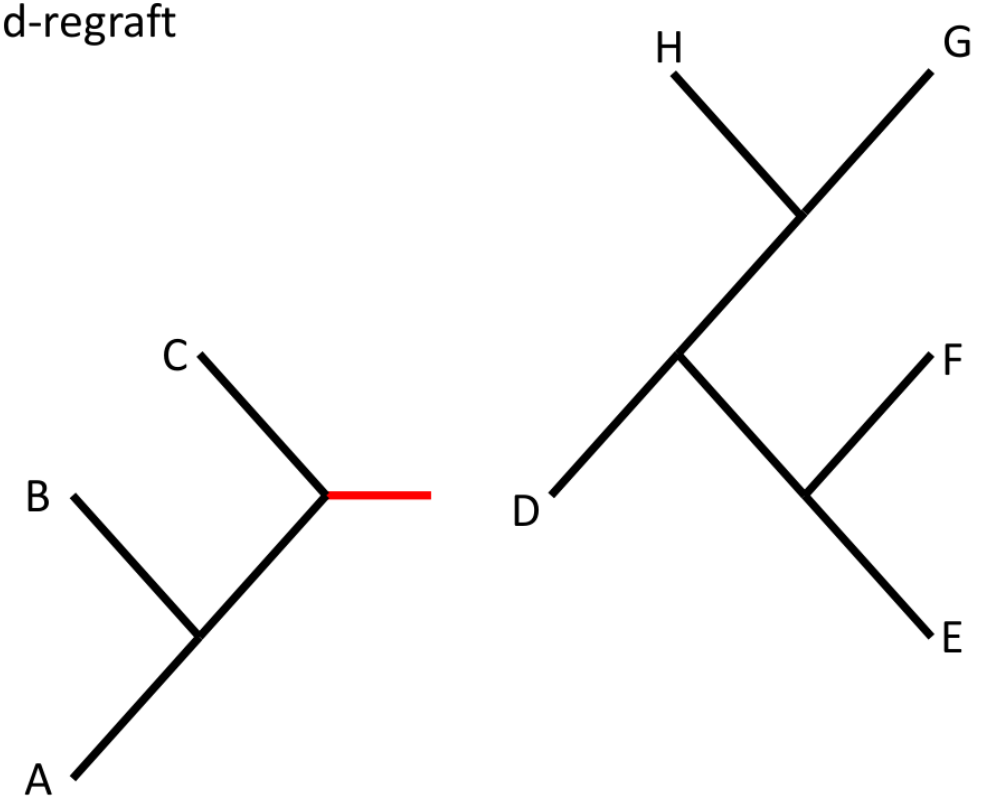
# Searching Tree-Space

Subtree prune-and-regraft



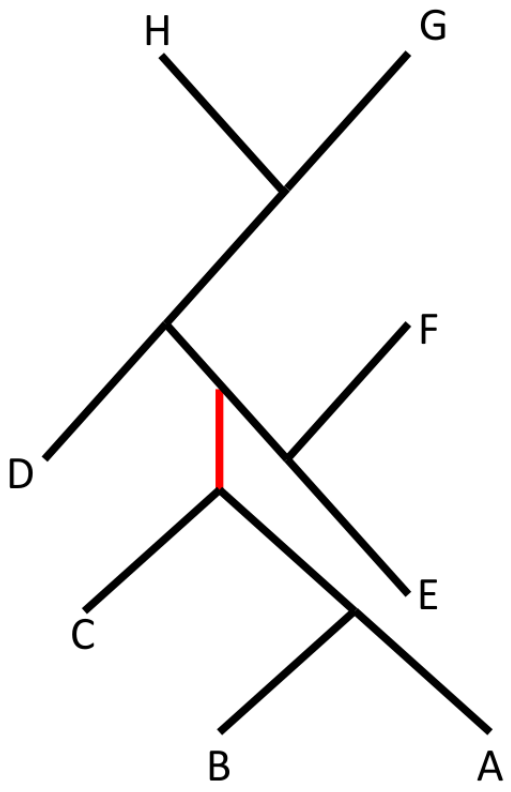
# Searching Tree-Space

Subtree prune-and-regraft



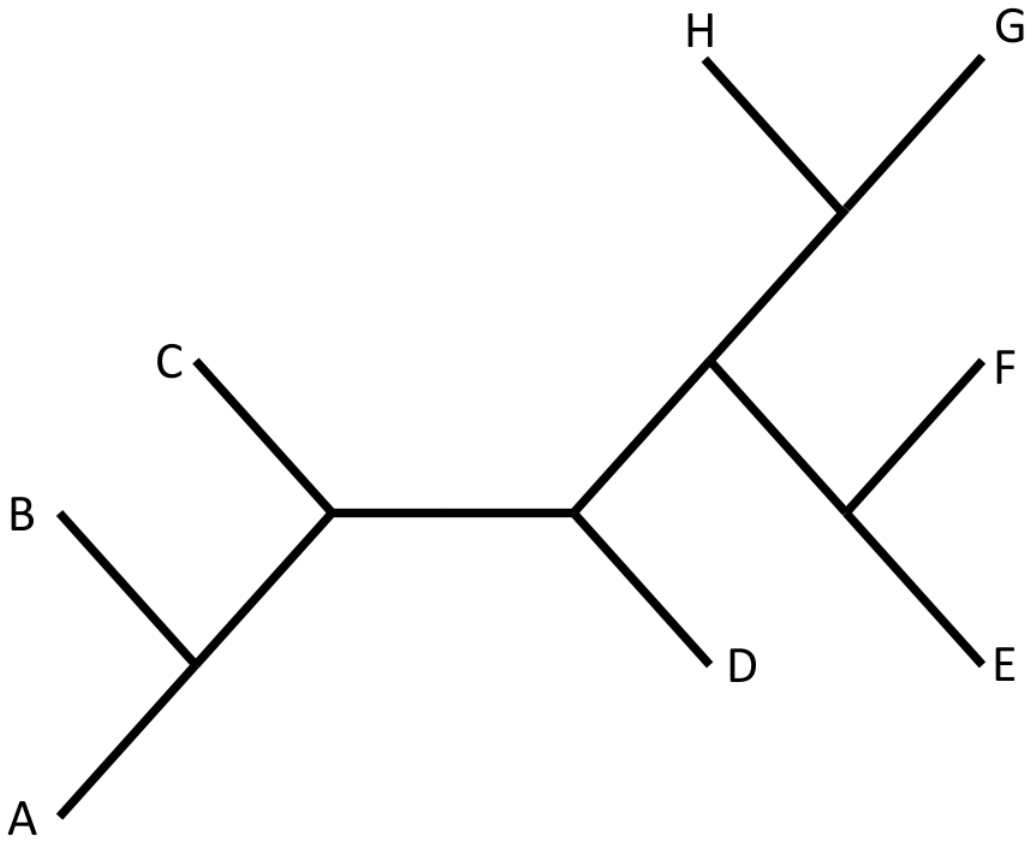
# Searching Tree-Space

Subtree prune-and-regraft



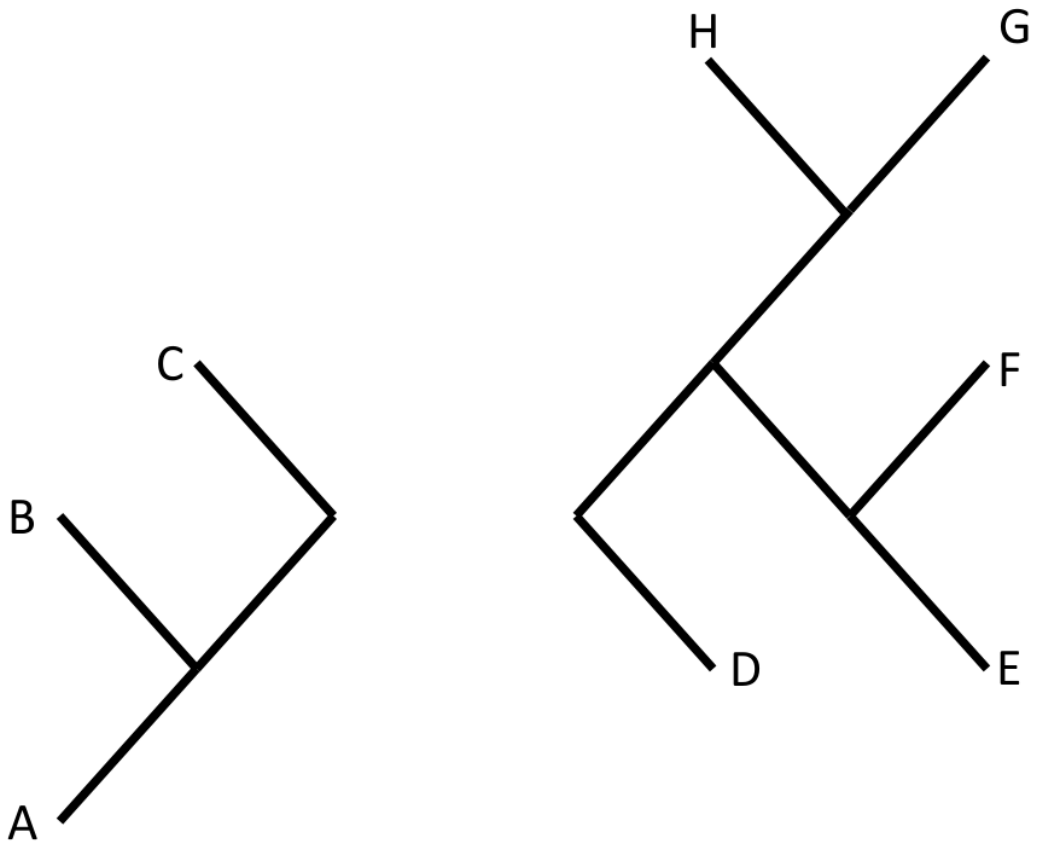
