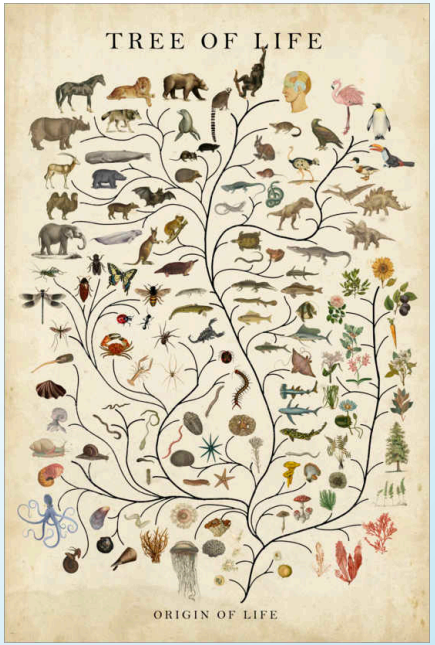
Phylogenetics: Recovering Evolutionary History

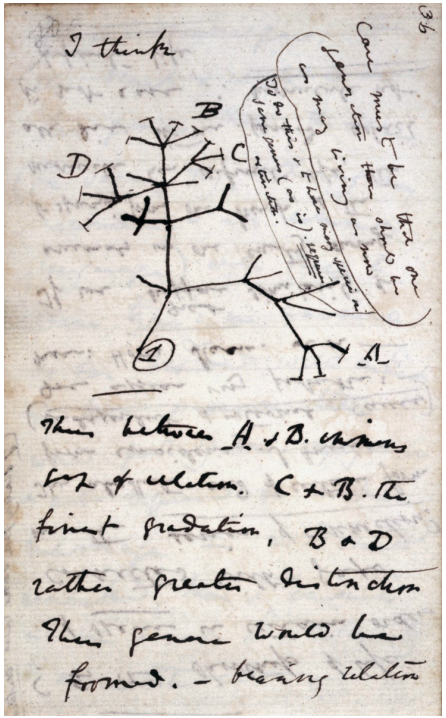
Hamim Zafar

BSE633



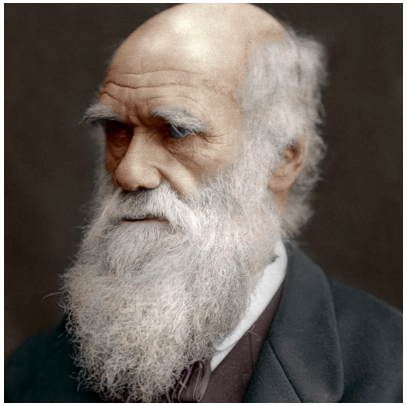
IIT Kanpur
Biological Sciences & Bioengineering

The Tree of Life

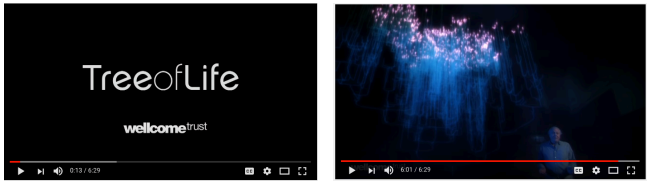


Charles Darwin's 1837 sketch, his first diagram of an evolutionary tree from his first Notebook on Transmutation of Species (1837).

> 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.
>
> — *Charles Darwin*

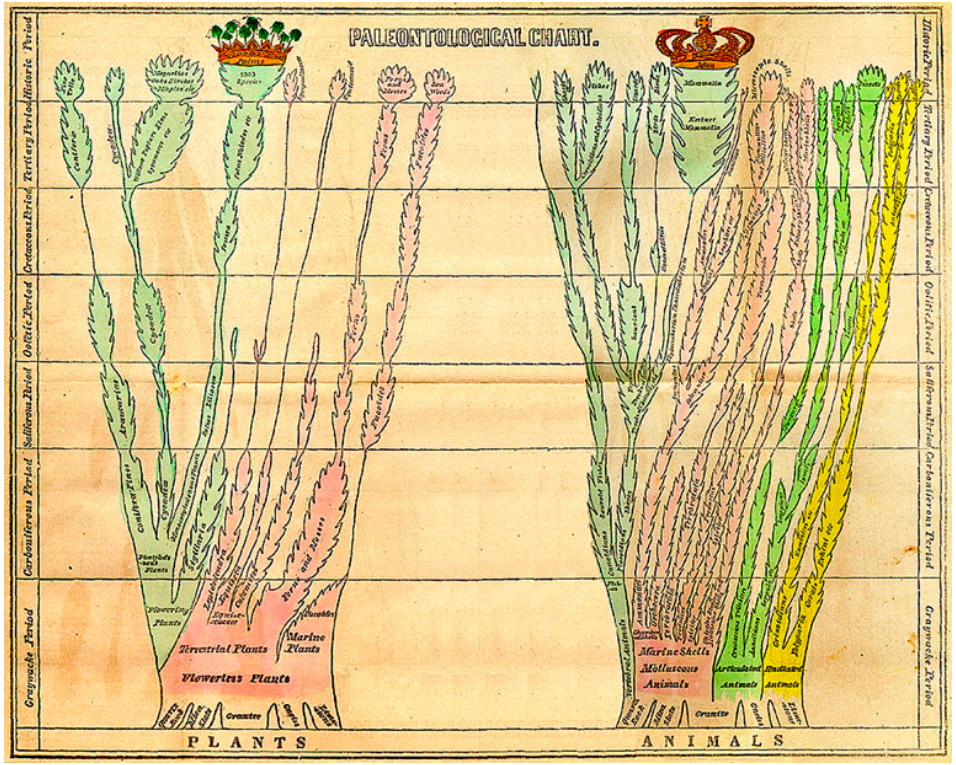


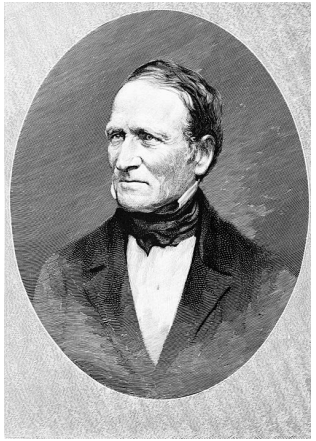
- All organisms on Earth had a common ancestor
- Any set of species is related



