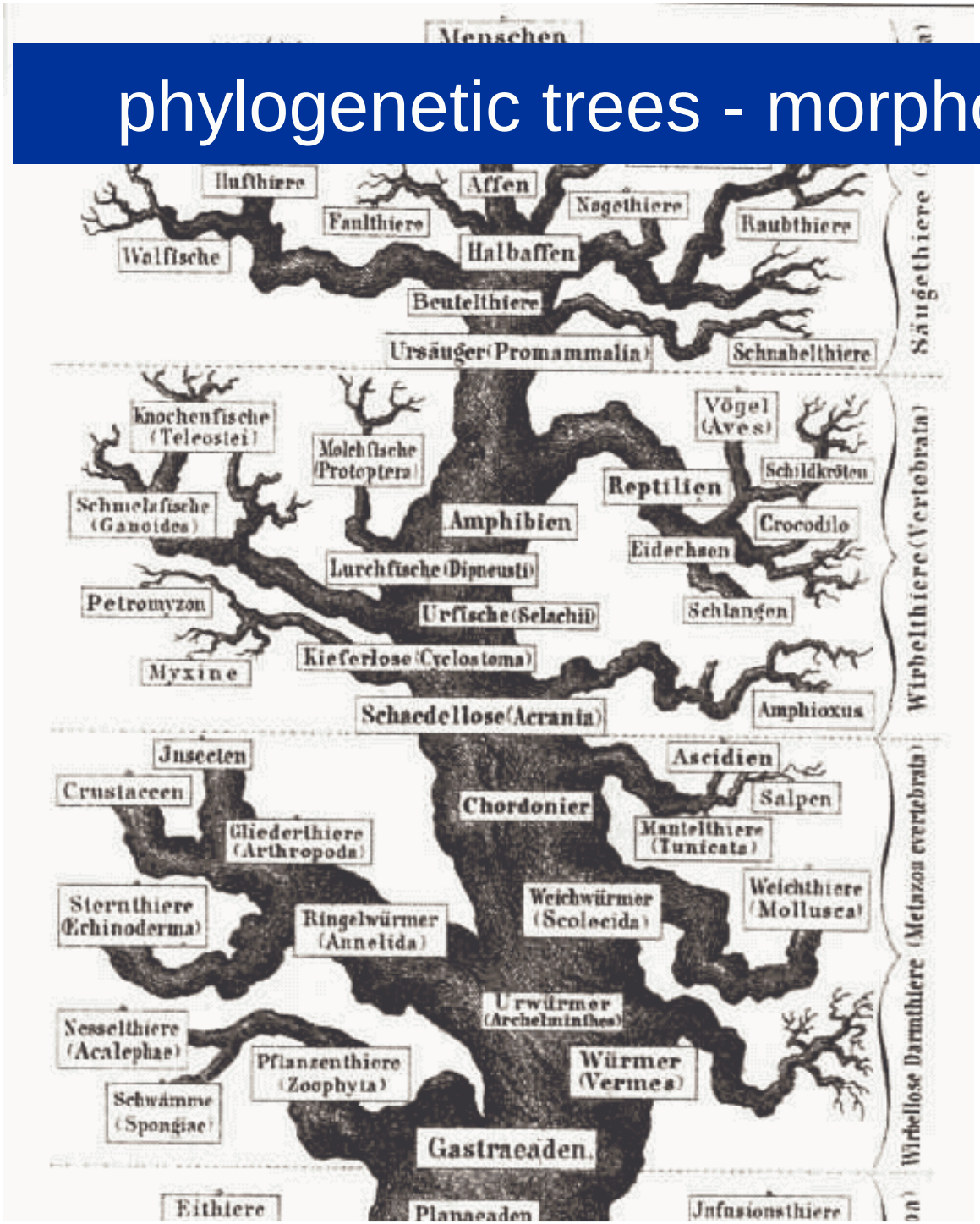
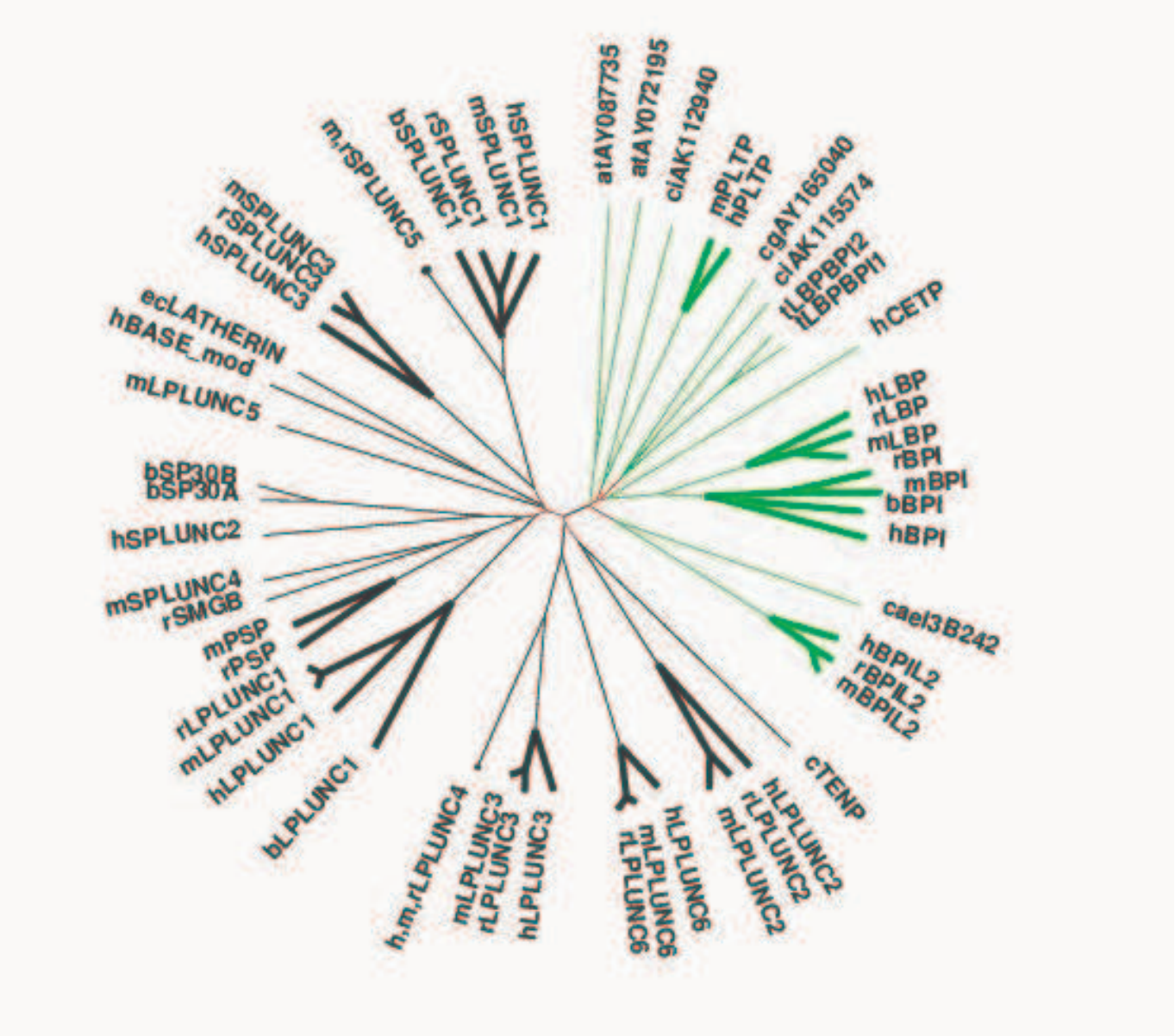
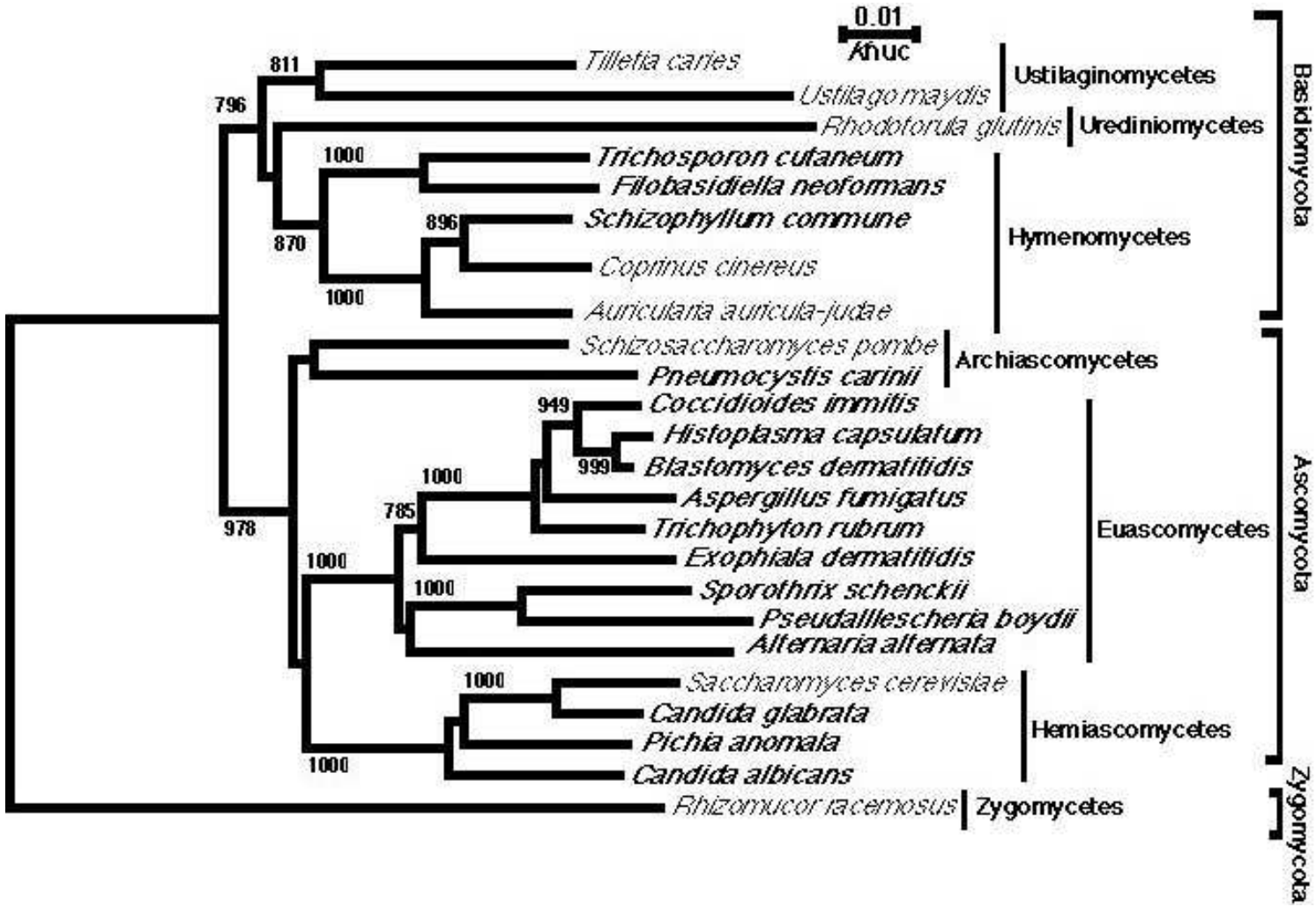
phylogenetic trees - morphological



