Introduction to Bioinformatics for Computer Scientists

Lecture 8

Beers

- On December 14
- Do we know where?
- Some of my lab members will also join

Exam Dates

- February 12 → we may have to change this date, I'll know after Christmas
- March 1
- If those dates don't work → ad hoc in Heidelberg

Plan for next lectures

- Today: Introduction to phylogenetics
- Lecture 9: Phylogenetic search algorithms
- Lecture 10 (Kassian): The phylogenetic Maximum Likelihood Model
- Lecture 11: Parallel Computing in Phylogenetics
- Lecture 12: (Andre) Discrete Operations on Trees
- Lecture 13: (Pavlos) Population genetics I
- Lecture 14: (Pavlos) Population genetics II
- Lecture 15: Wrap-up

The story so far

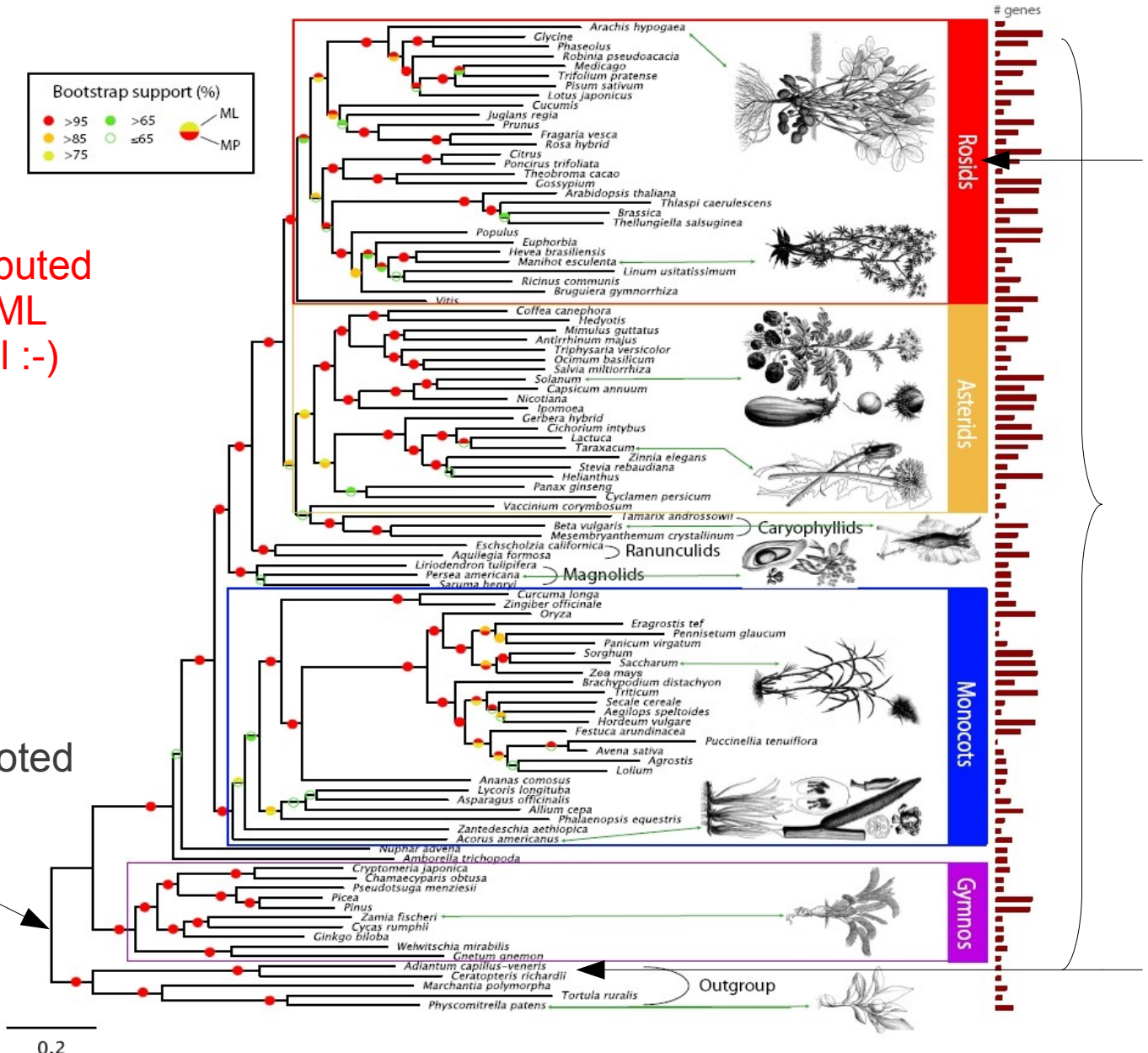
- Biological Terminology: RNA, DNA, genes, genomes, etc
- Pair-wise Sequence Alignment
- Sequence Comparison
- Genome Assembly
- Suffix Trees and Indexing
- Multiple Sequence Alignment

The story so far

- Biological Terminology: RNA, DNA, genes, genomes, etc
- Pair-wise Sequence Alignment
- Sequence Comparison
- Genome Assembly
- Suffix Trees and Indexing
- Multiple Sequence Alignment
- Phylogenetic Inference

A Phylogeny or Phylogenetic Tree

Tree computed
With RAxML
→ our tool :-)



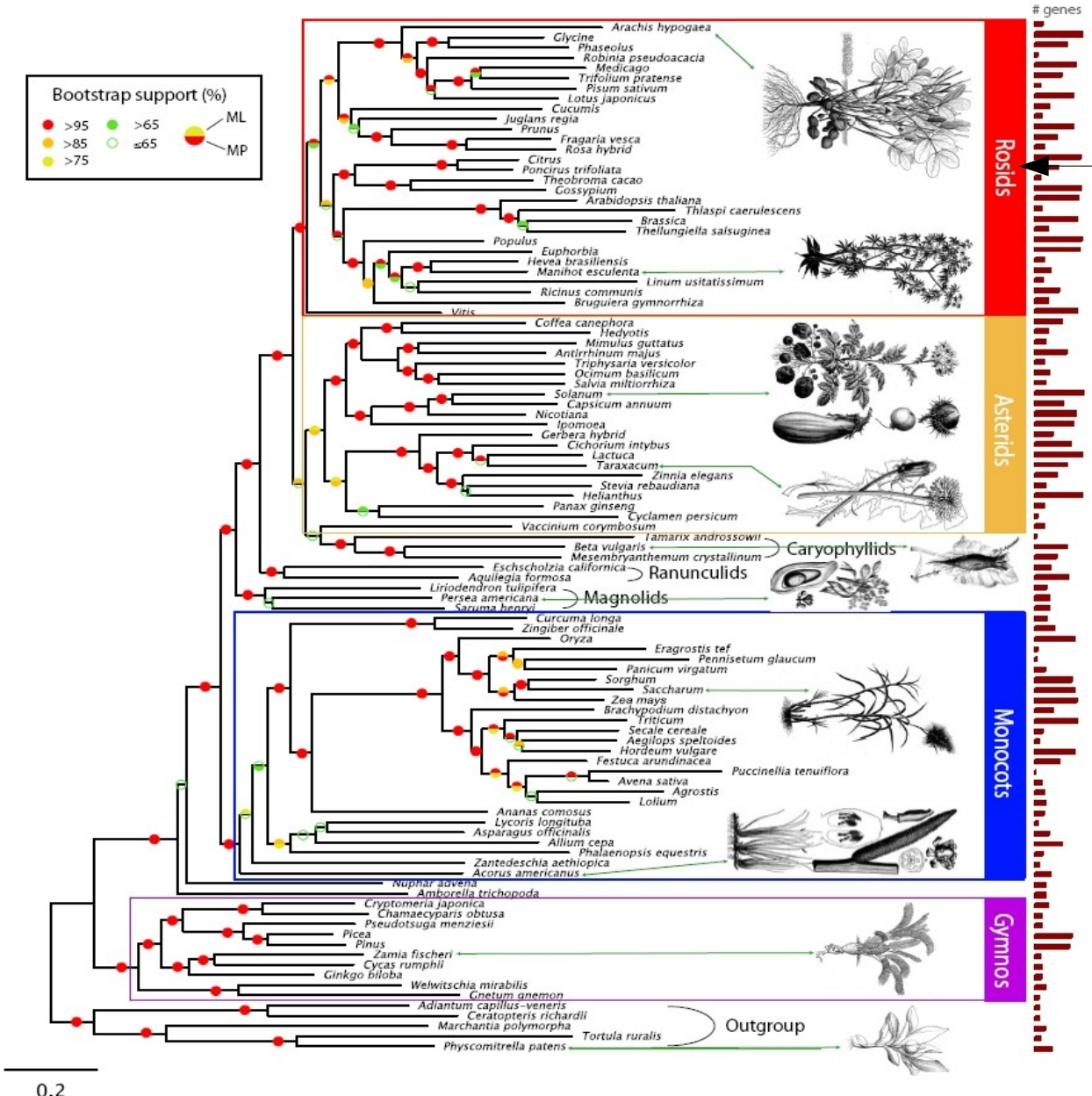
A taxonomic subclass

The ingroup

This tree is unrooted

The outgroup

A Phylogeny or Phylogenetic Tree



In Phylogenetics such a subtree is often also called *Lineage*!

Phylogeny

- An unrooted strictly binary tree
- Leafs are labeled by *extant* “übrig geblieben” (currently leaving) organisms represented by their DNA/Protein sequences
- Inner nodes represent *hypothetical common ancestors*
- *Outgroup*: one or more closely related, but different species → allows to root the tree

Taxon

- Used to denote clades/subtrees in phylogenies or taxonomies
- A group of one or more species that form a biological unit
- As defined by taxonomists

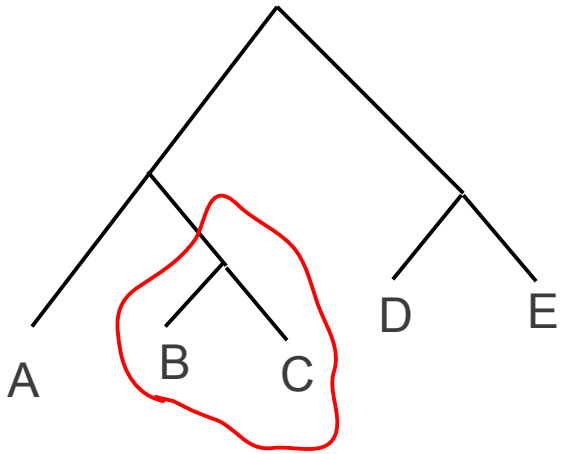
  → subject of controversial debates

  → part of the culture/fuzziness of Biology
- In phylogenetics we often refer to a single leaf as taxon

  → the plural of taxon is *taxa*

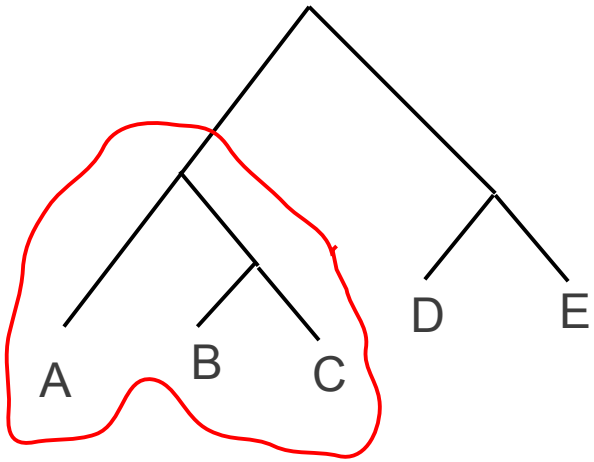
  → we often say that a tree with *n* leaves (sequences) has *n* taxa

Some more terminology



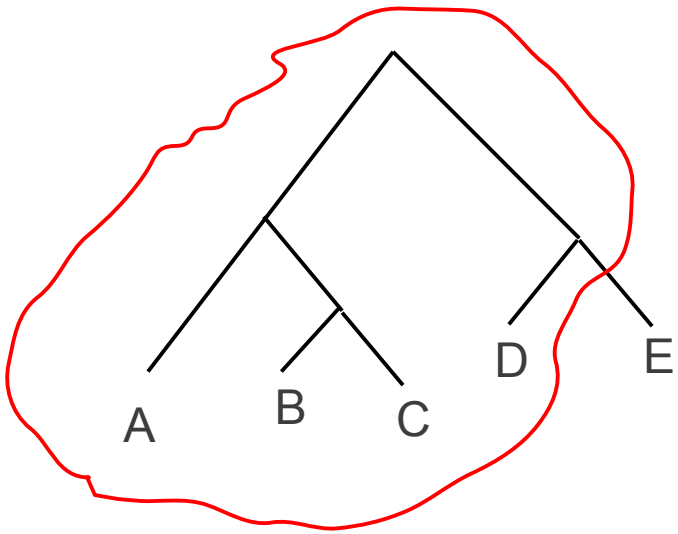
B and C are a *monophyletic* group; they are sister species

Some more terminology



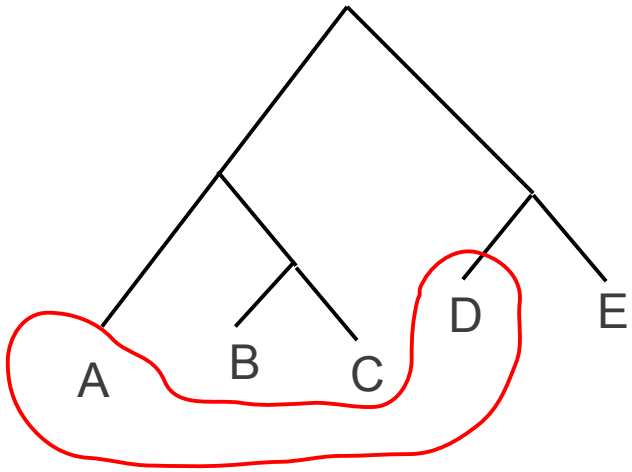
(A,B,C) is a *monophyletic* group; it is are sister to (D, E)

Some more terminology



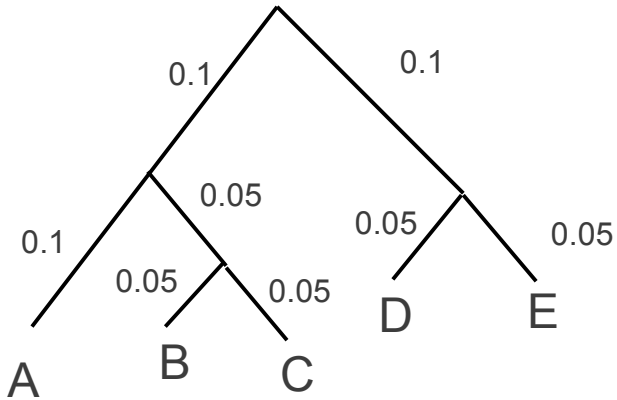
(A,B,C,D) is *paraphyletic* → E is excluded

Some more terminology



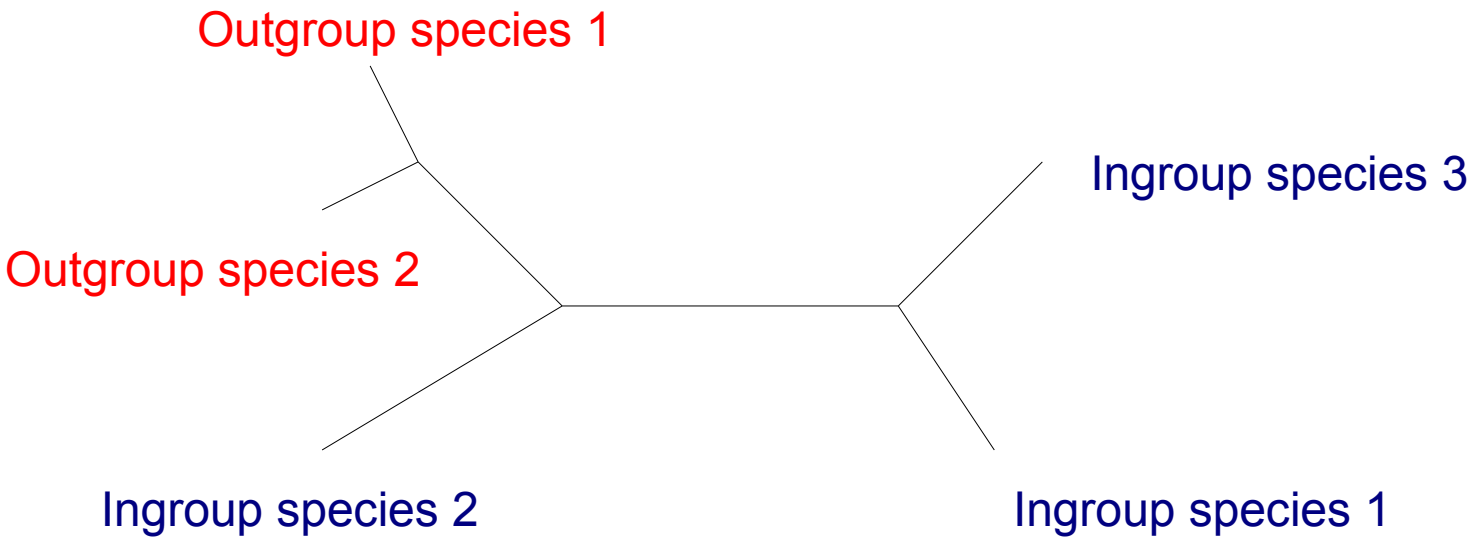
(A,D) is a *polyphyletic* group → their most recent common ancestor (MRCA) is excluded

Some more terminology

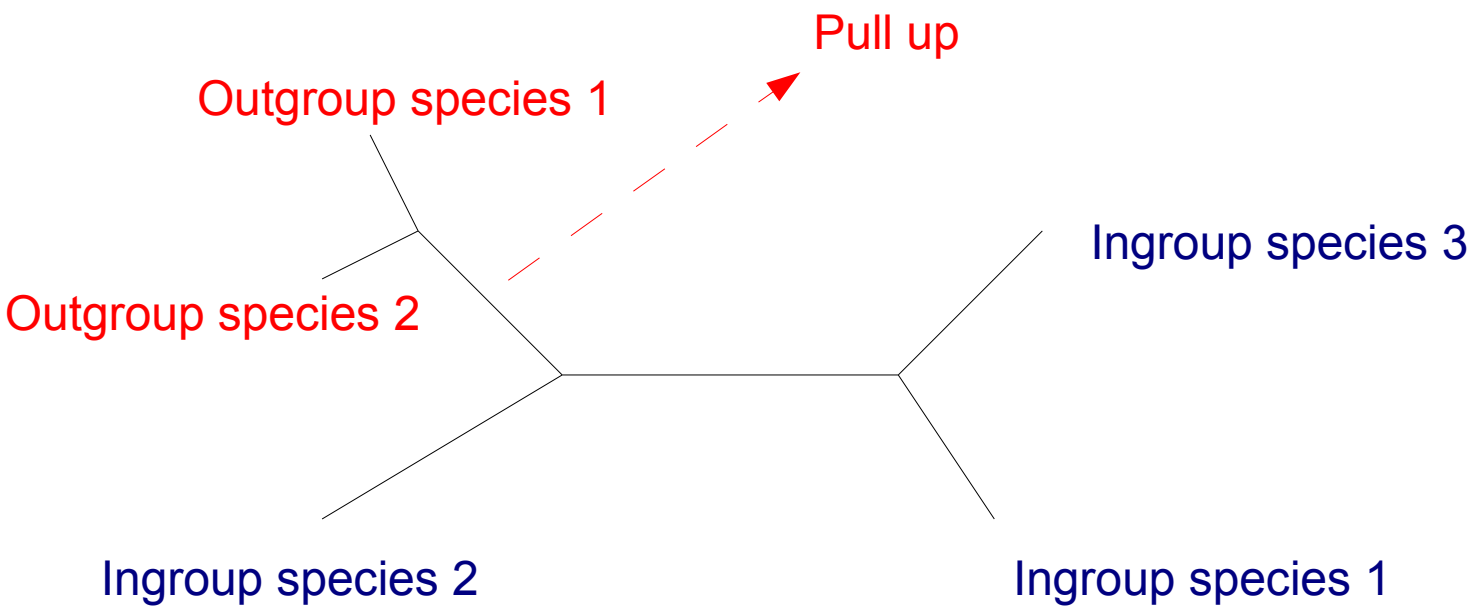


Tree-based or *patristic distance* between two taxa:
Sum over branch lengths along the path in the tree, e.g.:
A ↔ B: 0.2
A ↔ D: 0.35

