Comparative Genomics
trees/phylogenies/genomes

Berend Snel

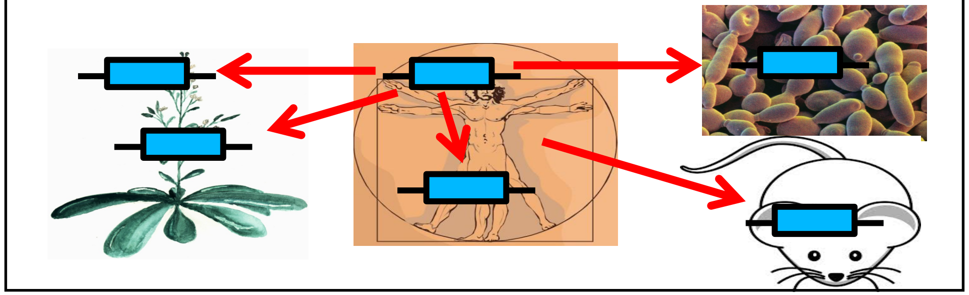
Theoretical Biology & Bioinformatics, Department of Biology, Faculty of Science

Utrecht Bioinformatics Center

Some of the slides in this lecture are courtesy of Jaap Heringa, Anders Gorm Pedersen , Can Kesmir , Bas Dutilh, Geert Kops and Michael Rosenberg

General idea of today

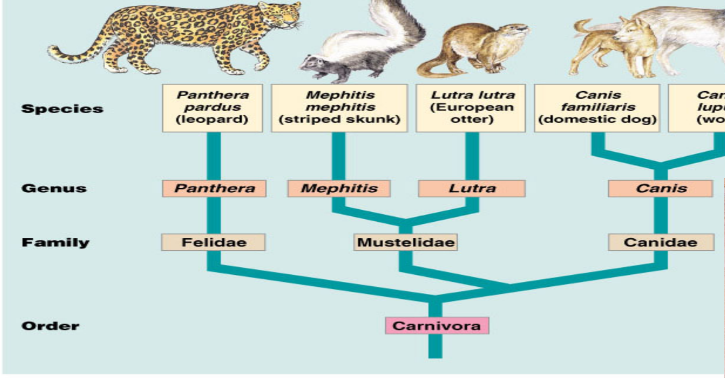
- Having all the genomes …
- What is the **relation** of my gene to *homologous* genes in the same and other organisms



Today

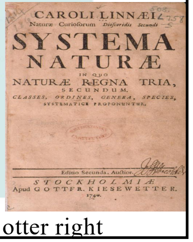
- What is a phylogenetic tree?
  - History
  - What can you do with a phylogenetic tree?
- How to “read” simple phylogenetic trees
- How can you make a phylogeny?
- How can you root a phylogeny?
- How to interpret a phylogenetic tree
  - Duplications
- (Genome Duplications -> evolutionary Genomics)

Tree: depiction (formalization) of hierarchical classification


Species
Panthera pardus (leopard)
Mephitis mephitis (striped skunk)
Lutra lutra (European otter)
Canis familiaris (domestic dog)
Canis lupus (wolf)
Genus
Panthera
Mephitis
Lutra
Canis
Family
Felidae
Mustelidae
Canidae
Order
Carnivora

Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Linnaeus



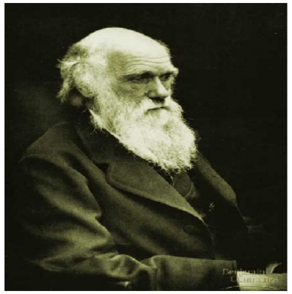
CAROLI LINNÆI
Naturæ Curioſorum *Dioſcoridis Secundi*
SYSTEMA
NATURÆ
IN QUO
NATURÆ REGNA TRIA,
SECUNDUM,
CLASSES, ORDINES, GENERA, SPECIES,
SYSTEMATICE PROPONUNTUR.

Editio Secunda, Auctior.

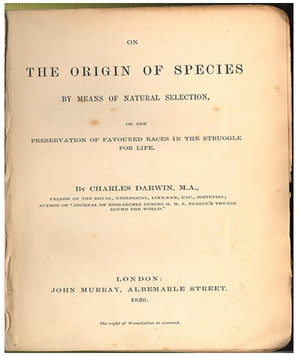
STOCKHOLMIÆ
Apud GOTTFR. KIESEWETTER.
1740.