https://www.youtube.com/watch?v=qabl5eIba2g

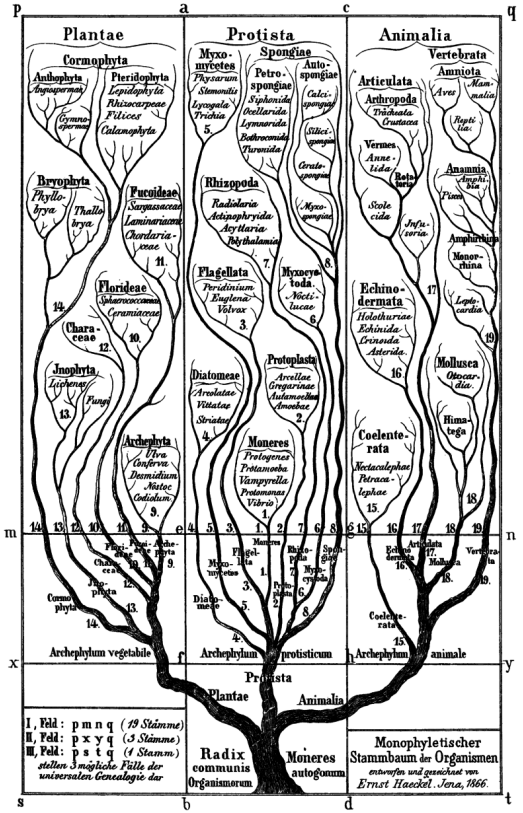
Early Phylogenetic Trees

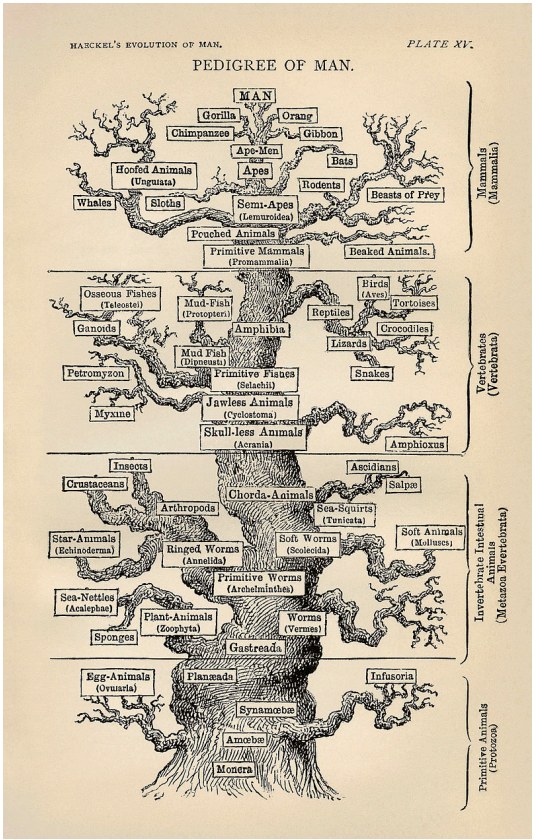




Paleontological Chart in the publication 'Elementary Geology' (1840) by Edward Hitchcock

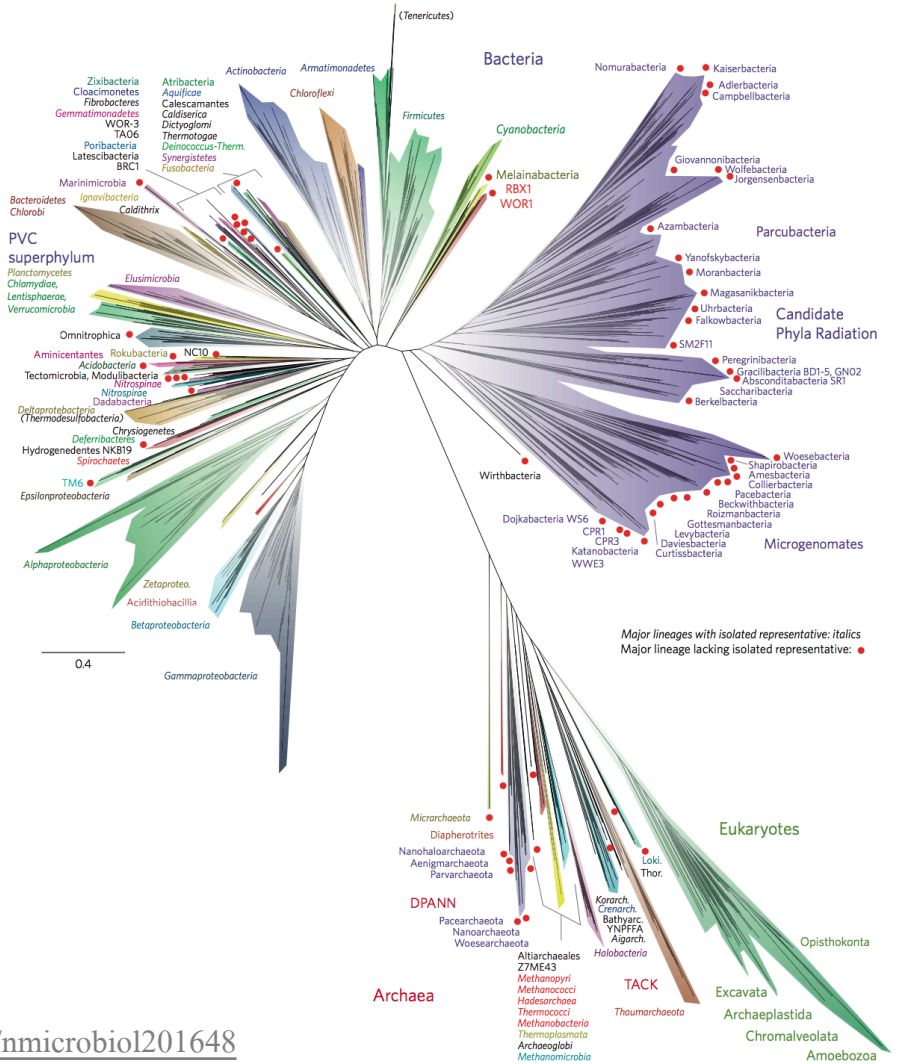
Ernst Haeckel Coined the term "Phylogeny"





Current View of the Tree of Life

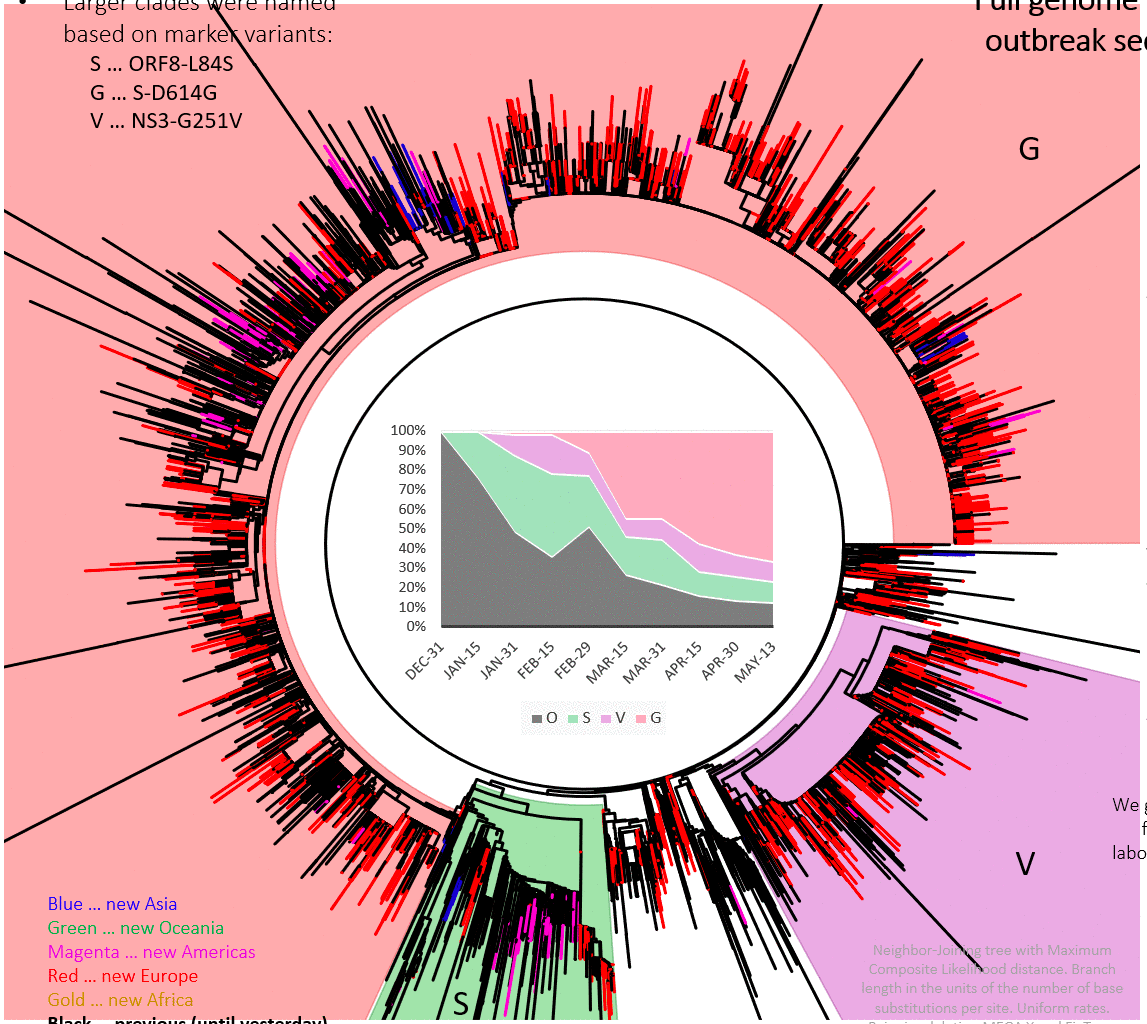
- Total diversity represented by sequenced genomes.
- Constructed from molecular information
- Includes 92 named bacterial phyla, 26 archaeal phyla and all five of the Eukaryotic supergroups



