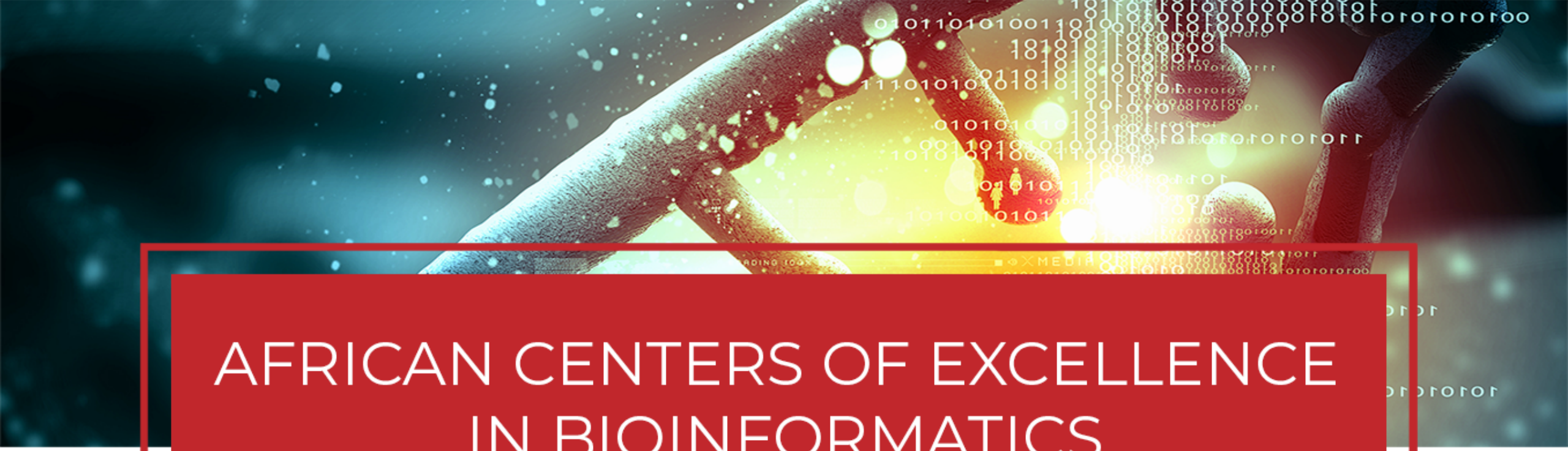
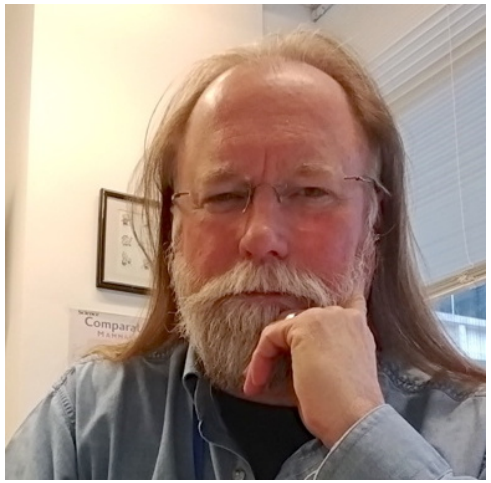


AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

INTRODUCTION TO PHYLOGENETICS 2: TREE INFERENCE AND MODELING EVOLUTION

Today's Instructor



Dr. Kurt Wollenberg, Ph.D. in Genetics

Ongoing Computational Biology projects:

- Hepatitis B molecular evolution
- CLAG protein family evolution

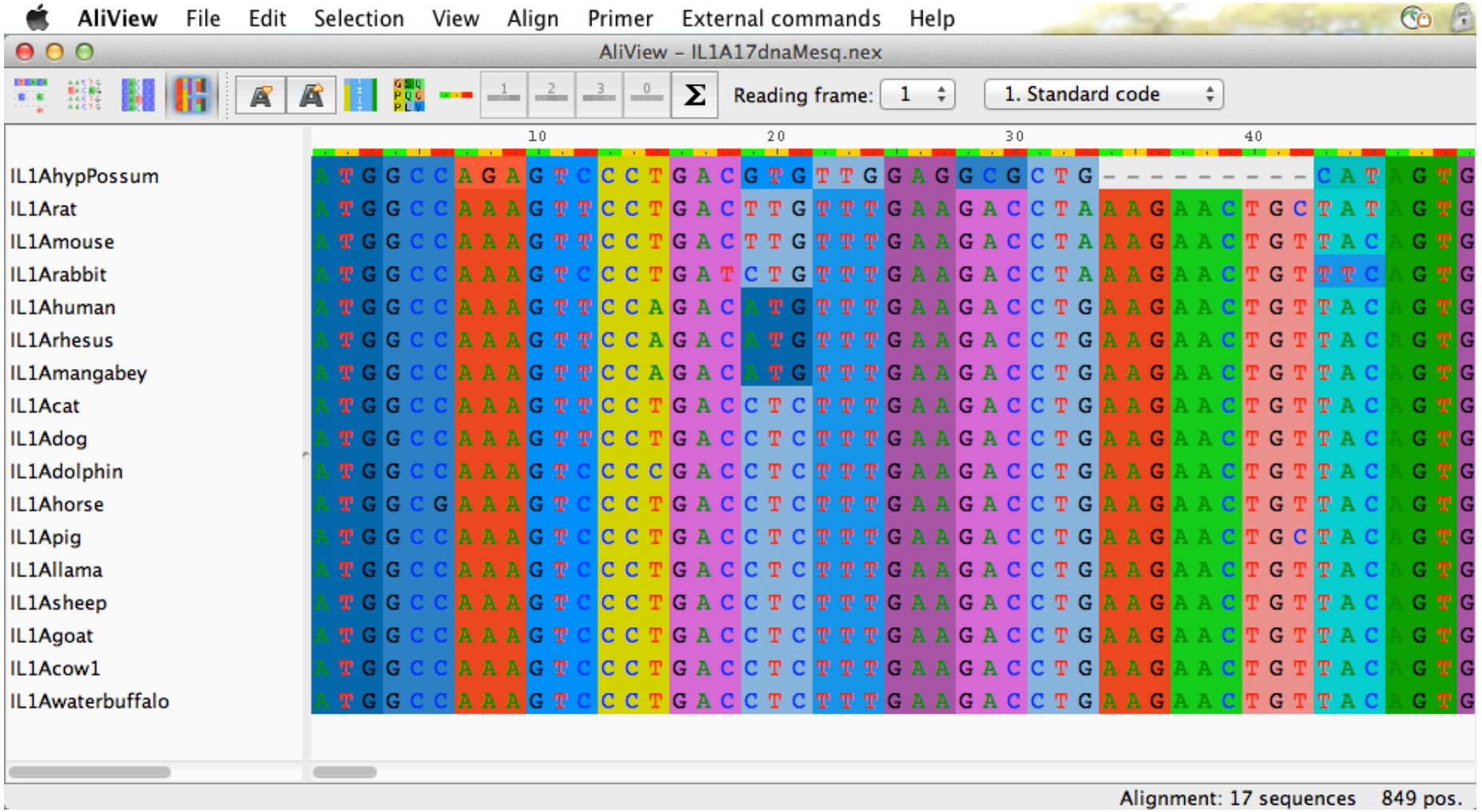
- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: bioinformatics@niaid.nih.gov
  - Instructor: kurt.wollenberg@nih.gov

Class Materials

- Directory on Uganda ACE server:
  - File directory: user@kla-ac-bio-03:/home/bcbb_teaching_files
  - Large data files
- NIAID github repository:
  - https://github.com/niaid/Principles-of-Sequence-Analysis-and-Phylogenetics
  - Code
  - Data files
  - Copies of lecture slides

INTRODUCTION TO PHYLOGENETICS

Multiple Sequence Alignment



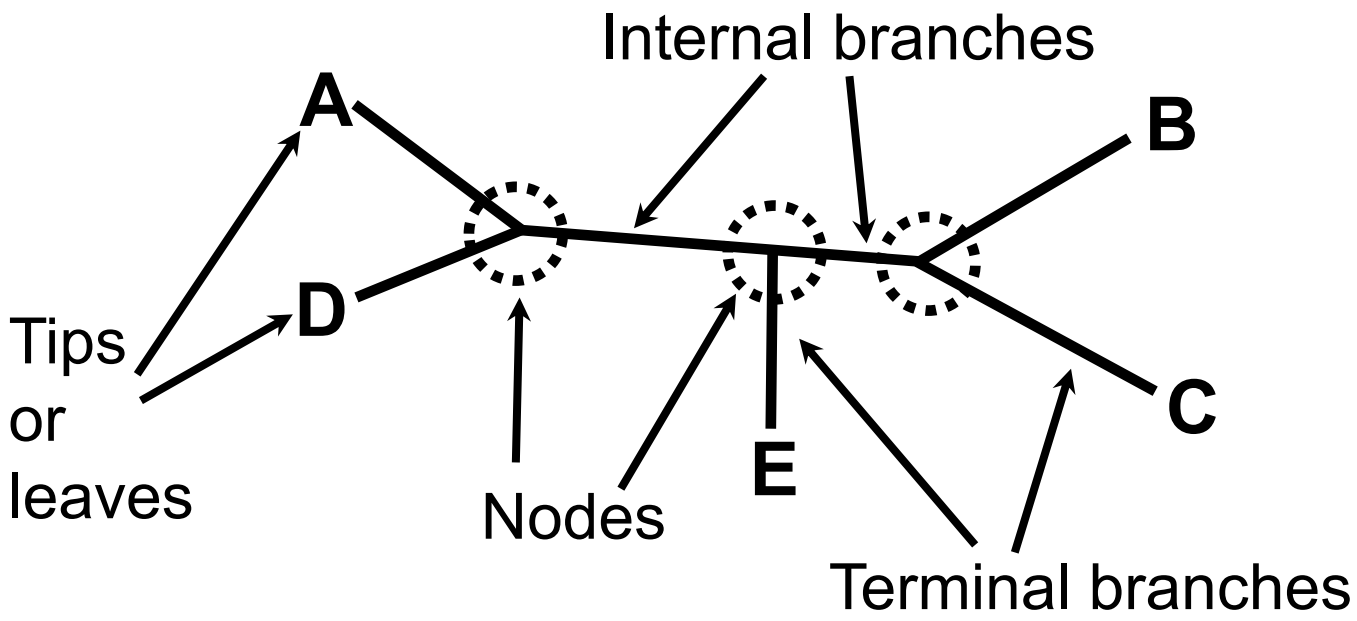
INTRODUCTION TO PHYLOGENETICS

What is a phylogenetic tree?