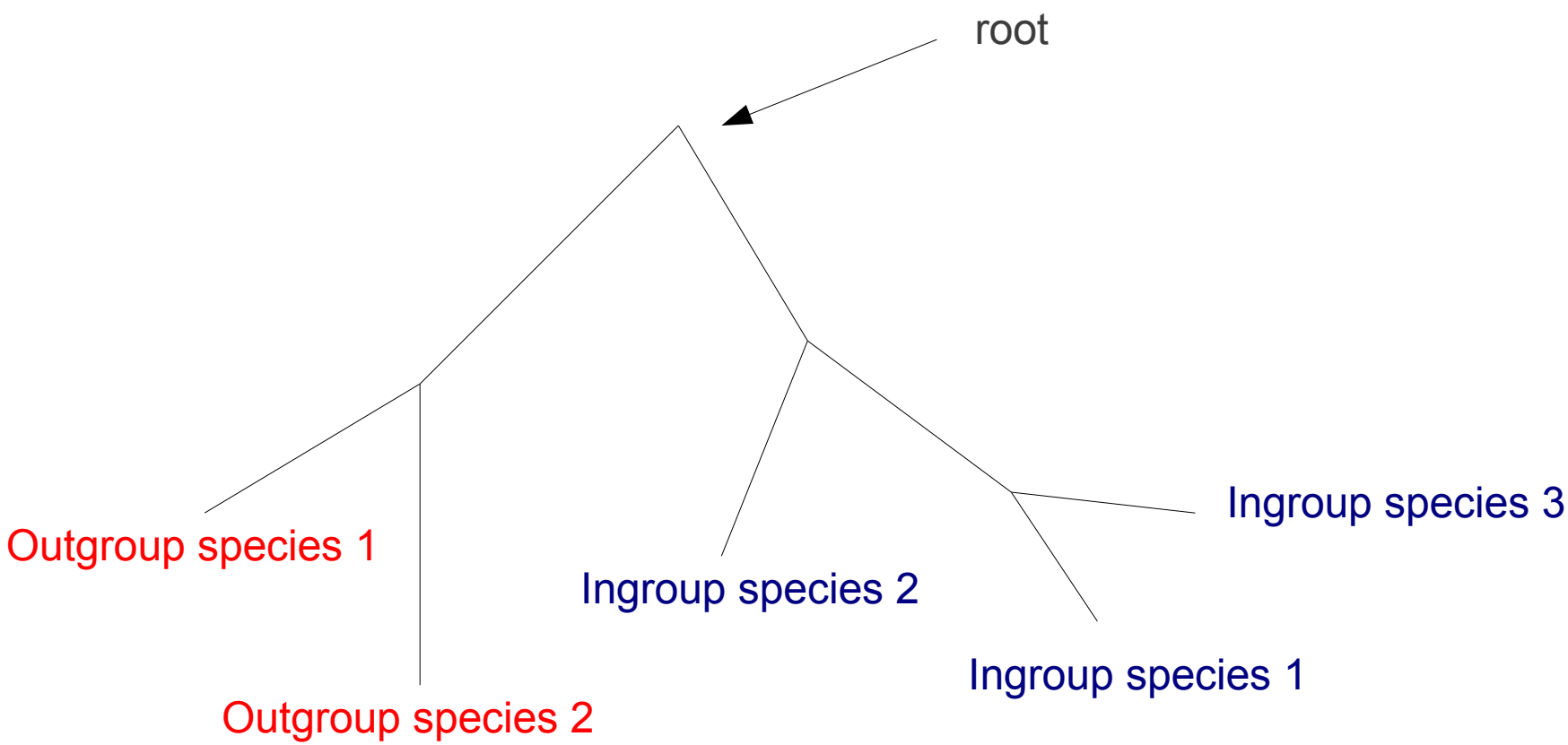
Tree Rooting



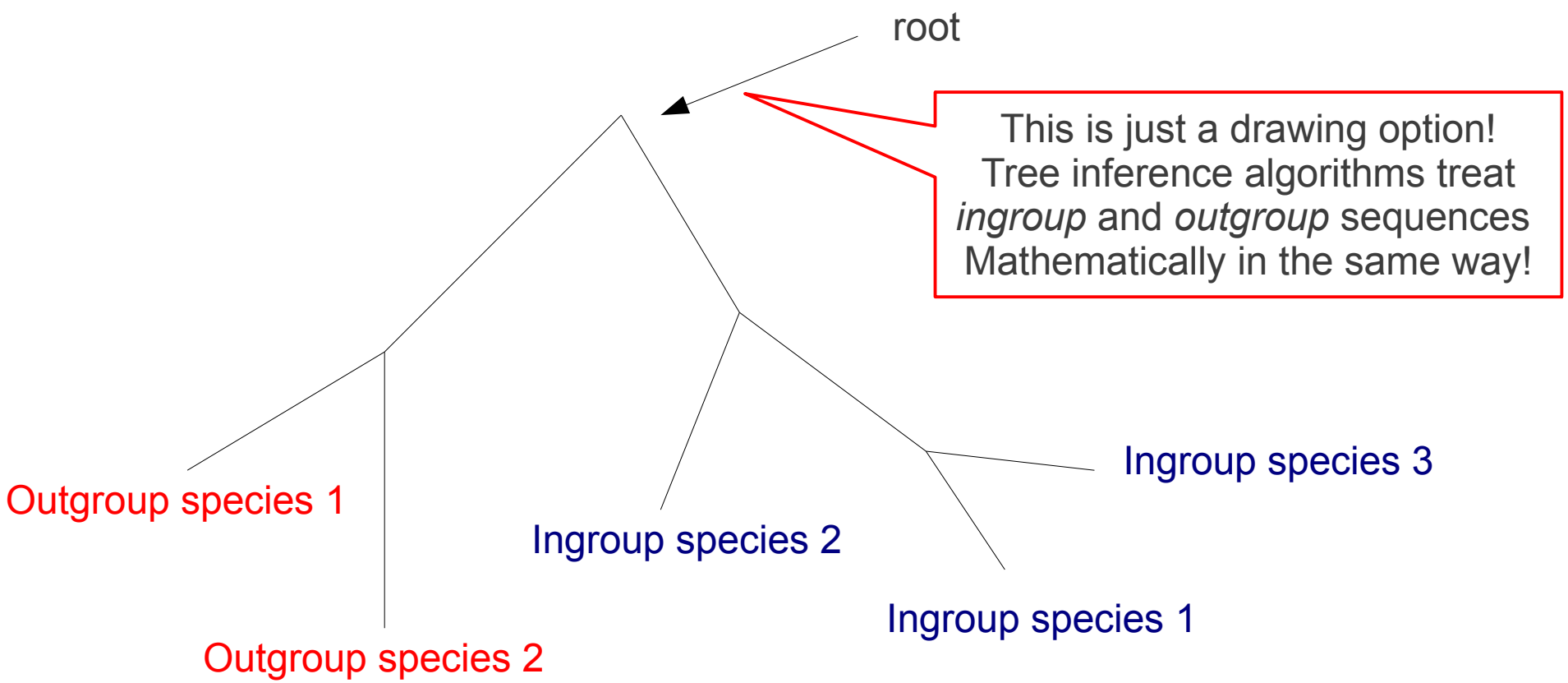
Tree Rooting



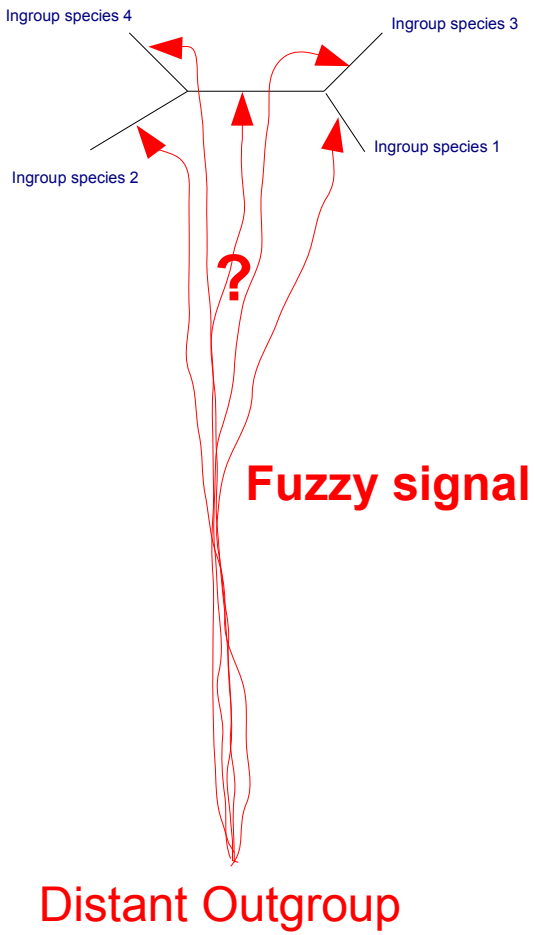
Tree Rooting

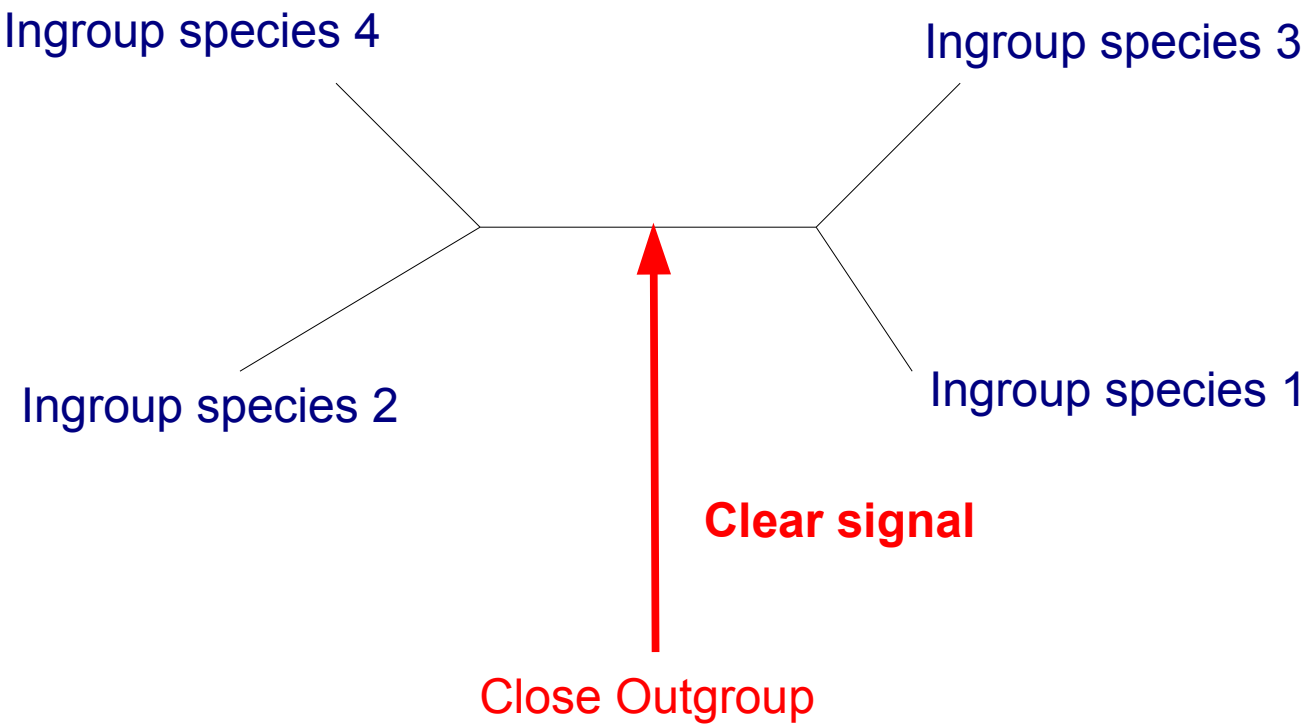


Tree Rooting

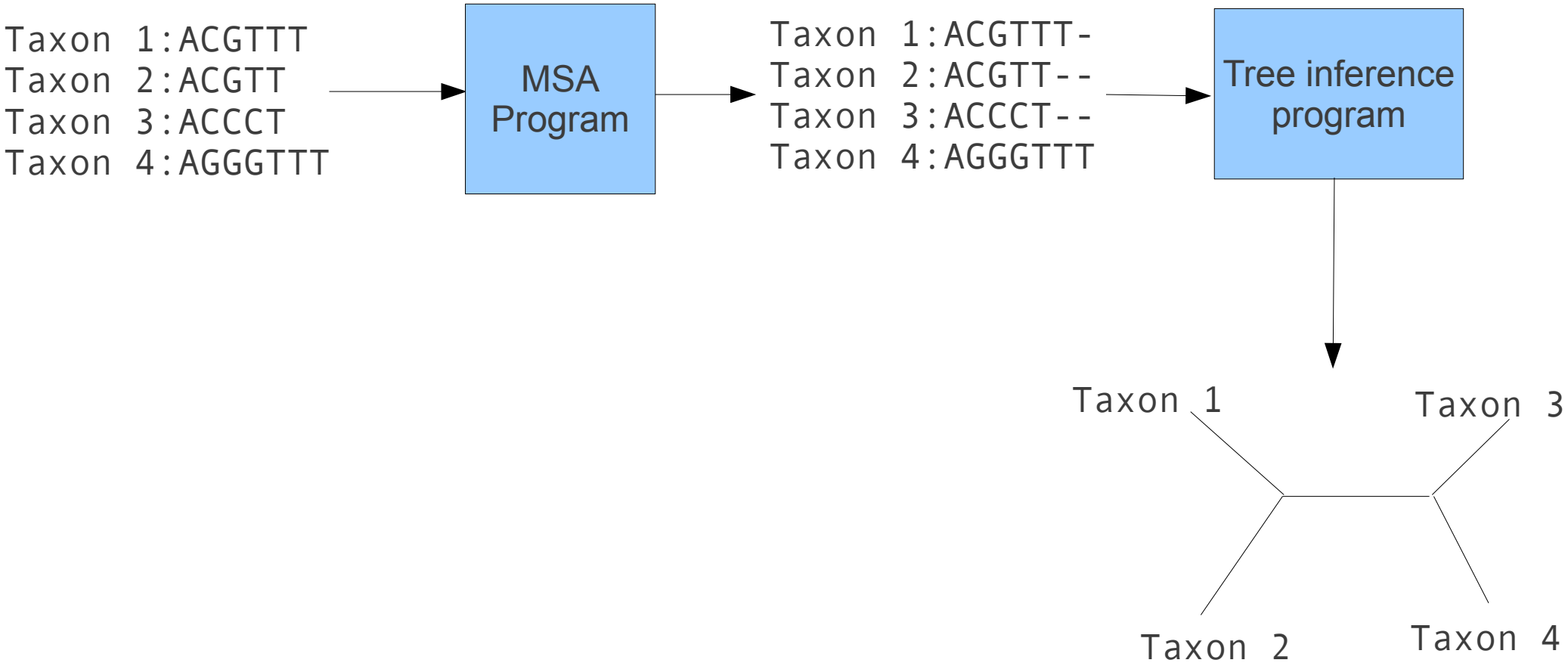


Outgroup Choice





Tree Inference



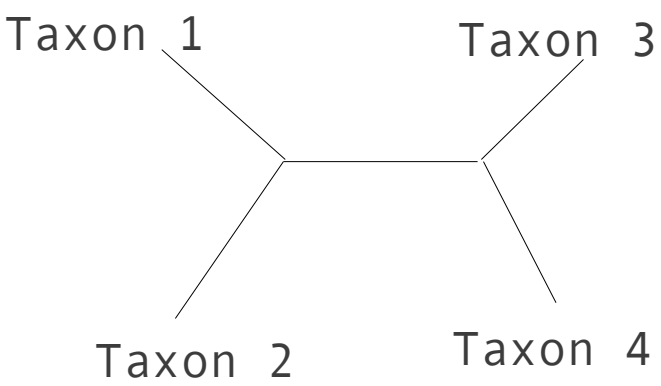
Tree Inference

```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

MSA Program

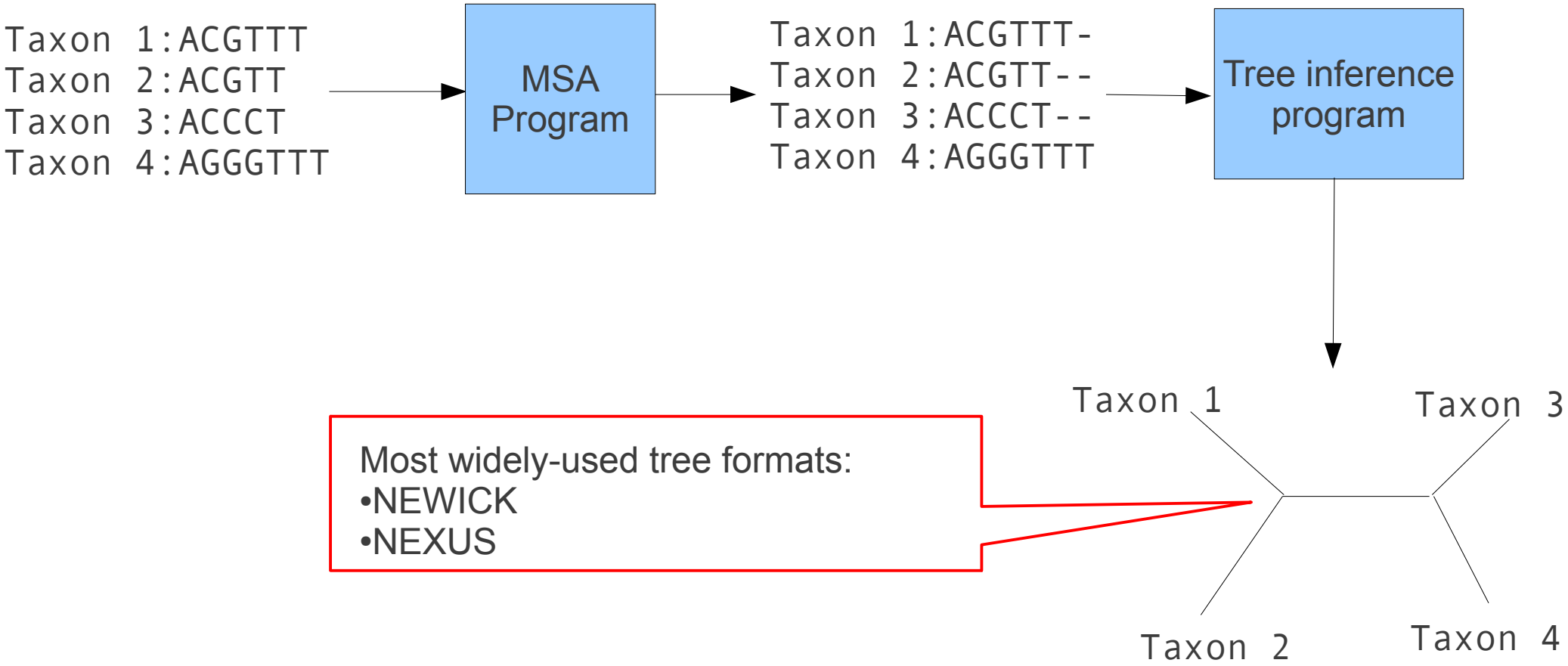
```
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT
```

Tree inference program



Most widely-used alignment formats:
- PHYLIP
- NEXUS
- FASTA

Tree Inference



Most widely-used tree formats:
- NEWICK
- NEXUS

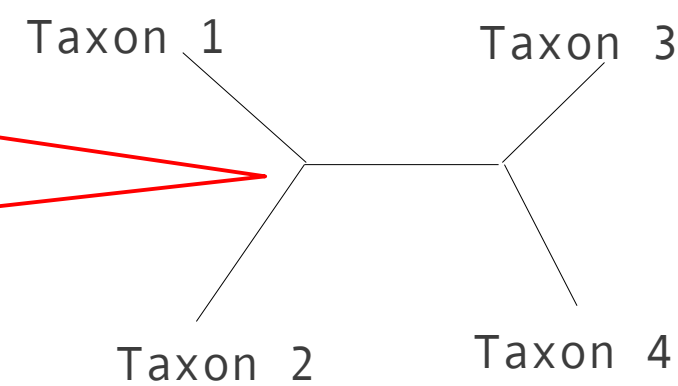
Tree Inference

```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

MSA Program

```
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT
```

Tree inference program



Newick example: Remember that this is an unrooted tree!