Source: https://www.nature.com/articles/nmicrobiol201648

SARS-CoV-2 Phylogeny

- Larger clades were named based on marker variants:
  S … ORF8-L84S
  G … S-D614G
  V … NS3-G251V



Full genome tree derived from all outbreak sequences 2020-05-13

Notable changes:
20,114 full genomes (+3,447) (excluding low coverage, out of 21,554 entries)

S clade 2,148 (+115):
38 USA/WA, 36 Scotland, 22 England, 10 Saudi Arabia, 5 Austria, 3 USA/MD, 1 Greece
G clade 13,321 (+2,603):
1685 England, 444 Scotland, 184 Austria, 89 USA/WA, 61 Saudi Arabia, 35 USA/VA, 28 Greece, 19 Japan, 17 USA/MD, 11 USA/ID, 9 USA/CT, 5 Italy, 4 USA/UN, 4 Ireland, 2 France, 2 USA/DC, 1 Bangladesh, 1 USA/NC, 1 USA/NY, 1 USA/AK
V clade 2,067 (+473):
416 England, 41 Scotland, 7 Greece, 5 USA/VA, 2 Austria, 2 USA/MD
Other clades 2,578 (+256): 184 England, 50 Scotland, 12 Austria, 4 Japan, 2 Greece, 1 Saudi Arabia, 1 USA/WA, 1 India, 1 Ireland

We gratefully acknowledge the Authors from Originating and Submitting laboratories of sequence data on which the analysis is based.

GISAID

by BII/GIS, A*STAR Singapore

Phylogenetic Tree

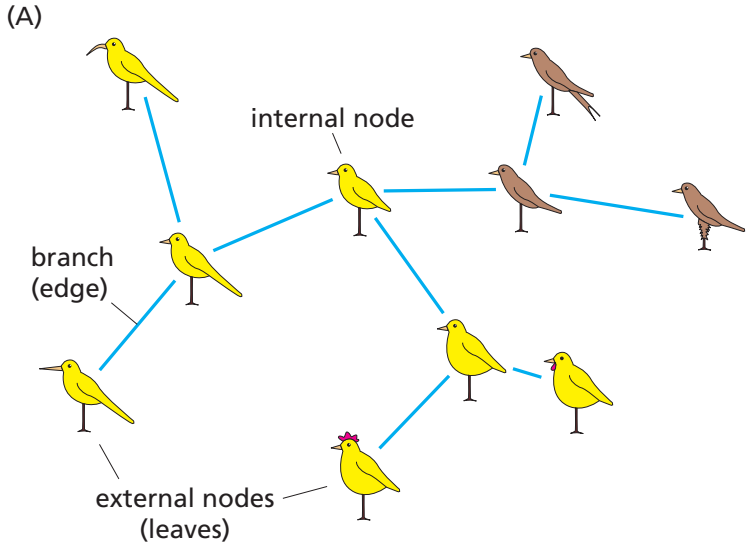
- When a group of aligned sequences shows significant similarity to each other, this can usually be taken as evidence that they are the result of divergent evolution from a common ancestral sequence.
- Sequence alignment will contain traces of the evolutionary history of these sequences.
- the evolutionary history of a set of sequences can be represented as a graphical structure called a **phylogenetic tree**.
- By studying sequences that have both a common ancestor and common function—known as **orthologous sequences** or **orthologs**—from different species, one can investigate the evolutionary relationships between species.

Structure and Interpretation of Phylogenetic Trees

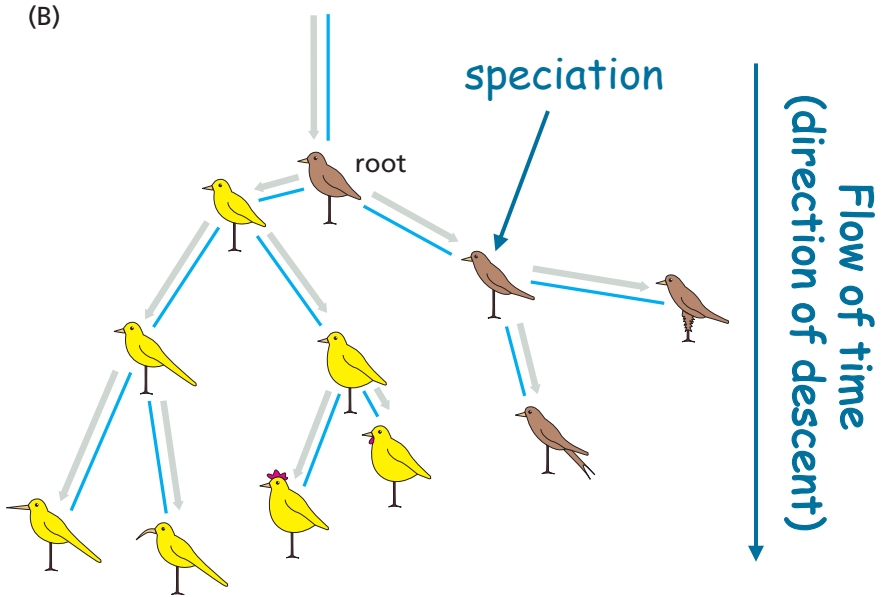
- A phylogenetic tree is a diagram that proposes a hypothesis for the reconstructed evolutionary relationships between a set of objects—which provide the data from which the tree is constructed.
  - 2D graphical structure
- These objects are referred to as the **taxa** (singular **taxon**) or **operational taxonomic units (OTUs)**
  - In phylogenies based on sequence data they are the individual genes or proteins. **Gene tree**
- When orthologous sequences from different species are being used with the aim of determining relationships between species, the taxa are labeled with the species name. Such trees are called **species trees**

