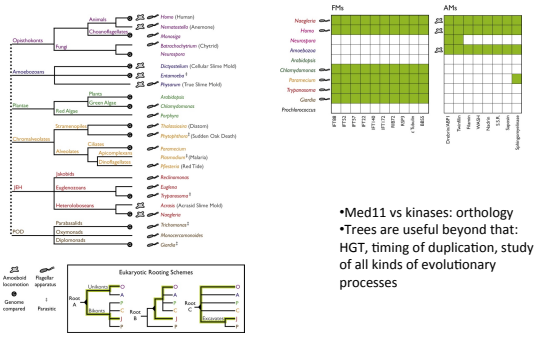
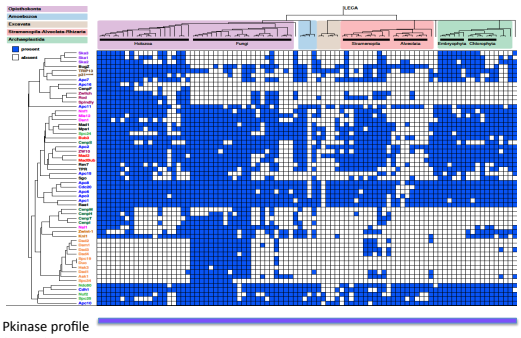
Phylogenetic trees & orthology

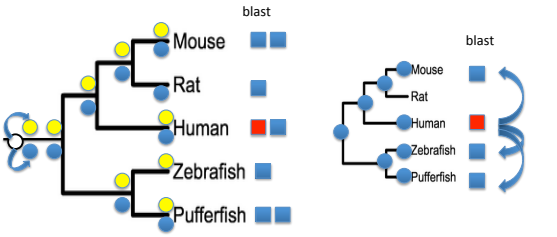


Fritz-Laylin et al. cell 2010

•Med11 vs kinases: orthology
•Trees are useful beyond that: HGT, timing of duplication, study of all kinds of evolutionary processes



Pkinase profile from pfam

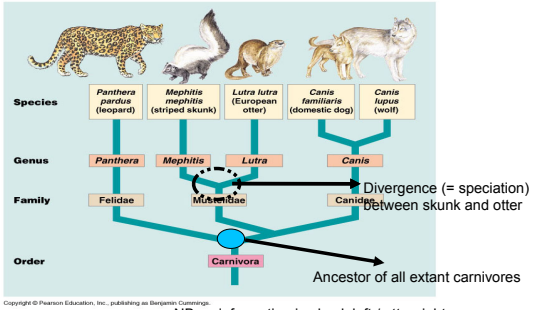


Gene originates before common ancestor. Duplicates. … evolves normally (has decent length e.g. 200AA and globular fold). Few losses.

Gene originates in common ancestor … evolves normally (has decent length e.g. 200AA and globular fold). Few losses.

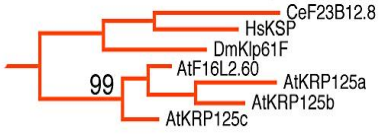
Problem: distinguishing paralogs from orthologs

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



NB no information in skunk left / otter right

Duplications in addition to speciations simple example (kinase)

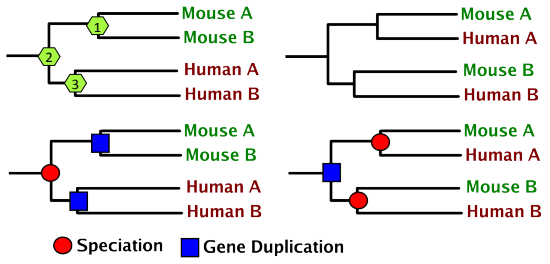


What are the nodes?

Two genes per species: how to differentiate between one ancient or two recent duplications?