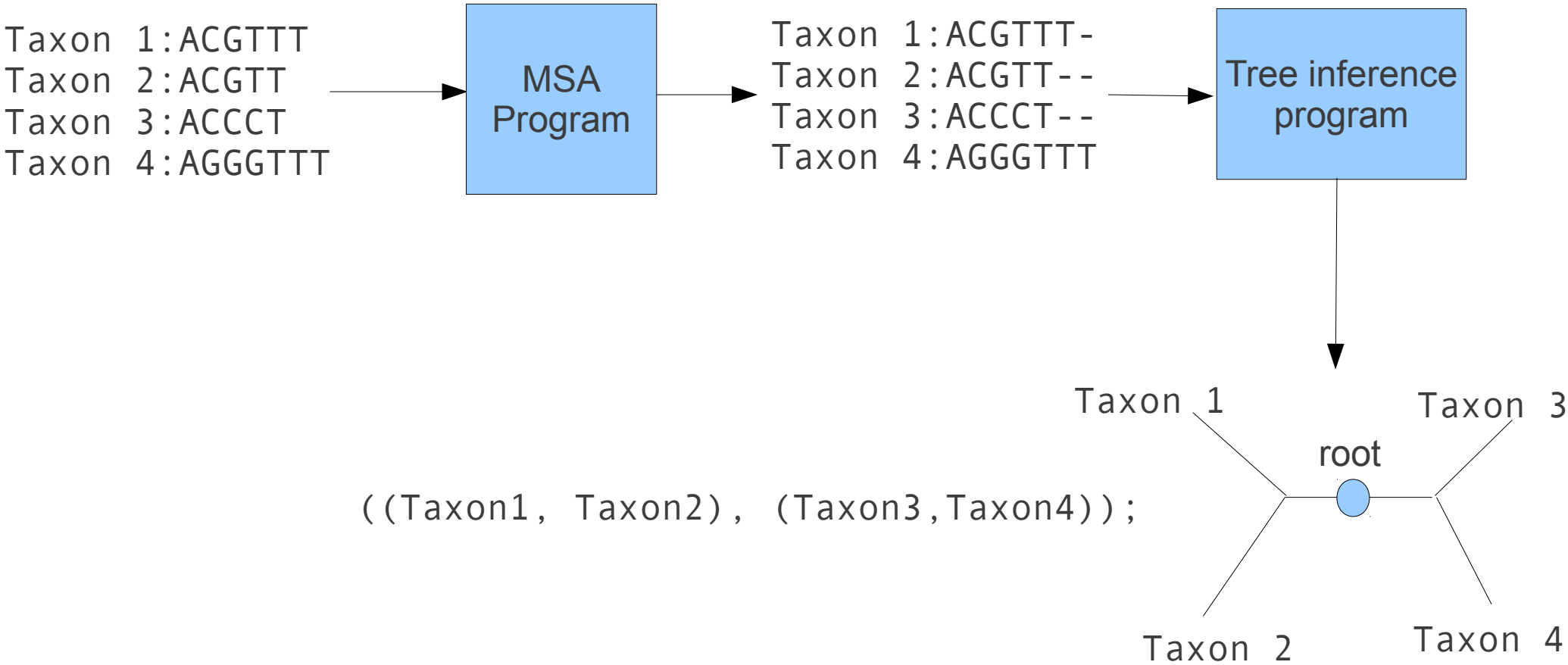
```
(Taxon1, Taxon2, (Taxon3,Taxon4));
```

or

```
((Taxon1, Taxon2), Taxon3,Taxon4);
```

Top level trifurcation

Tree Inference



((Taxon1, Taxon2), (Taxon3,Taxon4));

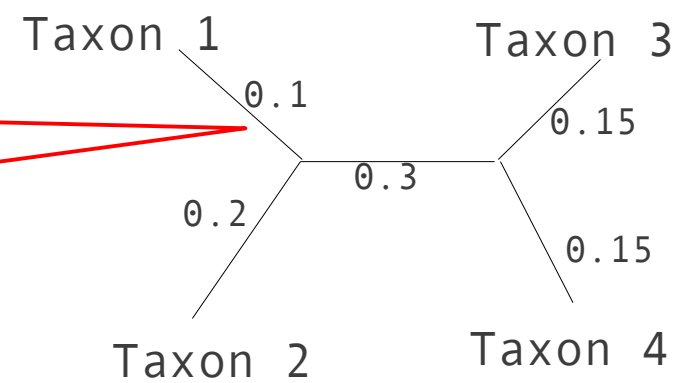
Tree Inference

```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

MSA Program

```
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT
```

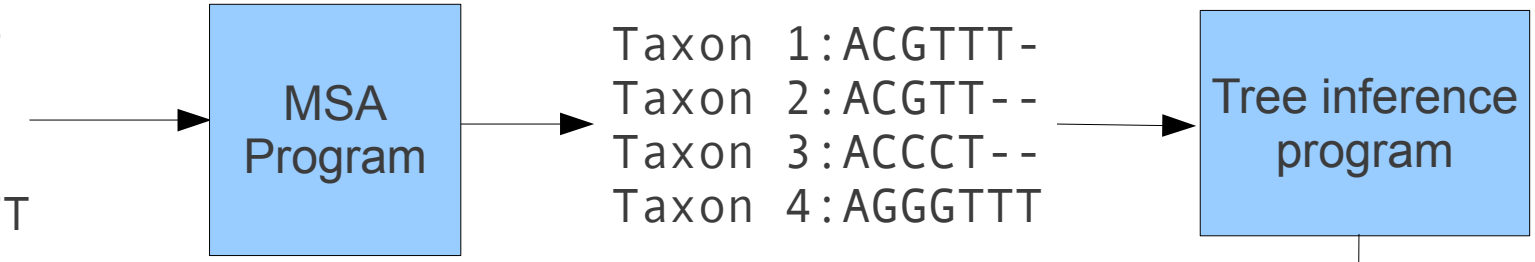
Tree inference program



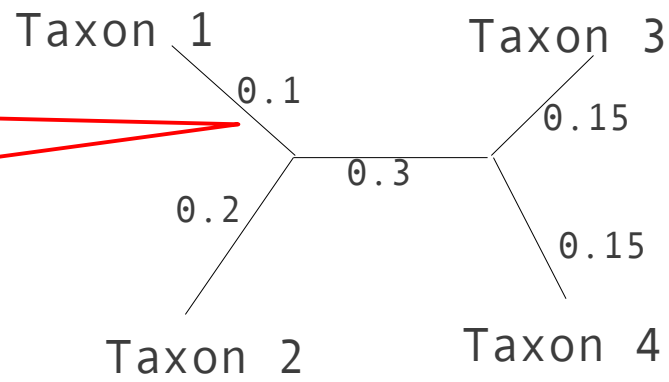
Trees may have ***relative*** branch lengths, depending on the tree inference method that was used

Tree Inference

```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```



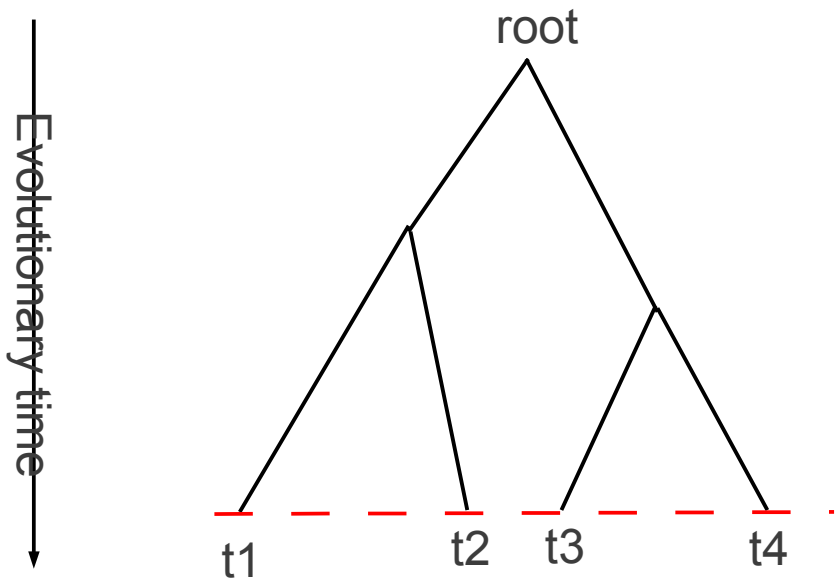

```
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT
```

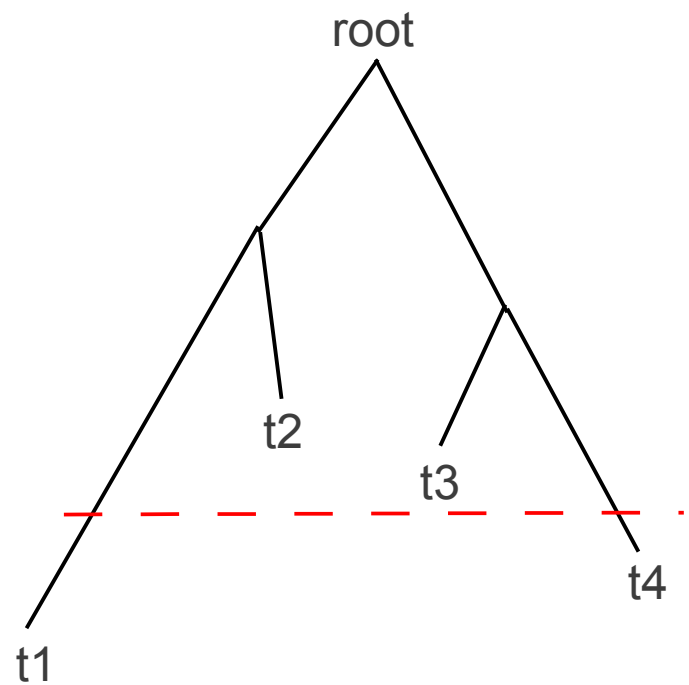
Trees may have ***relative*** branch lengths, depending on the tree inference method that was used

```
(Taxon1:0.1, Taxon2:0.2, (Taxon3:0.15,Taxon4:0.15):0.3);
```

