Bioinformatics

Phylogenetic trees

David Gilbert

Bioinformatics Research Centre

www.brc.dcs.gla.ac.uk

Department of Computing Science, University of Glasgow

Overview

- Phylogentics
- Trees
  - Definitions
  - Properties
- Molecular clock[s]
- Tree [re]construction
- Distance methods
  - UPGMA
  - Neighbour-joining
- Character methods
  - Maximum parsimony
  - Maximum likelihood
- Assessing trees
  - Bootstrapping trees

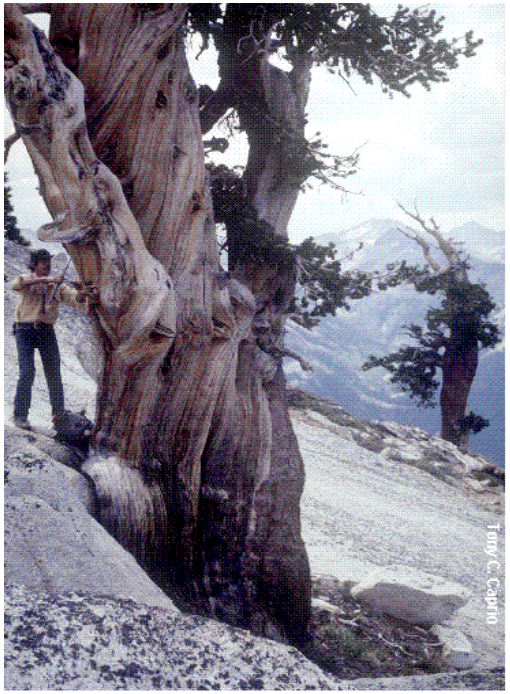
- Mostly based on Chapter 7 ‘Building Phylogenetic Trees’ from R. Durbin et al ‘Biological Sequence Analysis’ CUP 1988

Phylogenetic Trees

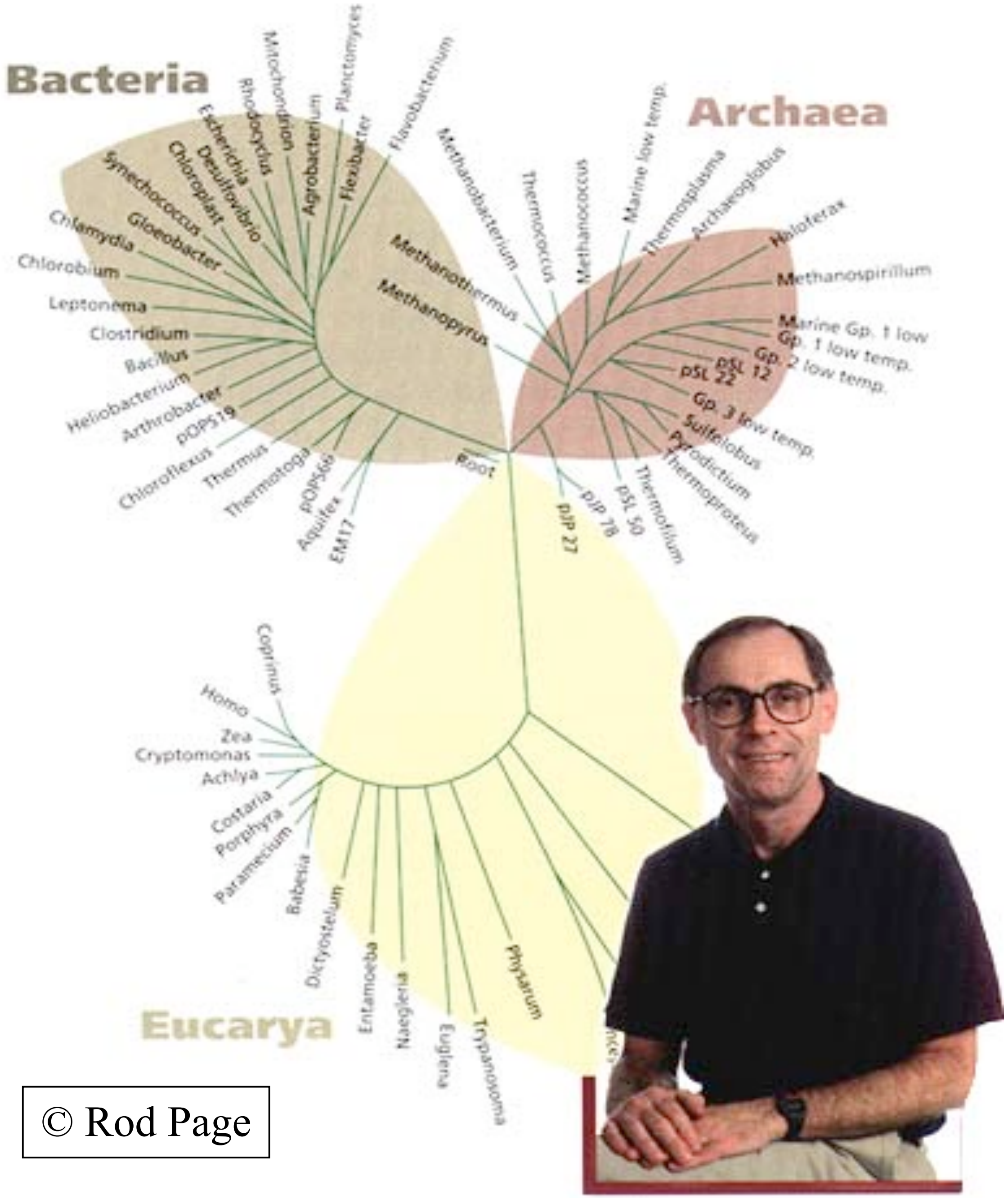
- Phylogeny : *The evolutionary history and line of descent of a species*
- Phylogenetic Tree : A diagram setting out the genealogy of a species
- Purpose
  - to reconstruct the correct genealogical ties between related objects
  - To estimate the time of divergence between them since they last shared a common ancestor
- Objects typically are protein or nucleic acid sequences

What is phylogeny good for?

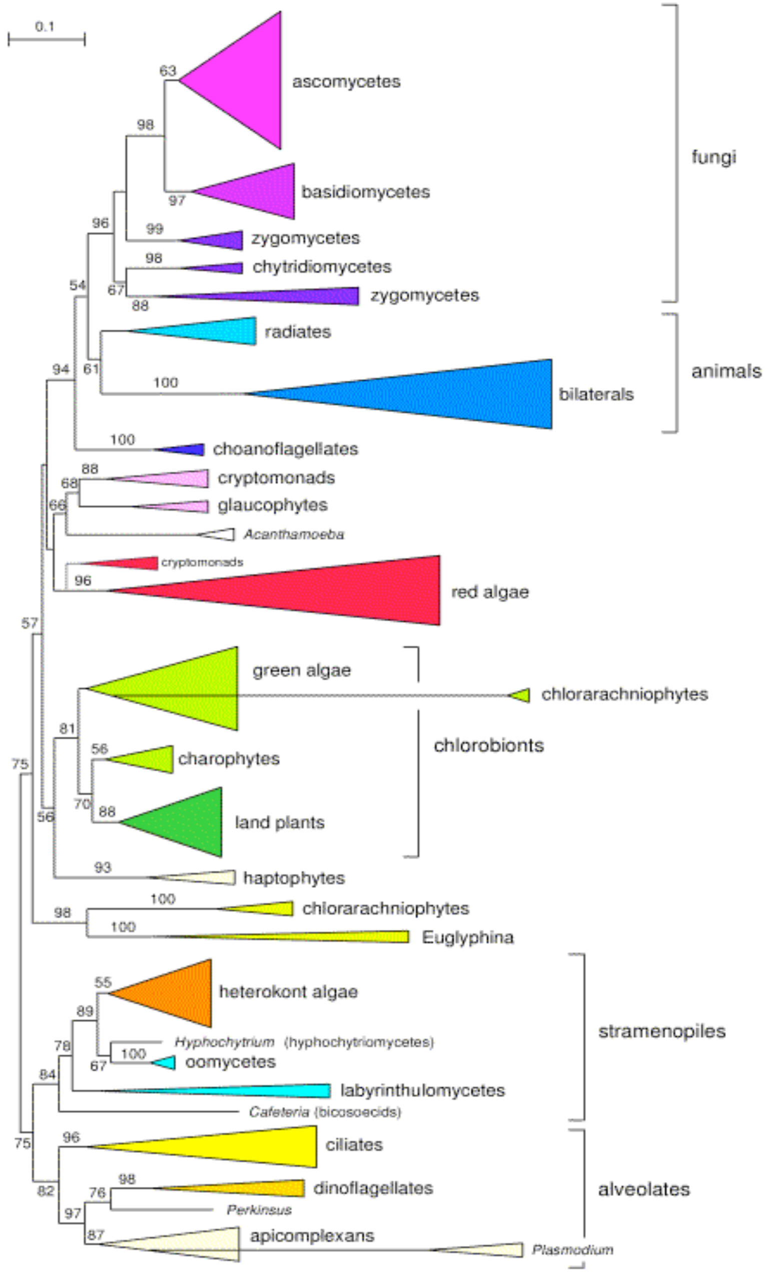
- Evolutionary history ("tree of life")
- Population history
- Rates of evolutionary change
- Origins of diseases
- Prediction of sequence function
- Can be applied to organisms, sequences, viruses, languages, etc.



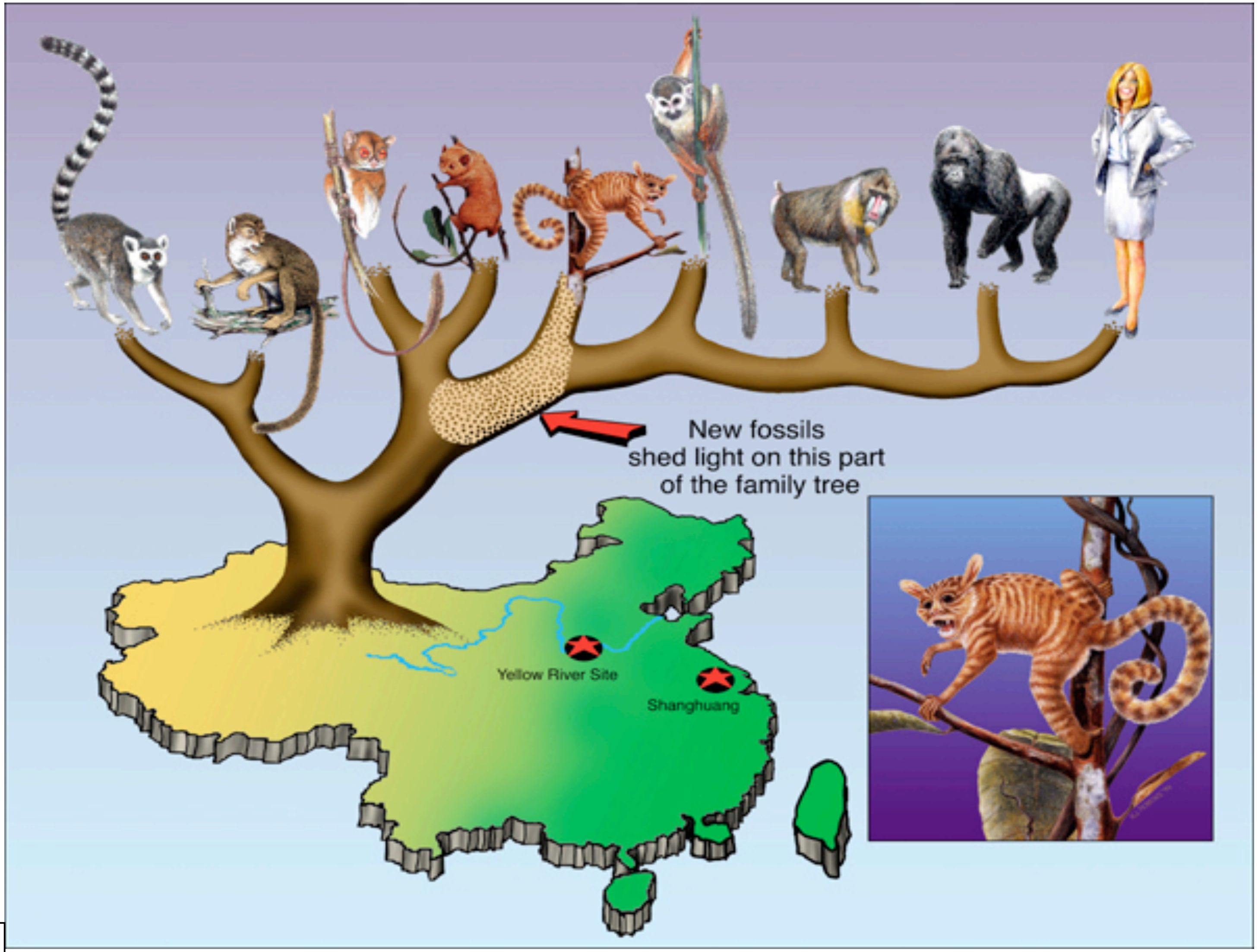
© Rod Page


Bacteria
Archaea
Eucarya
Mitochondrion
Planctomyces
Rhodocyclus
Agrobacterium
Flavobacterium
Escherichia
Desulfovibrio
Flexibacter
Chloroplast
Synechococcus
Gloeobacter
Chlamydia
Chlorobium
Leptonema
Clostridium
Bacillus
Heliobacterium
Arthrobacter
pOPS19
Chloroflexus
Thermus
Thermotoga
pOPS66
Aquifex
EM17
Root
Methanobacterium
Thermococcus
Methanococcus
Marine low temp.
Thermoplasma
Archaeoglobus
Haloferax
Methanospirillum
Methanothermus
Methanopyrus
Marine Gp. 1 low temp.
Gp. 1 low temp.
Gp. 2 low temp.
pSL 22
pSL 12
Gp. 3 low temp.
Sulfolobus
Pyrodictium
Thermoproteus
Thermofilum
pSL 50
pJP 78
pJP 27
Coprinus
Homo
Zea
Cryptomonas
Achlya
Costaria
Porphyra
Paramecium
Babesia
Dictyostelium
Entamoeba
Naegleria
Euglena
Trypanosoma
Physarum