NB no information in skunk left / otter right

Theory of evolution



Charles Darwin
1809-1882



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

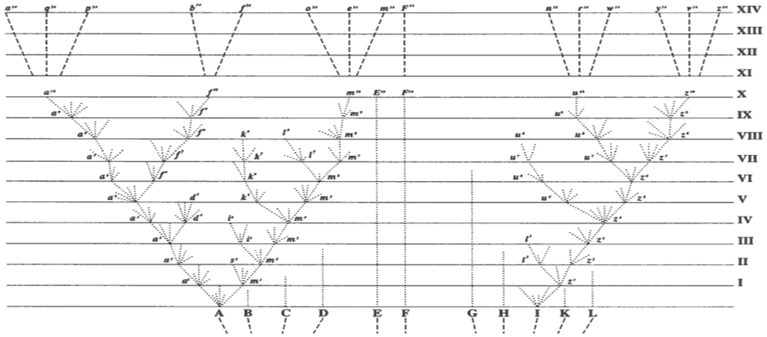
PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

BY CHARLES DARWIN, M.A.,
FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

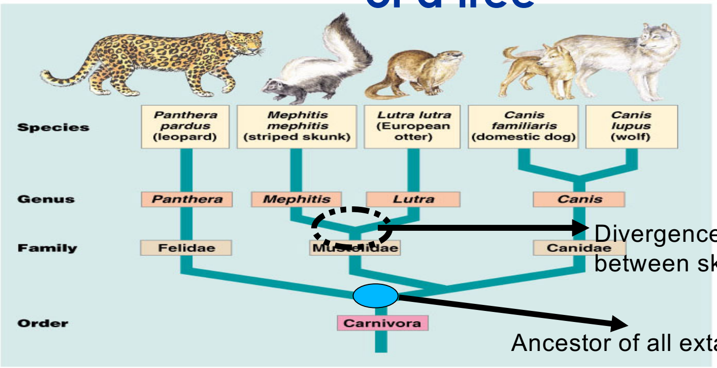
LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

The only figure in Darwin's "On the Origin of Species" is...



Phylogenetic tree: historical pattern of relationships among organisms: interpretation of a tree


Species
Genus
Family
Order
Panthera pardus (leopard)
Mephitis mephitis (striped skunk)
Lutra lutra (European otter)
Canis familiaris (domestic dog)
Canis lupus (wolf)
Panthera
Mephitis
Lutra
Canis
Felidae
Mustelidae
Canidae
Carnivora

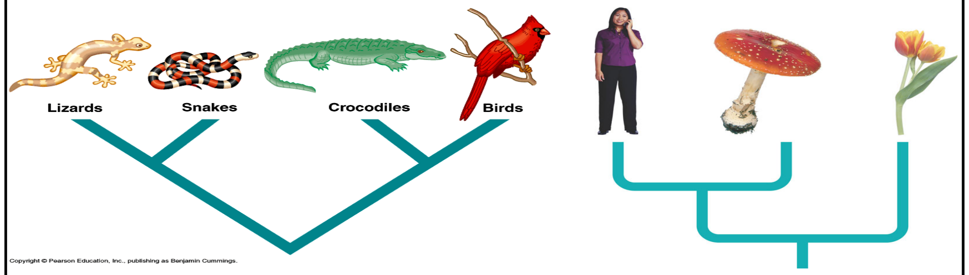
Divergence (= speciation) between skunk and otter

Ancestor of all extant carnivores

Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

NB still no information in skunk left / otter right

(molecular) Phylogenetic insights changed taxonomy



Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

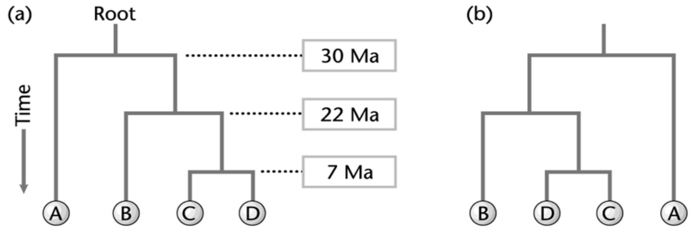
cladogram

?

Today

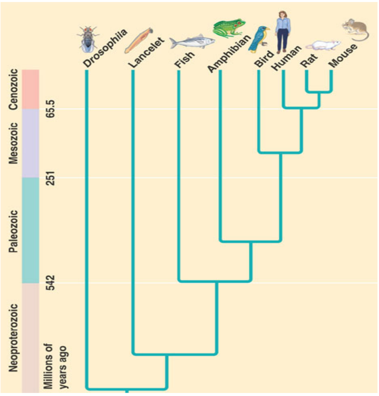
- What is a phylogenetic tree?
- How to "read" simple phylogenetic trees
  - Types of trees
  - Unrooted vs rooted
  - Molecular clock vs no molecular clock
- How to make a phylogeny
- How to root a phylogeny
- Afternoon Lecture

Rooted trees with molecular clock

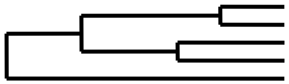


?