- Reconstruction of biological history
- Based on similarities and differences among homologous attributes (characters) of the entities under scrutiny
- Molecular characters (sequences, usually) are most often found only in extant organisms

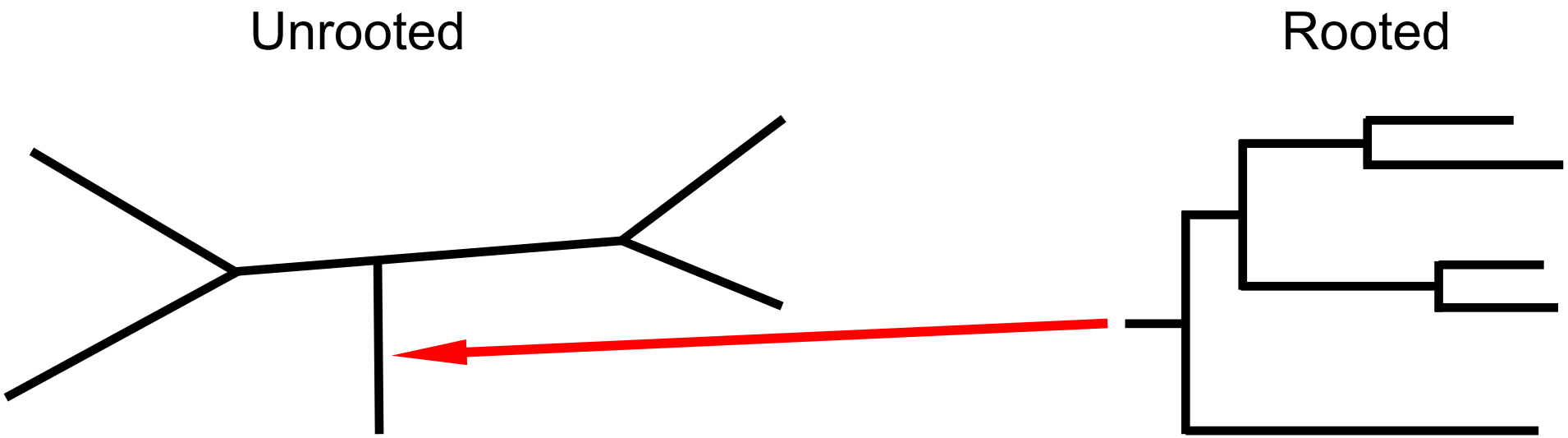
INTRODUCTION TO PHYLOGENETICS

What is a phylogenetic tree?



INTRODUCTION TO PHYLOGENETICS

What is a phylogenetic tree?



INTRODUCTION TO PHYLOGENETICS

Two approaches to tree building

- Application of an algorithm to build the best tree from the data
- Evaluation of multiple possible best trees using an optimality criterion

INTRODUCTION TO PHYLOGENETICS

The algorithm approach: Distance Methods

- Distance calculated based on a specific substitution model (J-C, Kimura, BLOSUM64, etc.)
- Distances from each sequence to all others are calculated and stored in a matrix
- Tree then calculated from the distance matrix using a specific tree-building algorithm
- Neighbor Joining is the most commonly used algorithm

INTRODUCTION TO PHYLOGENETICS

Distance Methods: Neighbor-joining

	A	B	C	D	E	R	R/(N-2)
A	-	0.1715	0.2147	0.3091	0.2326	0.9279	0.3093
B	-0.4766	-	0.2991	0.3399	0.2058	1.0163	0.3388
C	-0.4905	-0.4356	-	0.2795	0.3943	1.1876	0.3959
D	-0.4527	-0.4514	-0.5689	-	0.4289	1.3574	0.4525
E	-0.4972	-0.5535	-0.4221	-0.4441	-	1.2616	0.4205

C to Node 1 distance = 0.2795/2 + (0.3959 - 0.4525)/2 = 0.1114
D to Node 1 distance = 0.2795 - 0.1114 = 0.1681

A to Node 1 distance = (0.2147 + 0.3091 - 0.2795)/2 = 0.1222
B to Node 1 distance = (0.2991 + 0.3399 - 0.2795)/2 = 0.1798
E to Node 1 distance = (0.3943 + 0.4298 - 0.2795)/2 = 0.2719

Hillis, Moritz, and Mable 1996, p. 489

INTRODUCTION TO PHYLOGENETICS

Distance Methods: Neighbor-joining

	A	B	E	Node 1	R	R/(N-2)
A	-	0.1715	0.2326	0.1222	0.5263	0.2631
B	-0.3701	-	0.2058	0.1798	0.5571	0.2785
E	-0.3856	-0.4278	-	0.2719	0.7103	0.3551
Node 1	-0.4278	-0.3856	-0.3701	-	0.5739	0.2869

A to Node 2 distance = 0.1222/2 + (0.2631 – 0.2869)/2 = 0.0492
Node 1 to Node 2 distance = 0.1222 – 0.0492 = 0.0730

B to Node 2 distance = (0.1715 + 0.1798 - 0.1222)/2 = 0.1146
E to Node 2 distance = (0.2326 + 0.2719 - 0.1222)/2 = 0.1912

Hillis, Moritz, and Mable 1996, p. 489

INTRODUCTION TO PHYLOGENETICS

Distance Methods: Neighbor-joining

	B	E	Node 2	R	R/(N-2)
B	-	0.2058	0.1146	0.3204	0.3204
E	-0.5116	-	0.1912	0.3970	0.3970
Node 2	-0.5116	-0.5116	-	0.3058	0.3058

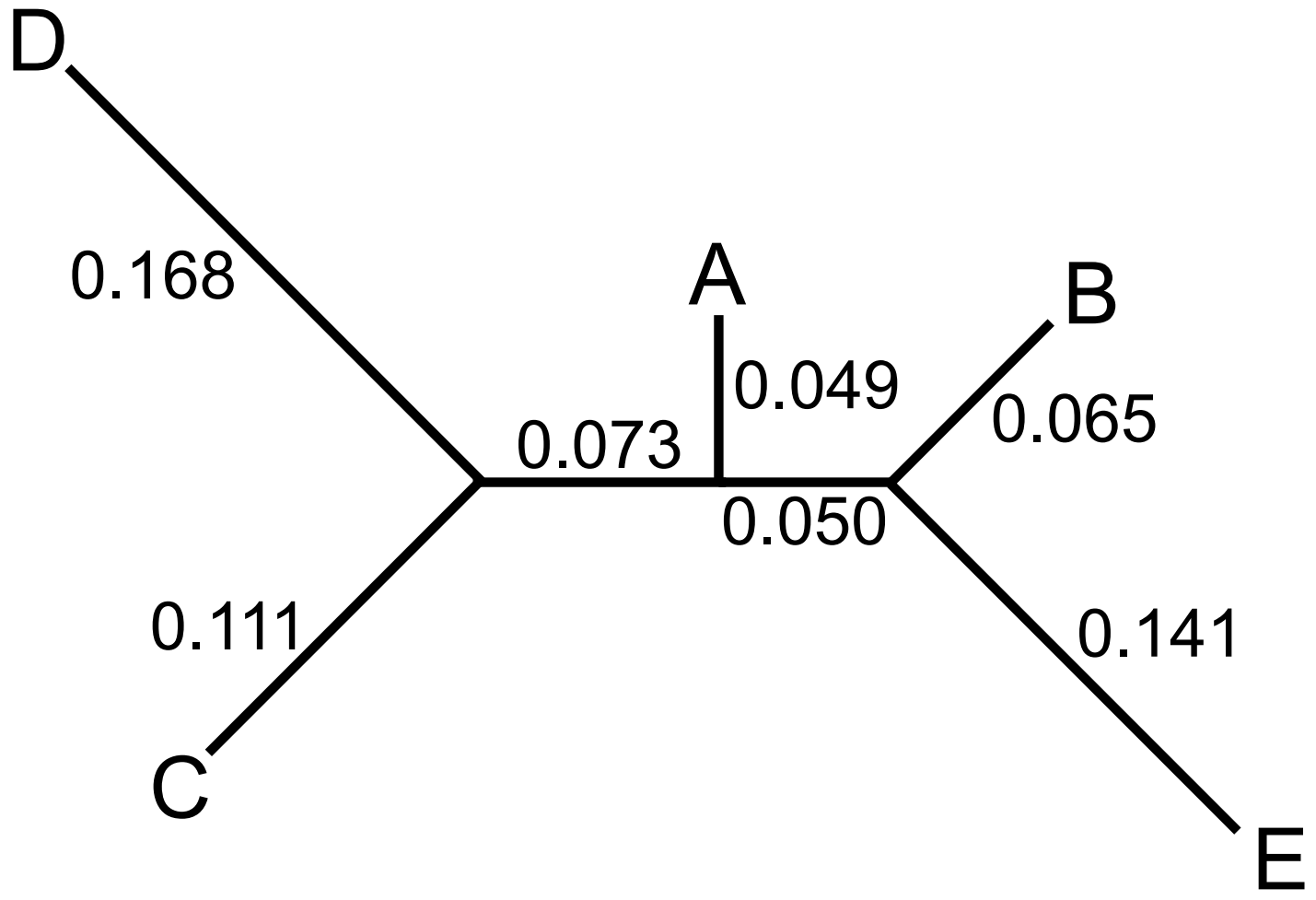
B to Node 3 distance = 0.1146/2 + (0.3204 – 0.3058)/2 = 0.0646
Node 2 to Node 3 distance = 0.1146 – 0.0646 = 0.0500

