Introduction to Bioinformatics for Computer Scientists

Lecture 6

Plan for next lectures

- Today: Introduction to phylogenetics
- Lecture 7 (Alexis): Phylogenetic search algorithms

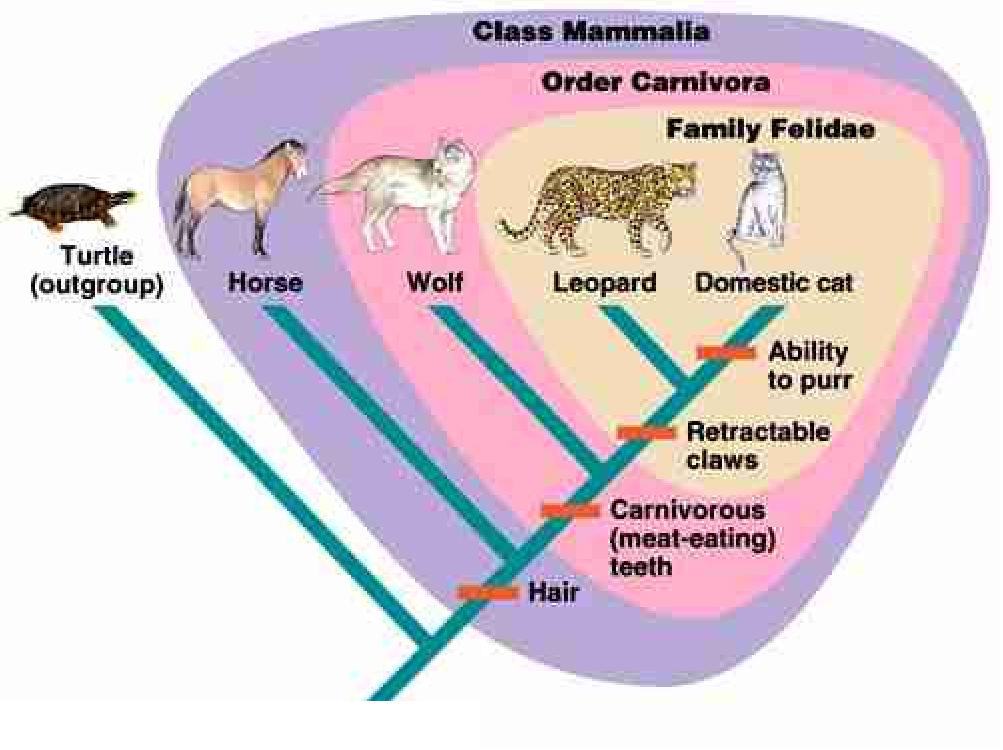
The story so far

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

The story so far

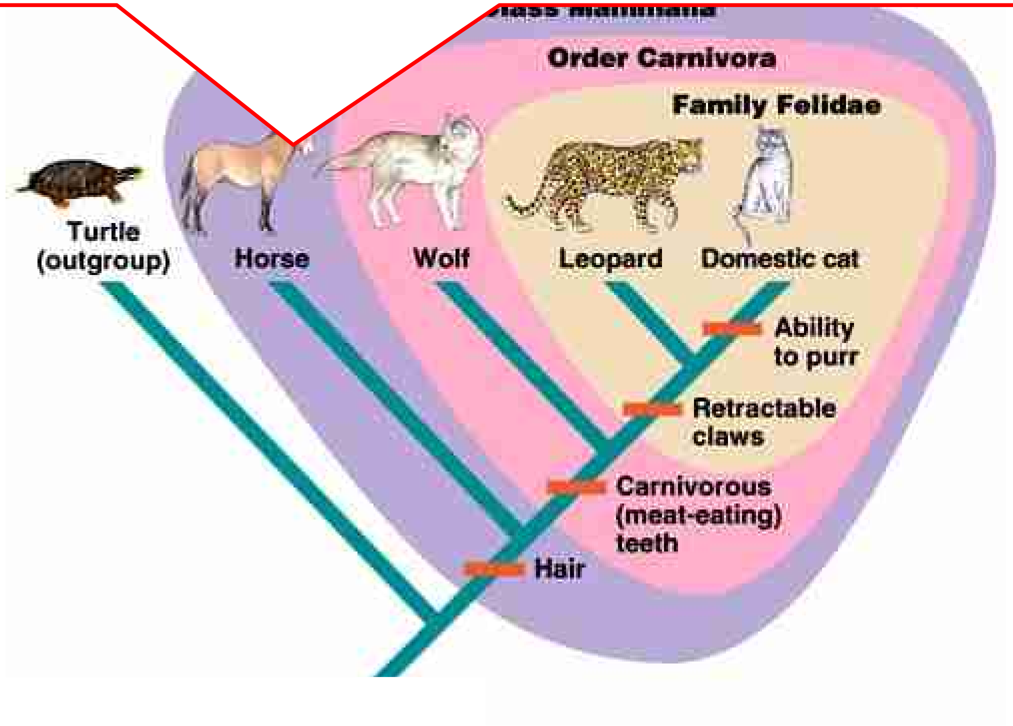
- Biological Terminology: RNA, DNA, genes, genomes, etc
- Pair-wise Sequence Alignment
- Sequence Comparison
- Genome Assembly
- Multiple Sequence Alignment
- Phylogenetic Inference

A Taxonomy


Class Mammalia
Order Carnivora
Family Felidae
Turtle (outgroup)
Horse
Wolf
Leopard
Domestic cat
Ability to purr
Retractable claws
Carnivorous (meat-eating) teeth
Hair

A Taxonomy

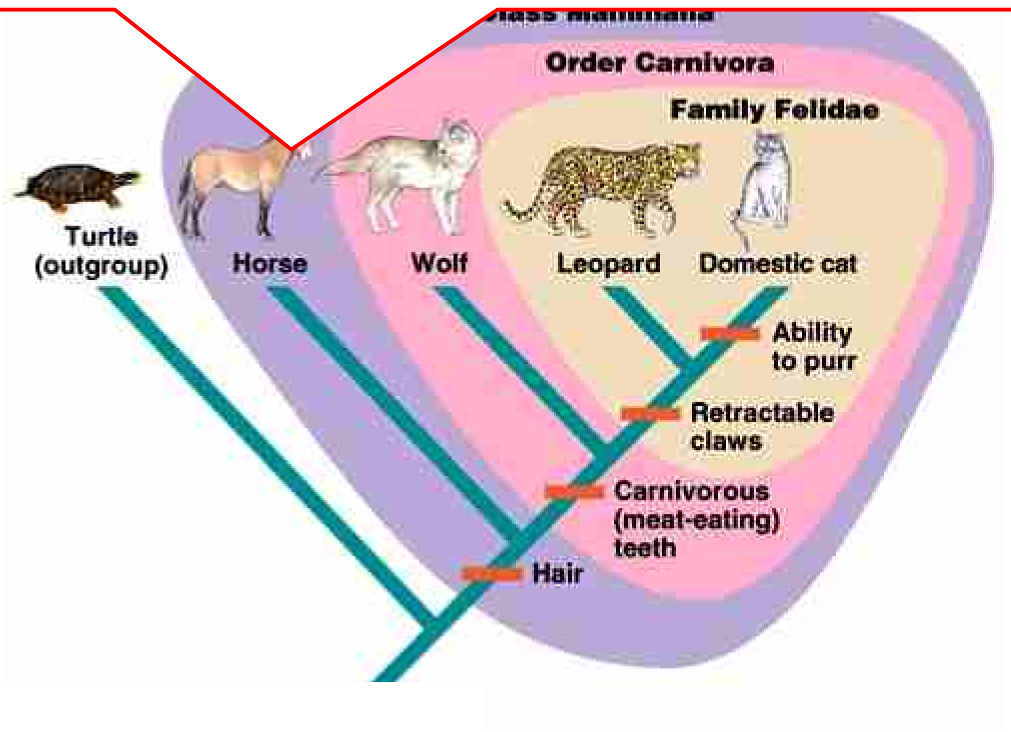
First systematic classification of living beings by Aristotele *384 -382* BC
Some terms still in use today, e.g., classification of animals into
Vertebrates versus *Invertebrates*



A Taxonomy

First systematic classification of living beings by Aristotele *384 -382* BC
Some terms still in use today, e.g., classification of animals into
Vertebrates versus *Invertebrates*

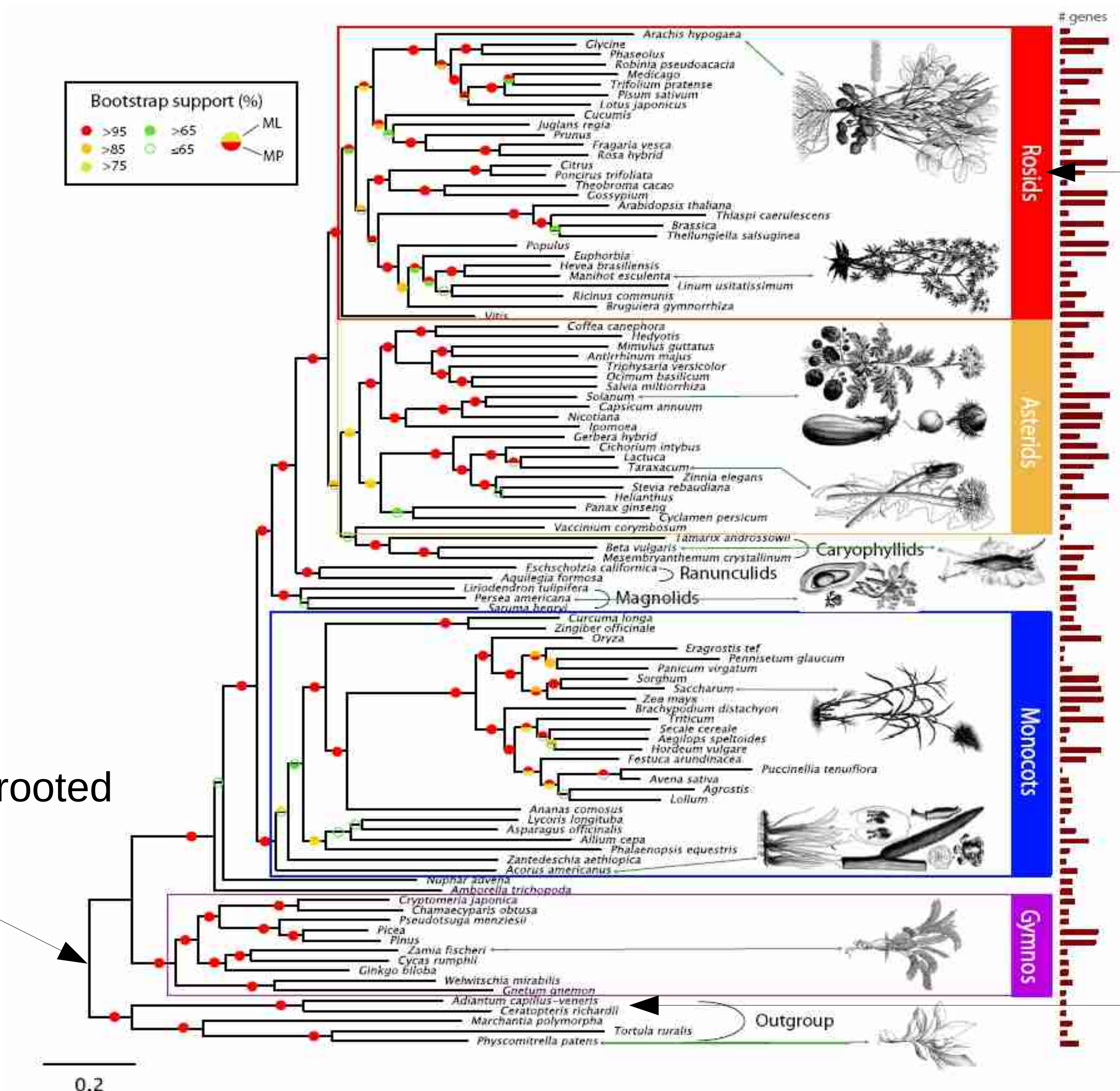
Wirbeltiere



Taxonomy

- Group biological organisms (species) into groups with similar characteristics
- Define characteristics of groups at different hierarchy levels, e.g., animals > mammals > great apes
- Taxonomic ranks
    - Domain → three domains of life
    - Kingdom
    - Phylum
    - Class
    - Order
    - Family
    - Genus
    - Species