phylogenetic trees - genetic



phylogenetic trees - genetic



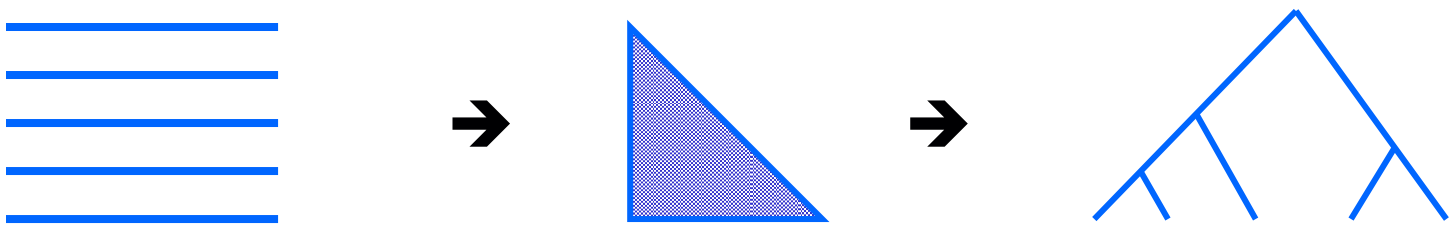
Phylogeny

Charlie Chaplin once entered a Charlie Chaplin look-alike contest in Monte Carlo, Monaco.
He placed third.

phylogeny

general approach

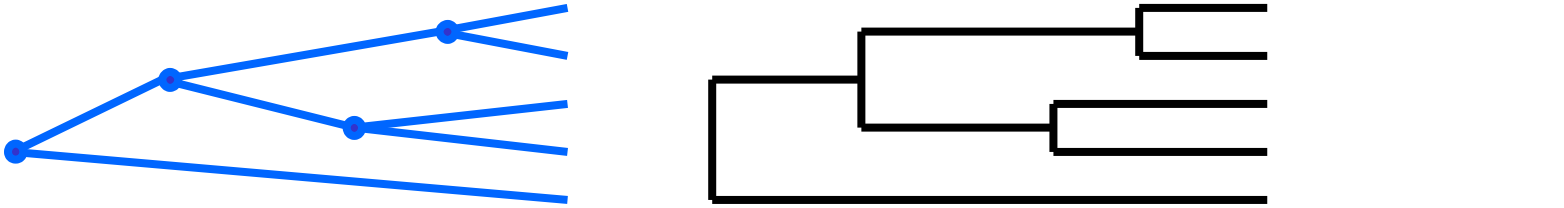
> choose sequences (taxa)
> multiple alignment
⇒ character matches
⇒ distances
> find tree (topology + edge lengths)



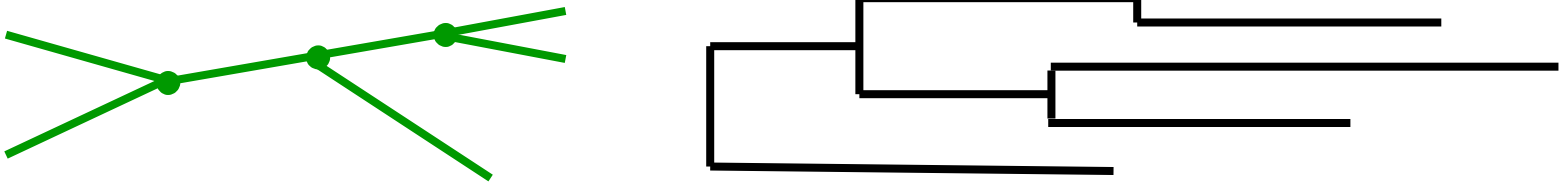
trees

> leaves: taxa '*operational taxonomic units*'
internal: hypothetical

> binary bifurcation, independent visual representation



> rooted vs. unrooted



> branch length (edges) ∼ evolutionary model / clock

> too many trees !!
assumptions on tree or heuristics

homologous genes

homologous genes ~ corresponding
- orthologous in different species
- paralogous in same species (genome) α β γ δ ε …