Figure 25.13 Campbell & Reece

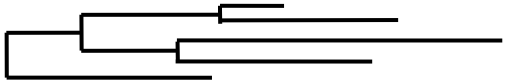


Evolutionary clock speeds



Uniform clock: leads to identical distances from root to leaves

(ultrametric tree)

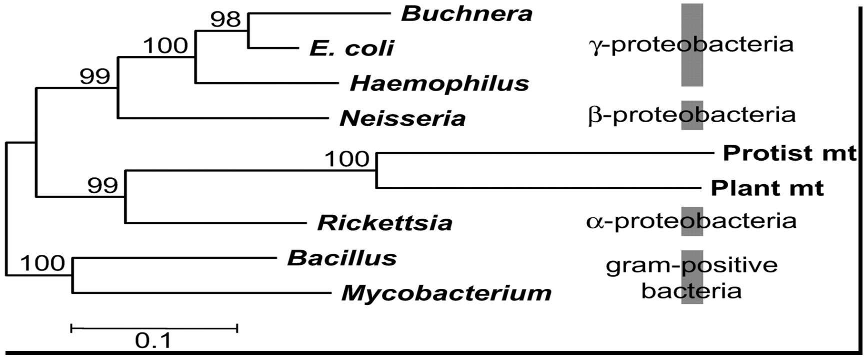


Non-uniform evolutionary clock: leaves have different distances …

(additive tree)

Root unknown: unrooted trees

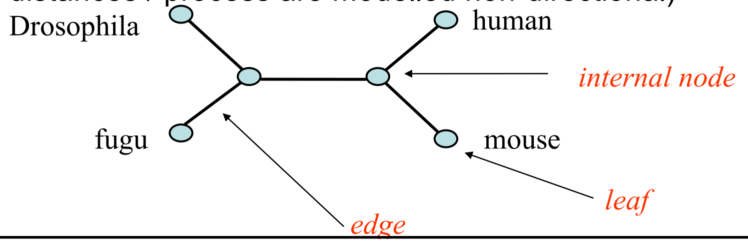
Unequal rates between species are a very real phenomenon



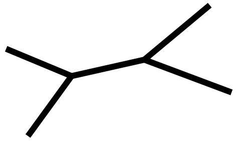
phylogram

No molecular clock means that a phylogenetic reconstruction method will infer only relations and no direction

- We “loose” flow of time
- Unambiguous representation: Unrooted tree
- NB most methods infer unrooted trees (because distances / process are modelled non-directional)

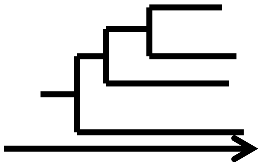


Radial tree (always explicitly unrooted)



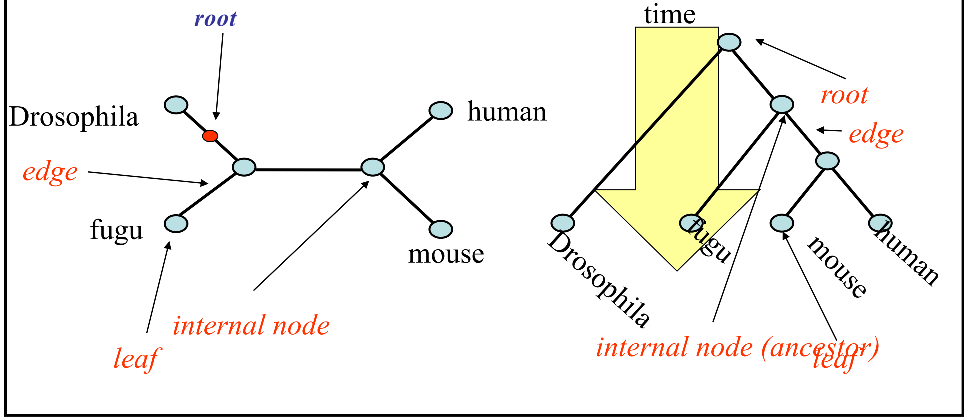
Phylogram (explicitly rooted)

Rooted = directed tree

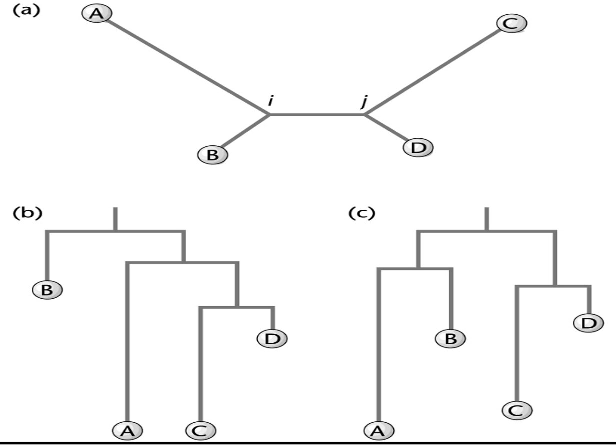


… both are phylogenies

Introduce a root to go from unrooted to rooted (or vice versa)



One unrooted tree can be turned into multiple rooted trees



?

