Phylogenetic Analysis

Andrea Silverj, PhD

Postdoc @University of Calgary

Department of Biological Sciences

Gregory Lab

01-12-2025, BWG10

# What is phylogenetics?

The diversity of living beings

• All life shares a common genetic history.
• Phylogeny is the framework for studying its diversity.

The Tree of Life

• A branching diagram that illustrates relationships of biological entities.
• A fundamental tool to investigate evolutionary processes.

Why it matters?

• Helps classify organisms based on shared ancestry.
• Reveals patterns of diversification and adaptation.
• Provides insights into gene and genome evolution.
• Essential for comparative genomics, ecology, and conservation.
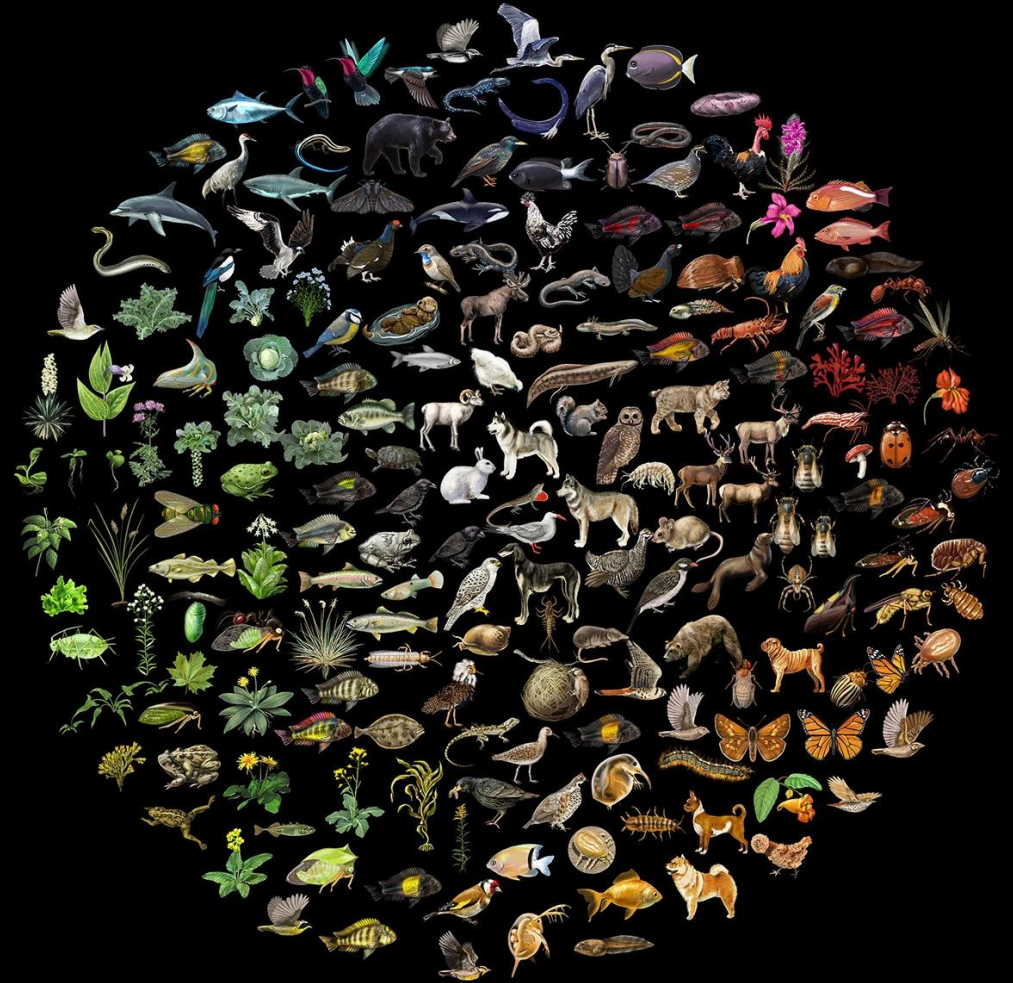
The first phylogenetic tree
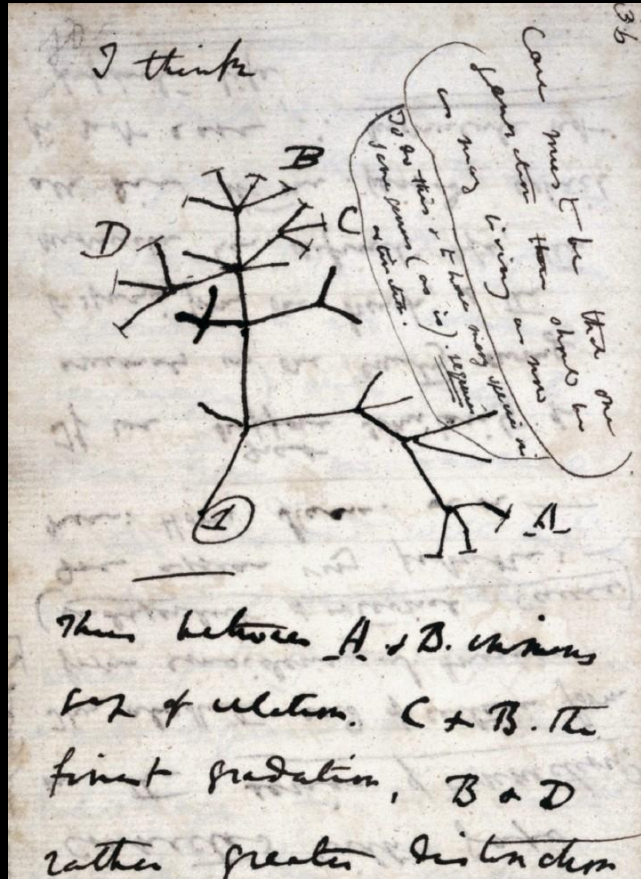
Darwin manuscript ms.dar.00121.38