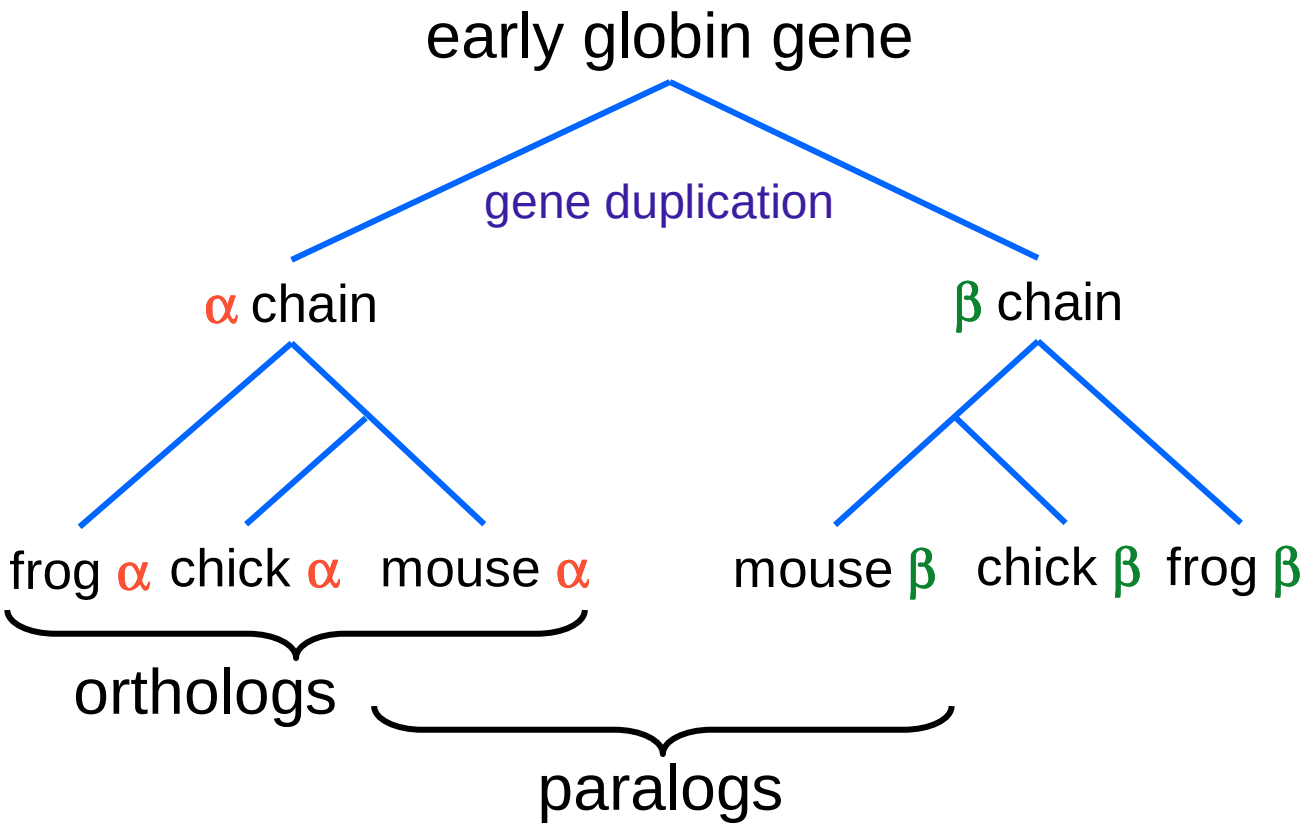


table of contents by methods available:

General-purpose ● Parsimony ● Distance matrix ● Computation of distances ● Maximum likelihood, Bayesian ● Quartets ● Genetic algorithms ● Evolutionary Parsimony ● Interactive tree manipulation ● Looking for hybridization ● Bootstrapping ● Compatibility analysis ● Consensus trees, distances ● Tree-based alignment ● Gene duplication ● Biogeographic analysis ● Comparative method analysis ● Simulation of trees ● Examination of shapes ● Clocks, dating, stratigraphy ● Prediction of data from trees ● Tree plotting/drawing ● Teaching ● Web or e-mail servers

http://evolution.genetics.washington.edu/phylip/software.html

phylogeny

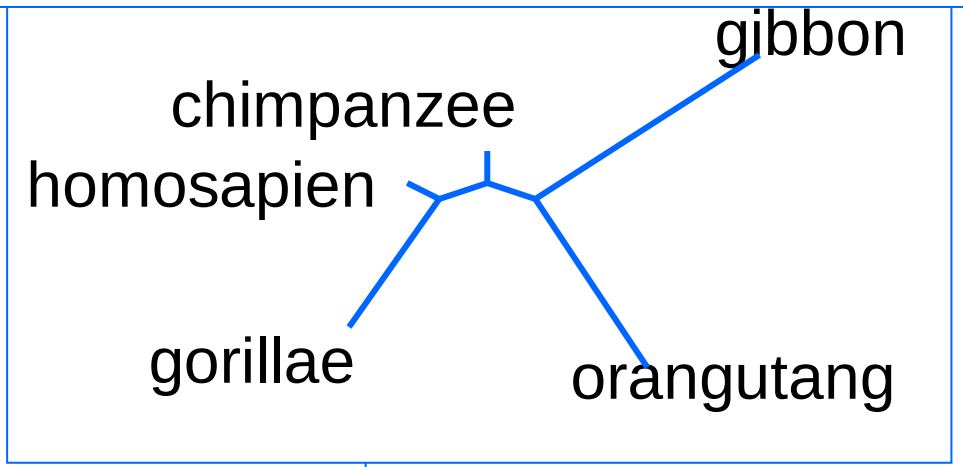
```
chimpanzee AAGCTTCACC GGCGCAATTA TCCTCATAAT CGCCCACGGA CTTACATCCT
gibbon          T  A    T    CCG                          A  C  T
gorillae                    G  T  T       T                    A
homosapien                 G C   T
orangutang                  CC  C      G   T      T        T
```

```
chimpanzee
gibbon       0.21
gorillae     0.13   0.29
homosapien   0.06   0.24   0.11
orangutang   0.15   0.26   0.25   0.18
```

```
Neighbor-joining method

   ┌----gorillae
   |
--3homosapien
   |
   |   ┌-chimpanzee
   └--2
      |   ┌---------gibbon
      └--1
          └------orangutang

 remember: this is an unrooted tree!
```



http://www.dina.dk/~sestoft/bsa/dinaws/neighbour.html

contents

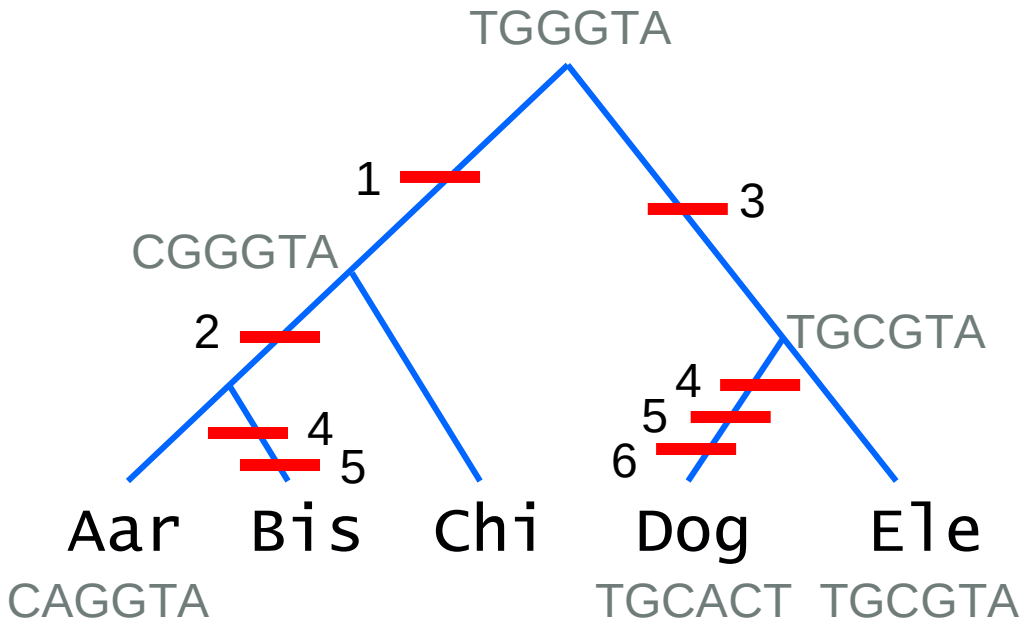
this lecture

- Character based
  - Sankoff
  - Maximum Likelihood Method
  - Heuristic Approaches
- Distance based
  - Neighbour Joining
  - UPGMA
  - Heuristic Approaches