Exercise 1

Today

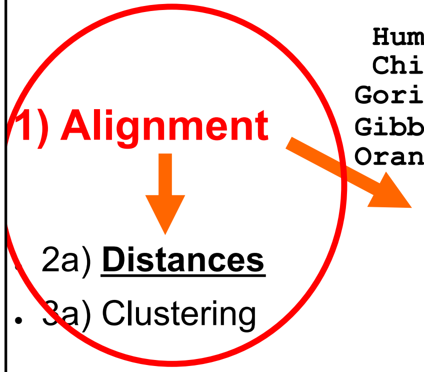
- What is a phylogenetic tree?
- How to “read” simple phylogenetic trees
- How to make a phylogeny
  - Distance methods
  - parsimony
- How to root a phylogeny
- Afternoon Lecture

Trees vs blast, phylogeny vs homology

- Blast/hmm/psi-blast tell you
  - How likely it is that two (parts) of a sequence are homologous or not (and how high the similarity between a profile and a sequence of between two sequences is)
  - Which portions of the sequences are significantly similar; which section of which sequence is homologous to which section of which other sequence.
  - Homologous is a yes/no thing
- Trees/phylogeny tell you
  - How the sequences are related, i.e. In which order they diverged

How to make a molecular phylogenetic tree

	1	2	3	4	5	6	7
Human	c	c	t	t	g	a	a
Chimp	c	c	t	t	g	a	a
Gorilla	c	c	t	a	g	t	a
Gibbon	t	c	a	a	g	a	a
Orangutan	t	c	a	a	g	a	t



- 2b) Explict model of sequence evolution plus best fitting tree, choice between **Parsimony** and **Maximum likelihood**

Phylogenetic tree by Distance methods (Clustering)

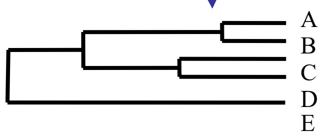
```
A  a t g a c c c c g
B  a t g a c c c c a
C  a t g a c g t c t
D  a t g a c g c g t
E  t t g t t c a a t
```

Multiple alignment

5×5 matrix

```
A 0
B 1 0
C 3 3 0
D 3 3 2 0
E 6 6 6 6 0
  A B C D E
```

Evolutionary Distance matrix



Phylogenetic tree

?

Clustering algorithm: UPGMA (assumes ultrametric trees)

Initialisation:

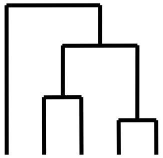
- Fill distance matrix with pairwise distances
- Start with N clusters of 1 element (gene) each

Iteration:

- Merge cluster C_i and C_j for which d_{ij} is minimal
- Place internal node connecting C_i and C_j at $d_{ij}/2$
- Delete C_i and C_j ; replace by new C with group average distances

Termination:

- When only two clusters i, j remain, put root at $d_{ij}/2$



```
A 0
B 1 0
C 5 4 0
D 7 8 2 0
E 9 9 9 9 0
  A B C D E
```

b

UPGMA

Iteration:

- Merge cluster C_i and C_j for which d_{ij} is minimal
- Place internal node connecting C_i and C_j at $d_{ij}/2$
- Delete C_i and C_j ; replace by new C with group average distances

Termination:

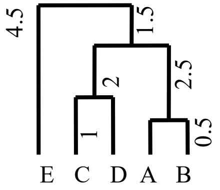
- When only two clusters i, j remain, put root at $d_{ij}/2$

```
A 0
B 1 0
C 5 4 0
D 7 8 2 0
E 9 9 9 9 0
  A B C D E
```