E to Node 3 distance = (0.2058 0.1912 - 0.1146)/2 = 0.1412

Hillis, Moritz, and Mable 1996, p. 489

INTRODUCTION TO PHYLOGENETICS

Distance Methods: Neighbor-joining



INTRODUCTION TO PHYLOGENETICS

The optimality criterion approach

- Build a tree or trees
- Evaluate the tree(s) using a specific numerical optimality criterion
- Most common optimality criteria
  - Maximum parsimony
  - Maximum likelihood
- Explore tree space to find the optimal tree

INTRODUCTION TO PHYLOGENETICS

Optimality Criterion: Parsimony

Occam's Razor: The simplest explanation is the preferred explanation.

The tree requiring the minimal number of changes is the optimal tree.

A step is any change in the data from one state to another.

INTRODUCTION TO PHYLOGENETICS

Optimality Criterion: Maximum Likelihood

The tree score is the logarithm of the likelihood of the tree.

The likelihood of the tree is the probability of the data given the tree structure.

$$L(\text{Tree}) = \text{Prob}(\text{Data}|\text{Tree}) = \prod_i \text{Prob}(\text{Data}^{(i)}|\text{Tree})$$

INTRODUCTION TO PHYLOGENETICS

Maximum Likelihood: a review

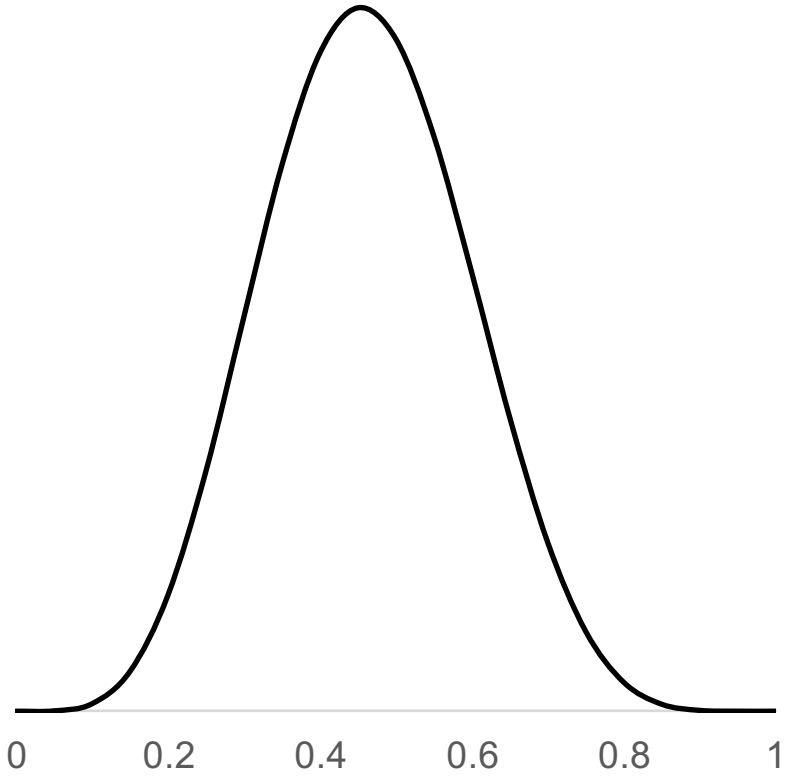
Consider a series of independent events, such as tossing a coin. For a sequence of tosses such as HHTTHTHHTTT we can calculate the probability of the data, which is the likelihood of the data given the underlying probability of the event

$$L = \text{Prob}(\text{Data}|p) = p^5(1-p)^6$$

INTRODUCTION TO PHYLOGENETICS

Maximum Likelihood: a review

Likelihood curve for HHTTHTHHTTT



$L = \text{Prob}(\text{Data}|p) = p^5(1-p)^6$

INTRODUCTION TO PHYLOGENETICS

Maximum Likelihood: a review

This curve has an obvious maximum value. We can calculate it (the maximum likelihood estimate of p, the probability of Heads given this series of flips) from $\frac{dL}{dp} = 0$ but using log-likelihoods (*lnL*) simplifies the math

$$L = p^5(1-p)^6;\ \ln L = 5\ln p + 6\ln(1-p)$$

$$\frac{d(\ln L)}{dp} = \frac{5}{p} - \frac{6}{(1-p)} = 0\ giving\ \hat{p} = \frac{5}{11} = 0.4545$$

INTRODUCTION TO PHYLOGENETICS

The optimality criterion approach

- Build the initial tree
  - Construct a neighbor-joining tree
  - Stepwise addition
- Calculate the tree score
  - Count steps (parsimony)
  - Calculate likelihood of the data given the tree
- Explore tree space
  - Branch swapping
    - Tree bisection and reconnection (TBR)
- Is this the best tree? (Stopping criteria)

INTRODUCTION TO PHYLOGENETICS

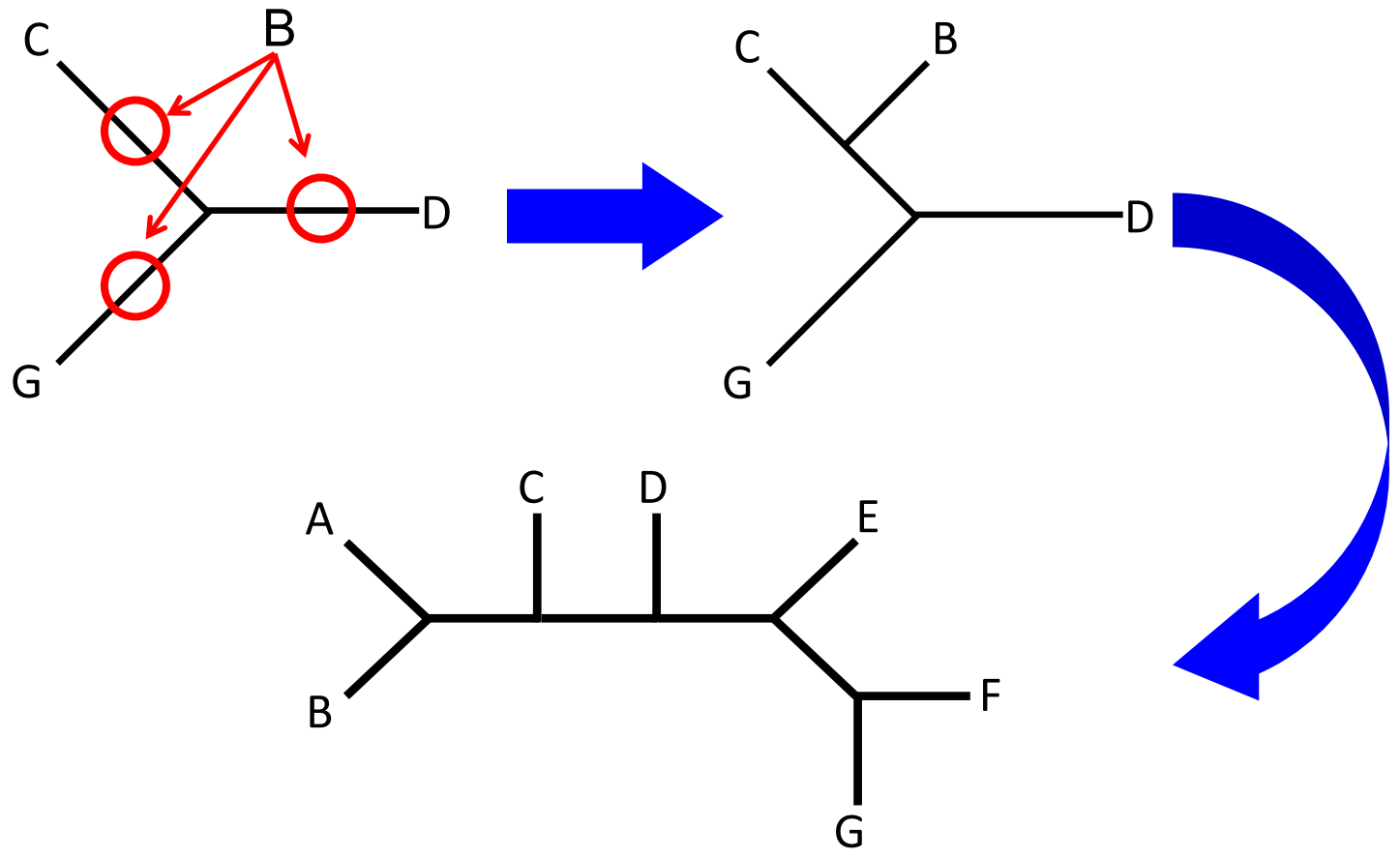
The optimality criterion approach

Building the initial tree

- Stepwise addition
  - Choose three taxa and join
  - Random, or closest
- Select a new taxon to add
- Calculate the optimal 4-taxa tree
- Repeat until all taxa are joined

