Bioinformatics

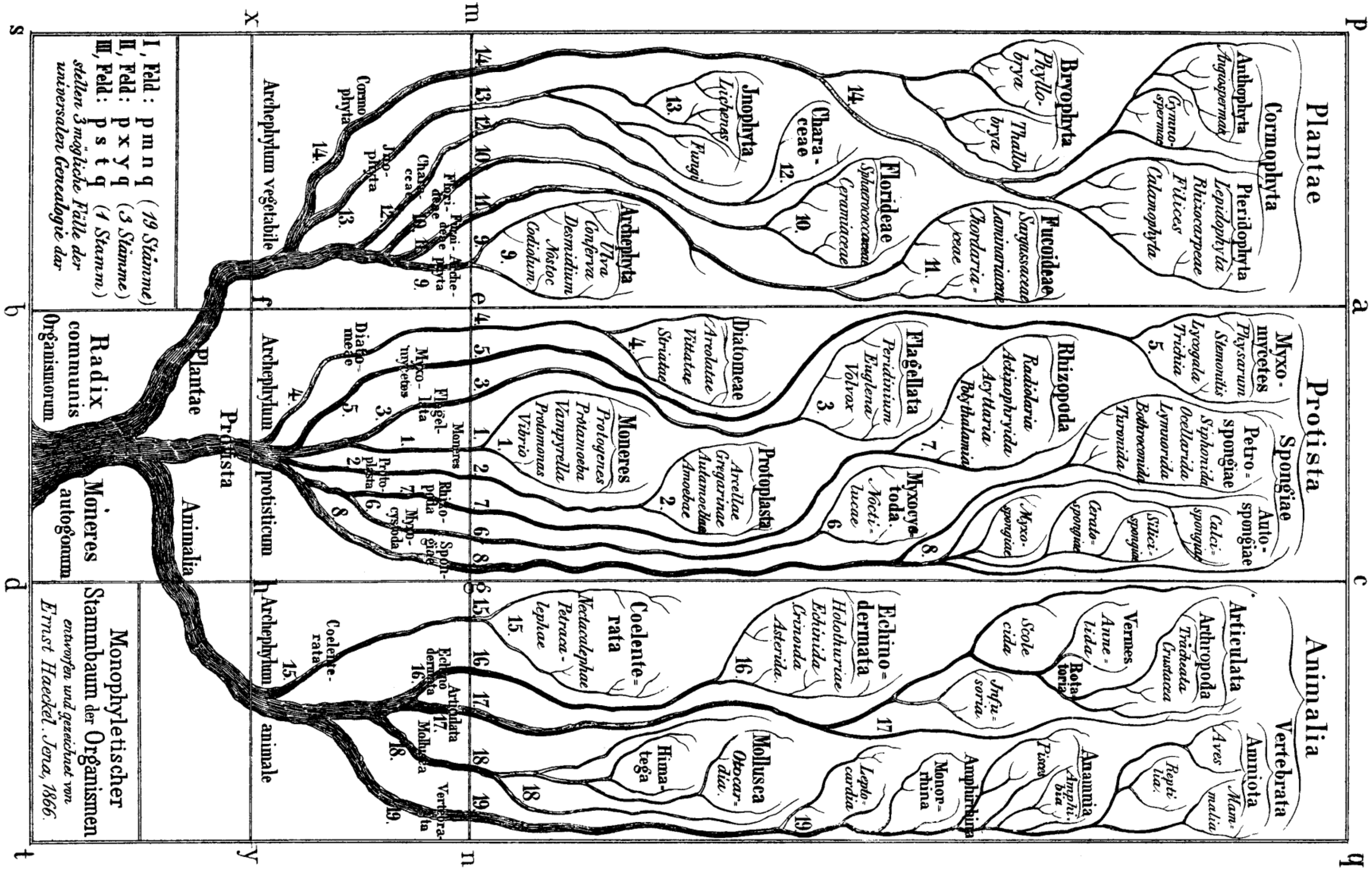
Ahmet Sacan

Phylogenetic analysis

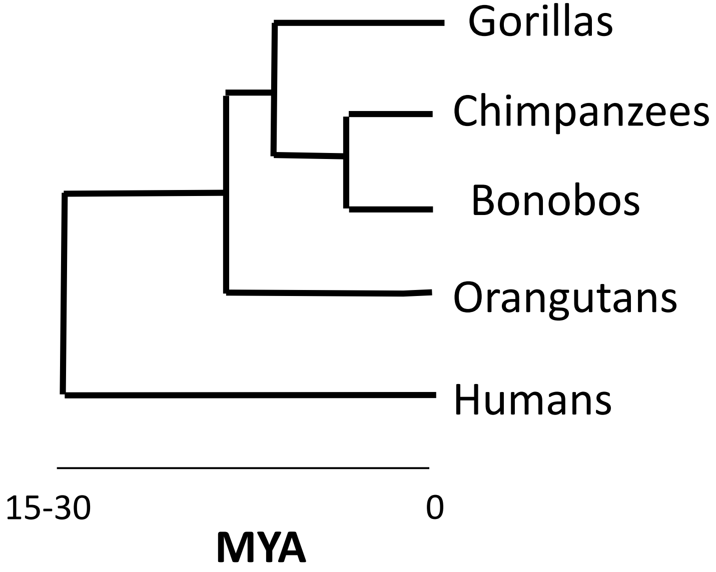
Phylogeny

- Definition:
  - the line of descent or evolutionary development of any plant or animal species
  - the origin and evolution of a division, group or race of animals or plants
- Goals:
  - Understand the evolutionary history
  - Assist in epidemiology
    - infectious diseases
    - genetic defects
  - Aid in functional prediction of genes
  - Understand microbial evolution
  - Understand adaptive immunity