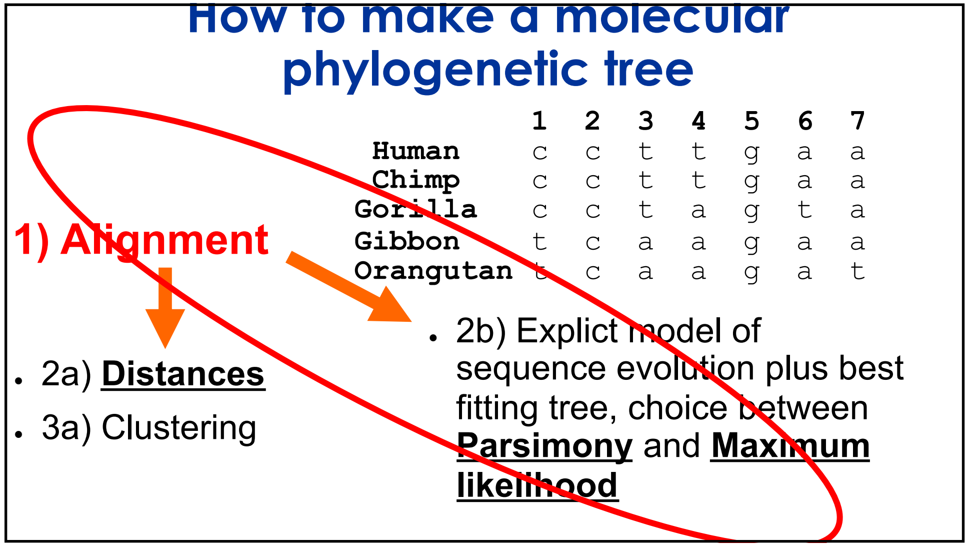
```
AB  0
 C  4.5  0
 D  7.5  2  0
 E  9    9  9  0
    AB   C  D  E
```

```
AB  0
CD  6  0
 E  9  9  0
   AB CD E
```

```
ABCD  0
   E  9   0
   ABCD  E
```



Termination



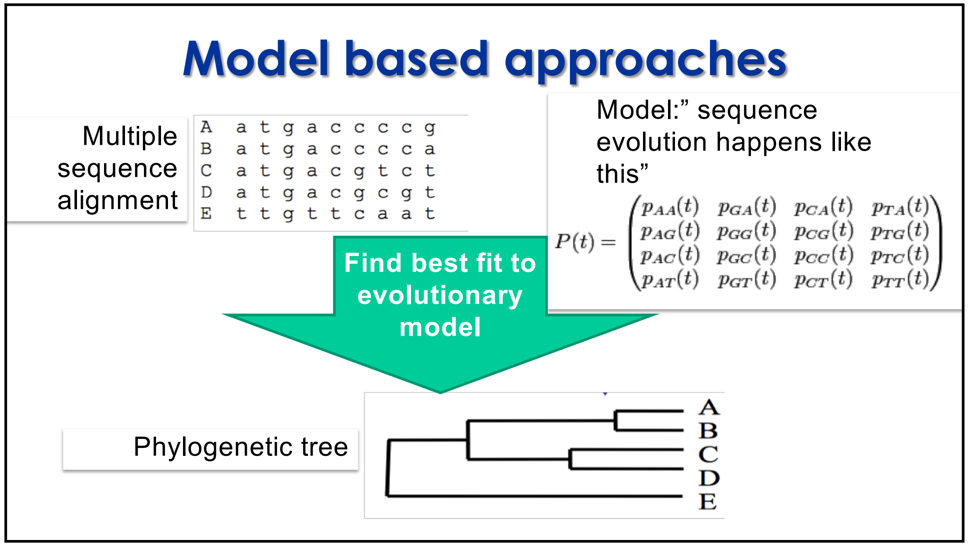
How to make a molecular phylogenetic tree

1) Alignment

	1	2	3	4	5	6	7
Human	c	c	t	t	g	a	a
Chimp	c	c	t	t	g	a	a
Gorilla	c	c	t	a	g	t	a
Gibbon	t	c	a	a	g	a	a
Orangutan	t	c	a	a	g	a	t

- 2a) **Distances**
- 3a) Clustering

- 2b) Explict model of sequence evolution plus best fitting tree, choice between **Parsimony** and **Maximum likelihood**