INTRODUCTION TO PHYLOGENETICS

The optimality criterion approach



INTRODUCTION TO PHYLOGENETICS

The optimality criterion approach

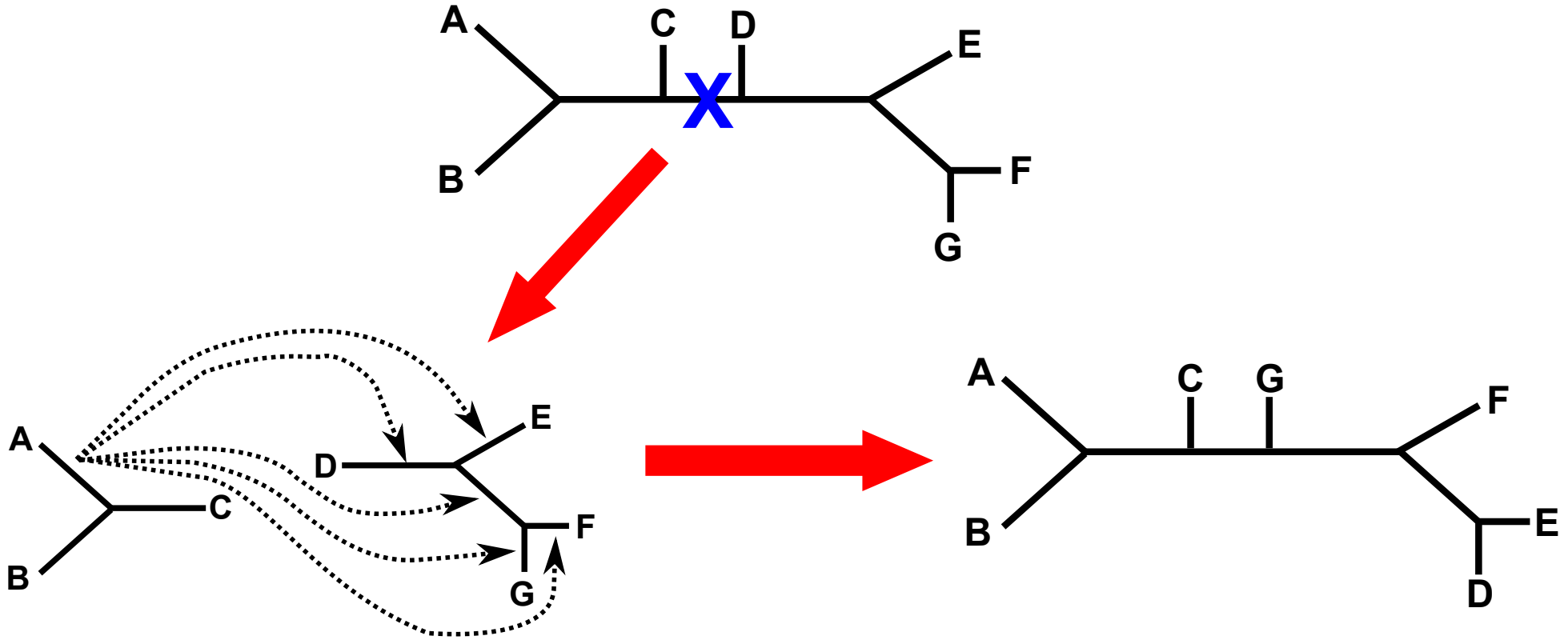
Exploring tree space: Branch swapping

- Nearest neighbor interchange
- Subtree pruning and regrafting
- Tree bisection and reconnection

INTRODUCTION TO PHYLOGENETICS

The optimality criterion approach

Branch swapping: Tree bisection and reconnection



INTRODUCTION TO PHYLOGENETICS

The optimality criterion approach

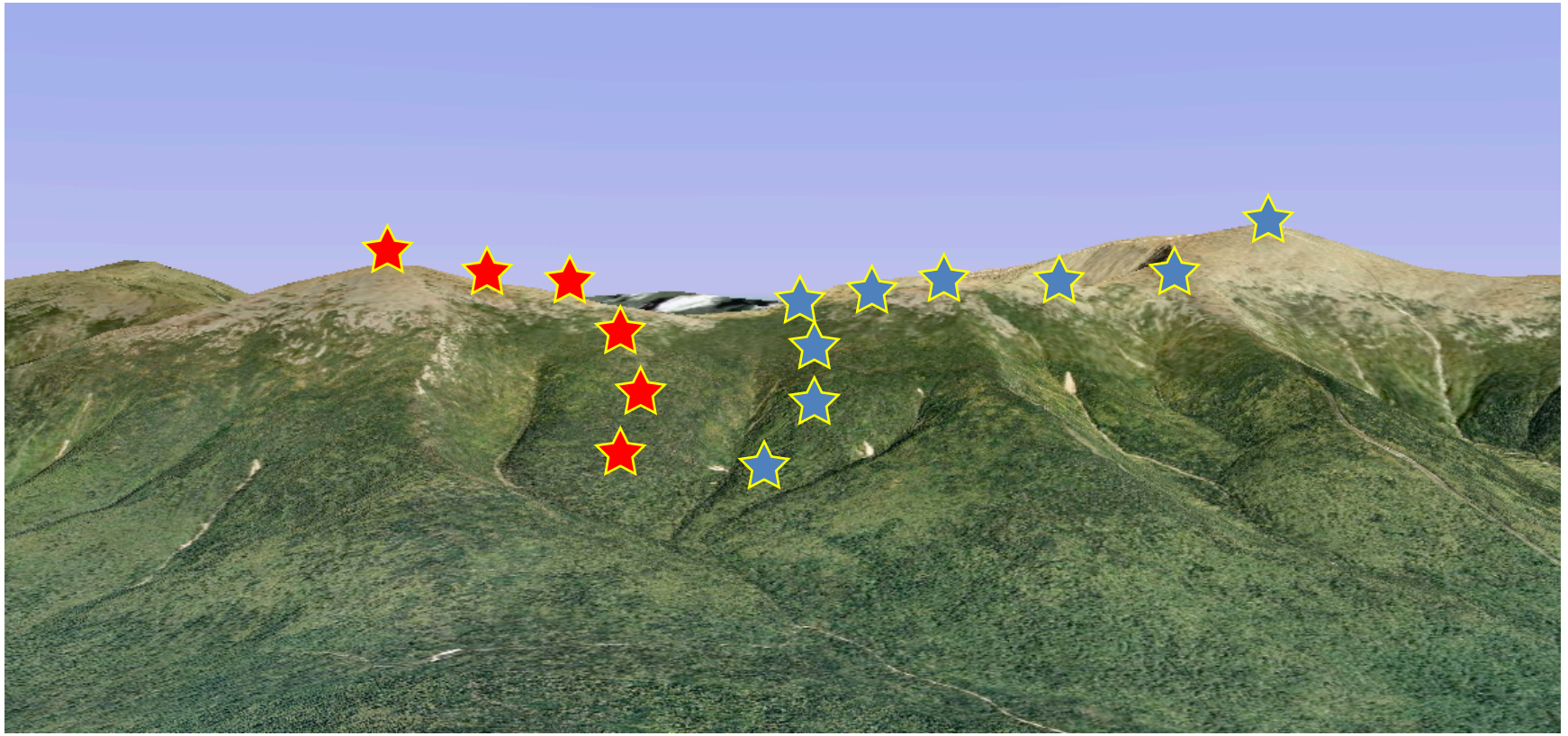
Exploring tree space

Beware! Hill climbing can often lead to local maxima rather than a global solution.