© Rod Page

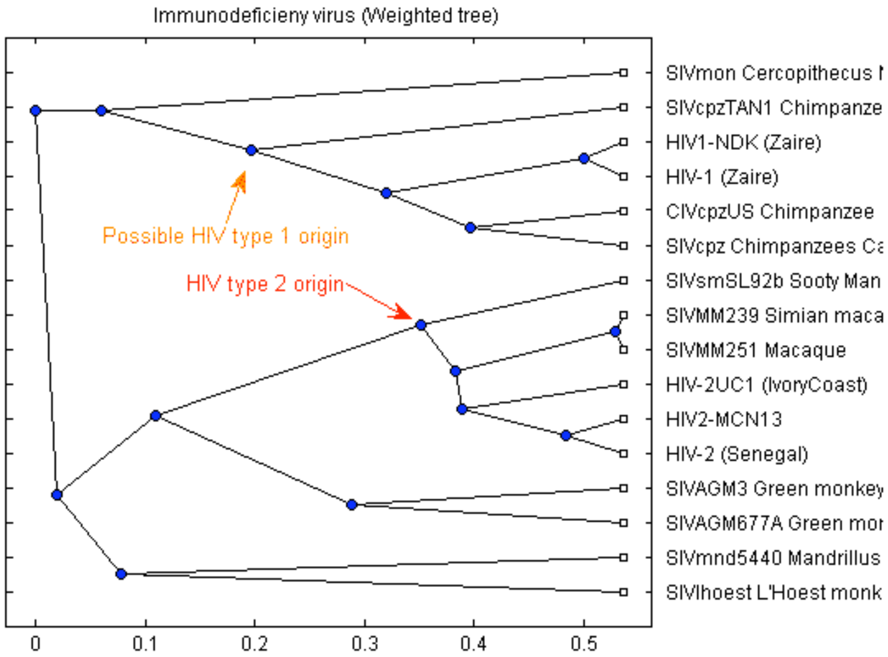

0.1
ascomycetes
basidiomycetes
zygomycetes
chytridiomycetes
zygomycetes
fungi
radiates
bilaterals
animals
choanoflagellates
cryptomonads
glaucophytes
Acanthamoeba
cryptomonads
red algae
green algae
chlorarachniophytes
chlorobionts
charophytes
land plants
haptophytes
chlorarachniophytes
Euglyphina
heterokont algae
Hyphochytrium (hyphochytriomycetes)
oomycetes
labyrinthulomycetes
Cafeteria (bicosoecids)
stramenopiles
ciliates
dinoflagellates
Perkinsus
apicomplexans
Plasmodium
alveolates

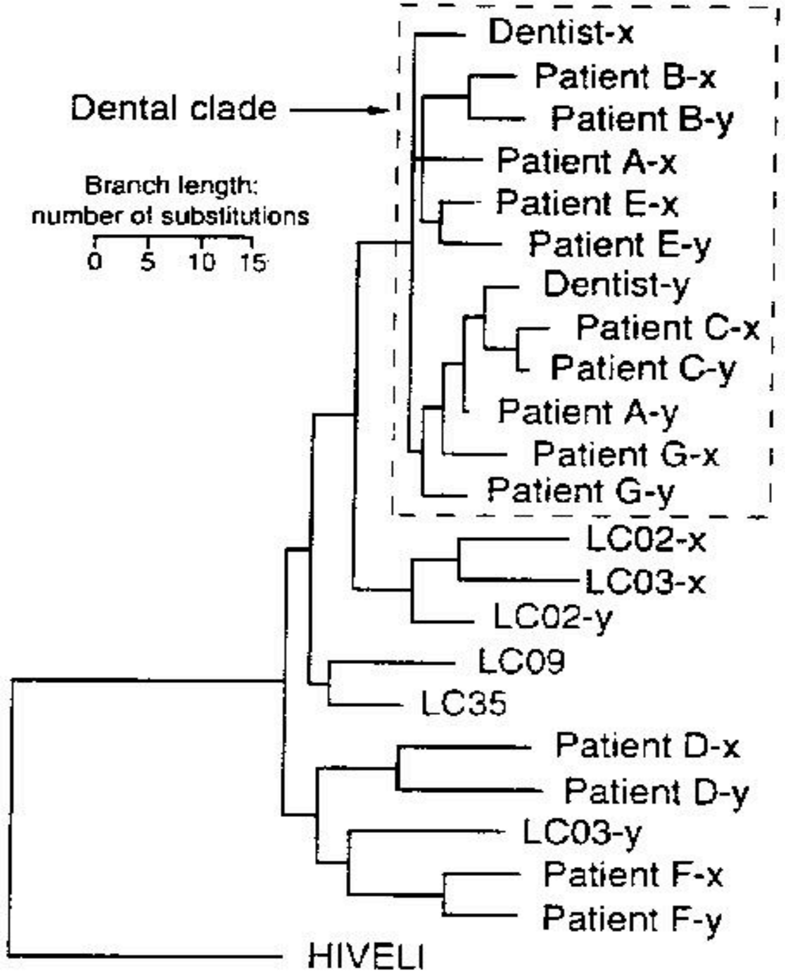
Our place in Nature



© Rod Page

HIV: where did it come from and how is it transmitted?