Model based approaches

Multiple sequence alignment

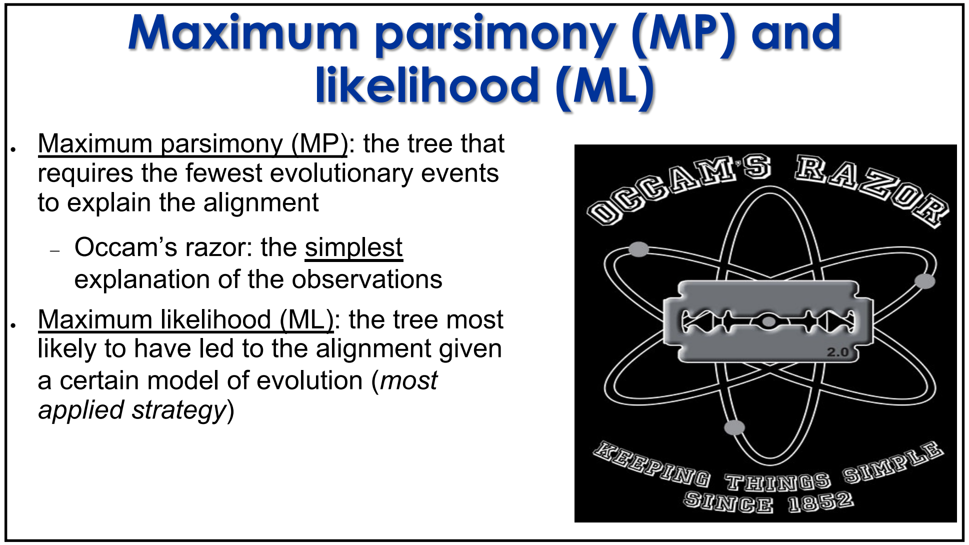
```
A  a t g a c c c c g
B  a t g a c c c c a
C  a t g a c g t c t
D  a t g a c g c g t
E  t t g t t c a a t
```

Model:” sequence evolution happens like this”

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{GA}(t) & p_{CA}(t) & p_{TA}(t) \\ p_{AG}(t) & p_{GG}(t) & p_{CG}(t) & p_{TG}(t) \\ p_{AC}(t) & p_{GC}(t) & p_{CC}(t) & p_{TC}(t) \\ p_{AT}(t) & p_{GT}(t) & p_{CT}(t) & p_{TT}(t) \end{pmatrix}$$

Find best fit to evolutionary model

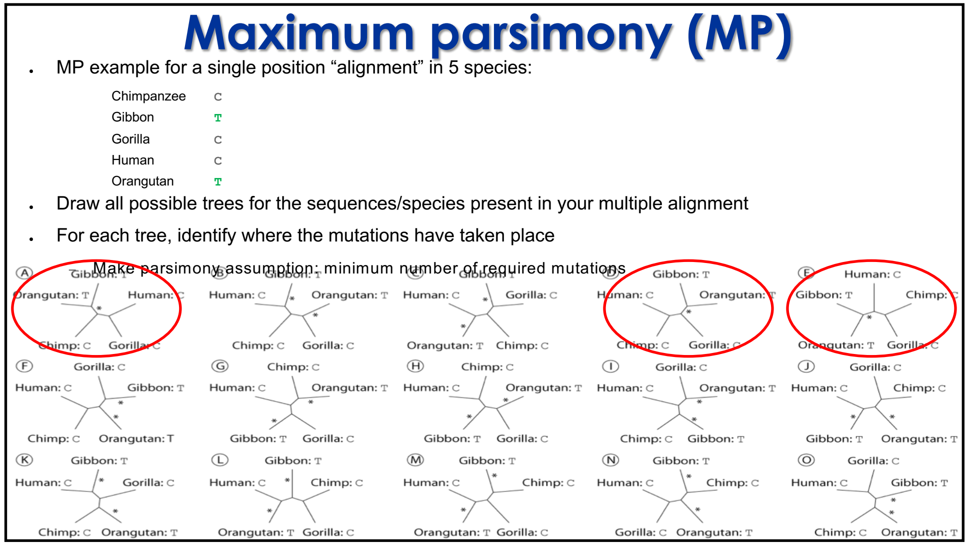
Phylogenetic tree



Maximum parsimony (MP) and likelihood (ML)

- Maximum parsimony (MP): the tree that requires the fewest evolutionary events to explain the alignment
  - Occam’s razor: the simplest explanation of the observations
- Maximum likelihood (ML): the tree most likely to have led to the alignment given a certain model of evolution (*most applied strategy*)

OCCAM'S RAZOR
2.0
KEEPING THINGS SIMPLE
SINCE 1852



Maximum parsimony (MP)

- MP example for a single position “alignment” in 5 species:

Chimpanzee	C
Gibbon	T
Gorilla	C
Human	C
Orangutan	T

- Draw all possible trees for the sequences/species present in your multiple alignment
- For each tree, identify where the mutations have taken place
  - Make parsimony assumption: minimum number of required mutations

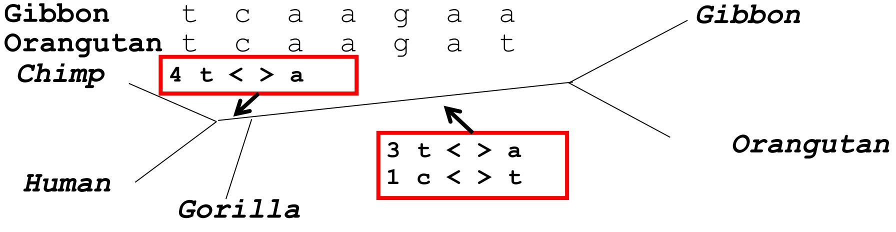
Maximum parsimony (MP)

