

## Phylogenetic Trees

## Phylogeny

- PHYLOGENY (coined 1866 Haeckel)
  1. the line of descent or evolutionary development of any plant or animal species
  2. the origin and evolution of a division, group or race of animals or plants

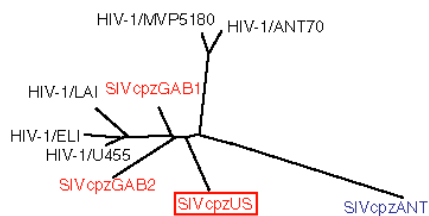
## Goals

- Understand evolutionary history
  - Origin of Europeans
- Assist in epidemiology
  - of infectious diseases
  - of genetic defects
- Aid in prediction of function of novel genes
- Biodiversity studies
- Understanding microbial ecologies

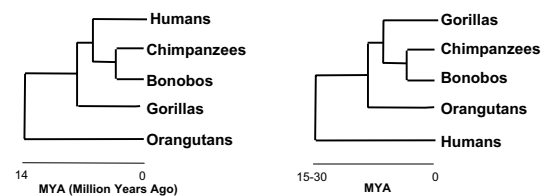
## Mitochondria and Phylogeny

- **Mitochondrial DNA (mtDNA):** Extra-nuclear DNA, transmitted through maternal lineage.
  - Allows tracing of a single genetic line
  - No recombinations
- 16.5 Kb circular DNA contains genes: coding for 13 proteins, 22 tRNA genes, 2 rRNA genes.
- mtDNA has a pointwise mutation substitution rate 10 times faster than nuclear DNA: provides a way to infer relationships between closely related individuals

## HIV-1 Origins



## Which species are the closest living relatives of modern humans?

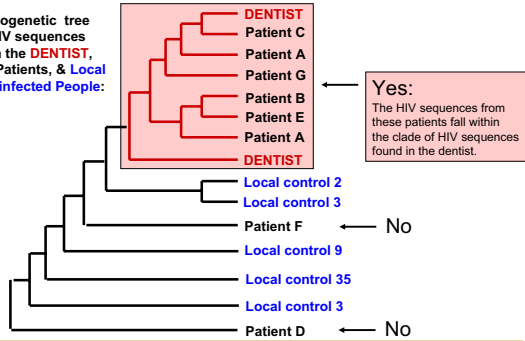


Mitochondrial DNA, most nuclear DNA-encoded genes, and DNA/DNA hybridization all show that bonobos and chimpanzees are related more closely to humans than either are to gorillas.

The pre-molecular view was that the great apes (chimpanzees, gorillas and orangutans) formed a clade separate from humans, and that humans diverged from the apes at least 15-30 MYA.

## Did the Florida Dentist infect his patients with HIV?

Phylogenetic tree of HIV sequences from the DENTIST, his Patients, & Local HIV-infected People:



From Ou et al. (1992) and Page & Holmes (1998)

## Gene Tree vs. Species Tree

- The evolutionary history of genes reflects that of species that carry them, except if :
  - horizontal transfer = gene transfer between species (e.g. bacteria, mitochondria)
  - Gene duplication : orthology/ paralogy

## Orthology / Paralogy

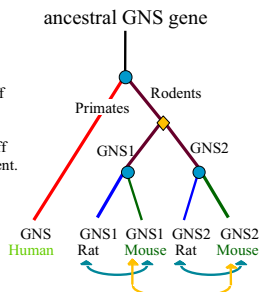
- speciation
- ◆ duplication

*Homology*: two genes are homologous iff they have a common ancestor.

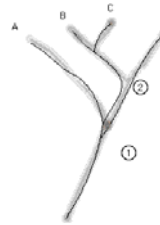
↔ *Orthology*: two genes are orthologous iff they diverged following a speciation event.

↔ *Paralogy*: two genes are paralogous iff they diverged following a duplication event.