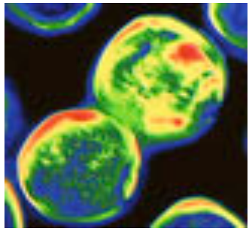


Yersinia
pestis

Arabidopsis
thaliana

Buchnerasp.
APS

Aquifex
aeolicus

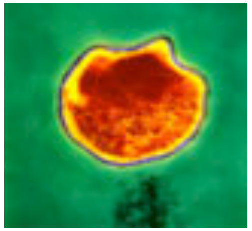

Archaeoglobus
fulgidus

Borrelia
burgorferi

Mycobacterium
tuberculosis

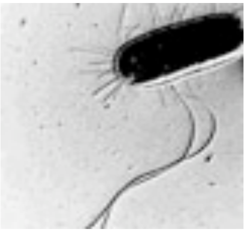
Caenorhabitis
elegans

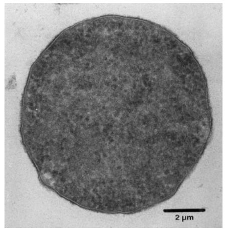
Campylobacter
jejuni

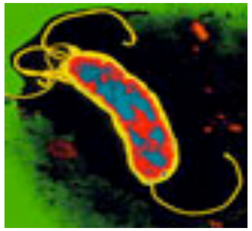

Chlamydia
pneumoniae

Vibrio
cholerae

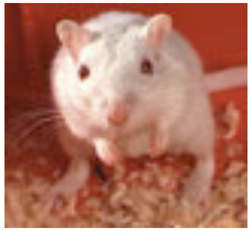
Drosophila
melanogaster

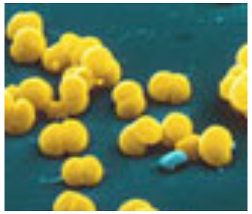

Escherichia
coli

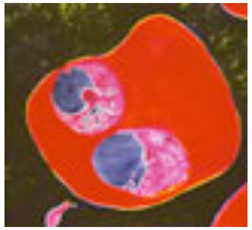

Thermoplasma
acidophilum


Helicobacter
pylori

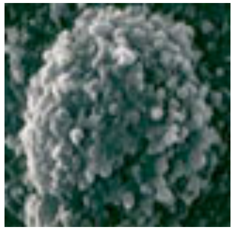
Mycobacterium
leprae


mouse

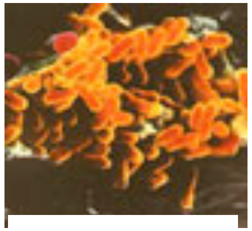

Neisseria
meningitidis
Z2491

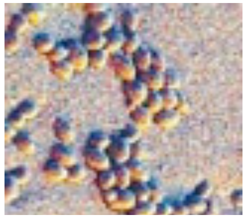

Plasmodium
falciparum

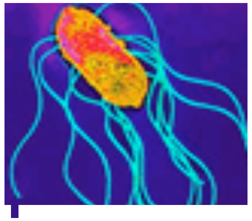
Pseudomonas
aeruginosa

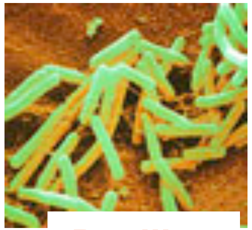

Ureaplasma
urealyticum

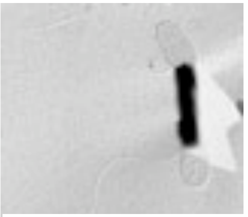
rat

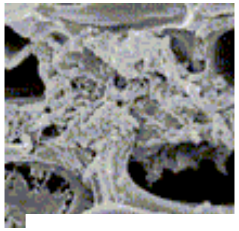

Rickettsia
prowazekii


Saccharomyces
cerevisiae

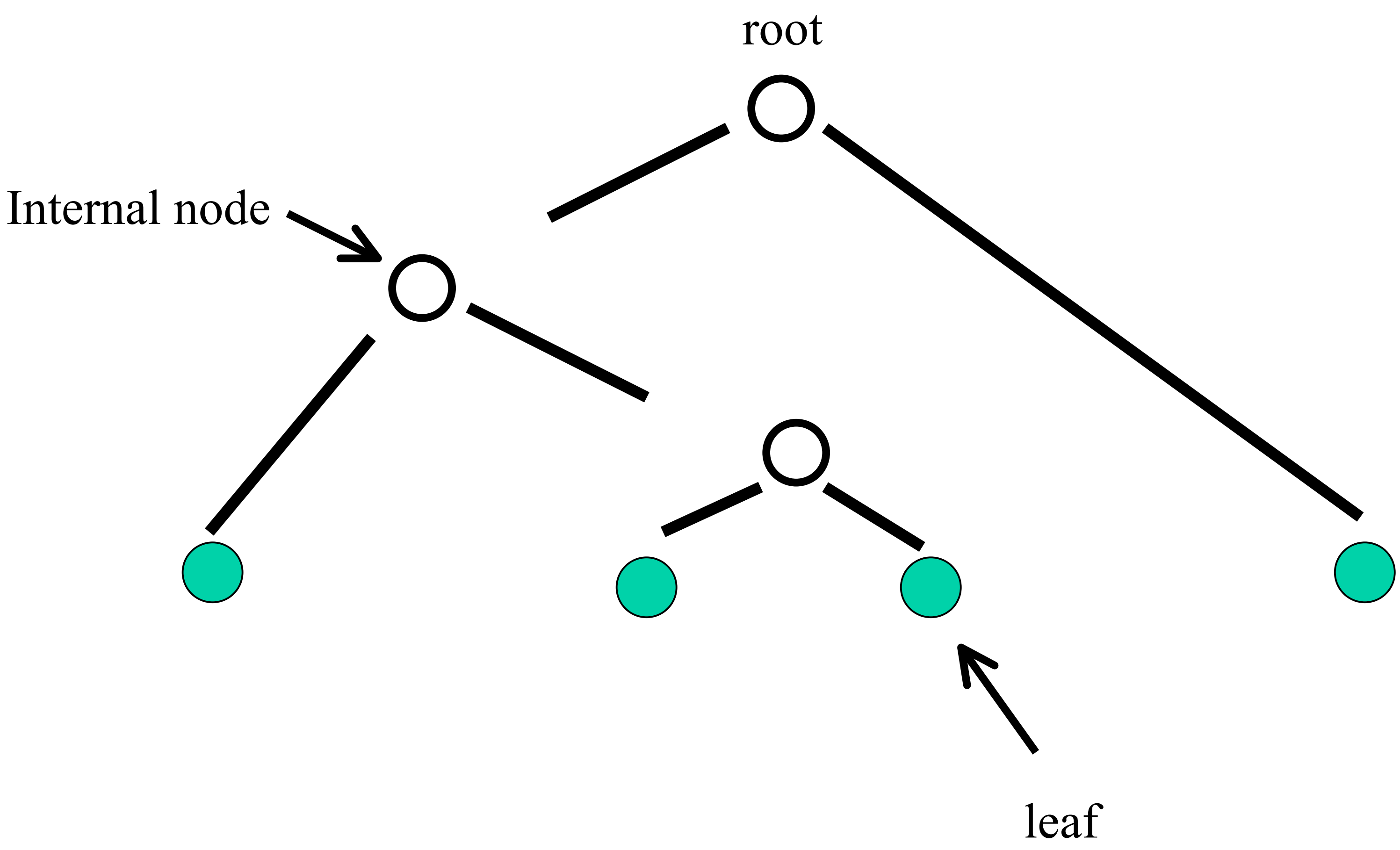

Salmonella
enterica


Bacillus
subtilis


Thermotoga
maritima

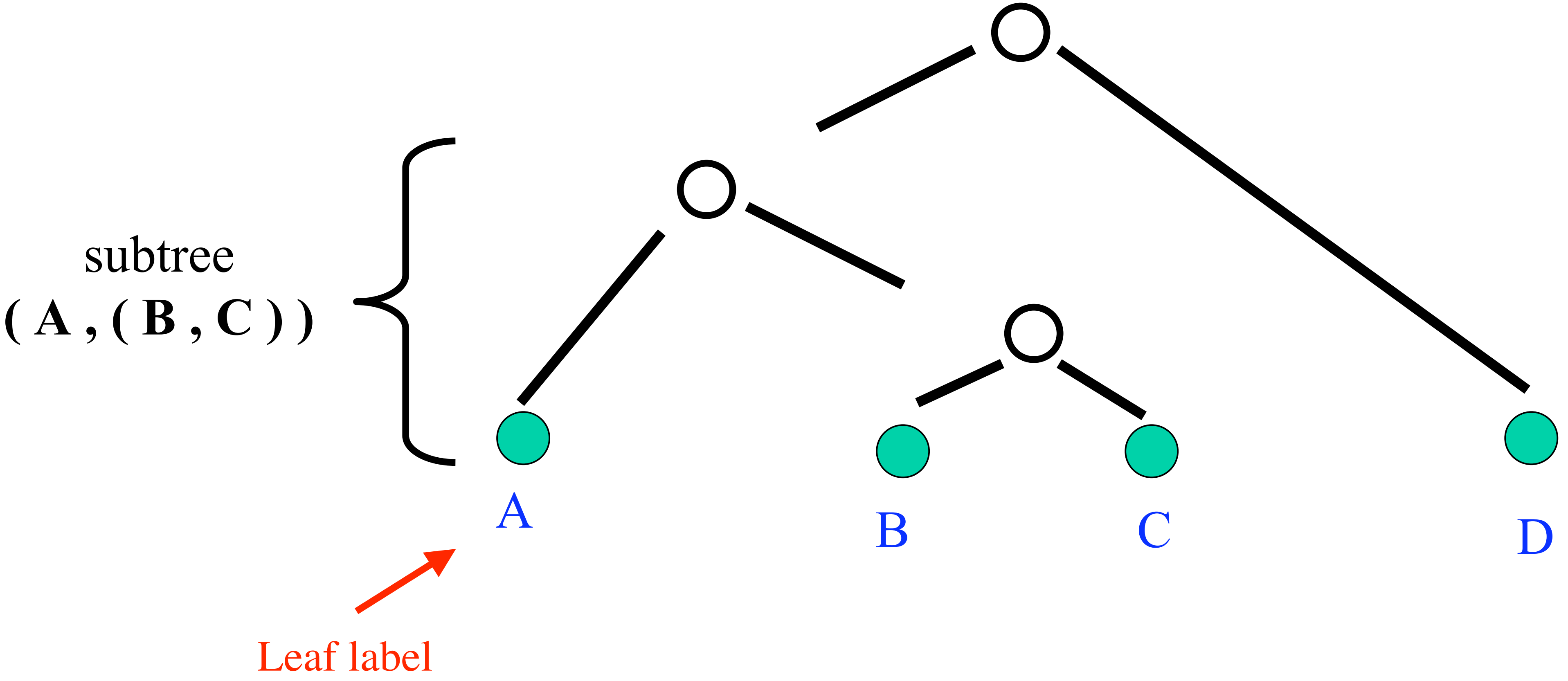

Xylella
fastidiosa

Tree terminology



Tree as text

((A , (B , C)) , D)



Definition of a tree

- A tree is either

  A **leaf**

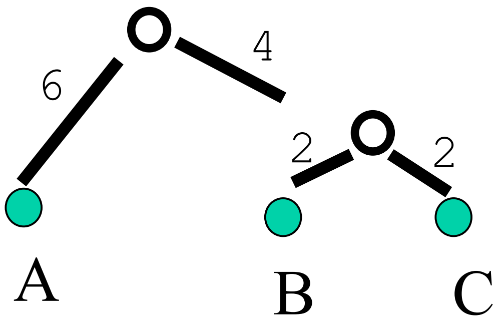
  or

  (LeftTree , RightTree)

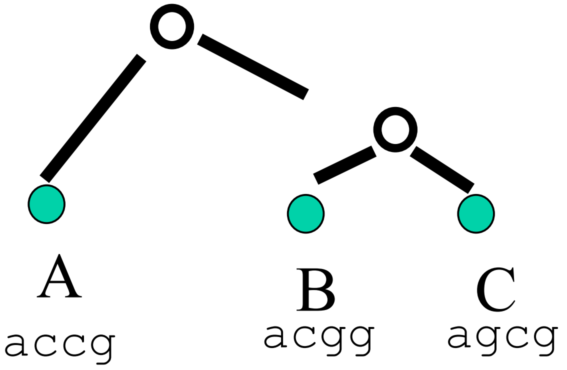
  where both LeftTree and RightTree are trees

Trees with labels

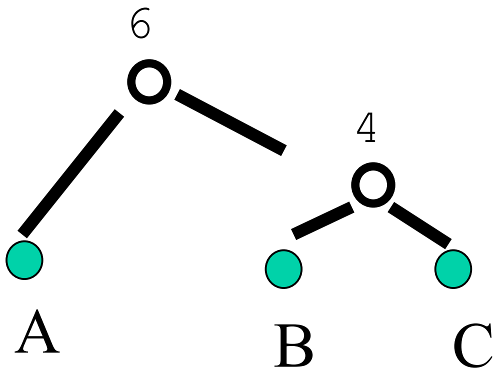
- We can add data to
  - – Leaves
  - – Branches
  - – Nodes



(A:6 , (B : 2 , C : 2):4)



(A[accg**] , (B [**acgg**] , C [**agcg**]))**



6:(A , 4:(B, C))

Isomorphic trees

((A , (B , C)) , D)

((A , (C , B)) , D)

Etc…

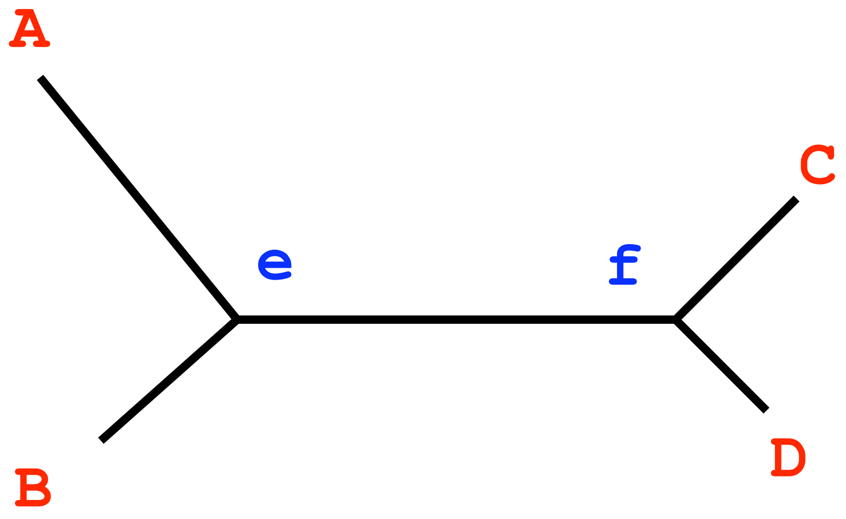
Derive [all] the isomorphic trees and draw them!

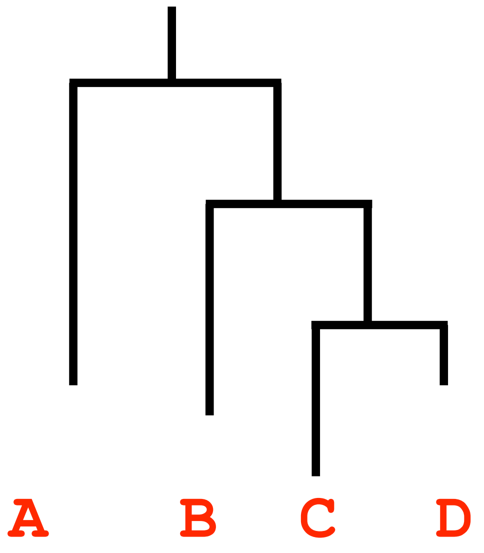
Isomorphic trees

- Isotrees(T1,T2) /Two trees T1 and T2 are *isomorphic/* if

```
   T1 is a leaf and T2 is a leaf and T1= =T2
 OR
   T1 = (L1,R1) , T2 = (L2,R2) AND
             Isotrees(L1, L2)  AND Isotrees(R1, R2)
             OR
             Isotrees(L1, R2)  AND Isotrees(R1, L2)
```