- How many trees are there?
  - # unrooted trees $N_U = (2n - 5)!!$ $= (2n - 5) \times (2n - 7) \times \ldots \times 1$
  - # rooted trees $N_R = (2n - 3)!!$ $= (2n - 3) \times (2n - 5) \times \ldots \times 1$
  - E.g.
    - For **5** species 15 unrooted trees possible,
    - for **50** species, 2.84 *x* 10^{74} unrooted trees possible
  - (do you know how many species exist ?)
- For parsimony & maximum-likehood phylogeny: “heuristic searches”

Most parsimonious tree

The MP tree has the minimum number of required mutations

	1	2	3	4	5	6	7
Human	c	c	t	t	g	a	a
Chimp	c	c	t	t	g	a	a
Gorilla	c	c	t	a	g	t	a
Gibbon	t	c	a	a	g	a	a
Orangutan	t	c	a	a	g	a	t



NB unrooted tree! = Mutation modelled in two directions

Maximum likelihood

- If *data* = alignment, *hypothesis* = tree, and under a given *evolutionary model:*
- compute “likelihood” that the *hypothesis* (=tree), given a *model* (e.g. substitution matrix), results in the observed *data* (= multiple sequence alignment).
- maximum likelihood selects the *hypothesis* (tree) that maximises the observed *data*
- CPU intensive method
- Best approach to find the “true” tree

Parsimony, Maximum Likelihood or Neighbor-Joining?

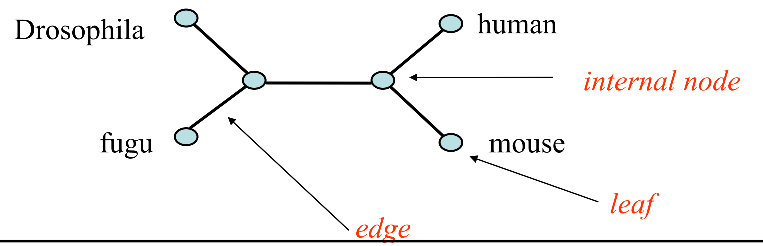
1. ML (PhyML, RaxML, IQ-tree) and bayesian methods (MrBayes and PhyloBase) are thought to be most accurate
2. Data is of greater importance than method
3. one must remember that a phylogenetic tree is a *hypothesis* of the true evolutionary history.
4. As a hypothesis it could be right or wrong or a bit of both.

Today

- What is a phylogenetic tree?
- How to "read" simple phylogenetic trees
- How to make a phylogeny
- How to root a phylogeny
- Afternoon Lecture

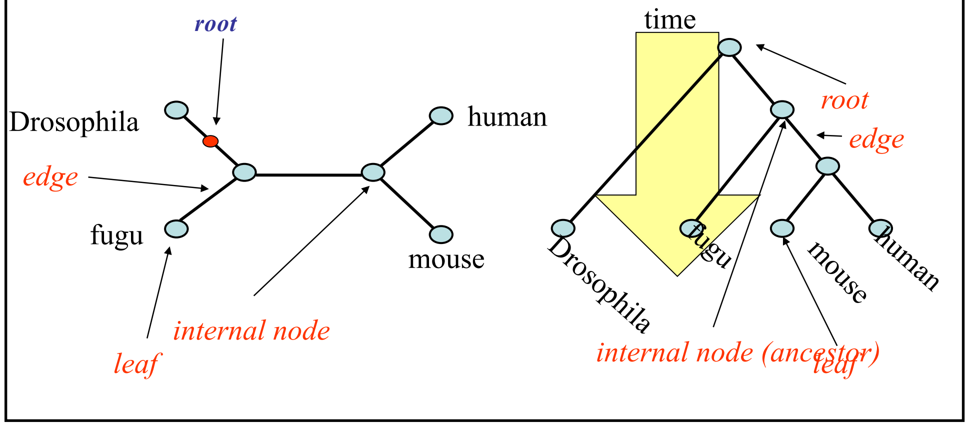
Unrooted trees

- Problem for interpretation in what order did my species diverge a tree without flow of time
- Representation: Unrooted tree
- NB most methods infer unrooted trees (because distances / process are not modeled directional)