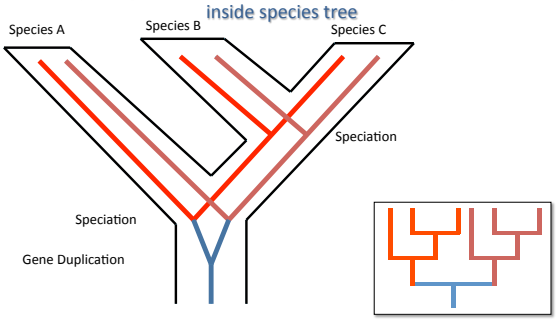
- Two genes in Human chromosomes (human A & Human B) & two genes in mouse chromosomes (Mouse A & Mouse B)

Duplications, Speciations

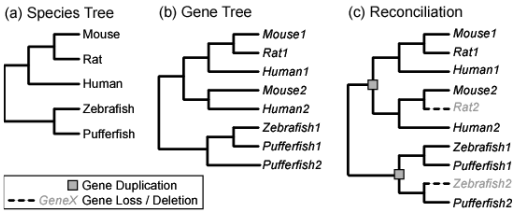


?

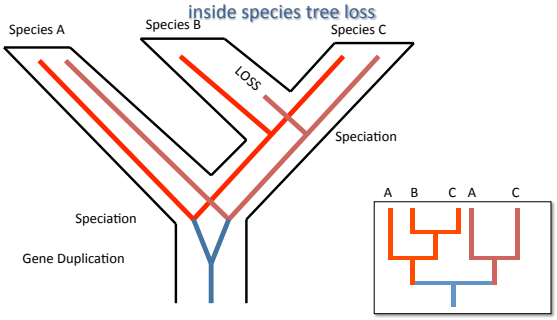
Interpreting the tree: duplications vs speciations, gene tree inside species tree



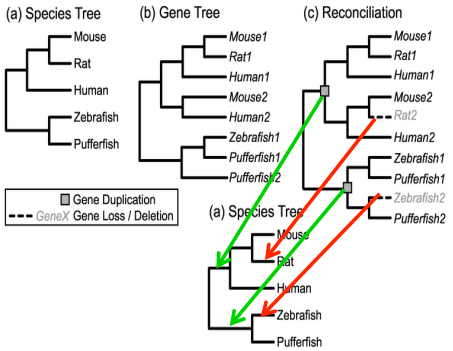
Interpreting the tree: gene trees vs species trees



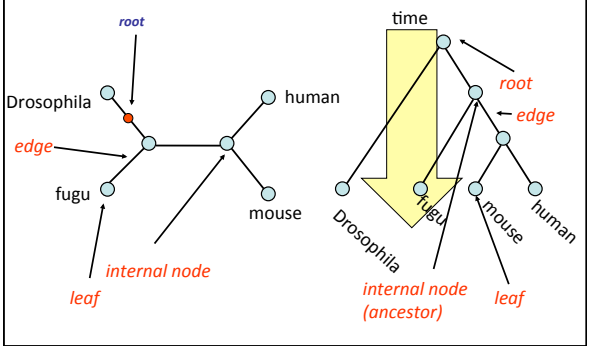
Interpreting the tree: duplications vs speciations, gene tree inside species tree loss



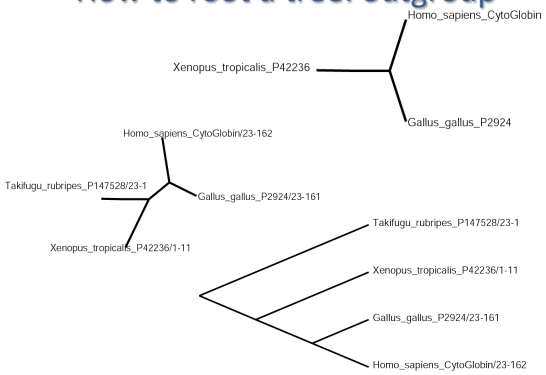
Tree reconciliation gives gene family dynamics on a species tree