Trees can be unrooted





Are there alternative rootings?
Draw them…

Evolution - basic concepts

- Mutation in DNA a natural evolutionary process

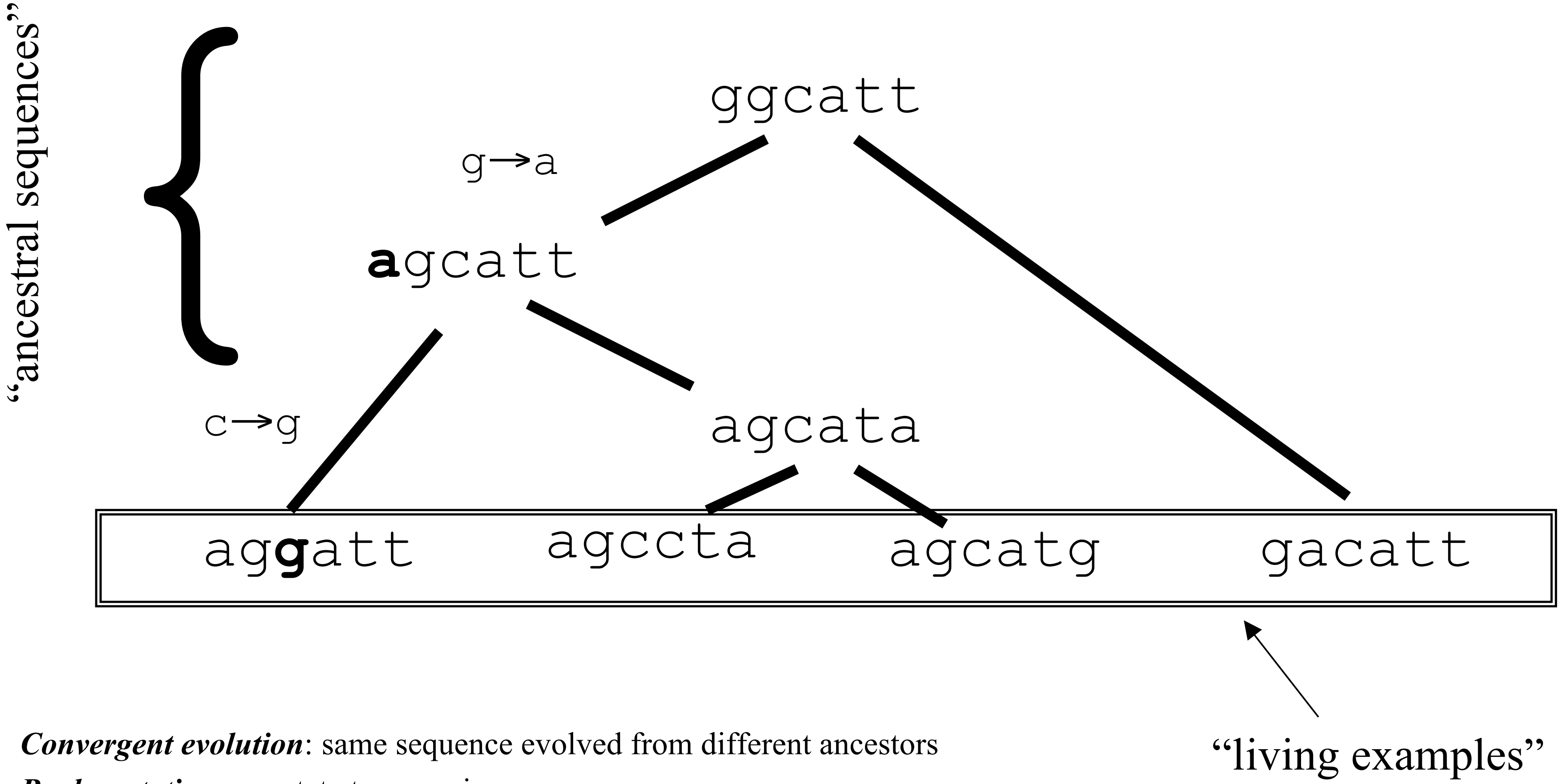
- DNA *replication* errors: (nucleotide)
  - substitutions
  - insertions
  - deletions

} **indels**

- Similarity between sequences
  - clue to common evolutionary origin, or
  - clue to common function

- This is a simplistic story: in fact the altered *function* of the expressed protein will determine if the organism will survive to reproduce, and hence pass on [transmit] the altered gene

Evolution - related sequences



Convergent evolution: same sequence evolved from different ancestors

Back mutations - mutate to a previous sequence

Definitions

- **Phylum** (phyla pl): A primary division of a kingdom, as of the animal kingdom, ranking next above a class in size.
- **Phylogeny**:
  - the sequence of events involved in the evolutionary development of a species or taxonomic group of organisms
  - Evolutionary relationships within and between taxonomic levels, particularly the patterns of lines of descent. Phylogenetics -The taxonomical classification of organisms based on their degree of evolutionary relatedness. Phylogenetic tree - A variety of dendrogram (diagram) in which organisms are shown arranged on branches that link them according to their relatedness and evolutionary descent.
- **Phylogentics**: study of evolutionary relationships
- **Phylogenetic analysis**: the means of *inferring* or estimating these relationships
- **Taxonomy**: The science of naming and classifying organisms 1. [n] practice of classifying plants and animals according to their presumed natural relationships
- 2. [n] (biology) study of the general principles of scientific classification
- 3. [n] a classification of organisms into groups based on similarities of structure or origin etc
- **Taxon**: any named group of organisms
- **Species**: taxonomic group whose members can interbreed (but more or less able to…)

Definitions (cont)

- **Clade**: a group of biological taxa or species that share features inherited from a common ancestor; A monophyletic taxon; a group of organisms which includes the most recent common ancestor of all of its members and all of the descendants of that most recent common ancestor. From the Greek word "klados", meaning branch or twig.

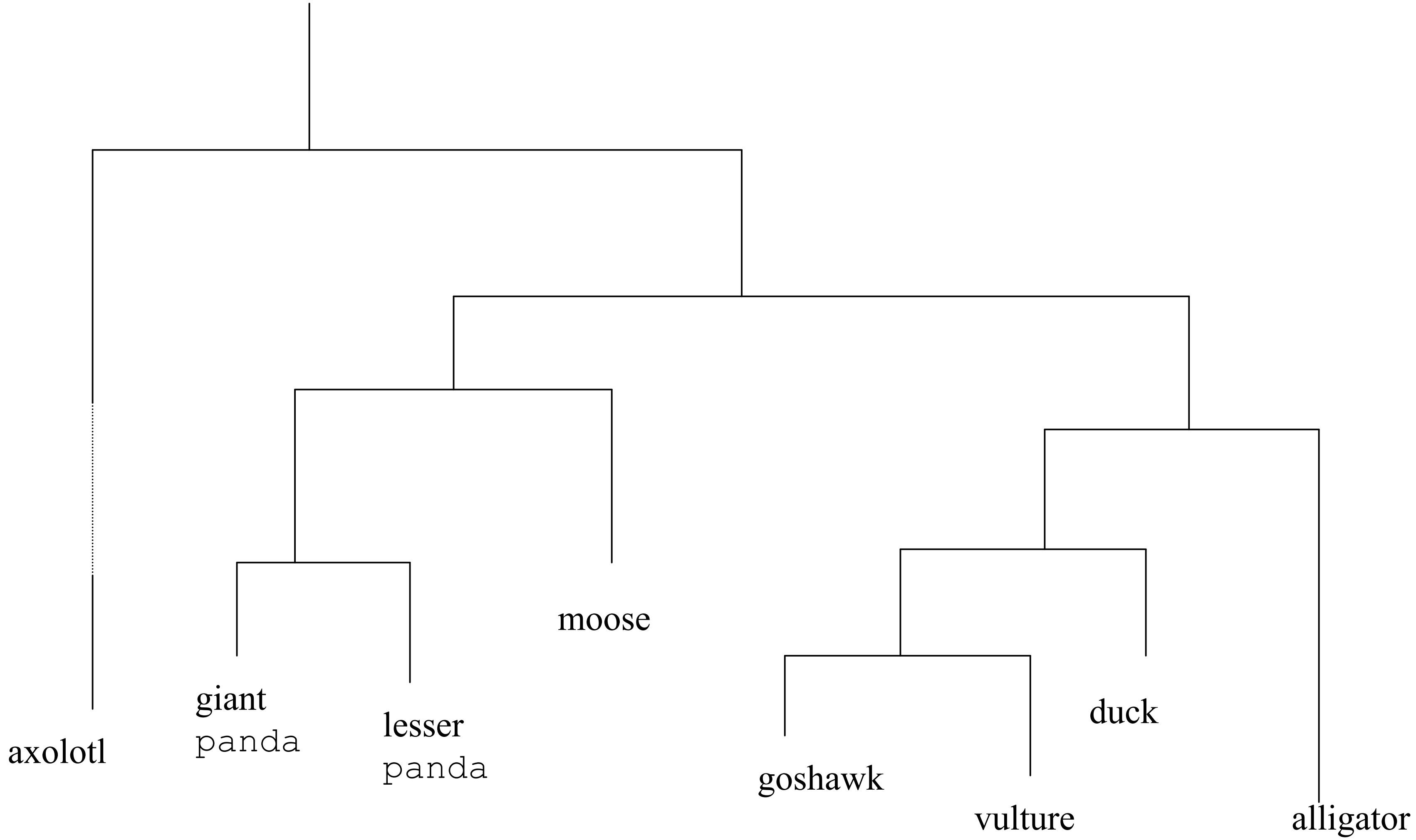
- **Tree** : A data structure consisting of nodes which may contain other nodes via its branches. Unlike a tree in nature, the root node is usually represented at the top of the structure and does not have a parent node. All other nodes have a single parent. Nodes having no child nodes are called leaf nodes

- **Dendrogram**: Any branching diagram (or tree) (cf. cladogram, phylogram, phenogram); A dendrogram is a 'tree-like' diagram that summaries the process of clustering. Similar cases are joined by links whose position in the diagram is determined by the level of similarity between the cases. A treelike figure used to represent graphically a hierarchy. (dendron - greek ‘tree’)

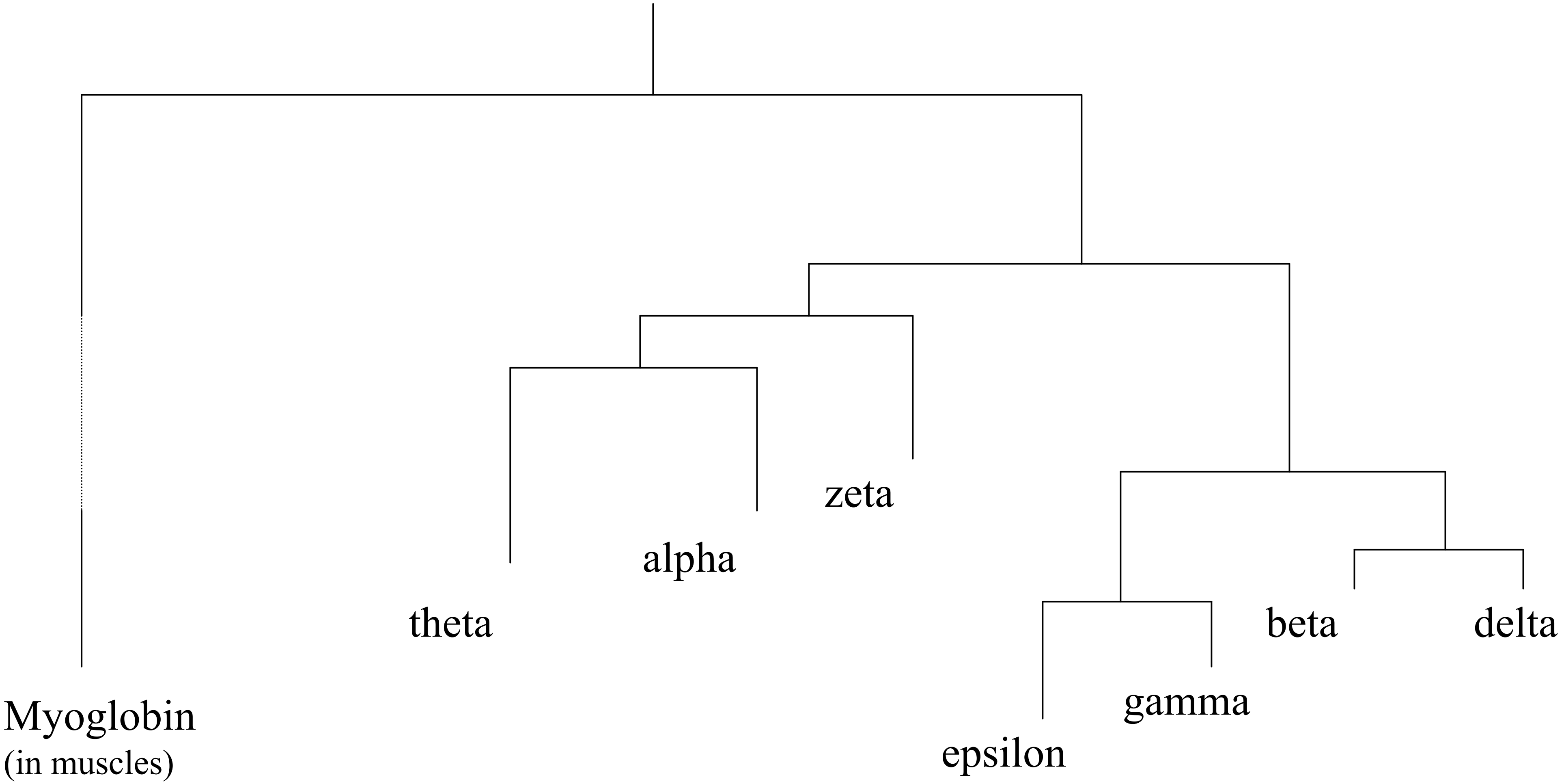
Homologues

- **Homologues**: sequences that have common origins but may or may not share common activity
- **Orthologues**: homologues produced by speciation. Tend to have similar function
- **Paralogues**: homologues produced by gene duplication. Tend to have different functions
- **Xenologues**: homologues resulting from horizontal gene transfer between 2 organisms.