Darwin's tree

On the Origin of Species (1859), Charles Darwin
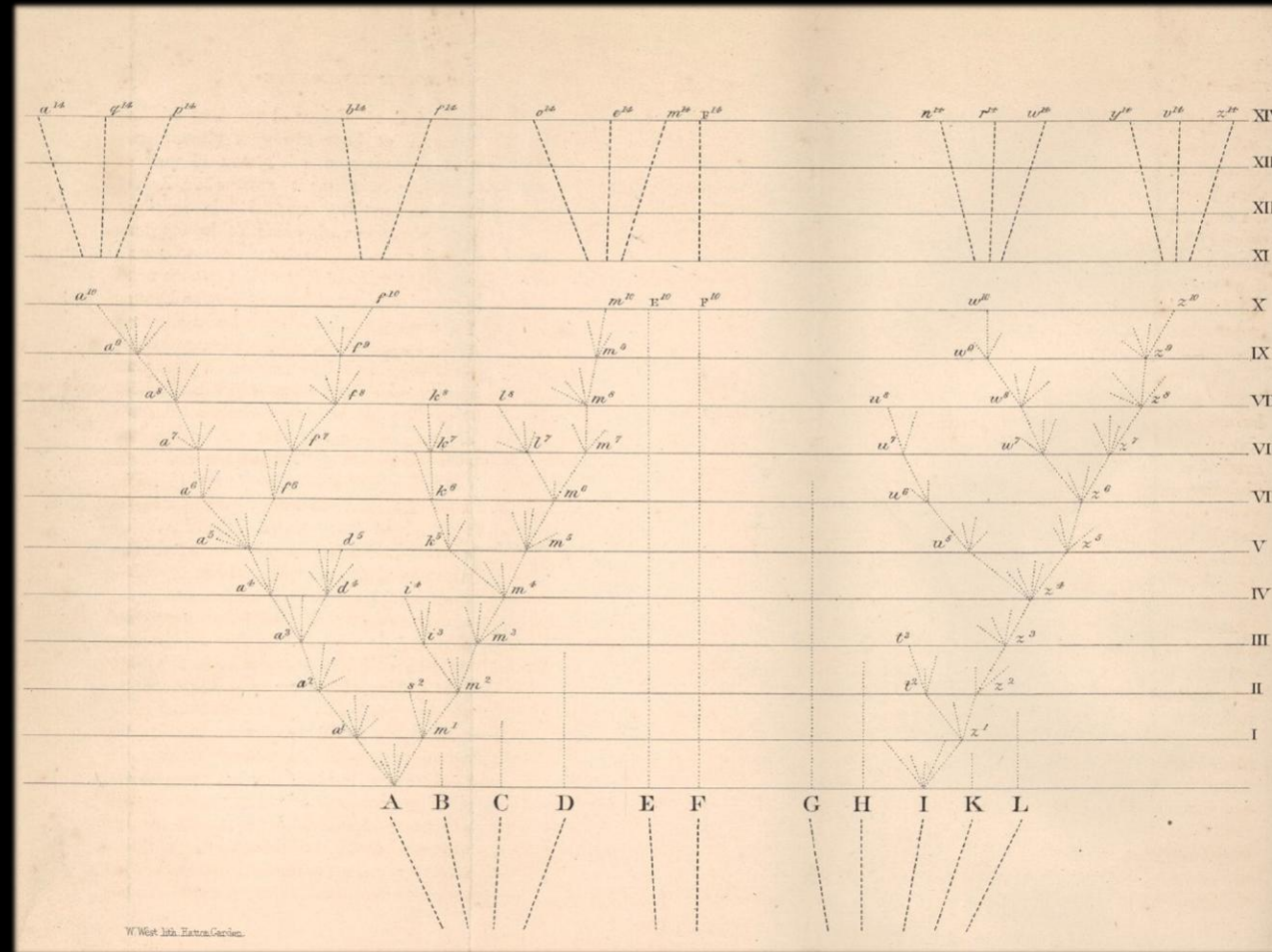
• Proposed that all living organisms share a common ancestor.