Structure and Interpretation of Phylogenetic Trees

Unrooted, binary species tree



Rooted, binary species tree



Branch: evolutionary relationships between species

Leaves: existing species/extinct species whose lineage died out without leaving any descendants

Internal nodes: ancestral states that are hypothesized to have occurred during evolution

Structure and Interpretation of Phylogenetic Trees

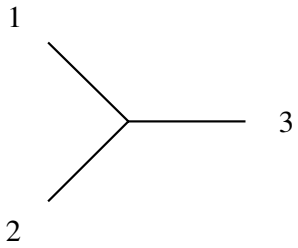
- **Binary tree/Bifurcating tree**: Every internal node is of degree 3 (connects to 3 others) and every leaf is of degree 1 (connects to only one other node)
- **Multifurcating tree:** Can have some internal nodes of higher degree.
- In a binary tree on n taxa, how may nodes, branches, internal nodes and internal branches are there?
- Two components in representing a phylogenetic tree
  - **tree topology** or the way it branches
  - **branch lengths**

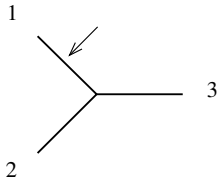
Number of Tree Topologies

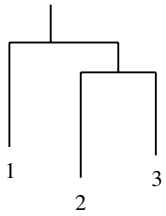
Number of Taxa	Number of unrooted trees	Number of rooted trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

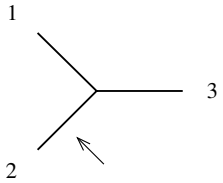
$$n \qquad \frac{(2n-5)!}{2^{n-3}(n-3)!} \qquad \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

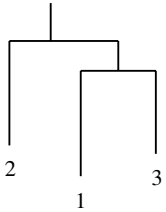
Different possible rooted tree

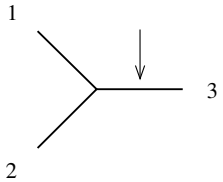


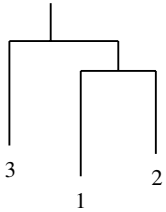


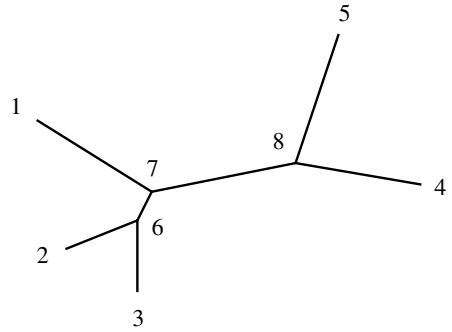






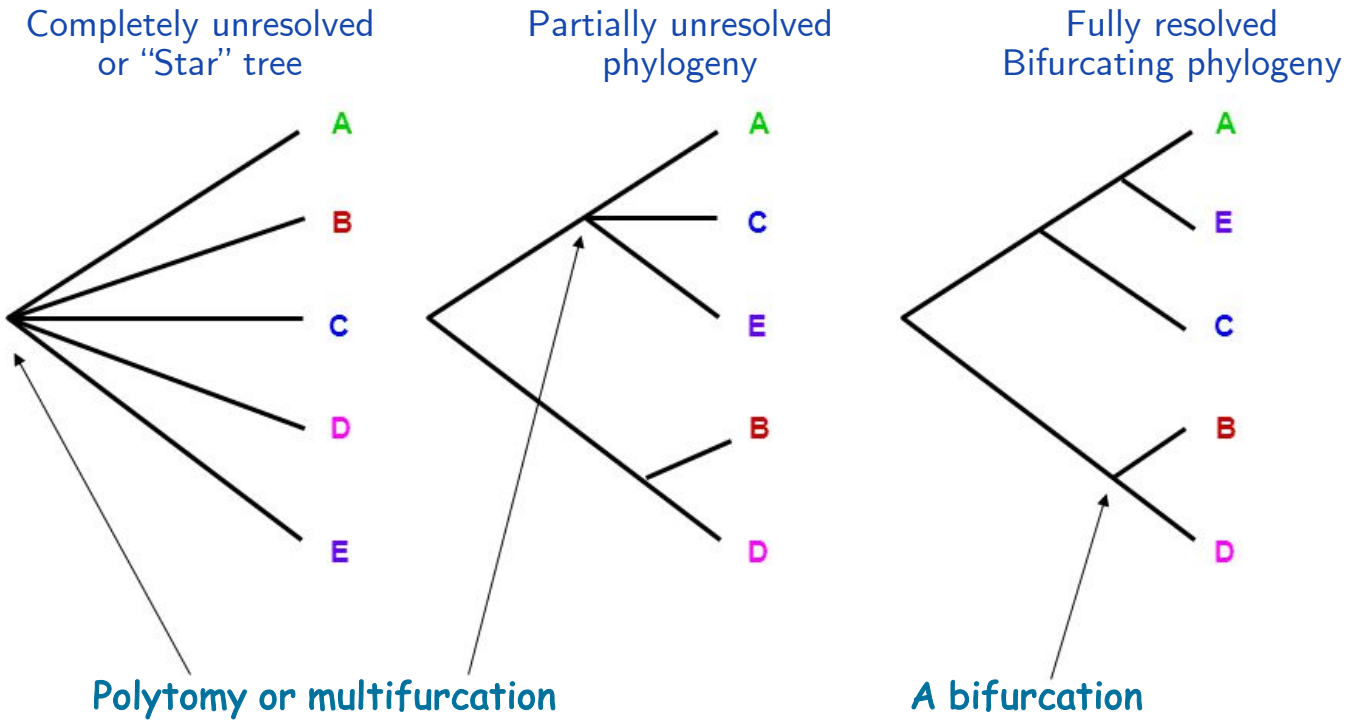






How many rooted tree?

Tree Resolution

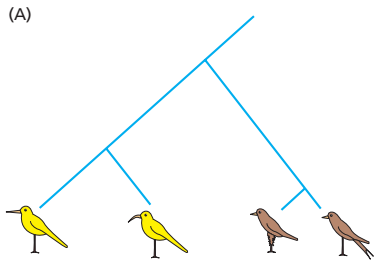


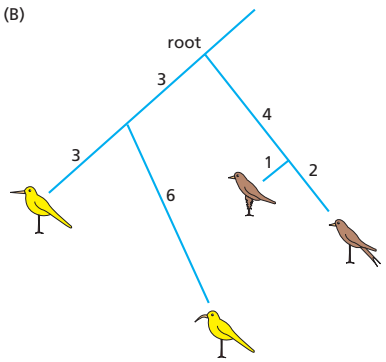
- A **polytomy** in a tree can be **resolved** (not necessarily fully) in many ways, thus producing trees with higher resolution
- A binary tree can be turned into a partially resolved tree by **contracting** edges