Tree Shapes

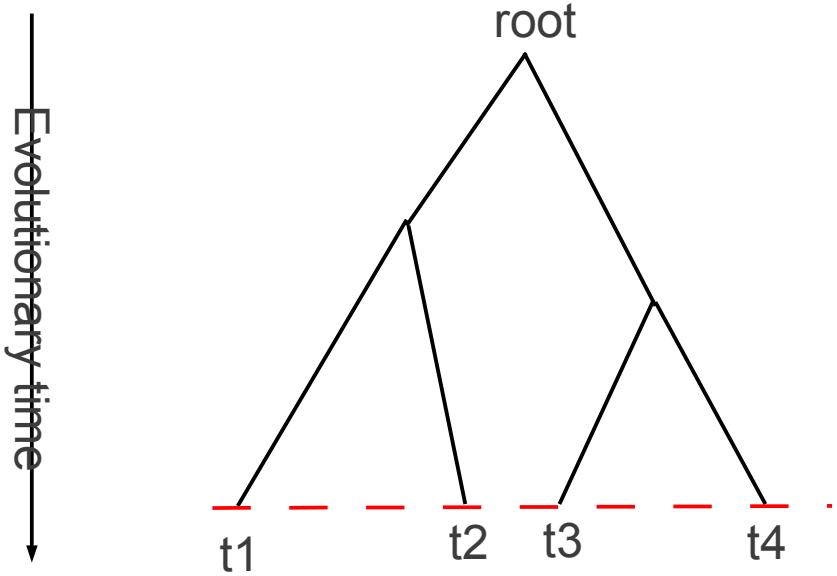


Ultrametric tree

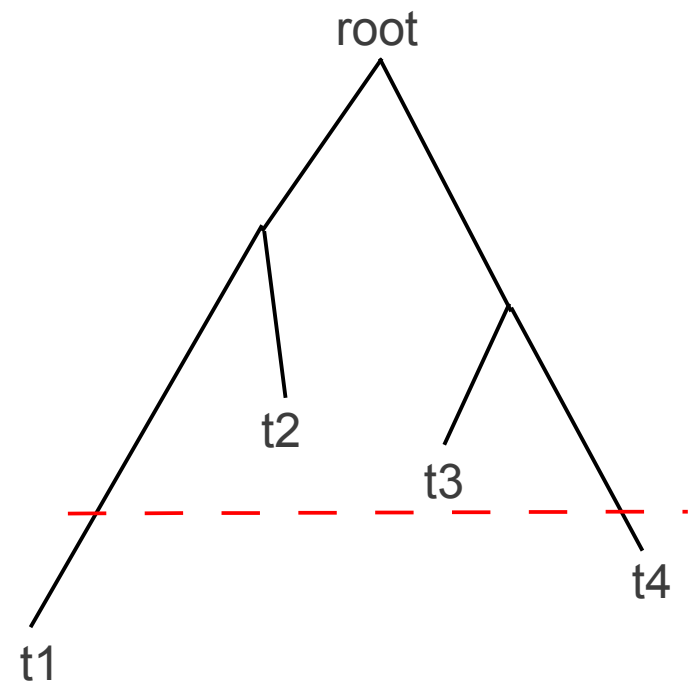


Non-ultrametric tree

Tree Shapes



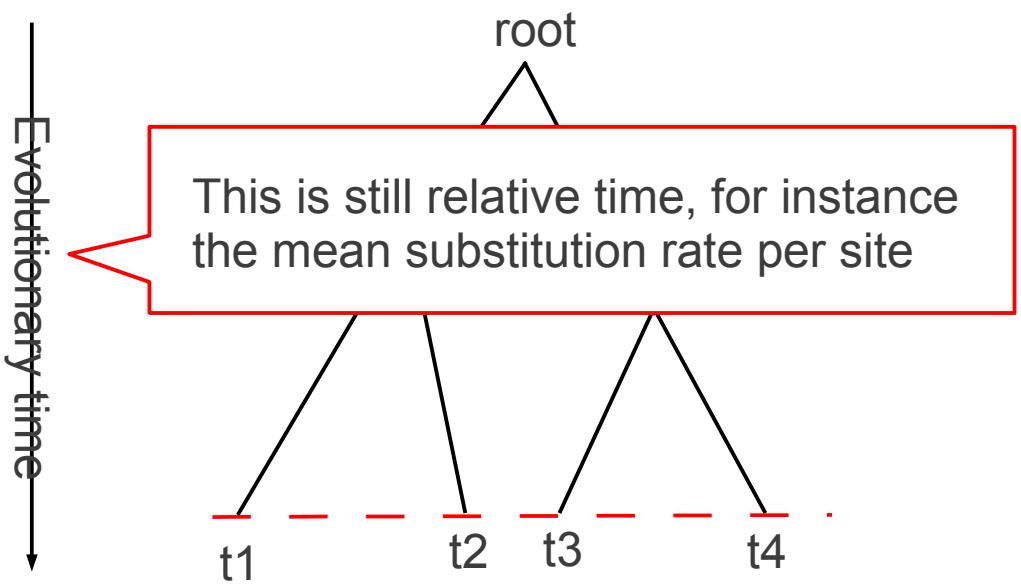
Ultrametric tree



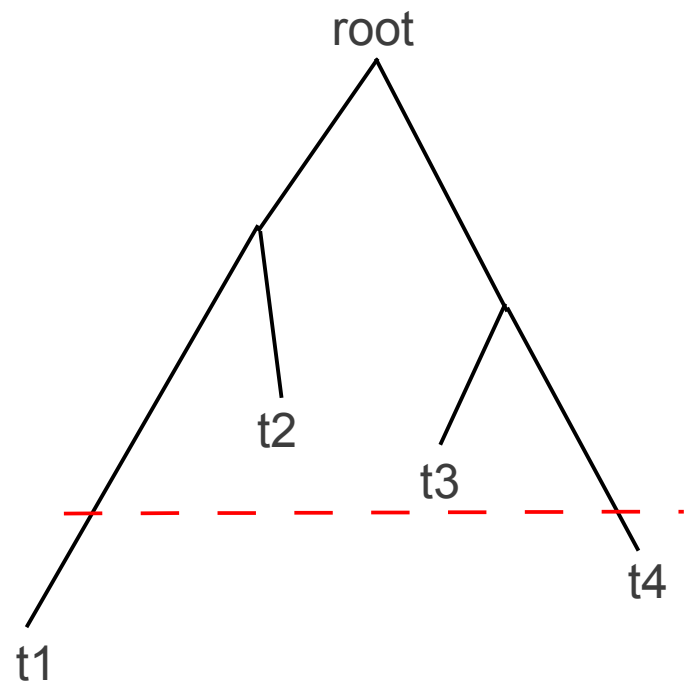
Non-ultrametric tree

Most tree inference models/algorithms/programs produce non-ultrametric trees

Tree Shapes

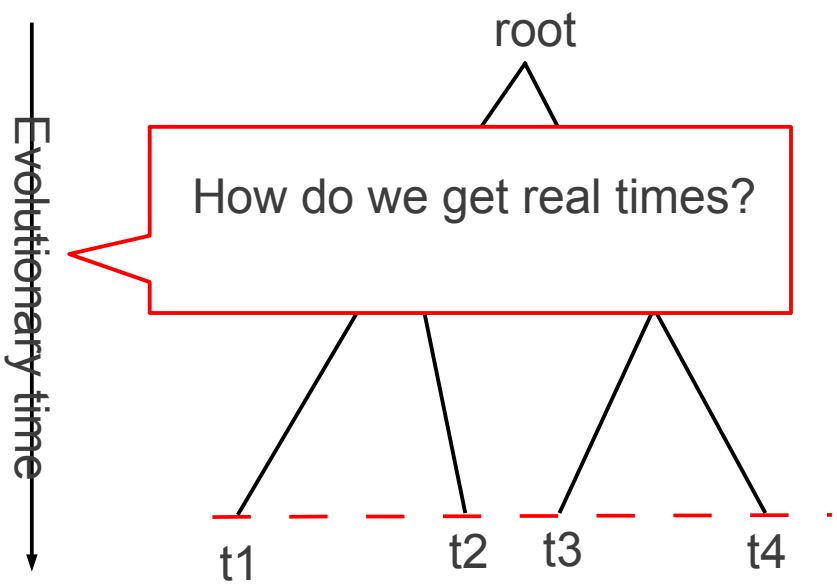


Ultrametric tree

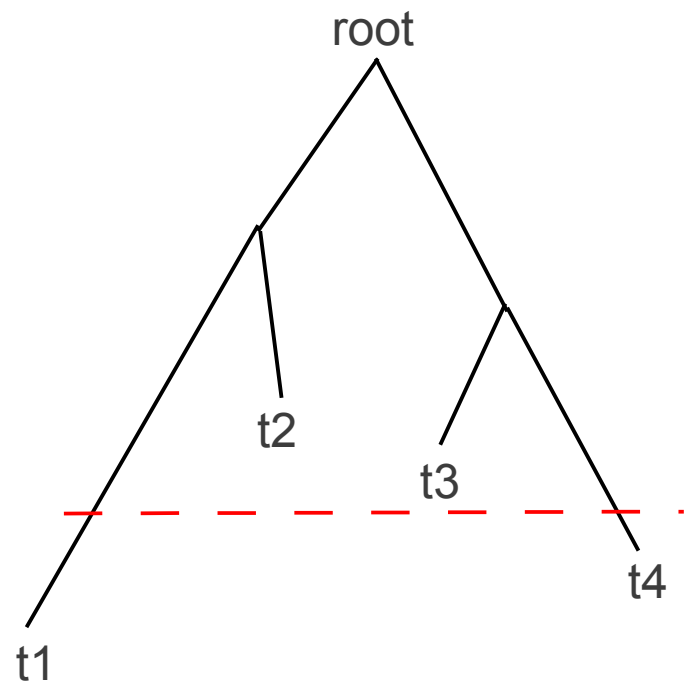


Non-ultrametric tree

Tree Shapes

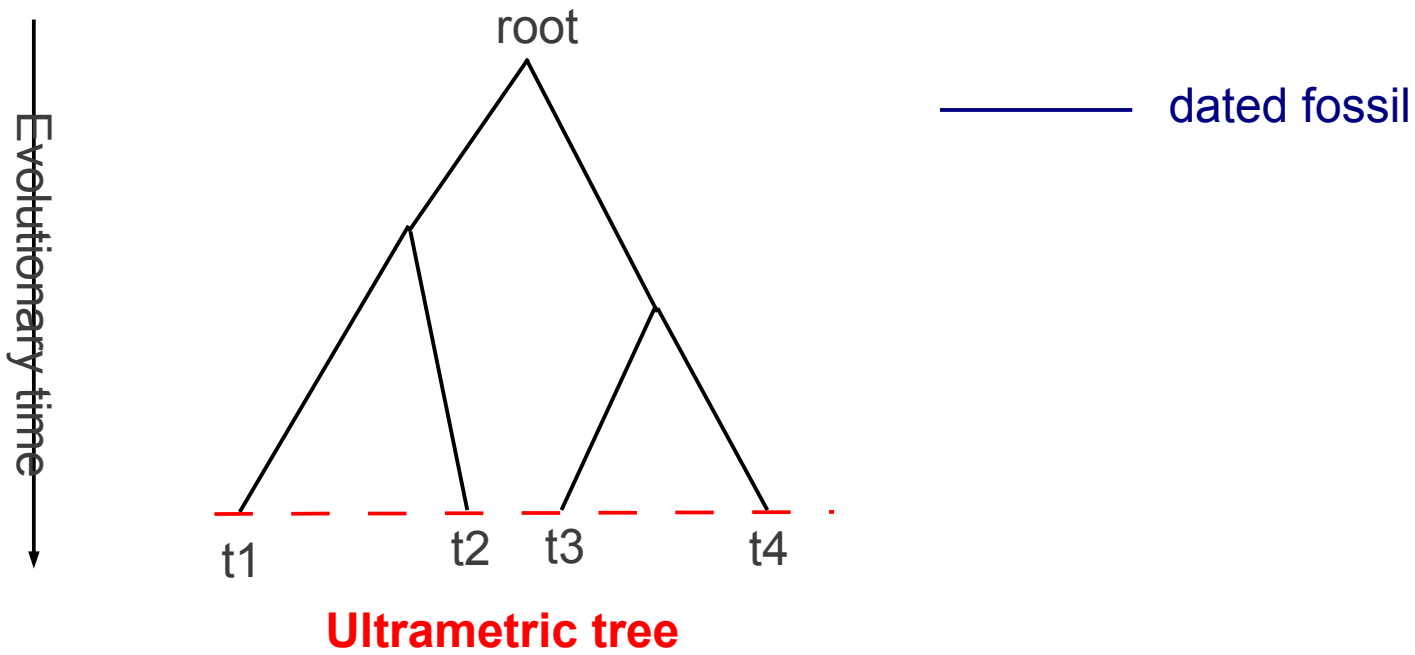


Ultrametric tree

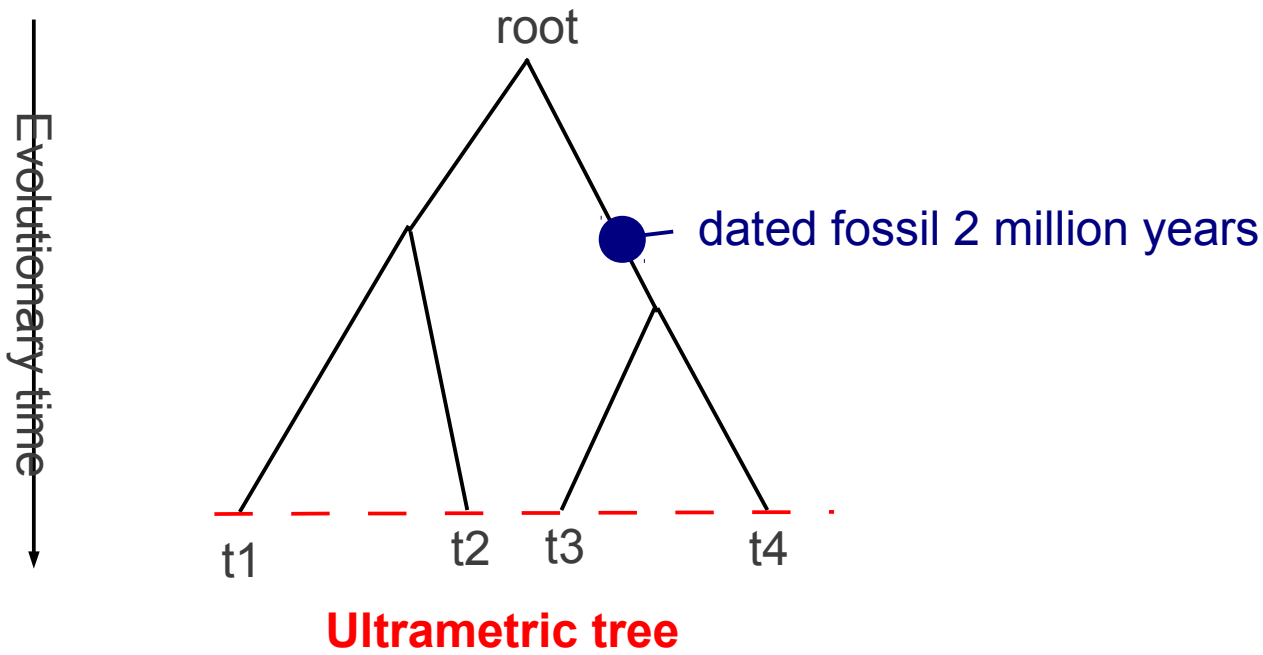


Non-ultrametric tree

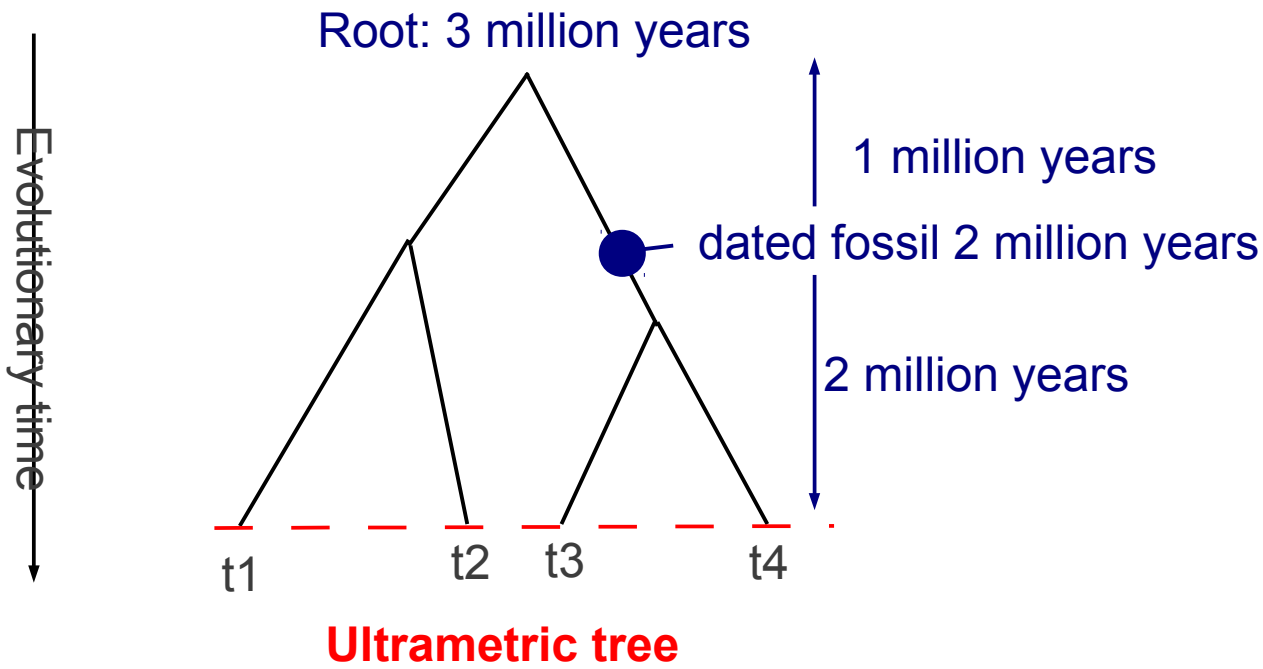
Dating Trees



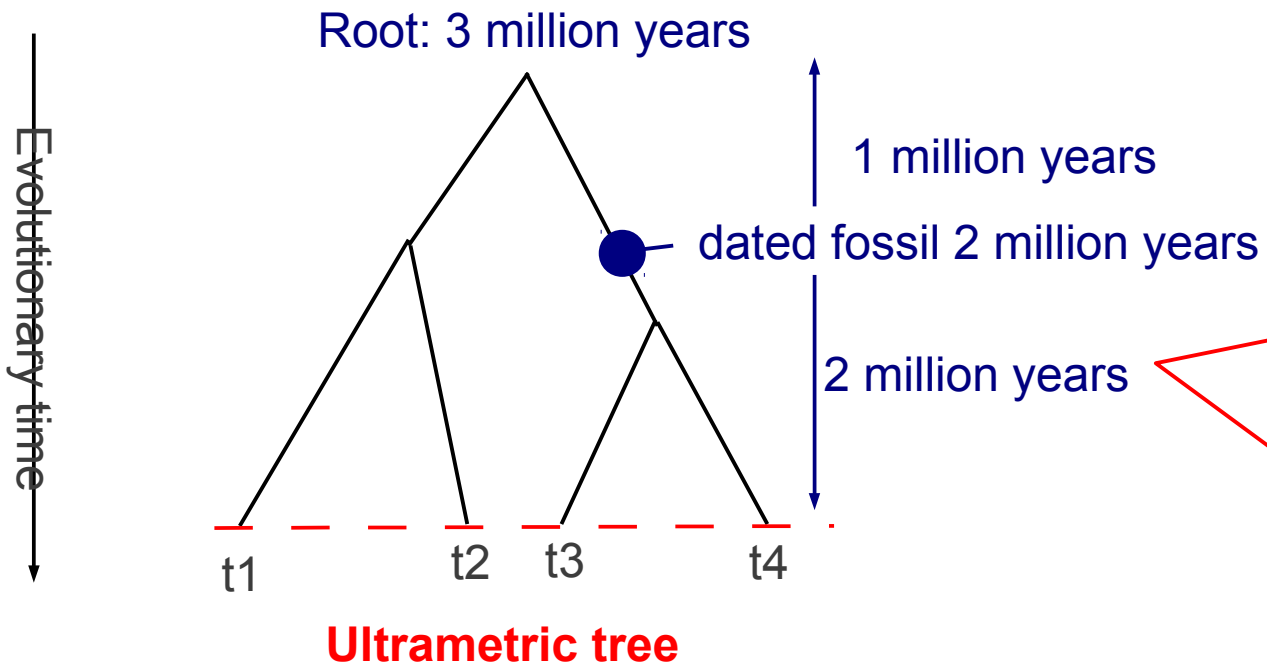
Dating Trees



Dating Trees

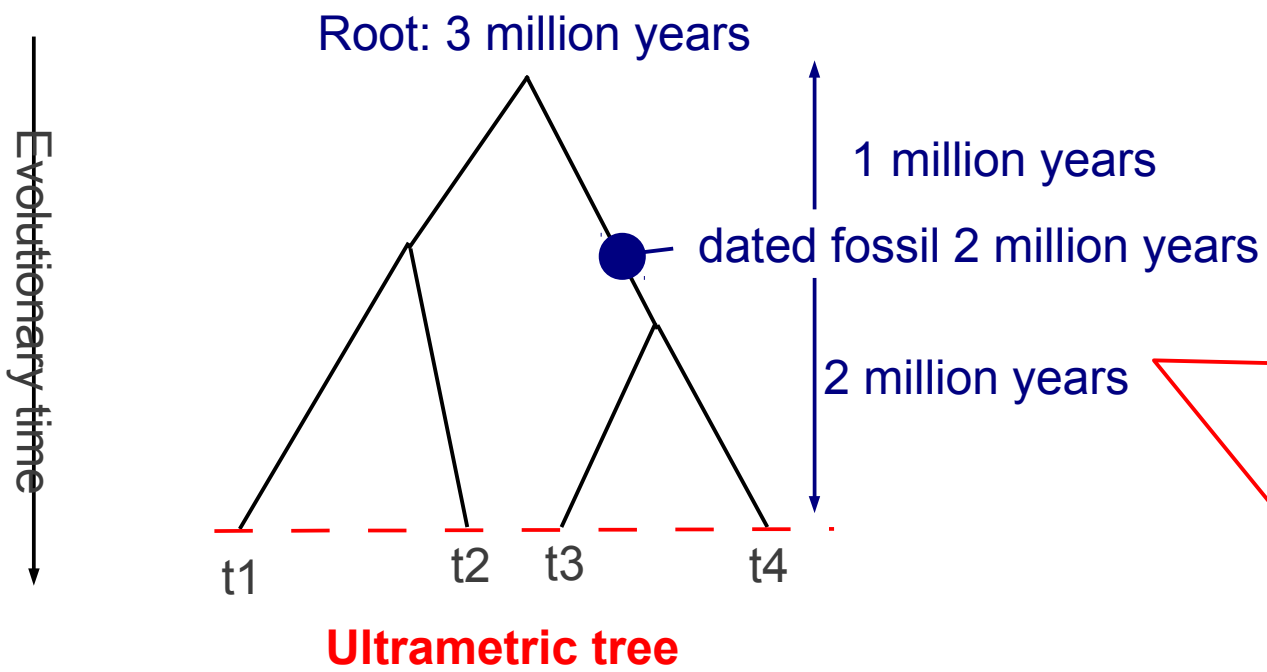


Dating Trees



We need a rooted & ultrametric tree!
→ rooting with outgroups
→ ultrametricity with programs for divergence time estimation
→ active research area
→ e.g., DPPDIV
→ most codes rely on the phylogenetic likelihood function and Bayesian stats.

Dating Trees



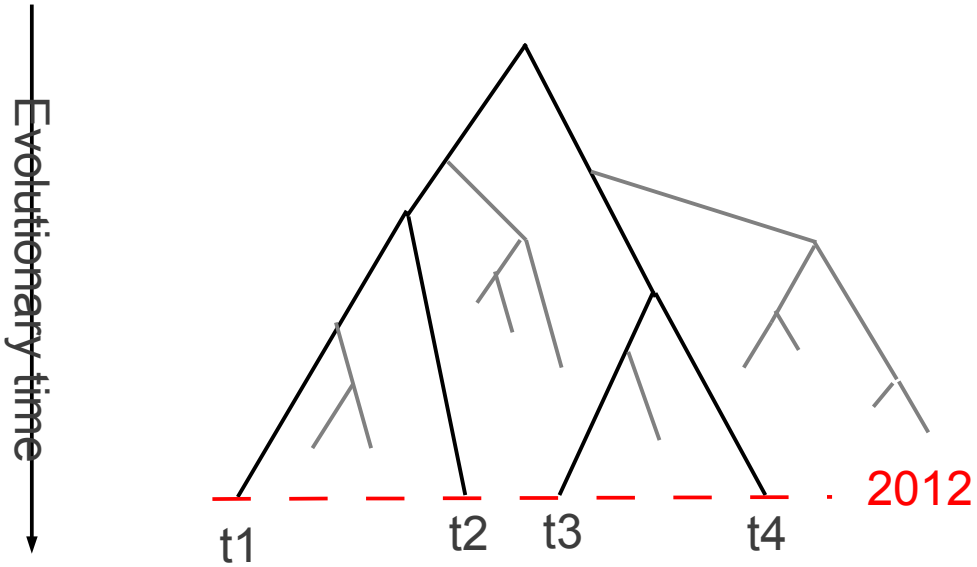
But how do we place the fossil?
→ typically no DNA data available

Fossil placement:
→ ad hoc using empirical knowledge
→ computationally using morphological data

The input for a phylogenetic analysis need not be molecular data!

We can also use sequences of morphological traits ("Merkmale")!

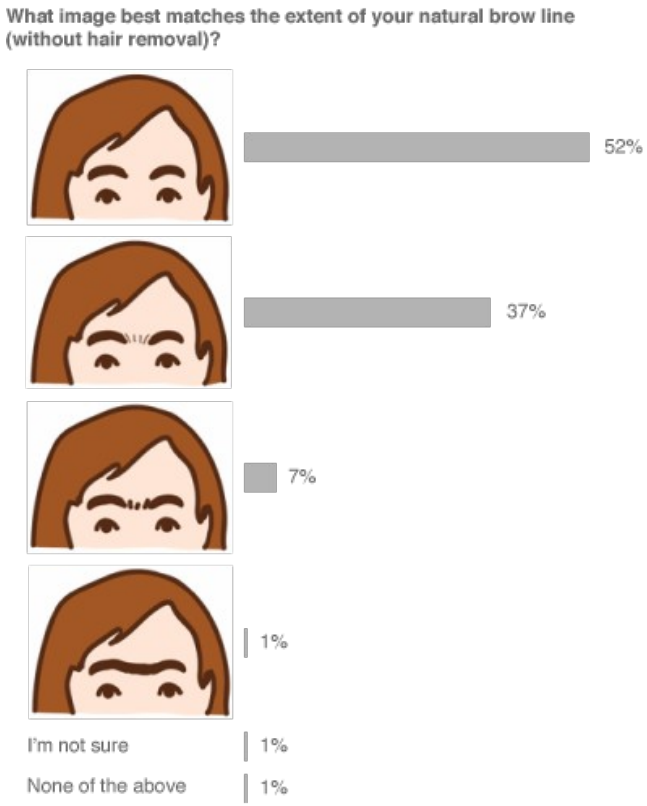
Remember that we deal with extant species!



Ultrametric tree

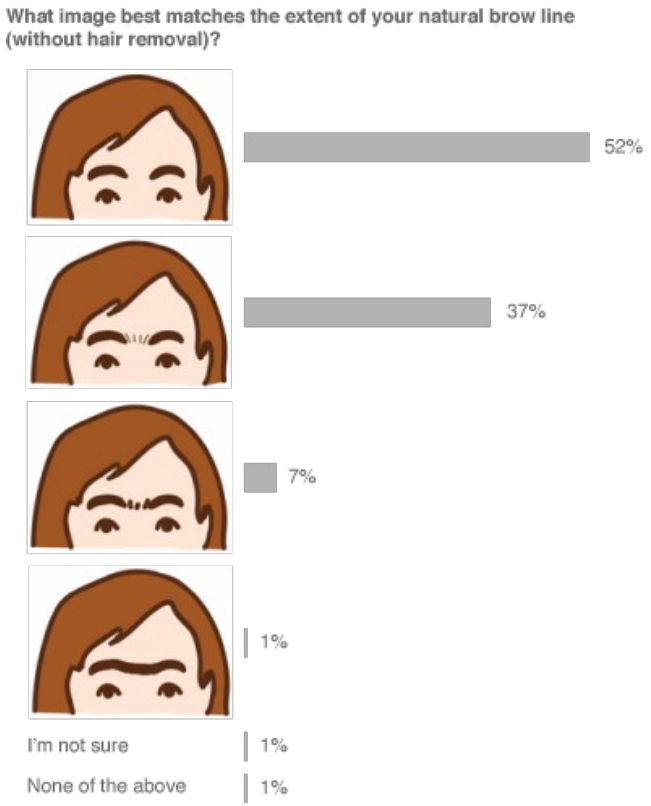
Morphological Traits

```
t1: 1000
t2: 0100
t3: 0010
T4: 0001

or:

t1: 0
t2: 1
t3: 2
t4: 3
```



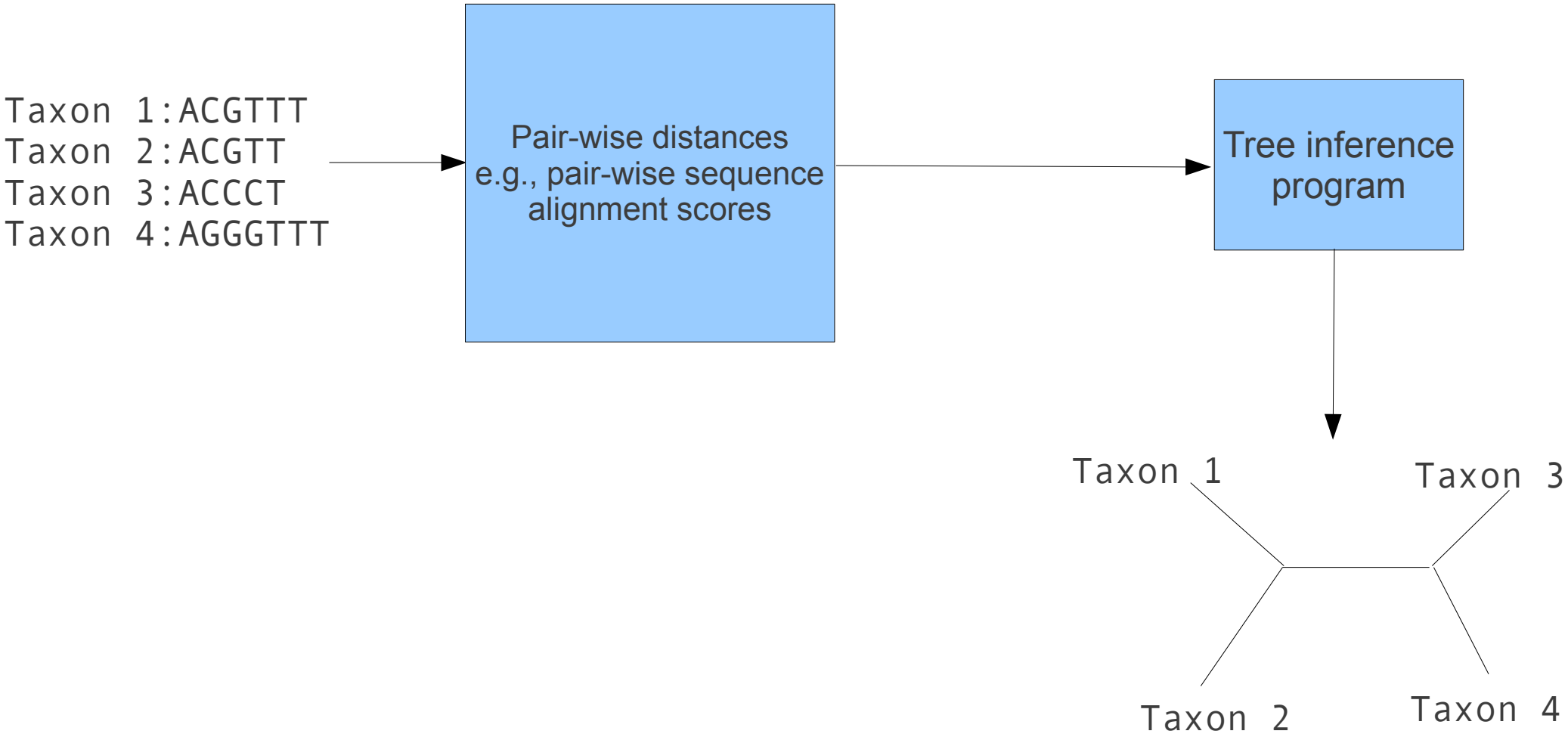
Morphological Traits

```
t1: 1000
t2: 0100
t3: 0010
T4: 0001

or:

t1: 0
t2: 1
t3: 2
t4: 3
```