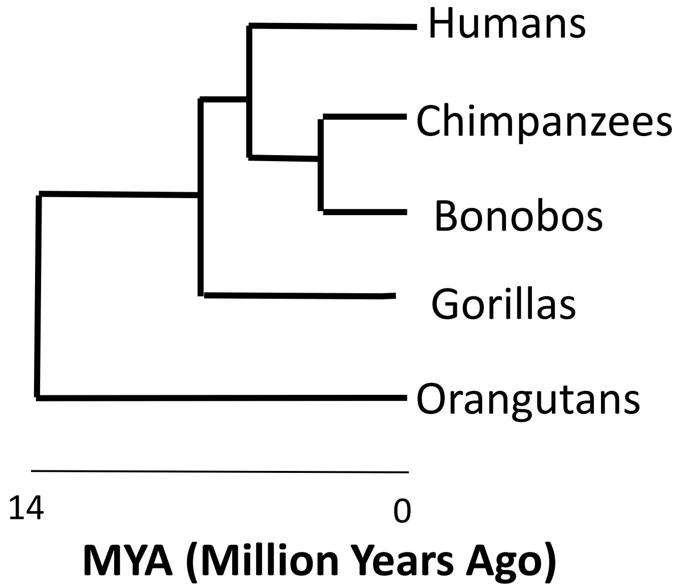
Ernst Haeckel, 1866



Closest living relatives of human?



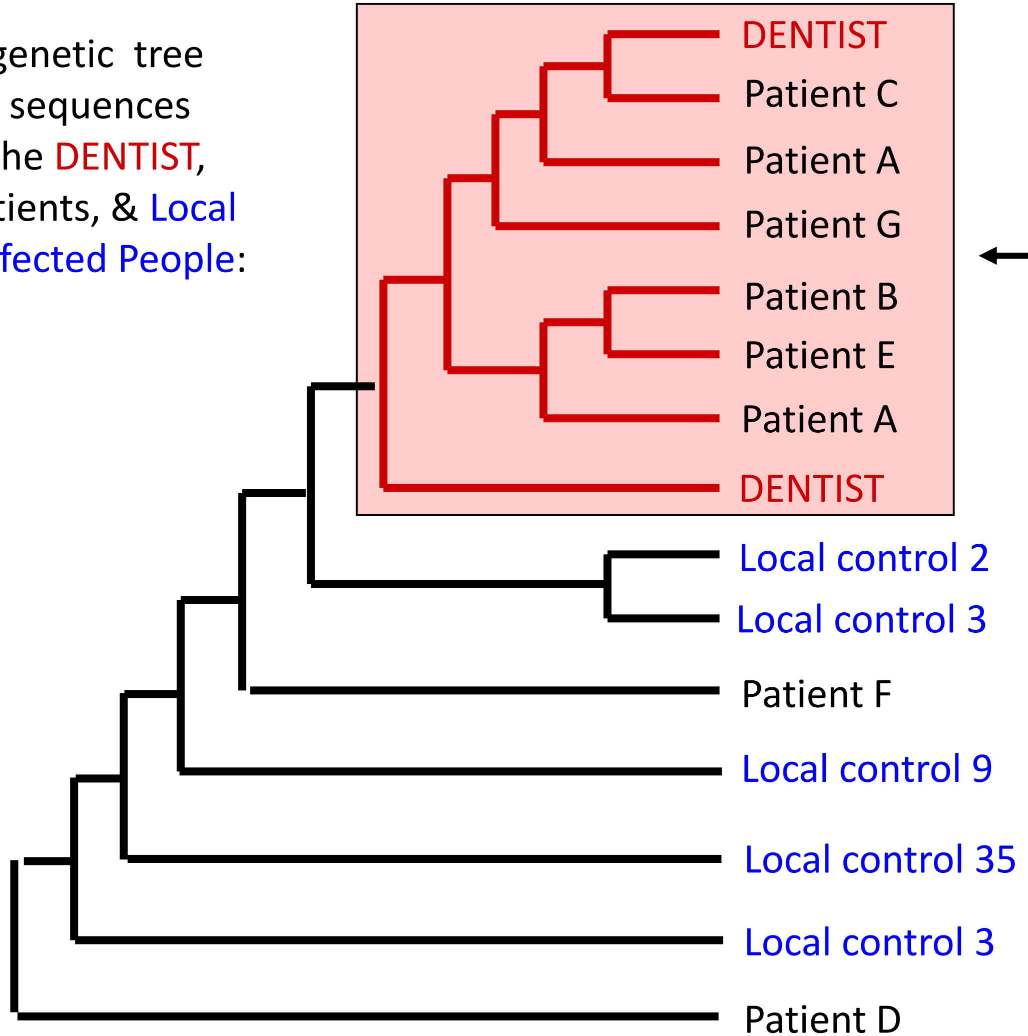
pre-molecular view



mitochondrial/nuclear DNA,
DNA-DNA hybridization

Forensic Analysis of HIV Sequence Data

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



Yes:

The HIV sequences from these patients fall within the clade of HIV sequences found in the dentist.