A Phylogeny or Phylogenetic Tree



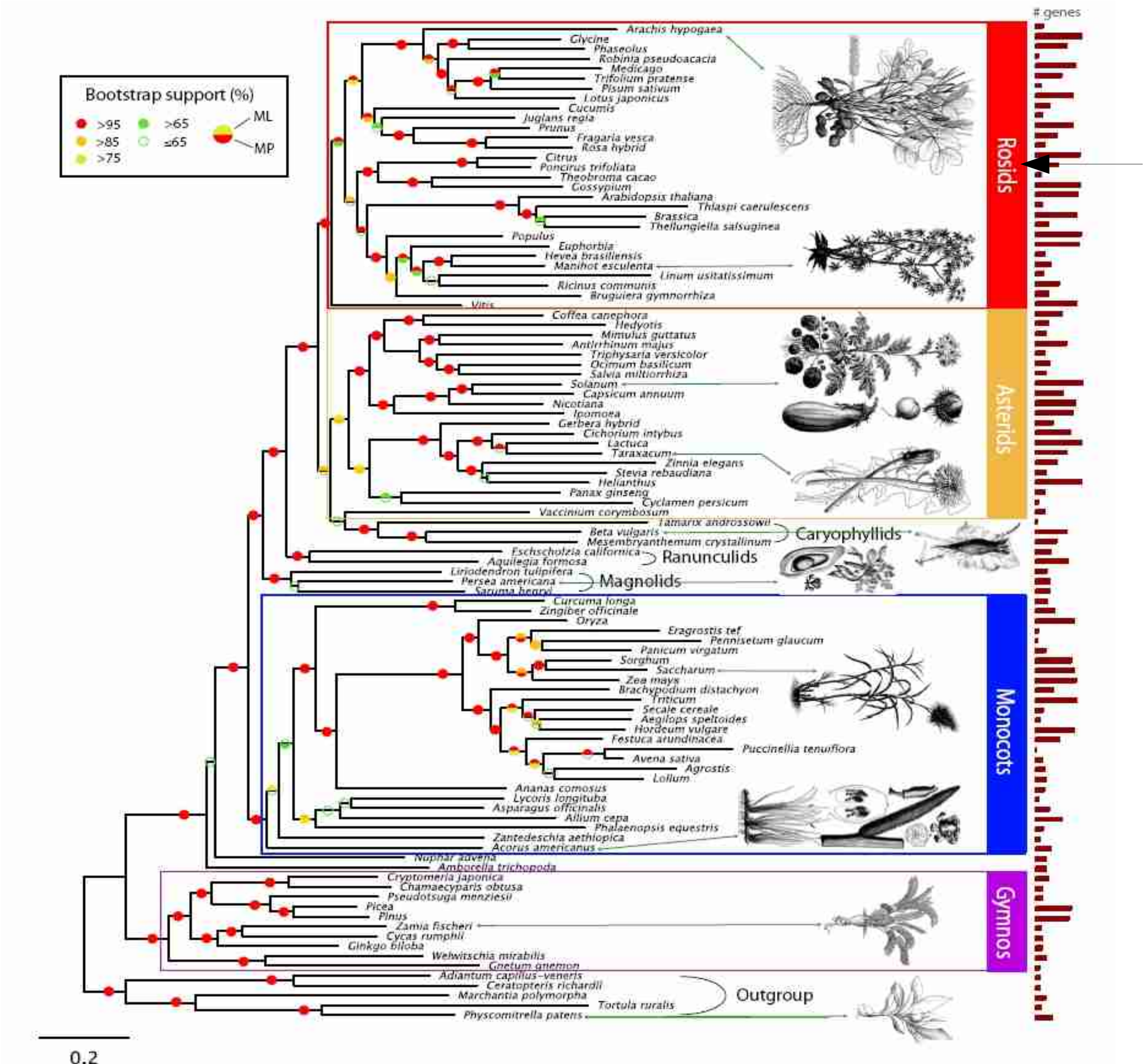
A taxonomic subclass

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 living) organisms represented by their DNA/Protein sequences

  → we can also sequence ancient DNA, see, for instance, the neandertal genome: “The complete genome sequence of a Neanderthal from the Altai Mountains”, *Nature* 2013

  → depends on temperature, time, and other environmental conditions

  → up to 300,000 years back, see

  http://www.pnas.org/content/110/39/15758.abstract
- 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

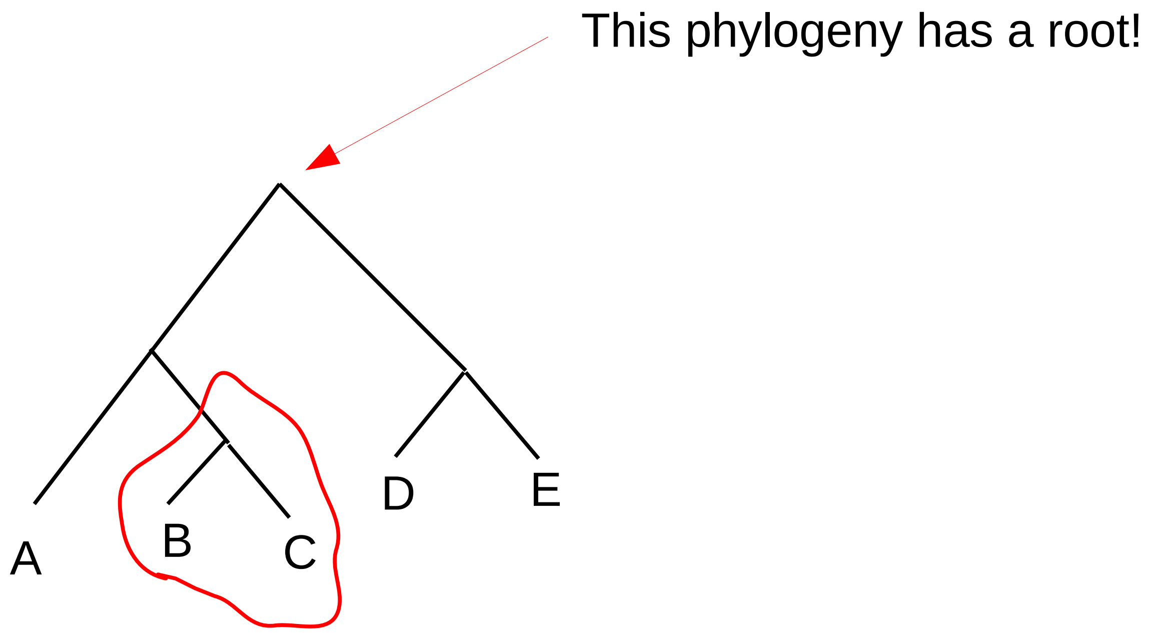
  → subject of controversial debates

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

  → the plural of taxon is *taxa*

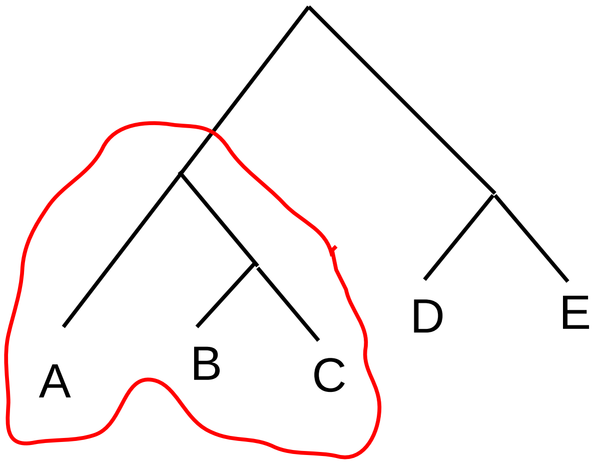
  → we often say that a tree with *n* leaves (sequences) has *n* taxa

Some more terminology



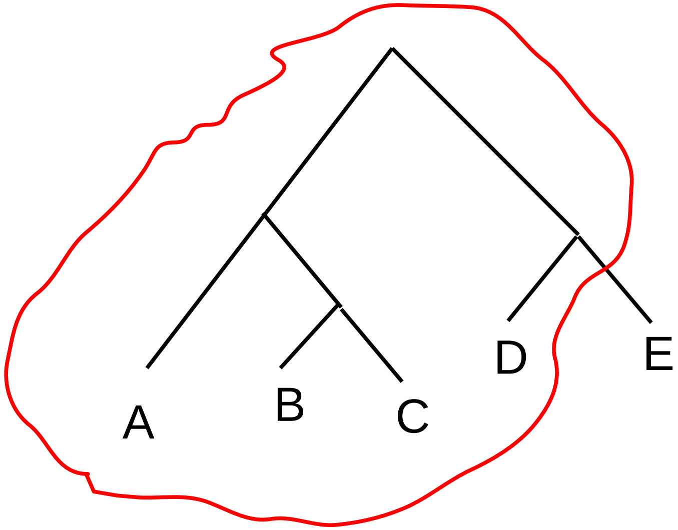
B and **C** are a *monophyletic* group; they are sister species

Some more terminology



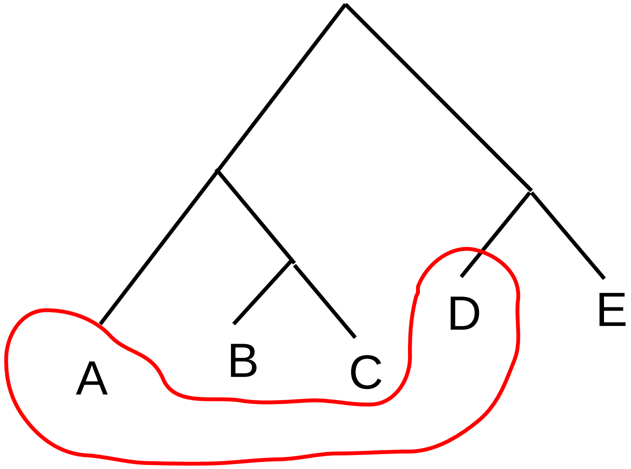
(A,B,C) is a *monophyletic* group; it is sister to **(D, E)**

Some more terminology



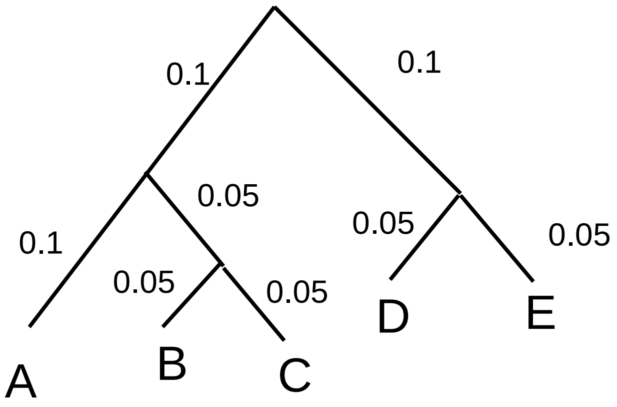
(A,B,C,D) is *paraphyletic* → **E** is excluded

Some more terminology



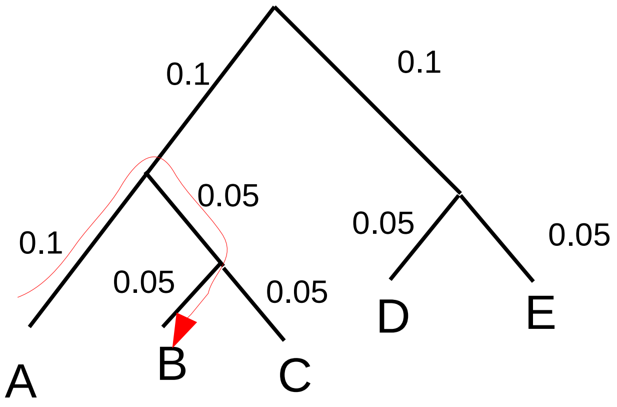
(A,D) is a *polyphyletic* group → their most recent common ancestor (MRCA) is excluded

Some more terminology



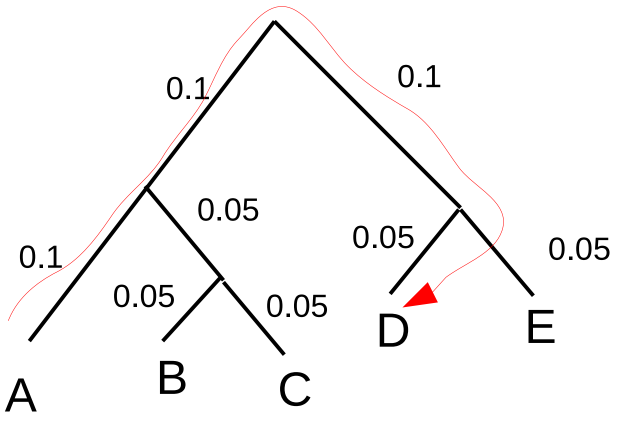
Tree-based or *patristic distance* between two taxa:
Sum over branch lengths along the path in the tree, e.g.:

Some more terminology



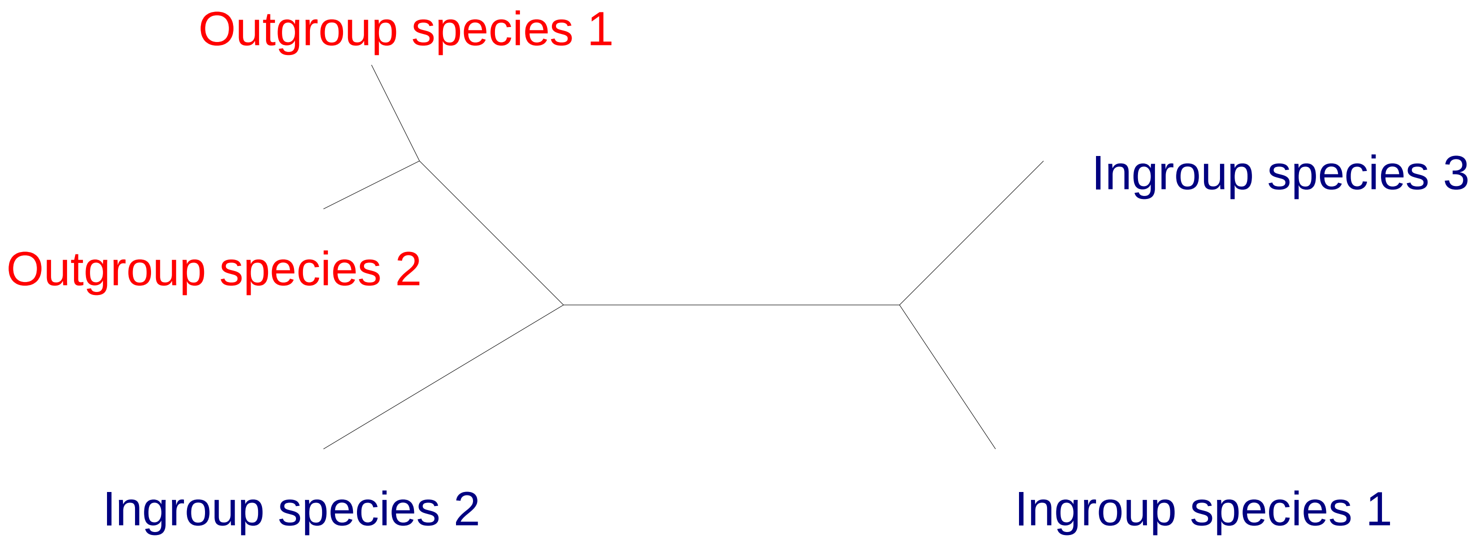
Tree-based or *patristic distance* between two taxa:
Sum over branch lengths along the path in the tree, e.g.:
A ↔ **B**: 0.2

Some more terminology

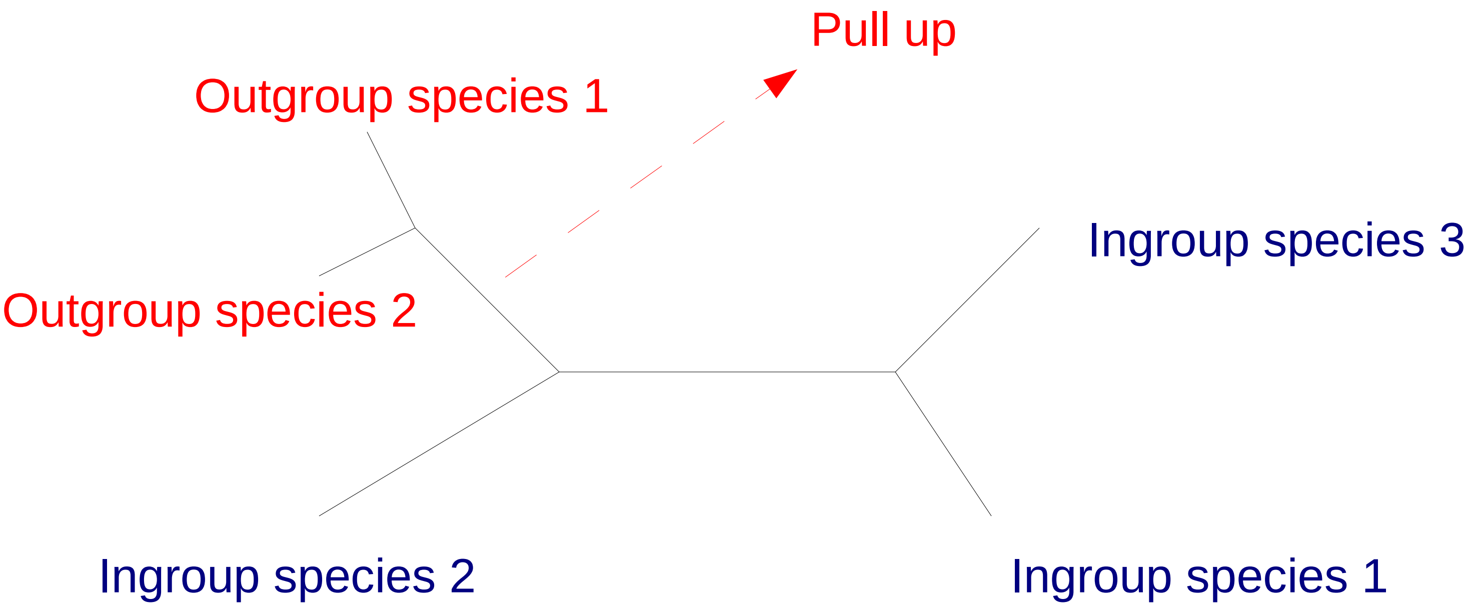


Tree-based or *patristic distance* between two taxa:
Sum over branch lengths along the path in the tree, e.g.:
A ↔ **B**: 0.2
A ↔ **D**: 0.35