INTRODUCTION TO PHYLOGENETICS

The optimality criterion approach

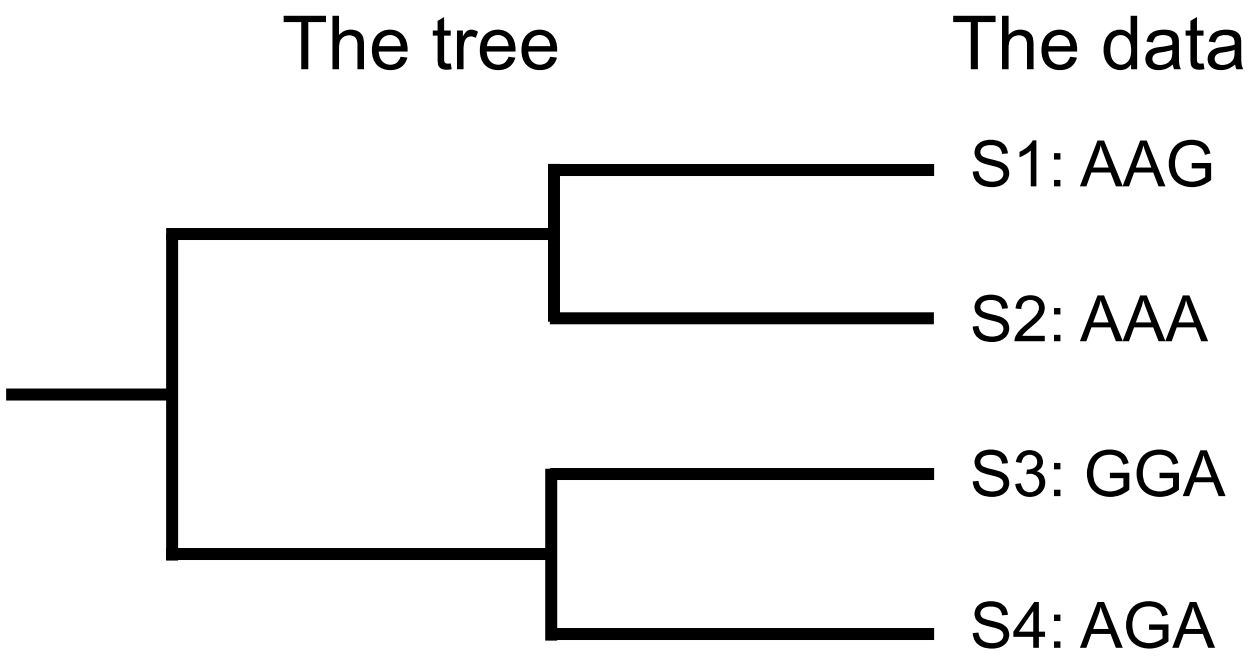
Exploring tree space



INTRODUCTION TO PHYLOGENETICS

Optimality Criterion: Parsimony

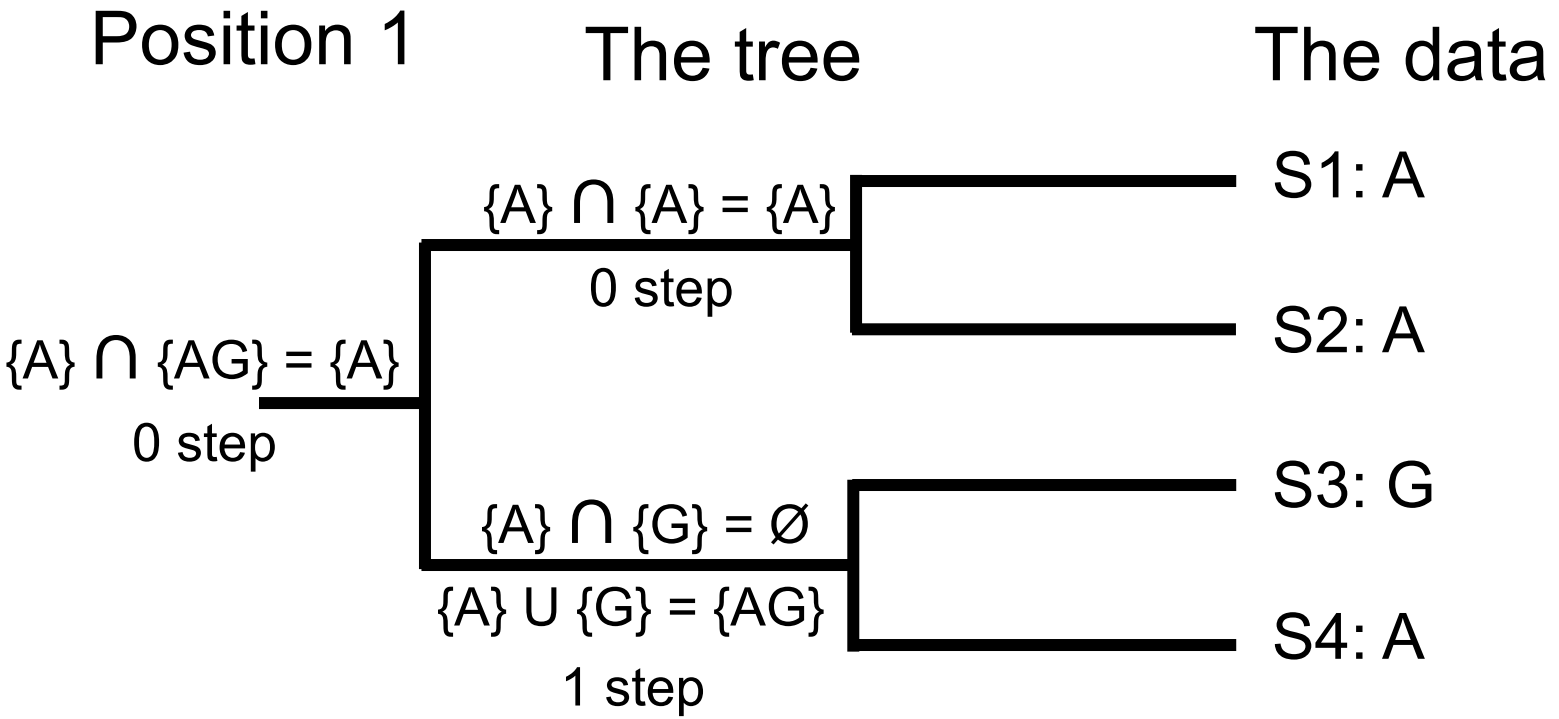
Counting changes (Fitch parsimony)



INTRODUCTION TO PHYLOGENETICS

Optimality Criterion: Parsimony

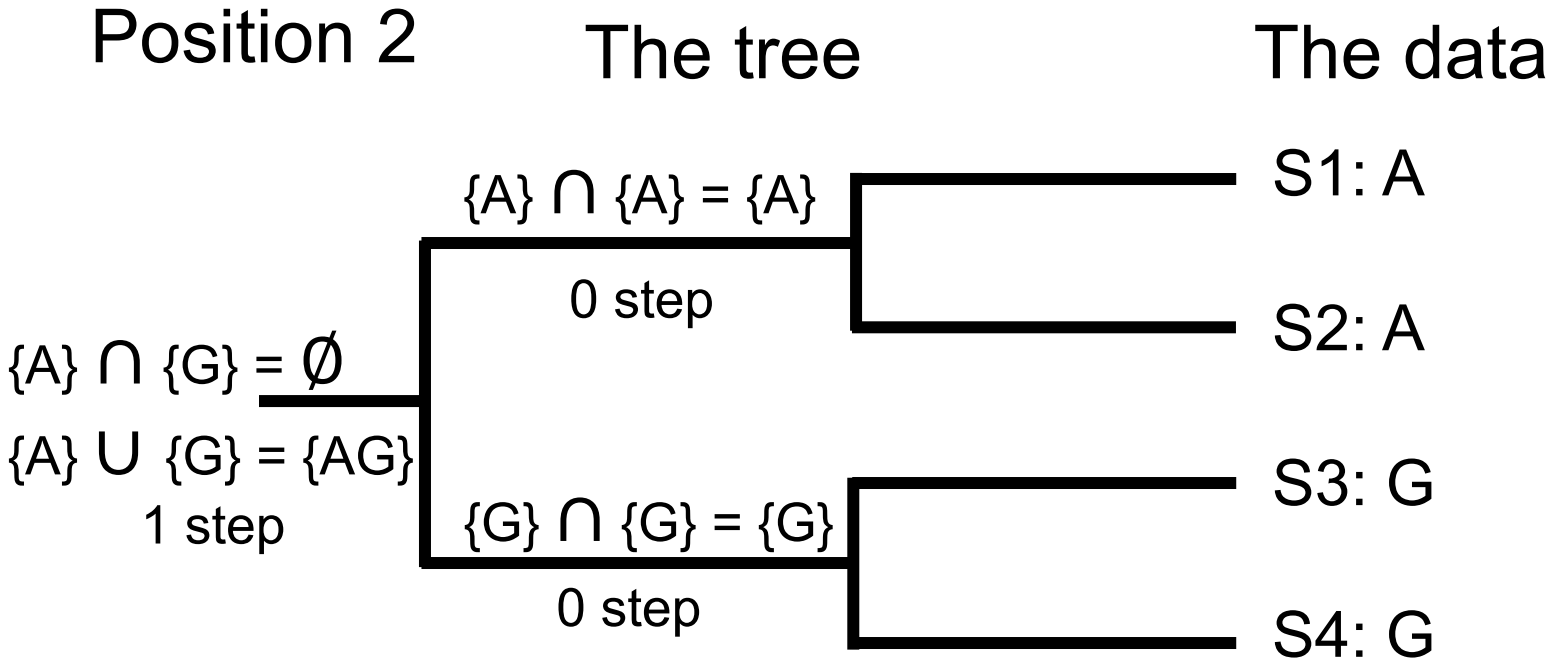
Counting changes (Fitch parsimony)