Variety of Types of Tree

Rooted cladogram

- branch lengths have no meaning
- ancestors only implied by the internal nodes



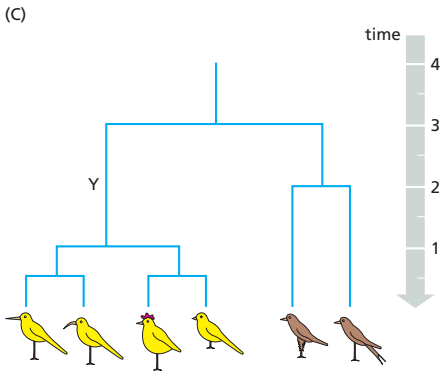


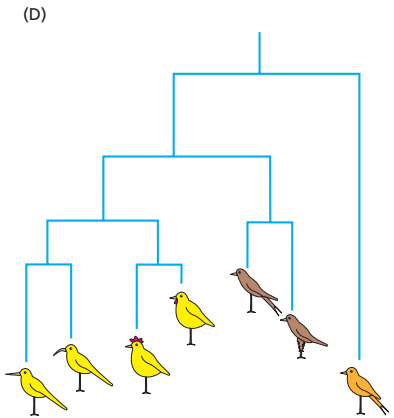
Additive tree

- branch lengths: measure of evolution
- proportional to the number of mutations per site

Ultrametric tree

- **molecular clock**: same constant rate of mutation assumed along all branches
- evolutionary distance from a common ancestor to all its descendants is the same



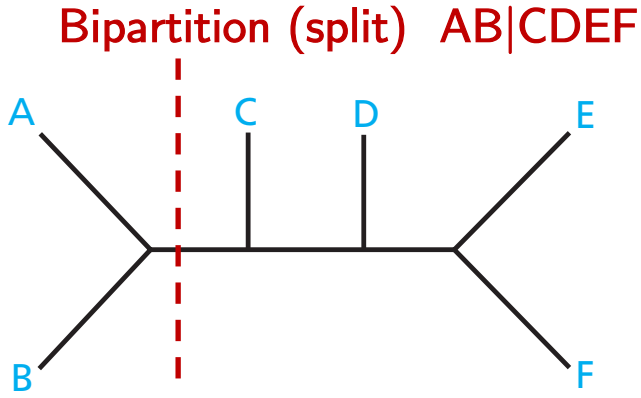


Additive tree rooted at an outgroup

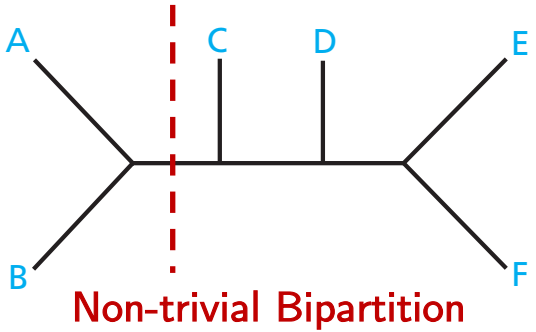
Root using a group of homologous sequences from species or genes that are distantly related to the main set of species or genes under study

Representing Tree Topology using Splits

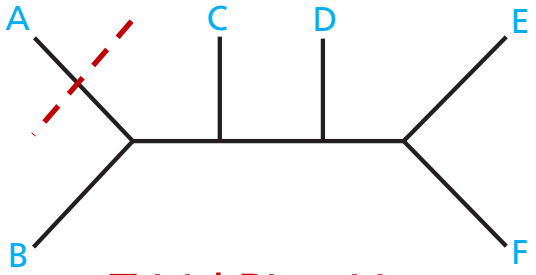
- The graphical views of trees are convenient for human visual interpretation, but not for other tasks such as comparison.
- Subdivide or split it into a collection of subgroups. Every branch in a tree connects two nodes, and if that branch is removed, the tree is divided into two parts
- Such a division is called a **split** or **bipartition**, and a tree contains as many splits as there are branches.



Representing Tree Topology using Splits

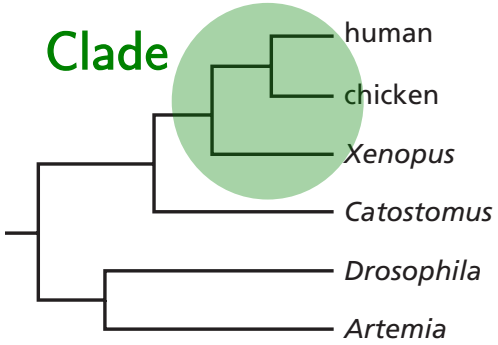


Non-trivial Bipartition



Trivial Bipartition

9 bipartitions (3 nontrivial)

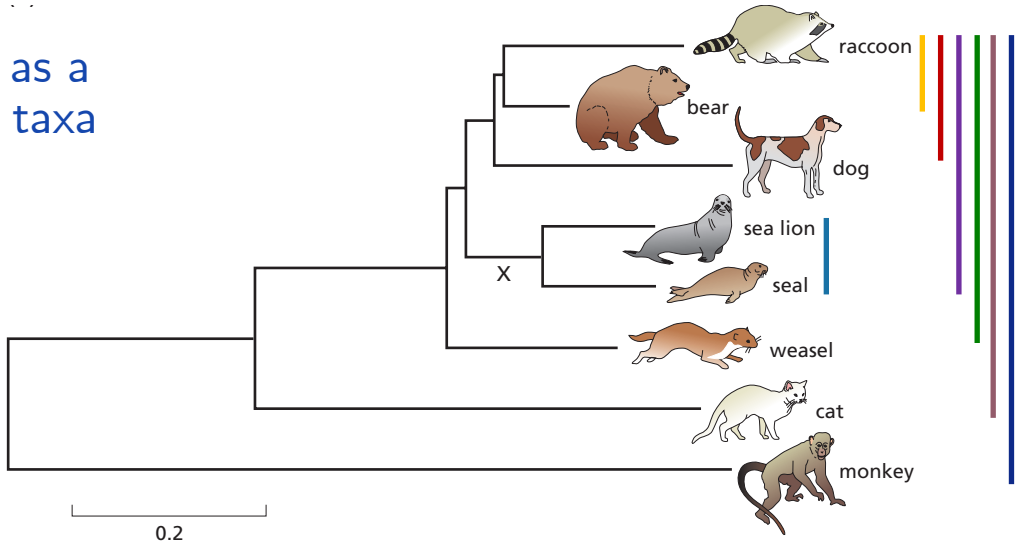


11 clades (4 nontrivial)

Clade in Rooted Tree: Group of leaves that form a **monophyletic** group meaning they have a common ancestor that is not a common ancestor for any other leaf in the **tree**.