• Introduced the <u>Tree of Life</u> concept to illustrate evolutionary relationships.

"The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth"
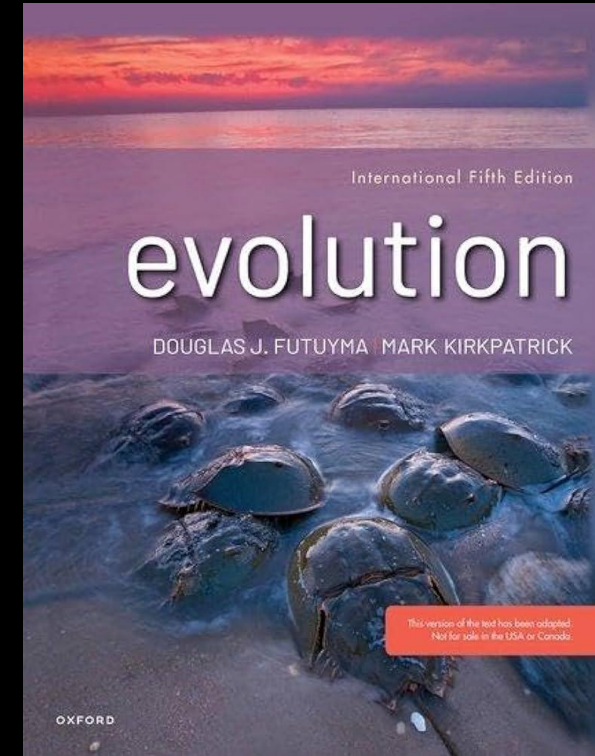
How many figures in "On the Origin of Species" ?

?

What is producing tree-like patterns?

Three main components:

1) Variation

2) Heritability

3) Differential reproductive success (selection)

Classification under an evolutionary framework

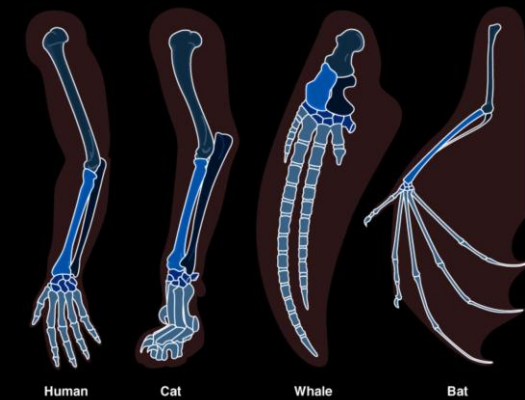1) You need features to compare!

- Morphological characters: bones, teeth, floral structures.

- Behavioral traits: mating behaviors, and ecological interactions.

- Cultural and linguistic data: Used in disciplines like anthropological phylogenetics.

2) These features need to have the same origin, or being homologous (similarities in traits between different biological entities that are due to inheritance from a common ancestor)

Analogy

Bird wing

Insect wing

Homology

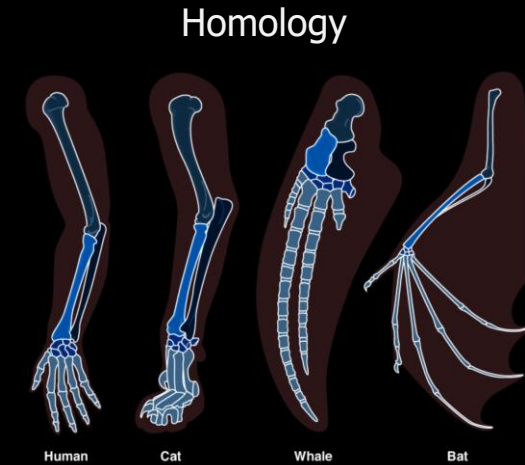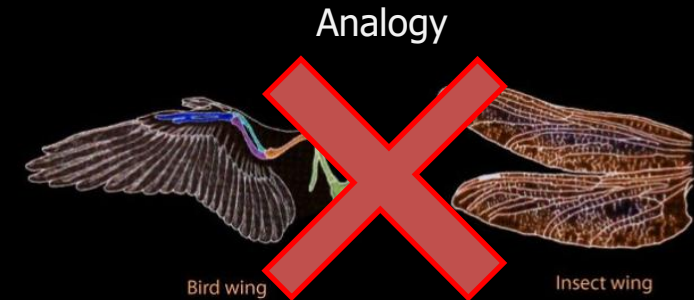Human   Cat   Whale   Bat

Classification under an evolutionary framework

1) You need features to compare!

- Morphological characters: bones, teeth, floral structures.

- Behavioral traits: mating behaviors, and ecological interactions.

