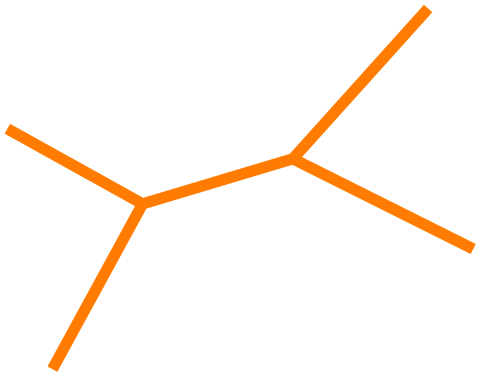
Comparative Genomics
trees/phylogenies/genomes

Berend Snel

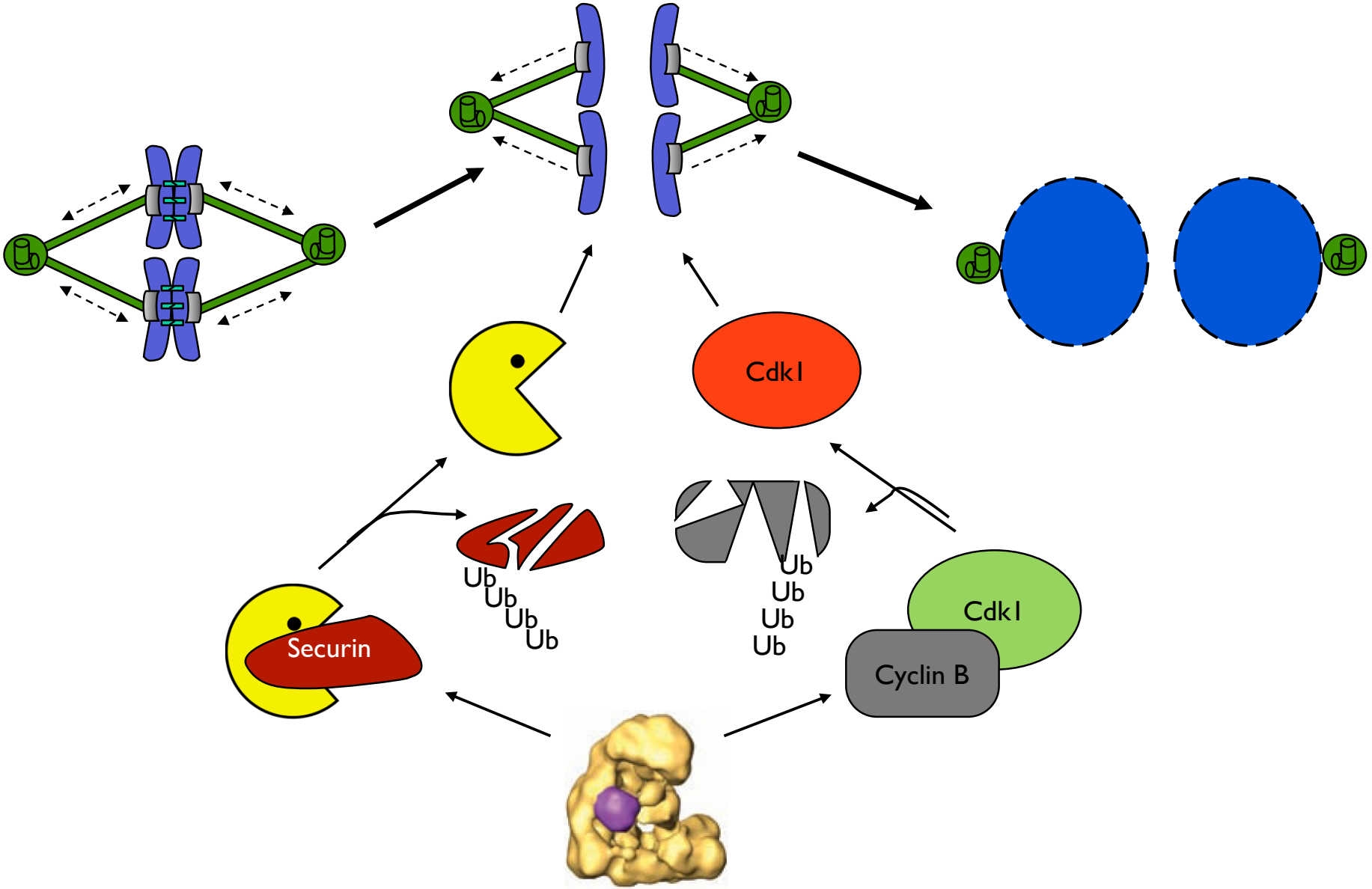
Theoretical Biology & Bioinformatics,
Department of Biology, Faculty of Science

Utrecht
Bioinformatics
Center



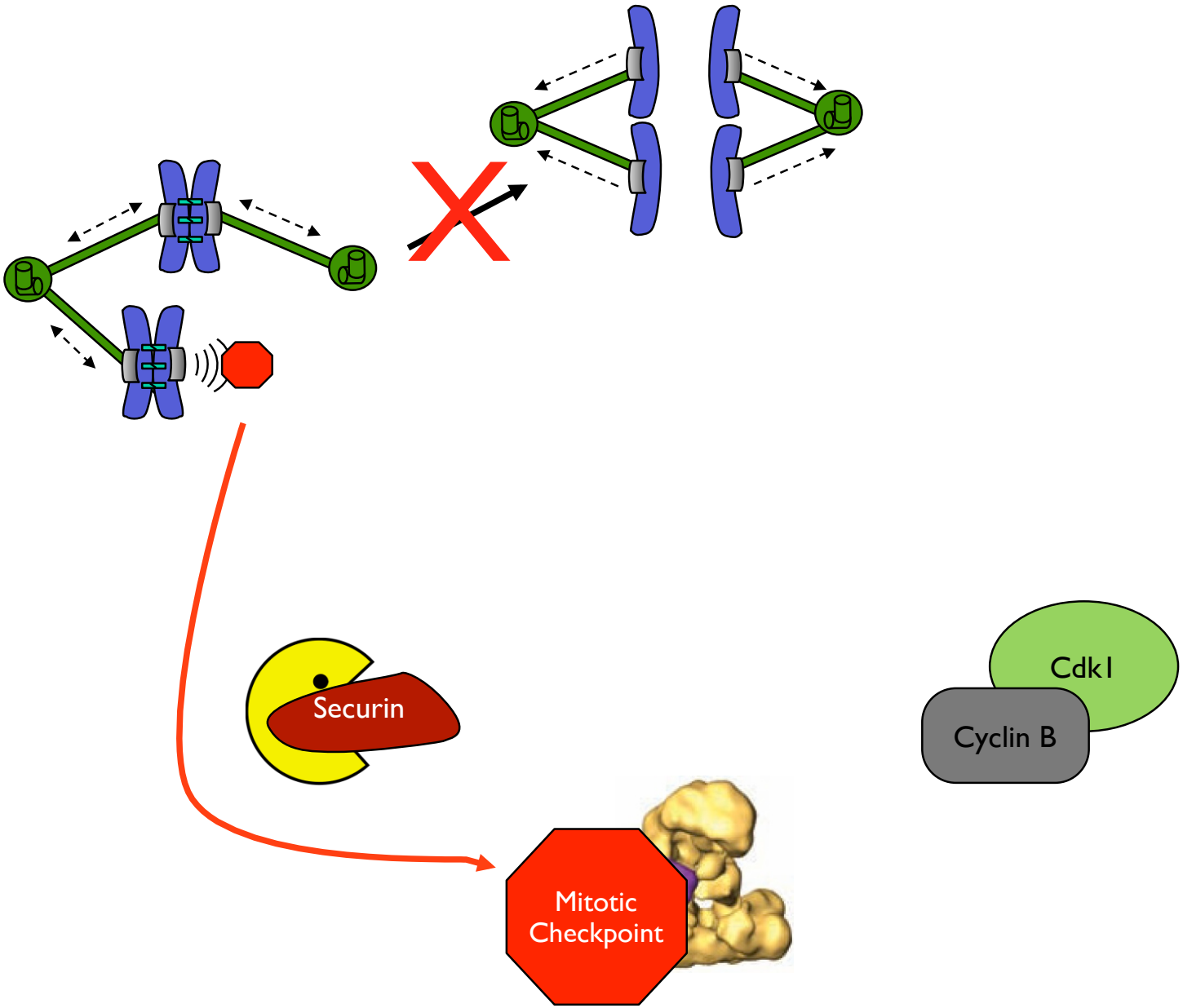
Some of the slides in this lecture are courtesy of Jaap Heringa, Anders Gorm Pedersen , Can Kesmir , Geert Kops and Michael Rosenberg

Case story Mitotic Checkpoint
Initiating Anaphase

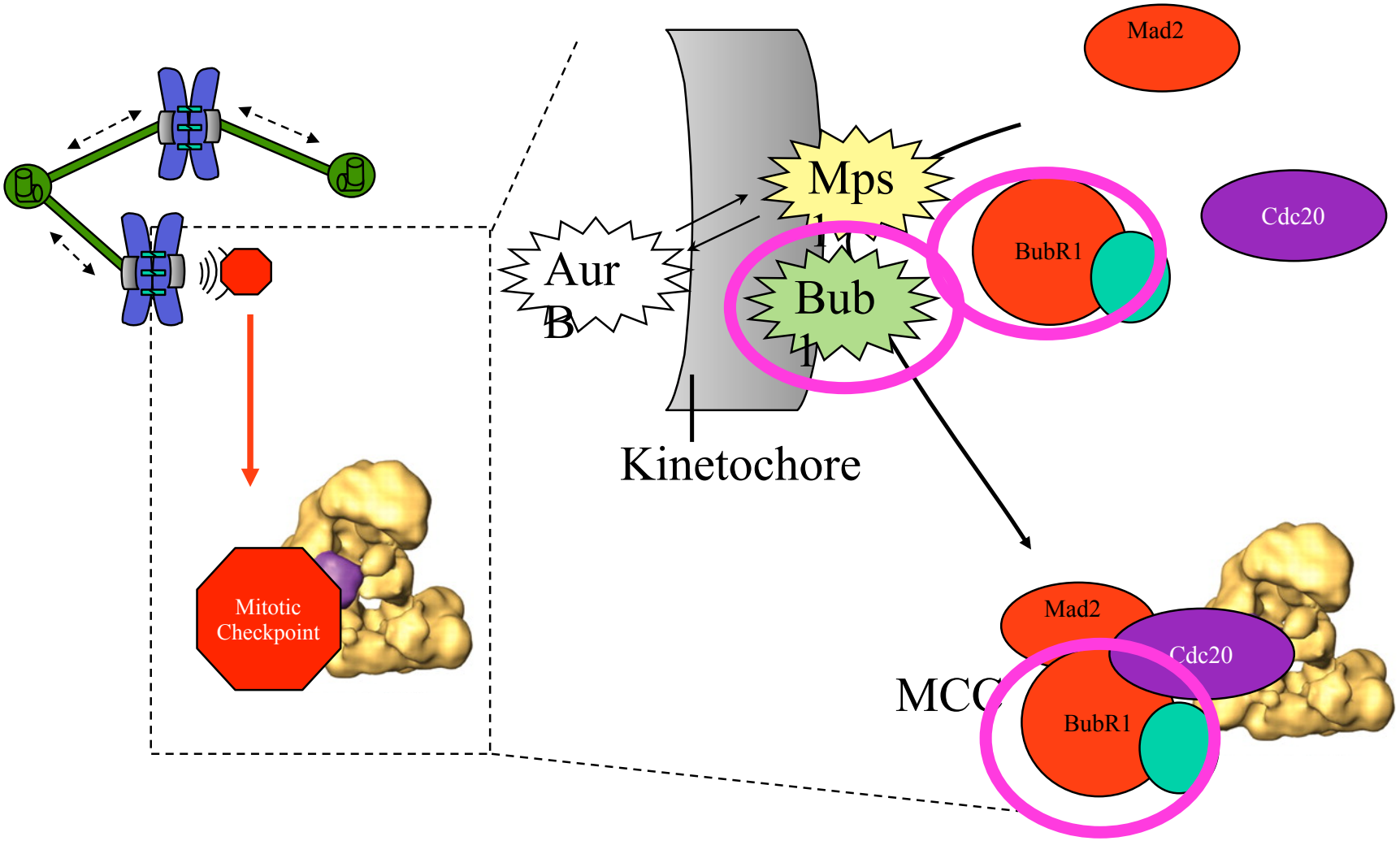


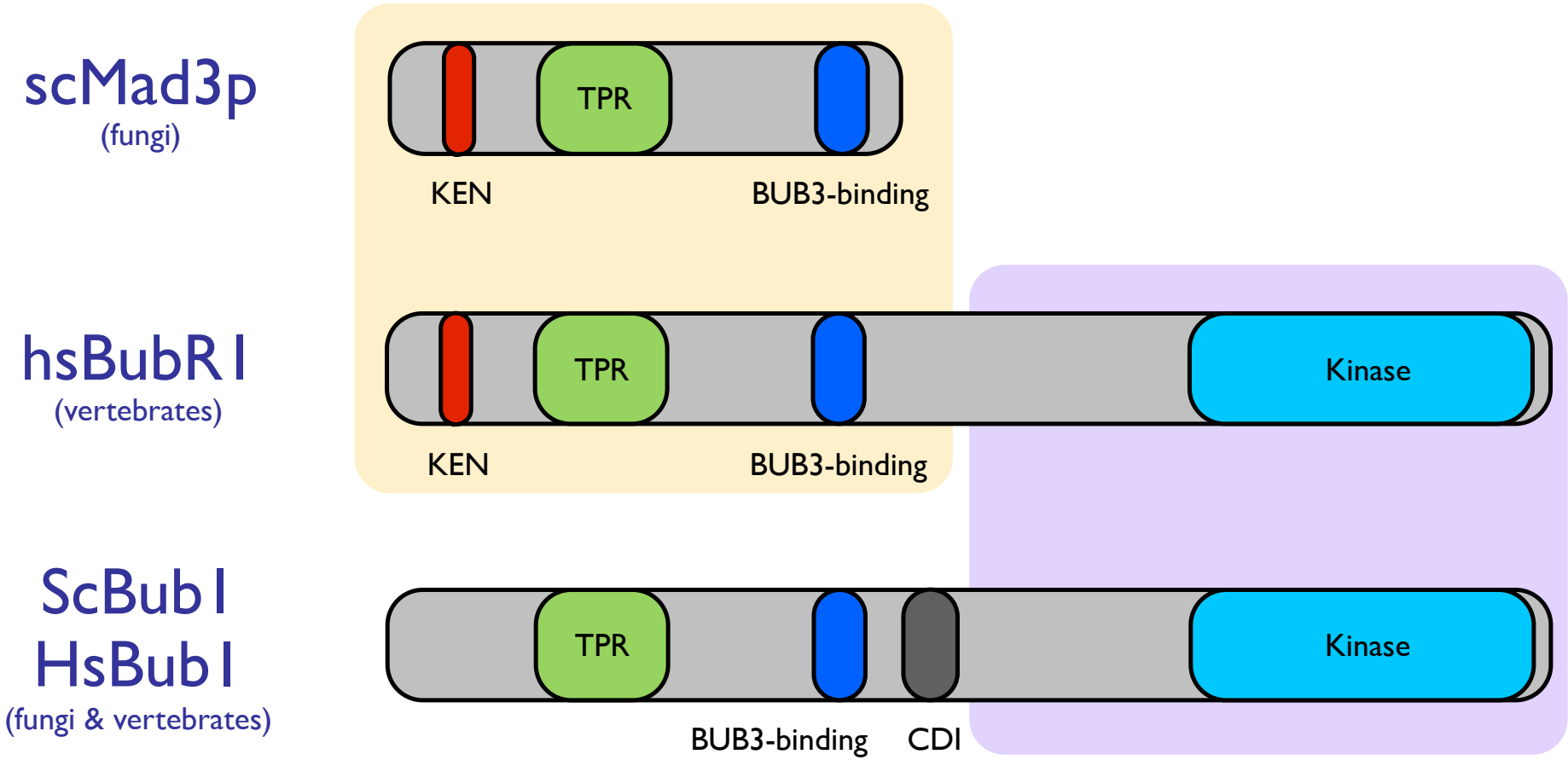
Anaphase Promoting Complex/Cyclosome (APC/C)