The Newick or New Hampshire Format

Each split is written as a bracketed list of the taxa



(monkey,(cat,(weasel,((seal,sea_lion),(dog,(bear,raccoon))))));

End of Tree

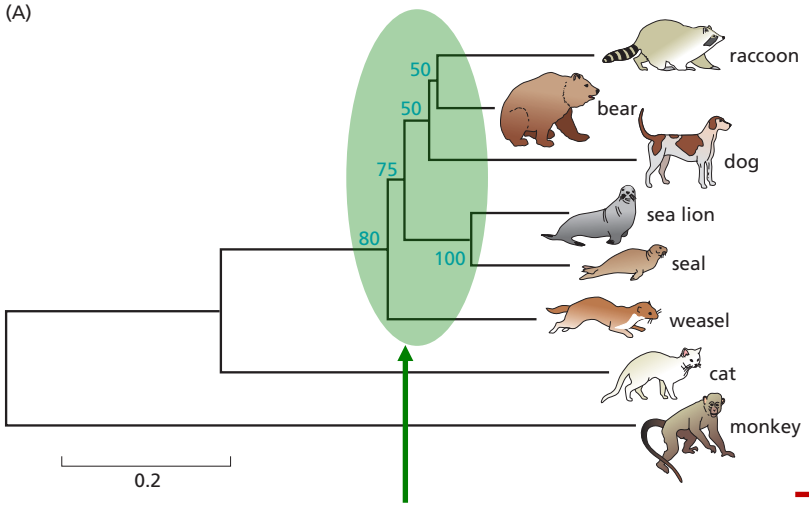
More information, such as branch lengths, can be added

(monkey:1,(cat:0.47,(weasel:0.18,((seal:0.1,sea_lion :0.1):0.08,(dog:0.2,(bear:0.1,raccoon:0.2):0.01):0.1):0.15):0.12):0.2);

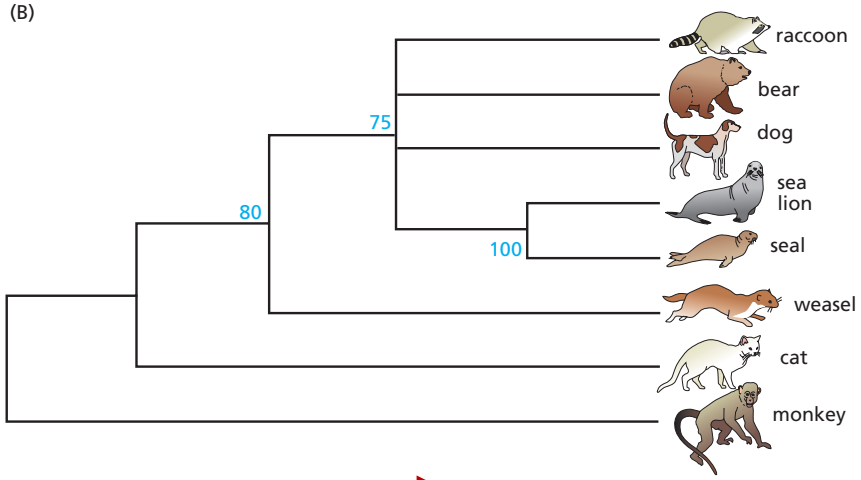
Consensus and Condensed Trees

- A set of trees might be regarded as equally representative of the data.
  - Trees may differ in their topology and branch lengths.
- When the trees have been produced by several methods of tree construction applied to a single set of data, or by the same method applied to several different datasets, all trees are treated equally in the analysis.
- The **bootstrap** procedure assigns values to individual branches that indicate whether their associated splits are well supported by the data.
- All internal branches that are not highly supported by the bootstrap are removed. Such a tree is called a **condensed tree**
  - Multifurcating internal nodes.

Consensus and Condensed Trees



Bootstrap support values



Contracting edges with poor support (bootstrap value < 60%)

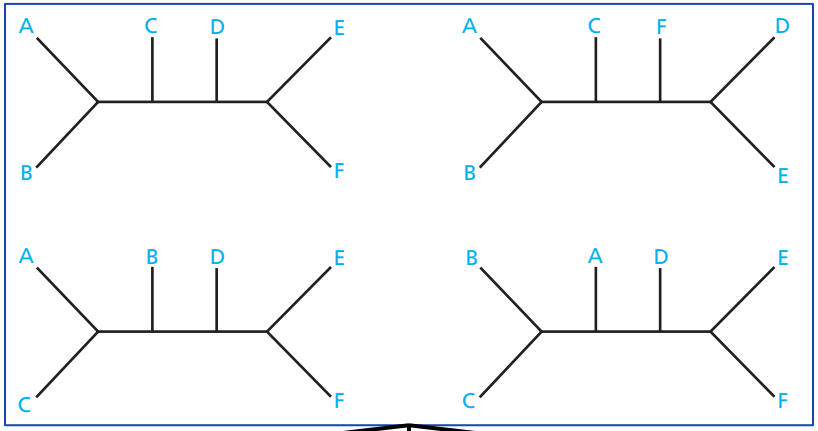
Consensus Tree

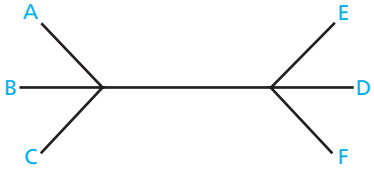
- **Consensus features:** features that are always (or frequently) observed when several equally well-supported trees are obtained from the same data
- Can also be useful when trees obtained using different data are expected to reveal essentially the same evolutionary history.
- Certain sequences may always group together
- The grouping itself may be the most informative feature

- **Consensus tree:** Retains only sufficiently commonly occurring topological features

Consensus Tree

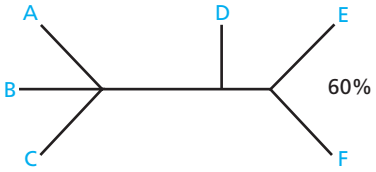
Consensus of sets of trees





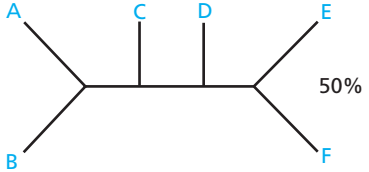
Strict Consensus

Splits that occur in all trees



Majority-rule Consensus (60%)

Splits that occur in at least 60% of the trees and majority



Majority-rule Consensus (50%)

Splits that occur in at least 50% of the trees and majority

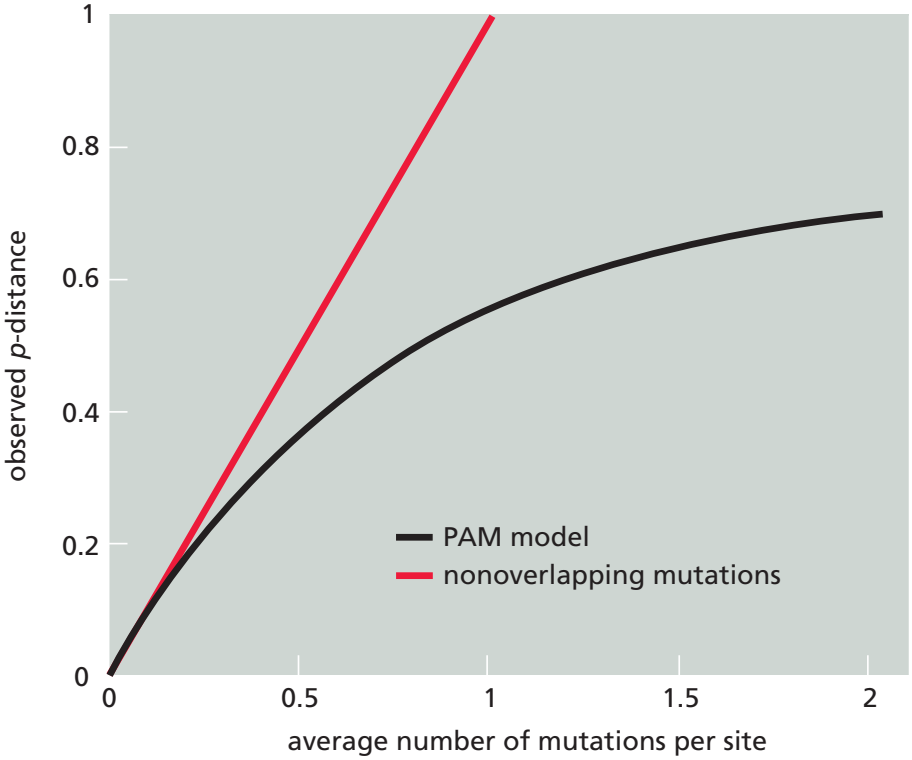
Molecular Evolution and its Consequences