INTRODUCTION TO PHYLOGENETICS

Optimality Criterion: Parsimony

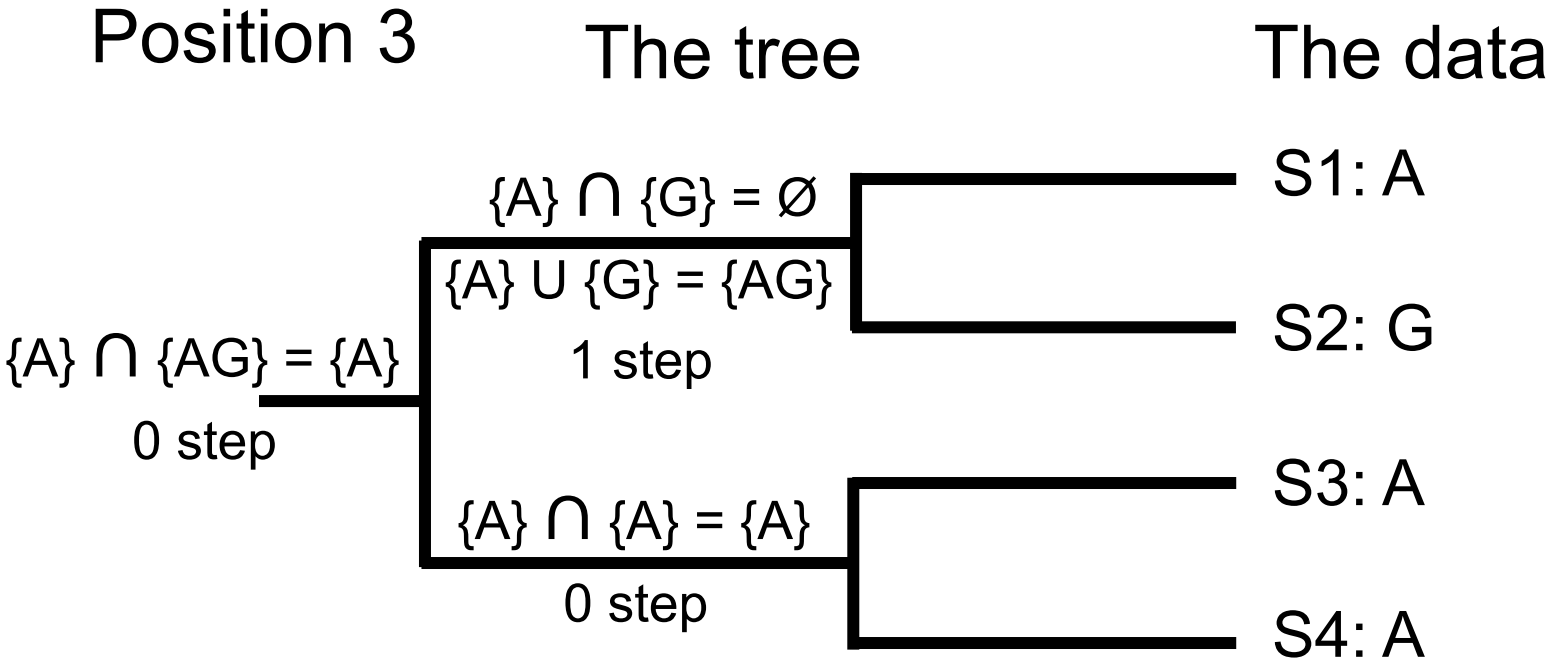
Counting changes (Fitch parsimony)



INTRODUCTION TO PHYLOGENETICS

Optimality Criterion: Parsimony

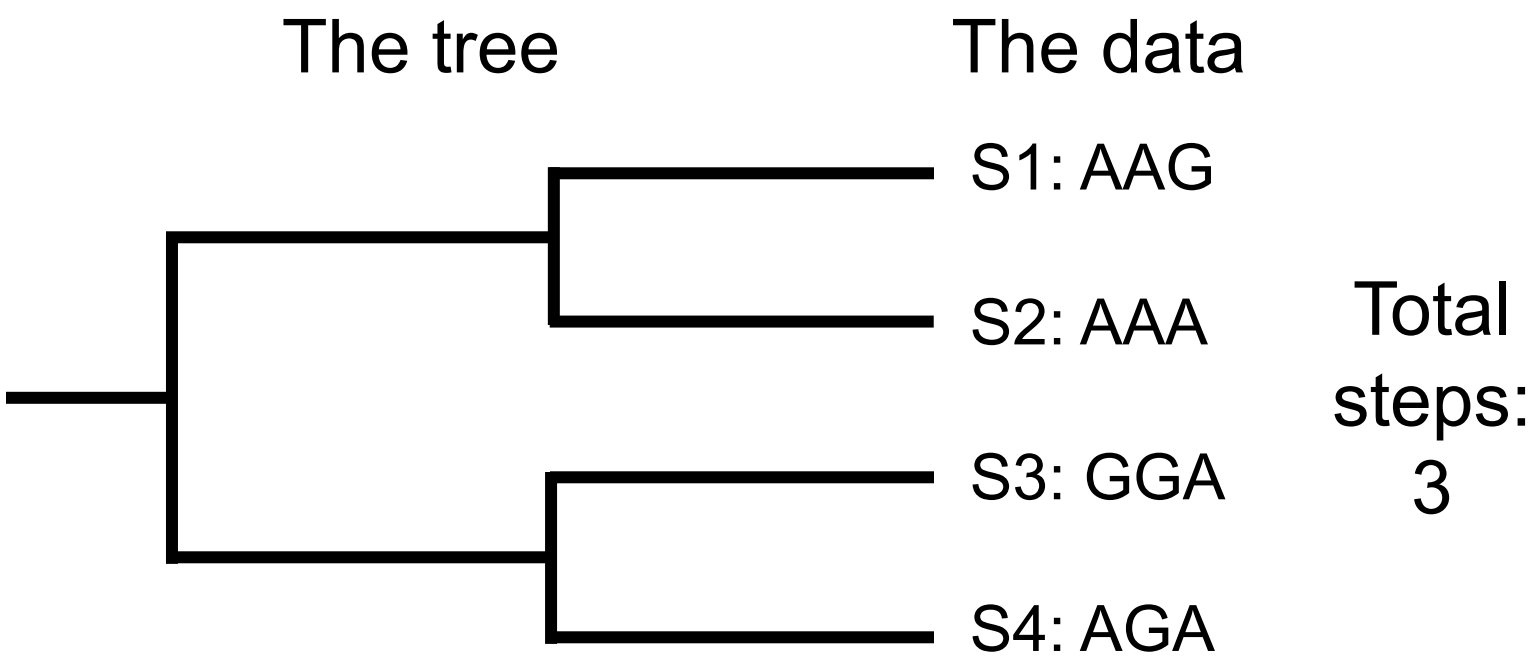
Counting changes (Fitch parsimony)



INTRODUCTION TO PHYLOGENETICS

Optimality Criterion: Parsimony

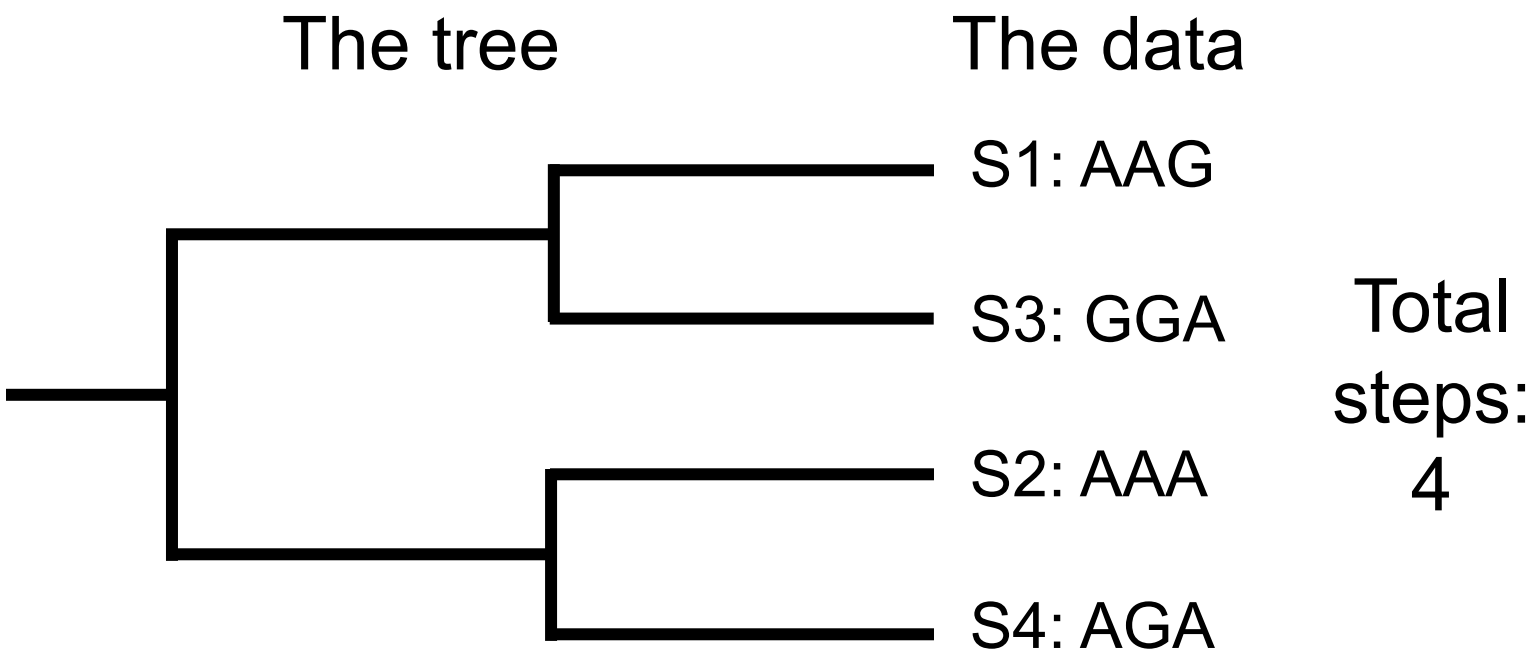
Counting changes (Fitch parsimony)