No

No

- Page & Holmes. Molecular Evolution: a phylogenetic approach.

Genetic anthropology

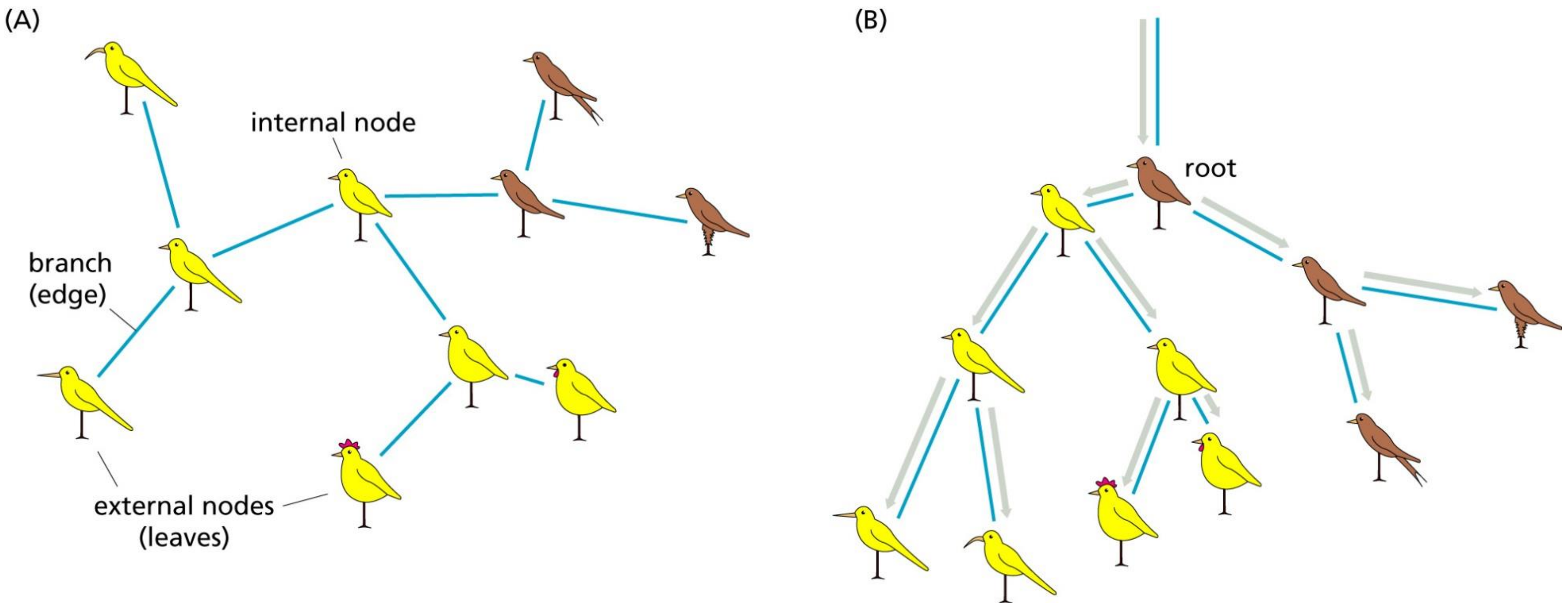
- Mithochondrial DNA (mtDNA)
  - Transmitted through maternal lineage
  - 16.5 Kb circular DNA
  - 13 proteins, 22 tRNA, 2 rRNA
  - mutation rate 10x faster than nuclear DNA: useful for comparing closely related individuals
- Y-chromosome
  - Paternal lineage

Projects

- Genographic Project
  - Largest DNA database for genetic anthropology
  - 5000 key populations, 100,000 people
- HapMap
  - haplotype map of the human genome
  - common patterns of genetic variation
- Journey of Mankind
  - Global journey of modern man over 160,000 years.