Introduce a root

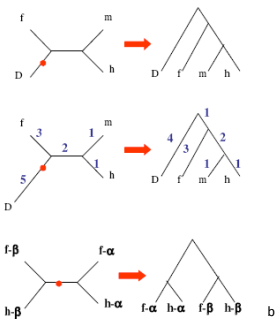


How to root a tree: outgroup

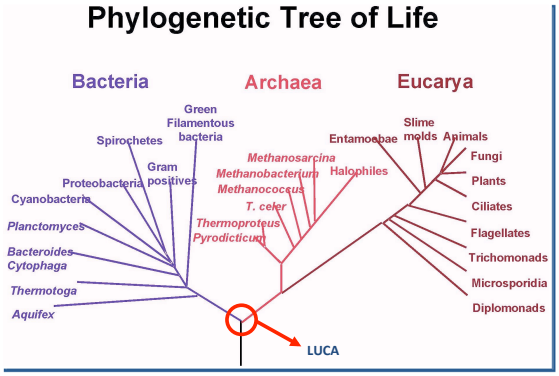


Interpreting the tree

- **Outgroup**. place root between distant homologouss sequence and rest group (b)
- **Midpoint.** place root at midpoint of longest path (sum of branches between any two leafs) NB njplot
- **Gene duplication**. Place root between paralogous gene copies (b)
- NB all affected by rates !

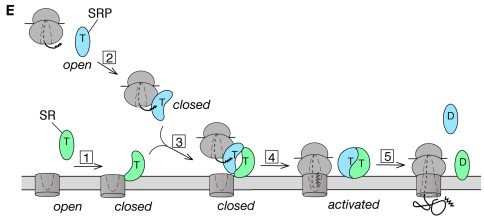


Phylogenetic Tree of Life



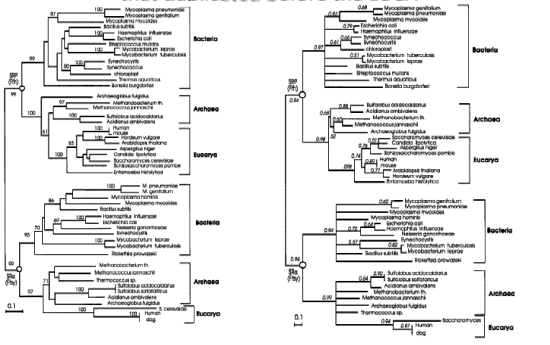
"three kingdoms"

How to root the tree of life?
1: Find paralogs that duplicated before the LUCA

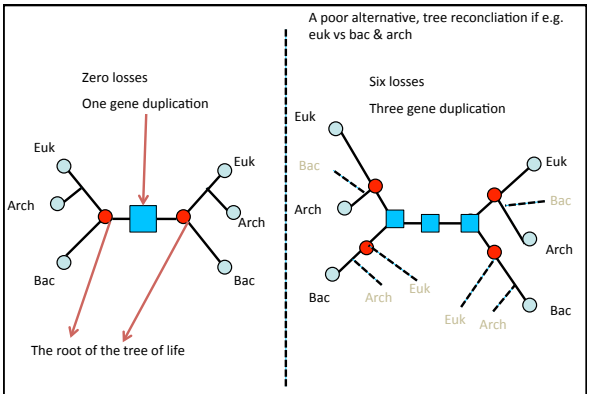


6 found so far

How to root the tree of life? 2: Make a tree of paralogs that duplicated before the LUCA