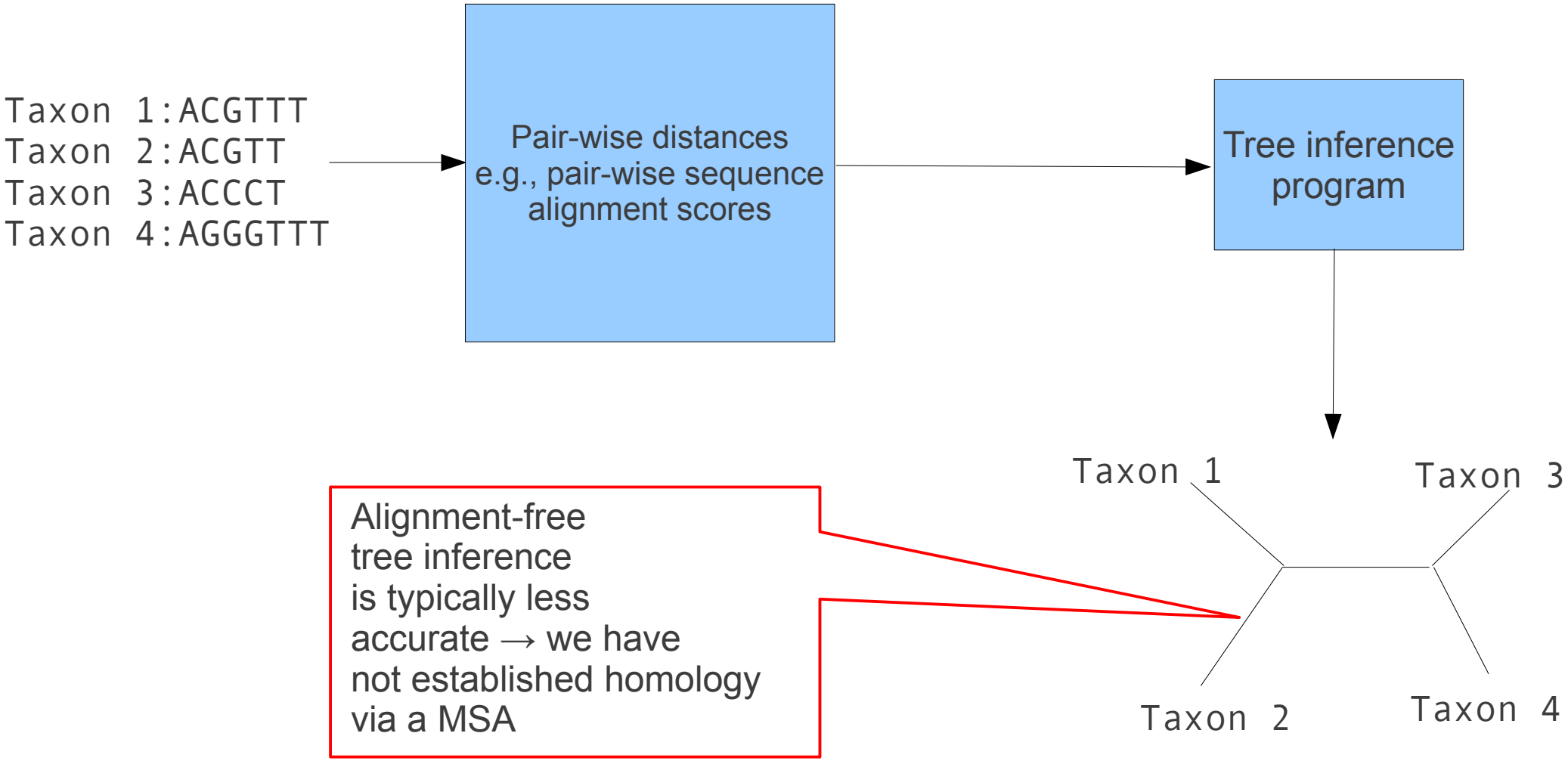


Traits need not be discrete: they can also be continuous: e.g. bone ratios

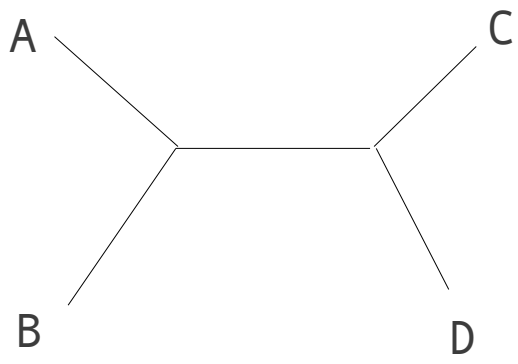
Alignment-Free Tree Inference

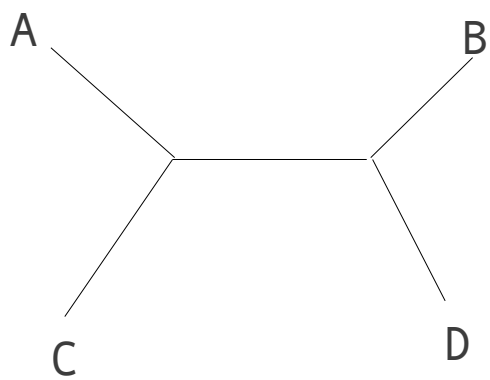


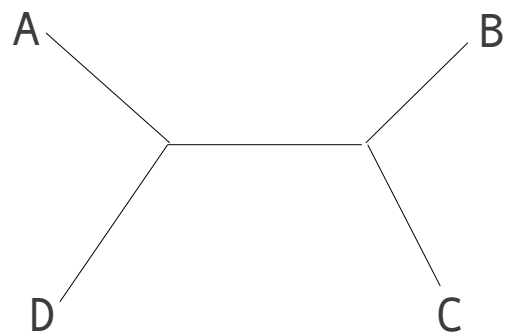
Alignment-Free Tree Inference



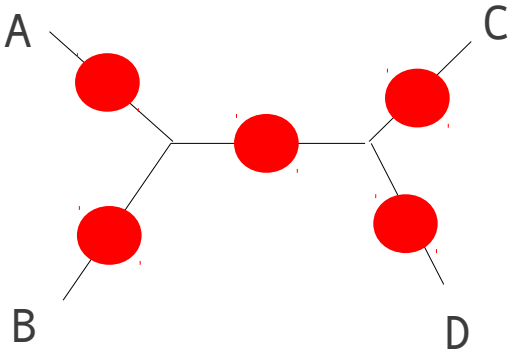
How many unrooted 4-taxon trees exist?

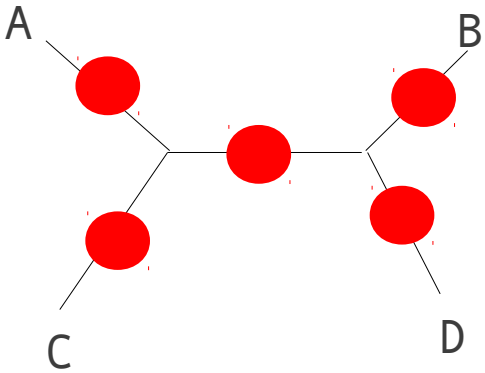


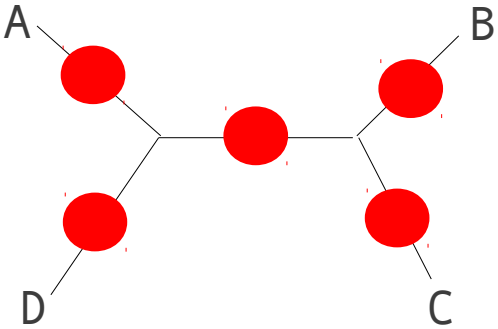




How many rooted 4-taxon trees exist?



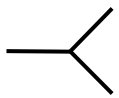




Tree Counts

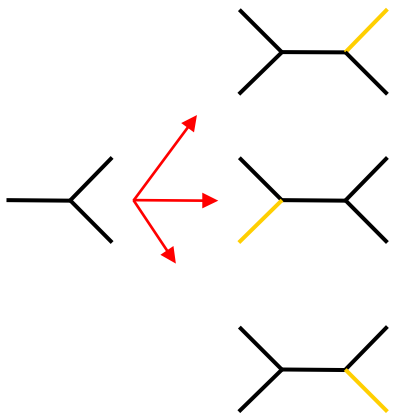
- Unrooted binary trees
  - 4 taxa → 3 distinct trees
  - A tree with *n* taxa has *n-2* inner nodes
  - And *2n-3* branches
- Rooted binary trees
  - 4 taxa → 3 unrooted trees * 5 branches each (rooting points) = 15 trees
  - *n-1* inner nodes
  - *2n-2* branches

The number of trees



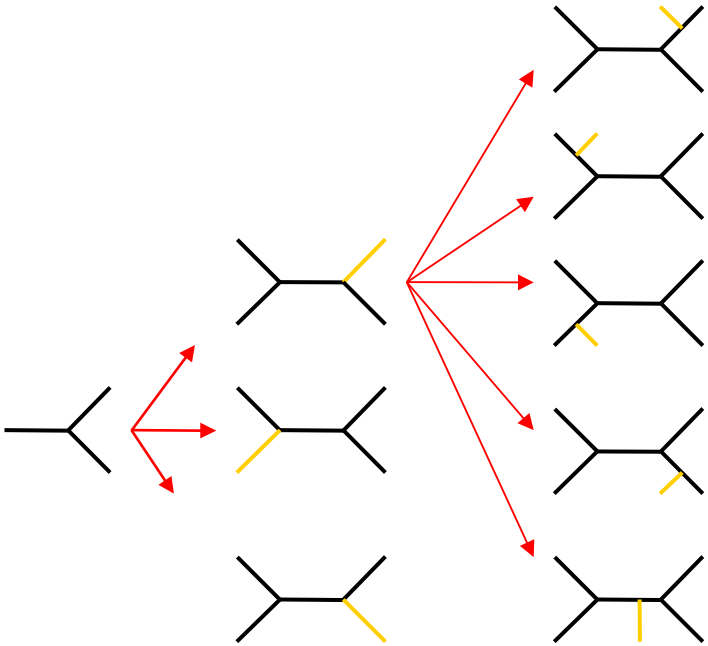
3 taxa = 1 tree

The number of trees



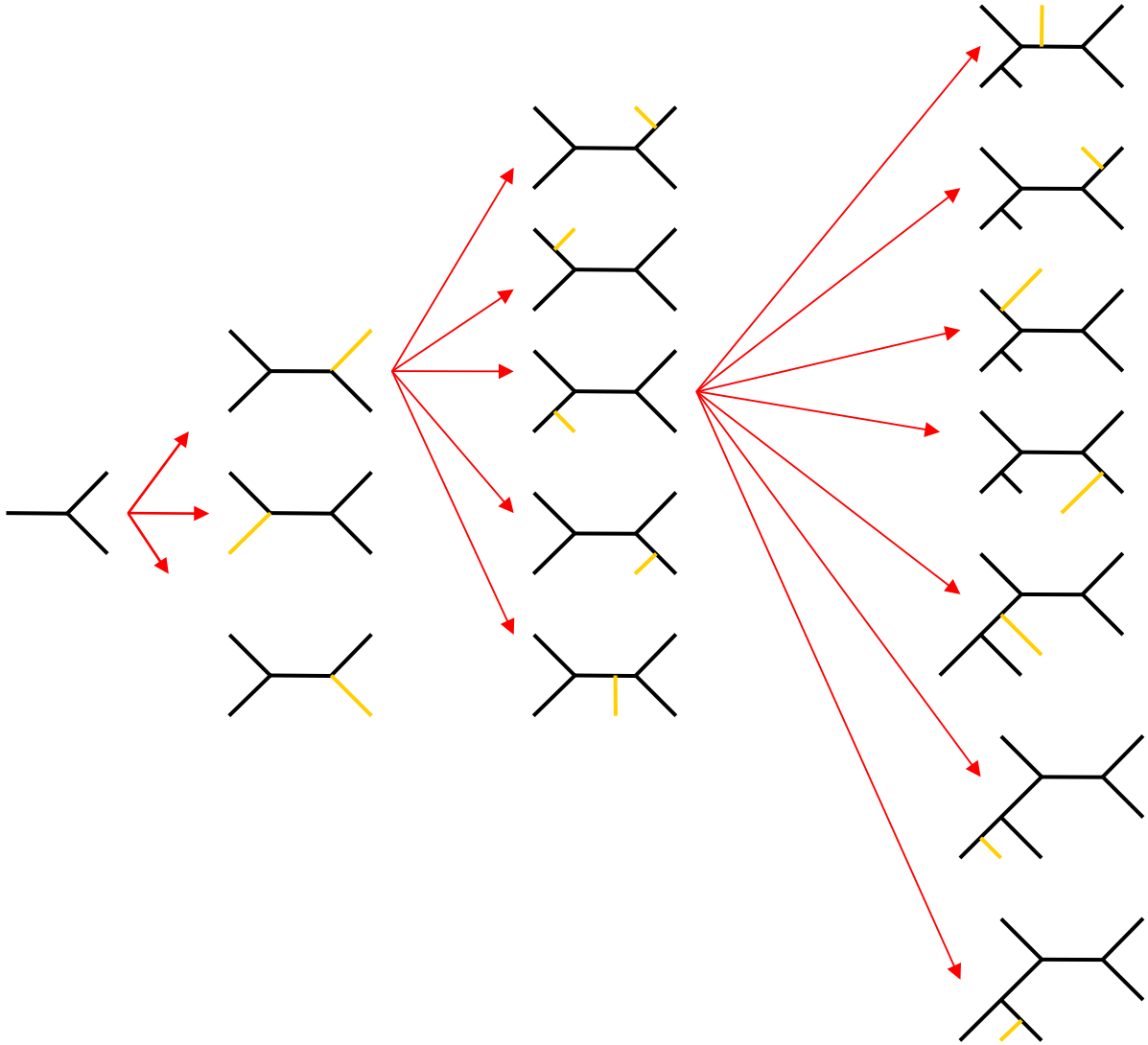
4 taxa: 3 trees
u: # trees of size 4-1 := 1
V: # branches in a tree of size 4-1 := 3
Number of unrooted binary trees with 4 taxa: u * v = 3

The number of trees



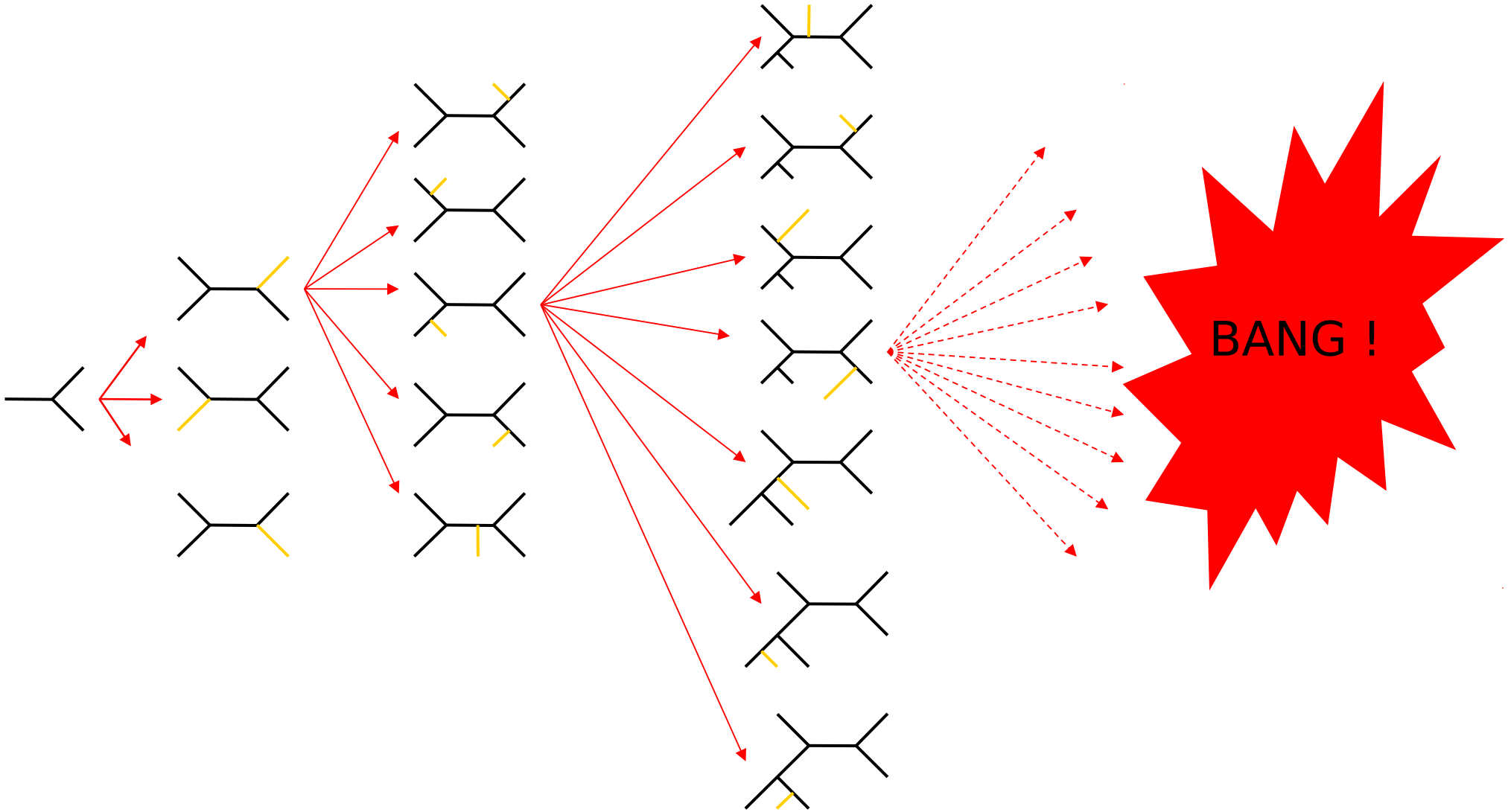
5 taxa: 15 trees
u = 3
v = 5
Number of unrooted trees with 5 taxa: 3 * 5 = 15

The number of trees



6 taxa: 105 trees
u = 15
v = 7
u * v = 105

The number of trees explodes!