Tree Rooting



Tree Rooting



Pull up

Outgroup species 1

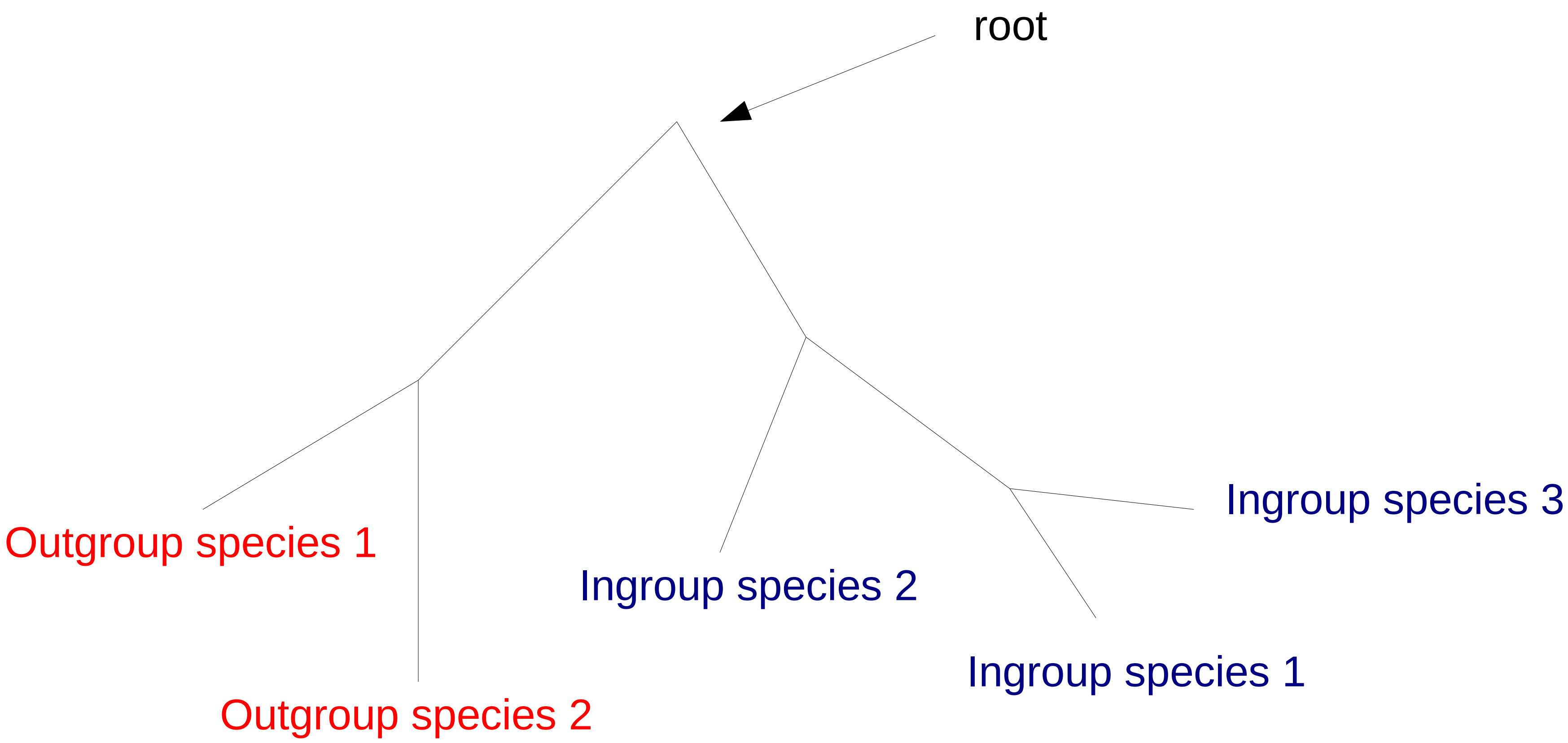
Outgroup species 2

Ingroup species 3

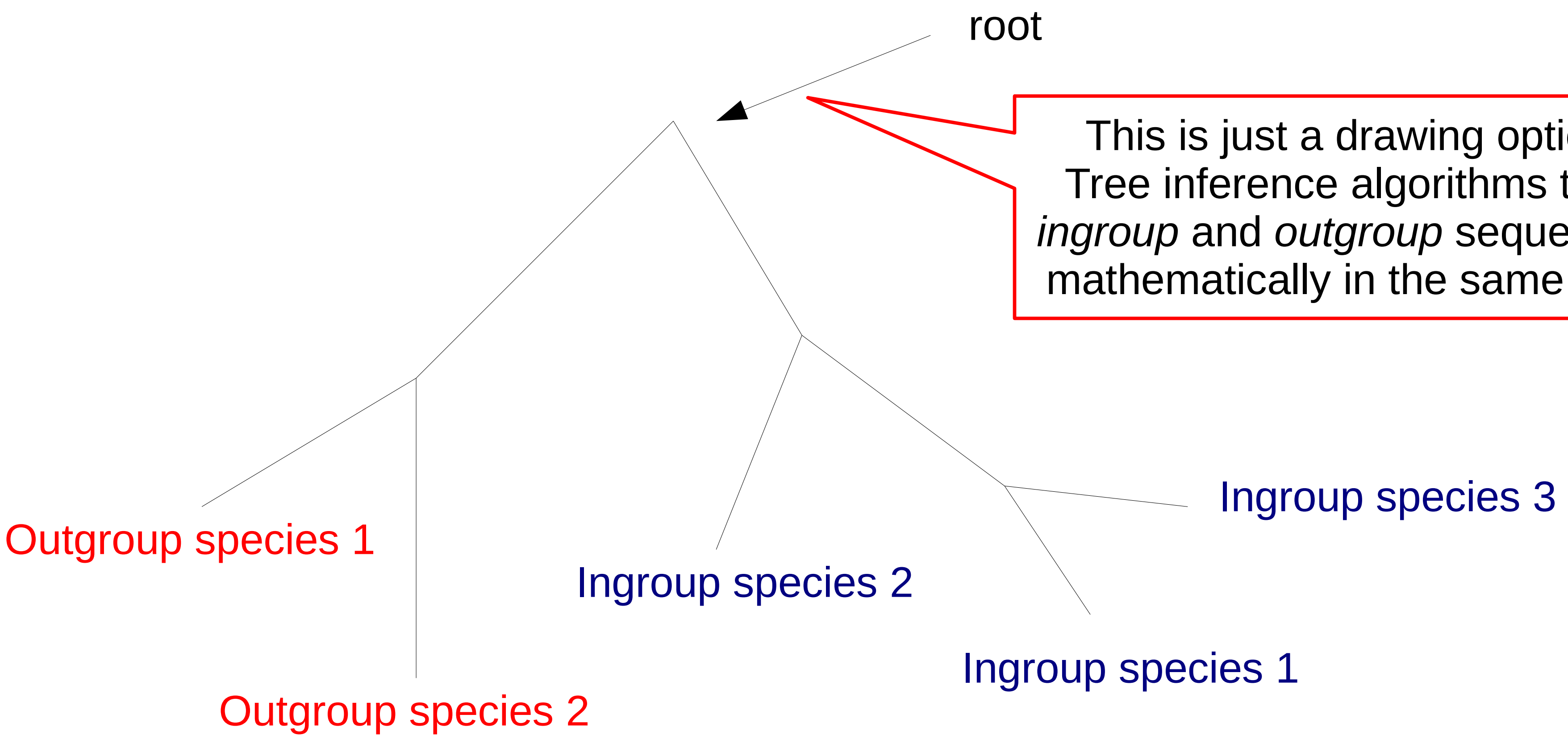
Ingroup species 2

Ingroup species 1

Tree Rooting

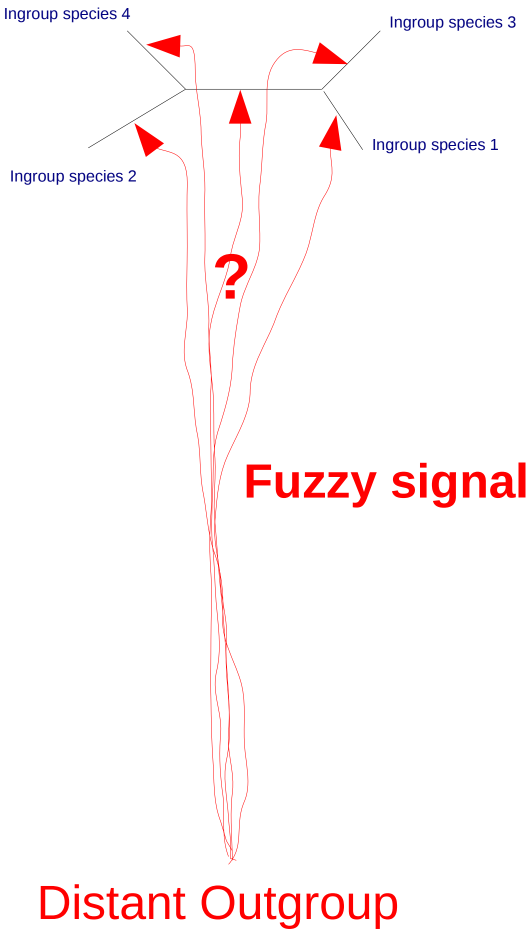


Tree Rooting



This is just a drawing option! Tree inference algorithms treat *ingroup* and *outgroup* sequences mathematically in the same way!

Outgroup Choice



Ingroup species 4

Ingroup species 3

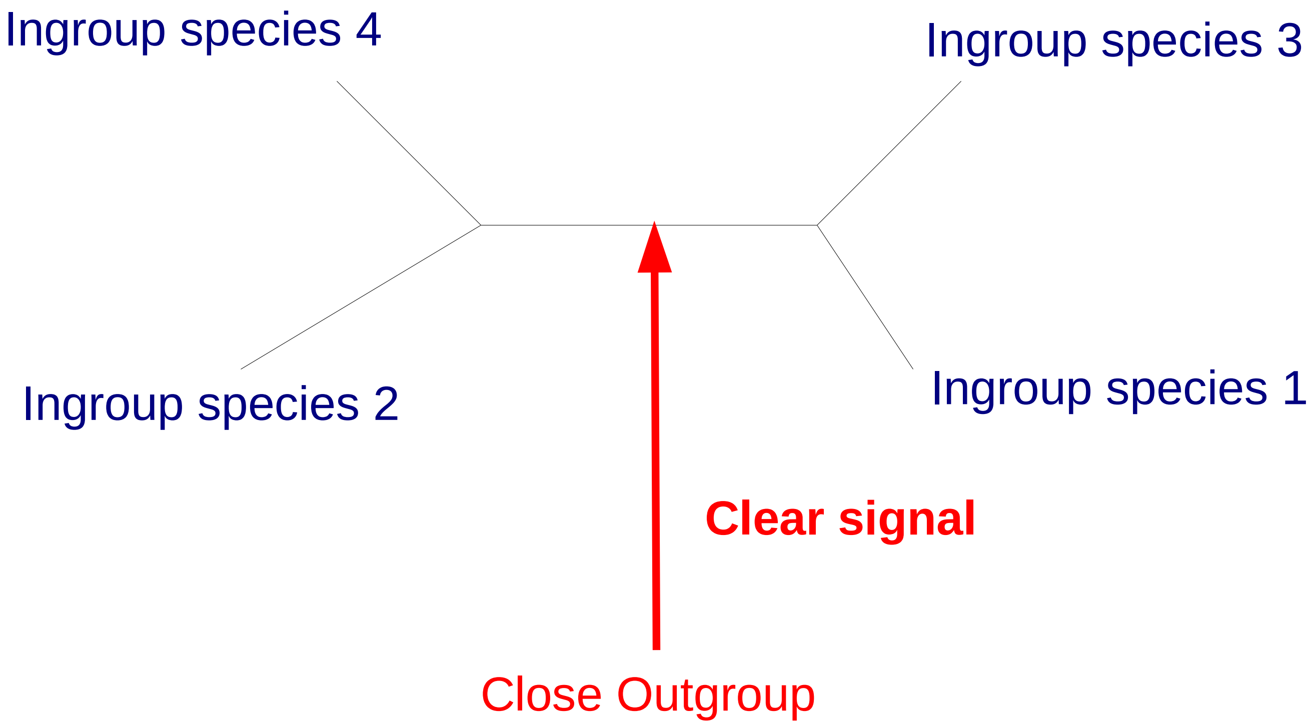
Ingroup species 1

Ingroup species 2

?

Fuzzy signal

Distant Outgroup



Ingroup species 4

Ingroup species 3

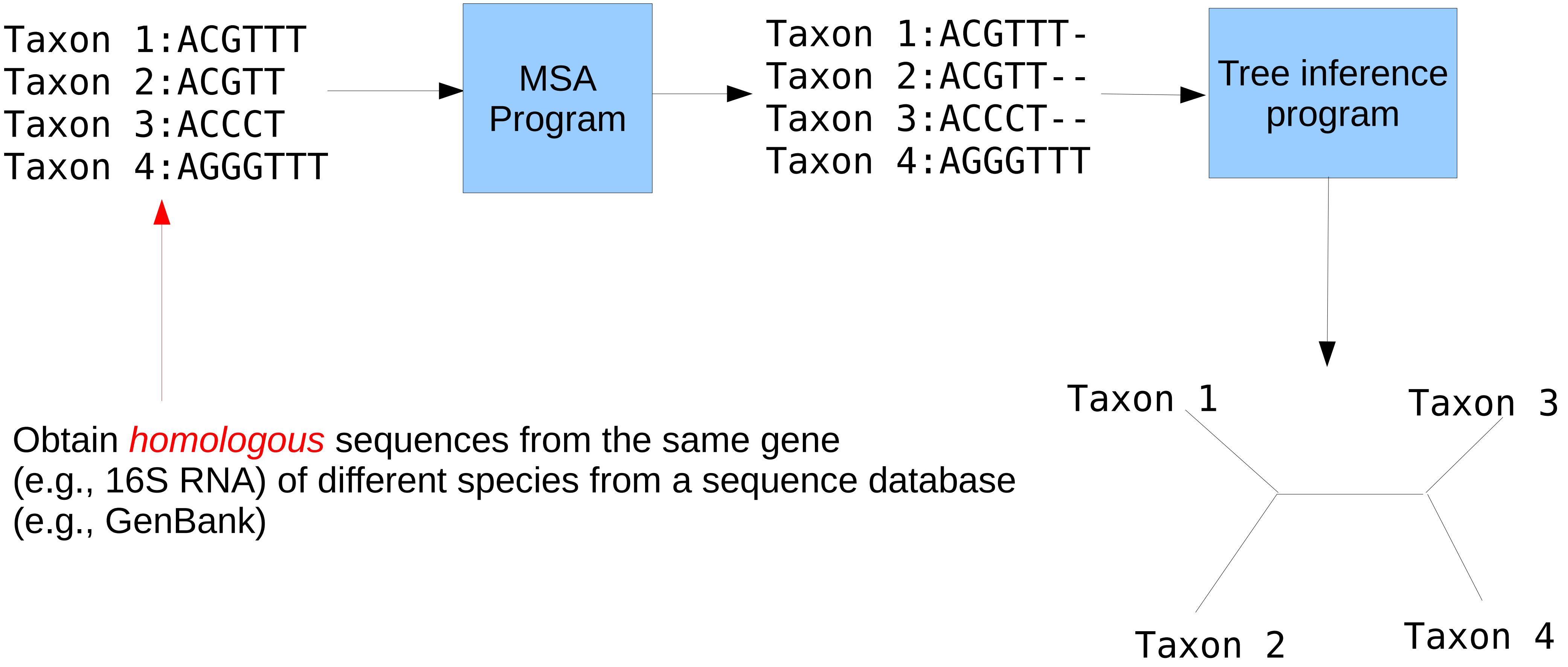
Ingroup species 2

Ingroup species 1

Clear signal

Close Outgroup

Tree Inference



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

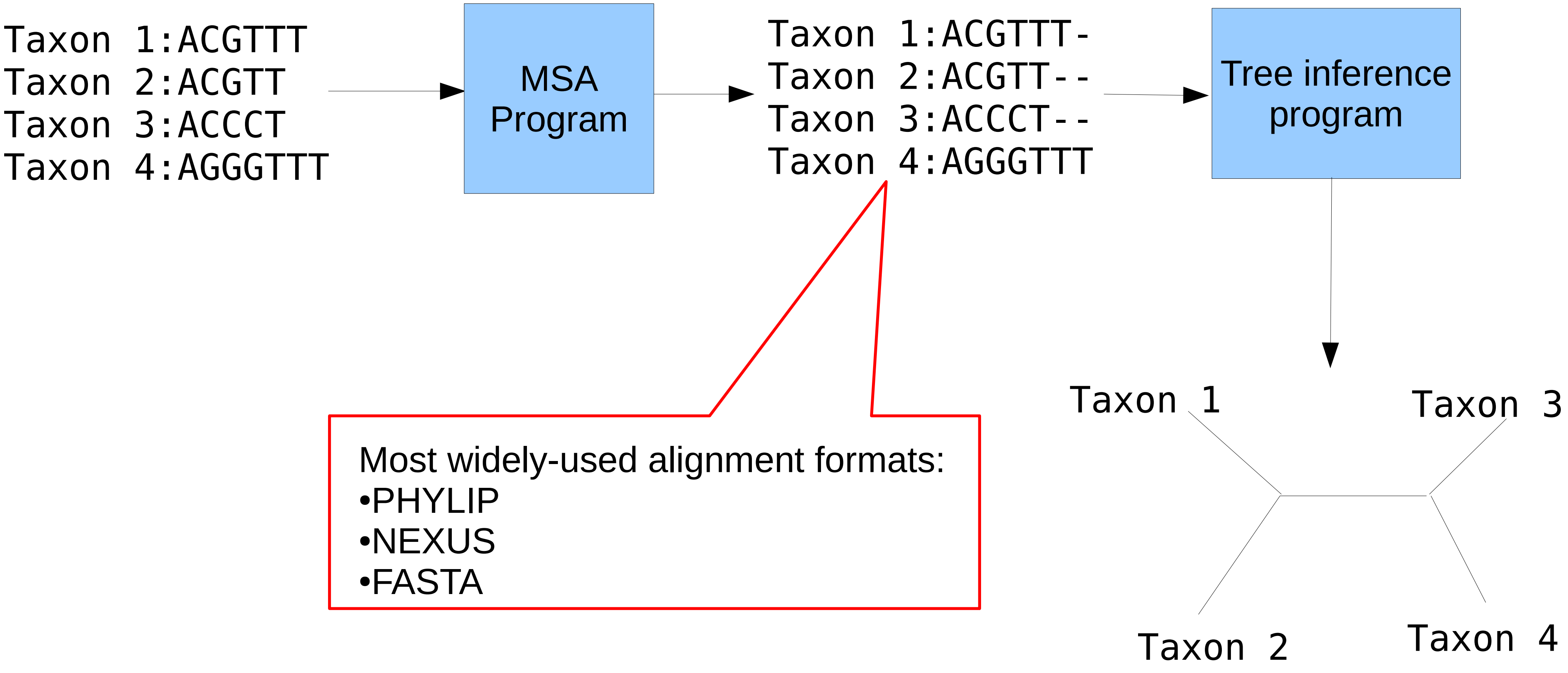
MSA Program

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

Tree inference program

Obtain *homologous* sequences from the same gene (e.g., 16S RNA) of different species from a sequence database (e.g., GenBank)