⚠ Orthology → functional equivalence



## Effect of alleles



- Different alleles (variations of a gene) of a gene can exist within a species
- As a result of speciation, the alleles get distributed across the species
- Individuals within a species may be more closely related to individuals of another species (based on this locus alone)
- Gene tree differs from species tree

## Building Phylogenies: Phenotype Information has problems

- Can be difficult to observe
  - Bacteria
- Difficult to compare diverse species
  - Plants, bacteria, animals
- Convergent evolution

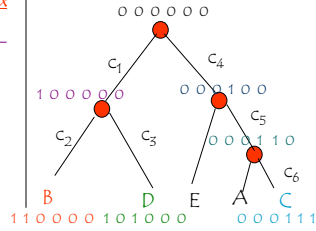
## Data for Building Phylogenies

- Characteristics
  - Traits (continuous or discrete)
  - Biomolecular features
  - character state matrix
- Numerical distance estimates
  - distance matrix

## Example of Character-based Phylogeny

A character state matrix

Taxon	c <sub>1</sub>	c <sub>2</sub>	c <sub>3</sub>	c <sub>4</sub>	c <sub>5</sub>	c <sub>6</sub>
A	0	0	0	1	1	0
B	1	1	0	0	0	0
C	0	0	0	1	1	1
D	1	0	1	0	0	0
E	0	0	0	1	0	0



## Different Kinds of Trees

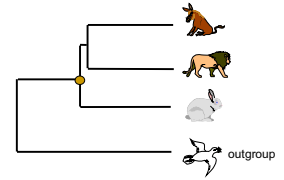
- Order of evolution
  - Rooted: indicates direction of evolution
  - Unrooted: only reflects the distance
- Rate of evolution
  - Edge lengths: distance (scaled trees)
    - Molecular clock: constant rate of evolution
  - Unscaled trees

## Rooted and Unrooted Trees

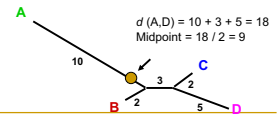
- Most phylogenetic methods produce **unrooted** trees. This is because they detect differences between sequences, but have no means to orient residue changes relatively to time.
- Two means to root an unrooted tree :
  - The **outgroup** method : include in the analysis a group of sequences known *a priori* to be external to the group under study; the root is by necessity on the branch joining the outgroup to other sequences.
  - Make the **molecular clock** hypothesis : all lineages are supposed to have evolved with the same speed since divergence from their common ancestor. Root the tree at the midway point between the two most distant taxa in the tree, as determined by branch lengths. The root is at the equidistant point from all tree leaves.

## Rooting unrooted trees

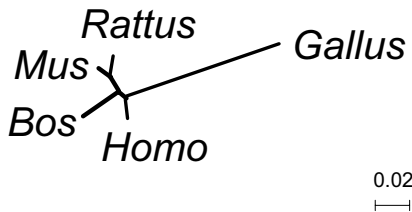
By outgroup:



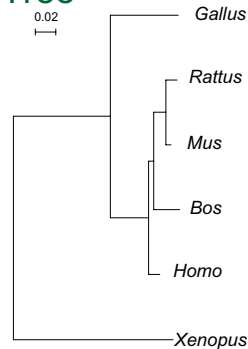
By midpoint or distance:

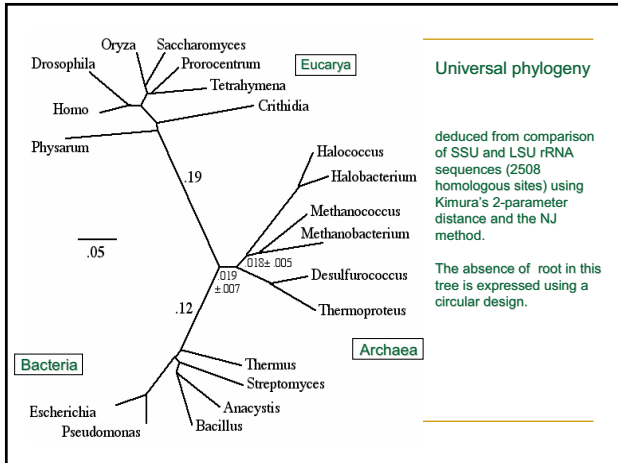


## Unrooted Tree



## Rooted Tree





## Tree building Methods

- Character-based methods
  - Maximum parsimony
  - Maximum likelihood
- Distance-based methods
  - UPGMA
  - NJ