# Searching Tree-Space

Tree bisection and reconnection



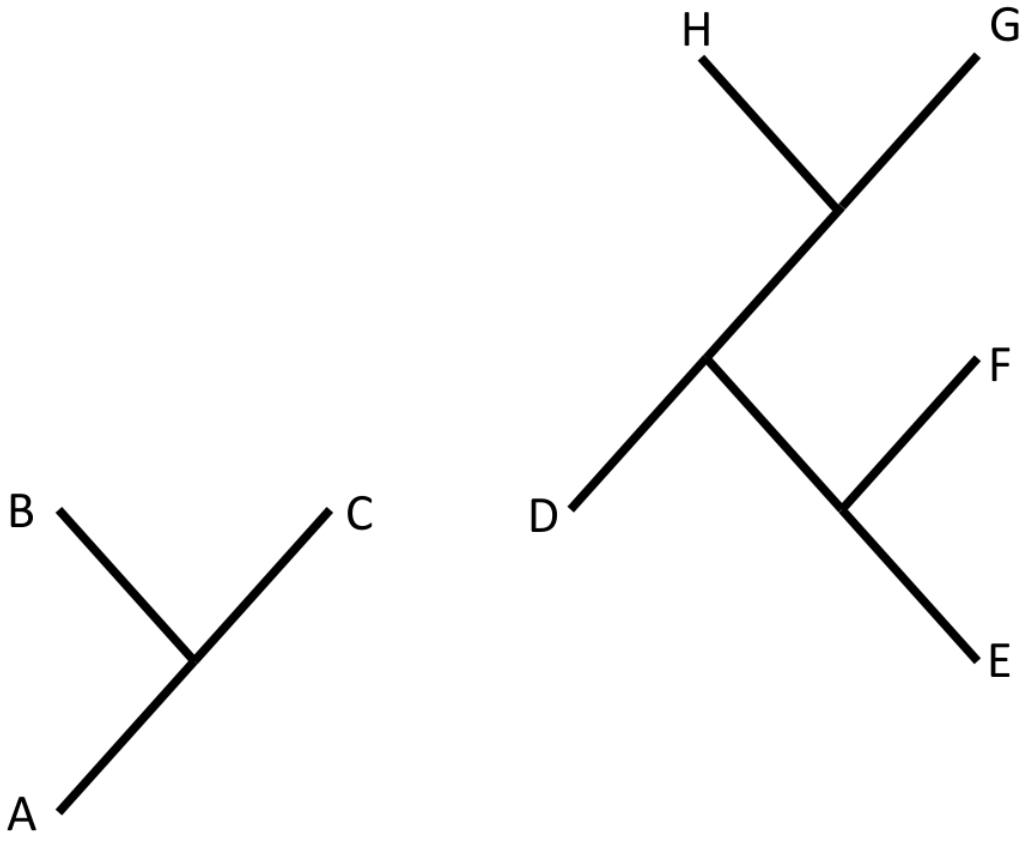
# Searching Tree-Space

Tree bisection and reconnection



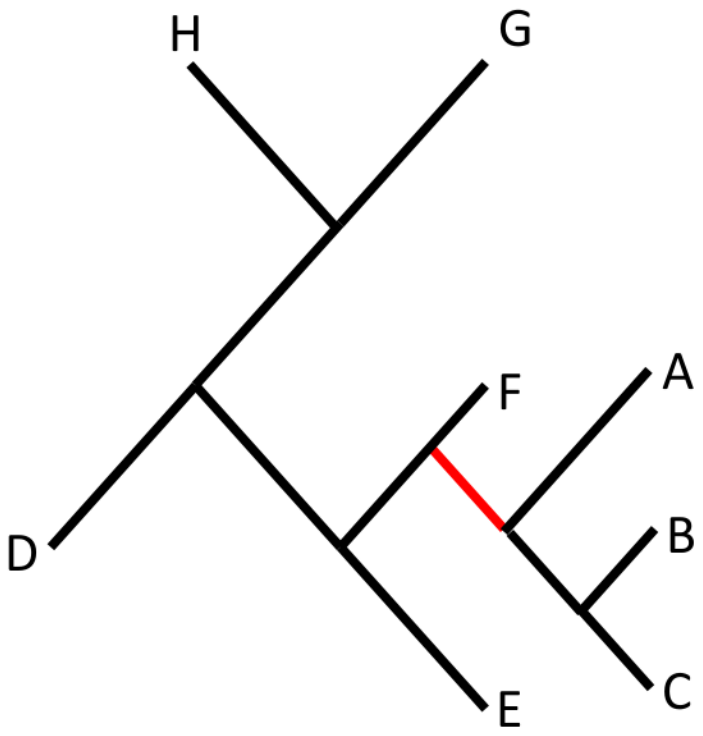
# Searching Tree-Space

Tree bisection and reconnection



# Searching Tree-Space

Tree bisection and reconnection



# Conclusion

# Summary

- Phylogenetics are a useful tool to investigate the relations between sequences

# Summary

- Phylogenetics are a useful tool to investigate the relations between sequences
- There are some tricks to interpretation of trees.

# Summary

- Phylogenetics are a useful tool to investigate the relations between sequences
- There are some tricks to interpretation of trees.
- Inferring a phylogeny requires: data, alignment, trimming, method selection.

# Summary

- Phylogenetics are a useful tool to investigate the relations between sequences
- There are some tricks to interpretation of trees.
- Inferring a phylogeny requires: data, alignment, trimming, method selection.
- Parsimony is simplest but easily misled.