character based

parsimony: minimal number of changes
5 species, 6 characters

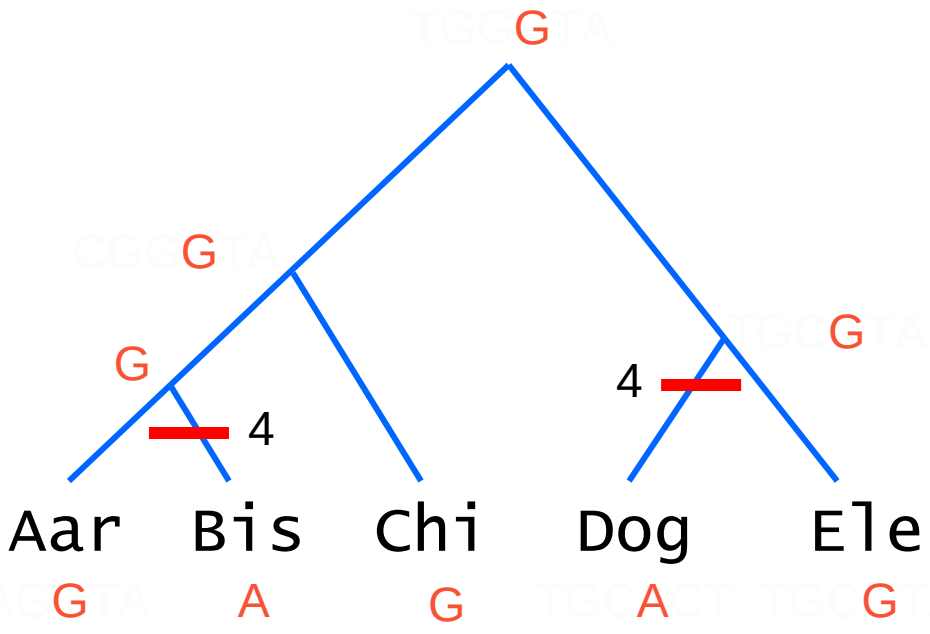
character states:	1	2	3	4	5	6
Aardvark	C	A	G	G	T	A
Bison	C	A	G	A	C	A
Chimp	C	G	G	G	T	A
Dog	T	G	C	A	C	T
Elephant	T	G	C	G	T	A



small parsimony: topology given, find labeling
large parsimony: also find optimal topology

character based

small parsimony: topology given
find labeling for character 4



character states:	1	2	3	4	5	6
Aardvark	C	A	G	G	T	A
Bison	C	A	G	A	C	A
Chimp	C	G	G	G	T	A
Dog	T	G	C	A	C	T
Elephant	T	G	C	G	T	A

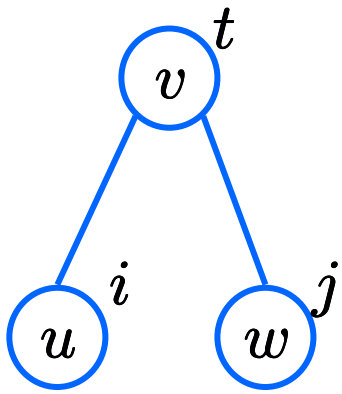
character based Sankoff

small parsimony (single character)

given topology: rooted tree

- k species - k states [in leaves]
- cost for changing state C_{ij}

bottom-up tree evaluation:
assign a cost vector $S_t(v)$ at each node
= minimal cost of state t at node v



✓ leaf

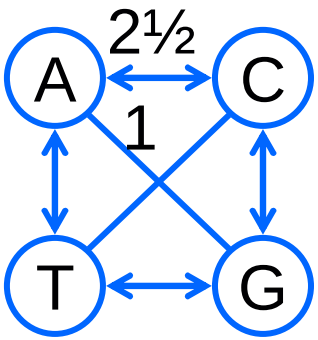
$$S_t(v) = \begin{cases} 0 & \text{state}(v) = t \\ \infty & \text{otherwise} \end{cases}$$

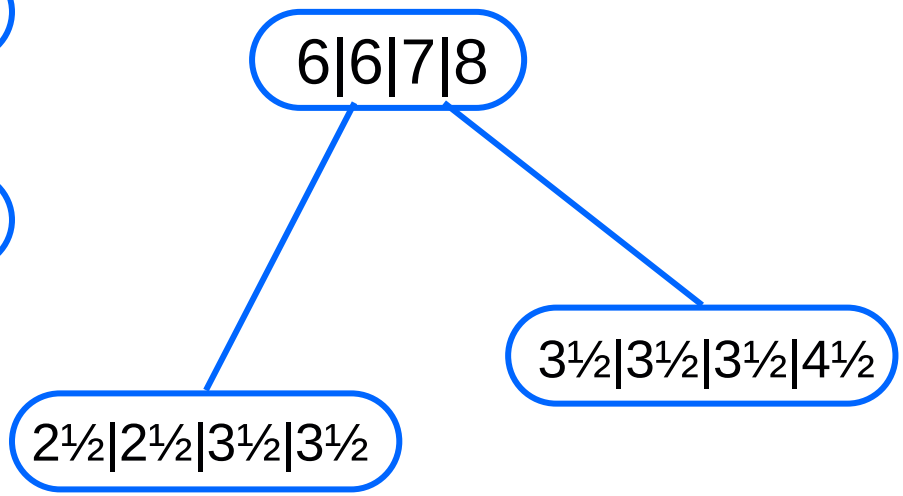
✓ internal

$$S_t(v) = \min_i \{C_{ti} + S_i(u)\} + \min_j \{C_{tj} + S_j(w)\}$$

to be repeated for each character

character based Sankoff





A|C|G|T

$S_t(v)$

A 6
C 6
G 7
T 8

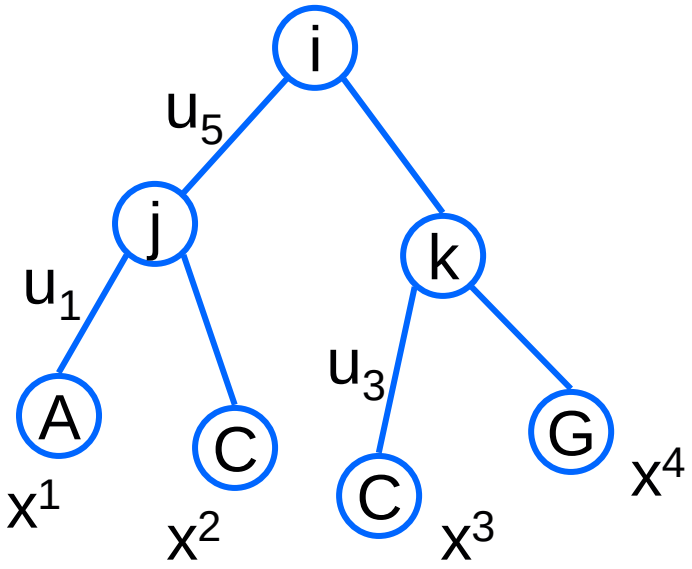
min { 2½+0, 2½+2½, 3½+1, 3½+2½ }
+ min { 3½+0, 3½+2½, 3½+1, 4½+2½ }

min { 2½+2½, 2½+0, 3½+2½, 3½+1 }
+ min { 3½+2½, 3½+0, 3½+2½, 4½+1 }

min { 2½+1, 2½+2½, 3½+0, 3½+2½ }
+ min { 3½+1, 3½+2½, 3½+0, 4½+2½ }

min { 2½+2½, 2½+1, 3½+2½, 3½+0 }
+ min { 3½+2½, 3½+1, 3½+2½, 4½+0 }

maximum likelihood



A	A	G	x^1
C	T	C	x^2
C	G	G	x^3
G	G	A	x^4

four taxa, three sites

site specific probability, multiply over all sites
probabilities $p_{ij}(t)$ given by evolutionary model

$$P(D|T) = \sum_{ijk} p_{ij}(u_5) \dots p_{kC}(u_3) \dots$$

evolutionary models

continuous time Markov model

Jules-Cantor

instantaneous change

$q_{xx} = -¾\alpha \quad q_{xy} = ¼\alpha$ evolutionary rate α

- $p_{xx}(t) = ¼ + ¾ \exp(-t\alpha)$
- $p_{xy}(t) = ¼ - ¼ \exp(-t\alpha)$

Kimura

transitions A – C – G – T – A

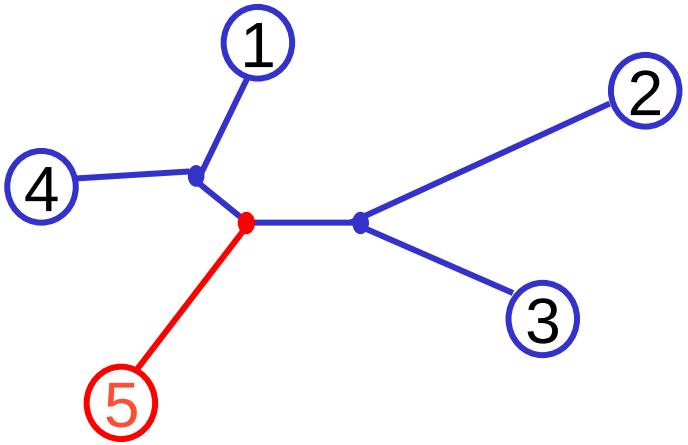
transversions A – G, T – C

Felsenstein, HKY, ...