Tree Inference



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

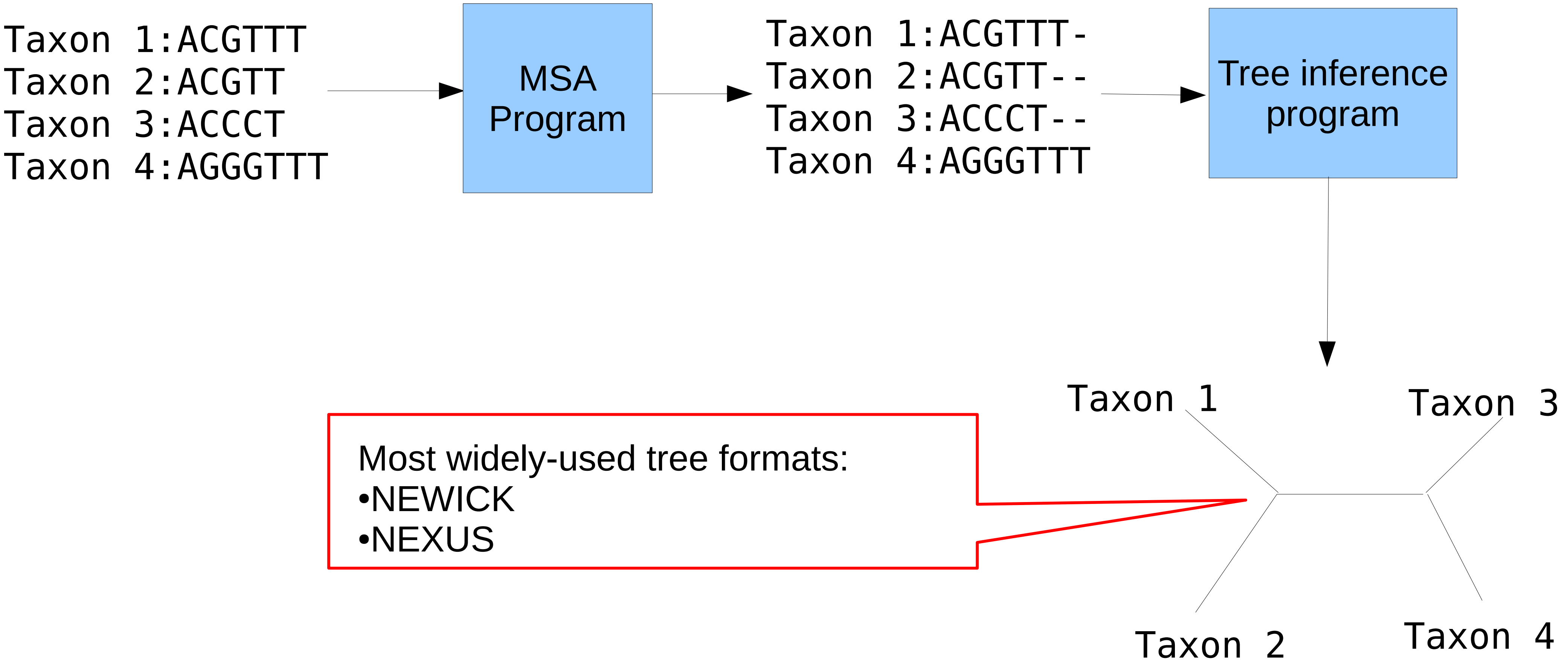
MSA Program

```
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



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

MSA Program

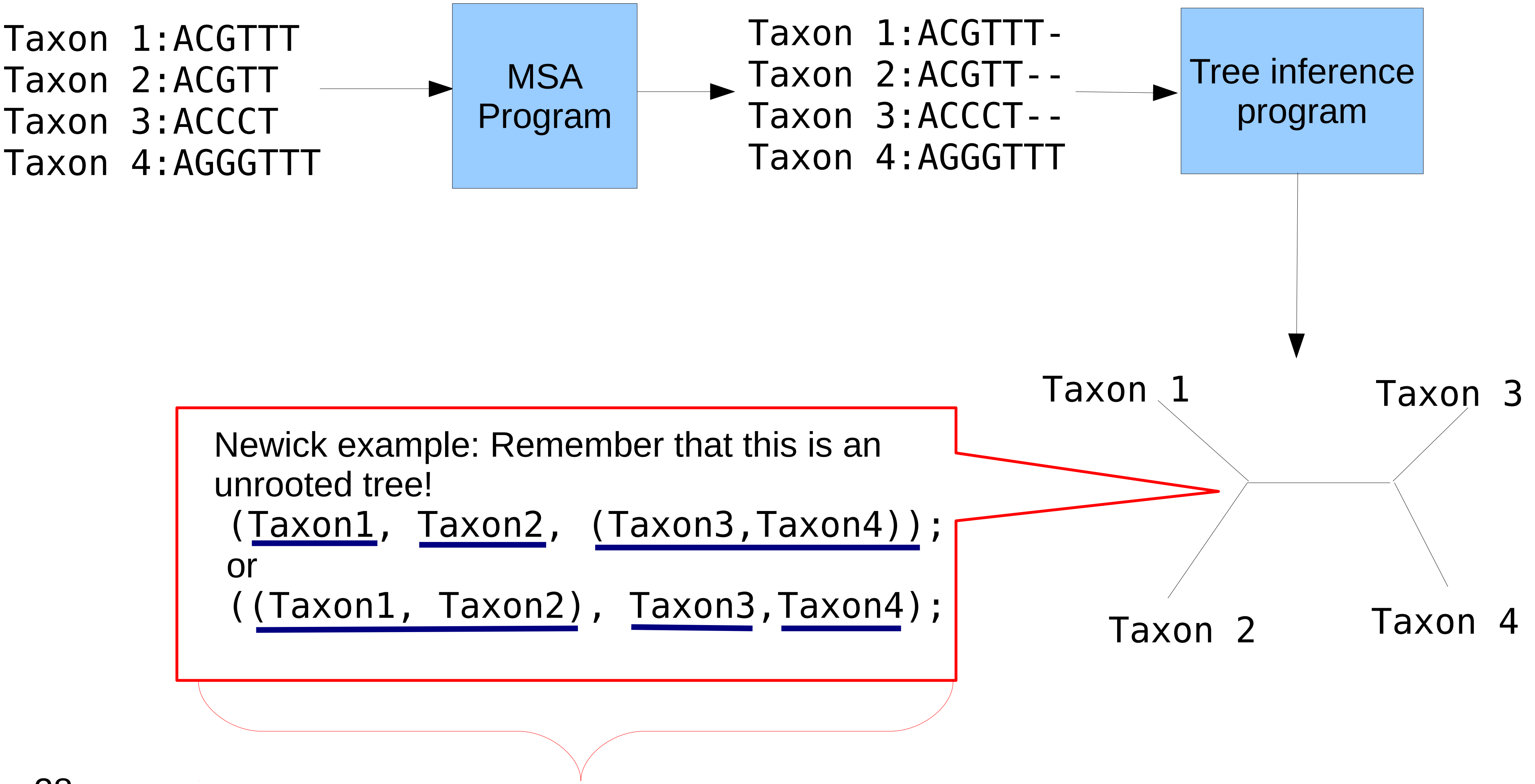
```
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT
```

Tree inference program

Taxon 1
Taxon 3
Taxon 2
Taxon 4

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

Taxon 1

Taxon 3

Taxon 2

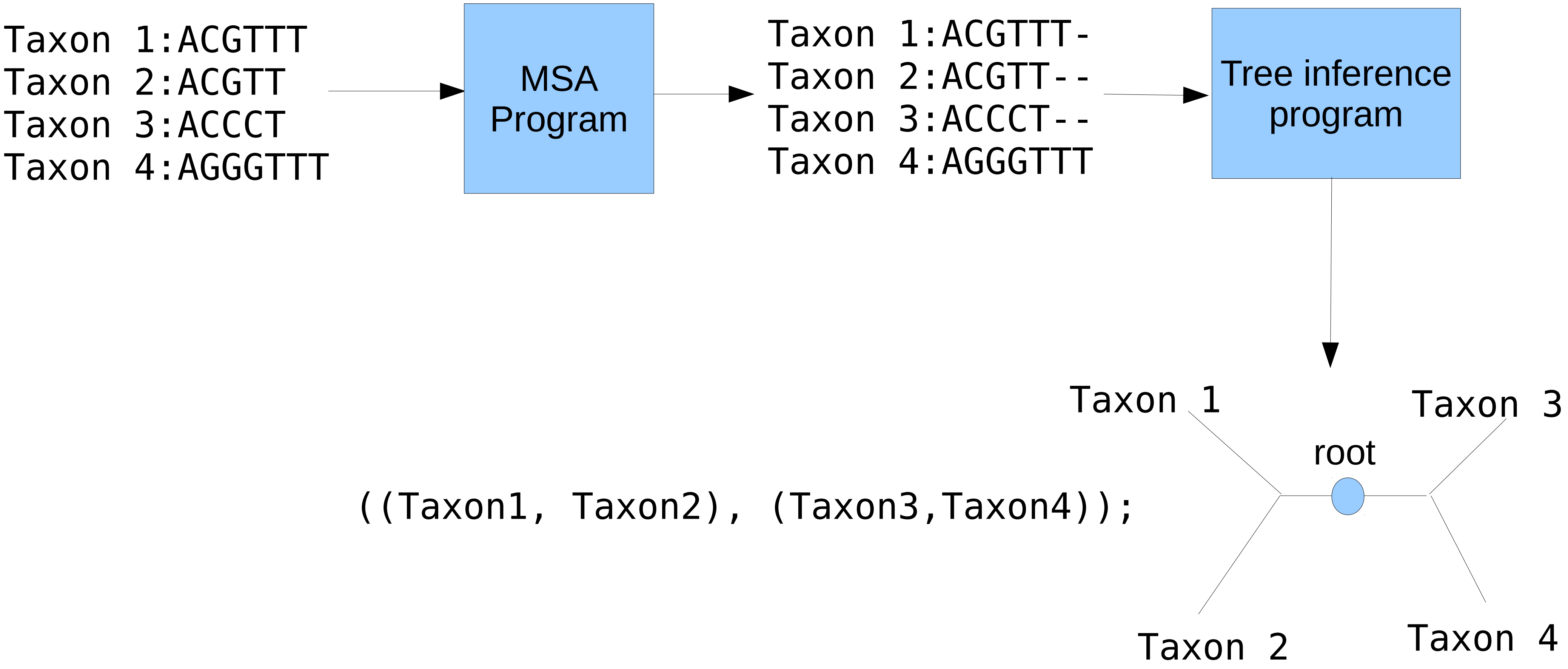
Taxon 4

Newick example: Remember that this is an unrooted tree!

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

Top level trifurcation

Tree Inference



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

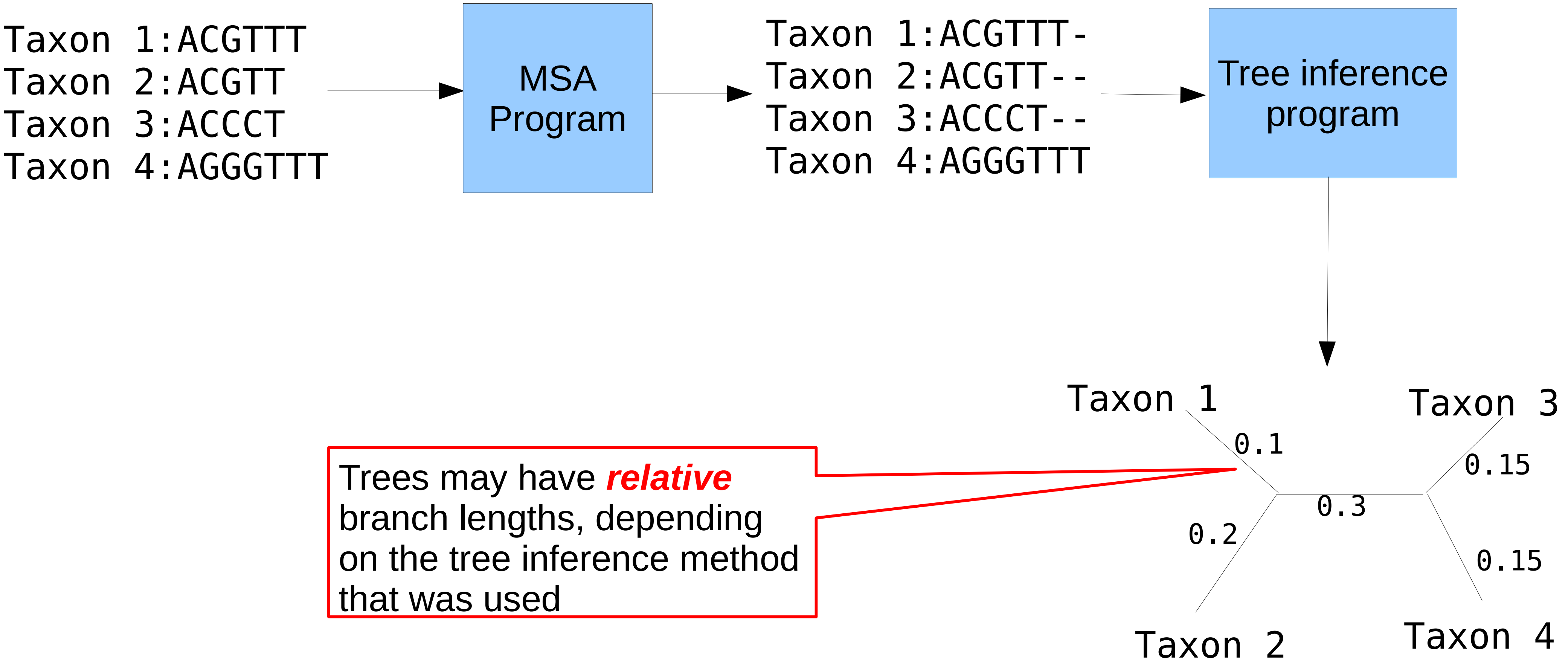
MSA Program

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

Tree inference program

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

Tree Inference



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

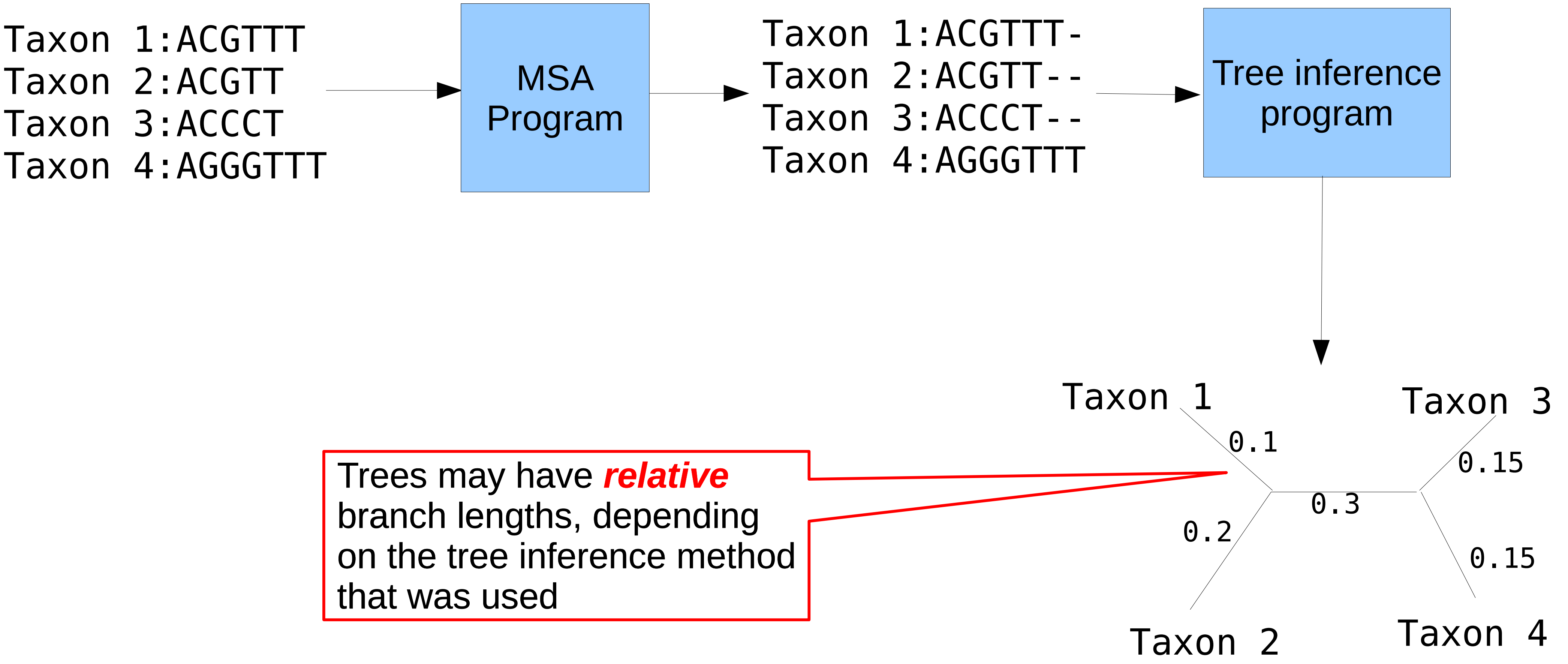
MSA Program