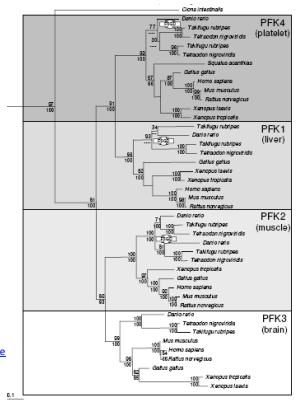


Griblado 1998 J Mol Evol

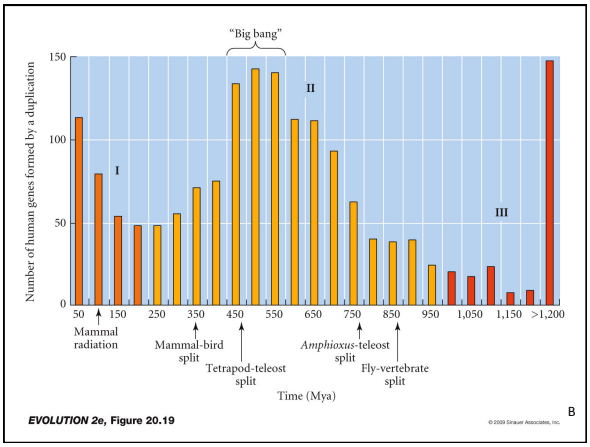


Interpreting the tree Example: vertebrate duplications

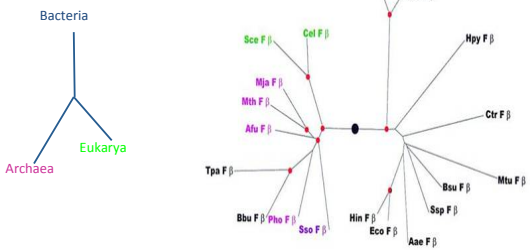
- Tetraploidy?



Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates.
Steinke D, **Hoegg** S, **Brinkmann** H, Meyer A.
BMC Biol. 2006 Jun 6;4:16.



Interpreting the tree: Horizontal Gene Transfer (HGT)



HGT between eukaryotes

Proc Natl Acad Sci U S A. 2011 Sep 13;108(37):15258-63. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Richards TA, Soanes DM, Jones MD, Vasieva O, Leonard G, Paszkiewicz K, Foster PG, Hall N, Talbot NJ.

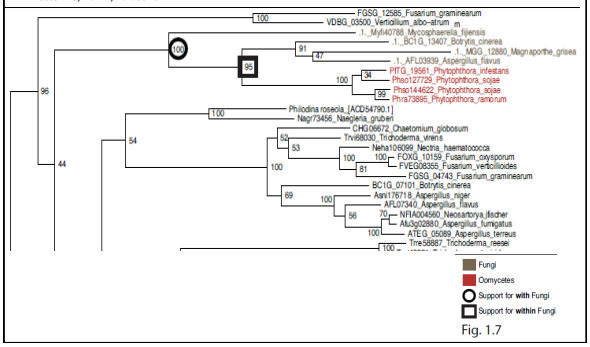
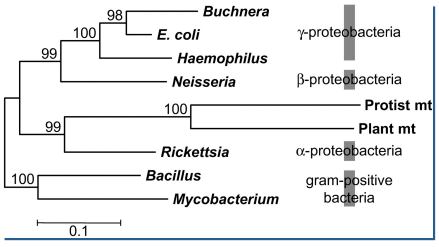


Fig. 1.7

Also endosymbiosis …



- More detail in Separate lecture …

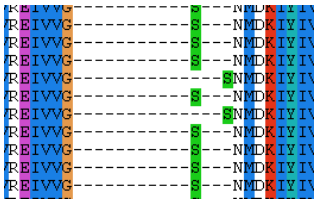
So annotating gene tree can give

- Timing of duplications & gene loss
- Horizontal gene transfer
- History of endosymbiosis
- A root to a tree in the absence of an outgroup sequence / species

Phylogenetic gene trees: how to make them

- Homology: *are* two pieces of sequence related; Trees: ±when did they diverge (= *how* are they related)
- Start from a multiple sequence alignment
- All multiple sequence programs alignments make a global alignment, thus feed it regions that you know are homologous → Domains !
- MUSCLE / clustal / t_coffee /**MAFFT** / clustalΩ
- Visual inspection of alignments (gaps, fragments/ complete sequences, weird things e.g. A)

Visual inspection of alignments: ?!