- Cultural and linguistic data: Used in disciplines like anthropological phylogenetics.

2) These features need to have the same origin, or being homologous (similarities in traits between different biological entities that are due to inheritance from a common ancestor)

Analogy

Bird wing   Insect wing

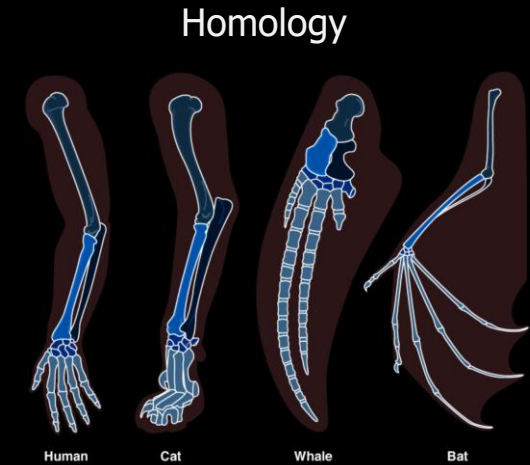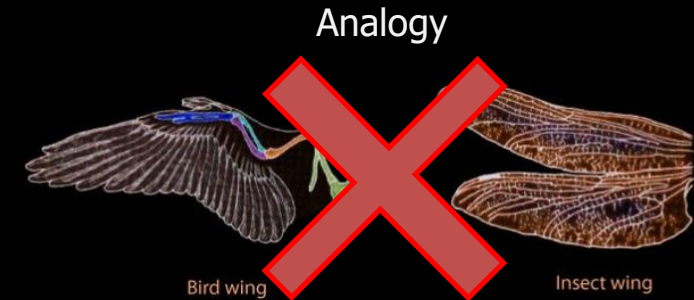Homology

Human   Cat   Whale   Bat

Classification under an evolutionary framework

1) You need features to compare!

- Morphological characters: bones, teeth, floral structures.

- Behavioral traits: mating behaviors, and ecological interactions.

- Cultural and linguistic data: Used in disciplines like anthropological phylogenetics.

2) These features need to have the same origin, or being homologous (similarities in traits between different biological entities that are due to inheritance from a common ancestor)
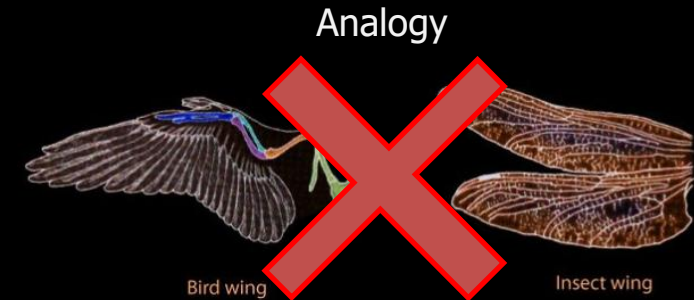
Can you think of a class of very abundant features that we might use?

Analogy

Bird wing

Insect wing

Homology

Human   Cat   Whale   Bat

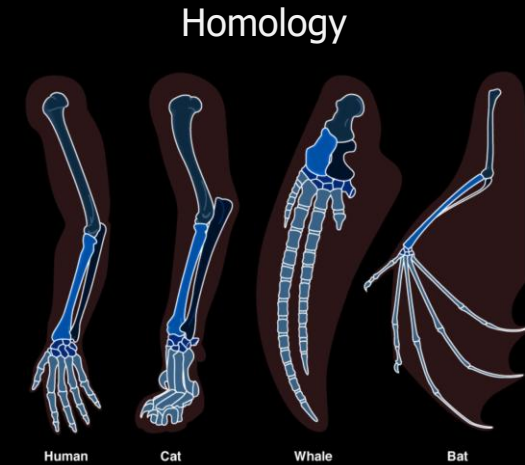# Classification under an evolutionary framework

1) You need features to compare!

- Morphological characters: bones, teeth, floral structures.

- Behavioral traits: mating behaviors, and ecological interactions.

- Cultural and linguistic data: Used in disciplines like anthropological phylogenetics.

2) These features need to have the same origin, or being homologous (similarities in traits between different biological entities that are due to inheritance from a common ancestor)

Can you think of a class of very abundant features that we might use?

Molecular traits: DNA, RNA, proteins

Analogy

Bird wing   Insect wing

Homology

Human   Cat   Whale   Bat

How can I build a phylogeny using molecules?