# Summary

- Phylogenetics are a useful tool to investigate the relations between sequences
- There are some tricks to interpretation of trees.
- Inferring a phylogeny requires: data, alignment, trimming, method selection.
- Parsimony is simplest but easily misled.
- Distance, ML, and Bayesian need an evolutionary model.

# Summary

- Phylogenetics are a useful tool to investigate the relations between sequences
- There are some tricks to interpretation of trees.
- Inferring a phylogeny requires: data, alignment, trimming, method selection.
- Parsimony is simplest but easily misled.
- Distance, ML, and Bayesian need an evolutionary model.
- Distance methods are fast but naive.

# Summary

- Phylogenetics are a useful tool to investigate the relations between sequences
- There are some tricks to interpretation of trees.
- Inferring a phylogeny requires: data, alignment, trimming, method selection.
- Parsimony is simplest but easily misled.
- Distance, ML, and Bayesian need an evolutionary model.
- Distance methods are fast but naive.
- ML and Bayesian methods treat phylogenetics as a statistics problem.

# Summary

- Phylogenetics are a useful tool to investigate the relations between sequences
- There are some tricks to interpretation of trees.
- Inferring a phylogeny requires: data, alignment, trimming, method selection.
- Parsimony is simplest but easily misled.
- Distance, ML, and Bayesian need an evolutionary model.
- Distance methods are fast but naive.
- ML and Bayesian methods treat phylogenetics as a statistics problem.
- Allow probabilistic reconstruction of ancestral states and population parameters.

# Summary

- Phylogenetics are a useful tool to investigate the relations between sequences
- There are some tricks to interpretation of trees.
- Inferring a phylogeny requires: data, alignment, trimming, method selection.
- Parsimony is simplest but easily misled.
- Distance, ML, and Bayesian need an evolutionary model.
- Distance methods are fast but naive.
- ML and Bayesian methods treat phylogenetics as a statistics problem.
- Allow probabilistic reconstruction of ancestral states and population parameters.
- Tree topology space is non-trivial to search.

Questions?

# References i


Abadi, S., Azouri, D., Pupko, T., and Mayrose, I. (2019).
**Model selection may not be a mandatory step for phylogeny reconstruction.**
*Nature Communications*, 10(1):934.

Barbrook, A. C., Howe, C. J., Blake, N., and Robinson, P. (1998).
**The phylogeny of the canterbury tales.**
*Nature*, 394(6696):839.

Holmes, E. C., Dudas, G., Rambaut, A., and Andersen, K. G. (2016).
**The evolution of ebola virus: Insights from the 2013–2016 epidemic.**
*Nature*, 538(7624):193.

# References ii


Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., et al. (2016).
**A new view of the tree of life.**
*Nature microbiology*, 1(5):16048.

Leonard, G. (2010).
**Development of fusion and duplication finder blast (fdfblast): a systematic tool to detect differentially distributed gene fusions and resolve trifurcations in the tree of life.**

Marwick, B. (2012).
**A cladistic evaluation of ancient thai bronze buddha images: six tests for a phylogenetic signal in the griswold collection.**
*Connecting empires*, pages 159–176.

# References iii


Nickle, D. C., Heath, L., Jensen, M. A., Gilbert, P. B., Mullins, J. I., and Pond, S. L. K. (2007).
**Hiv-specific probabilistic models of protein evolution.**
*PLoS One*, 2(6):e503.

Richards, T. A., Soanes, D. M., Foster, P. G., Leonard, G., Thornton, C. R., and Talbot, N. J. (2009).
**Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi.**
*The Plant Cell*, 21(7):1897–1911.

Ryu, C.-K., Kim, H.-J., Ji, S.-H., Woo, G., and Cho, H.-G. (2008).
**Detecting and tracing plagiarized documents by reconstruction plagiarism-evolution tree.**
In *2008 8th IEEE International Conference on Computer and Information Technology*, pages 119–124. IEEE.

# References iv

Skelton, C. (2008).
**Methods of using phylogenetic systematics to reconstruct the history of the linear b script.**
*Archaeometry*, 50(1):158–176.