In practice

- Neighbor-joining for very big or very quick trees
- For reliable trees maximum likelihood and/or bayesian methods
- All these methods give unrooted trees

How to make a 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 (neighbor-joining, UPGMA, Fitch-Margoliash)

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

How to make a 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)

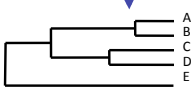
```
A a t g a c c c c g
B a t g a c c c c a
C a t g a c g t c t
D a t g a c g c g t
E t t g t t c a a t
```

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

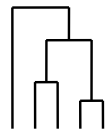
?

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 3 3 0
D 3 3 2 0
E 6 6 6 6 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

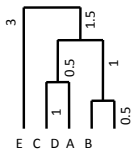
Termination:
- When only two clusters i, j remain, put root at $d_{ij}/2$

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

```
AB 0
 C 3 0
 D 3 2 0
 E 6 6 6 0
  AB C D E
```

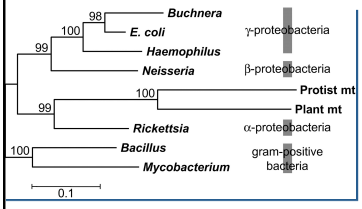
```
AB 0
CD 3 0
 E 6 6  0
  AB CD E
```

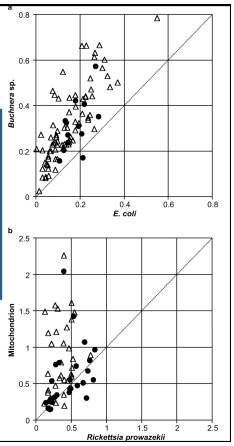
```
ABCD 0
   E 6   0
    ABCD E
```



Termination

Unequal rates between species are a very real phenomenon





Thus no molecular clock, thus phylogeny methods only infer relations, thus resulting trees are unrooted

- Neighbor-joining can deal with these kind of trees but I will only explain this if time permits …

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 approaches

Multiple sequence alignment

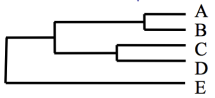
```
A  a t g a c c c c g
B  a t g a c c c c a
C  a t g a c g t c t
D  a t g a c g c g t
E  t t g t t c a a t
```

Model:" sequence evolution happens like this"

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

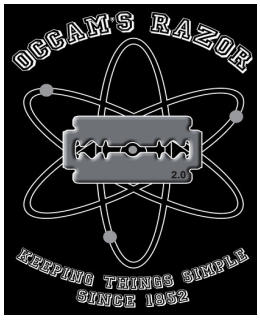
Find best fit to evolutionary model

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

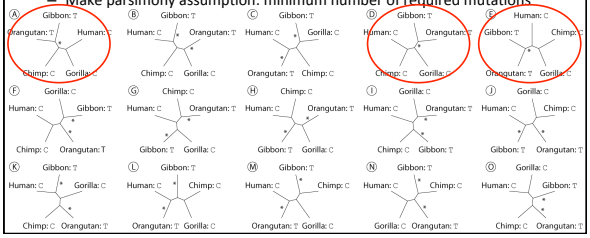


Maximum parsimony (MP)

- 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



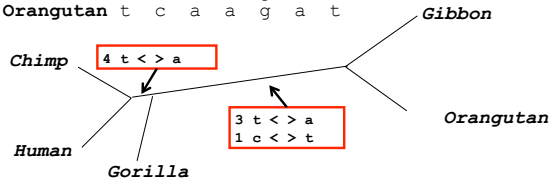
Maximum parsimony (MP)

- 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

Dollo parsimony

- The Dollo parsimony method is based on the assumption that a complex character that has been lost during evolution of a particular lineage cannot be regained.
- Dollo parsimony is the method of choice for reconstructing evolution of the gene repertoire of eukaryotic organisms because although multiple, independent losses of a gene in different lineages are common, multiple gains of the same gene are improbable