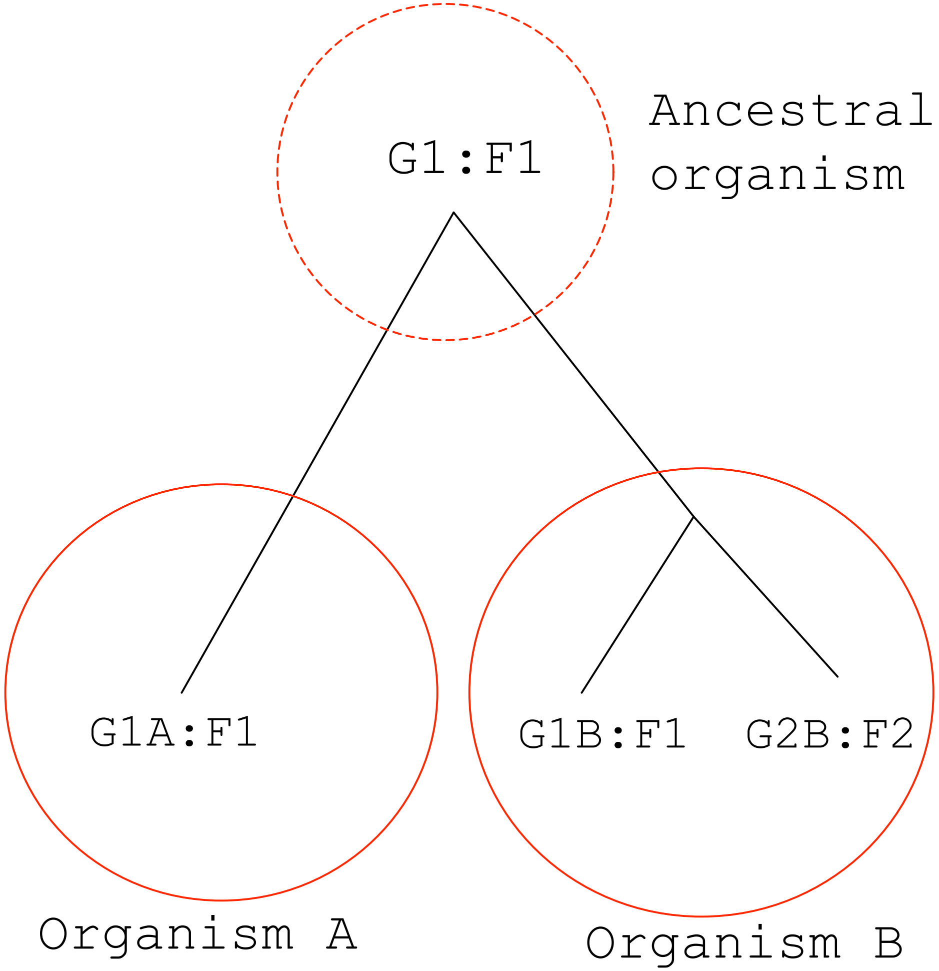
Tree of orthologues based on a set of α-haemoglobins



Tree of paralogues: human haemoglobins and myoglobin



Orthologues & paralogues



Orthologues
G1A:F1 & G1B:F1

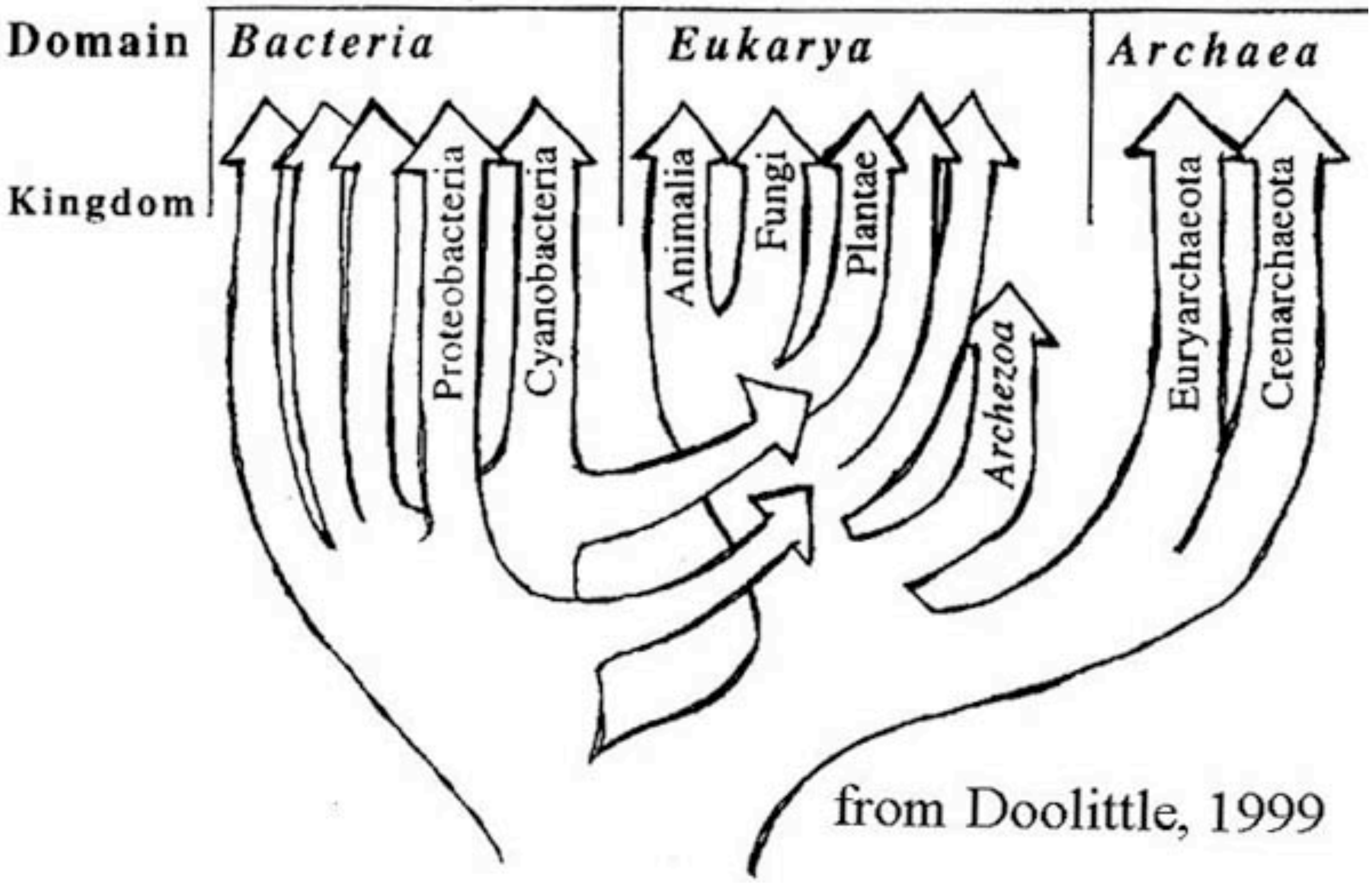
Paralogues
G1B:F1 & G2B:F2

Horizontal (lateral) gene transfer!

Bacteria can take up and spread genes in different ways

- Uptake of naked DNA directly from surroundings
- Obtain genes from infecting viruses
- Take up genes through cross-species mating

Horizontal gene transfer



from Doolittle, 1999

Phylogenetic Trees

Approaches to reconstructing phylogenetic trees

- Distance based methods
- Maximum parsimony methods
- Maximum likelihood methods

Phylogenetic analysis - 4 steps

1. Alignment / distance computation
2. Determine the data model
3. Tree building
4. Tree evaluation

Tree building methods

1. **Distance-based**: Transform the data into pairwise distances (dissimilarities), and then use a matrix during tree building. E.g. data from from immunology, nucleic acid hybridization, and breeding experiments, are automatically expressed as pair-wise distances but most character data need to be mathematically transformed into distances.

2. **Character-based**: Use the aligned characters, such as DNA or protein sequences, directly during tree inference – based on substitutions.

Tree building algorithms

- Distance-based:
  - input = evolutionary distance data (sequence edit dist; melting temp DNA hybridisations, strength of antibody cross reactions..)
  - Goal: construct weighted tree whose pairwise distances agree with evolutionary distances
  - Ultrametric distance data - elegant solution: UPGMA
  - Additive data - efficient solution: NJ
  - FM - Fitch-Margoliash
  - ME - minimum evolution
  - Not additive - reconstruction not guaranteed; find trees whose distances “best approximate” given data
- Rooting trees

Metric, ultrametric

- A *metric* on a set of objects *O* is given by the assignment of a real number *d(x,y)* for every pair of objects *x,y* in *O* such that

  $d(x,y) > 0$ for $x \neq y$

  $d(x,y) = 0$ for $x=y$

  $d(x,y) = d(y,x)$ for all *x,y*

  $d(x,y) \leq d(x,z) + d(y,z)$ for all *x,y,z* (triangle inequality)

- In addition, for an *ultrametric*

  $d(x,y) \leq max(d(x,z) , d(y,z))$

- An *ultrametric tree* is characterised by the condition

  $d(x,y) \leq d(x,z) = d(y,z)$ (for any 3 points their pairwise distances are all equal, or 2 are equal and 1 is smaller)

Ages {Fred:20, Mary:22, Jane:24} e.g. add, subtract – are these metric / ultrametric / …?

How many pairwise comparisons?

	A	B	C	D	E
A					
B					
C					
D					
E					

- How many for whole table?
- *Identity:* $(x,y) = 0$ for $x=y$
- *Symmetry:* $d(x,y) = d(y,x)$ for all x,y

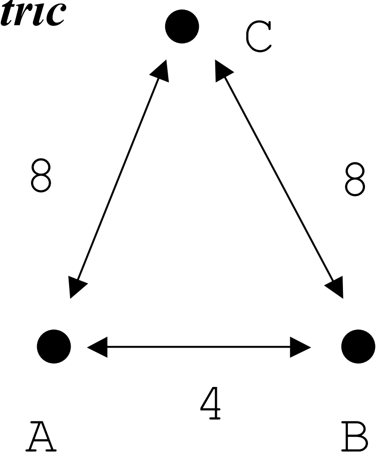
Metric, Ultrametric trees

?

metric

ultrametric

Distances



Distance matrix

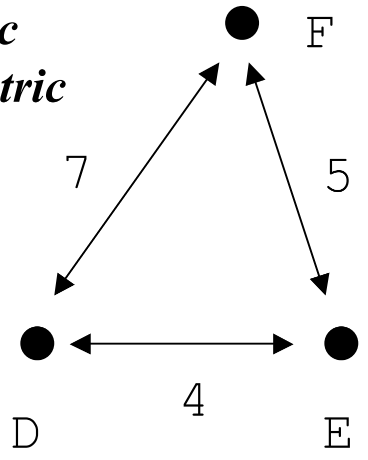
	A	B	C
A			
B			
C			

Tree

?

metric

ultrametric



	D	E	F
D			
E			
F			

Ultrametric trees

Def: Given D a symmetric matrix n by n of real numbers; an *ultrametric tree* for D is a rooted tree T with the following properties:

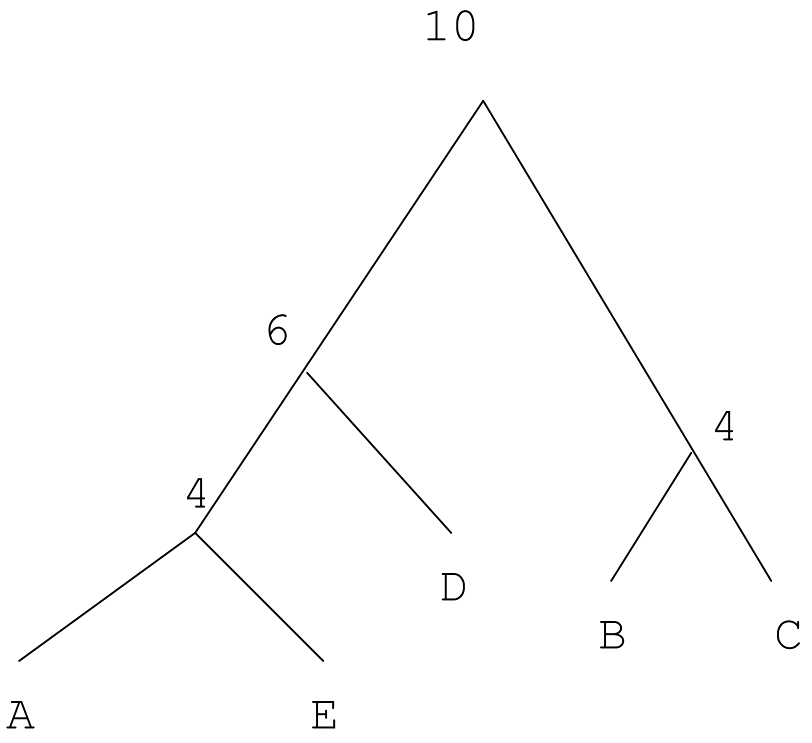
1. T contains n leaves, each labelled by a unique row D
2. Each internal node of T is labelled by one entry from D and has at least 2 children
3. Along any path from the root to the leaf, the numbers labelling the internal nodes *strictly decrease*
4. For any two leaves i,j of T, D(i,j) is the label of the least common ancestor of i and j in T

Def: A *min-ultrametric tree* for D is a rooted tree T with all the properties of an ultrametric tree except that (3) is changed to:

3. Along any path from the root to the leaf, the numbers labelling the internal nodes *strictly increase*

Symmetric matrix & Ultrametric tree

	A	B	C	D	E
A	0	10	10	6	4
B		0	4	10	10
C			0	10	10
D				0	6
E					0



Obtaining pairwise distances

- Melting temp DNA hybridisations,
- Strength of antibody cross reactions
- Protein structure comparison
- Sequences: edit distance and variations
  - d(i,j) : fraction *f* of sites *u* where residues x_{iu} x_{ju} differ (assume an alignment)
  - For 2 unrelated sequences : random substitutions cause *f* to approach fraction of differences expected by chance, hence:
  - Jukes-Cantor distance d(i,j) = -3/4 * log(1-4*f*/3)
    - Tends to ∞ as equilibrium value of *f* (75% residues differ) reached

Gene sequence distance programs

- Blast (!) [but ‘score’ *increases* with similarity…]
- Clustal[w] – fasta format, but outputs a tree
- Phylip – its own format – sequences must be of the same length…hence use clustalw to align the sequences first [output in Phylip format, no tree]

Clustering methods: UPGMA

Unweighted Pair Group Method using Arithmetic Averages

Define the distance d(i,j) between two clusters Ci and Cj to be the average distance between pairs of sequences from each cluster:

$$d(i,j) = \frac{1}{|C_i| * |C_j|} \sum_{p \in C_i, q \in C_j} d(p,q)$$

where |Ci| and |Cj| denote the number of sequences in clusters i and j respectively

Note that if Ck = Ci ∪Cj and Cl is any other cluster then

$$d(k,l) = \frac{d(i,l) * |C_i| + d(j,l) * |C_j|}{|C_i| + |C_j|}$$

UPGMA algorithm

Initialisation:

- Assign each sequence i to its own cluster Ci
- Define one leaf T for each sequence, and place it height zero

Iterate

- Determine the two clusters i,j for which d(i,j) [distance i-j] is minimal. (If there are more than two such pairs, pick one at random)
- Define a new cluster k by Ck = Ci ∪ Cj, and define d(k,l) for all l by

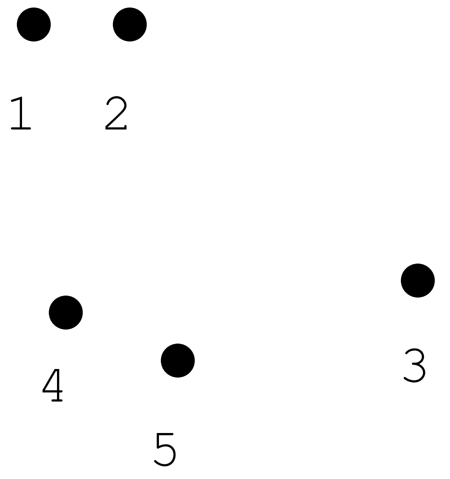
$$d(k,l) = \frac{d(i,l) * |C_i| + d(j,l) * |C_j|}{|C_i| + |C_j|}$$

- Define a node k with daughter nodes i and j, and place it at height d(i,j)/2
- Add k to the current clusters and remove i and j