- The Darwinian concept of evolution by natural selection concentrates on the consequences of evolutionary changes for the fitness of the organism: its ability to survive and transmit its genes to the next generation by producing offsprings
- Fitness depends on the properties of the organism as a whole, and thus change at the DNA sequence level will be constrained by considerations of how it affects protein expression and function, and how these affect cellular properties and whole organism physiology and behavior
- In the case of sequence-based phylogenetic reconstruction, the changes that occur in genomic DNA are the main focus and, where relevant, their effects on amino acid sequence
- Changes may affect one or a few nucleotides, entire genes, and even whole genomes

Multiple Mutations at a Site

- Most related sequences have many positions that have mutated several times
  - Even apparently conserved bases may in the past have mutated to a base that subsequently mutated back to the original base; any such pairs of mutations that have occurred are undetectable from the sequence alignment
- The **p distance** between two sequences is the fraction of nonidentical alignment positions
  - estimates the evolutionary distance
- The p distance is almost always an underestimate of the number of mutations that actually occurred
- This distance is, therefore, **corrected** to reflect the correct evolutionary distance

Multiple Mutations at a Site

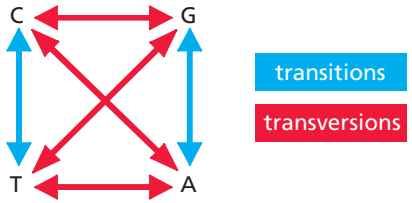


The number of observed mutations is often significantly less than the actual number of mutations because of overlapping mutations.

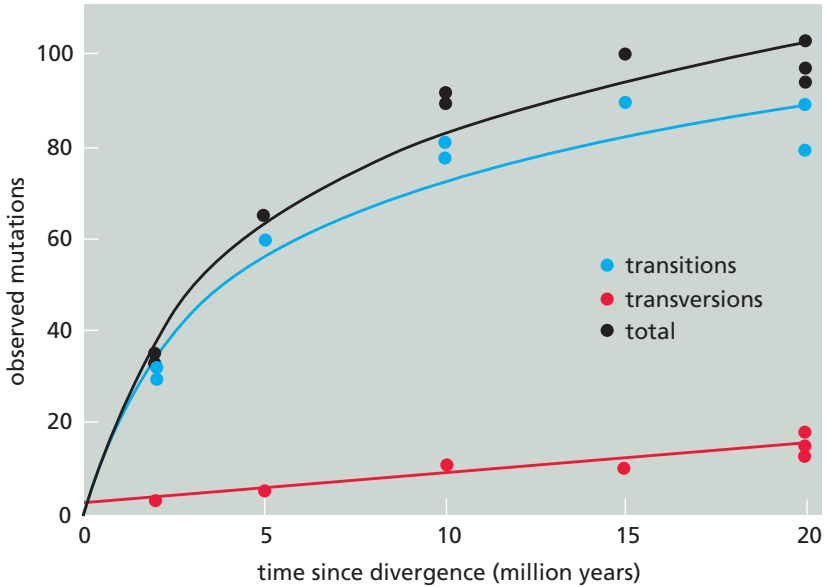
The red line represents the *p*-distance—the fraction of nonidentical sites in an alignment— that would be observed if each site only received one mutation at most. The observed *p*-distance in an alignment is plotted (black line) against the average number of mutations at each site as calculated by the PAM model.

Transitions vs. Transversions

The rate of accepted mutation is usually not the same for all types of base substitutions



Transition mutations have little effect on the DNA structure, and hence are much more commonly observed



transition/transversion ratio, R:

The number of transitions per transversion during the evolution of the sequences being studied

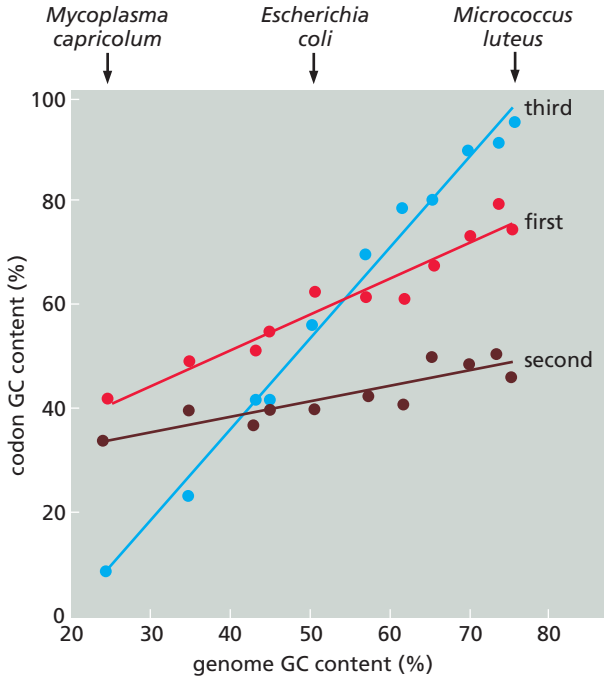
value of R relates to mutations accepted during evolution

Synonymous vs. Nonsynonymous Substitutions

- Another factor that affects the acceptance rates of mutations in protein-coding sequences is the **effect of the mutation on the amino acid sequence**, and thus, potentially, on the function of the protein
- Nucleotide mutations that do not change the encoded amino acid are called **synonymous mutations** (most changes at the third codon position are synonymous)
- Synonymous mutations are generally considered to be neutral (to have no effect)
- Nucleotide mutations that alter the encoded amino acid are called **nonsynonymous mutations**
  - The protein product will have a different sequence, and may thus have altered properties

Synonymous vs. Nonsynonymous Substitutions

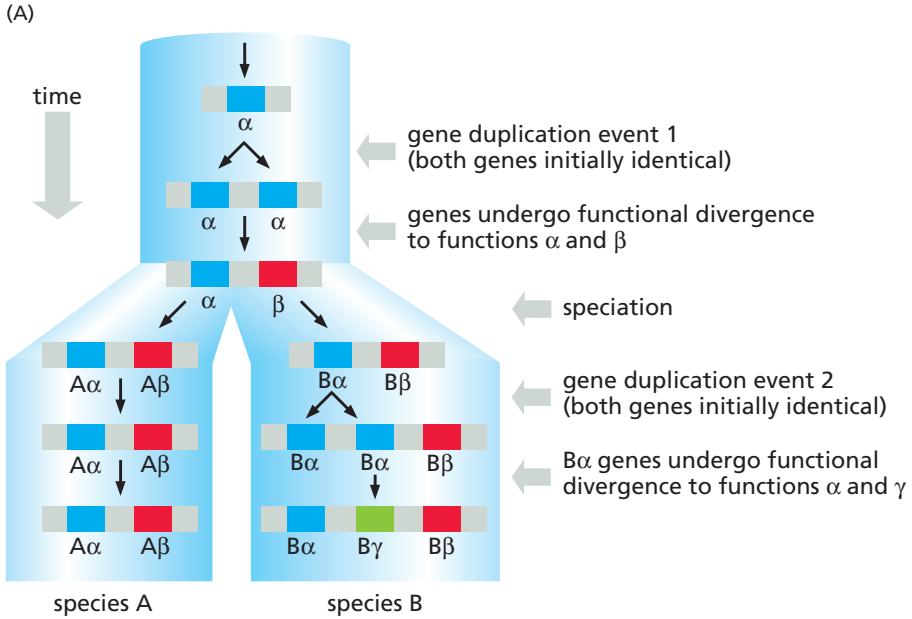
- Because nucleotide substitutions at the third codon position are almost always synonymous, the accepted mutation rate at these sites would be expected to be higher than at the first and second positions
- The fact that almost all third-position mutations are synonymous is also shown by the phenomenon of **biased mutation pressure**
- While all three codon positions adapt to some extent to the compositional bias of the genome, the third position adapts most.
- Results in a more extreme percentage GC value at the third codon position than for the overall 40 genome.