INTRODUCTION TO PHYLOGENETICS

Optimality Criterion: Parsimony

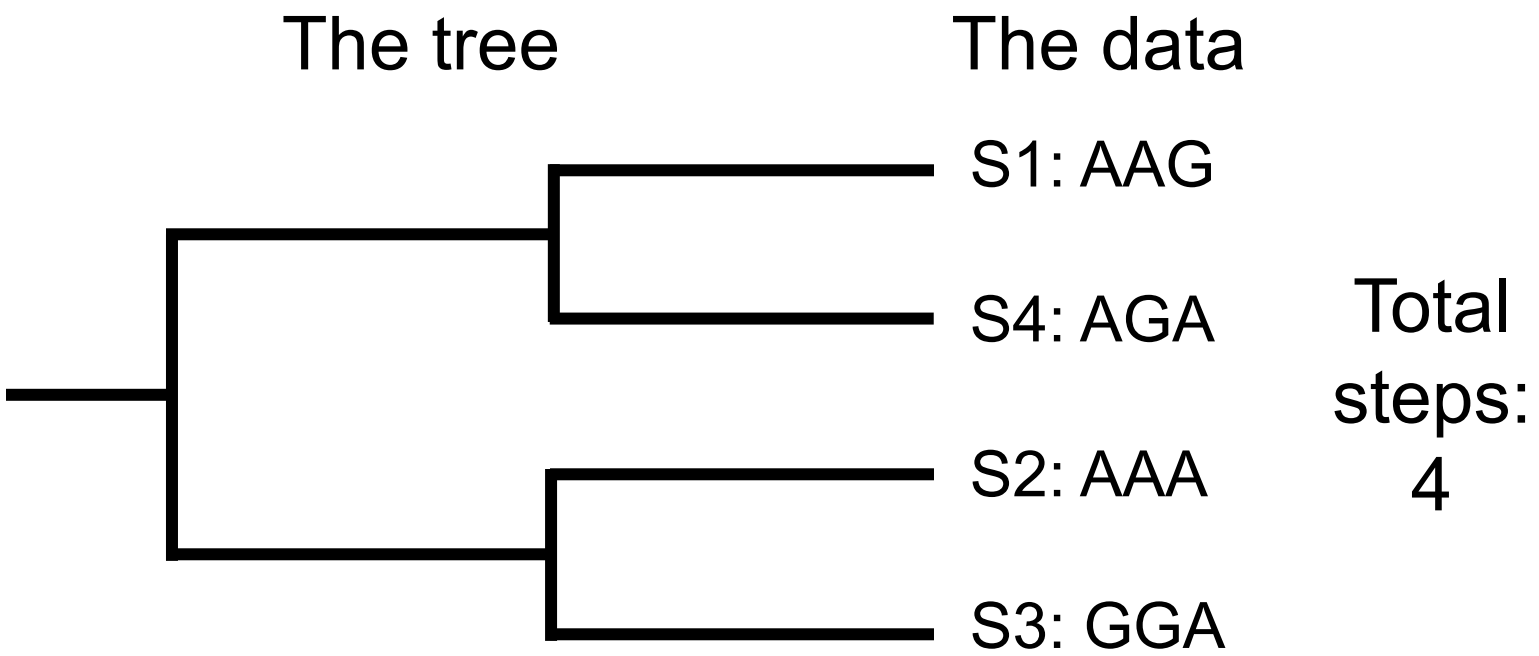
Counting changes (Fitch parsimony)



INTRODUCTION TO PHYLOGENETICS

Optimality Criterion: Parsimony

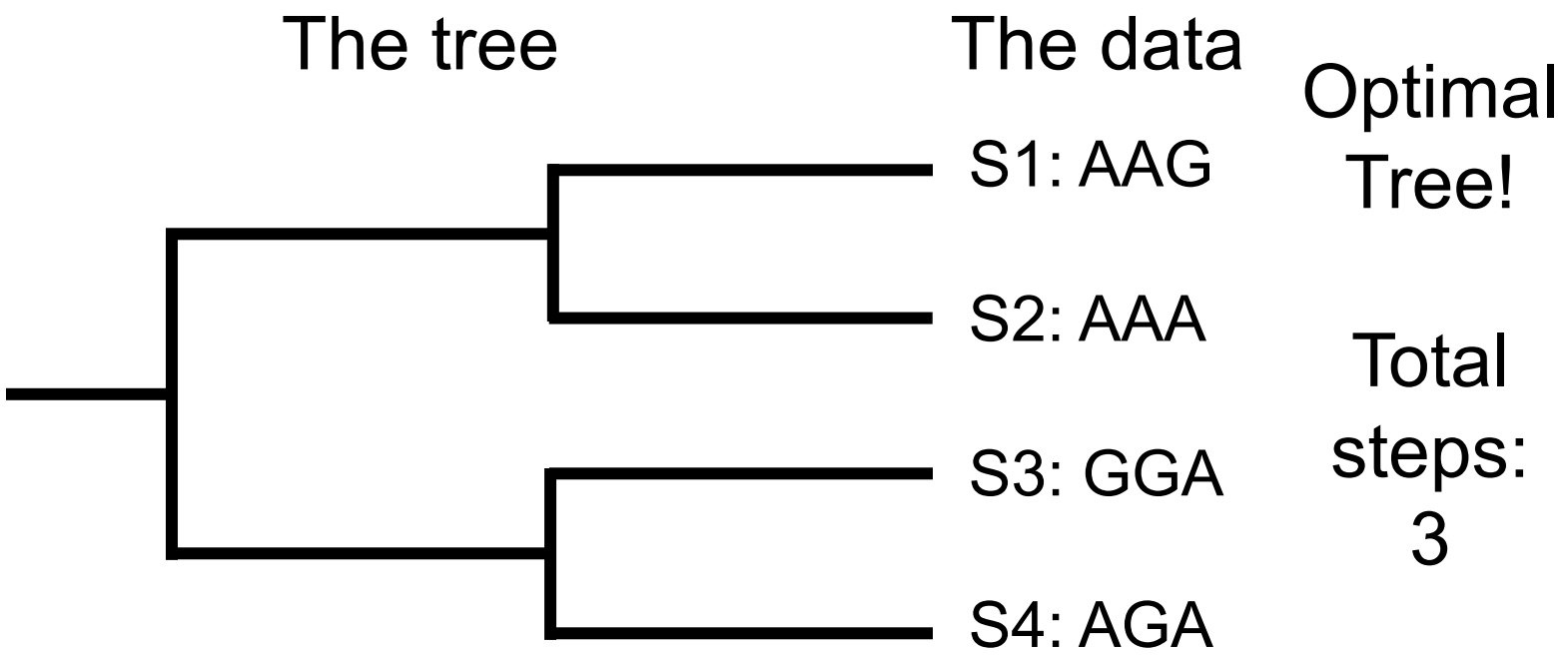
Counting changes (Fitch parsimony)



INTRODUCTION TO PHYLOGENETICS

Optimality Criterion: Parsimony

Counting changes (Fitch parsimony)



INTRODUCTION TO PHYLOGENETICS

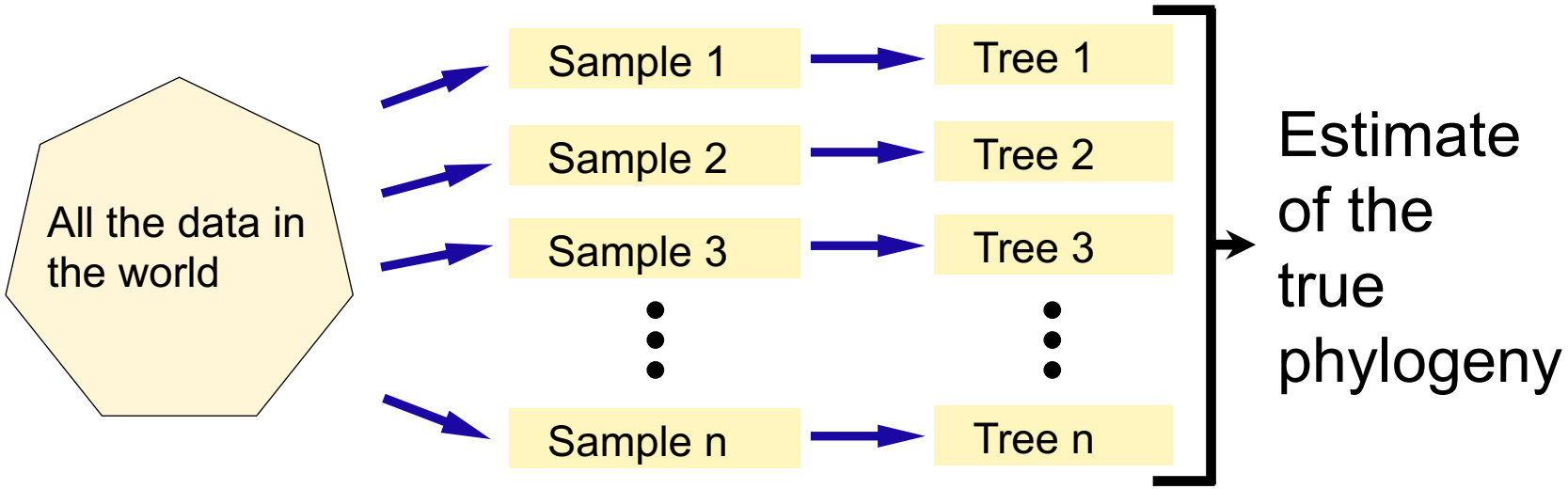
Bootstrapping: How reliable are my trees?