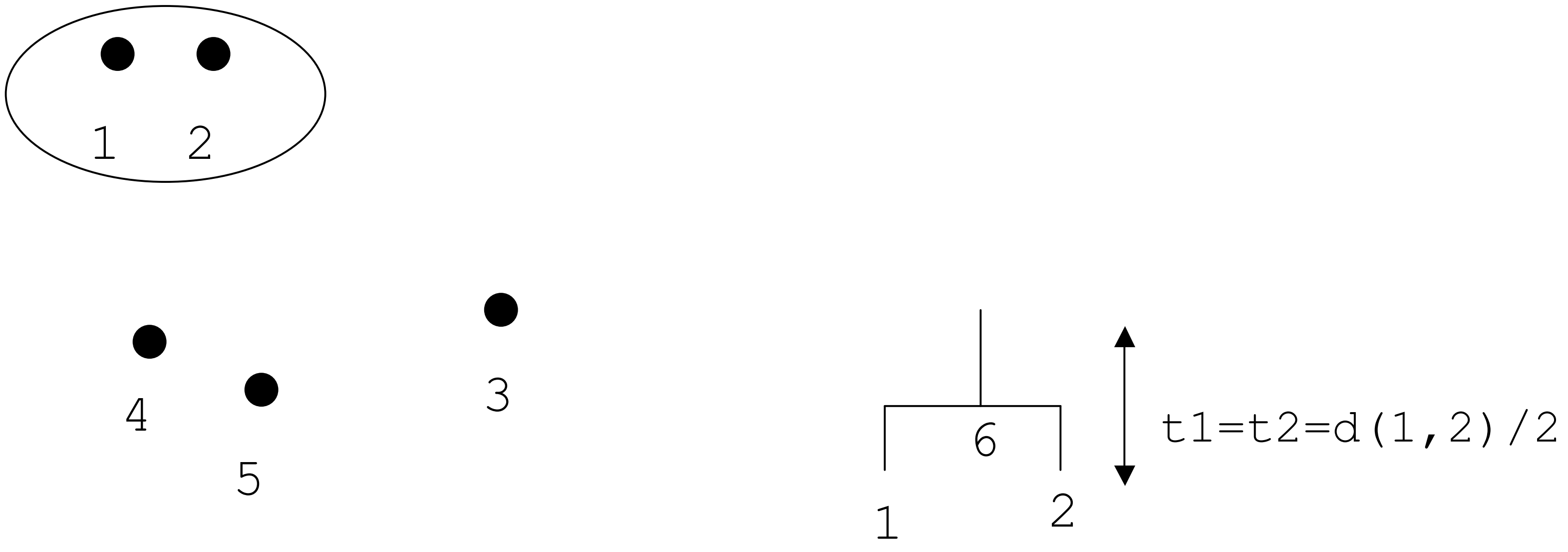
Termination

- When only two clusters i,j remain, place the root at height d(i,j)/2

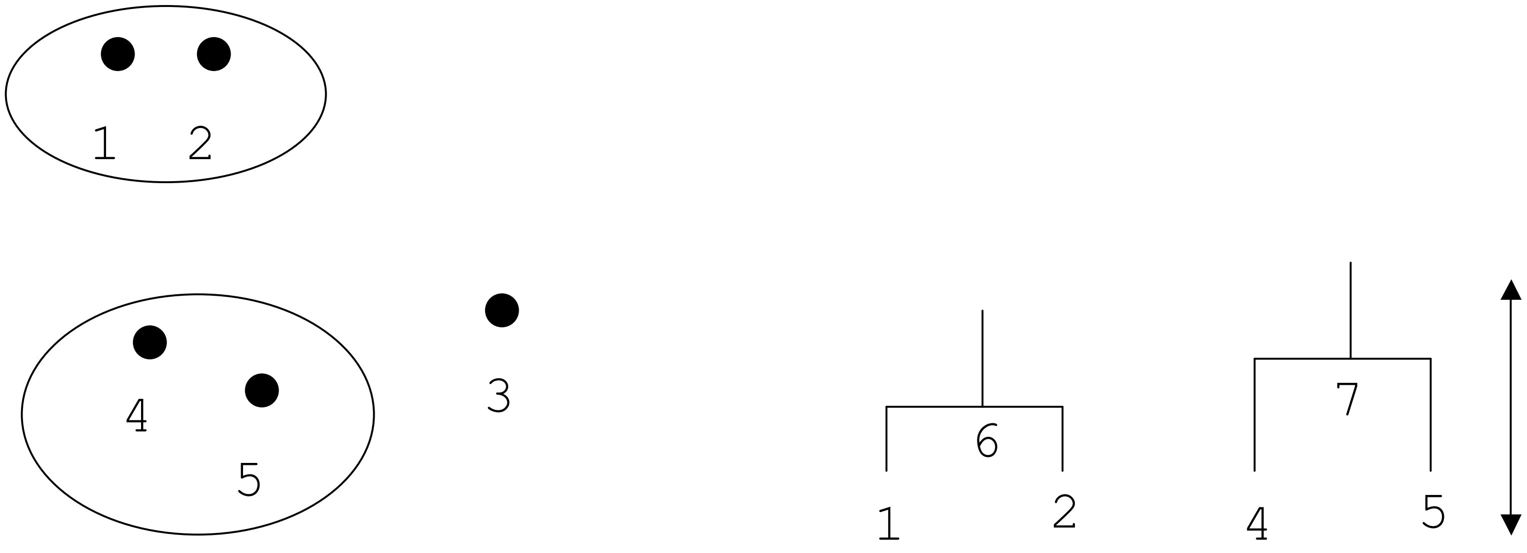
UPGMA example



UPGMA example

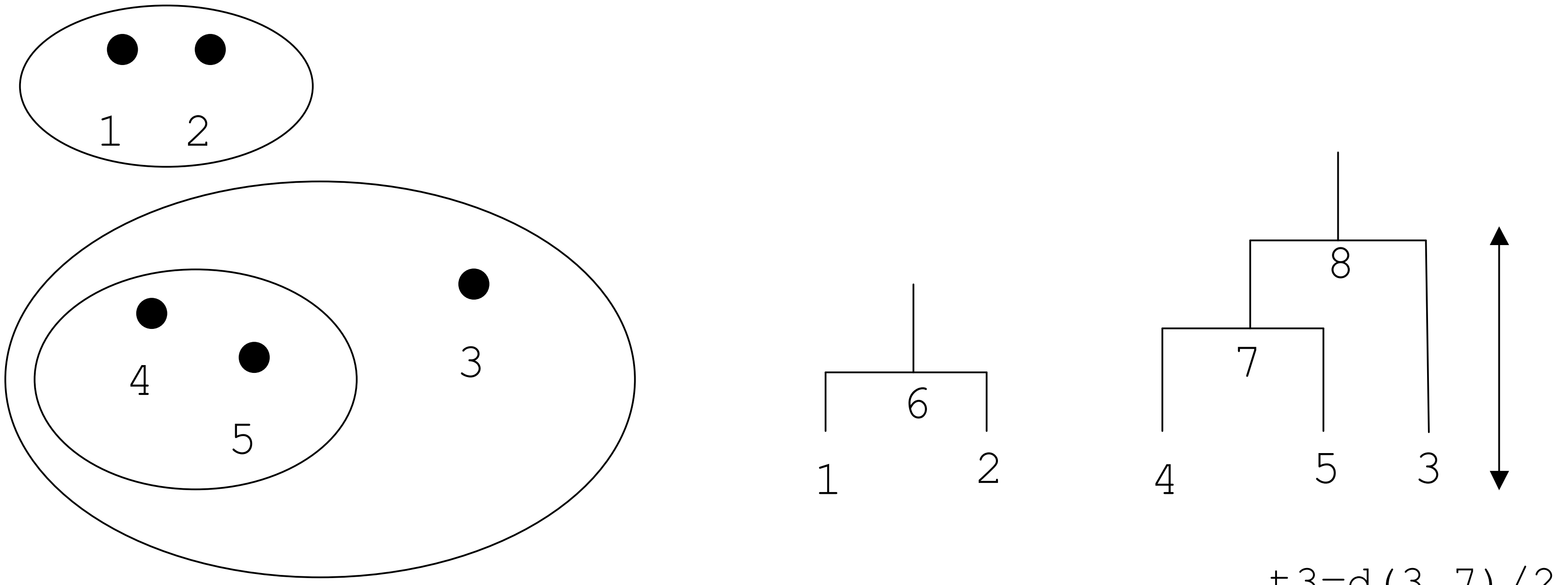


UPGMA example



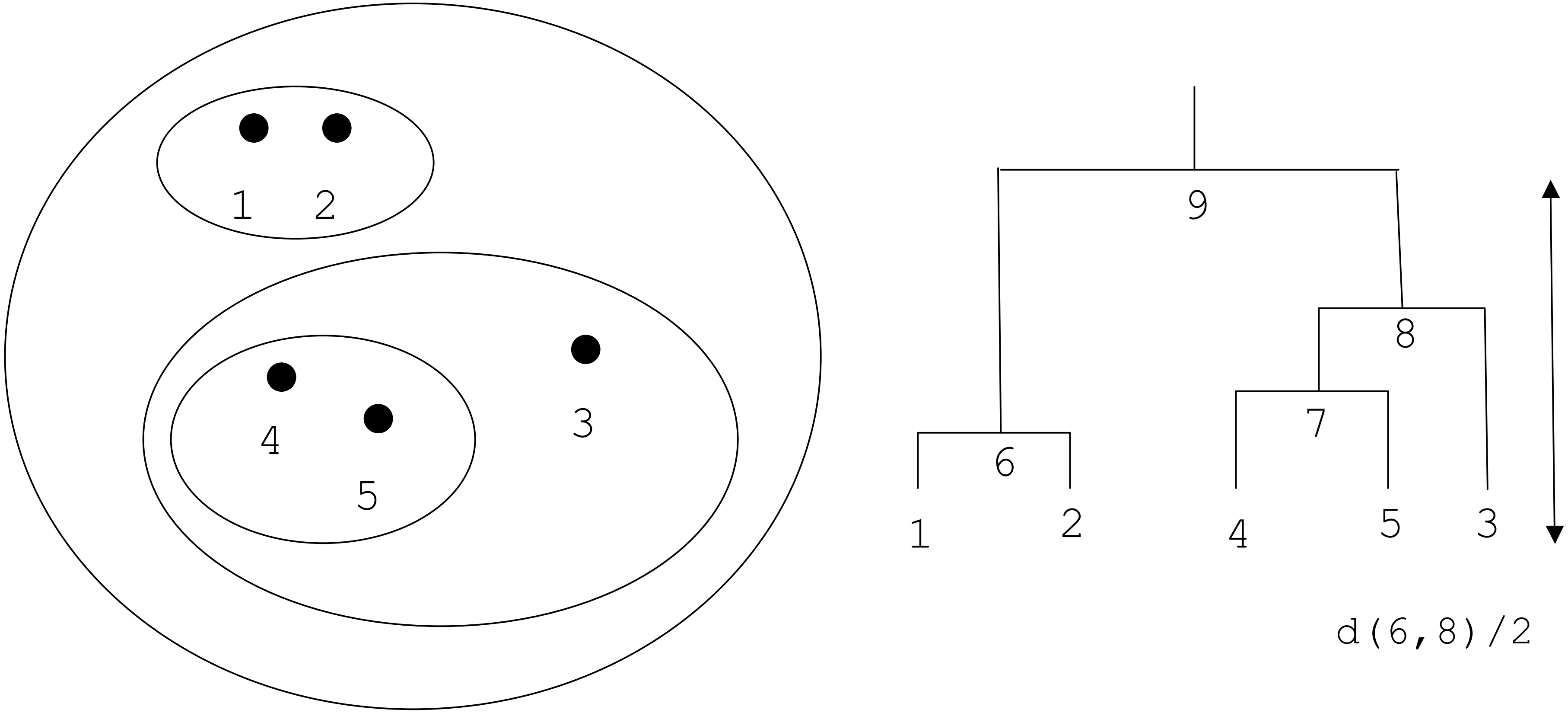
t4=t5=d(4,5)/2

UPGMA example



t3=d(3,7)/2

UPGMA example

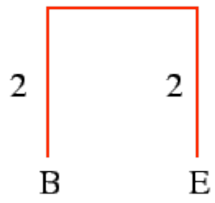


UPGMA tree construction example

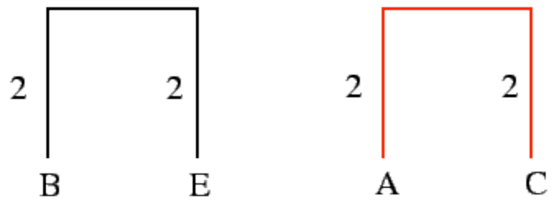
	A	**B**	**C**	**D**	**E**
A	0	8	4	6	8
B		0	8	8	4
C			0	6	8
D				0	8
E					0

Like this?

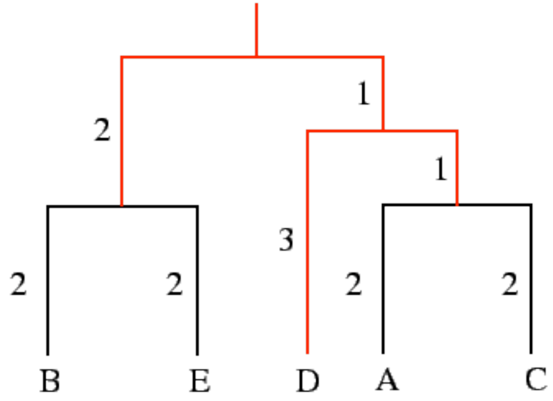
	A	B	C	D	E
A	0	8	4	6	8
B		0	8	8	4
C			0	6	8
D				0	8
E					0



	A	$\binom{B}{E}$	C	D
A	0	8	4	6
(B E)		0	8	8
C			0	6
D				0



	$\binom{A}{C}$	$\binom{B}{E}$	D
(A C)	0	8	6
(B E)		0	8
C			0



Tree (re)construction

Given: set of sequences

Assume that

1 the pairwise sequence alignments provide a measure for the evolutionary distance between the sequences

2. the the resulting distance matrix constitutes an ultrameric (the ideal case),

Reconstruct the phylogenetic, ultrametric tree by the general clustering procedure:

Always pick the closest pair from the distance matrix and merge these two objects into one.

Schemes differ in how the distance between a newly formed object and the other objects is defined.

If object x has been formed by merging y and z, and u is another object:

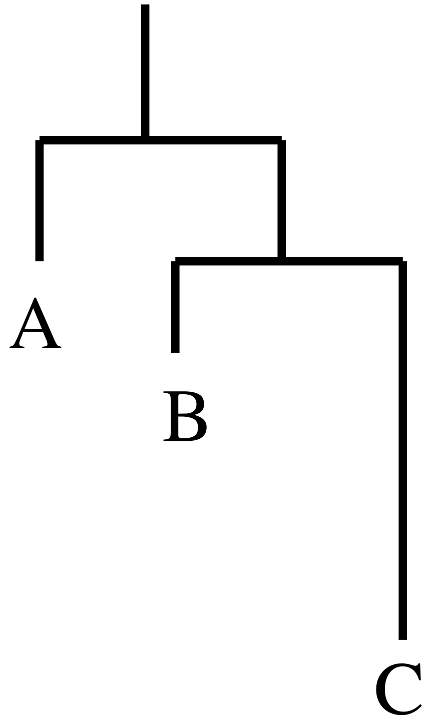
- Single linkage clustering: $d(x,u) = min(d(y,u) , d(z,u))$
- maximal linkage clustering: $d(x,u) = max(d(y,u) , d(z,u))$
- average linkage clustering: $d(x,u) = (d(y,u) + d(z,u)) / 2$

Assumptions underlying UPGMA

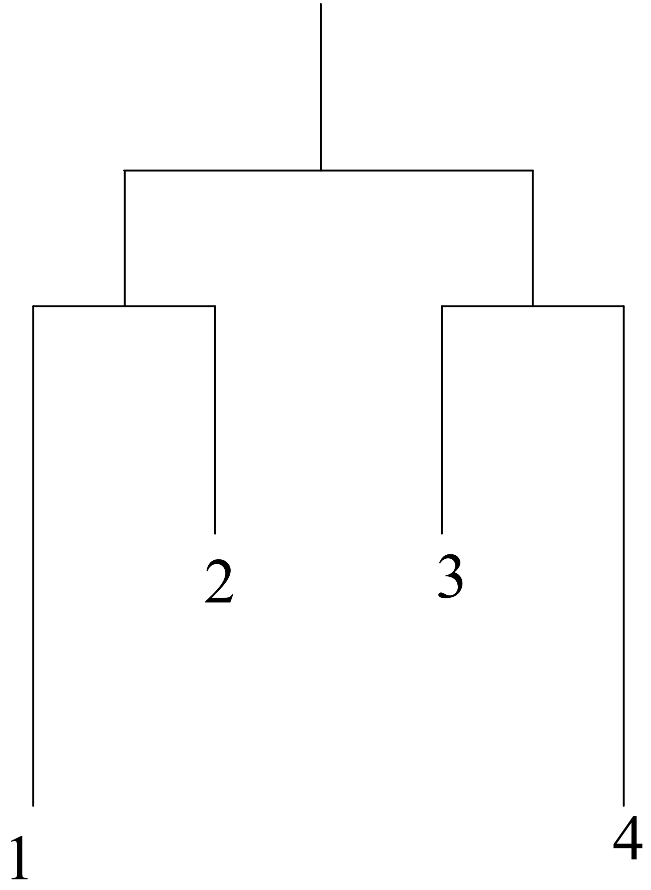
- ***Molecular clock with constant rate***
  - – Edge lengths correspond to times measured by clock
  - – Divergences of (sequences) assumed to occur at constant rate at all points in the tree, I.e.
  - – Sum of times down a path to the leaves from any node is the same, whatever choice of path
- ***Additivity of edge lengths***
  - – Distance between any pair of leaves = sum of the lengths of the edges on the paths connecting them
  - – (automatic when UPGMA tree *constructed*)

Incorrect tree construction by UPGMA???

	A	B	C
A	0	3	6
B		0	5
C			0



Incorrect tree construction by UPGMA



What will UPGMA give?

Incorrect tree reconstruction by UPGMA

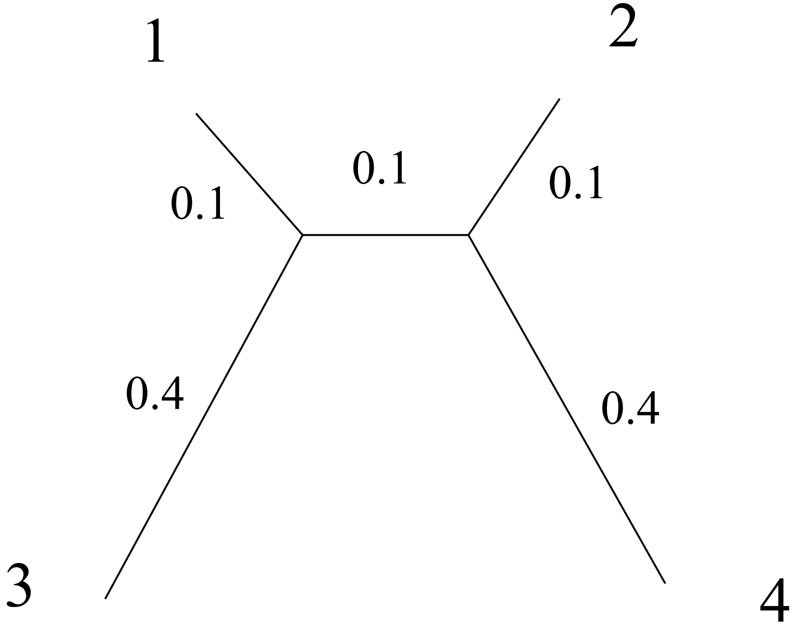
	A	**B**	**C**	**D**	**E**
A	0	16	16	10	6
B		0	6	16	16
C			0	16	16
D				0	10
E					0

Construct the tree by UPGMA

Addivity & neighbour-joining

- Possible to have *molecular clock property fail* but *addivity hold*
- Use e.g. ***neighbour-joining*** algorithm
- Does *not* produce rooted trees…

Tree whose closest leaves are not neighbours.



d(1,2) =?, d(1,3)=?

Which are *neighbouring* and which are *closest*?