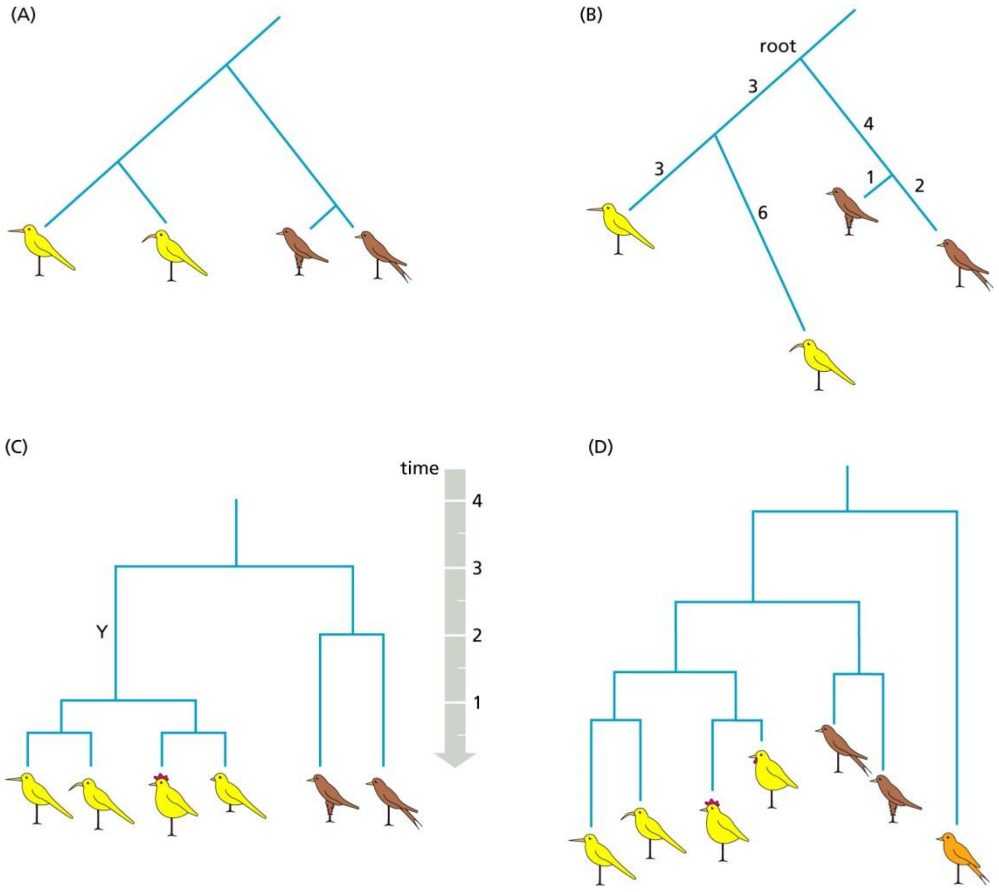
Rooted/Unrooted Trees

- Rooted: indicates direction of evolution
- Unrooted: only reflects the distance



Types of rooted trees

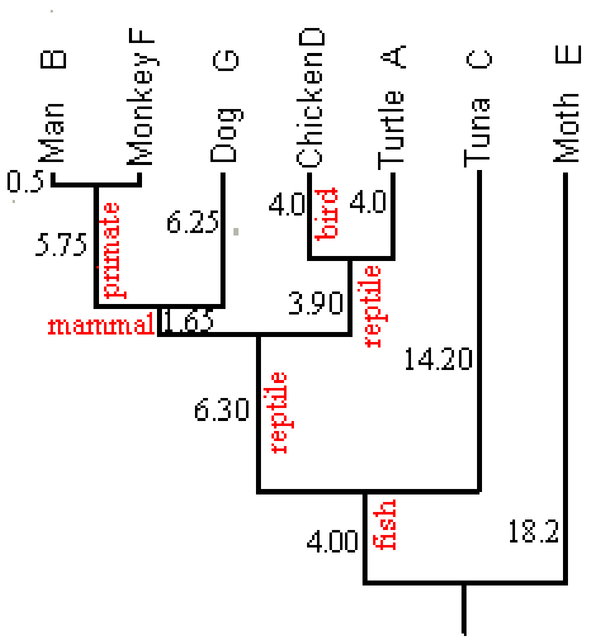
- Cladogram
  - branch lengths irrelevant
- Additive tree
  - branch lengths measure evolutionary divergence
  - but no information on time
- Ultrametric tree
  - molecular clock: constant mutation rate in all branches
- Example of rooting an unrooted tree using an outgroup.



Example: UPGMA clustering

- **UPGMA: Unweighted Pair Group Method with Arithmetic Mean**

		Turtle	Man	Tuna	Chicken	Moth	Monkey	Dog
		A	B	C	D	E	F	G
Turtle	A	0	19	27	8	33	18	13
Man	B	19	0	31	18	36	1	13
Tuna	C	27	31	0	26	41	32	29
Chicken	D	8	18	26	0	31	17	14
Moth	E	33	36	41	31	0	35	28
Monkey	F	18	1	32	17	35	0	12
Dog	G	13	13	29	14	28	12	0



- Fitch & Margoliash. Construction of Phylogenetic Trees. Science, 1967.

		Turtle	Man	Tuna	Chicken	Moth	Monkey	Dog
		A	B	C	D	E	F	G
Turtle	A	0	19	27	8	33	18	13
Man	B	19	0	31	18	36	1	13
Tuna	C	27	31	0	26	41	32	29
Chicken	D	8	18	26	0	31	17	14
Moth	E	33	36	41	31	0	35	28
Monkey	F	18	1	32	17	35	0	12
Dog	G	13	13	29	14	28	12	0

		Turtle	Man	Tuna	Chicken	Moth	Monkey	Dog
		A	B	C	D	E	F	G
Turtle	A	0	19	27	8	33	18	13
Man	B	19	0	31	18	36	1	13
Tuna	C	27	31	0	26	41	32	29
Chicken	D	8	18	26	0	31	17	14
Moth	E	33	36	41	31	0	35	28
Monkey	F	18	1	32	17	35	0	12
Dog	G	13	13	29	14	28	12	0

		Turtle	Man	Tuna	Chicken	Moth	Monkey	Dog
		A	B	C	D	E	F	G
Turtle	A	0	19	27	8	33	18	13
Man	B	19	0	31	18	36	1	13
Tuna	C	27	31	0	26	41	32	29
Chicken	D	8	18	26	0	31	17	14
Moth	E	33	36	41	31	0	35	28
Monkey	F	18	1	32	17	35	0	12
Dog	G	13	13	29	14	28	12	0

		Turtle	Man	Tuna	Chicken	Moth	Monkey	Dog
		A	B	C	D	E	F	G
Turtle	A	0	19	27	8	33	18	13
Man	B	19	0	31	18	36	1	13
Tuna	C	27	31	0	26	41	32	29
Chicken	D	8	18	26	0	31	17	14
Moth	E	33	36	41	31	0	35	28
Monkey	F	18	1	32	17	35	0	12
Dog	G	13	13	29	14	28	12	0