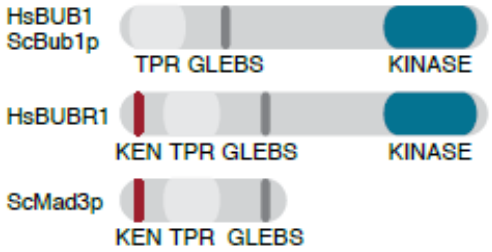
The Mitotic Checkpoint



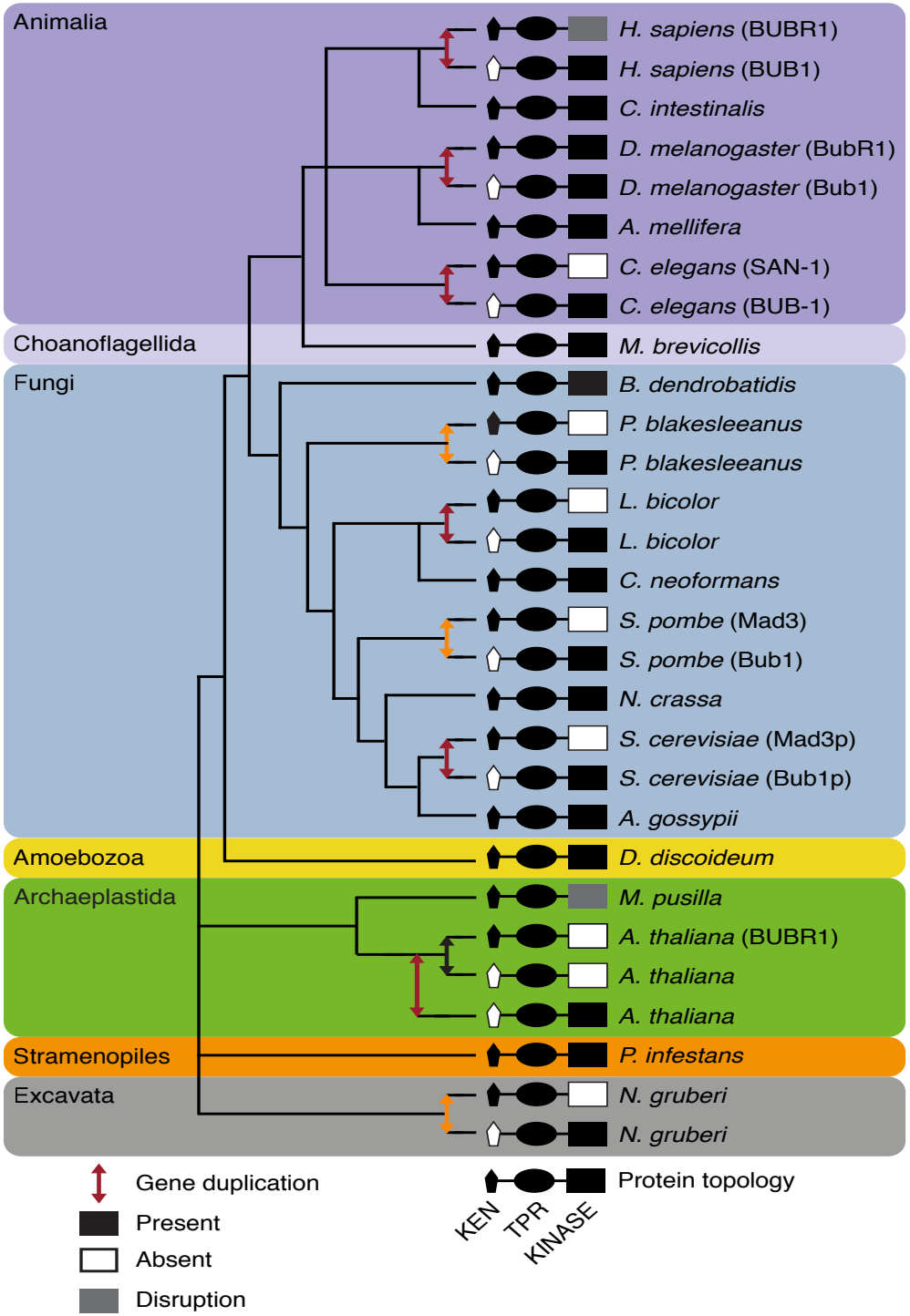
The Mitotic Checkpoint Complex (MCC)




scMad3p
(fungi)
TPR
KEN
BUB3-binding
hsBubR1
(vertebrates)
TPR
Kinase
KEN
BUB3-binding
ScBub1
HsBub1
(fungi & vertebrates)
TPR
Kinase
BUB3-binding
CDI

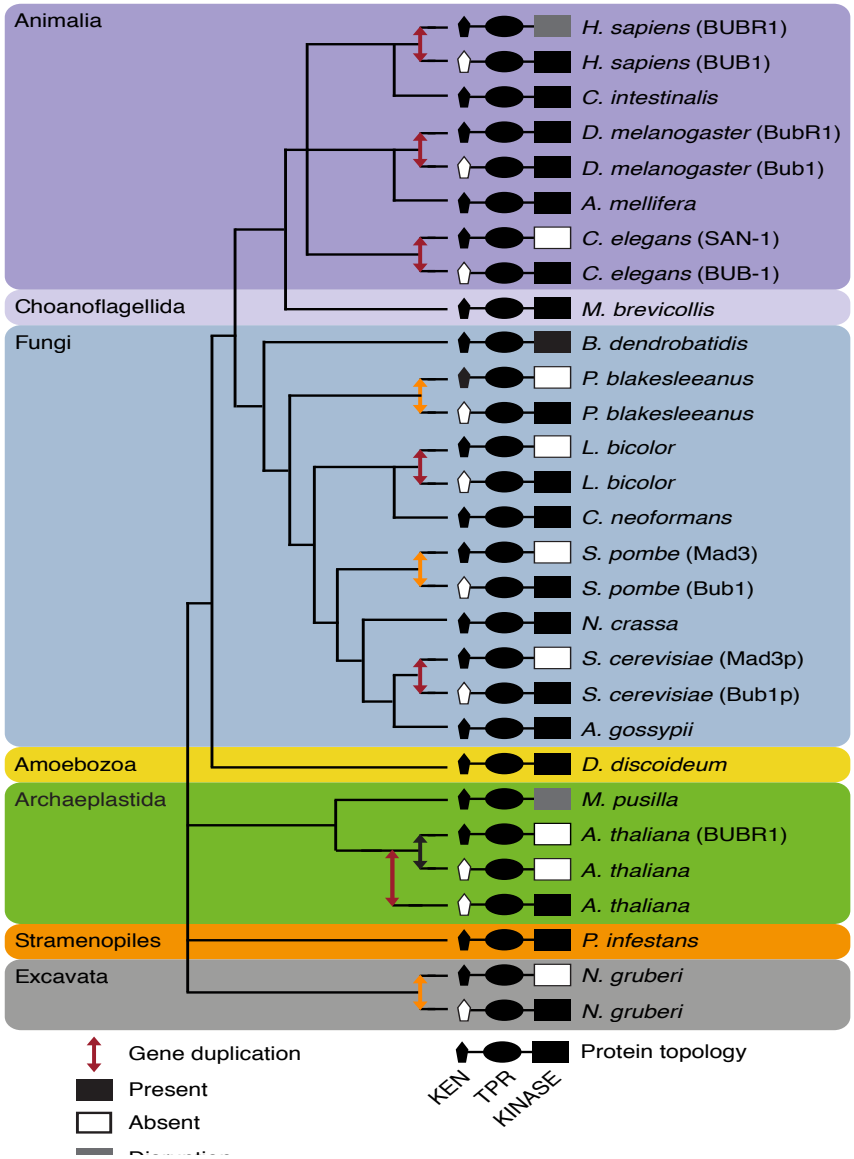


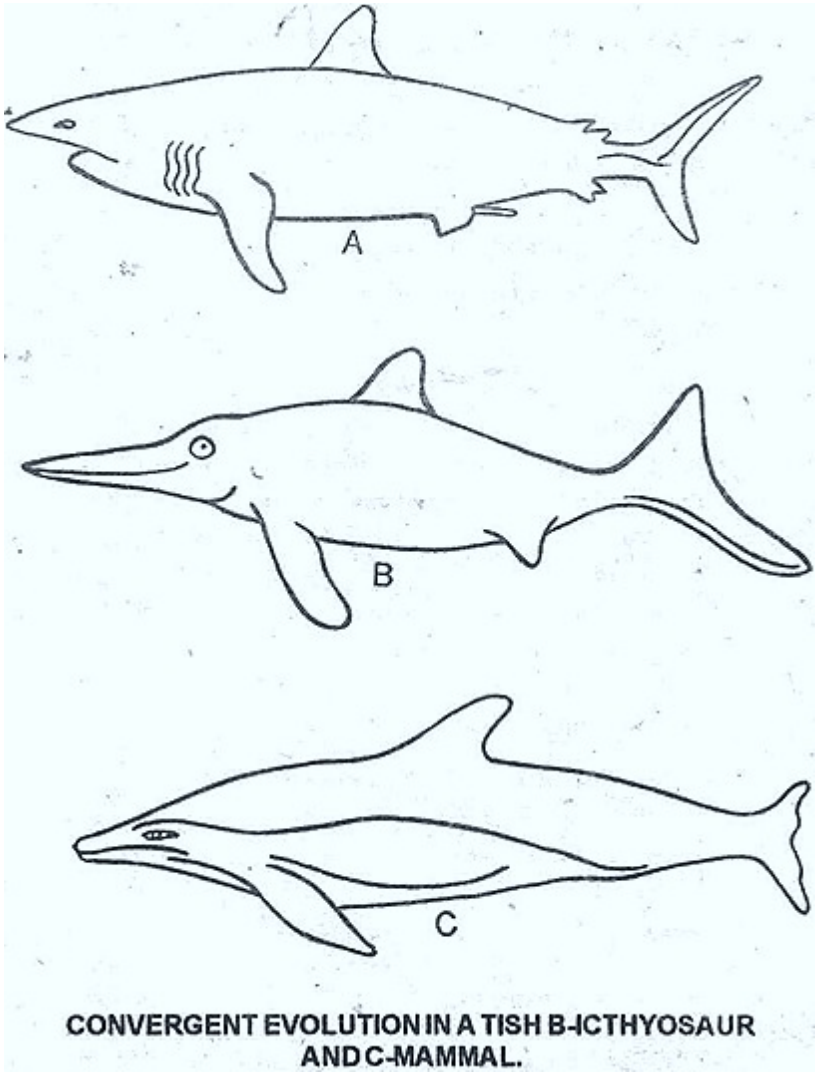
9 independent duplications.
7 cases where a mad3-like and a bub1-like protein arose out of a bubmad-like ancestor.



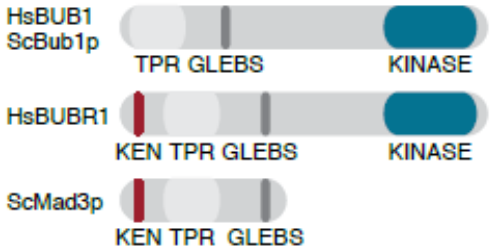
Suijkerbuijk *et al.*, Snel and Kops *Dev Cell* 2012

Recurrent (convergent/parallel) evolution in molecular systems!