```
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
```

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

Newick format with branch lengths:

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

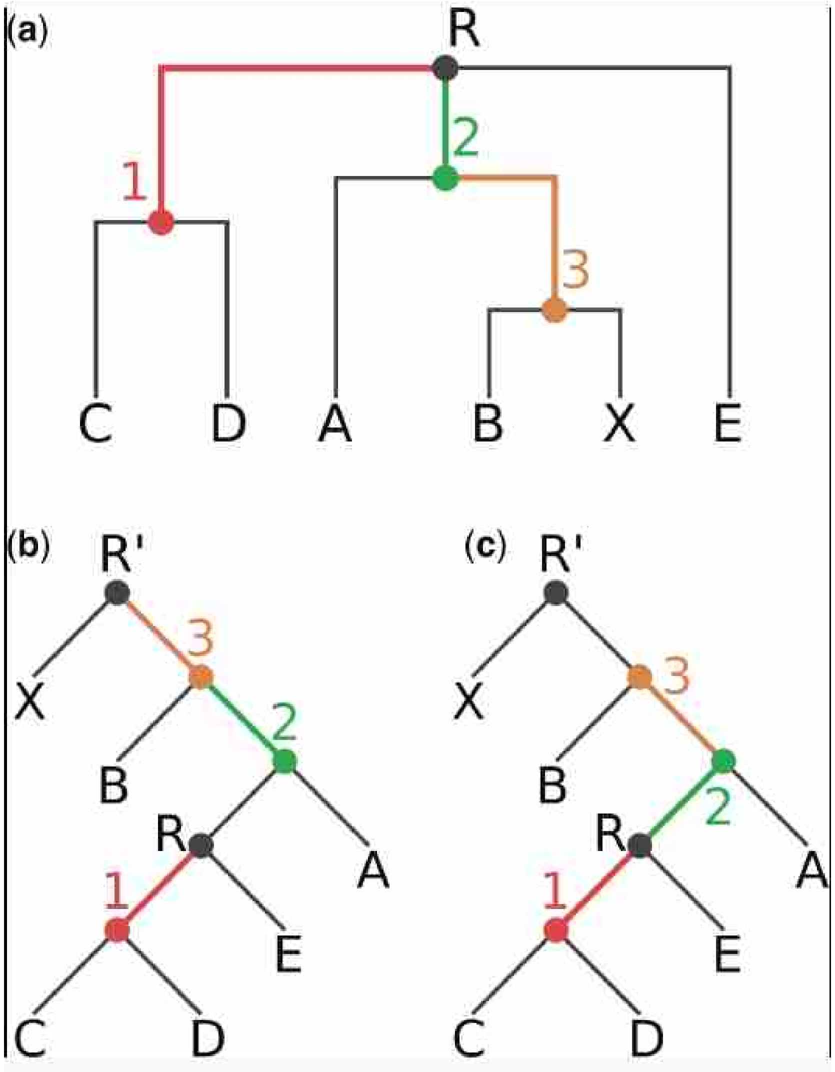
Problems with Newick tree format

- Except for branch length values: no way to associate meta-data to branch lengths
- However, there is important meta-data, e.g., branch support: how well is a branch in the tree supported?

  → ad hoc solution: represent branch support values as node meta-data!

  → this causes problems

Problems with Newick tree format

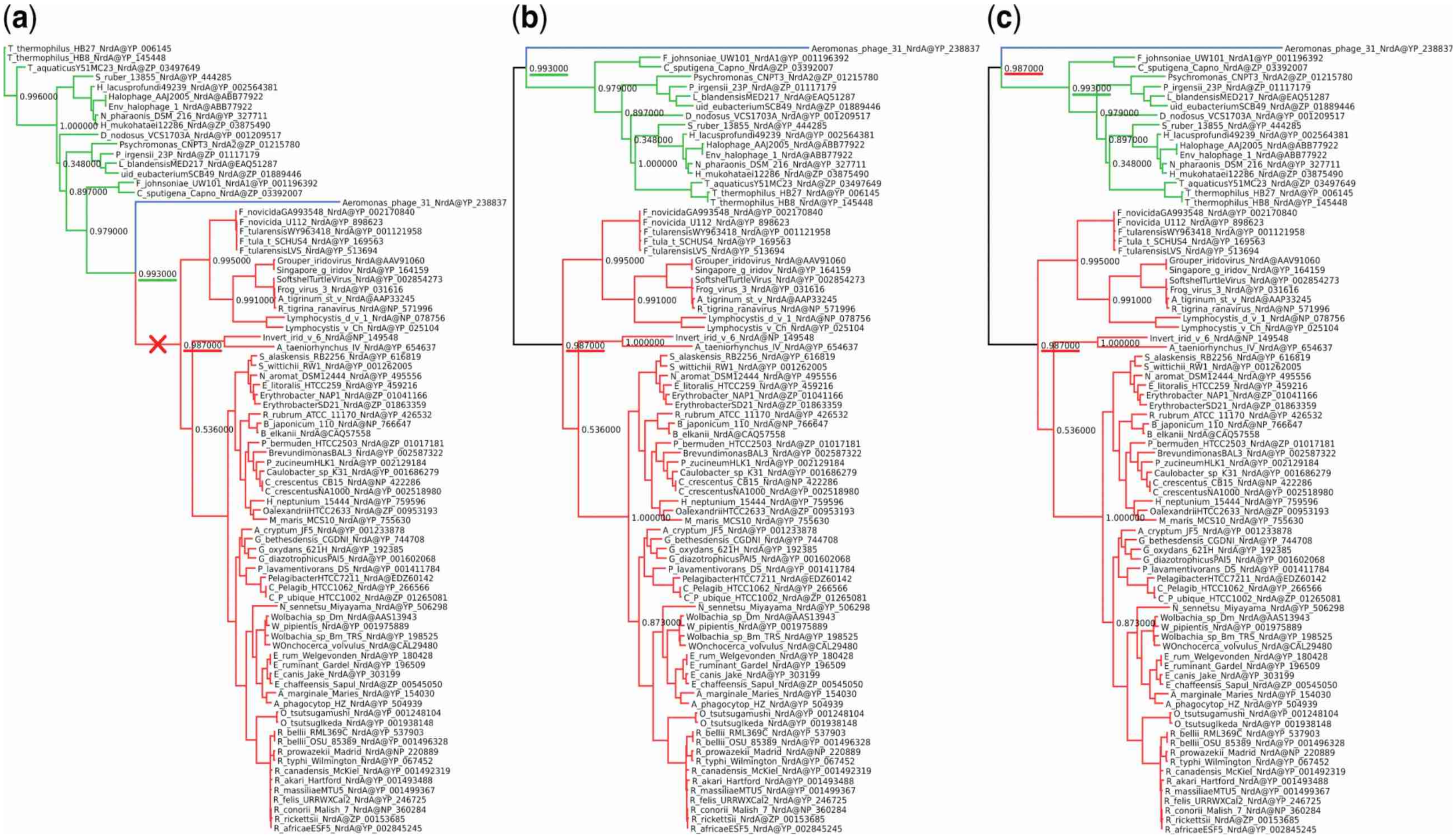


Branch support values represented as node meta-data can be assigned incorrectly to branches after re-rooting.

About 50% of the tools we checked had this Problem. For details see: https://academic.oup.com/mbe/article/34/6/1535/3077051

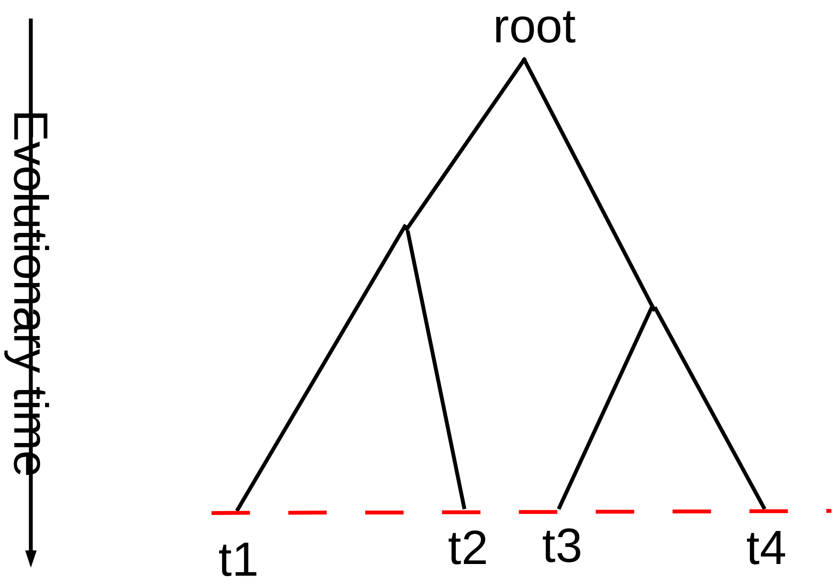
Which representation is correct?

A real example

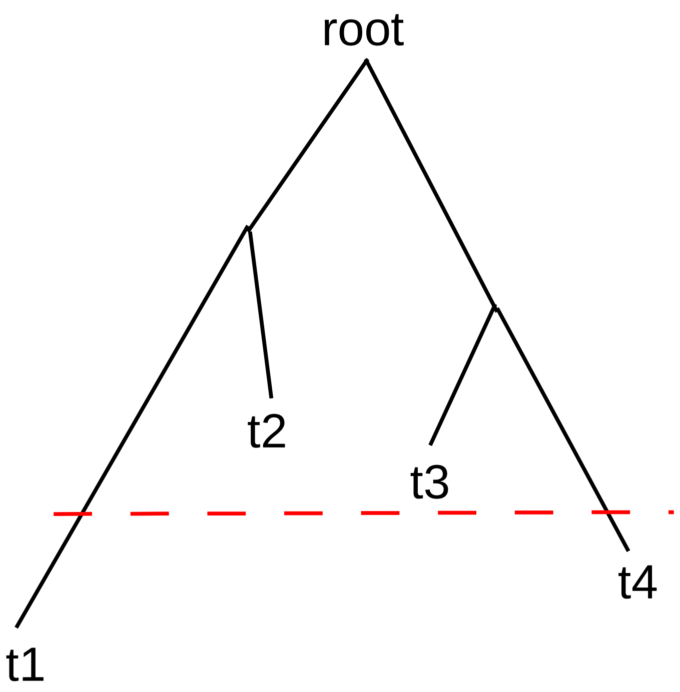

(a)
T_thermophilus_HB27_NrdA@YP_006145
T_thermophilus_HB8_NrdA@YP_145448
T_aquaticusY51MC23_NrdA@ZP_03497649
S_ruber_13855_NrdA@YP_444285
H_lacusprofundi49239_NrdA@YP_002564381
Halophage_AAJ2005_NrdA@ABB77922
Env_halophage_1_NrdA@ABB77922
N_pharaonis_DSM_216_NrdA@YP_327711
H_mukohataei12286_NrdA@ZP_03875490
D_nodosus_VCS1703A_NrdA@YP_001209517
Psychromonas_CNPT3_NrdA2@ZP_01215780
P_irgensii_23P_NrdA@ZP_01117179
L_blandensisMED217_NrdA@EAQ51287
uid_eubacteriumSCB49_NrdA@ZP_01889446
F_johnsoniae_UW101_NrdA1@YP_001196392
C_sputigena_Capno_NrdA@ZP_03392007
Aeromonas_phage_31_NrdA@YP_238837
F_novicidaGA993548_NrdA@YP_002170840
F_novicida_U112_NrdA@YP_898623
F_tularensisWY963418_NrdA@YP_001121958
F_tula_t_SCHUS4_NrdA@YP_169563
F_tularensisLVS_NrdA@YP_513694
Grouper_iridovirus_NrdA@AAV91060
Singapore_g_iridov_NrdA@YP_164159
SoftshelTurtleVirus_NrdA@YP_002854273