Some Numbers

Number of Organisms	Number of alternative Trees
3	1
4	3
5	15
6	105
7	945
10	2.027.025
15	7.905.853.580.625
20	$2.21 * 10^{20}$
50	$2.84 * 10^{76}$

Table 2.1: Number of possible trees for phylogenies with 3–50 organisms

Equation for the number of unrooted trees

- Simple proof via induction

$$\prod_{i=3}^{n}(2i-5)$$

- The number of rooted trees for *n* taxa simply is the number of unrooted trees for *n+1* taxa
- The additional ($n+1^{th}$) taxon represents all possible rootings for the # of unrooted trees with *n* taxa

trees with 2000 tips

```
stamatak@exelixis:~/Desktop/GIT/TreeCounter$ ./treeCounter -n 2000


GNU GPL tree number calculator released June 2011 by Alexandros Stamatakis

Number of unrooted binary trees for 2000 taxa: 30049638174211656151632910065681814981377232074237013089504954043012636525258308210827685996688247000464352735214265634288295
89150234460006314939691306329704360561848618774654822779912235368092334555631999108345976931267565250128998674331877528114019609916315223670306091217357097623798477054676677
79532479718261438527333822672778425073725284991666968758440351057958702068650581768704466631812374290102143850643247136093449166702113596975694030066625264647926912455103149
42366195542824118277625114848758254581227914289801132648902674033761294712745767036267579086843169660718609847941818865957214557044744572288661729053583520744253688123124
01066131569488619609411956467362003425752413352775750858291610964225757276997679914082833432101613274016528309938039045923276906900359729197099407393495634862038990107426872
82297597465537710225767267684285801187722495010621811734052320826539734296222735253659051586563138327203111984198746759973864631829032038325230859799799221610122721578080 5
24814583120684401676062393060097116167297155047284877996343375313489942303724373478791319890859537640701348494461138775725769524087024617201078742973804622750525457066893723
19418206440706891884003870590289772197516454495975821662130620506461776109948566373416818358498932907699338206780105243728461492403422961155182609778228619192672071295189 5
89360099591309742330723163825184281103305710174411568843051318658775443763085003114511107238370397074651822320404061547082730786299575493310312752086167006607912980142622 30
05651235227180638195093358726517286235890205200161443617560756542864714221266130044348070840675015892476731663415395405750744749949098314964730310804114018918497359128112 28
37877404988483405621024205664244638600938996508574296194726905430152812375265109658152846997970367921711290355680981807916958795161415928104952817985584729253444786442443 59
98085315372047968149694659917686145337010519859285771574824559433772423695825762426630169463204824951822559392874031776234338810486046309751915569238711675130952134150988 16
71546430786235260623778640683868042469025274911393192768026115159905826038867331729307136739034036186374639806057648364746702744467278808853370742544219227266777470033294 03
32010382880351126890262551830967919483586789293701637681753048206338943871497931152353698229625111163071482945992116208033026847620133356904410896681454361509051558775811 67
97700125639121511162374441704973717046040294811041148222864661319188219975713833683520725260552027698239746132184952492648970507903983602562556062898522888395613578741565 76
64889992608732866126306425432602489792291135600716405739845163752452433769437558573847255455643975996042559146401122211447552355731762399730577471839565312174165322959866 75
90129411612392407220932503696731248844915537592106506560154167207741592362408686676753482865129648887390597075788024733934634708481590116397727977474804173162687009167287 35
61216422684681606831989598012603764856153127816116895872151231233087600634733810972531184233396403909373783950668355787353078863586464005632994994906311874240290927792726 93
30032244537759579722487345689151145855707838505416816676674258113019580636219075007902950310882090972717481364369894739710799327777006763017306175665387397260377771730084 41
34394051236690555449324861650825399577950363267049478442934988531727973481777971465671751511788763964340693324580763461107342143281950499096808740273976889147045174720555 43
89693966687426014772418946932129024533173341882867731946535441133021008665750817132403426475804892186236634616079559437205163951540969498064862423097969472111696659615800 41
78839922322646284989423522769263911243607675089976717896834593378907423463545571935306656153799002779162656363618619740485909382340622354597697331372136593437175855906644 39
64613283001136726019340687064423394891992153043852815416596301185494236348635245857466428346090629162795649265847230036085555989897619162932481400945924898994684688623225 86
01705514689056498372760039274870695504637888197416994290491028531804107765160072633847216389037822700138435995997302657271988043246643123597585552171969605139210102265963 24
78878309774053331315517629781528807186526031763272648280944993704562580999380530584976995700802893798080149029010052938984722799471678048216894241591182842576964647865057 31
53251783023363072982516922103465842658944474649161238546897185079681729091390321828341111848213847677283165486532123173820041319905105189670222018870495856871805095907303 60
69304029372160389689176055876769553823180937058262570838983874090984686566342713975000132918351059433217298798252437075082720879598594371576676601557826996603431977526233 08
89899625878006280095609444169323779449554410336965862615562560106693903032038789709836737860870566414335851061116583145204245132085085899949323648316896711949516716195676 22
70709069738895888557956246664153656172354930180739400476052980172177139168678800027785196617307006128451730758250373564310206511244373082522962504045316059074134388187256 3
47791383066059093188025223100853401768402614015396169891920751471080337577088497401418345997539720598786820648791160649698581776011539720584982226989071813494326918018211 73
31880636539108936898117148913574566805428074851701758582666396335701893544498326697628350926579222017463721902731196417514899440100796368760178267471070199454732188878327 42
60889667243715747134206000937042513098936305374597842799804031329894172664922904257309583685344162156405572902820662240038632375263809102332698978388604237596256015679752 62
69507986398681042948323331602672165551781208992646778049357413263871374084238855465383361586434513054396242813972795597259951107063143059926154956229583202327080576811566 90
48958661052203005737252984721187478271367136660586692710948755639748584894759108197270338782844398644867434562009581619303147273459619004993184243379752436624893633212448 50
59719925236685292493053462527641378534132089431289015237380925560459870909127666623296787033288820591349495800740744731433888007245323217473096597419671144414531271327902 05
10100476710143506388579534784472553898015419233170275198961806351526825431731938329258919315301641305489723111286664654929719304792643282955671909288169209104233412200745 4
24204990087258504620805110487588305949599031118873666850941488217257345763552339640384813182131674083590069164000532622581847837650678044511777173286581898992153583094477 65
350341796875

Approximately 3.00 times 10^6328
```

A side-note
The `treeCounter` tool

- Evidently, the tree count can not be computed using normal integers

  → we need an arbitrary precision library

  → I used the GNU GMP (Multiple Precision Arithmetic) library

  → `treeCounter` available as open-source code at https://github.com/stamatak

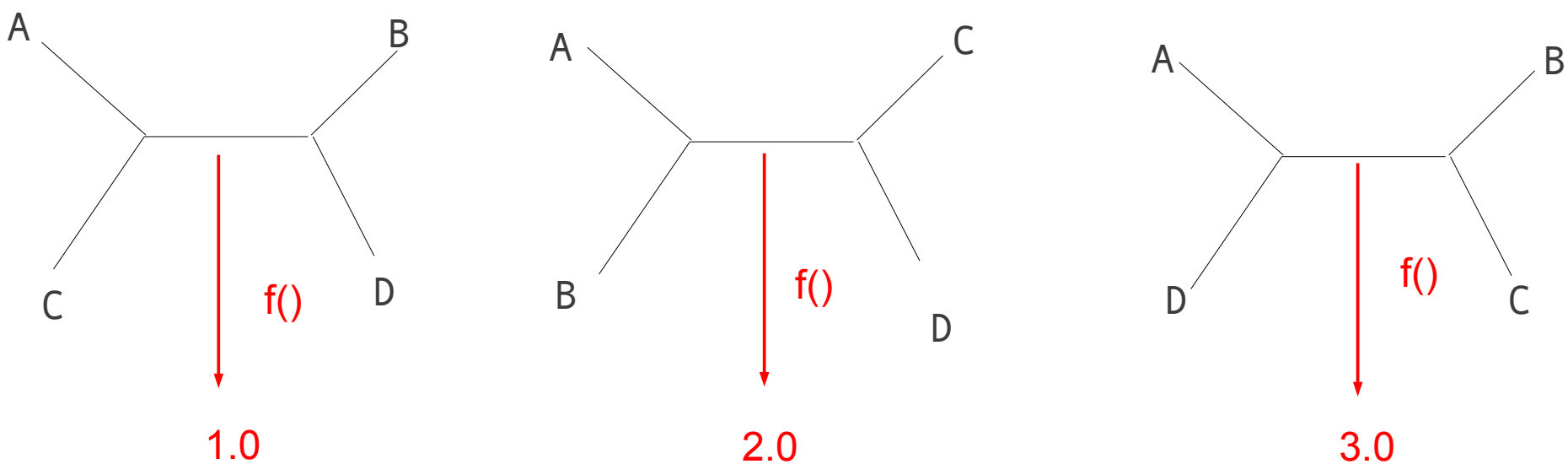
  → Has anybody already used GNU GMP?

Scoring Trees

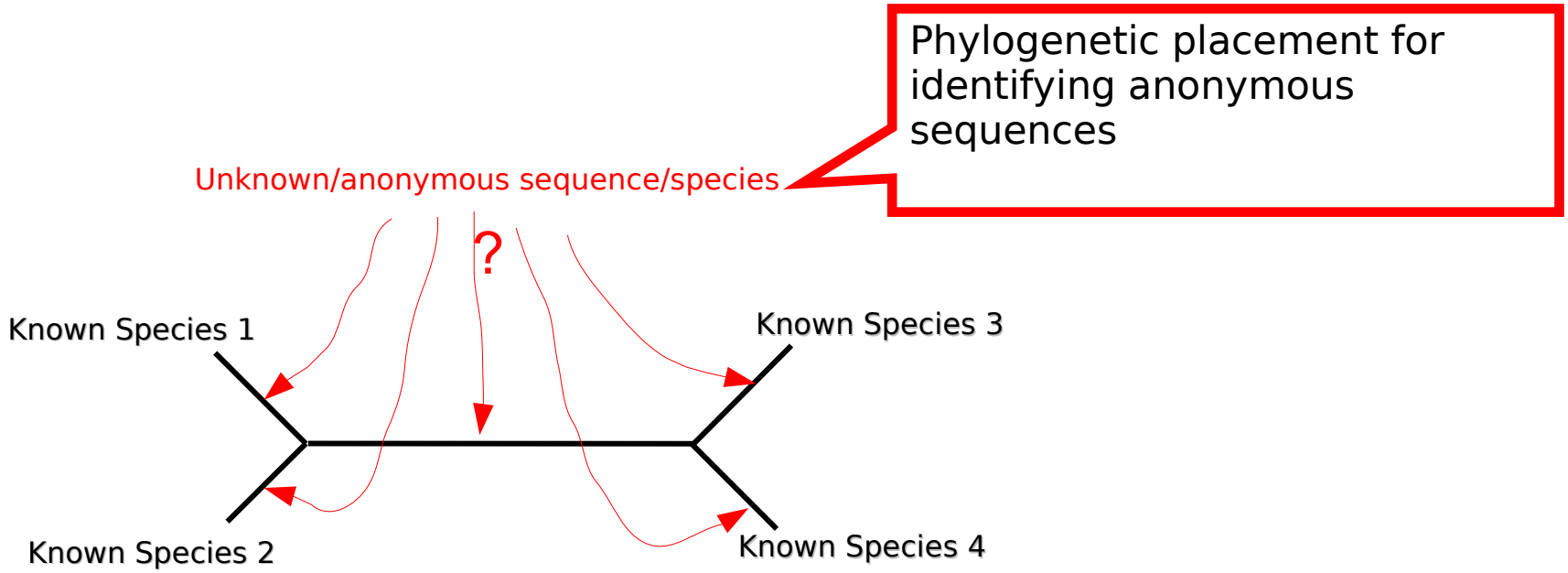
- Okay, so now we know how many **unrooted** candidate trees there exist for *n* taxa
- How do we chose among them?

  → we need some scoring criterion f() to evaluate them

  → finding the optimal tree under most criteria is NP-Hard

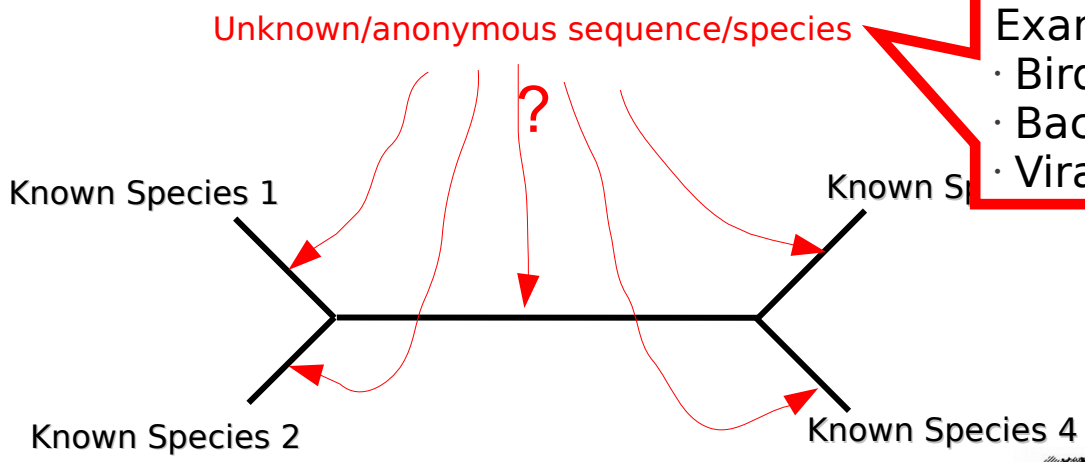


What can we do with Phylogenies?



reference phylogeny

What can we do with Phylogenies?



Phylogenetic placement for identifying anonymous sequences

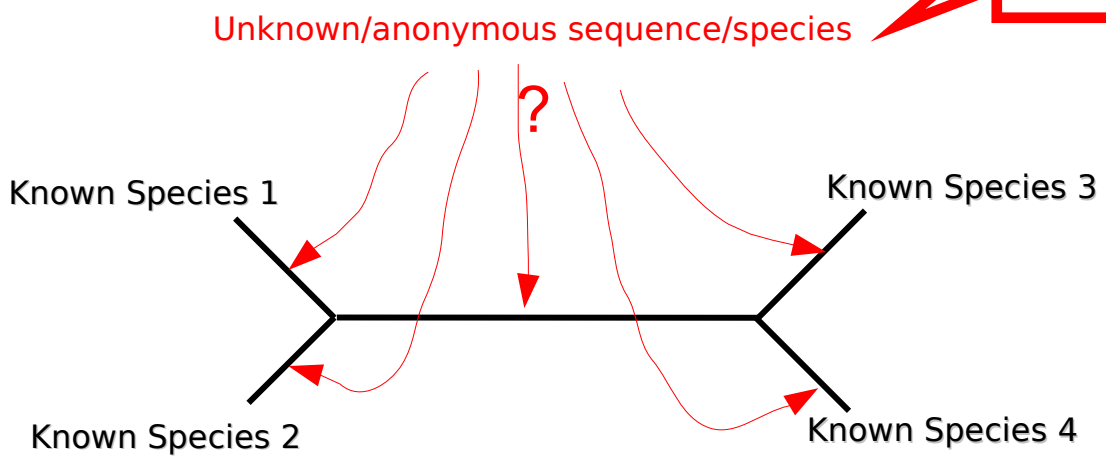
Examples:

- Bird strike
- Bacteria
- Viral strains

reference phylogeny

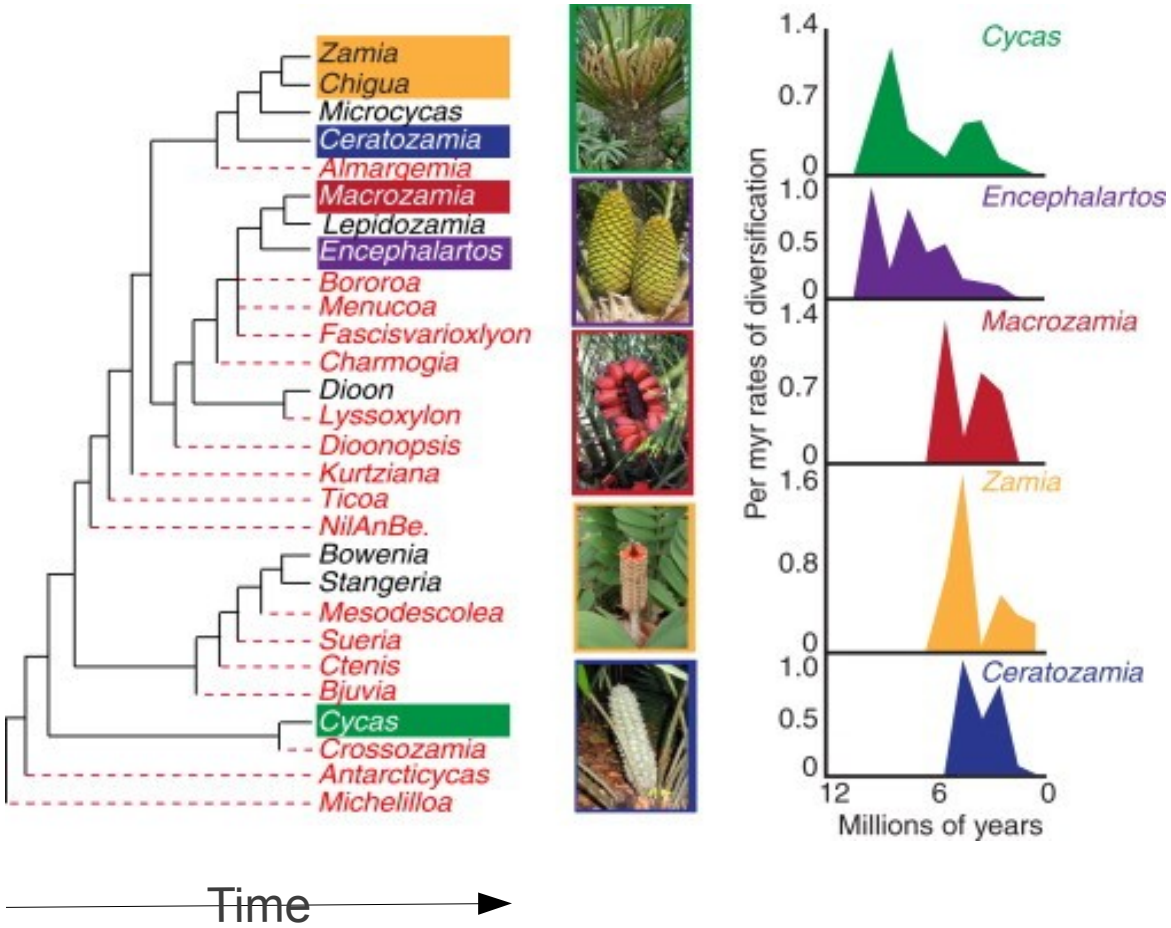
What can we do with Phylogenies?

Note that, this is similar to placing an outgroup into the tree!



reference phylogeny

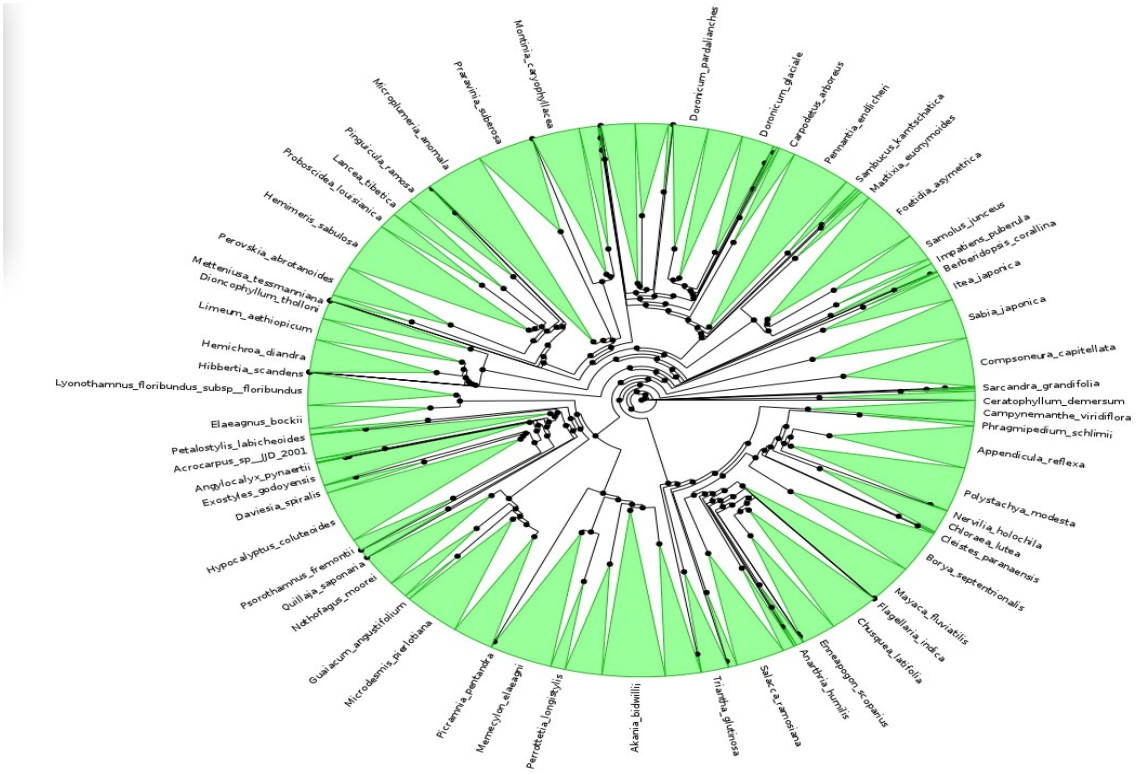
Diversification Rates



From: Charles C. Davis, Hanno Schaefer: “Plant Evolution: Pulses of Extinction and Speciation in Gymnosperm Diversity”, *Current Biology*, 2011.