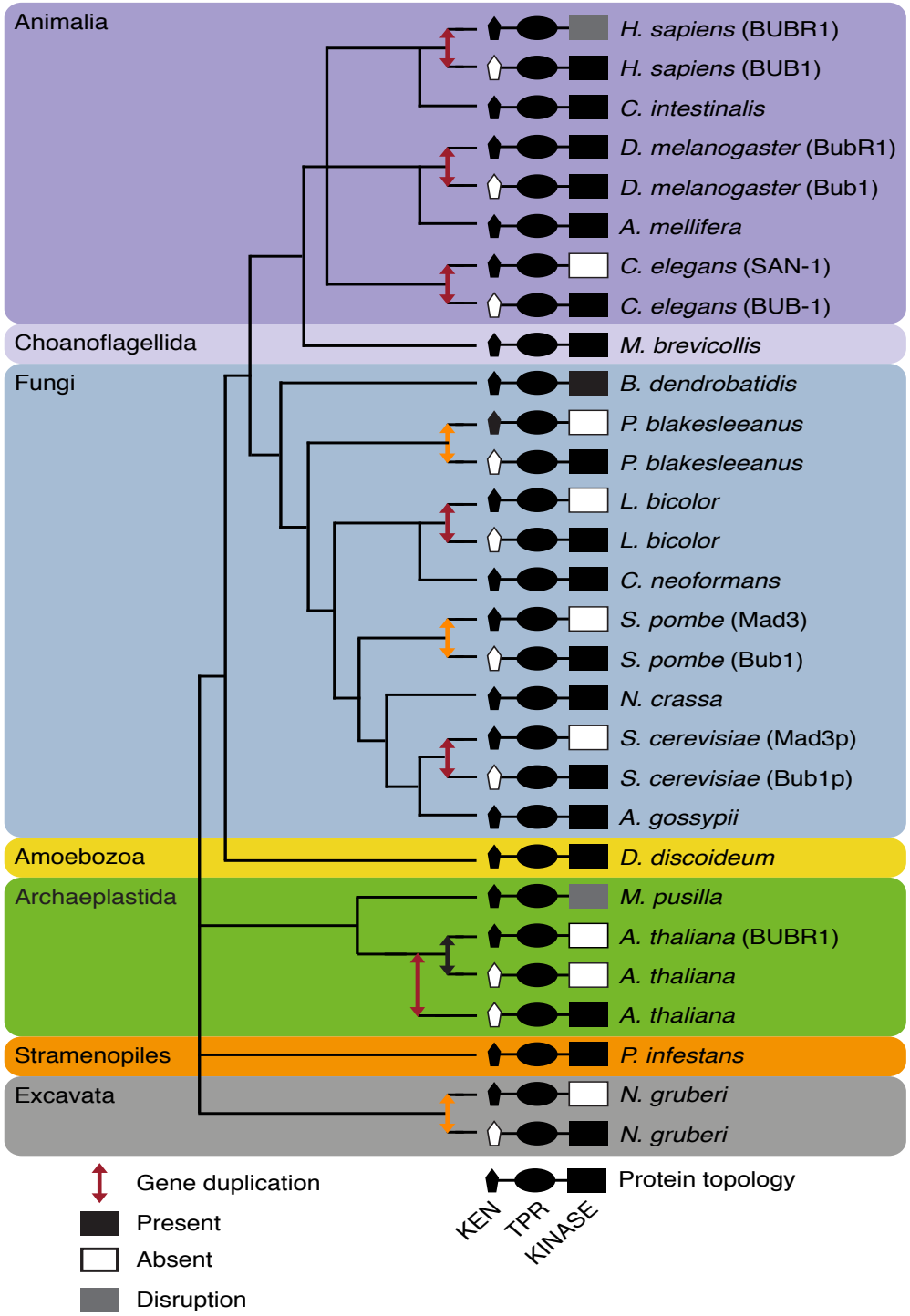




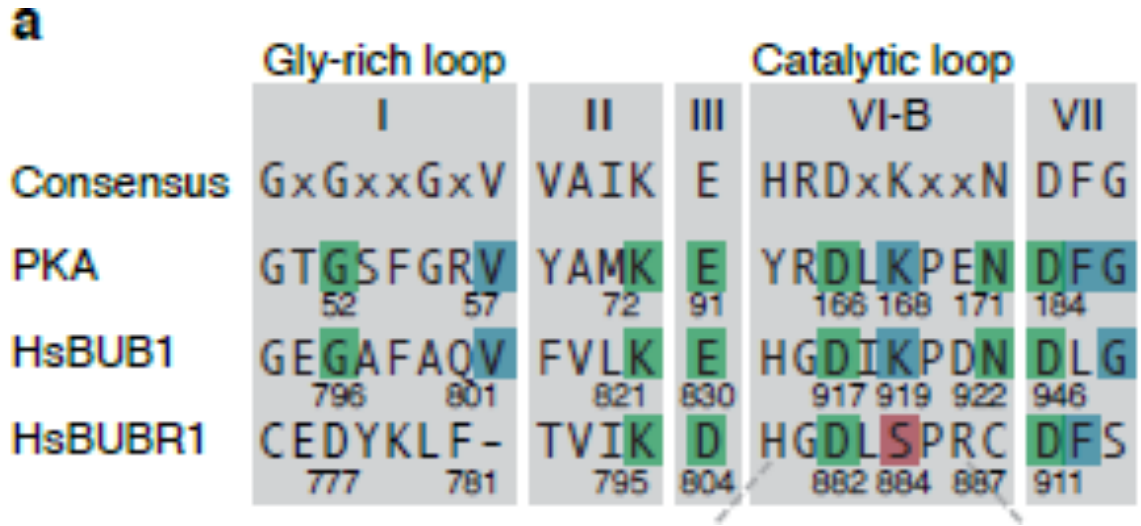
CONVERGENT EVOLUTION IN A TISH B-ICTHYOSAUR AND C-MAMMAL.



What about the kinase domain in human (and fly)



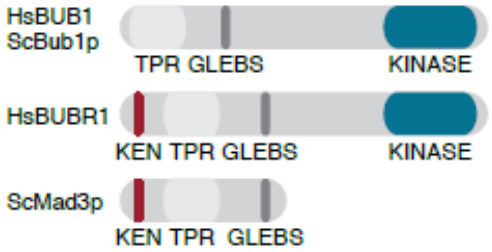
What about the kinase domain in human bubr1?



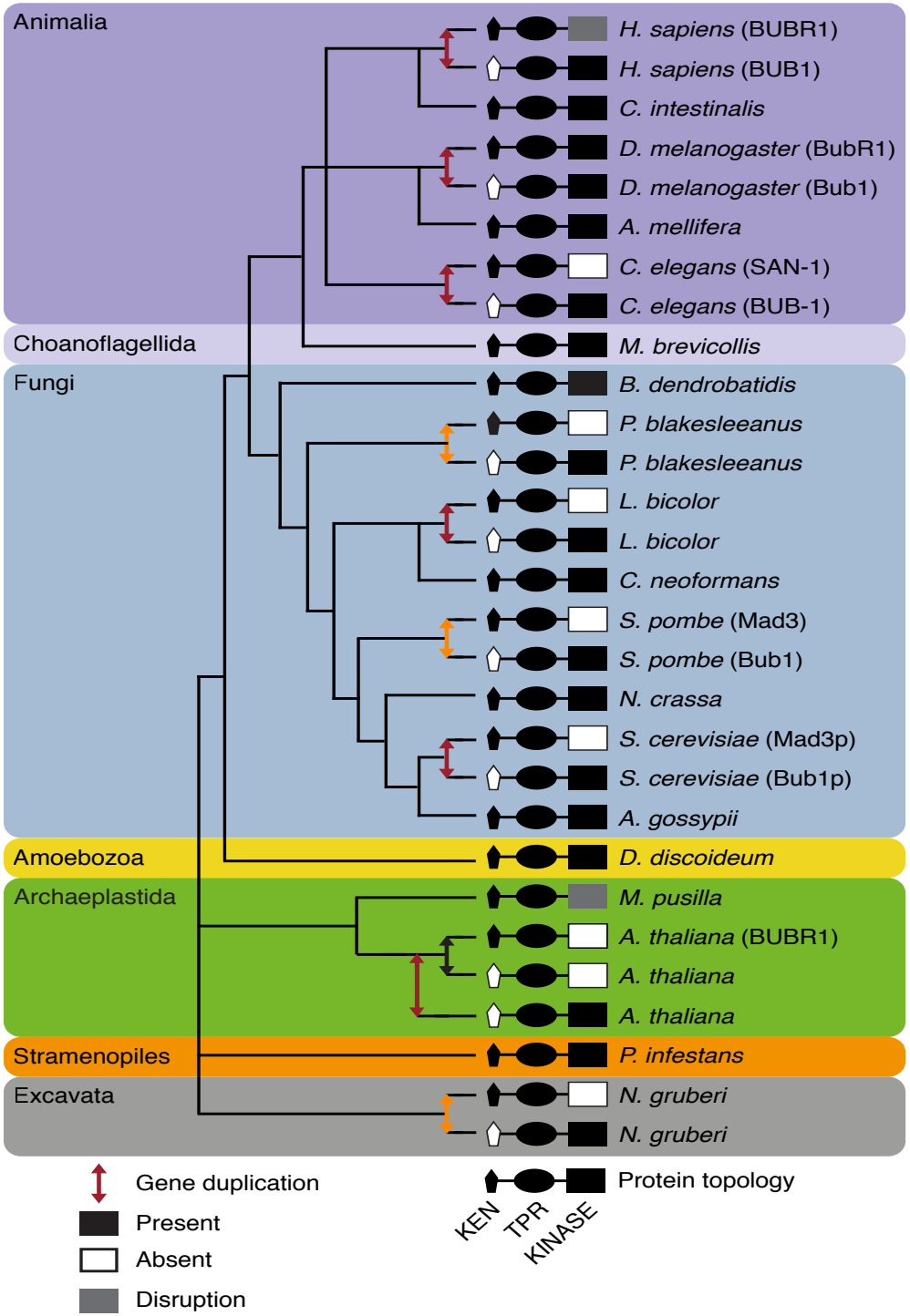
degeneration of motifs essential for catalysis

Further experiments showed vertebrates are not exception. The kinase domain of BubR1 lacks enzymatic activity.

"This explained the field's inability to identify substrates of BubR1, and dispelled a leading theory of SAC silencing based on inactivation of BubR1 after kinetochore-microtubule attachment."



9 independent duplications.
8 cases where a bub1-like protein and a protein without a (functional) kinase arose: mad3, bubr1



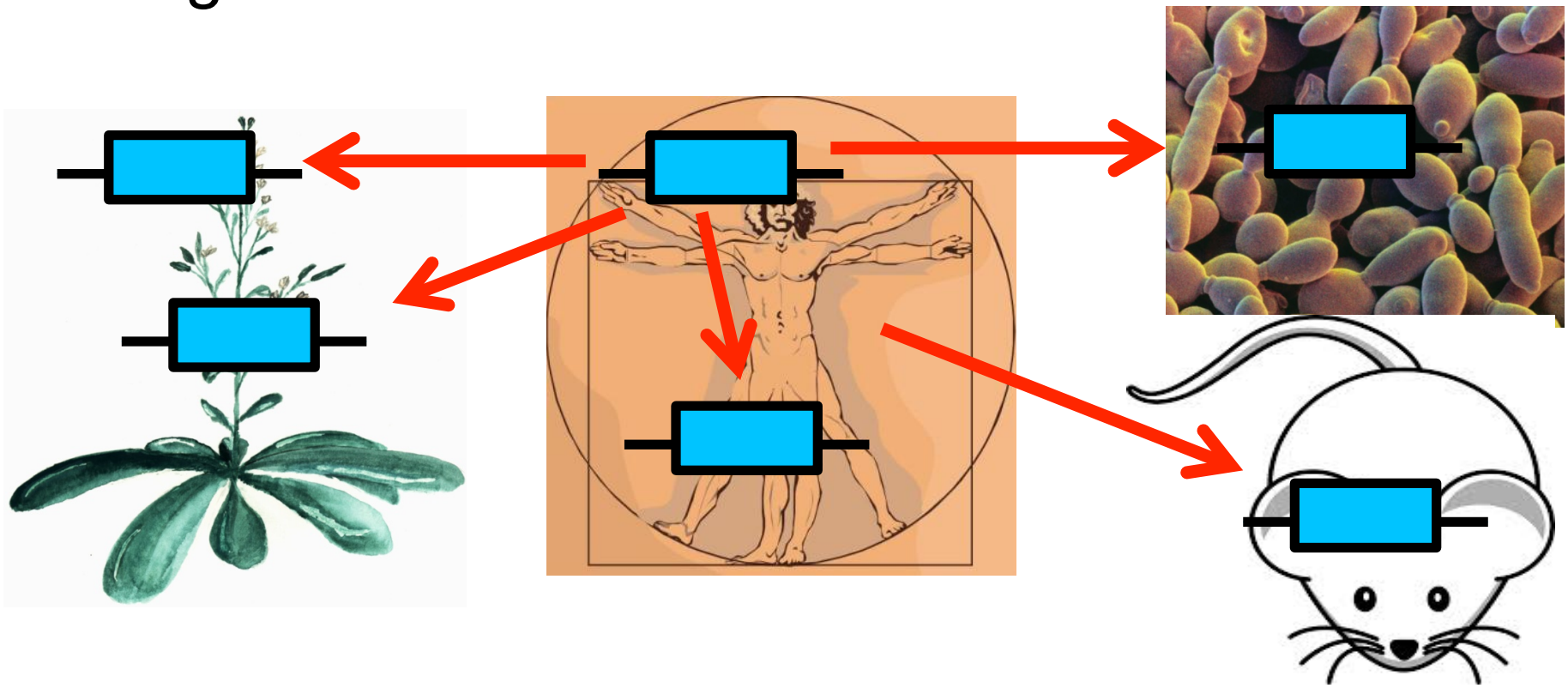
Suijkerbuijk *et al.*, Snel and Kops *Dev Cell* 2012

Evolutionary Cell Biology

- We can explain the relation between yeast MAD3, yeast & human BUB1 and human BUBR1
- We find a spectacular case of parallel evolution in a core cellular pathway,
- Strongly suggests recurrent functional specialization from a multifunctional ancestor, which was experimentally tested & validated
- *“This explained the field’s inability to identify substrates of BubR1, and dispelled a leading theory of SAC silencing based on inactivation of BubR1 after kinetochore-microtubule attachment.”*

General idea of today

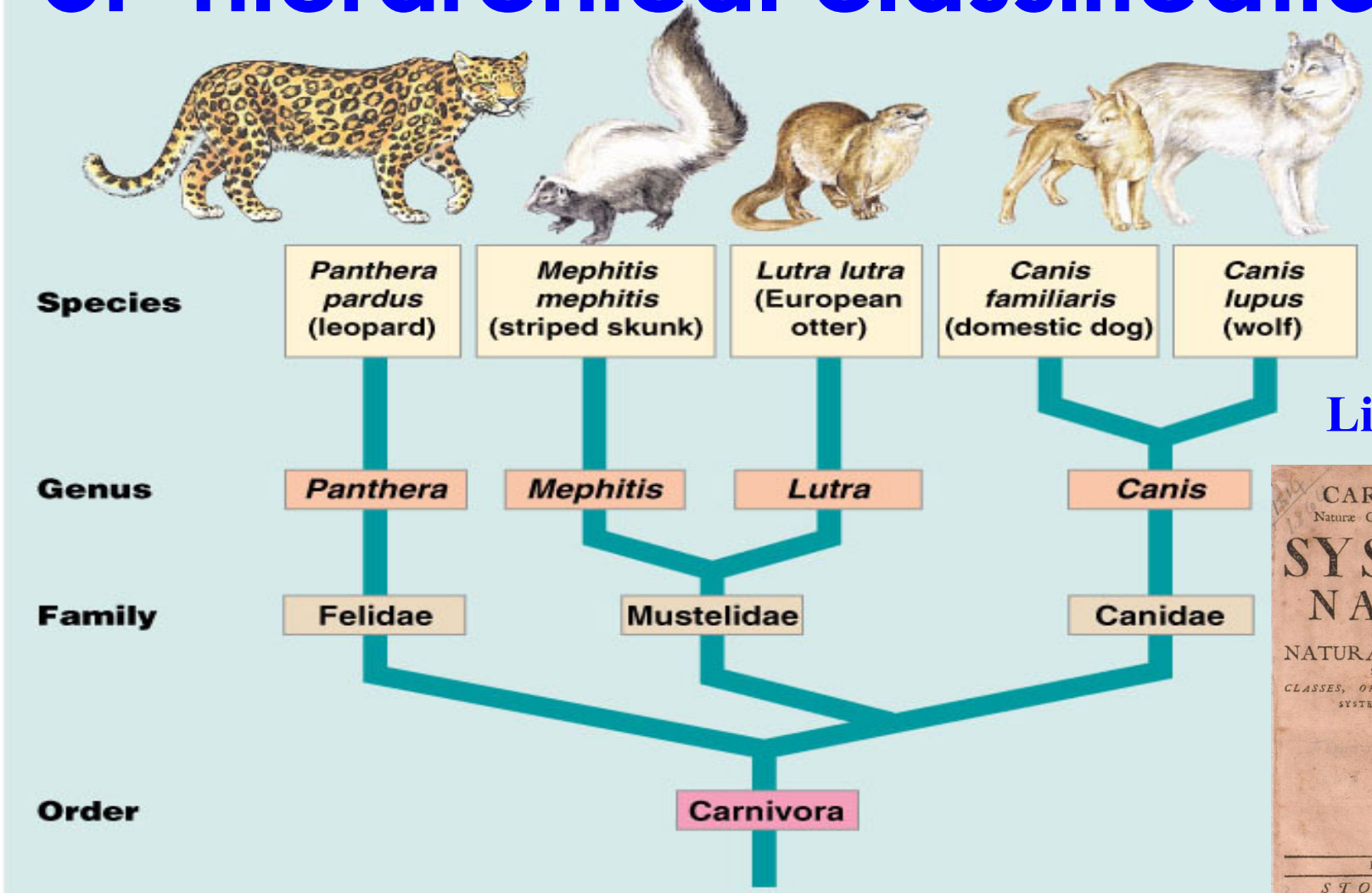
- Having all the genomes …
- What is the **relation** of my gene to *homologous* genes in the same and other organisms