b

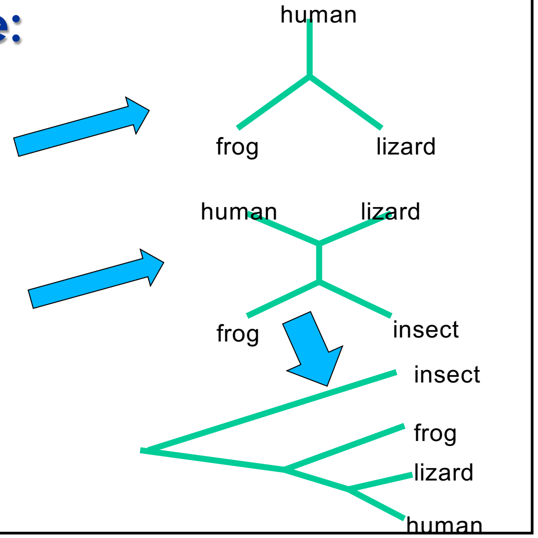
Introduce a root



How to root a tree: outgroup

```
 Human ccttgaa
  Frog ccttgat
Lizard ccttgac
```

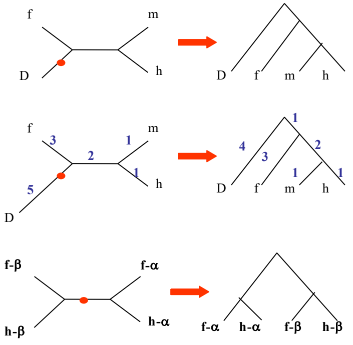
```
 Human ccttgaa
  Frog ccttgat
Lizard ccttgac
Insect aattgat
```



b

How to root a tree

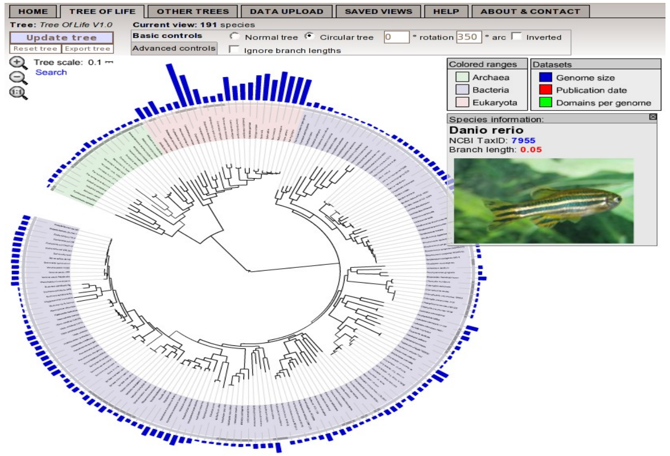
- Outgroup – place root between distant (still homolog) sequence and rest group
- Midpoint – place root at midpoint of longest path (sum of branches between any two leafs)
- Gene duplication – place root between paralogous gene copies



Implication of the fact that rooting is (like) a display choice

- If a tree is not rooted by the method (e.g. ML, NJ, MP), you are free to root it yourself … as long as you explain where you rooted it (and why)

iTOL



Tree: *Tree Of Life V1.0*

Current view: 191 species

○ Normal tree ● Circular tree 0 ° rotation 350 ° arc □ Inverted

□ Ignore branch lengths

Tree scale: 0.1

Colored ranges: Archaea, Bacteria, Eukaryota

Datasets: 🟦 Genome size, 🟥 Publication date, 🟩 Domains per genome

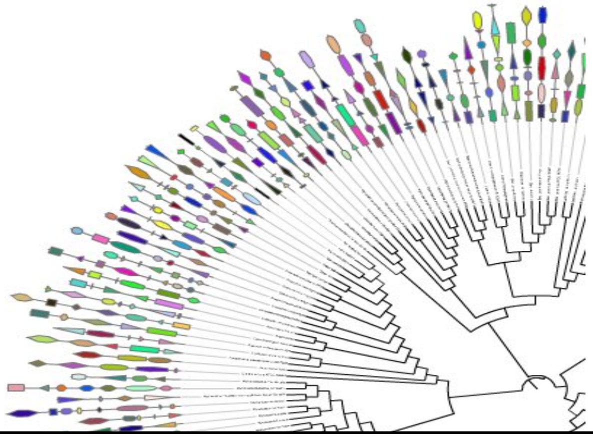
Species information:

Danio rerio

NCBI TaxID: 7955

Branch length: 0.05

iTOL



- This lecture was about reading trees and how they are constructed …
- After practicing with this we will have a lecture on recognizing gene duplications and gene loss in a gene tree (important for comparative genomic of eukaryotes)
- Thinking in trees also very relevant for recognizing horizontal gene transfer, host-pathogen co-evolution and many more things …

Further reading

- https://www.khanacademy.org/science/biology/her/tree-of-life/a/phylogenetic-trees

Exercises 2 & 3