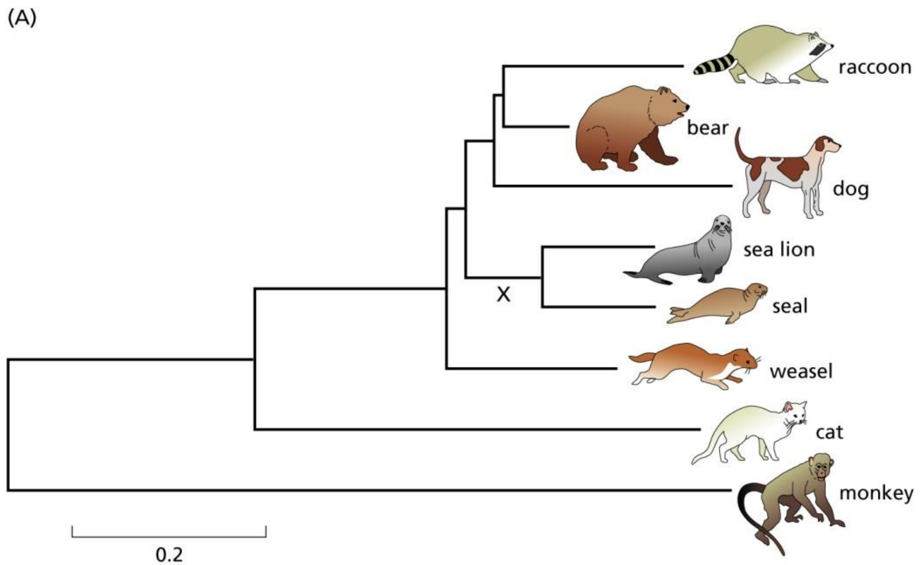
		Turtle	Man	Tuna	Chicken	Moth	Monkey	Dog
		A	B	C	D	E	F	G
Turtle	A	0	19	27	8	33	18	13
Man	B	19	0	31	18	36	1	13
Tuna	C	27	31	0	26	41	32	29
Chicken	D	8	18	26	0	31	17	14
Moth	E	33	36	41	31	0	35	28
Monkey	F	18	1	32	17	35	0	12
Dog	G	13	13	29	14	28	12	0

		Turtle	Man	Tuna	Chicken	Moth	Monkey	Dog
		A	B	C	D	E	F	G
Turtle	A	0	19	27	8	33	18	13
Man	B	19	0	31	18	36	1	13
Tuna	C	27	31	0	26	41	32	29
Chicken	D	8	18	26	0	31	17	14
Moth	E	33	36	41	31	0	35	28
Monkey	F	18	1	32	17	35	0	12
Dog	G	13	13	29	14	28	12	0

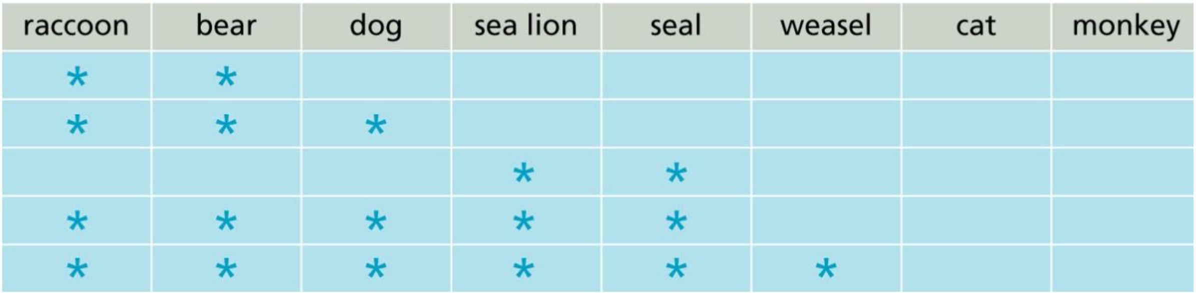
Dividing trees into splits/partitions

- Cutting each branch defines a split
- Newick/NewHampshire format:

```
{(
 {[
  ((raccoon,bear),dog),
  (sea lion,seal)
 ], weasel},
 cat),
monkey}
```



(B)