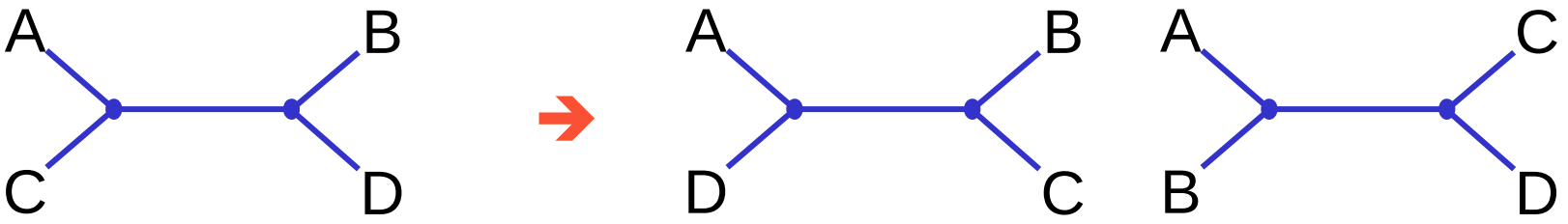
character based heuristics

large parsimony is NP complete ☹
search strategies, heuristics
space = ‘all’ trees

→ branch-and-bound
add taxon (species) at a time



→ nearest neighbour interchanges



II

distance based

metric space objects + distance d

- $d(x,y) \geq 0$
- $d(x,y) = 0$ iff $x=y$
- $d(x,y) = d(y,x)$ symmetric

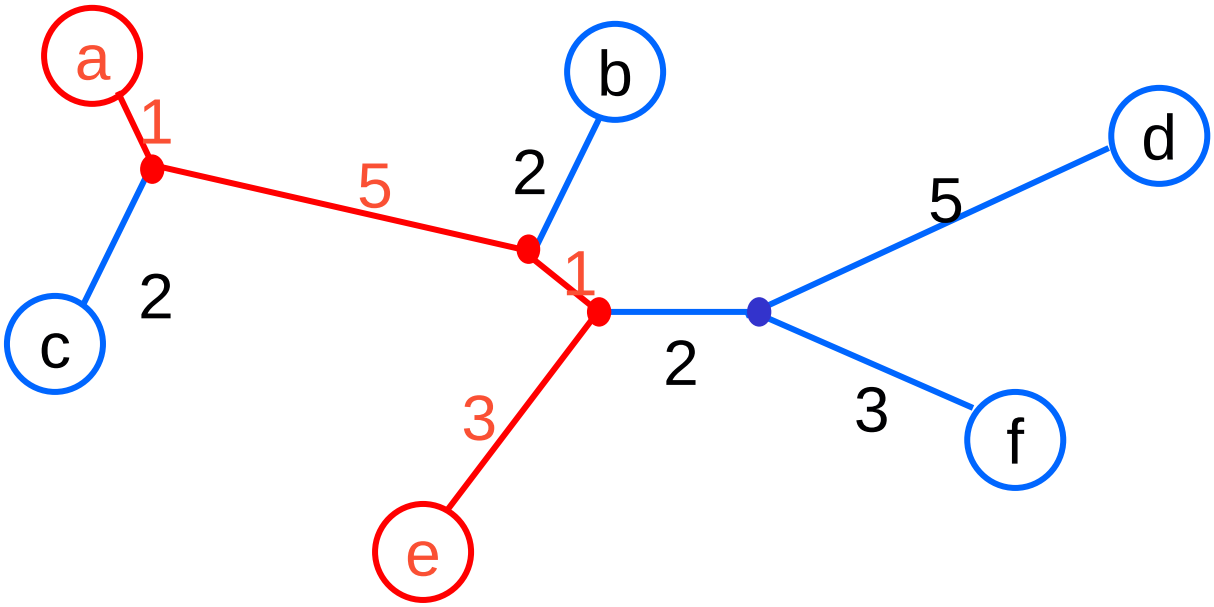
triangle inequality

- $d(x,z) \leq d(x,y) + d(y,z)$

additional properties $\Rightarrow$ specific algorithms

- triangle inequality
- additive (Neighbour Joining)
- ultrametric (UPGMA)

distance based additive

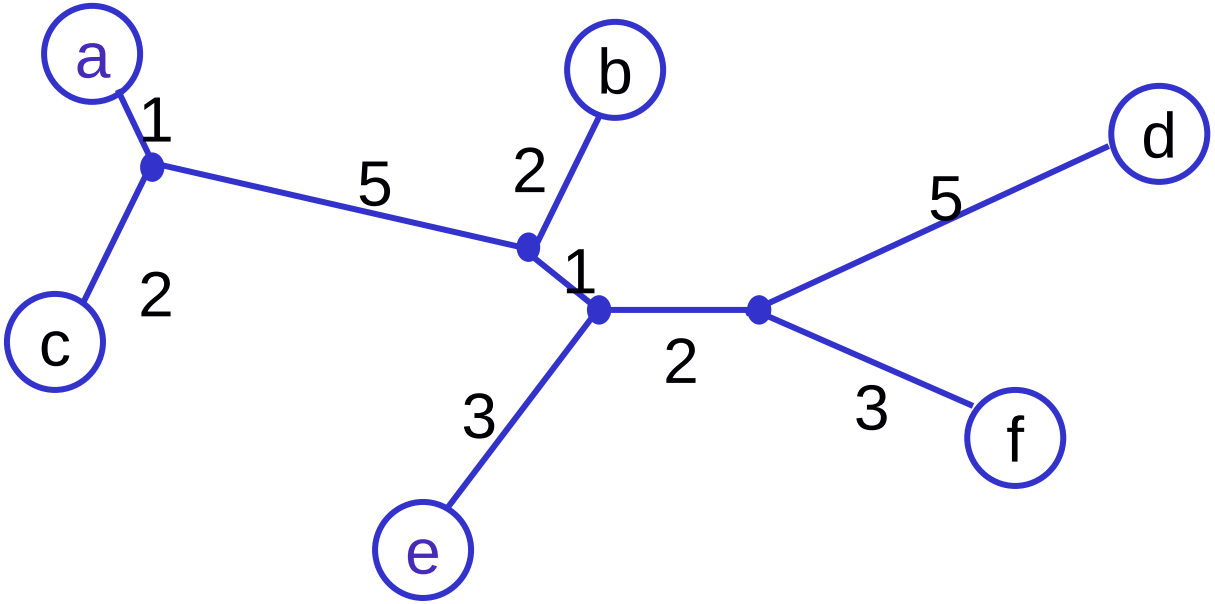


additive distance function = generated by a tree

d(a,e) = 1+5+1+3=10

- given a distance matrix
  how to recognize it is additive ?
- given an additive distance matrix
  how to (re)construct the tree ?

distance based additive



	a	b	c	d	e	f
a	-	8	3	14	10	12
b		-	9	10	6	8
c			-	15	11	13
d				-	10	8
e					-	8
f						-