Alignment (nucleotides)

```
species 1   AGGATCTGCAATTGCTCTTCTAATCTGTCTGATCAGGAT
species 2   AGG-------AATTGCTCTTCTAATCTGTCT---CAGGAT
species 3   AGGATCTGCAATTGC---TCTAATCTGTCTGATCAGGAT
species 4   AGAATCTGCAATTGCTCTTCTGATCTGTCTGATCACGAT
species 5   AGGATCTGC---TGCTCTTCTGATCTGTCTGATCAGGAT
```
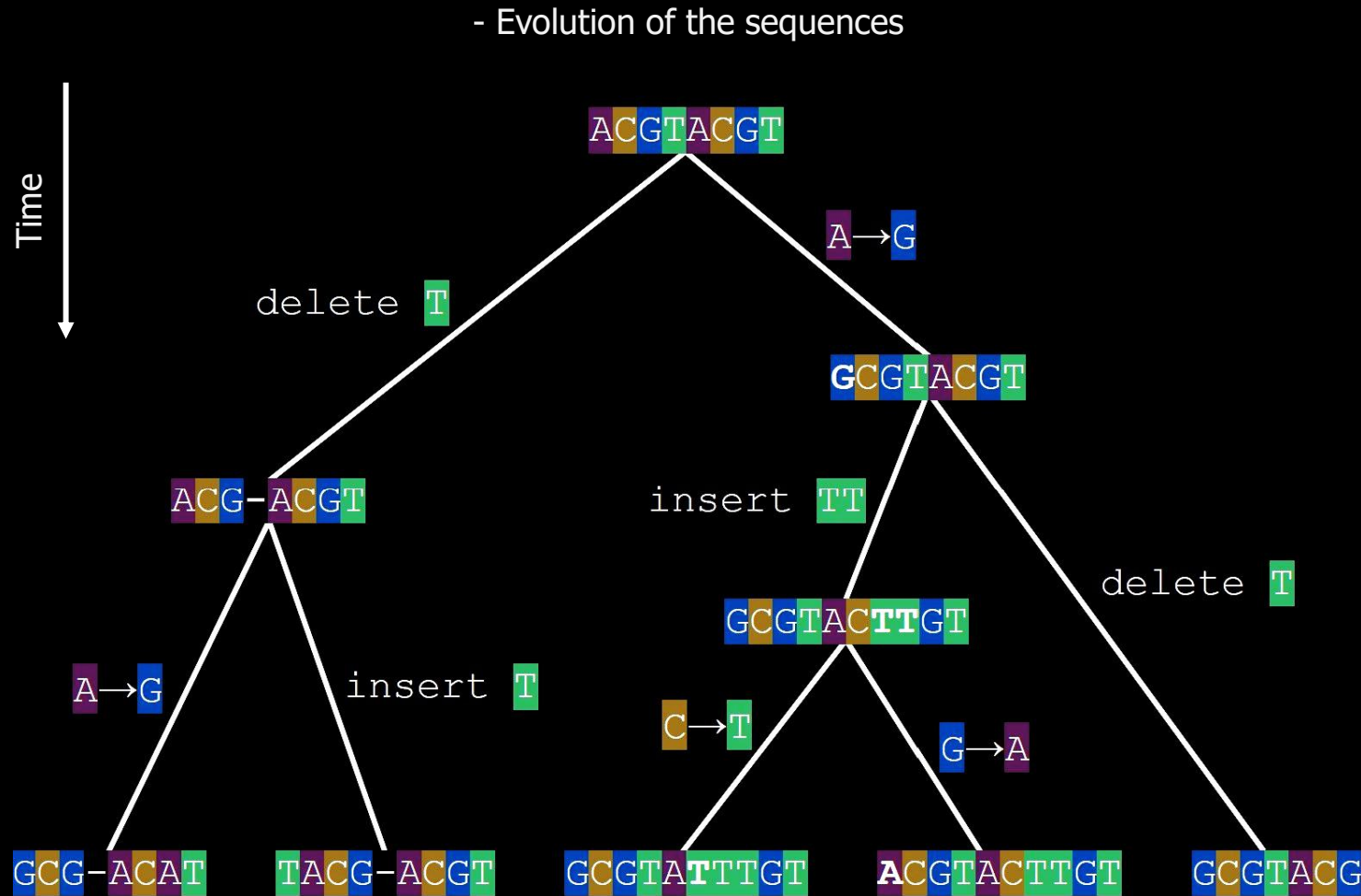
Alignment (nucleotides + amino acids)

```
species 1   AGGATCTGCAATTGCTCTTCTAATCTGTCTGATCAGGAT
            ValArgSerCysSerCysValArgSerCysValValSer
species 2   AGG-------AATTGCTCTTCTAATCTGTCT---CAGGAT
            Val-------CysSerCysValArgSerCys---ValSer
species 3   AGGATCTGCAATTGC---TCTAATCTGTCTGATCAGGAT
            ValArgSerCysSer---ValArgSerCysValValSer
species 4   AGAATCTGCAATTGCTCTTCTGATCTGTCTGATCACGAT
            TrpArgSerCysSerCysValCysSerCysValArgSer
species 5   AGGATCTGC---TGCTCTTCTGATCTGTCTGATCAGGAT
            ValArgSer---SerCysValCysSerCysValValSer
```

Types of alignment:

- Pairwise (1 VS 1)

- Multiple Sequence Alignment (MSA)

- <u>Global alignment</u>: Aligns entire sequences from end to end. This algorithm is ideal when query sequences are of similar length and are expected to share homology across their full length.