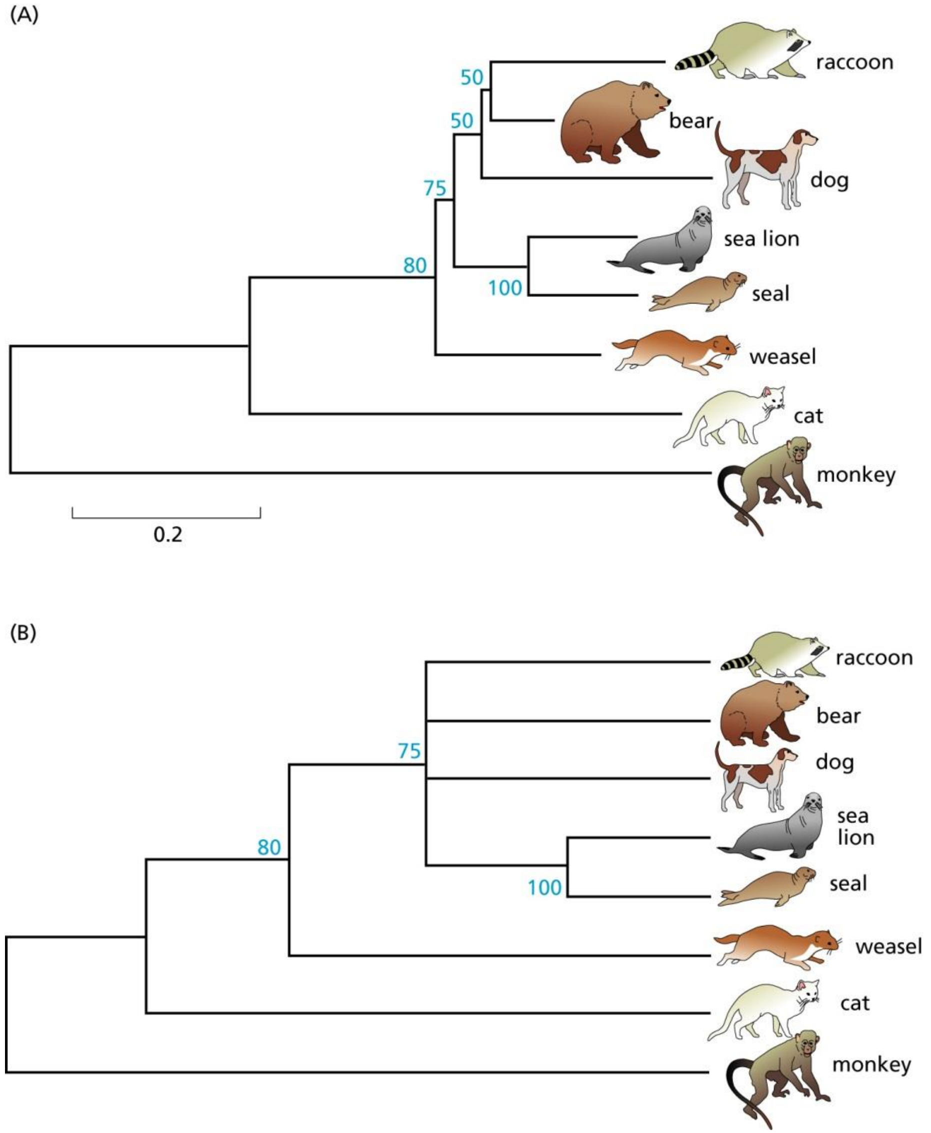


raccoon	bear	dog	sea lion	seal	weasel	cat	monkey
*	*						
*	*	*					
			*	*			
*	*	*	*	*			
*	*	*	*	*	*		

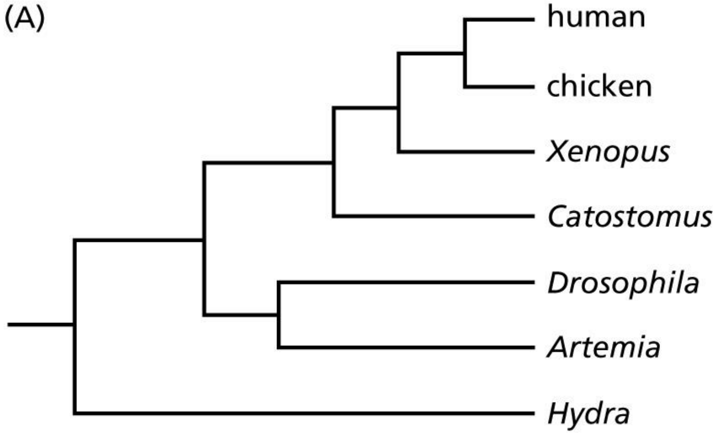
- (B:6.0,(A:5.0,C:3.0,E:4.0):5.0,D:11.0);

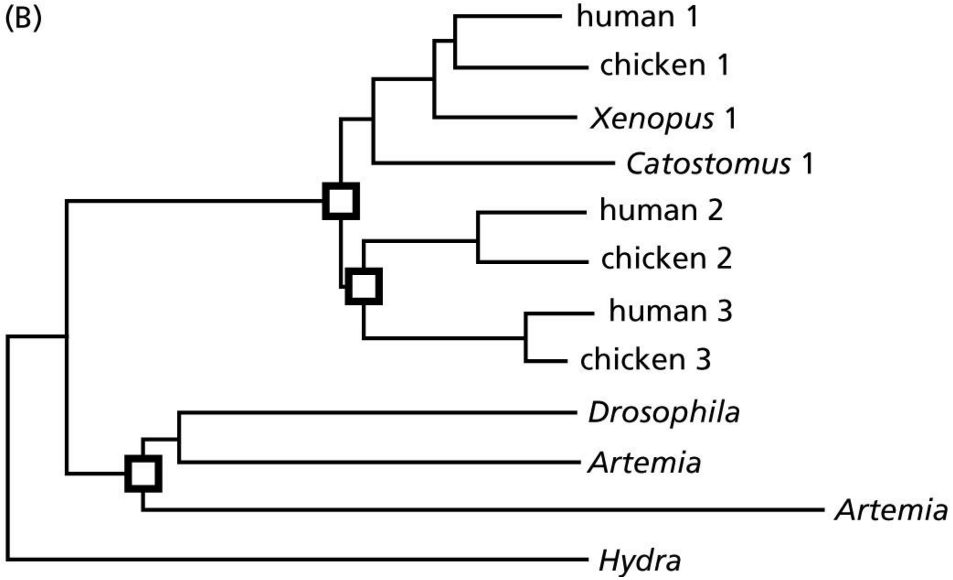
Bootstrap and condensed trees

- Bootstrap analysis estimates the support in the data for a given split
- A condensed tree can be obtained by removing all branches that are supported by less than e.g., 60% of bootstrap tests.