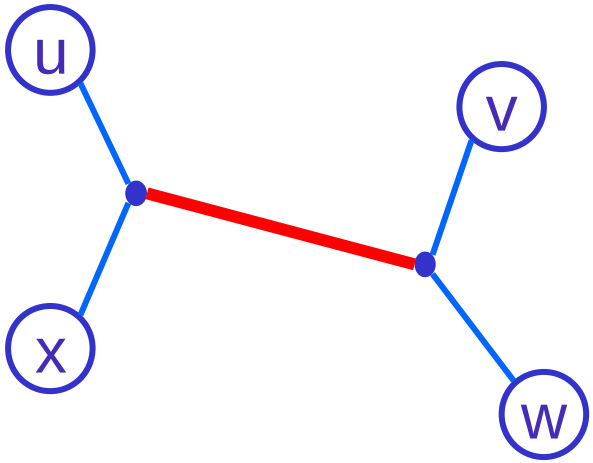
distance based additive

➔ given a distance matrix
how to recognize it is additive ?

four point condition ⇔ additive

$$d(u,v) + d(w,x) = d(u,w) + d(v,x) \geq d(u,x) + d(v,w)$$

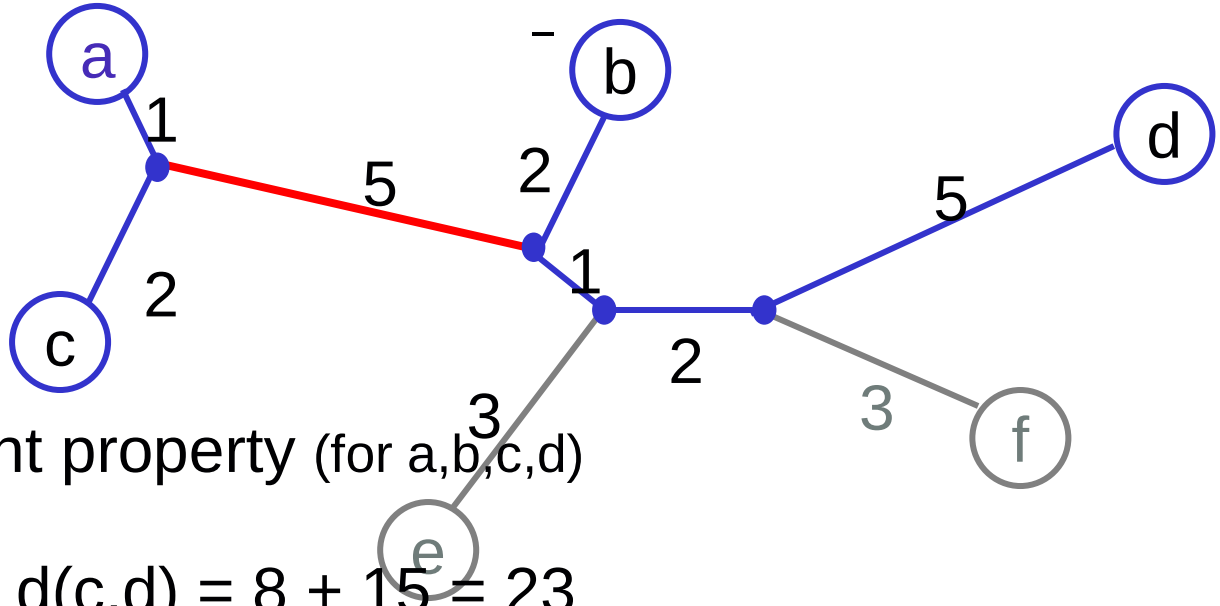
for some ordering of u,v,w,x



distance based additive

lsaev

	a	b	c	d	e	f
a	-	8	3	14	10	12
b		-	9	10	6	8
c			-	15	11	13
d				-	10	8
e					-	8
f						-



four point property (for a,b,c,d)

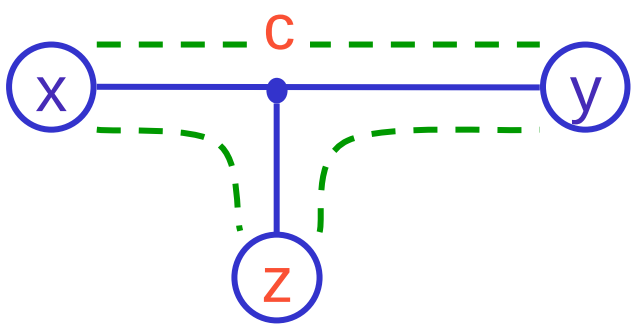
d(a,b) + d(c,d) = 8 + 15 = 23
d(a,c) + d(b,d) = 3 + 10 = 13
d(a,d) + d(b,c) = 14 + 9 = 23

distance based

➔ given an additive distance matrix
how to (re)construct the tree ?

intuition: adding the third point … where?

$$d(x,y) = \ell_{xc} + \ell_{cy}$$
$$d(x,z) = \ell_{xc} + \ell_{cz}$$
$$d(y,z) = \ell_{yc} + \ell_{cz}$$



$$\ell_{cy} = \frac{d(x,y) + d(y,z) - d(x,z)}{2}$$

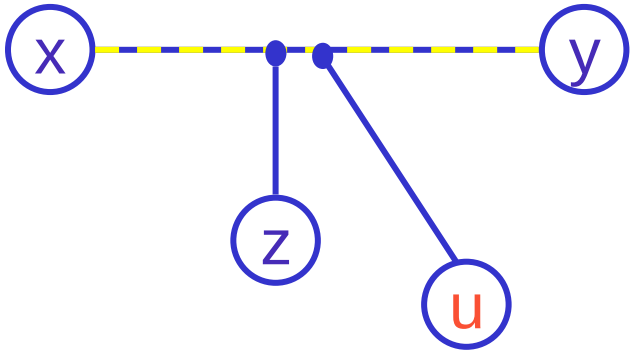
≥ 0 triangle inequality

and symmetric for cx, cz

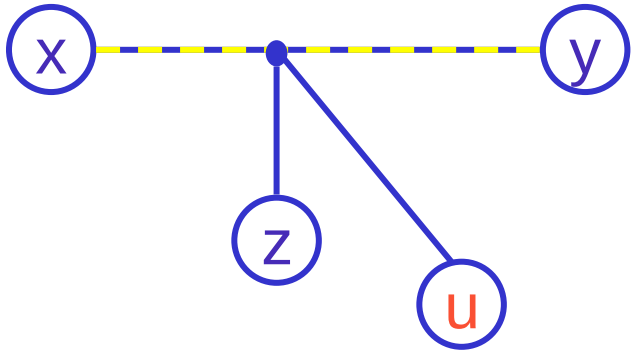
distance based

➔ given an additive distance matrix
how to (re)construct the tree ?

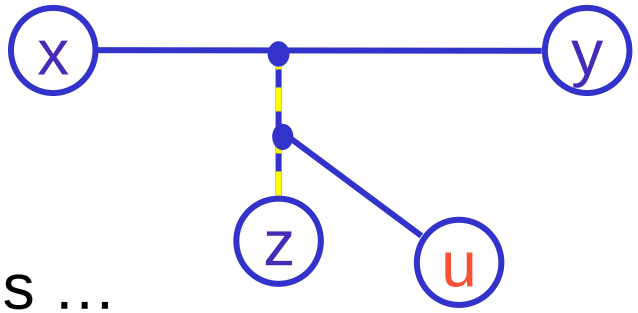
adding another point to (x,y)



ok



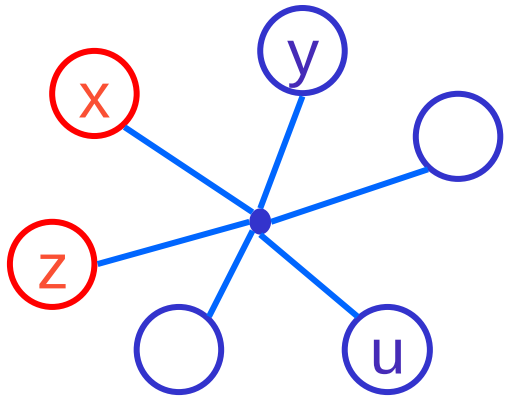
try again for (x,z)