Diversification Rates

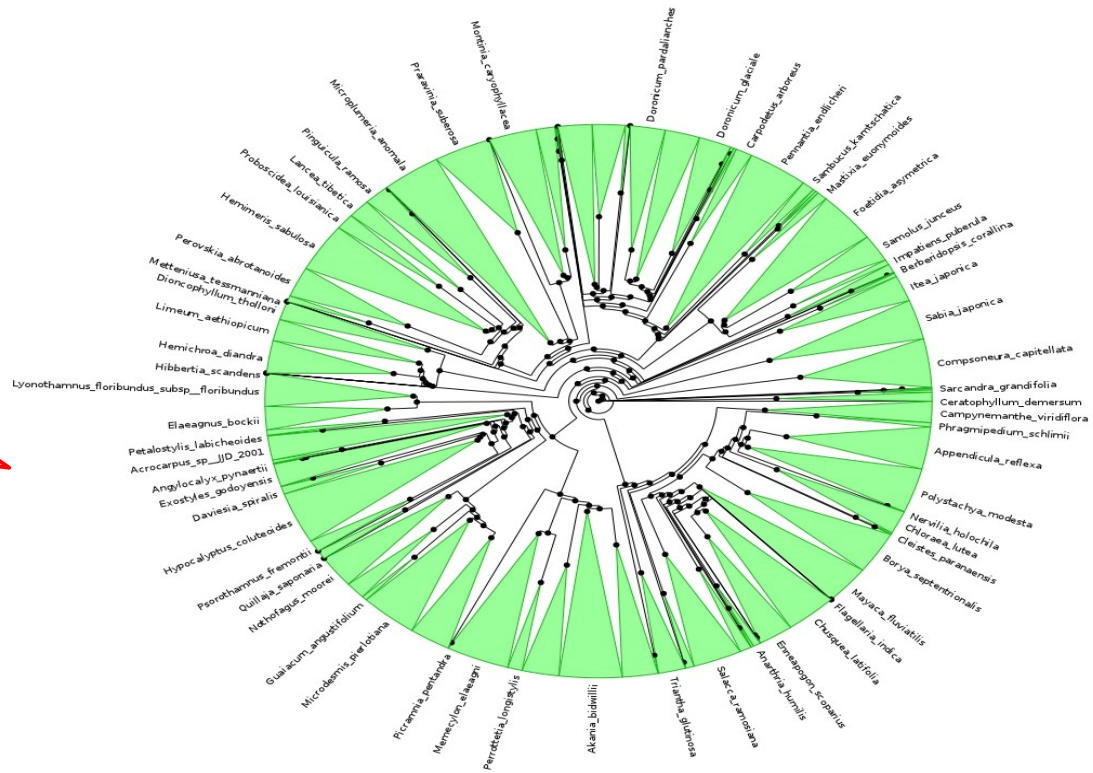
- With former PostDoc Stephen Smith: “Understanding angiosperm diversification using small and large phylogenetic trees”, *American Journal of Botany* 98 (3), 404-414, 2011.
- Largest tree of angiosperms computed to date
- 55,000 taxa



Diversification Rates

- With former PostDoc Stephen Smith: “Understanding angiosperm diversification using small and large phylogenetic trees”, *American Journal of Botany* 98 (3), 404-414, 2011.
- Largest tree of angiosperms computed to date
- 55,000 taxa

Visualizing big trees also represents a challenge → graph drawing & layout Algorithms.



Influenza Outbreaks

Snakebites

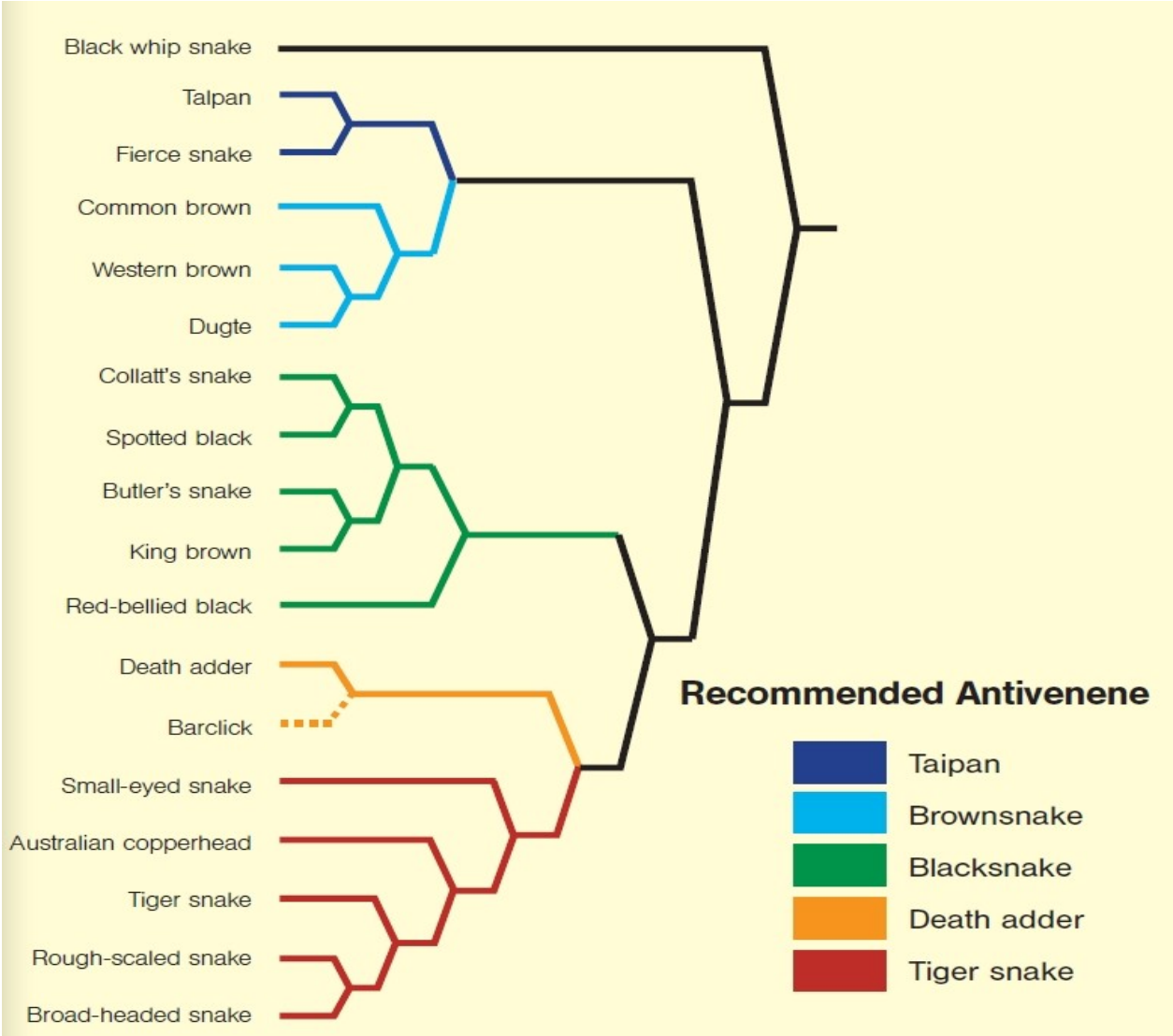
Australia has more poisonous snakes than any other continent, and many people die from snakebites each year. Developing effective antivenins is thus a high priority, but little is known about the venins of most species.

Phylogenetic analysis is helping with this task because venin properties correlate strongly with evolutionary relationships.

Although the red-bellied black snake looks very different from the king brown, it is actually closely related and can be treated with the same antivenin.

Conversely, the western brown looks very similar to the king brown, but it is only distantly related and thus responds best to different antivenin.

The phylogeny is also predictive: the recent demonstration that the poorly-known bardick is closely related to the death adder (orange lineage) predicts that the former is also highly dangerous and might respond to widely-available death adder antivenin.



Snakebites

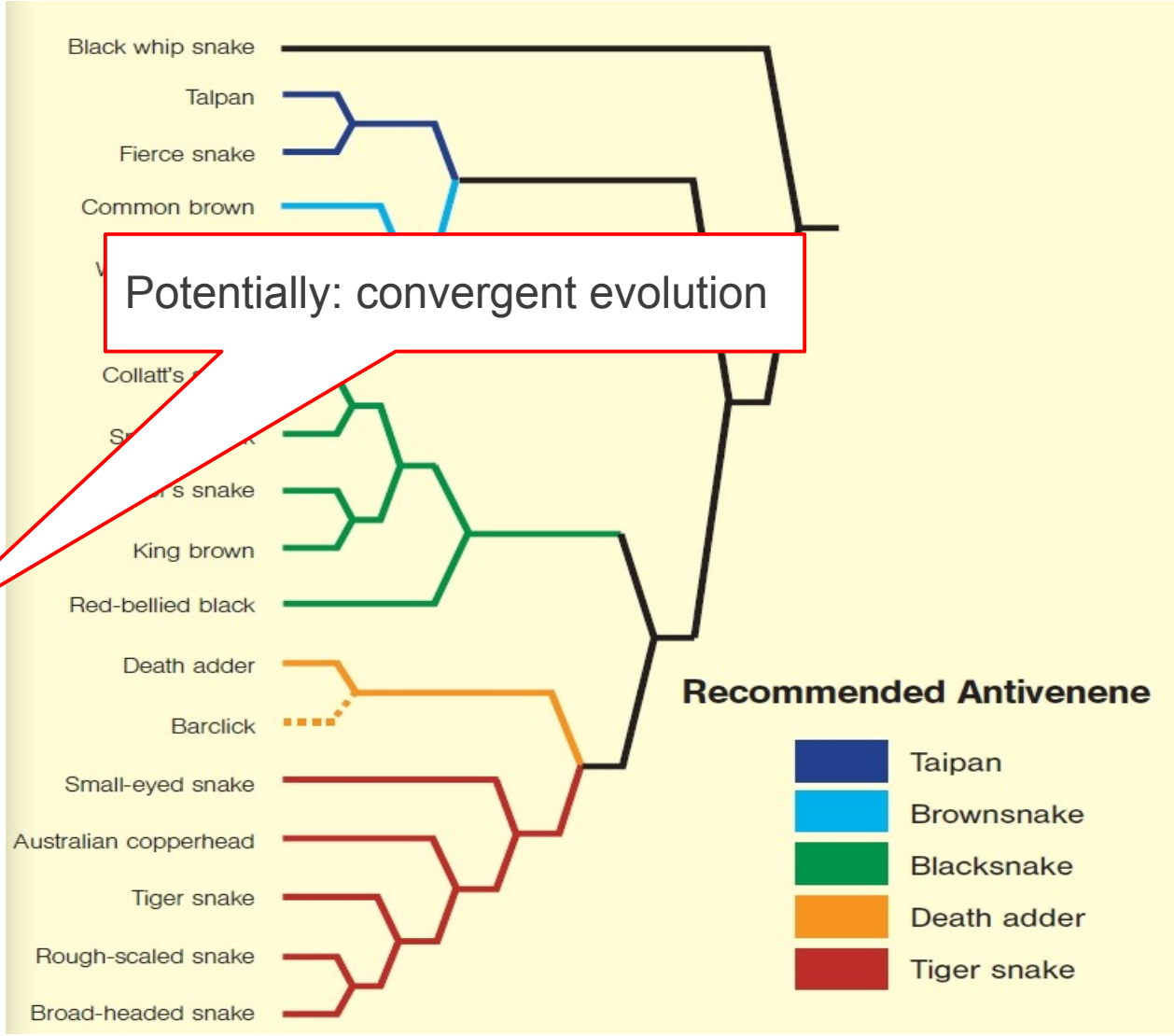
Australia has more poisonous snakes than any other continent, and many people die from snakebites each year. Developing effective antivenins is thus a high priority, but little is known about the venins of most species.

Phylogenetic analysis is helping with this task because venin properties correlate strongly with evolutionary relationships.

Although the red-bellied black snake looks very different from the king brown, it is actually closely related and can be treated with the same antivenin.

Conversely, the western brown looks very similar to the king brown, but it is only distantly related and thus responds best to different antivenin.

The phylogeny is also predictive: the recent demonstration that the poorly-known bardick is closely related to the death adder (orange lineage) predicts that the former is also highly dangerous and might respond to widely-available death adder antivenin.

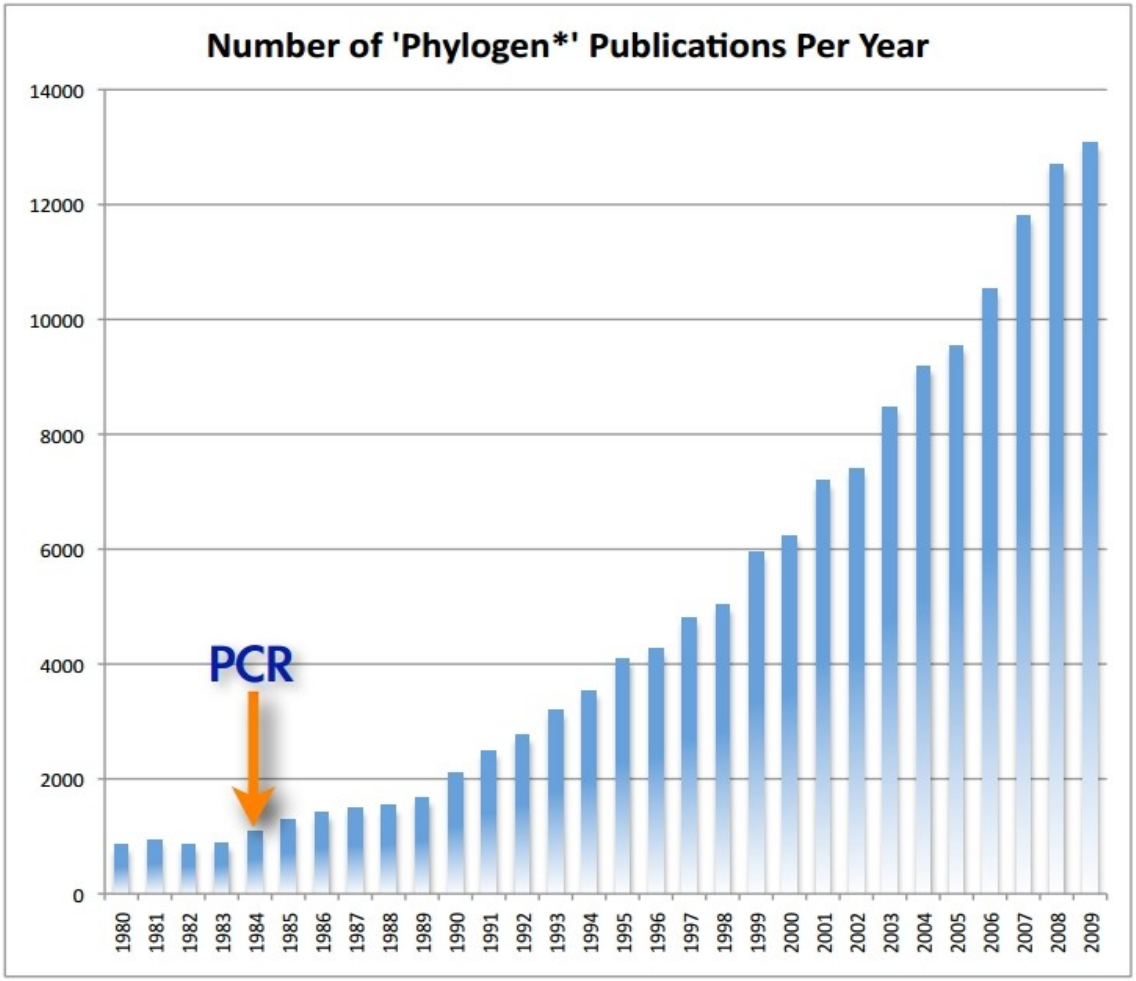


What can we do with phylogenetic trees?

- identifying unknown species
- divergence time estimates
- diversification rates
- viral outbreaks
- forensics → M.L. Metzker, D.P. Mindell, X.M. Liu, R.G. Ptak, R.A. Gibbs, D.M. Hillis: “Molecular evidence of HIV-1 transmission in a criminal case” PNAS: 99(22):14292-7, 2002.
- further reading → see two new links on course web site
  - Assembling the tree of life
  - Disentangling (“entwirren/entflechten”) the trails of evolution

“Nothing in Biology makes sense, except in the light of evolution”

Why?
Advances in:
•Sequencing technology
•Hardware
•Methods & Tools



Building Trees

- We distinguish between
  - *Distance-based methods*

    → use MSA to compute a matrix of pair-wise distances

    → build a tree using these distances

    → Heuristics (essentially hierarchical clustering methods)

      → *Neighbor Joining:* NJ

      → *Unweighted Pair Group Method with Arithmetic Mean:* UPGMA

    → least-squares method: explicit optimality criterion
  - *Character-based methods*

    → optimality criteria *f()* operate directly on the MSA

      → parsimony

      → maximum likelihood

      → Bayesian inference

    → take the current tree topology & MSA to calculate a score

    → the score tells us how well the MSA data fits the tree

Building Trees

- We distinguish between
  - *Distance-based methods*

    → use MSA to compute a matrix of pair-wise distances

    → build a tree using these distances

    → Heuristics (essentially hierarchical clustering methods)

    → *Neighbor Joining:* NJ

    → *Unweighted Pair Group Method with Arithmetic Mean:* UPGMA

    → least-squares method: explicit optimality criterion

  - *Character-based methods*

    → optimality criteria *f()* operate directly on the MSA

    → parsimony

    → maximum likelihood

    → Bayesian inference

    → take the current tree topology & MSA to calculate a score

    → the score tells us how well the MSA data fits the tree

Less accurate, but faster

Slow, but more accurate

Building Trees

- We distinguish between
  - *Distance-based methods*

    → use MSA to compute a matrix of pair-wise distances

    → build a tree using these distances

    → Heuristics (essentially hierarchical clustering methods)

    → *Neighbor Joining:* NJ

    → *Unweighted Pair Group Method with Arithmetic Mean:* UPGMA

    → least-squares method: explicit optimality criterion

  - *Character-based methods*

    → optimality criteria *f()* operate directly on the MSA

    → parsimony

    → maximum likelihood

    → Bayesian inference

    → take the current tree topology & MSA to calculate a score

    → the score tells us how well the MSA data fits the tree

Less accurate, but faster

Memory-intensive!

Slow, but more accurate

Building Trees

- We distinguish between
  - *Distance-based methods*

    → use MSA to compute a matrix of pair-wise distances

    → build a tree using these distances

    → Heuristics (essentially hierarchical clustering methods)

    → *Neighbor Joining:* NJ

    → *Unweighted Pair Group Method with Arithmetic* UPGMA

    → least-squares method: explicit optimality crite

  - *Character-based methods*

    → optimality criteria *f()* operate directly on

    → parsimony

    → maximum likelihood

    → Bayesian inference

    → take the current tree topology & MSA to calculate a score

    → the score tells us how well the MSA data fits the tree

Less accurate, but faster

Storing this matrix can also become problematic memory-wise
→ out-of-core/external memory algorithms
→ e.g.: NINJA tool for Neighbor joining
"Large-scale neighbor-joining with ninja"
T Wheeler,
Algorithms in Bioinformatics, 2009

Out-of-core Algorithms

- Definition from Wikipedia:

  Out-of-core or External memory algorithms are algorithms that are designed to process data that is too large to fit into a computer's main memory at one time. Such algorithms must be optimized to efficiently fetch and access data stored in slow bulk memory such as hard drive or tape drives.

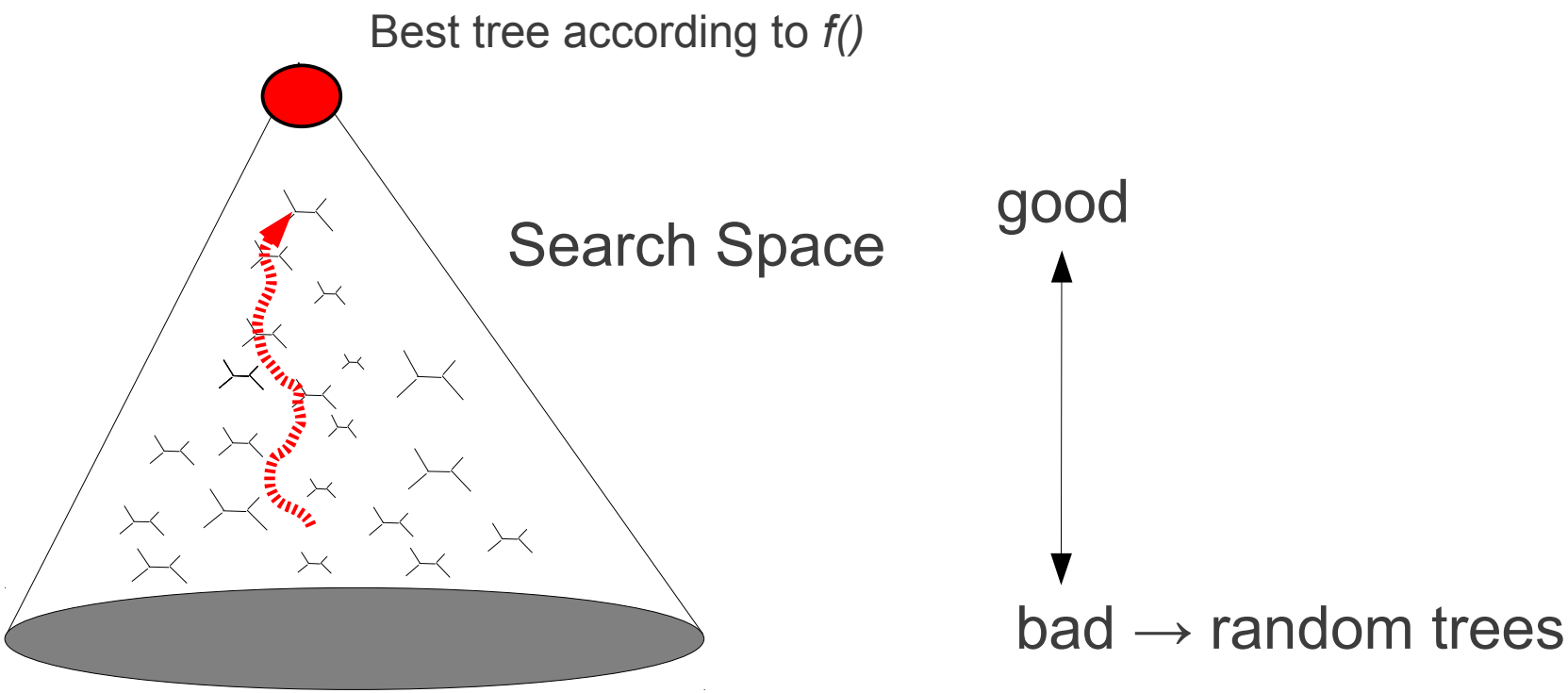
- We do the data transfer RAM ↔ disk explicitly from within the application code by using application-specific knowledge (e.g., about the data access patterns)
- This is to circumvent the paging procedure that would normally be initiated by the OS
- Out-of-core algorithms are typically much faster than the *application-agnostic* paging procedure carried out by the OS
- For an example from phylogenetics see:

  Fernando Izquierdo-Carrasco, Alexandros Stamatakis: "Computing the Phylogenetic Likelihood Function Out-of-Core", *IEEE HICOMB 2011 workshop*, Anchorage, USA, May 2011.