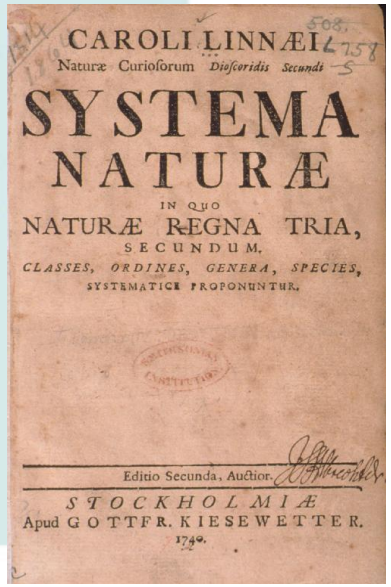
Today

- 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

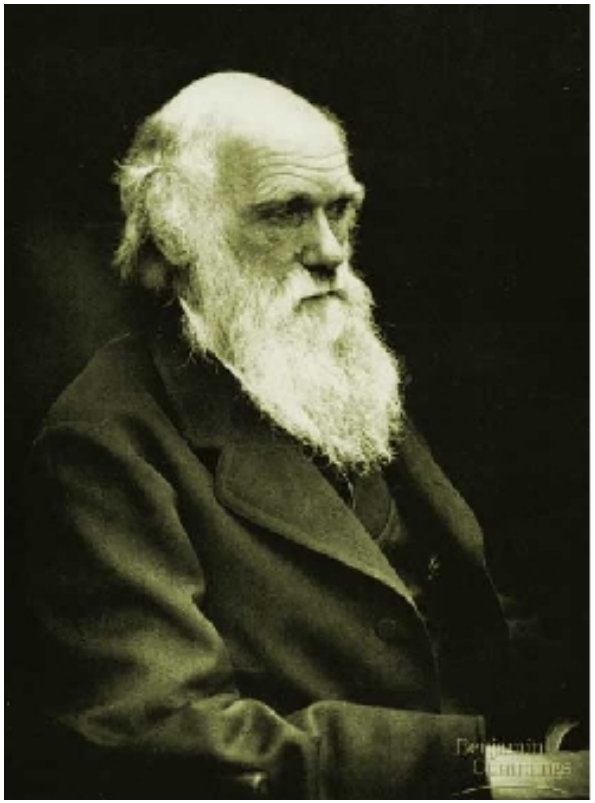


Linnaeus

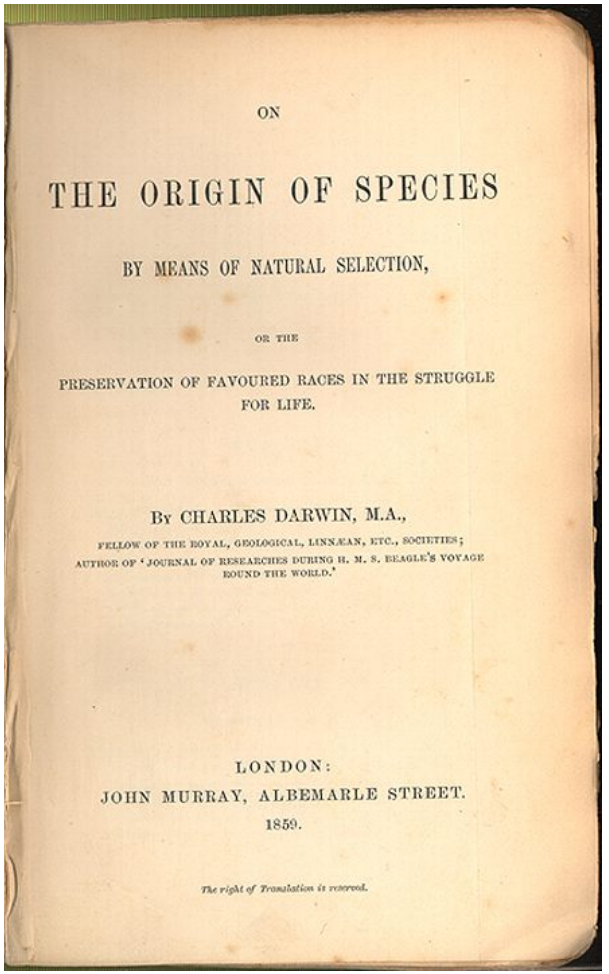


NB no information in skunk left / otter right

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.

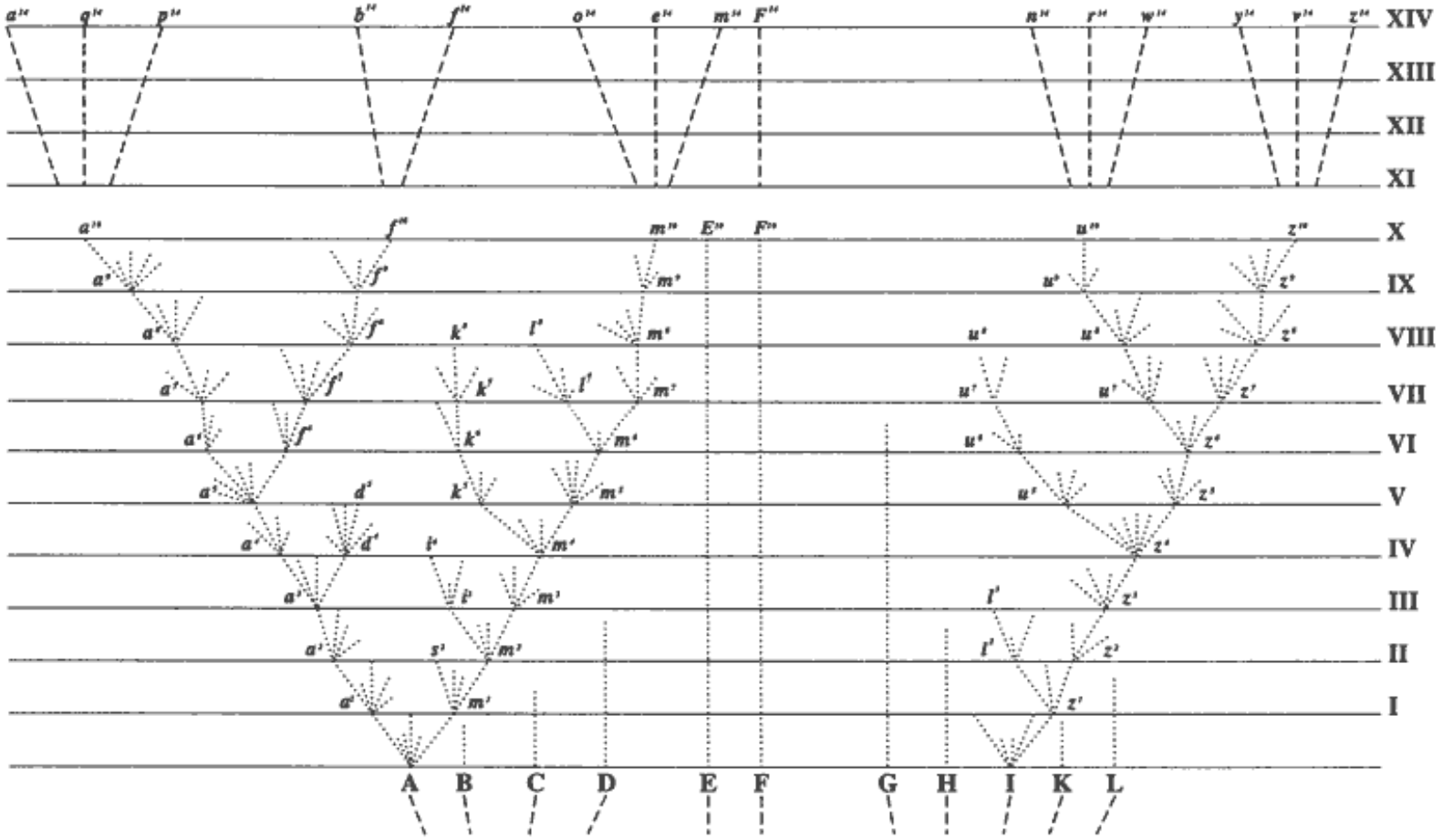
By CHARLES DARWIN, M.A.,

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

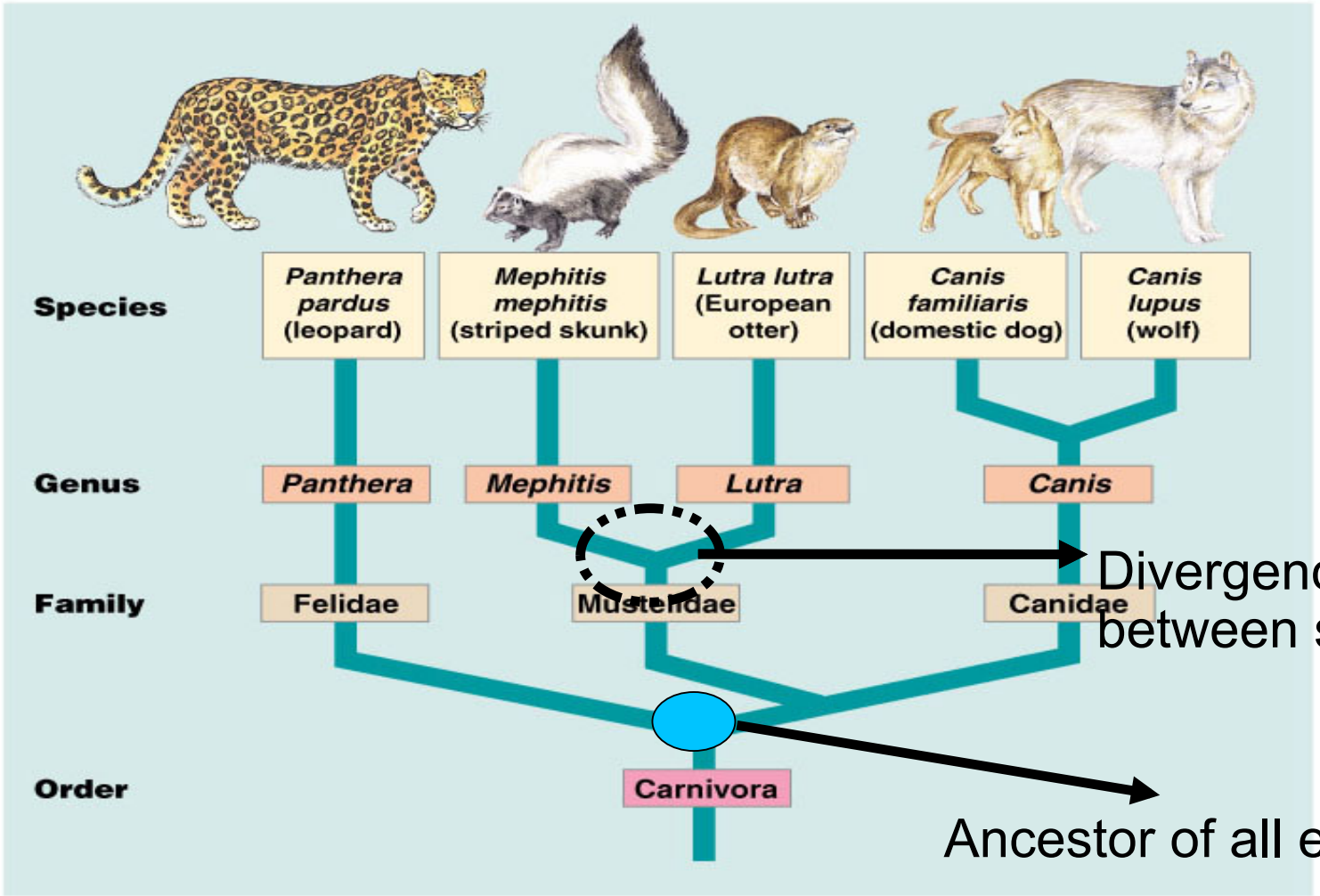
LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