Species tree vs. Gene Tree





- Species tree of 7 eukaryotes
  - Xenopus: a frog
  - Catostomus: a fish
  - Drosophila: a fruit fly
  - Artemia: the brine shrimp

- Gene tree of Na+K+ ion pump proteins
- Small squares: gene duplications
- Homology: Orthologs, paralogs, ohnologs, xenologs
- NCBI: Homologene

Choice of data for evolutionary reconstruction

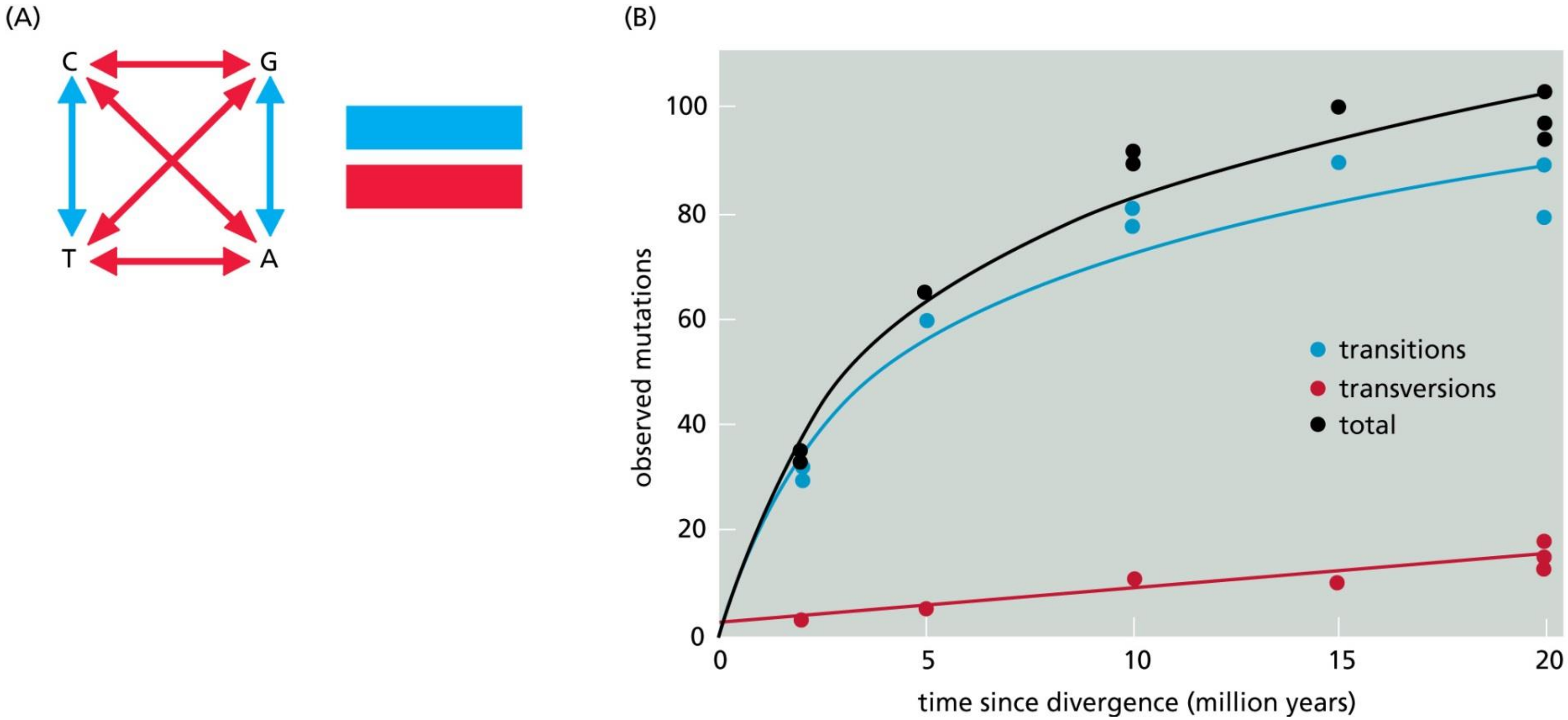
- Requirements for an ideal genomic region:
  - occurring in every species only once
  - doesn't include any horizontal gene transfer
  - rate of change fast enough to distinguish closely related species; but not too fast so remotely related species can still be accurately aligned. --> a single sequence with highly conserved and variable domains would do.
- Small ribosomal subunit rRNAs has been found to be well suited for evolutionary reconstruction of species

Evolutionary models estimate evolutionary distance

- p-distance
  - fraction of nonidentical alignment positions
  - underestimates the number of mutations that actually occurred.
- Causes of error:
  - Positions mutate several times.
  - Rate of substitution is not the same for all bases or at all locations
- Distance correction
  - Estimate the actual number of mutations that have occurred.