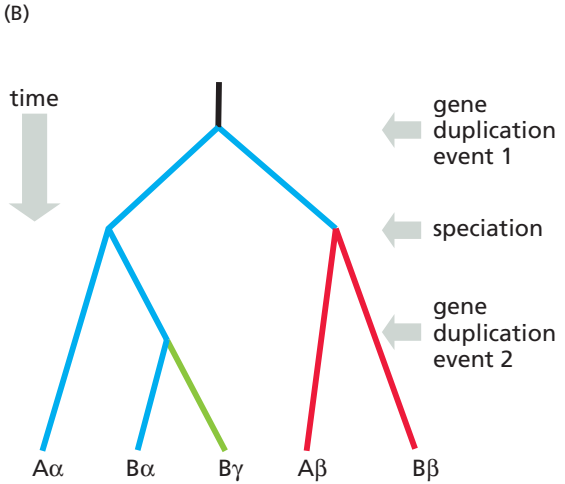


Types of Homology

- Homologous genes can arise through a variety of different biological processes.
- One is by **speciation**, which results in two homologous genes diverging in different lineages. Pairs of homologous genes derived this way are described as orthologous and called **orthologs**.
- Orthologs can be formally defined as pairs of genes whose last common ancestor occurred immediately before a speciation event
- Another way in which homologous genes arise is by **gene duplication**, the process by which a gene becomes copied, within the same genome.
- A pair of genes arising from a gene duplication event is described as **paralogous**, and are called **paralogs**, which can be more formally defined as a pair of genes whose most recent common ancestor occurred immediately before a gene duplication event

Orthology vs. Paralogy





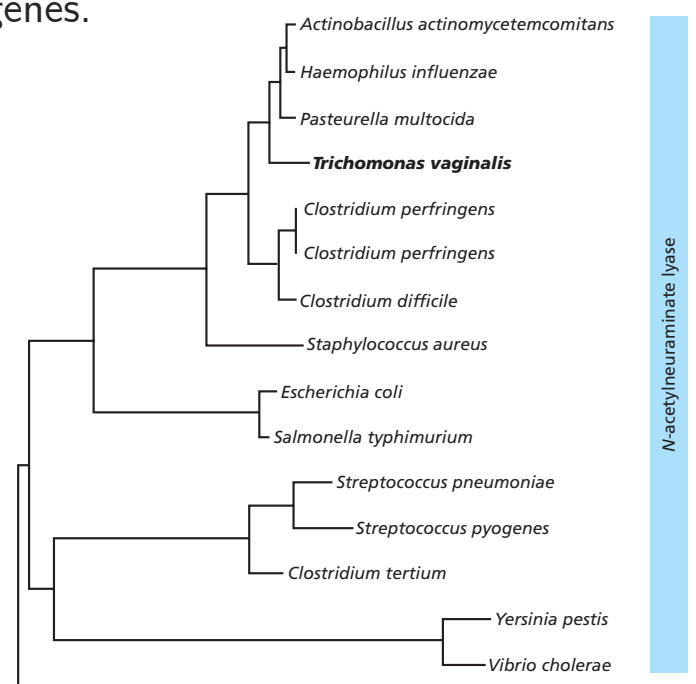
At the end of this period of evolution, all five genes in both species are homologous, with three orthologous pairs: $A\beta/B\beta, A\alpha/B\alpha, and\ A\alpha/B\gamma$. The $B\alpha$ and $B\gamma$ genes are paralogous. $A\alpha$ and $B\gamma$ are orthologs despite their different functions. To study the evolution of a the α function, we need to distinguish the $A\alpha/B\alpha$ from $A\alpha/B\gamma$. Sequence similarity would be expected to be greater for the $A\alpha/B\alpha$ pair as they will be evolving under almost identical evolutionary pressures.

Orthology

- Only orthologous sequences will identify the speciation times
- Species phylogenetic trees should ideally be constructed using only orthologous sequences
- Although orthologs often have the same function, not all proteins with identical or similar function are orthologs, or even homologs.
- Unrelated nonhomologous genes can sometimes develop equivalent functions as the result of convergent evolution, although sequence identity between such genes is usually very limited
- Sequence similarities that are not due to homology are known generally as **homoplasy**
- Convergent evolution is just one cause of homoplasy; others are parallel evolution and evolutionary reversal

Horizontal Gene Transfer (HGT)

- **horizontal gene transfer (HGT)/ lateral gene transfer (LGT):** involves a gene from one species being transferred into another species
- Shortly after HGT, the sequence of the gene in the donor and recipient species will be very similar. Such pairs of genes are called **xenologous** genes.
- If such genes are included in a standard phylogenetic analysis, the resultant tree will have the **gene from the recipient species appear in a much closer relationship to that of the donor** species than should be the case
- If an extra 24-residue N-terminal sequence is ignored, there is 80% identity at the amino acid level between the *Trichomonas* sequence and the sequences from the bacteria *Actinobacillus* and *Haemophilus*.



De novo Identification of Orthologues

- Tree Based
  - Identifies orthologues by aligning homologous sequences and reconstructing a tree to find those that are most plausibly related by speciation rather than by duplication or HGT
  - Requires entire gene families comprising hundreds of sequences
  - Gene family relationships may be further obscured if other processes causing gene-tree discordance are not accounted for
    - Incomplete Lineage Sorting
    - HGT
    - hybridization, introgression and non-allelic gene conversion
- Graph Based
  - Assumption: a gene in one species should be more similar to its orthologue than to any other gene in a second species and vice versa
  - all-against-all pairwise sequence comparisons mostly performed using BLAST for defining sequence similarity
  - computationally efficient

Alternative to de novo Prediction

- Use a set of reference orthologues and to identify their co-orthologues in newly sequenced species.
- Several dedicated databases offer orthologous sequences suitable for this cause, some spanning all domains of life
  - OrthoDB
  - OMA
  - Plaza
  - OrthoMam

- Computationally cheaper
- May alleviate errors associated with incomplete gene sampling.

Reading Materials

- Chapter 7 of the book "Understanding Bioinformatics", by M. Zvelebil and J.O. Baum. Published by Garland Science, 2008.
- Paschalia Kapli, Ziheng Yang & Maximilian J. Telford. “Phylogenetic tree building in the genomic age”. *Nature Reviews Genetics* (2020)