- <u>Local Alignment</u>: Aligns specific regions of sequences rather than the full length. It is ideal when sequences are globally dissimilar but contain localized similarities, such as motifs or conserved domains.

- Similarity scores; balance the number of gaps with number of mismatches

Sequences evolve on a tree

- Evolution of the sequences

ACGTACGT

Time

delete T

A→G

GCGTACGT

ACG-ACGT

insert TT

delete T

A→G

insert T

GCGTACTTGT

C→T

G→A

GCG-ACAT   TACG-ACGT   GCGTATTTGT   ACGTACTTGT   GCGTACG

- Sequences (sampled at present)

```
species 1    GCGACAT
species 2    TACGACGT
species 3    GCGTATTTGT
species 4    ACGTACTTGT
species 5    GCGTACG
```

- "True" alignment; nucleotides
are arranged to be orthologous

```
species 1    -GCG-AC-AT-
species 2    TACG-AC-GT-
species 3    -GCGTATTTGT
species 4    -ACGTACTTGT
species 5    -GCGTAC--G-
```

Do all changes occur with the same frequency?
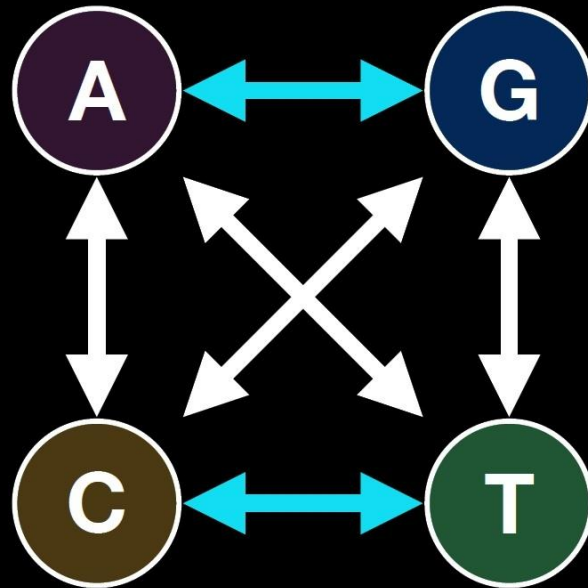
Transitions (Ts): Substitutions within the same nucleotide class A ↔ G (purines), C ↔ T (pyrimidines). More frequent than transversions due to structural similarity.

Transversions (Tv): Substitutions between different nucleotide classes (purine ↔ pyrimidine).
Examples: A ↔ C, A ↔ T, G ↔ C, G ↔ T. Less common due to greater structural change.

- Evolutionary models

- Most common (for nucleotides): General Time Reversible (GTR)

- Not only for nucleotides!

  - Codons (e.g. MG94)

  - Amino acids (e.g. PAM, WAG, LG)

Using the alignment differences to build trees

- <u>Distance-based methods</u> (e.g. UPGMA, Neighbor Joining)

Based on the total number of evolutionary changes between pairs of sequences. Starting from the alignment, these methods look at all possible pairs of the aligned sequences and count how many characters are different at each position. These pairwise differences are represented in a distance matrix. Best used for exploratory analysis of large datasets before conducting more intensive tree building using character base methods.

- Character-based methods (e.g. Maximum Parsimony, Maximum Likelihood, Bayesian Inference)
Compare all sequences by considering one character (nucleotides or amino acids) in the alignment at a time.
As character-based methods incorporate <u>evolutionary models</u> making these methods more accurate than distance-based methods.
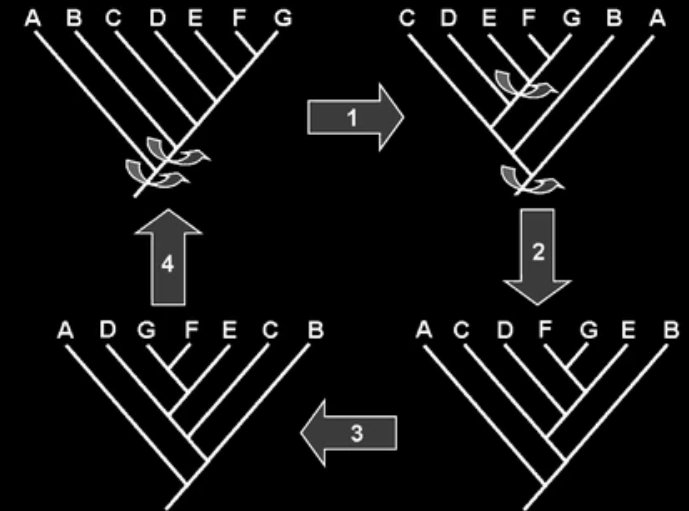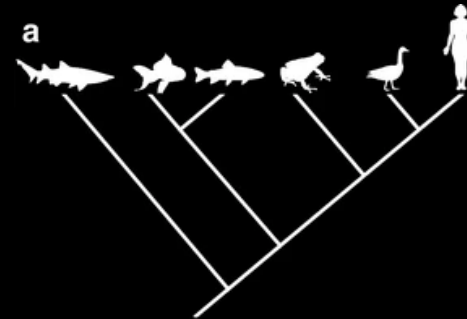Character-based phylogenetic inference is really about tree-scoring, not tree finding…