Distances

Hence do not pick closest leaves, but subtract the averaged distances to all other leaves to compensate for long edges

$$D(i,j) = d(i,j) - \left(r_i + r_j\right)$$

$$where$$

$$r_i = \frac{1}{|L| - 2} \sum_{k \in L} d(i,k)$$

Test for additivity

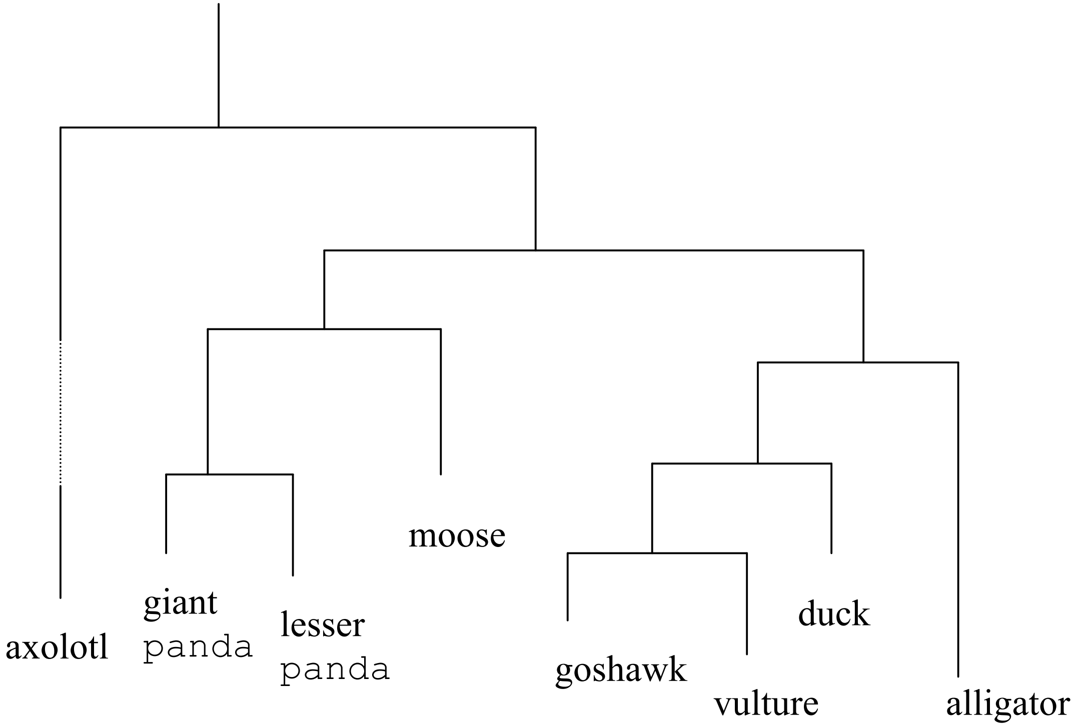
Four point condition:

- **For every set of 4 leaves *i,j,k,l***
  two of the distances
  d(I,j)+d(k,l) d(I,k)+d(j,l), d(I,l)+d(j,k)
  must be equal, and also larger to the third

Can use NJ even if lengths are not additive, but correct reconstruction not guaranteed.

Rooting trees

- Add an ***outgroup***
- E.g. add axolotl
- If no outgroup, then ad-hoc strategies
- See also ***TreeView*** by Rod Page (Glasgow)



Quick demo

http://www.dina.dk/~sestoft/bsa/Match7Applet.html

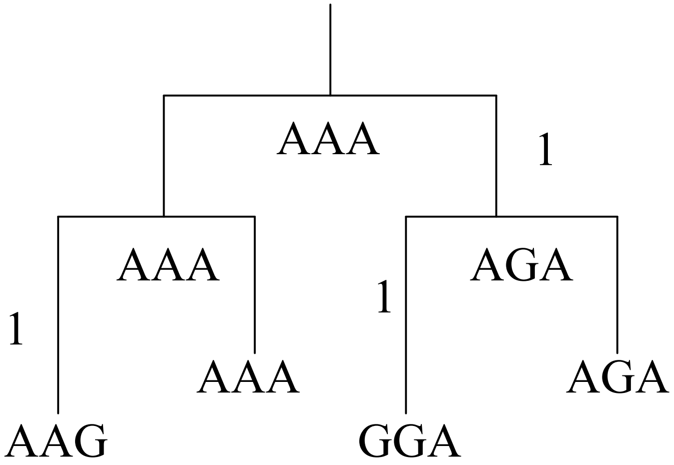
Parsimony

- Character-based method
- Finds the tree which can explain the observed sequences with a minimal number of substitutions

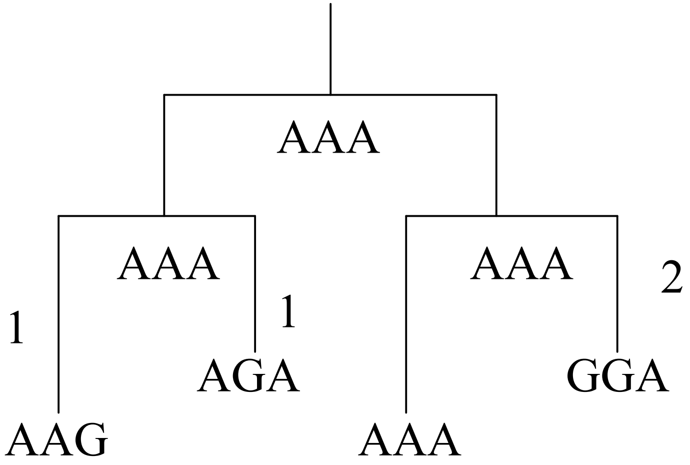
1. Assigns a *cost* to a given tree
2. Searches through all trees to find overall minimum of this cost

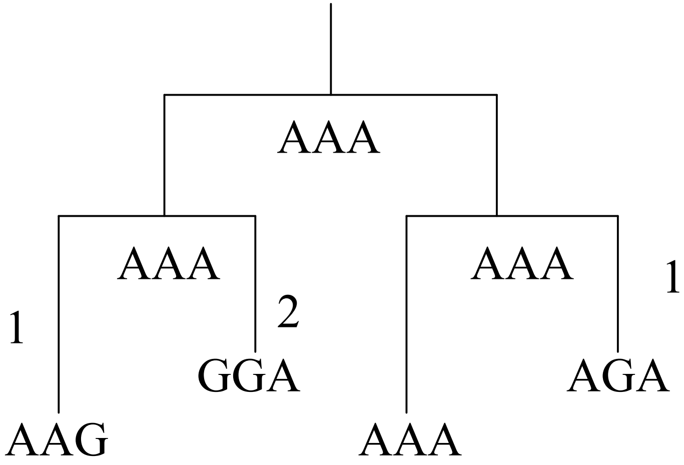
Parsimony example

AAG
AAA
GGA
AGA



Total number of changes?





Parsimony method

- Treats each site independently
- Given a topology & assignment of residues →leaves:
- Basic step = count minimal number of changes needed at 1 site

Weighted parsimony

- Count number of subsitutions *and*
- Add costs S(a,b) for each substitution of a by b
- Aim : minimise this cost
- Weighted parsimony = traditional parsimony when S(a,b) = 1 for all a≠b
- Algorithm starts at the leaves and works up to the root : *post-order traversal*

Traditional parsimony

- Just count the number of substitutions :
- Keep a list of minimal cost residues at each node, plus current cost

Parsimony : rooted / unrooted

- Parsimony forumlated for rooted trees
- Minimal cost in trad.parsimony independent of location of root
- Root can be removed - number of trees to be searched over is reduced
- But easier to count costs in rooted tree

Search stratgeies

- Improve over simple enumeration
- Stochastic methods: randomly swap branches on tree and chose altered tree if better than current
  - Not guaranteed to find overall best tree; adding sequences in different orders can give different trees
- Build tree by adding edges 1 at time [ditto]
- Branch and bound:
  - begin systematicall building trees with increasing number of leaves
  - Abandon avenue whenever current incomplete tree has cost > smallest cost obtained so far for complete tree

Assessing trees - bootstrap

- How much trees should be trusted
- Given dataset of alignment of sequences
  - Generate artificial dataset (same size) by picking columns from alignment at random with replacement (given column in original can appear several times)
  - Apply treebuilding to new dataset
- Repeat many (1000) times
- Frequency with which phylogenetic feature appears ⇒ measure of confidence in feature

Maximum likelihood

- Rank trees according to their
  - likelihood P(data|tree)
  - Posterior probability (Bayesian view) P(tree|data)
- Define & compute the probability of a set of data given a tree
- Need a model of evolution - mutation and selection events that change sequences along the edges of a tree

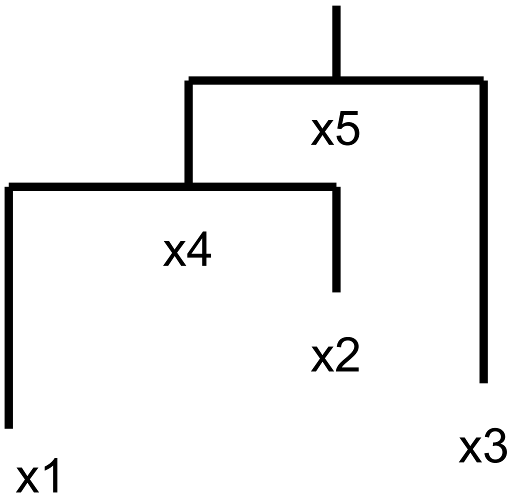
Maximum likelihood

- Maximum Likelihood
- Given a probabilistic model for nucleotide or amino-acid substitution, select the tree that has the highest probability of generating the observed data – sequences
  - Give data *D* and model *M*, find tree *T* such that $P(D|T, M)$ maximised
- Actually what we really want is $P(T|D, M)$
  - Not so easy to obtain as summation over all possible tree topologies required to obtain posterior

$$P(T \mid D, M) = \frac{P(D \mid T, M) P(T, M)}{\sum_{T} P(D \mid T, M) P(T, M)}$$

Phylogenetic Trees

- Assumptions
  - Different sites evolve independently
  - Diverged sequences evolve independently after diverging
- Consider following tree
  - *Likelihood corresponds to* $P(D|T, M) = P(X1,X2,X3|T,M)$



Phylogenetic Trees

- Note that
  - $P(D|T, M)$

    $= P(X1,X2,X3|T,M)$

    $=\sum_{X4,\ X5} P(X1,X2,X3,X4,X5|T,M)$

    $=\sum_{X4,\ X5} P(X1|X4)P(X2|X4)P(X3|X5)P(X4|X5)P(X5)\ |T,M$
- Efficient methods to obtain required summation for likelihood (Felsenstein Algorithm)

- Have now only obtained likelihood we require maximum of likelihood over all possible topologies and models (branch lengths)

- Approximate (stochastic)search methods required

Maximum likelihood tree

- Search over tree topolgies
- For each topology: search over all possible lengths of edges
- (2n-3)!! rooted binary trees with n leaves
- N=10 :: 2 million, N=20 :: 2.2x10^{20}
- Thus need optimisation techniques

Probabilistic models of evolution

- Residue subsitution
- Deletion & insertion of groups of residues
- Simple model:
  - – every site independent
  - – only substitutions

Maximising likelihood

- Small number of sequences (2..5) : can enumerate all trees
  - Write down likelihood as function of edge lengths
  - Maximise likelihood by numerical technique (Kishano et al 1990, protein sequences - use PAM matrices)

- Larger number of sequences: Felsenstein’s algorithm

- For protein sequences computationally demanding 20x20 substitution matrix

- Can use sampling methods

More realistic evolutionary models

- Different rates at different sites

- Models with gaps

Internet resources and programs

- Compilation of available programs: http://evolution.genetics.washington.edu/phylip/software.html
- http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html -- phylip , e.g. for nj
- http://evolution.genetics.washington.edu/phylip.html Phylip home page [widely used, poor interface…]
- http://www.hiv.lanl.gov/content/hiv-db/TREE_TUTORIAL/Tree-tutorial.html

Phylogenetic Trees

Differences between methods

- UPGMA does not employ models of evolution
- Maximum Likelihood & Maximum Parsimony employ models of evolution
- Maximum Likelihood claimed to be best for large evolutionary distances
  - Very time consuming to build trees

Summary

- Phylogenies
- Trees, representations, rooted, unrooted
- Metrics, ultrametrics, additivity, molecular clock
- Distance methods: UPGMA (ultrametric), NJ (additive; no root)…
- Tree rooting
- Some Programs