Frog_virus_3_NrdA@YP_031616
A_tigrinum_st_v_NrdA@AAP33245
R_tigrina_ranavirus_NrdA@NP_571996
Lymphocystis_d_v_1_NrdA@NP_078756
Lymphocystis_v_Ch_NrdA@YP_025104
Invert_irid_v_6_NrdA@NP_149548
A_taeniorhynchus_IV_NrdA@YP_654637
S_alaskensis_RB2256_NrdA@YP_616819
S_wittichii_RW1_NrdA@YP_001262005
N_aromat_DSM12444_NrdA@YP_495556
E_litoralis_HTCC259_NrdA@YP_459216
Erythrobacter_NAP1_NrdA@ZP_01041166
ErythrobacterSD21_NrdA@ZP_01863359
R_rubrum_ATCC_11170_NrdA@YP_426532
B_japonicum_110_NrdA@NP_766647
B_elkanii_NrdA@CAQ57558
P_bermuden_HTCC2503_NrdA@ZP_01017181
BrevundimonasBAL3_NrdA@YP_002587322
P_zucineumHLK1_NrdA@YP_002129184
Caulobacter_sp_K31_NrdA@YP_001686279
C_crescentus_CB15_NrdA@NP_422286
C_crescentusNA1000_NrdA@YP_002518980
H_neptunium_15444_NrdA@YP_759596
OalexandriiHTCC2633_NrdA@ZP_00953193
M_maris_MCS10_NrdA@YP_755630
A_cryptum_JF5_NrdA@YP_001233878
G_bethesdensis_CGDNI_NrdA@YP_744708
G_oxydans_621H_NrdA@YP_192385
G_diazotrophicusPAI5_NrdA@YP_001602068
P_lavamentivorans_DS_NrdA@YP_001411784
PelagibacterHTCC7211_NrdA@EDZ60142
C_Pelagib_HTCC1062_NrdA@YP_266566
C_P_ubique_HTCC1002_NrdA@ZP_01265081
N_sennetsu_Miyayama_NrdA@YP_506298
Wolbachia_sp_Dm_NrdA@AAS13943
W_pipientis_NrdA@YP_001975889
Wolbachia_sp_Bm_TRS_NrdA@YP_198525
WOnchocerca_volvulus_NrdA@CAL29480
E_rum_Welgevonden_NrdA@YP_180428
E_ruminant_Gardel_NrdA@YP_196509
E_canis_Jake_NrdA@YP_303199
E_chaffeensis_Sapul_NrdA@ZP_00545050
A_marginale_Maries_NrdA@YP_154030
A_phagocytop_HZ_NrdA@YP_504939
O_tsutsugamushi_NrdA@YP_001248104
O_tsutsugIkeda_NrdA@YP_001938148
R_bellii_RML369C_NrdA@YP_537903
R_bellii_OSU_85389_NrdA@YP_001496328
R_prowazekii_Madrid_NrdA@NP_220889
R_typhi_Wilmington_NrdA@YP_067452
R_canadensis_McKiel_NrdA@YP_001492319
R_akari_Hartford_NrdA@YP_001493488
R_massiliaeMTU5_NrdA@YP_001499367
R_felis_URRWXCal2_NrdA@YP_246725
R_conorii_Malish_7_NrdA@NP_360284
R_rickettsii_NrdA@ZP_00153685