## Distance Matrix Methods

- Given a pairwise distance matrix D
- Produce a tree such that the *path distance* between leaves i and j (sum of edge weights in the path between i and j) equals  $d_{ij}$
- Optimize the error between d and D
  - Least square error metric: LSQ
  - $LSQ(d,D) = \sum \sum (d_{ij} - D_{ij})^2$
  - NP-complete
- Heuristics (usually based on agglomerative (group by group) clustering)
  - UPGMA
  - NJ
  - Both assume additive distances
    - implies that distance is a metric
      - symmetry
      - triangle inequality
      - $d(x,y) = 0$  iff  $x = y$
      - $d(x,y) \geq 0$

## Distance Measures

- DNA sequences
  - Jukes and Cantor model
- Protein sequences
  - PAM matrix
- A site can undergo repeated changes
  - True vs. observed distances

## Example Tree and Additive Matrix

	a	b	c	d	e
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

There exists a tree with additive distances

## Additive Trees from Additive Matrices

- Verify that the distance matrix is additive
- Choose a pair of objects, which results in the first path in the tree.
- Choose a third object and establish the linear equations to let the object branch off the path.
- Choose a pair of leaves in the tree constructed so far and compute the point at which a newly chosen object is inserted.
  1. The new path branches off an existing node in the tree: Do the insertion step once more in the branching path.
  2. The new path branches off an edge in the tree: This insertion is finished.

### Example

	A	B	C	D	E
A	0	2	7	4	7
B		0	7	4	7
C			0	7	6
D				0	7
E					0

**NO!**

### Approximating Additive Matrices

In practice, the distance matrix between molecular sequences will not be additive.

An additive tree  $T$  whose distance matrix approximates the given one is used.

The methods for exact tree reconstruction provide an inventory for heuristics for tree construction based on approximating additive metrics.

Heuristics give exact results when operating on additive metrics.

### UPGMA

- Unweighted Pair-Group Method with Arithmetic Mean
  - Sokal and Michener 1958
- Agglomerative clustering
- Ultrametric tree
  - distances from root to all leaves are equal
- Cluster distances defined as

$$d_{AB} = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d_{ab}$$

### UPGMA Step 1

combine B and C

Choose two clusters with minimum distance and combine them

	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

### Updating distance matrices

	A	BC	D	E
A	0	11	8	7
BC		0	5	15
D			0	12
E				0

Distance of new cluster to nodes in the cluster is half of original distance

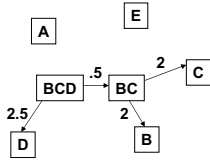
Distance of new cluster to other clusters is weighted mean of individual distances

### UPGMA step 2

combine BC and D

	A	BC	D	E
A	0	11	8	7
BC		0	5	15
D			0	12
E				0

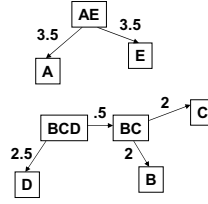
## Updating distance matrices



	A	BCD	E
A	0	10	7
BCD		0	14
E			0

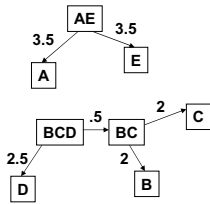
## UPGMA step 3

combine A and E



	A	BCD	E
A	0	10	7
BCD		0	14
E			0

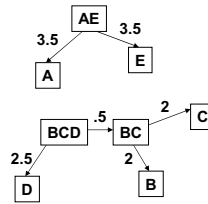
## Updating distance matrices



	AE	BCD
AE	0	12
BCD		0

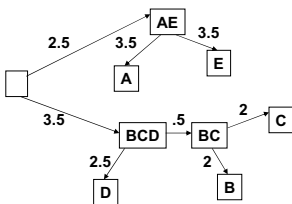
## UPGMA step 4

combine AE and BCD



	AE	BCD
AE	0	12
BCD		0

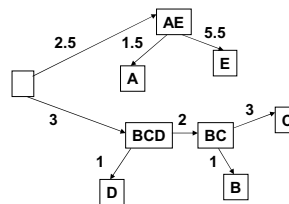
## UPGMA Result



	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

produced tree

## Actual tree



	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

actual tree