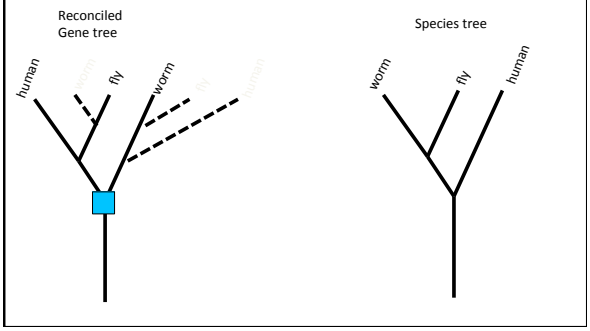
http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199297306.001.0001/acprof-9780199297306-chapter-11

Annotating gene tree can give all kinds of incredibly cool things but in reality gene trees are very noisy

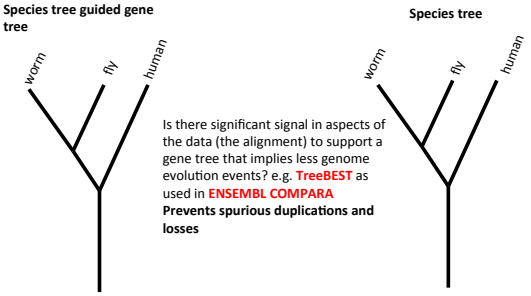
- For inferring the Tree of Life -> concatenated alignments of hundreds of genes, which allows to filter the alignment for only well behaved positions and provides sufficient data to see the history through the noise
- BUT for gene trees: If strict reconciliation / annotation gene trees would give e.g. many spurious duplications, B

Inconsitent gene tree and species tree -> reconciliation needed

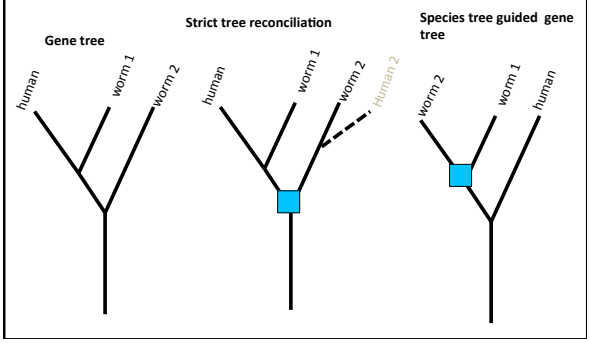
Strict tree reconciliation



…. Species tree guided tree building?



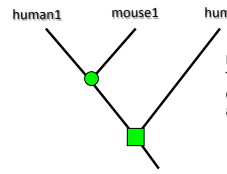
Example with gene duplicates in data



Gene Trees, Gene Duplications, and Orthology

- How to make trees
- Bootstrap
- Interpreting trees
- duplications vs speciations vs loss, timing of duplications, HGT
- Orthology

Jargon for interpretation: Orthology (and paralogy) as a specification of homology when discussing two species



Fitch 1970
Two genes in two species are orthologous if they derive from a single gene in their *last* common ancestor

"the corresponding gene"

"last common node is a speciation node"