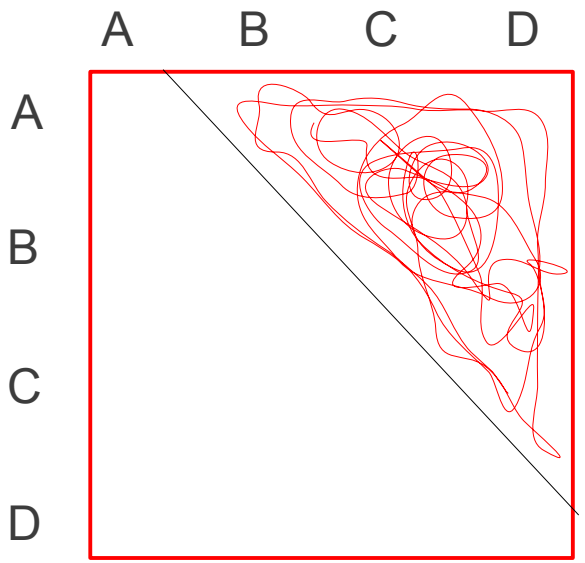
NP-Hardness

- Because of the super-exponential increase in the number of possible trees for *n* taxa ...
- all interesting criteria on trees are NP-hard:
  - Least squares
  - Parsimony → discrete criterion
  - Likelihood → statistical criterion
  - Bayesian → integrate likelihood over entire tree space

Search Space

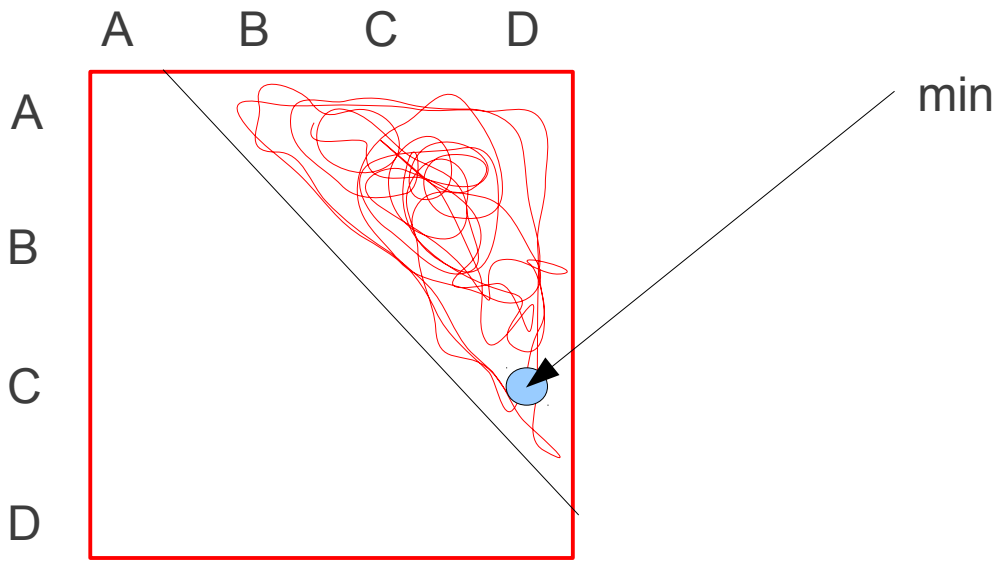


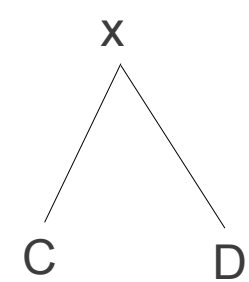
Neighbor Joining → Principle



Given a kind of distance matrix $D_{i,j}$ where $i,j=1...4$

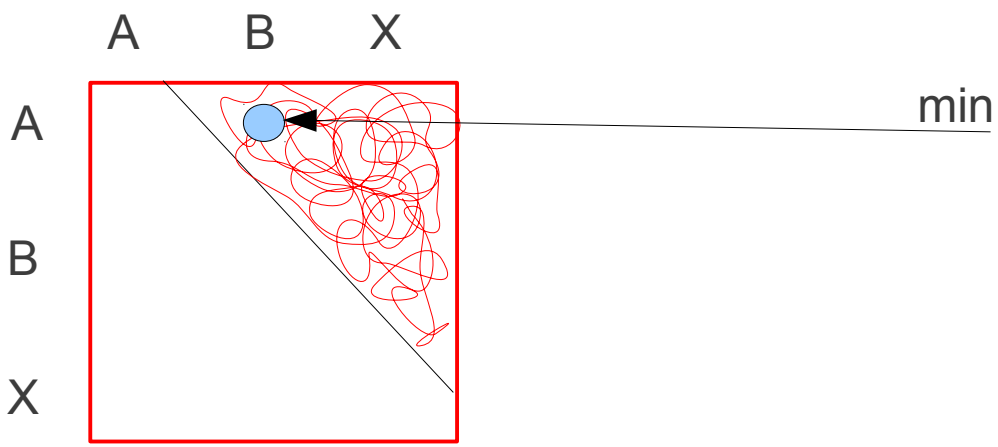
Neighbor Joining → Principle

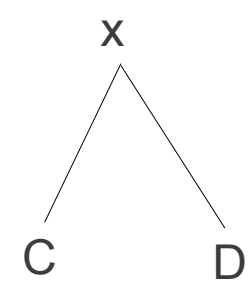




Given a kind of distance matrix $D_{i,j}$ where $i,j=1...4$
Find minimum and merge taxa

Neighbor Joining → Principle





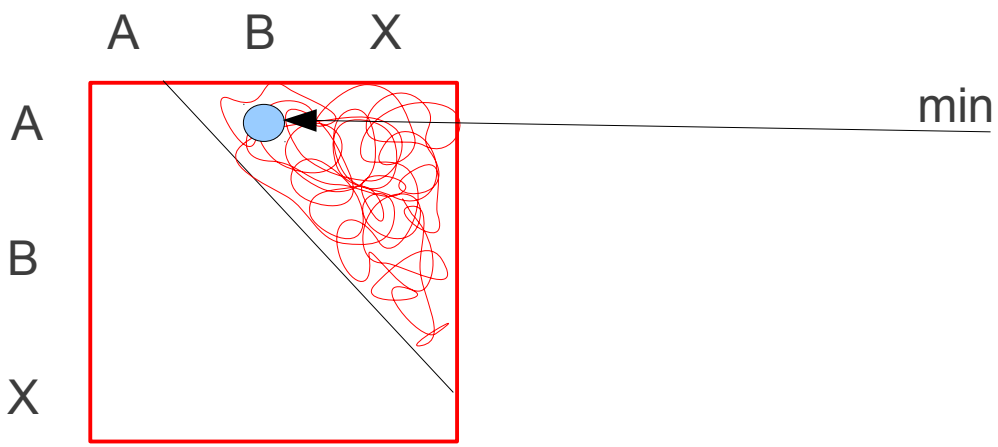
Given a kind of distance matrix $D_{i,j}$ where $i,j=1...4$

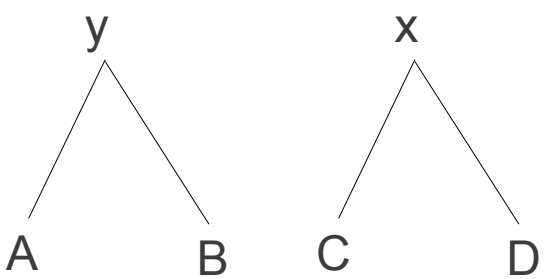
Find minimum and merge taxa

Compute a new distance matrix of size $n-1 = 3$

Find minimum

Neighbor Joining → Principle





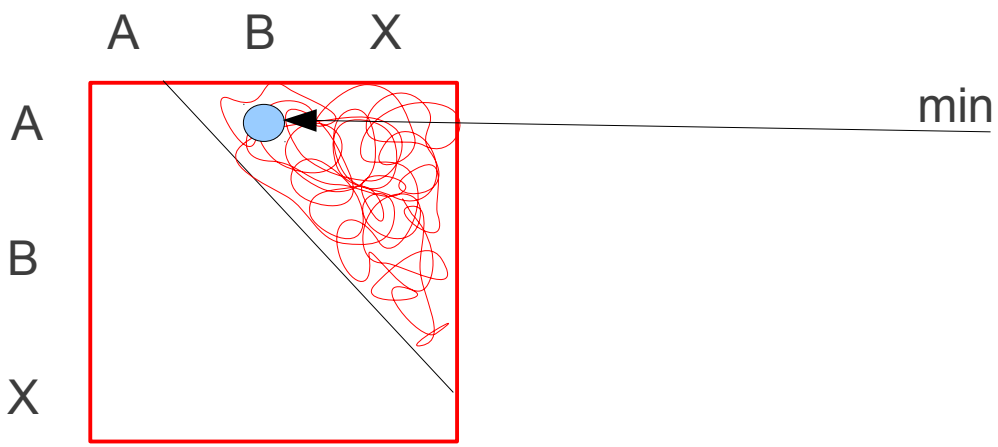
Given a kind of distance matrix $D_{i,j}$ where $i,j=1...4$

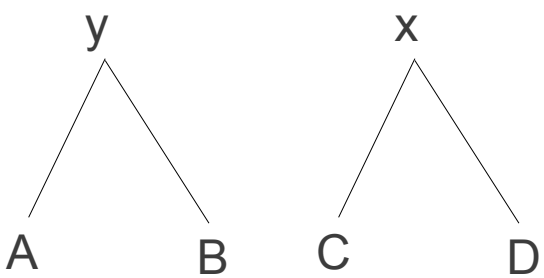
Find minimum and merge taxa

Compute a new distance matrix of size $n-1 = 3$

Find minimum and merge taxa

Neighbor Joining → Principle





Given a kind of distance matrix $D_{i,j}$ where $i,j=1...4$
Find minimum and merge taxa
Compute a new distance matrix of size $n-1 = 3$
Find minimum and merge taxa
Etc.
Space complexity: $O(n^2)$
Time complexity: $O(n^3)$
Key question: how do we compute distance between *X* and *A* or *X* and *B* respectively
→ for progressive alignment we may align the profile of *X* with all remaining sequences

Neighbor Joining Algorithm

- For each tip compute

  $u_i = \Sigma_j D_{ij} / (n-2)$

  → this is in principle the average distance to all other tips

  → the denominator is *n-2* instead of *n*

- Find the pair of tips, *(i, j)* for which $D_{ij}-u_i-u_j$ is minimal
- Connect the tips *(i,j)* to build a new ancestral node *X*
- The branch lengths from the ancestral node *X* to *i* and *j* are:

  $b_i = 0.5\, D_{ij} + 0.5\, (u_i-u_j)$

  $b_j = 0.5\, D_{ij} + 0.5\, (u_j-u_i)$

- Update the distance matrix:
  → Compute distance between the new node *X* and each remaining tip as follows:

  $D_{ij,k} = (D_{ik}+D_{jk}-D_{ij})/2$

- Replace tips *i* and *j* by the new node *X* which is now treated as a tip
- Repeat until only two nodes remain
  → connect the remaining two nodes with each other

Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

Distance matrix, usually denoted as *D*

i	u_i
A	(17+21+27)/2=32.5
B	(17+12+18)/2=23.5
C	(21+12+14)/2=23.5
D	(27+18+14)/2=29.5

Average distance

Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

i	u_i
A	(17+21+27)/2=32.5
B	(17+12+18)/2=23.5
C	(21+12+14)/2=23.5
D	(27+18+14)/2=29.5

	A	B	C	D
A	-	-39	-35	-35
B		-	-35	-35
C			-	-39
D				-

$D_{ij}-u_i-u_j$

Usually denoted as *Q* matrix

Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

i	u_i
A	(17+21+27)/2=32.5
B	(17+12+18)/2=23.5
C	(21+12+14)/2=23.5
D	(27+18+14)/2=29.5

	A	B	C	D
A	-	-39	-35	-35
B		-	-35	-35
C			-	-39
D				-

$D_{ij}-u_i-u_j$

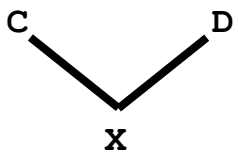
Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

i	u_i
A	(17+21+27)/2=32.5
B	(17+12+18)/2=23.5
C	(21+12+14)/2=23.5
D	(27+18+14)/2=29.5

	A	B	C	D
A	-	-39	-35	-35
B		-	-35	-35
C			-	-39
D				-

$D_{ij}-u_i-u_j$



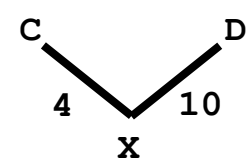
Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

i	u_i
A	(17+21+27)/2=32.5
B	(17+12+18)/2=23.5
C	(21+12+14)/2=23.5
D	(27+18+14)/2=29.5

	A	B	C	D
A	-	-39	-35	-35
B		-	-35	-35
C			-	-39
D				-

$D_{ij}-u_i-u_j$

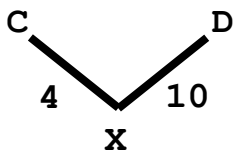


$b_C = 0.5 \times 14 + 0.5 \times (23.5-29.5) = 4$

$b_D = 0.5 \times 14 + 0.5 \times (29.5-23.5) = 10$

Neighbor Joining Algorithm

	A	B	C	D	X
A	-	17	21	27	
B		-	12	18	
C			-	14	
D				-	
X					-

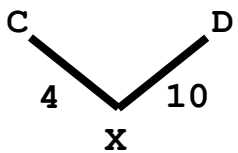


Neighbor Joining Algorithm

	A	B	C	D	X
A	-	17	21	27	
B		-	12	18	
C			-	14	
D				-	
X					-

$$D_{XA} = (D_{CA} + D_{DA} - D_{CD})/2$$
$$= (21 + 27 - 14)/2$$
$$= 17$$

$$D_{XB} = (D_{CB} + D_{DB} - D_{CD})/2$$
$$= (12 + 18 - 14)/2$$
$$= 8$$

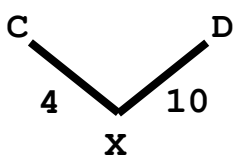


Neighbor Joining Algorithm

	A	B	C	D	X
A	-	17	21	27	17
B		-	12	18	8
C			-	14	
D				-	
X					-

$$D_{XA} = (D_{CA} + D_{DA} - D_{CD})/2$$
$$= (21 + 27 - 14)/2$$
$$= 17$$

$$D_{XB} = (D_{CB} + D_{DB} - D_{CD})/2$$
$$= (12 + 18 - 14)/2$$
$$= 8$$

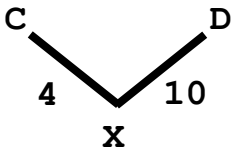


Neighbor Joining Algorithm

	A	B	X
A	-	17	17
B		-	8
X			-

$$D_{XA} = (D_{CA} + D_{DA} - D_{CD})/2$$
$$= (21 + 27 - 14)/2$$
$$= 17$$

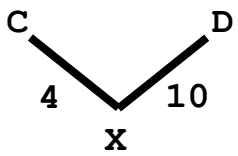
$$D_{XB} = (D_{CB} + D_{DB} - D_{CD})/2$$
$$= (12 + 18 - 14)/2$$
$$= 8$$



Neighbor Joining Algorithm

	A	B	X
A	-	17	17
B		-	8
X			-

i	u_i
A	(17+17)/1 = 34
B	(17+8)/1 = 25
X	(17+8)/1 = 25



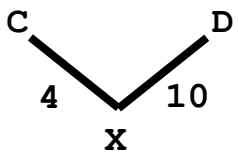
Neighbor Joining Algorithm

	A	B	X
A	-	17	17
B		-	8
X			-

i	u_i
A	(17+17)/1 = 34
B	(17+8)/1 = 25
X	(17+8)/1 = 25

	A	B	X
A	-	-42	-28
B		-	-28
X			-

$D_{ij}-u_i-u_j$



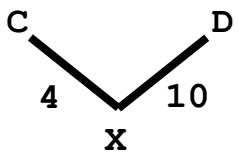
Neighbor Joining Algorithm

	A	B	X
A	-	17	17
B		-	8
X			-

i	u_i
A	(17+17)/1 = 34
B	(17+8)/1 = 25
X	(17+8)/1 = 25

	A	B	X
A	-	-42	-28
B		-	-28
X			-

$D_{ij}-u_i-u_j$



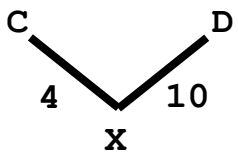
Neighbor Joining Algorithm

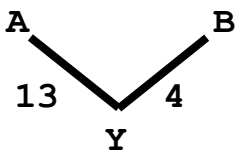
	A	B	X
A	-	17	17
B		-	8
X			-

i	u_i
A	(17+17)/1 = 34
B	(17+8)/1 = 25
X	(17+8)/1 = 25

	A	B	X
A	-	-42	-28
B		-	-28
X			-

$D_{ij}-u_i-u_j$



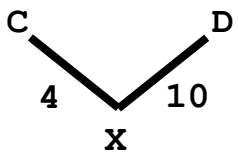


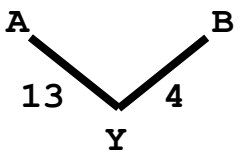
$b_A = 0.5 \times 17 + 0.5 \times (34-25) = 13$

$b_D = 0.5 \times 17 + 0.5 \times (25-34) = 4$

Neighbor Joining Algorithm

	A	B	X	Y
A	-	17	17	
B		-	8	
X			-	
Y				

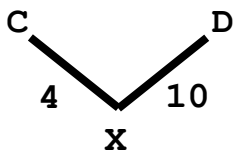


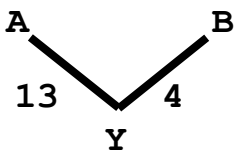


Neighbor Joining Algorithm

	A	B	X	Y
A	-	17	17	
B		-	8	
X			-	4
Y				

$$D_{YX} = (D_{AX} + D_{BX} - D_{AB})/2$$
$$= (17 + 8 - 17)/2$$
$$= 4$$





Neighbor Joining Algorithm

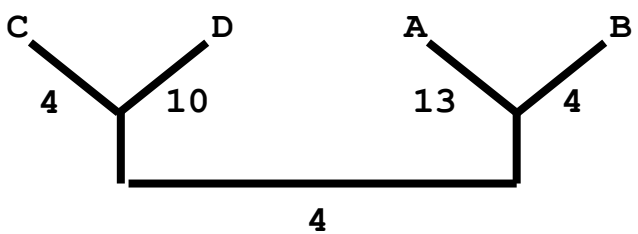
	X	Y
X	-	4
Y		-

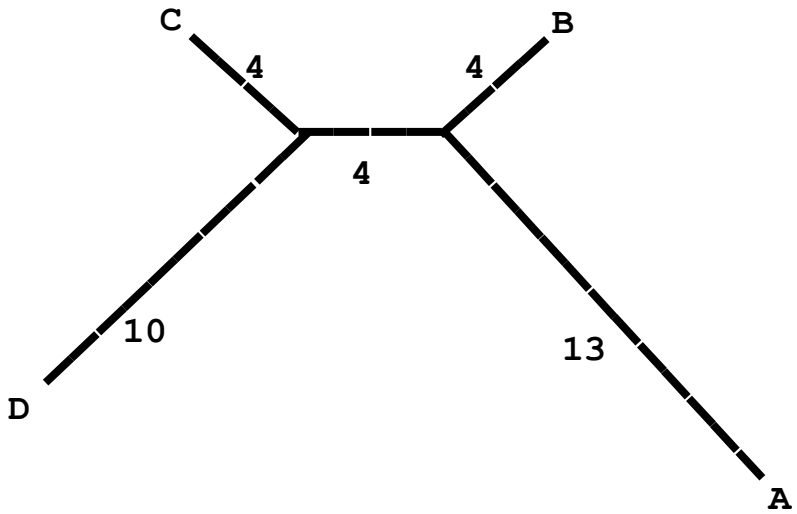
$$D_{YX} = (D_{AX} + D_{BX} - D_{AB})/2$$
$$= (17 + 8 - 17)/2$$
$$= 4$$

Neighbor Joining Algorithm

	X	Y
X	-	4
Y		-

$$D_{YX} = (D_{AX} + D_{BX} - D_{AB})/2$$
$$= (17 + 8 - 17)/2$$
$$= 4$$



Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-