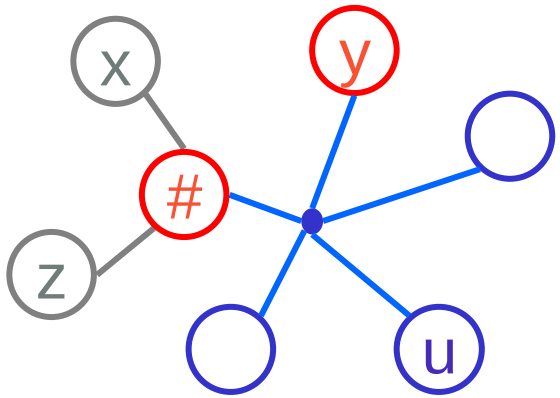


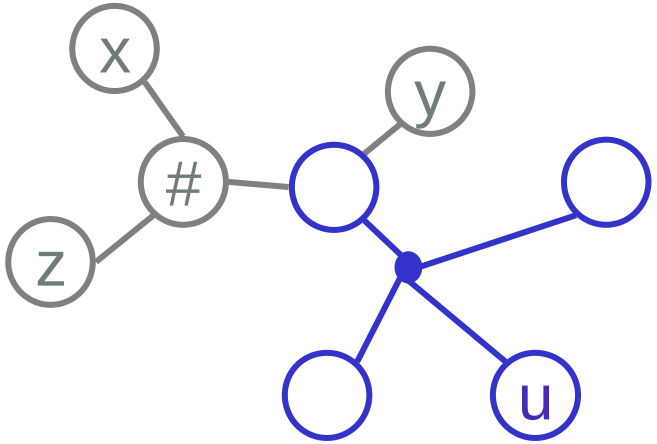
in reality errors in distances …

distance based neighbour joining

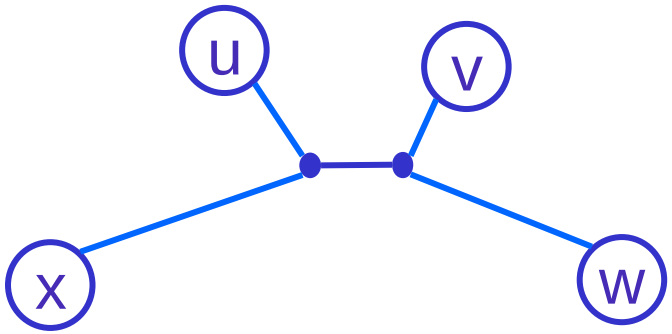
intuition: star graph, join closest neighbours







shortest distance $\not\Rightarrow$ neighbours



distance based neighbour joining

total distance to other nodes

$$r(x) = \frac{1}{N-2} \sum_z d(x,z)$$

pairs of nodes

$$D(x,y) = r(x) + r(y) - d(x,y)$$

- pick x,y maximal for D(x,y)
- join x,y into a new node z

distances of z to ‘old’ nodes:

d(z,x) = ½(d(x,y) + r(x) - r(y))

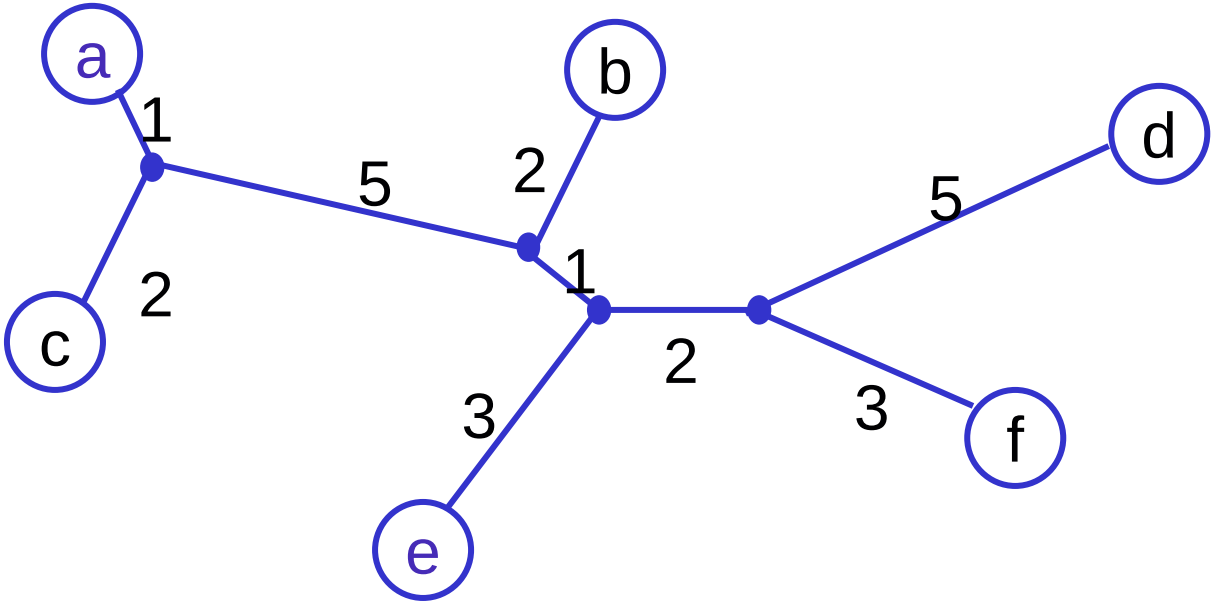
d(z,y) = …

thus d(x,z) + d(z,y) = d(x,y)

d(z,w) = ½(d(x,w) + d(y,w) - d(x,y))

thus ≥ 0

neighbour joining example

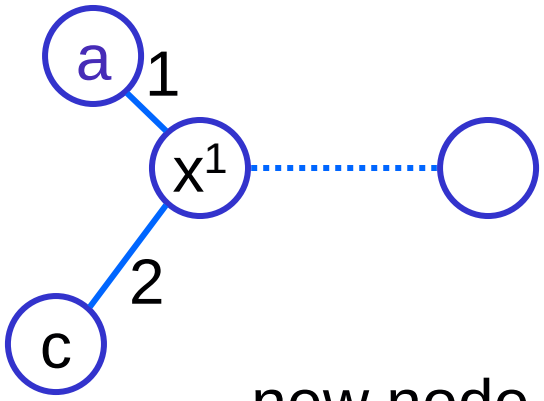


	a	b	c	d	e	f
a	-	8	3	14	10	12
b		-	9	10	6	8
c			-	15	11	13
d				-	10	8
e					-	8
f						-

neighbour joining example

	a	b	c	d	e	f	4·r
a	-	8	3	14	10	12	47
b		-	9	10	6	8	41
c			-	15	11	13	51
d				-	10	8	57
e					-	8	45
f						-	49

N-2=4



4·D(x,y)

	a	b	c	d	e	f
a	-	56	86	48	52	48
b		-	56	58	32	58
c			-	48	52	48
d				-	62	74
e					-	62
f						-

new node x^1 = {a,c}

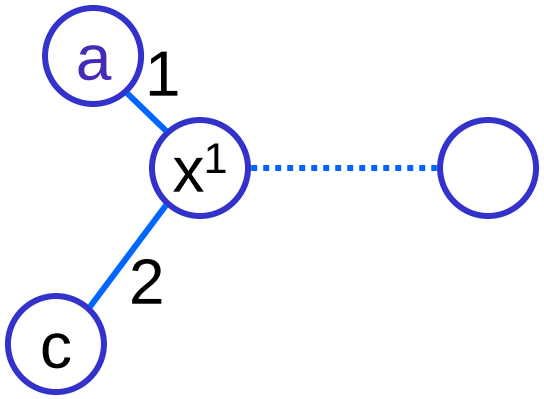
$d(a,x^1) = \frac{1}{2}(3 + {}^{47}/_{4} - {}^{51}/_{4}) = 1$

$d(c,x^1) = \frac{1}{2}(3 + {}^{51}/_{4} - {}^{47}/_{4}) = 2$

neighbour joining example

	a	b	c	d	e	f	x1
a	-	8	3	14	10	12	1
b		-	9	10	6	8	7
c			-	15	11	13	2
d				-	10	8	13
e					-	8	9
f						-	11
x1							-