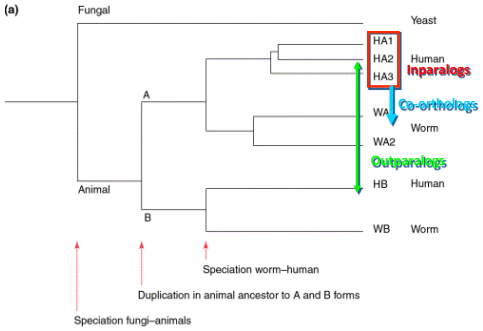
implied to have the same function

Orthology ~ annotating internal nodes as duplications or speciations

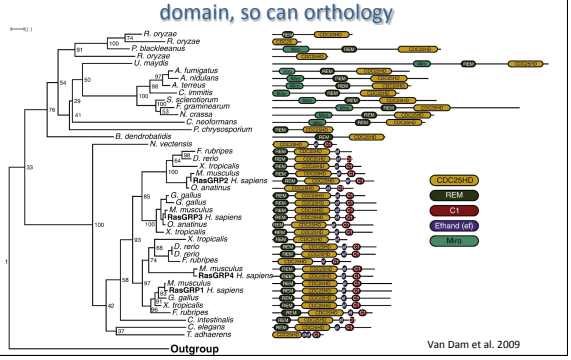
Genes can diverge by

- Speciation, or
- Duplication

Terminology: inparalogs, outparalogs, co-orthologs



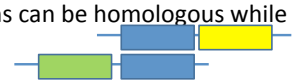
Orthologs can have different domain composition: (likely changed function); orthology is a specification of the homology relation and just like homology can span only a domain, so can orthology



Van Dam et al. 2009

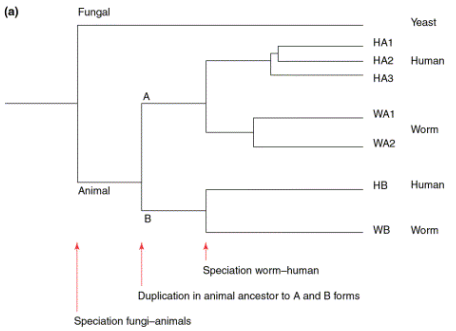
Gene / protein sequence evolution: what is homology

- Homologous residues = alignment
- Parts of proteins can be homologous while others are not



- Hence part of proteins can be orthologous while the rest is not

Depending on your comparison you want to have a reference speciation



Re-cap 1

- Sometimes sequence similarity is the bottle neck for finding orthologs e.g. med11, apc15???, spindly
  - Fulfill separated by speciation and bi-directional best hit criterion
  - are occasionally found via experiments rather than sequence

Recap 2: inparalogs. You have to deal with them.

- When comparing plant or plasmodium proteins to human or yeast proteins, plenty of time for duplications to make genes that are still co-orthologs. Such duplications are thus very frequent, also at shorter time scales (i.e. vertebrates vs invertebrates, flowering plants vs green algae).

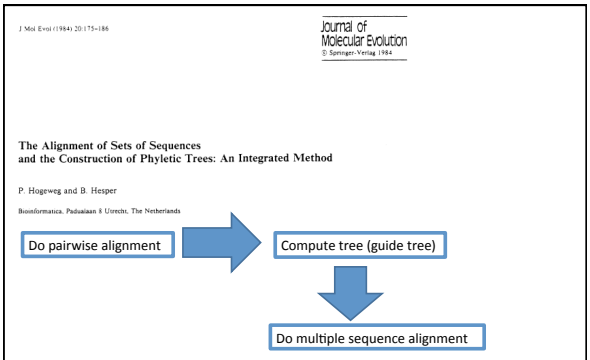
Molecular evolution is recursive / iterative 1

- To get a good alignment you need a good substitution matrix, to get a good substitution matrix you need a good alignment:
- "PAM matrix: were based on 1572 observed mutations in the phylogenetic trees of 71 families of closely related proteins. The proteins to be studied were selected on the basis of having high similarity with their predecessors. **The protein alignments included were required to display at least 85% identity**"
- For PAM_n $M_n = M_1^n$

http://en.wikipedia.org/wiki/Point_accepted_mutation

Molecular evolution is recursive / iterative 2

- To get a good tree you need a multiple sequence alignment but to get a good multiple sequence alignment you need a tree



As implemented in clustalx, MAFFT etc. !