Reading trees

Clades

• Phylogenies indicate both relatedness and historical descent.
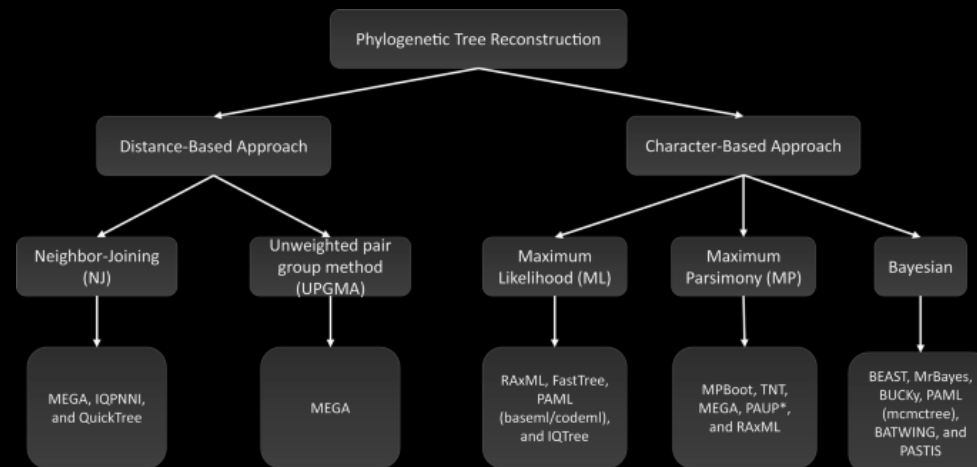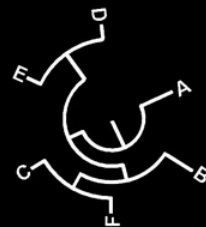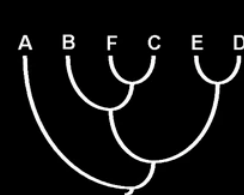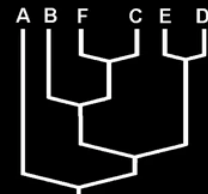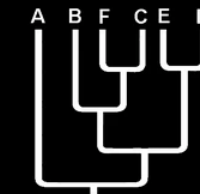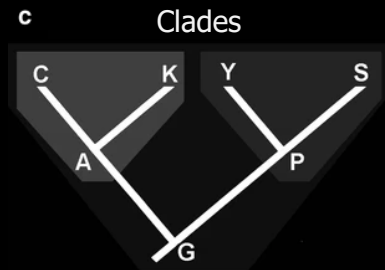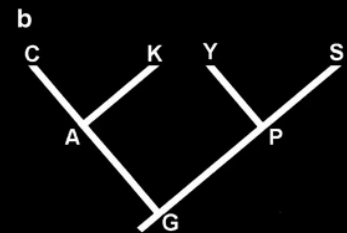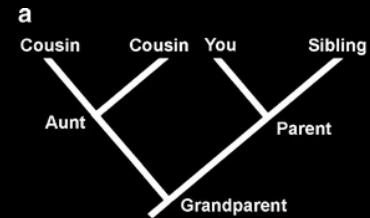
• Different ways to represent trees.

• Trees are not ladders!

• Many rotations, same tree.