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

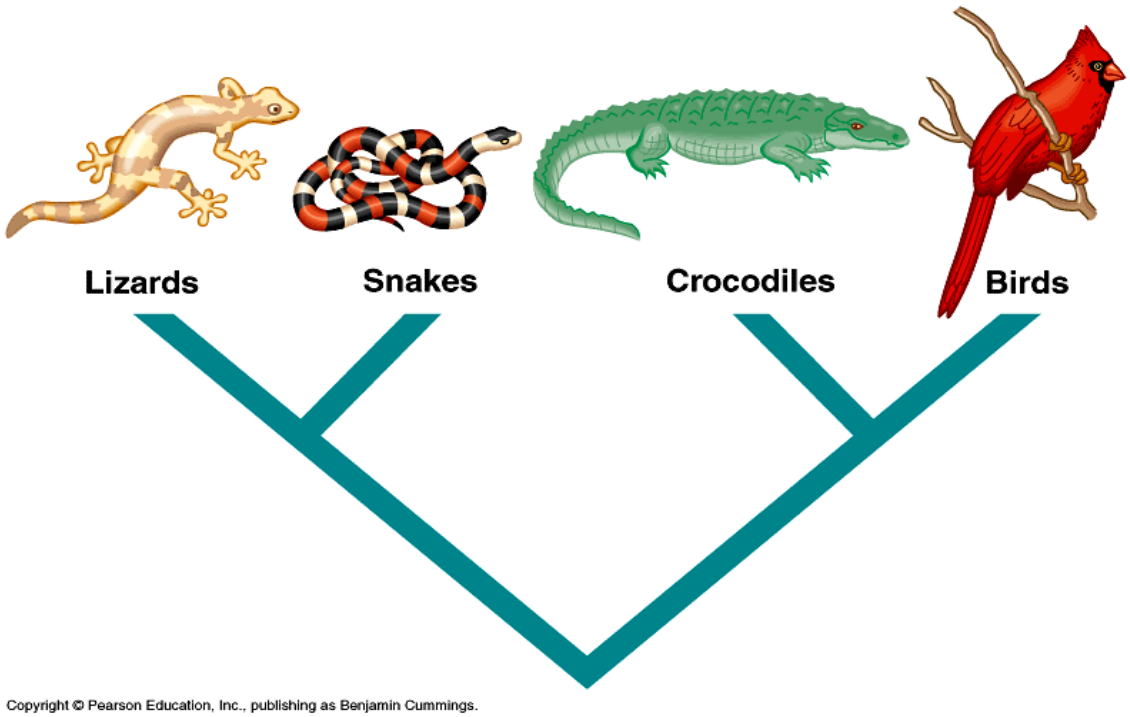


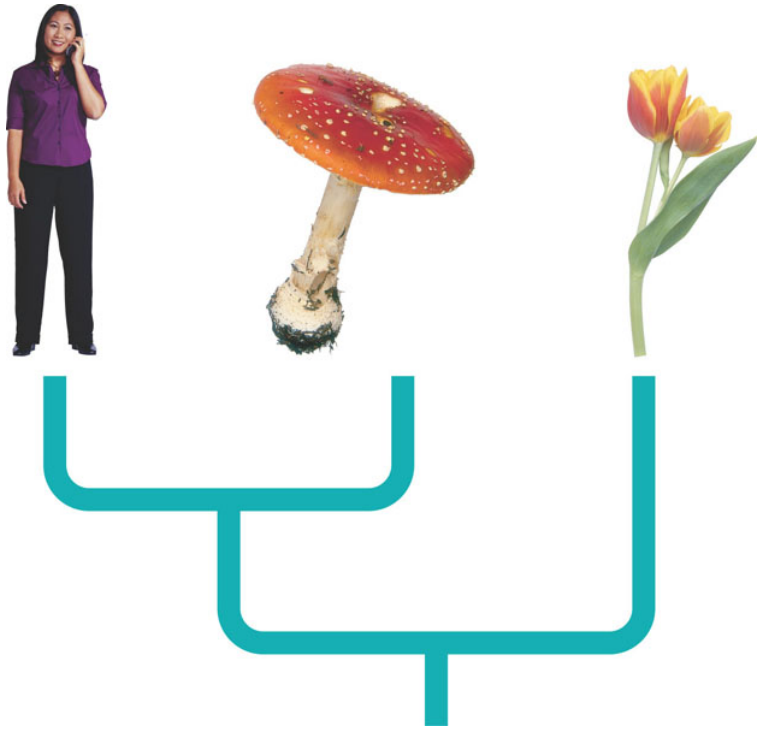
Phylogenetic tree: historical pattern of relationships among organisms: interpretation of a tree



NB still no information in skunk left / otter right

(molecular) Phylogenetic insights changed taxonomy





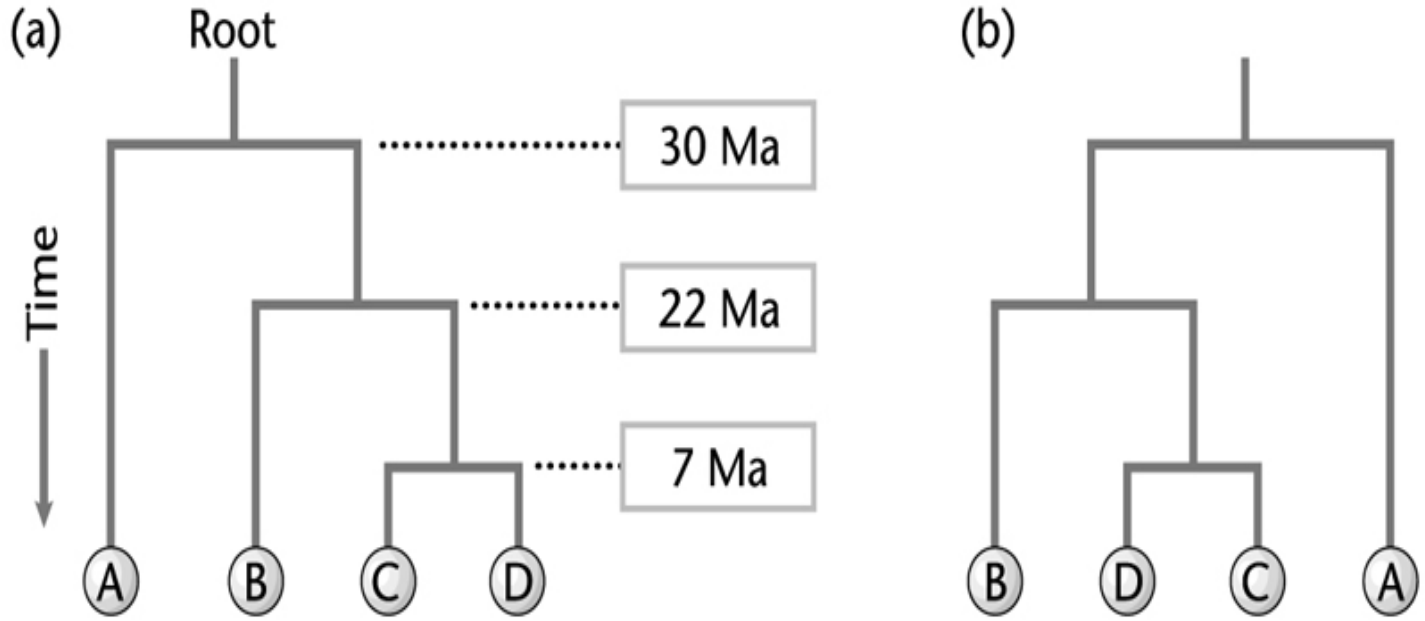
cladogram

?

Today

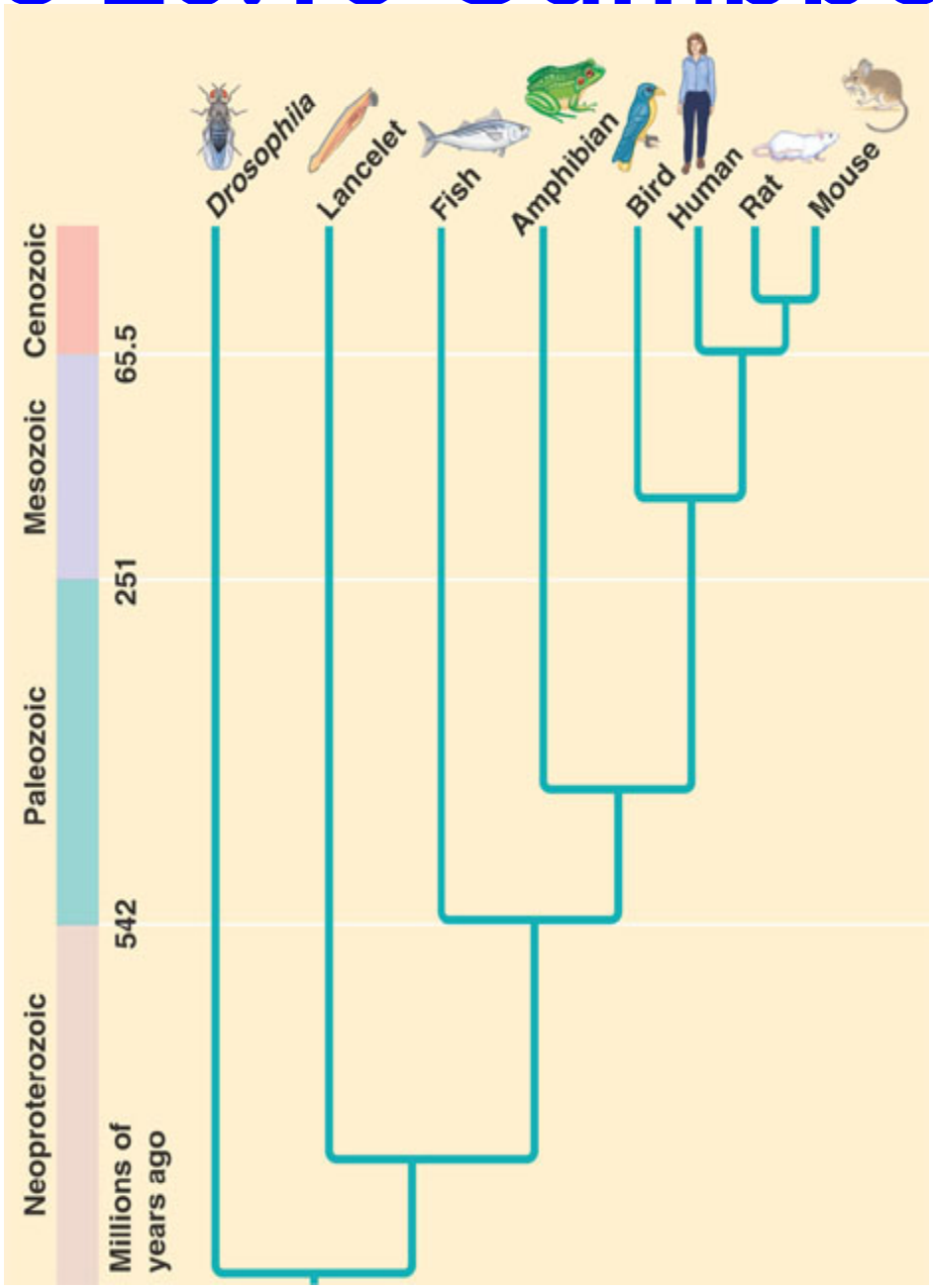
- What is a phylogenetic tree?
- How to “read” simple phylogenetic trees
  - Types of trees
  - Unrooted vs rooted
  - Molecular clock vs no molecular clock
- How to make a phylogeny
- How to root a phylogeny
- Afternoon Lecture

Rooted trees with molecular clock

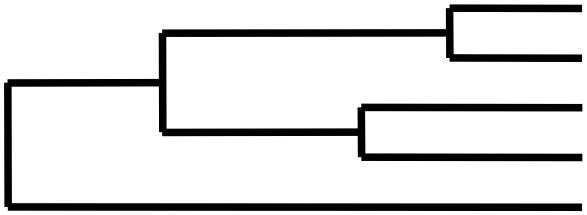


?

Figure 25.13 Campbell & Reece

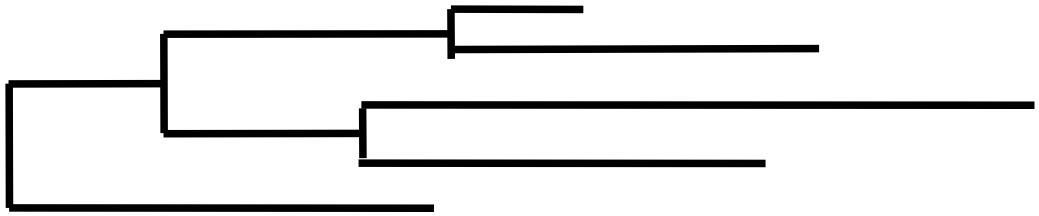


Evolutionary clock speeds



Uniform clock: leads to identical distances from root to leaves

(ultrametric tree)

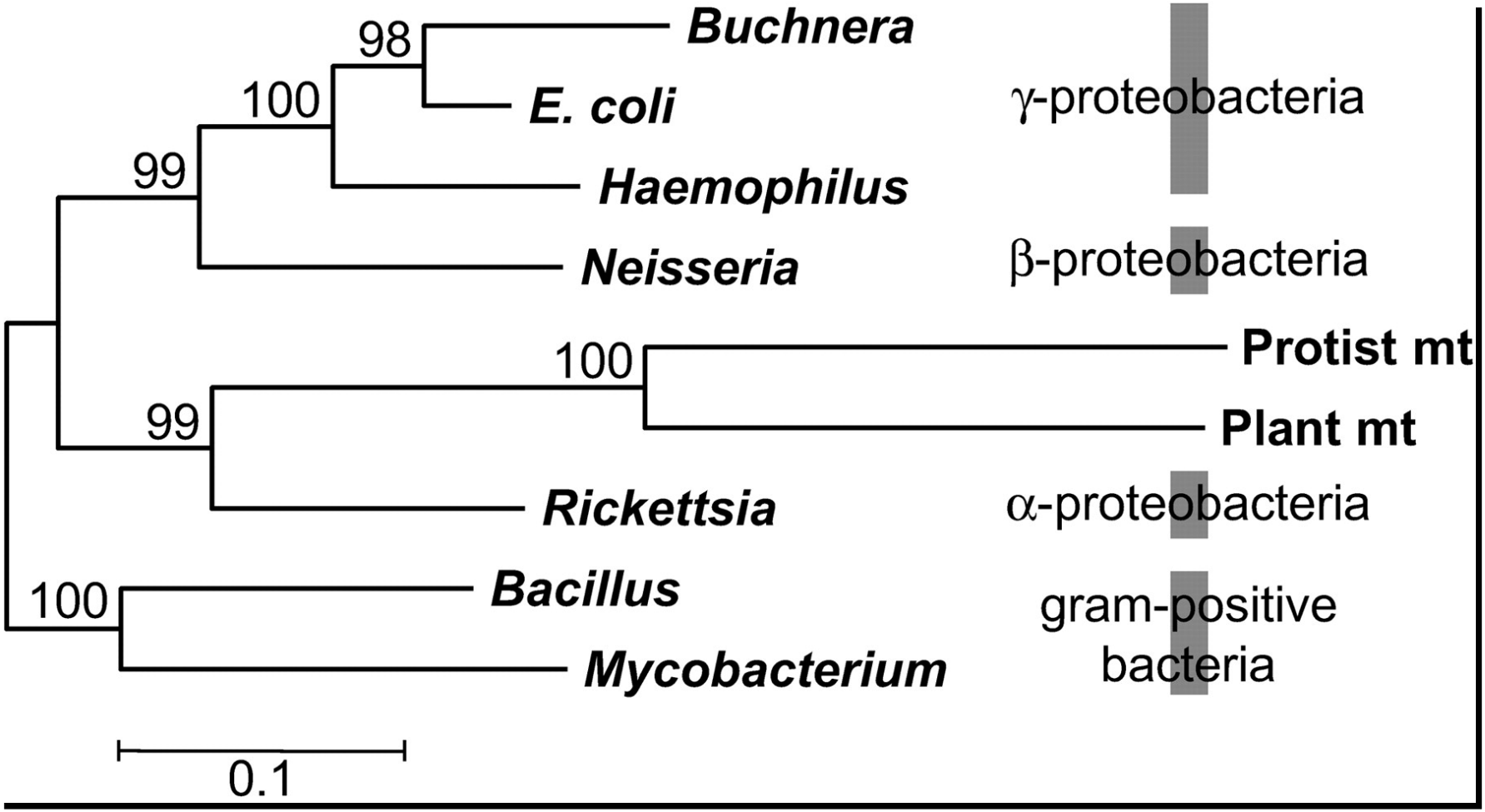


Non-uniform evolutionary clock: leaves have different distances …

(additive tree)