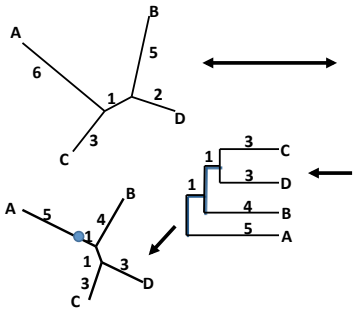
Molecular evolution is recursive / iterative 3: generalized

- To study the a evolution of a gene you need a model / framework of the evolution of the gene, but to get an idea of a proper framework / model of the evolution of a gene, you need the need to study the evolution of a gene
- Thus: heuristics & build on previous results. Start from stuff you trust (alignment of highly identical sequences), and/or only the use the general but flawed overview (e.g. guide tree). Then iterate
- Not yet so automatically solved for evolutionary history of a gene and its homologs as it is for other case …

Science is recursive / iterative

- To study something you need a model / framework of the thing, but to get an idea of a proper framework / model of the thing you need to study the thing
- Thus: heuristics & build on previous results. Start from stuff you trust, and/or only the use the general but flawed overview Then iterate
- This needs to happen in your brains

An additive tree which is wrongly reconstructed by UPGMA



	A	B	C	D
A	x	12	9	9
B	12	x	9	7
C	9	9	x	6
D	9	7	6	x

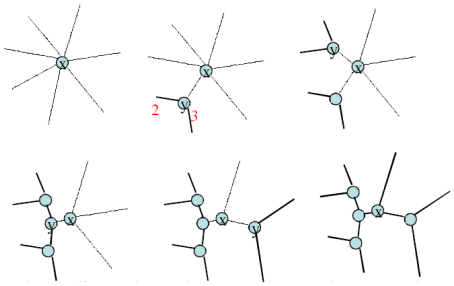
	A	B	CD
A	x	12	9
B	12	x	8
CD	9	8	x

	A	BCD
A	x	10
BCD	10	x

Neighbour-Joining (Saitou and Nei, 1987)

- Global measure. keeps total branch length minimal
- At each step, join two nodes such that distances are minimal (criterion of minimal evolution)
- Leads to unrooted tree

Neighbour-Joining



At each step all possible "neighbour joinings" are checked and the one corresponding to the minimal total tree length (calculated by adding all branch lengths) is taken.

Neighbour-Joining

	A	B	C	D	r
A	x	12	9	9	30
B	12	x	9	7	28
C	9	9	x	6	24
D	9	7	6	x	22

$M_{ab} = d_{ab} - (r_a + r_b)/(N-2)$

$Mab = 12 - (30+28)/(4-2)) = -17$

r= net divergence

	A	B	C	D
A	x	-17	-18	-17
B		x	-17	-18
C			x	-17
D				x

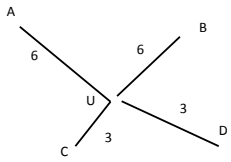
$AC \rightarrow U$

$d_{au} = d_{ao}/2 + (r_a - r_c)/(2(N-2))$
$= 9/2 + (30-24)/(2*2) = 6$

$d_{cu} = d_{ac} - d_{au} = 9 - 6 = 3$

$d_{bu} = (d_{ab} + d_{bc} - d_{ac}) / 2 = (12 + 9 - 9) / 2 = 6$

$d_{du} = (d_{ad} + d_{cd} - d_{ac}) / 2 = (9 + 6 - 9) / 2 = 3$



	U	B	D	r
U	x	6	3	9
B	6	x	7	13
D	3	7	x	10

	U	B	D
U	x	-16	-16
B		x	-16
D			x

e.g. UB →V

$D_{vu} = d_{ub} / 2 + (r_u - r_b) / (2(N-2)) = 6/2 + (9-13)/(2*1) = 3 - 2 = 1$

$D_{vb} = d_{ub} - d_{uv} = 6 - 1 = 5$

$D_{dv} = (d_{ud} + d_{bd} - d_{ub})/2 = (3+7-6)/2 = 2$