Original Science/Evolution Review | Open access | Published: 12 February 2008
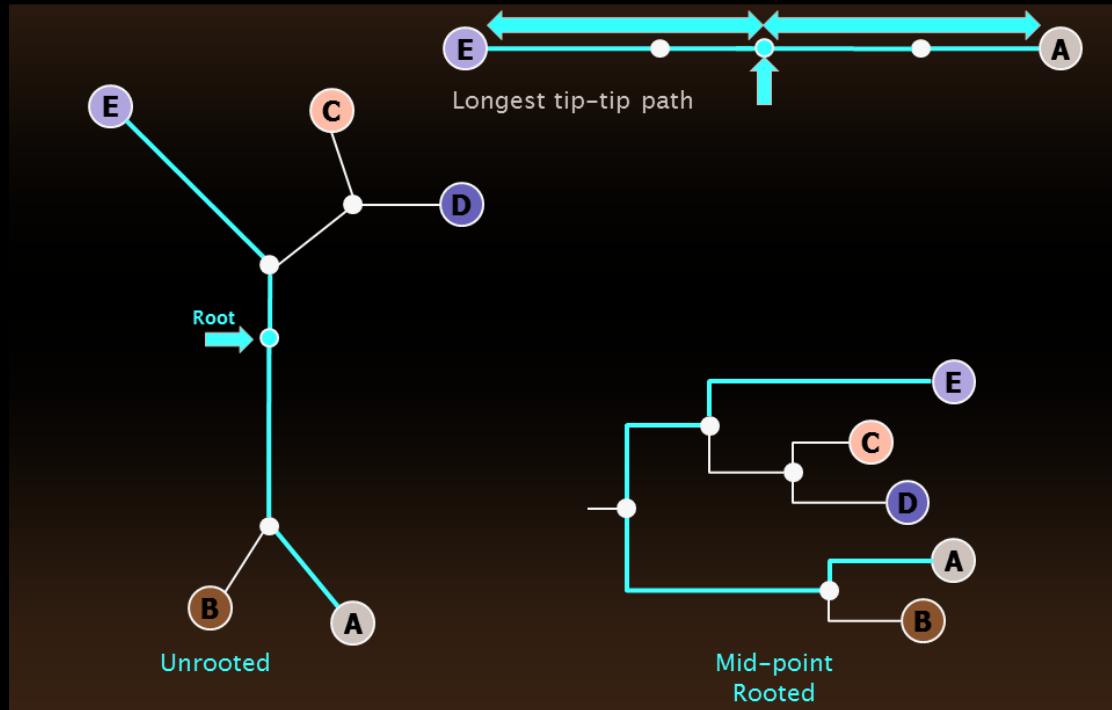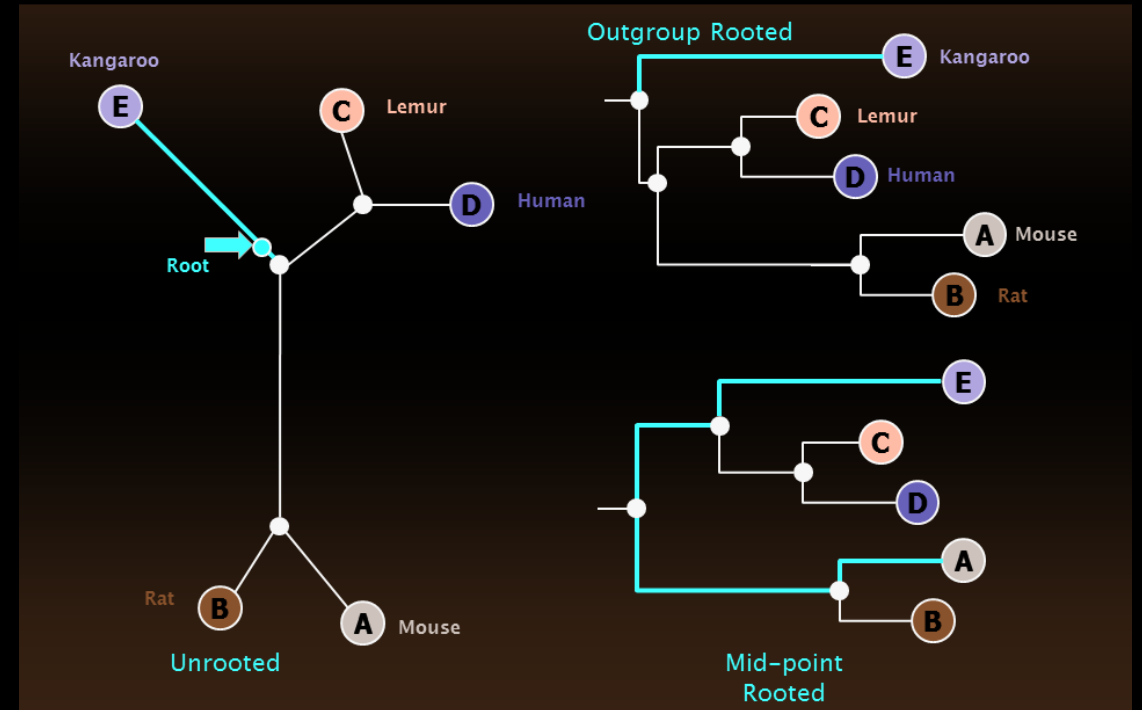
Understanding Evolutionary Trees

T. Ryan Gregory ✉

Evolution: Education and Outreach **1**, 121–137 (2008) | Cite this article

Rooting a tree

- Mid-point rooting

- Outgroup rooting

Quick quiz

Quick quiz

Steps of a phylogenetic analysis

1) Identify orthologous genes

2) Align genes

3) Trim alignments (optional)

4) Concatenate single-genes alignments

5) Model selection

6) Tree inference

https://github.com/andrea-silverj/BioinfoTutorials/tree/main/Phylogenetics101