Root unknown: unrooted trees

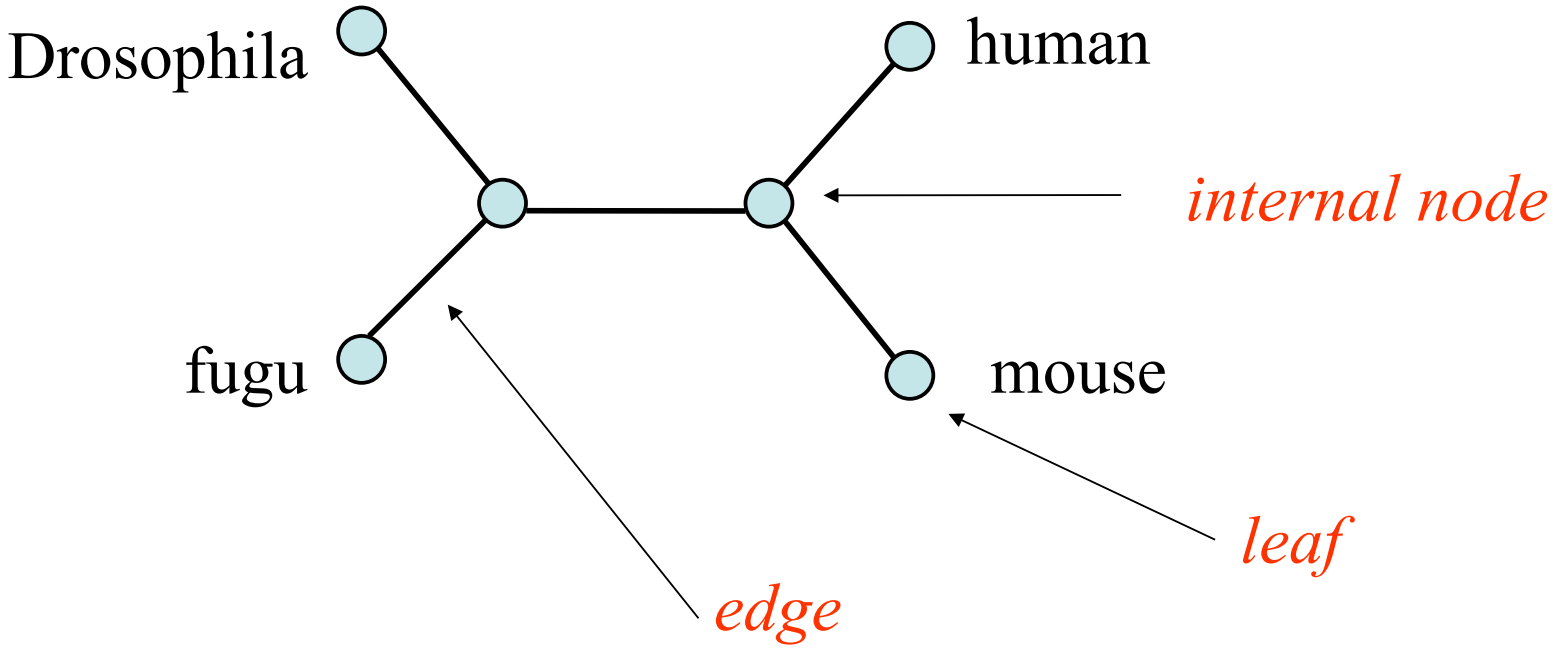
Unequal rates between species are a very real phenomenon



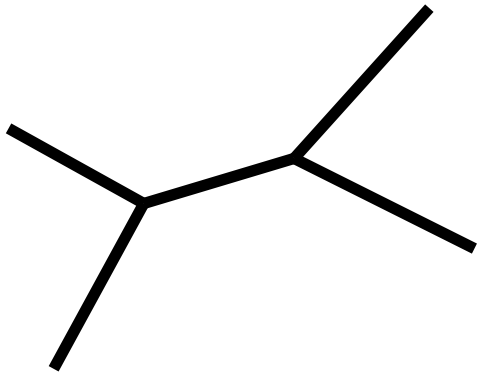
phylogram

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)

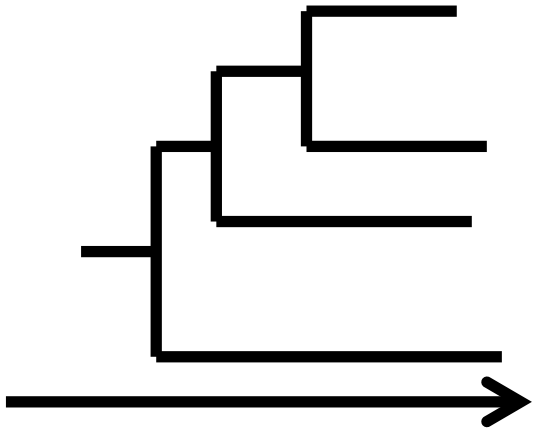


Radial tree (always explicitly unrooted)



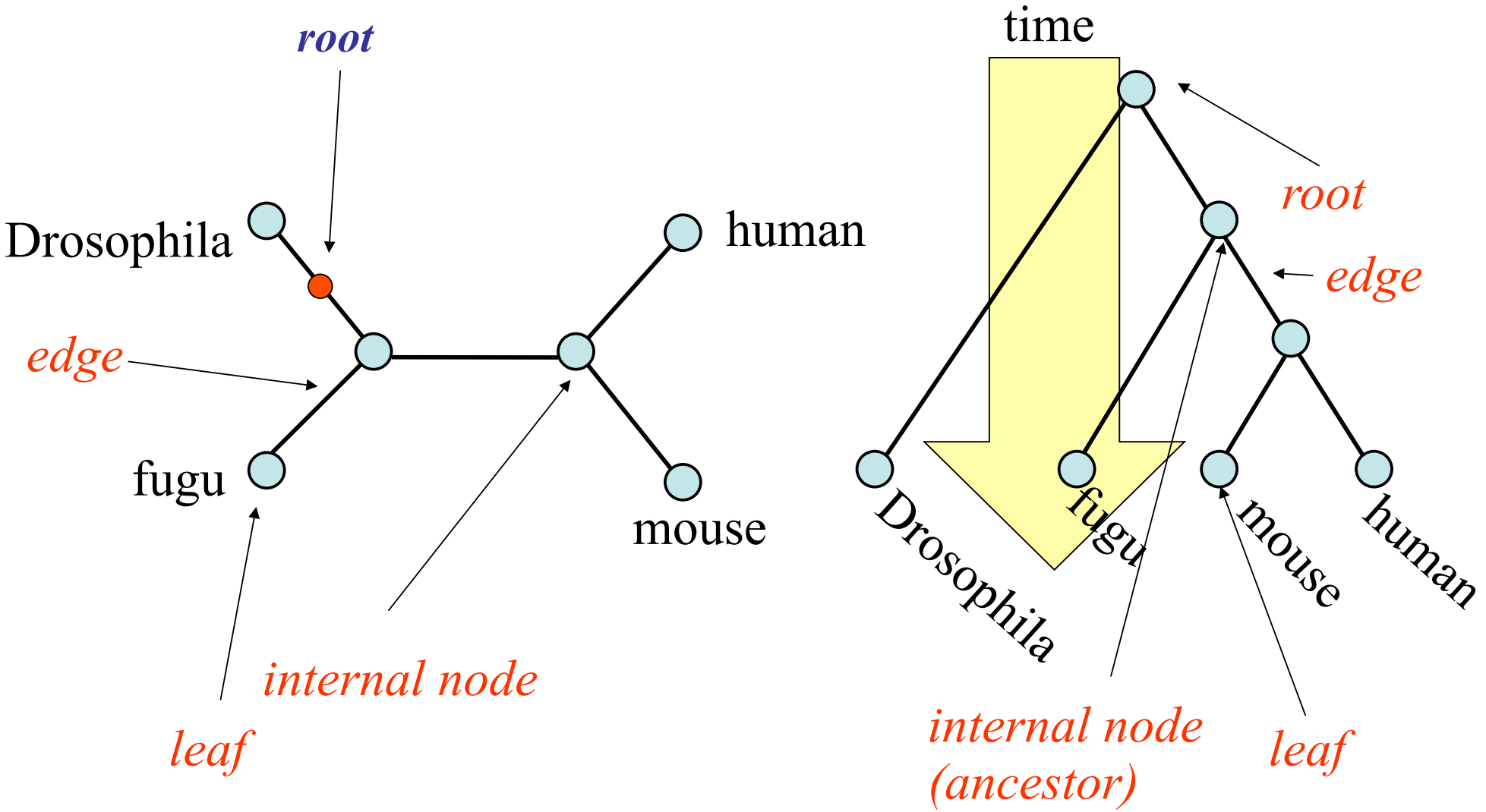
Phylogram (explicitly rooted)

Rooted = directed tree

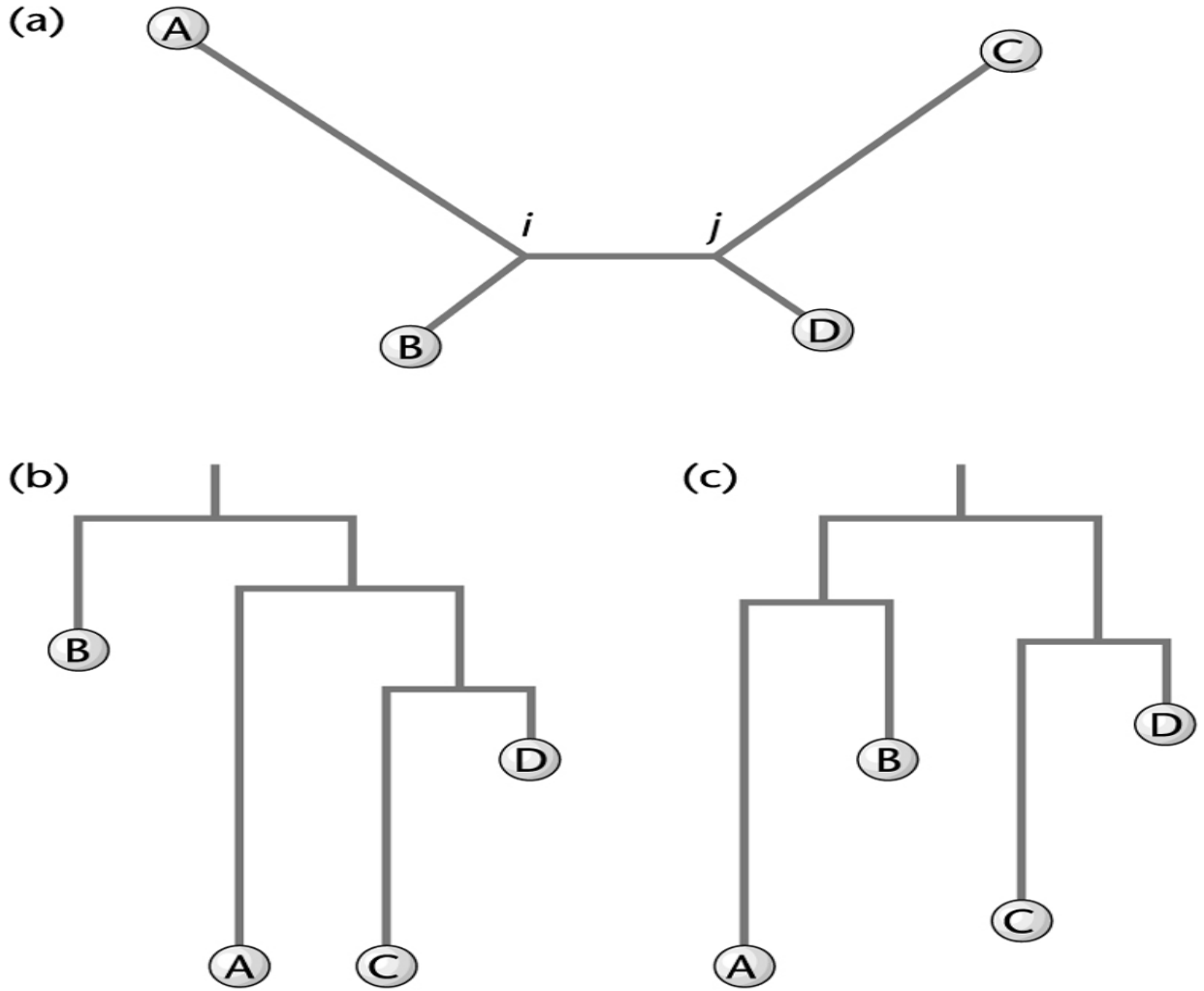


… both are phylogenies

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



One unrooted tree can be turned into multiple rooted trees



?