- Parametric bootstrapping: generate replicate data sets based on a set of parameters that describe the original data.
- Nonparametric bootstrapping: generate replicate data sets by sampling with replacement from the original data.

INTRODUCTION TO PHYLOGENETICS

Tree Reliability: Nonparametric Bootstrapping

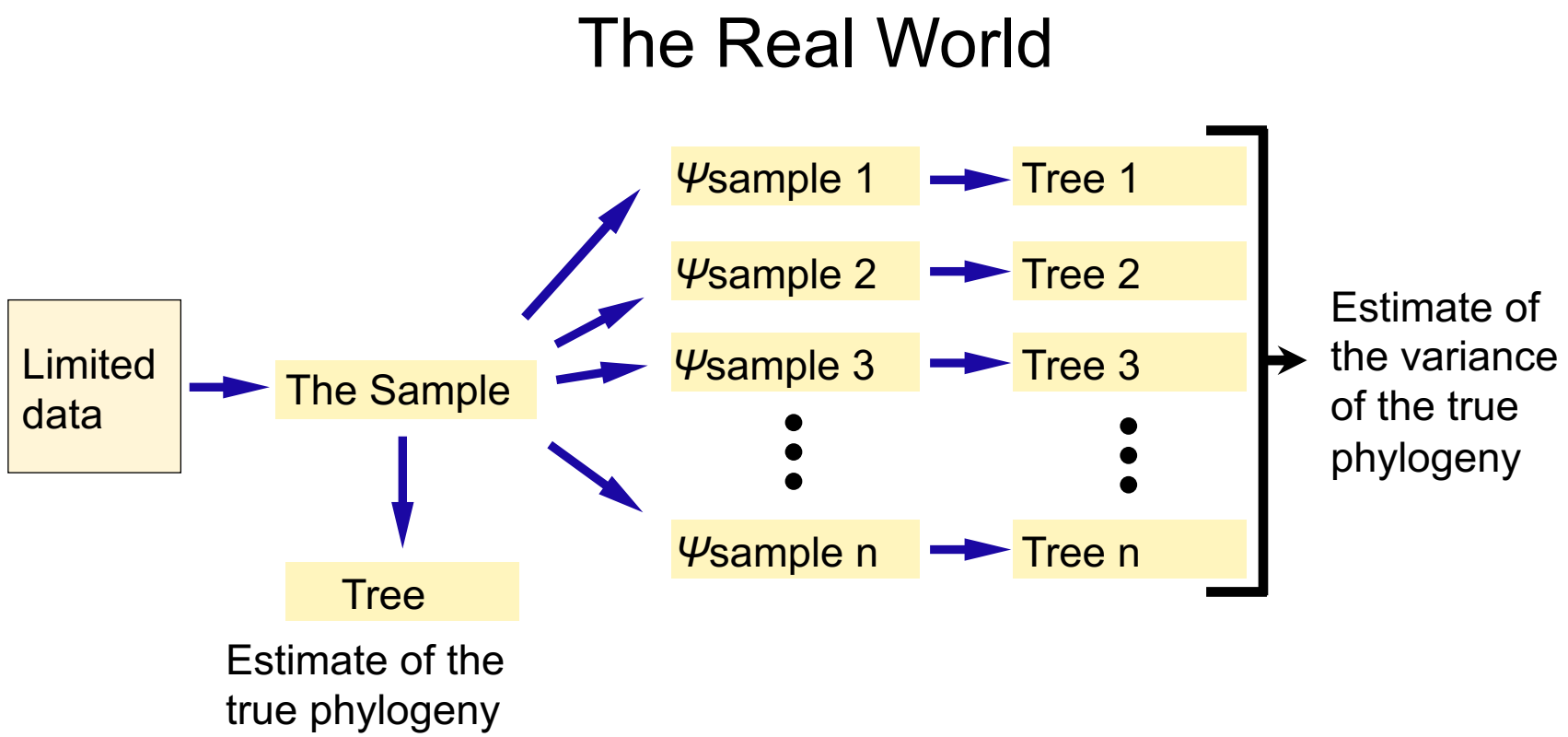
The Ideal World



Build replicates by resampling from unlimited data

INTRODUCTION TO PHYLOGENETICS

Tree Reliability: Nonparametric Bootstrapping



Build pseudoreplicates of unlimited data by sampling with replacement from limited data

INTRODUCTION TO PHYLOGENETICS

Molecular Clocks

- As protein data (gel electrophoresis and molecular sequences) accumulated it appeared that proteins were varying in a regular manner across different lineages of organisms.
- This lead Zuckerkandl and Pauling (1965) to postulate that a “molecular clock” existed
- The “ticking” of the clock is amino acid or nucleotide substitution
- Regular, clock-like substitution would make building phylogenies much simpler
- Trees built from sequences that change in a clock-like manner will have tip-to-tip distances that sum over the branches. These types of trees are also called *ultrametric*.

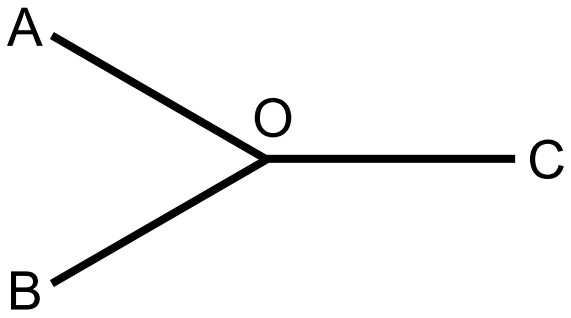
INTRODUCTION TO PHYLOGENETICS

Molecular Clocks

Are our sequences changing in a clock-like manner?

The Relative-Rate Test (Sarich and Wilson 1973)

For three sequences A, B, and C



$$K_{AC} = K_{OA} + K_{OC}$$
$$K_{BC} = K_{OB} + K_{OC}$$
$$K_{AB} = K_{OA} + K_{OB}$$

$$K_{OA} = (K_{AC} + K_{AB} - K_{BC})/2$$
$$K_{OB} = (K_{AB} + K_{BC} - K_{AC})/2$$
$$K_{OC} = (K_{AC} + K_{BC} - K_{AB})/2$$

INTRODUCTION TO PHYLOGENETICS

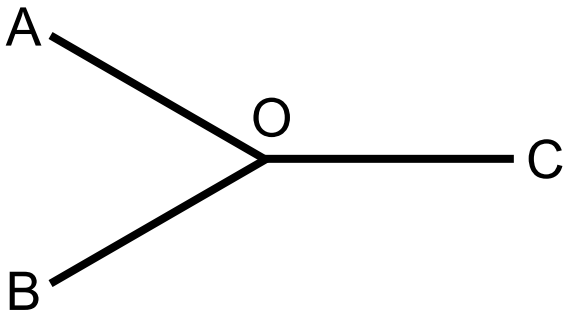
Molecular Clocks

Are our sequences changing in a clock-like manner?

The Relative-Rate Test (Sarich and Wilson 1973)

For three sequences A, B, and C

$$K_{OA} = \frac{(K_{AC} + K_{AB} - K_{BC})}{2} \; ; \; K_{OB} = \frac{(K_{AB} + K_{BC} - K_{AC})}{2} \; ; K_{OC} = \frac{(K_{AC} + K_{BC} - K_{AB})}{2}$$



- If substitution is clock-like, then $K_{OA} = K_{OB}$, or $d = K_{OA} - K_{OB} = K_{AC} - K_{BC}$
- If $|d| > 2$ x SE, then $d > 0$ and substitution is not clock-like
- $V(d) = V(K_{AC}) + V(K_{BC}) - 2V(K_{OC})$
- Calculation of variances depends on substitution model being used

INTRODUCTION TO PHYLOGENETICS

Molecular Clocks

- Data indicate that substitution does not occur at the same rate across divergent lineages (rodent vs primate, for example)
- Substitution can be clock-like when lineages are closely related (local clocks)
- Sources of rate variation among lineages
  - Differences in efficiency of DNA repair
  - Differences in generation time
  - Differences in metabolic rate
  - Selection on nonsynonymous substitutions

INTRODUCTION TO PHYLOGENETICS

Software for Phylogenetic Analysis

- A list of phylogenetics software
  - http://evolution.genetics.washington.edu/phylip/software.html
- Multi-method packages
  - MEGAX
  - Phylip
  - DAMBE
  - R/phangorn

INTRODUCTION TO PHYLOGENETICS

Software for Phylogenetic Analysis

- Single-method software
  - PhyML
  - RaxML
  - Garli
- Phylogenetic tree processing
  - FigTree
  - Dendroscope3
  - iTOL

INTRODUCTION TO PHYLOGENETICS

Thank you