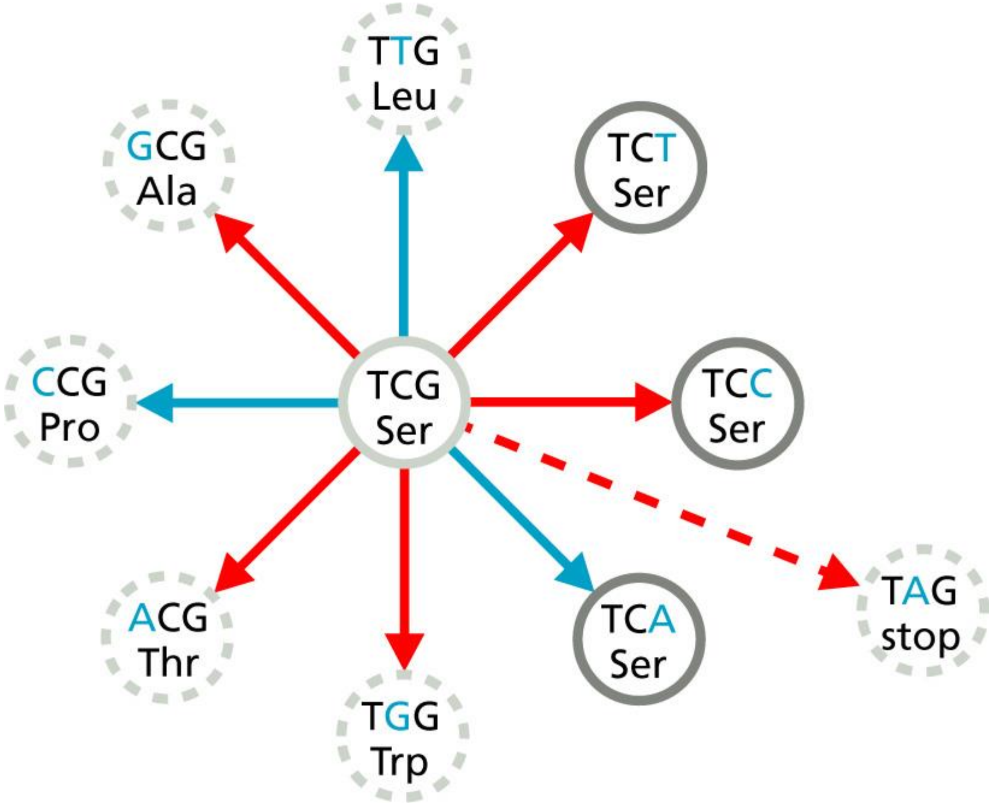
Transitions are more common than transversions

- Blue: transition mutations
- Red: transversion mutations



Coding sequence mutations have higher selective pressure

- Different codon positions have different mutation rates
- Synonymous/nonsynonymous ratio can help identify the type of selection (positive/negative)

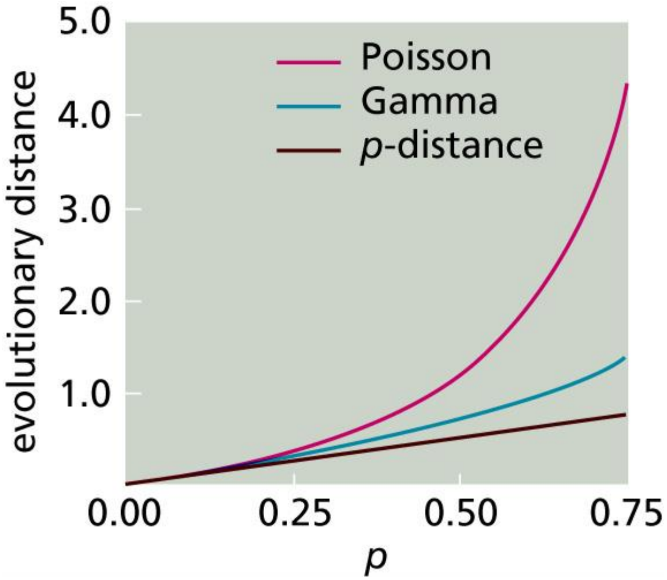


Poisson Distribution

- A discrete probability distribution that describes the probability of a given number of events occurring (in a fixed time/space).
- Examples:
  - Number of emails you receive per day.
  - Number of patients arriving in emergency room between midnight and 6am.
  - Number of goals in a soccer game.
- Assumptions:
  - Events are independent. (Occurrence of one event does not affect the probability of another)
  - λ : rate at which events occur is constant
- $P(k\ events\ in\ interval) = \frac{\lambda^k e^{-\lambda}}{k!}$
- $P(k = 0\ events\ in\ interval) = e^{-\lambda}$

Poisson model corrects for multiple mutations at the same site

- Assume probability of mutation at a site follows a Poisson distribution, with uniform mutation rate r per site per time point.
- After time t, average number of mutations at each site is rt. Number of mutations is then: Poisson($\lambda = rt$)
- Probability of n mutations having occurred at time t:
  - $e^{-rt}(rt)^n/n!$
- Probability of no mutations at a site:
  - e^{-rt}
- No mutations in either sequence:
  - e^{-2rt}
  - The sequences are an evolutionary distance $2rt = d$ from each other.
  - $e^{-d} = 1 - p$
  - $d = -\ln(1 - p)$



Other corrections

- Gamma correction takes into account differing mutation rates at different positions.
- Jukes-Cantor model handles mutations giving rise to identical sites; and is a widely accepted correction for nucleotide sequences.
- Other corrections involve:
  - distinguishing rates of transitions/transversions
  - unequal base compositions