N-2=4



new node x^1 = {a,c}

$d(x^1,b) = ½(\ 8 + \ 9 - 3\ ) = \ 7$

$d(x^1,d) = ½(\ 14+15 - 3\ ) = 13$

distance based ultrametric

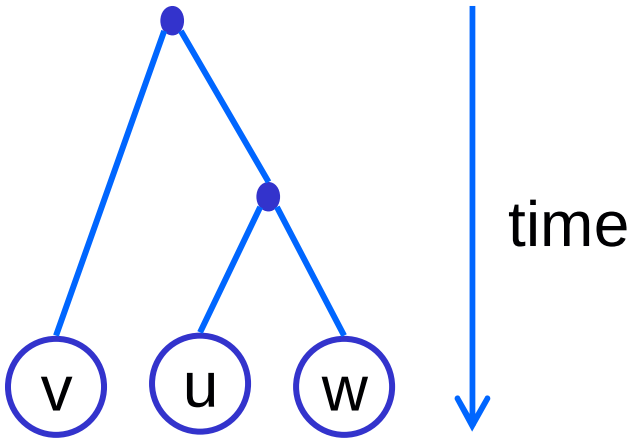
ultrametric distances

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

$\Leftrightarrow$ three point condition

$d(u,v) = d(v,w) \geq d(u,w)$

for some ordering of u,v,w

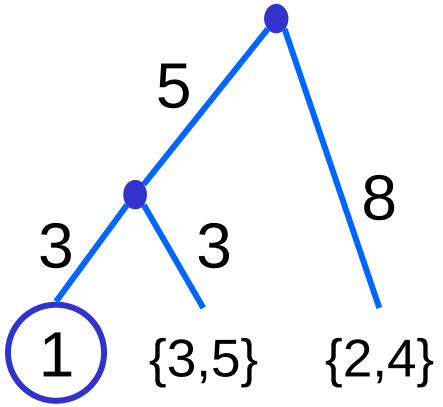


(uniform) molecular clock

distance based ultrametric

sort according to distance to node 1

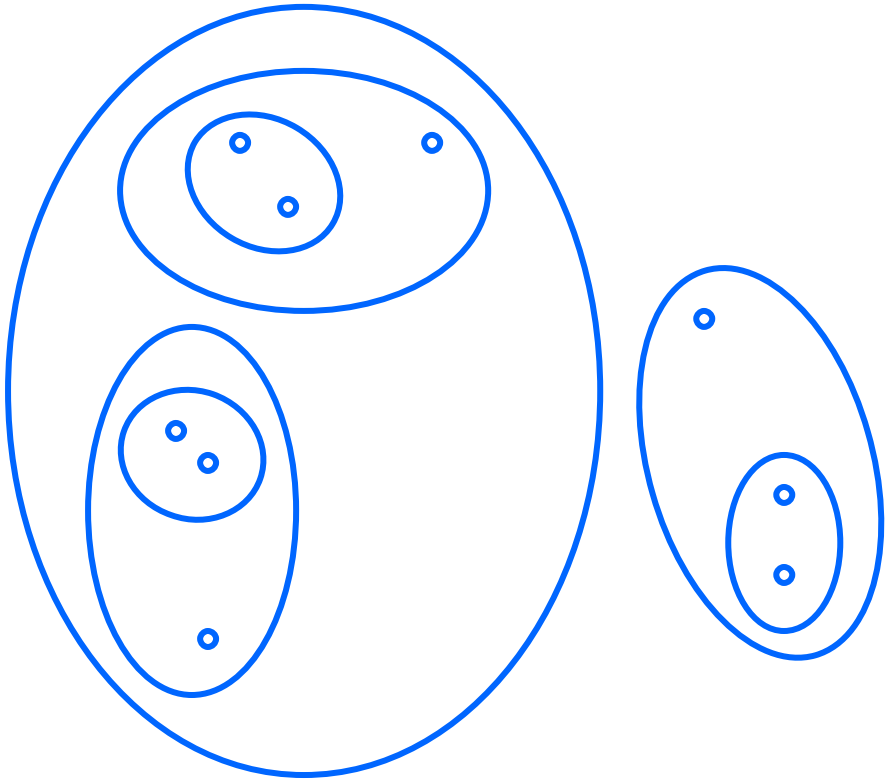
	1	2	3	4	5
1	-	16	6	16	6
2		-	16	8	16
3			-	16	2
4				-	16
5					-



recursively solve for {3,5} and {2,4} (easy)

- ✓ conceptually simple, but
- ✓ tricky to implement efficiently
- ✓ can be adapted to the additive case
- ✓ even to character methods [says Gusfield]
  combinatorial rather than numeric
- ✓ UPGMA more general: clustering

distance based clustering



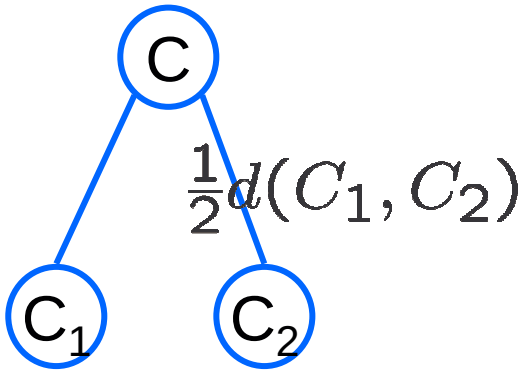
distance based UPGMA

unweighted pair group method using arithmetic averages

based on clustering

$$d(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{\substack{x \in C_1 \\ y \in C_2}} d(x, y)$$

- choose minimal distance custers
- join them
- compute new distances



distance based UPGMA

new distances from old

$C = C_1 \cup C_2$

$$d(C, C') = \frac{1}{|C| \cdot |C'|} \sum_{\substack{x \in C \\ y \in C'}} d(x, y)$$

$$d(C, C') = \frac{|C_1|}{|C|} d(C_1, C) + \frac{|C_2|}{|C|} d(C_2, C)$$

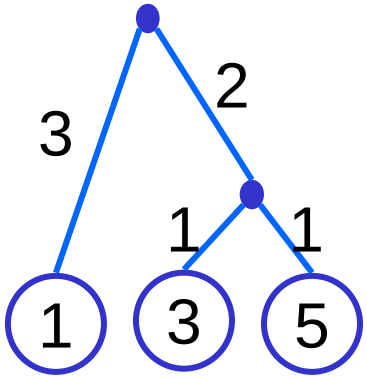
UPGMA example

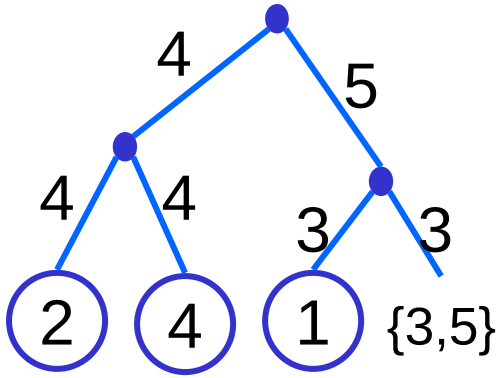
	1	2	3	4	5
1	-	16	6	16	6
2		-	16	8	16
3			-	16	2
4				-	16
5					-

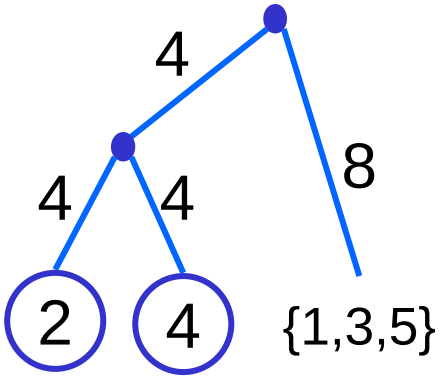
	1	2	4	35
1	-	16	16	6
2		-	8	16
4			-	16
35				-

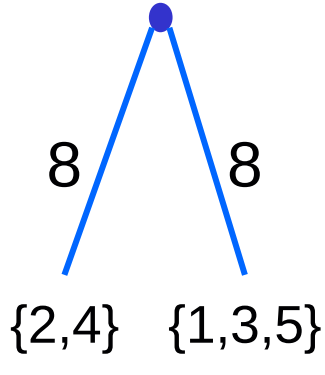
	2	4	135
2	-	8	16
4		-	16
135			-

	24	135
24	-	16
135		-









distance based heuristics

comparing distances: matrix M vs. tree d

$$\rho(M,d) = \sum_{xy} (M(x,y) - d(x,y))^2$$

optimization problem
- M given (matrix)
- d from suitable tree (to be found)

search ‘tree space’
genetic algorithms, etc.