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

1) Alignment

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

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

Phylogenetic tree by Distance methods (Clustering)

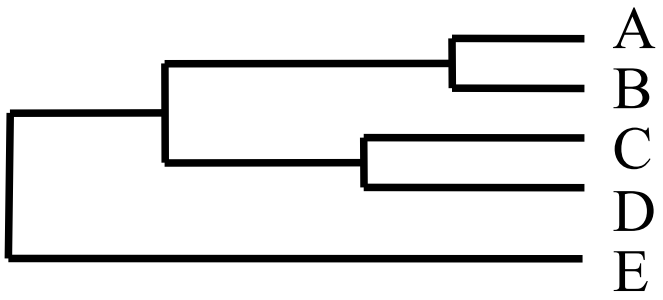
```
A  a t g a c c c c g
B  a t g a c c c c a
C  a t g a c g t c t
D  a t g a c g c g t
E  t t g t t c a a t
```

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:

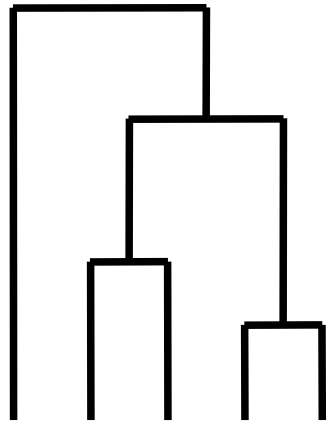
- Fill distance matrix with pairwise distances
- Start with N clusters of 1 element (gene) each

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:

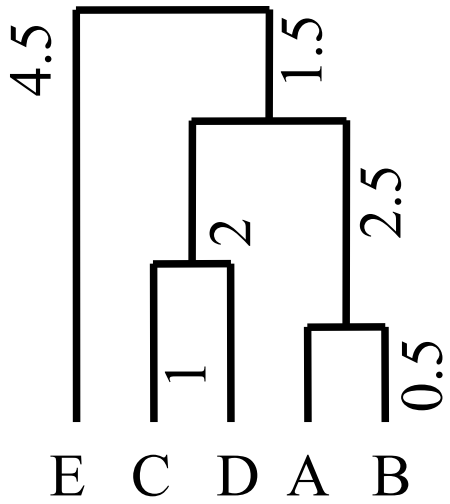
- When only two clusters i, j remain, put root at $d_{ij}/2$

```
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	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

1) Alignment

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

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

Model based

Multiple sequence alignment

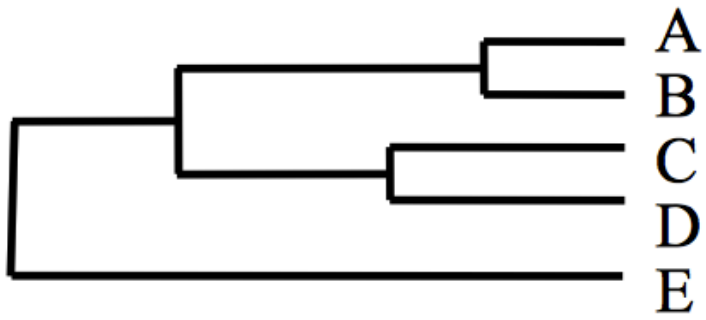
```
A  a t g a c c c c
B  a t g a c c c c
C  a t g a c g t c
D  a t g a c g c g
E  t t g t t c a a
```

Model:" sequence evolution happens like this"

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

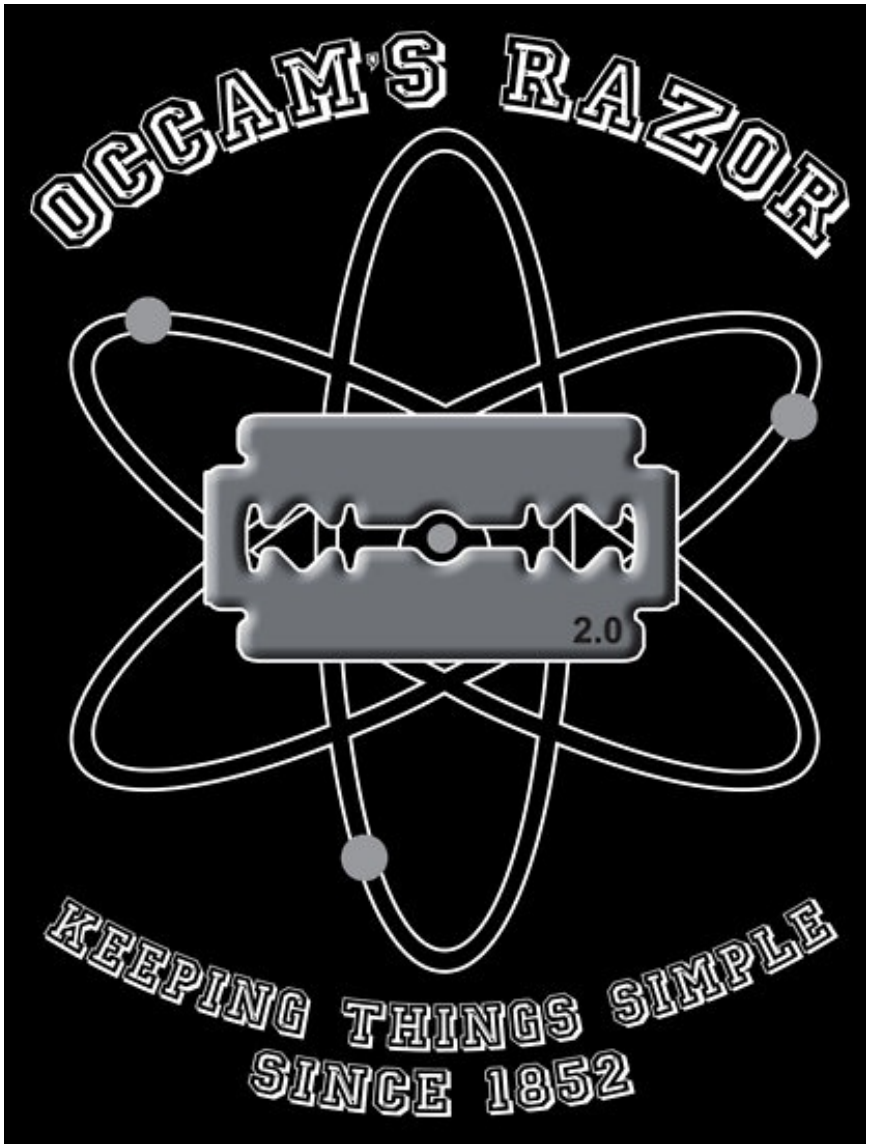
Find best fit to evolutionary model

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

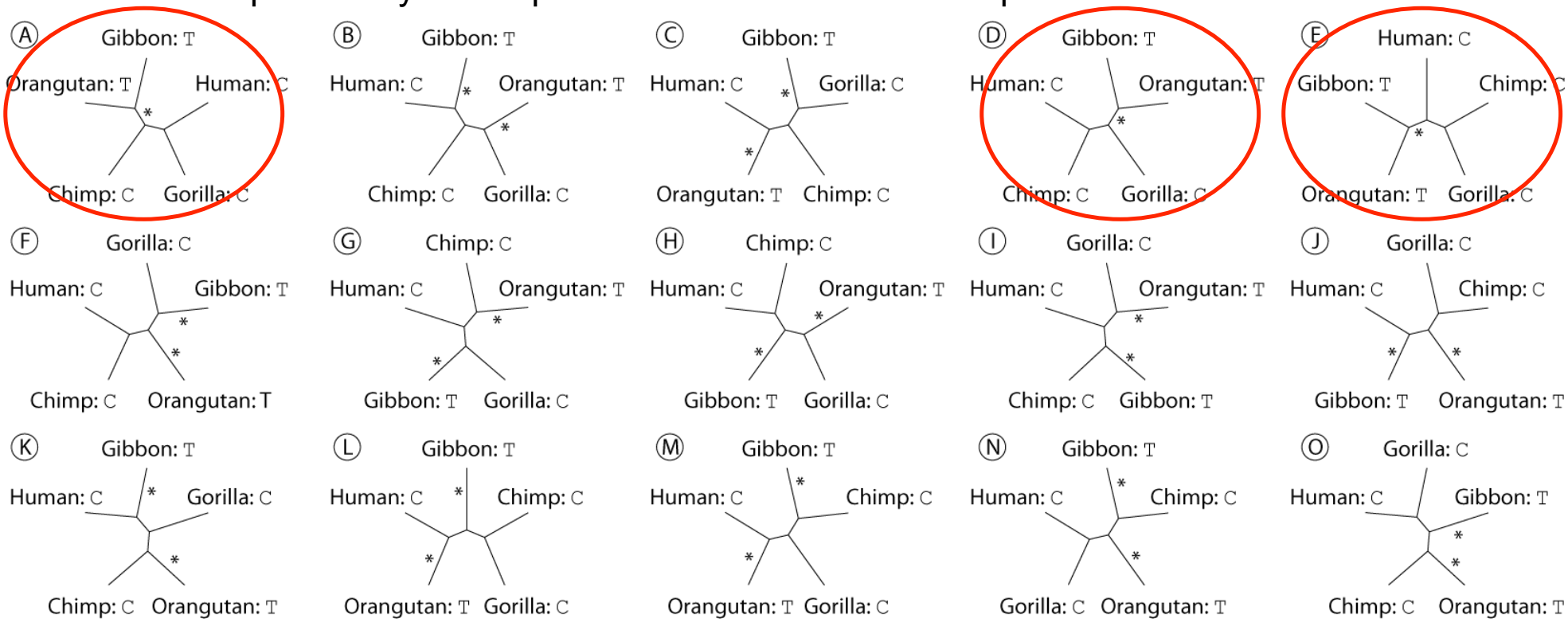


Maximum parsimony (MP)

- 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



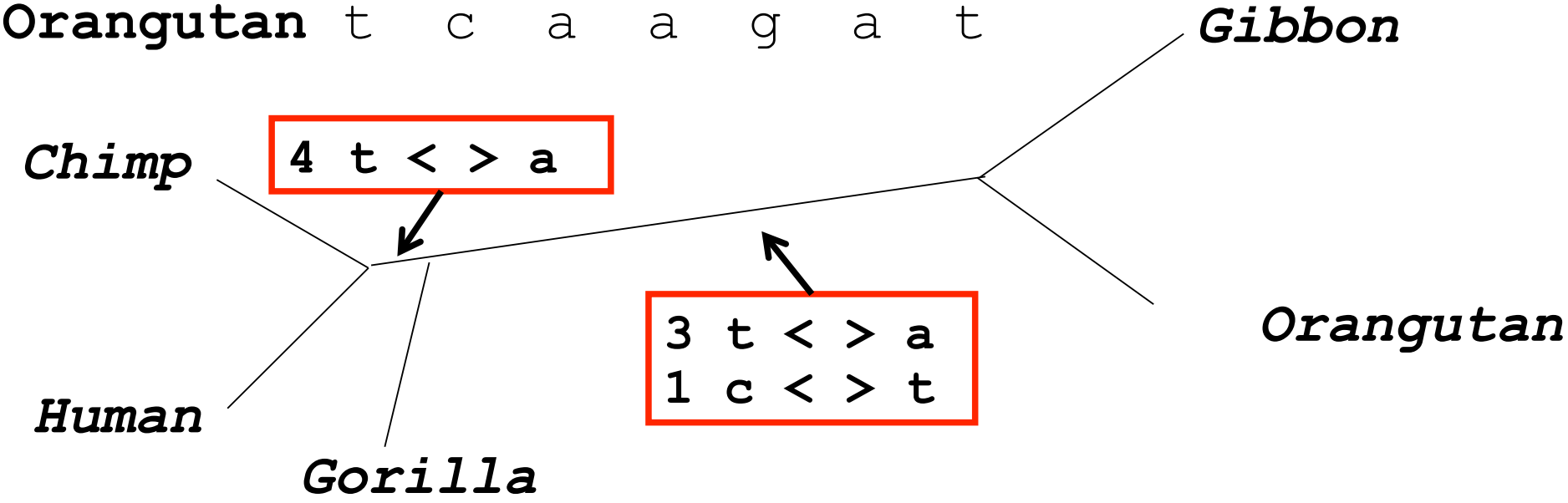
Maximum parsimony (MP)

- 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, 2.84×10^{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) 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.

Be careful:
Garbage in garbage out!

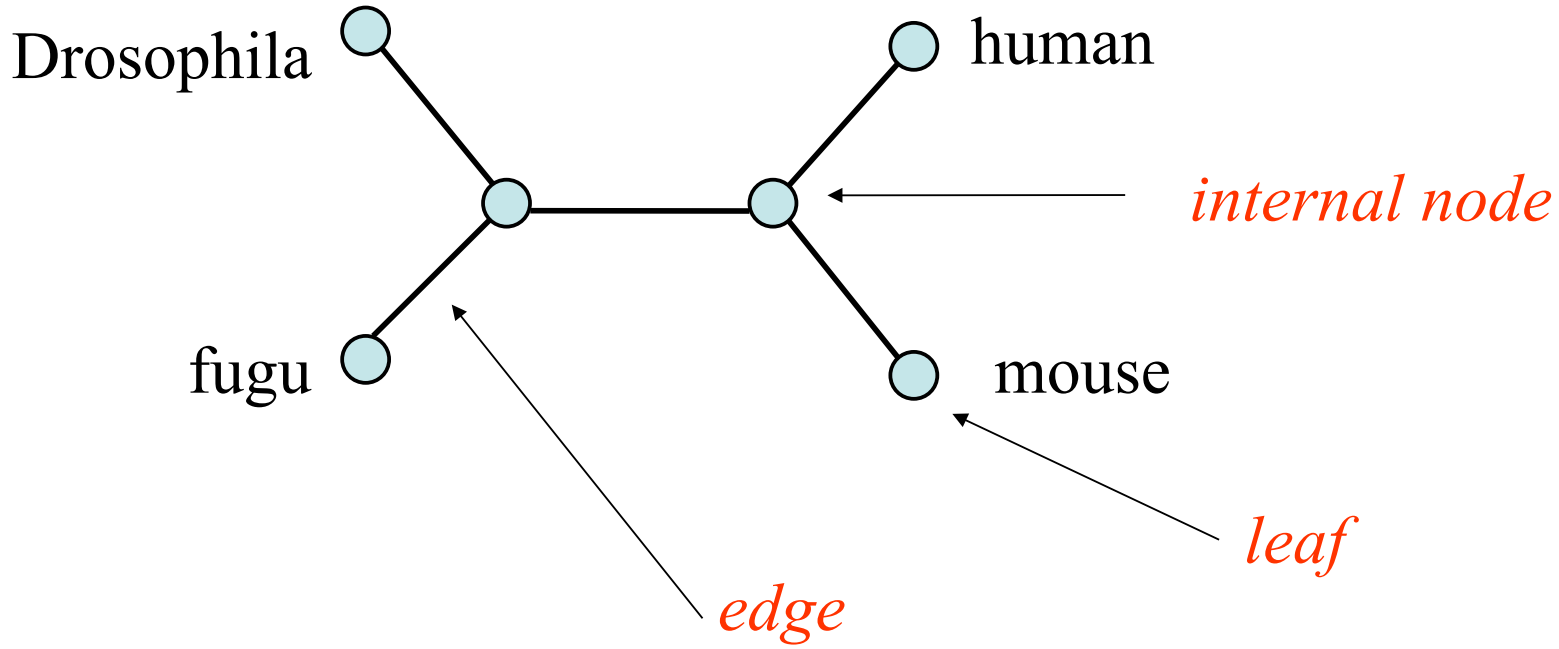
- Non homologous sequences will be aligned by e.g. clustalx *and* any phylogeny program will make a tree
- Similarly unaligned sequences or very poorly sequences will nevertheless be turned into a tree by any phylogeny program

Today

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

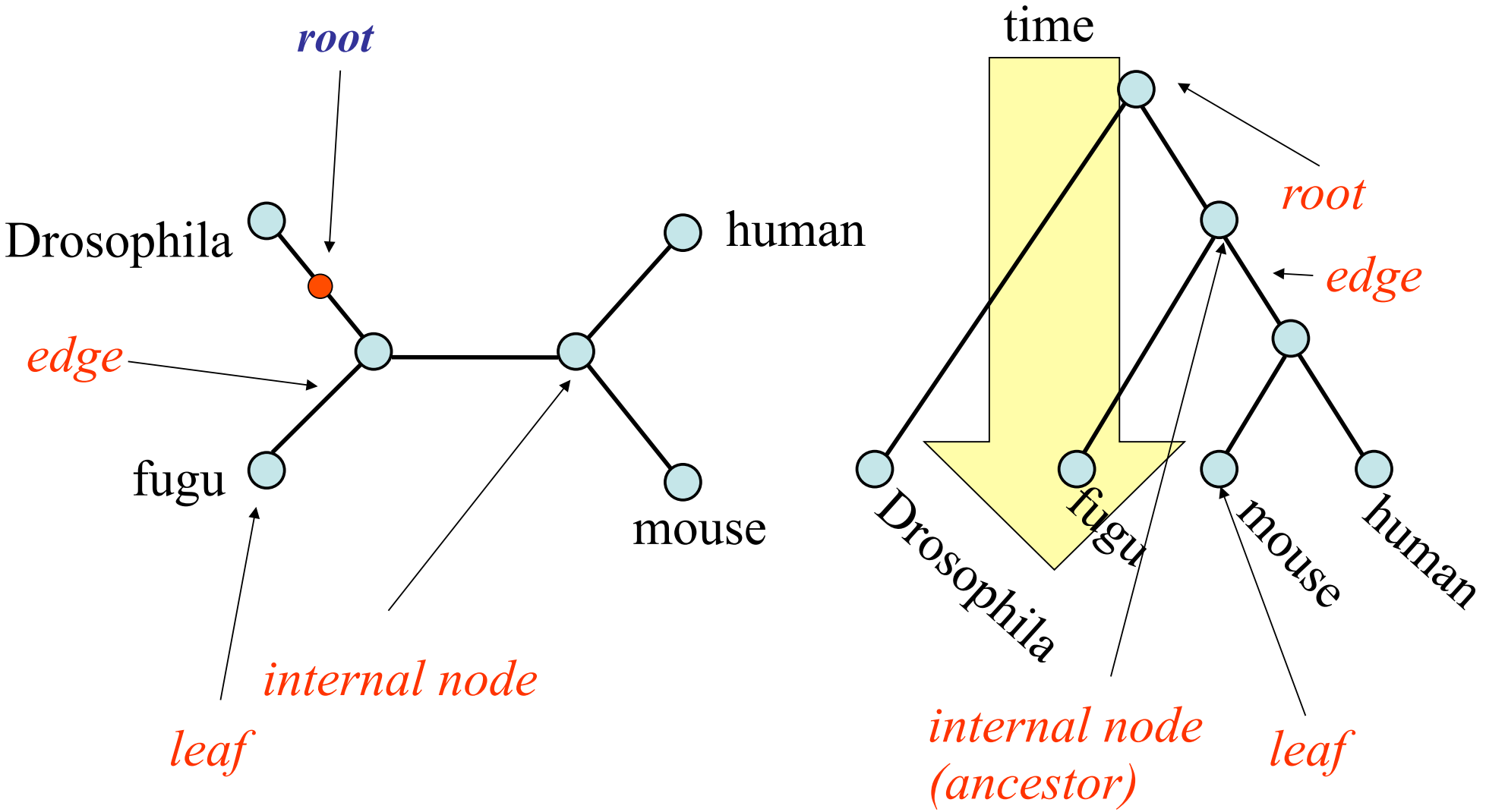
Unrooted trees

- 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

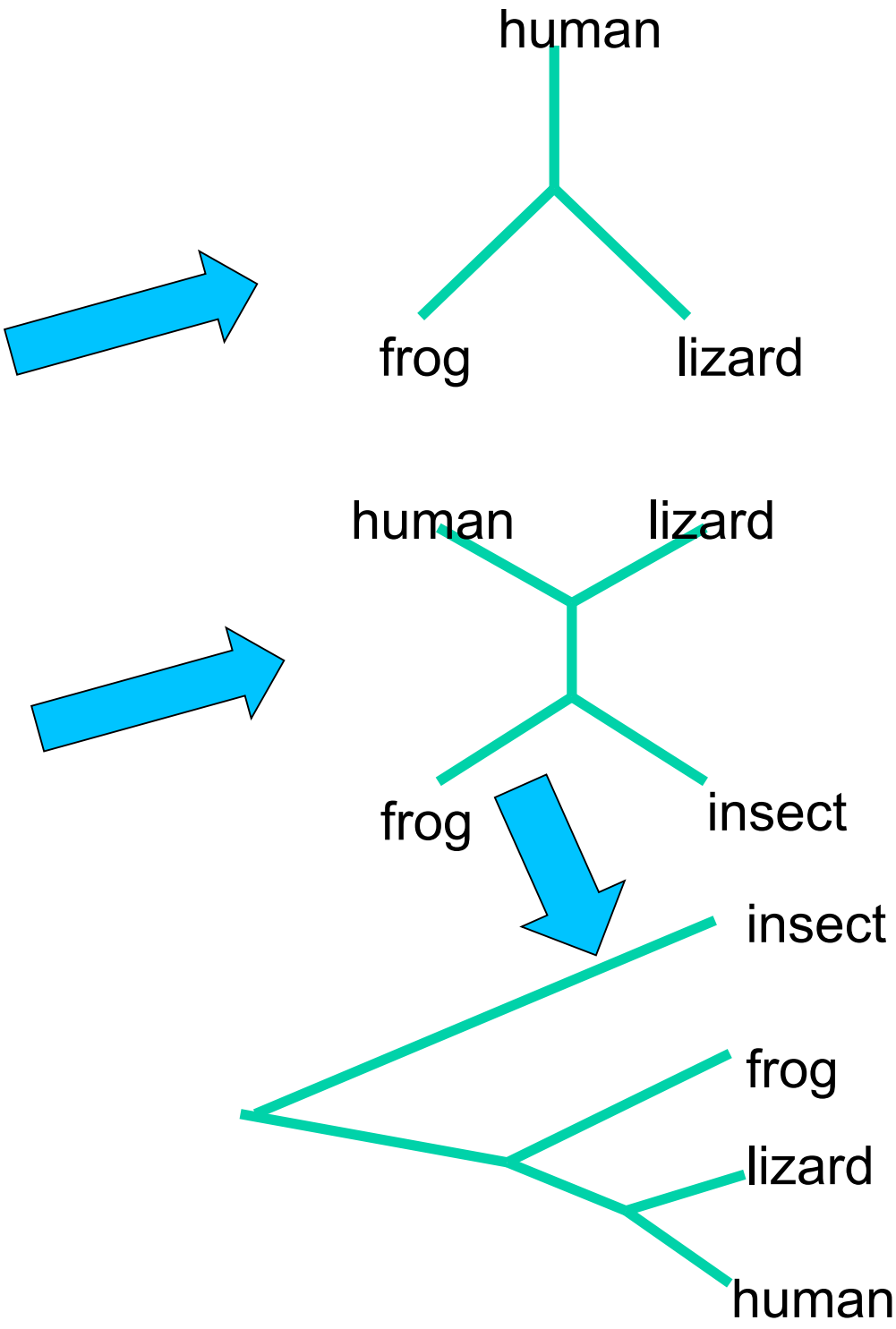
Introduce a root



How to root a tree: outgroup

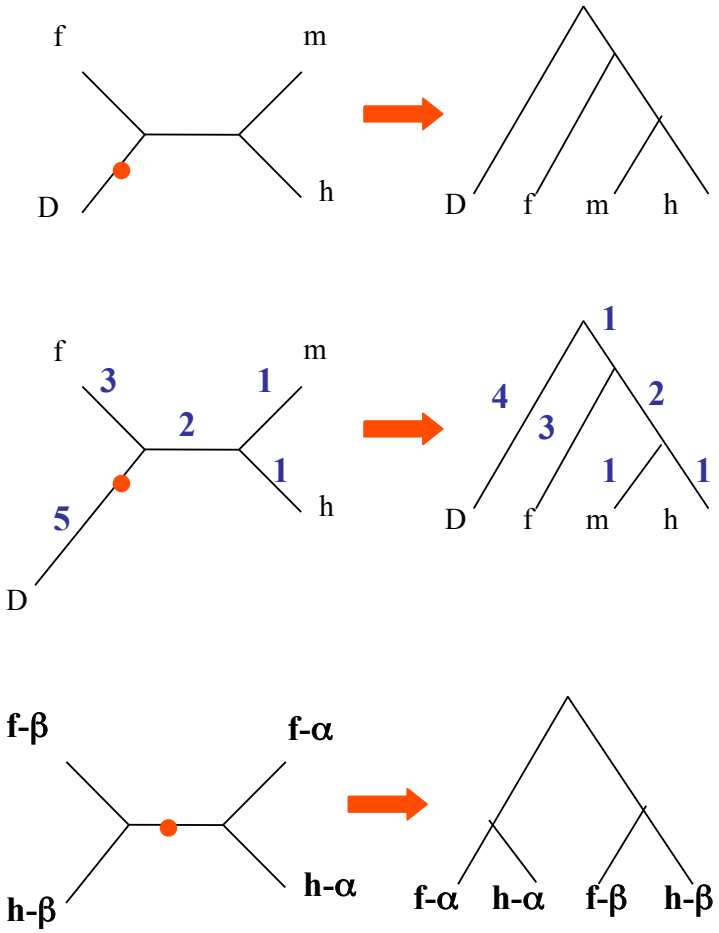
```
Human ccttgaa
  Frog ccttgat
Lizard ccttgac
```

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



How to root a tree

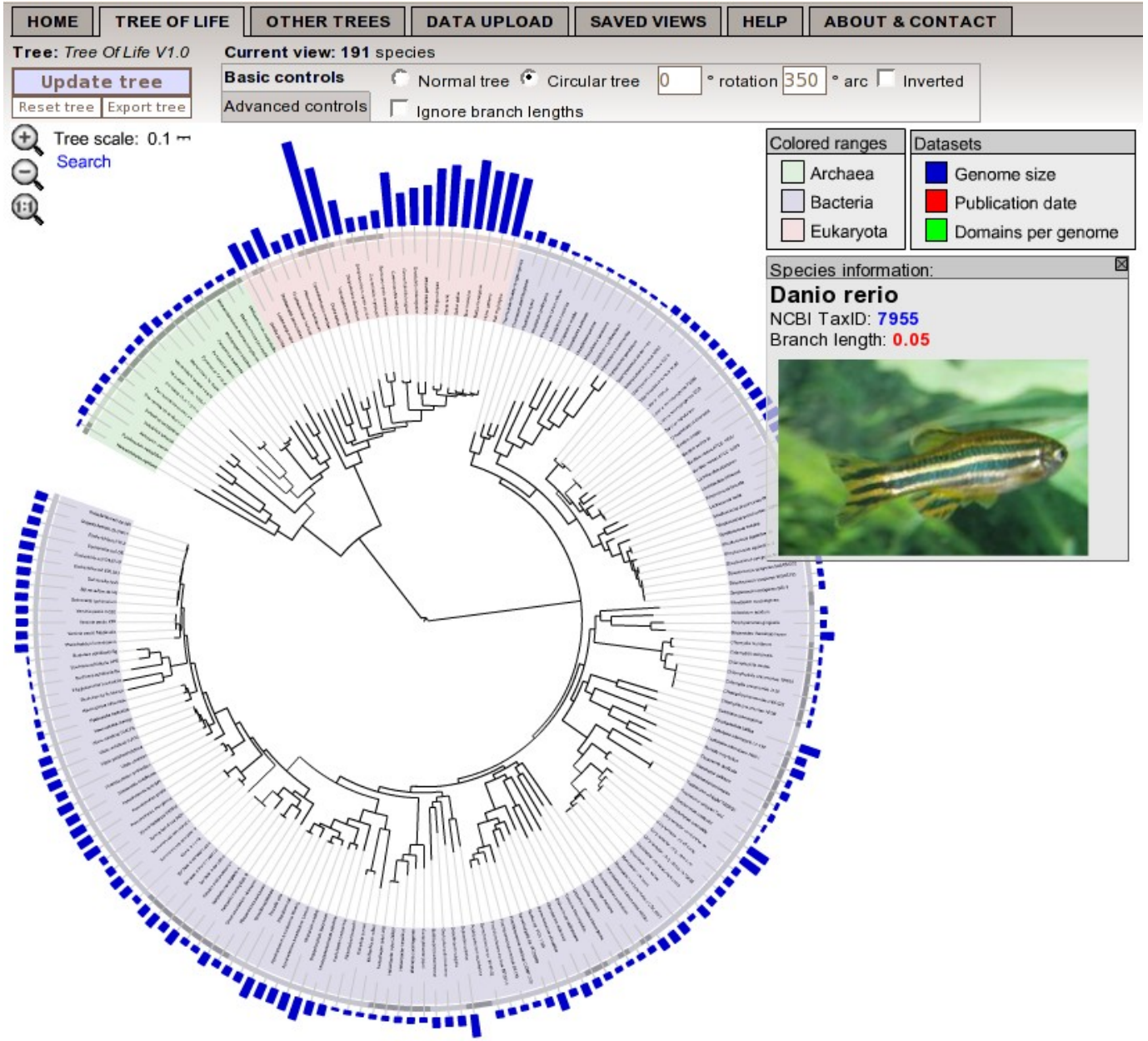
- Outgroup – place root between distant (still homolog) sequence and rest group
- Midpoint – place root at midpoint of longest path (sum of branches between any two leafs)
- Gene duplication – place root between paralogous gene copies



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


HOME
TREE OF LIFE
OTHER TREES
DATA UPLOAD
SAVED VIEWS
HELP
ABOUT & CONTACT
Tree: Tree Of Life V1.0
Current view: 191 species
Update tree
Reset tree
Export tree
Basic controls
Normal tree
Circular tree
0
° rotation
350
° arc
Inverted
Advanced controls
Ignore branch lengths
Tree scale: 0.1
Search
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