R_africaeESF5_NrdA@YP_002845245
0.996000
1.000000
0.348000
0.897000
0.979000
0.993000
0.995000
0.991000
✕
0.987000
0.536000
(b)
Aeromonas_phage_31_NrdA@YP_238837
F_johnsoniae_UW101_NrdA1@YP_001196392
C_sputigena_Capno_NrdA@ZP_03392007
Psychromonas_CNPT3_NrdA2@ZP_01215780
P_irgensii_23P_NrdA@ZP_01117179
L_blandensisMED217_NrdA@EAQ51287
uid_eubacteriumSCB49_NrdA@ZP_01889446
D_nodosus_VCS1703A_NrdA@YP_001209517
S_ruber_13855_NrdA@YP_444285
H_lacusprofundi49239_NrdA@YP_002564381
Halophage_AAJ2005_NrdA@ABB77922
Env_halophage_1_NrdA@ABB77922
N_pharaonis_DSM_216_NrdA@YP_327711
H_mukohataei12286_NrdA@ZP_03875490
T_aquaticusY51MC23_NrdA@ZP_03497649
T_thermophilus_HB27_NrdA@YP_006145
T_thermophilus_HB8_NrdA@YP_145448
0.993000
0.979000
0.897000
0.348000
1.000000
0.995000
0.991000
0.987000
1.000000
0.536000
1.000000
0.873000
(c)
Aeromonas_phage_31_NrdA@YP_238837
0.987000
0.993000
0.979000
0.897000
0.348000
0.995000
0.991000
0.987000
1.000000
0.536000
1.000000
0.873000

a) original tree
b) re-rooted tree with shifted support values
c) re-rooted tree with correct support values

Tree Shapes

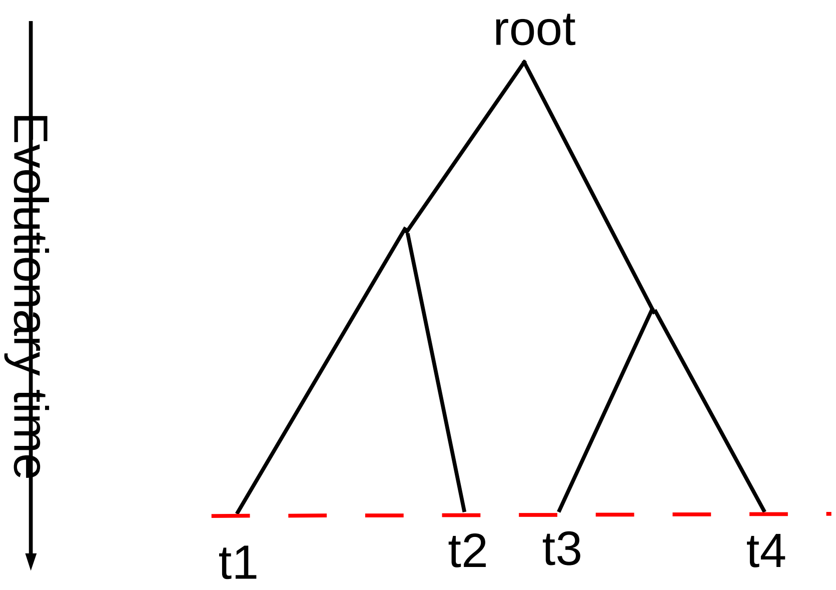


Ultrametric tree

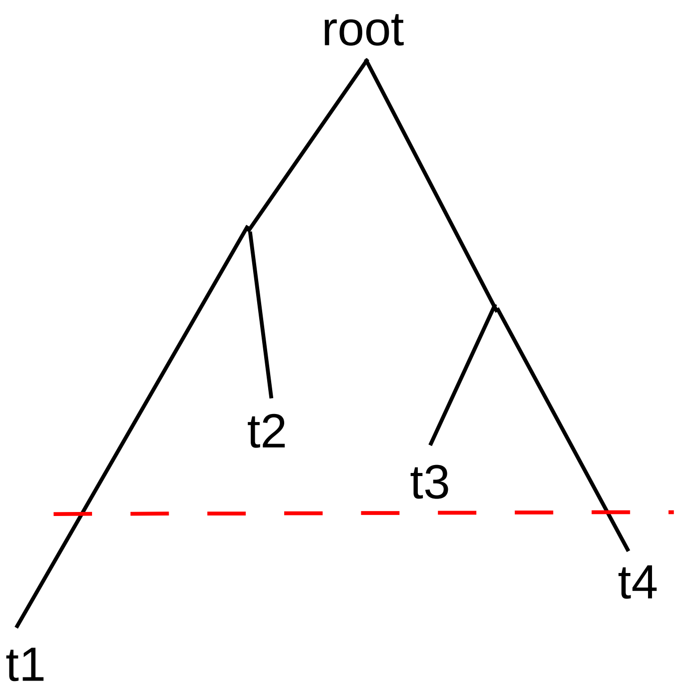


Non-ultrametric tree

Tree Shapes



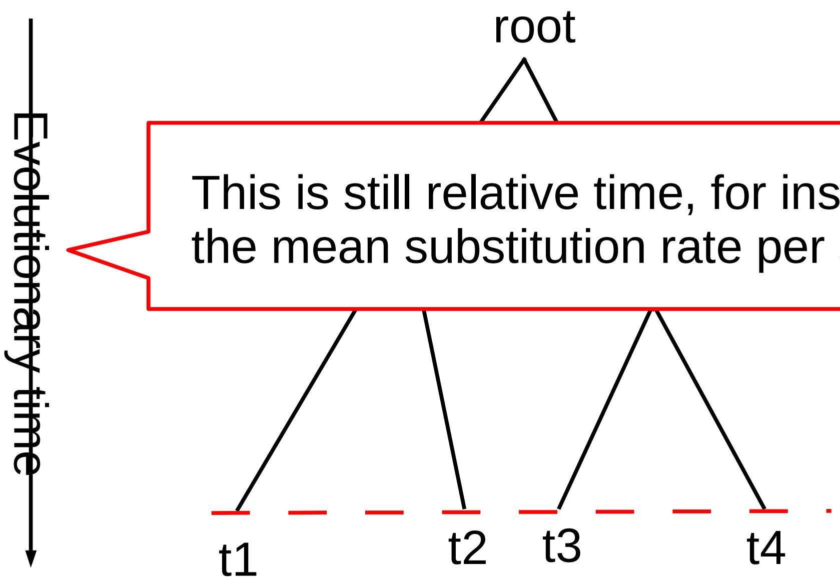
Ultrametric tree



Non-ultrametric tree

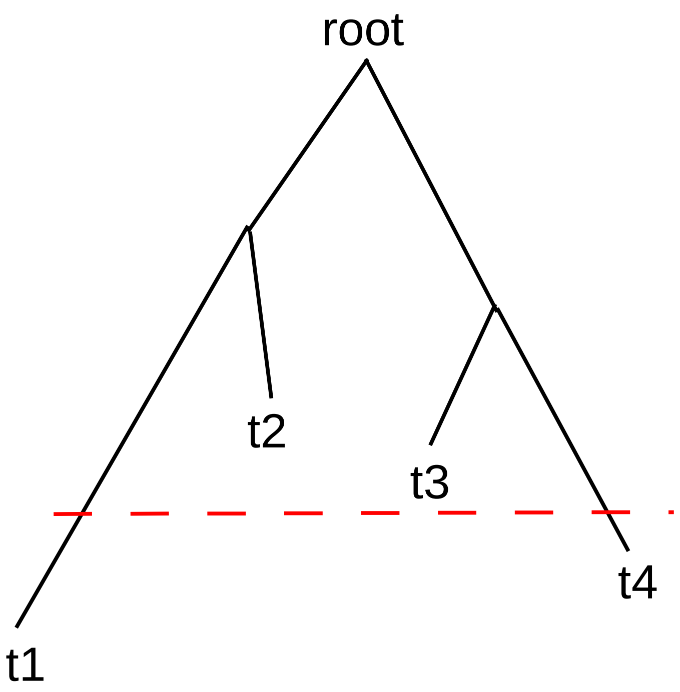
Most tree inference models/algorithms/programs produce non-ultrametric trees

Tree Shapes



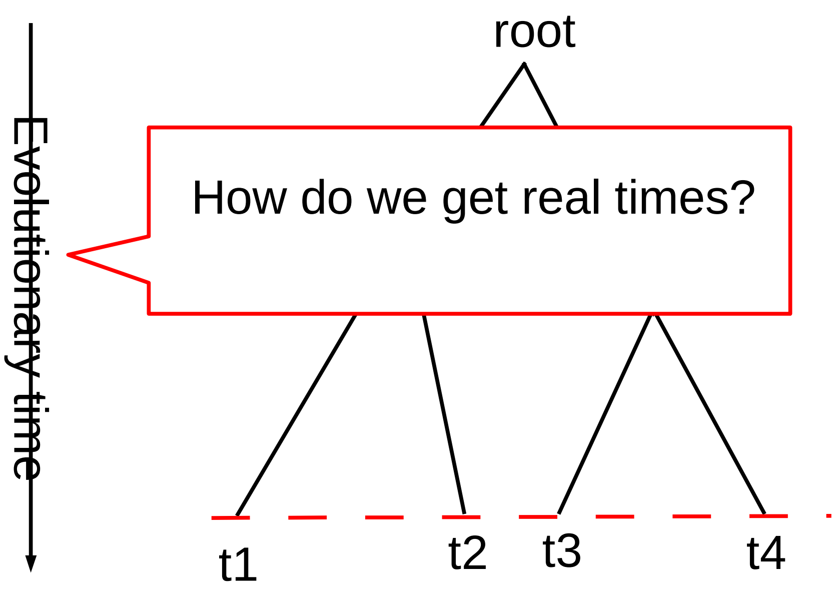
This is still relative time, for instance the mean substitution rate per site

Ultrametric tree



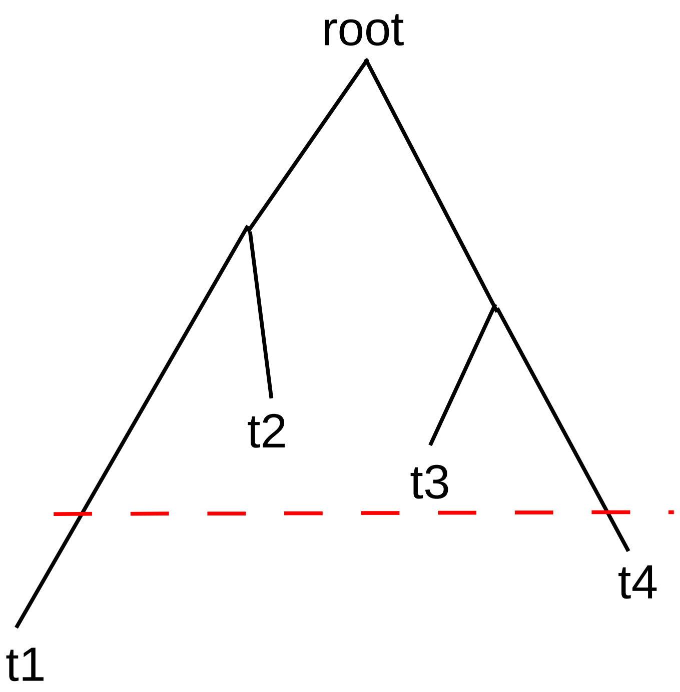
Non-ultrametric tree

Tree Shapes



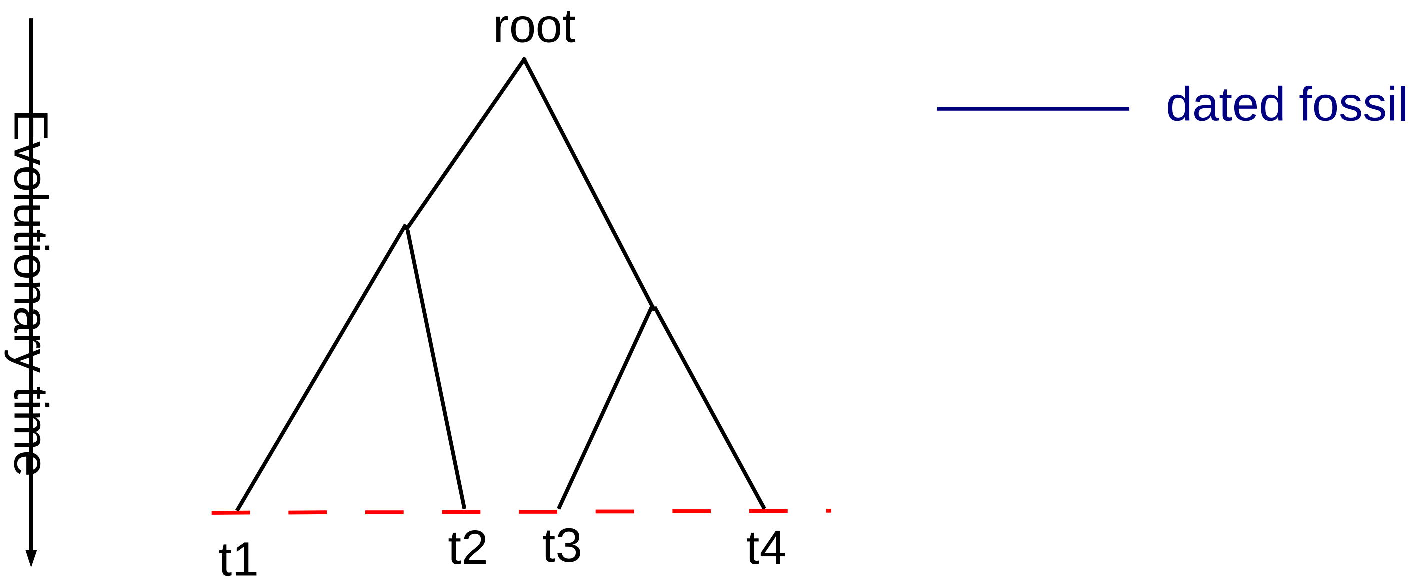
How do we get real times?

Ultrametric tree



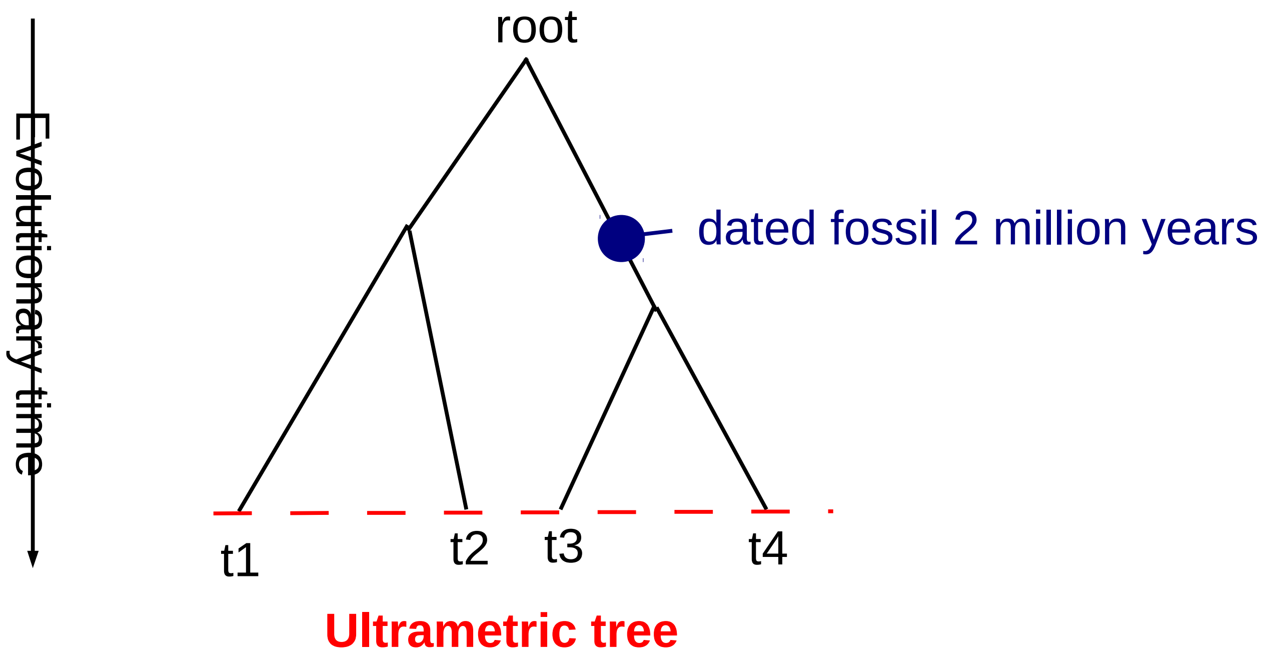
Non-ultrametric tree

Dating Trees

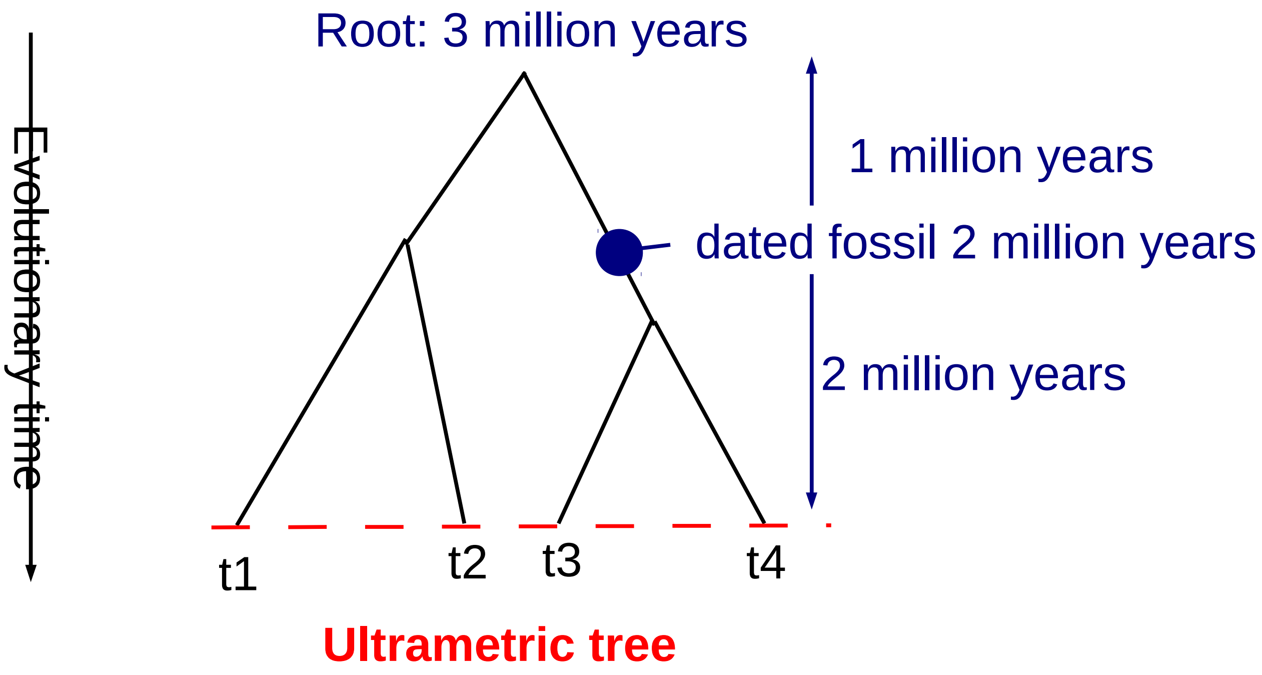


Ultrametric tree

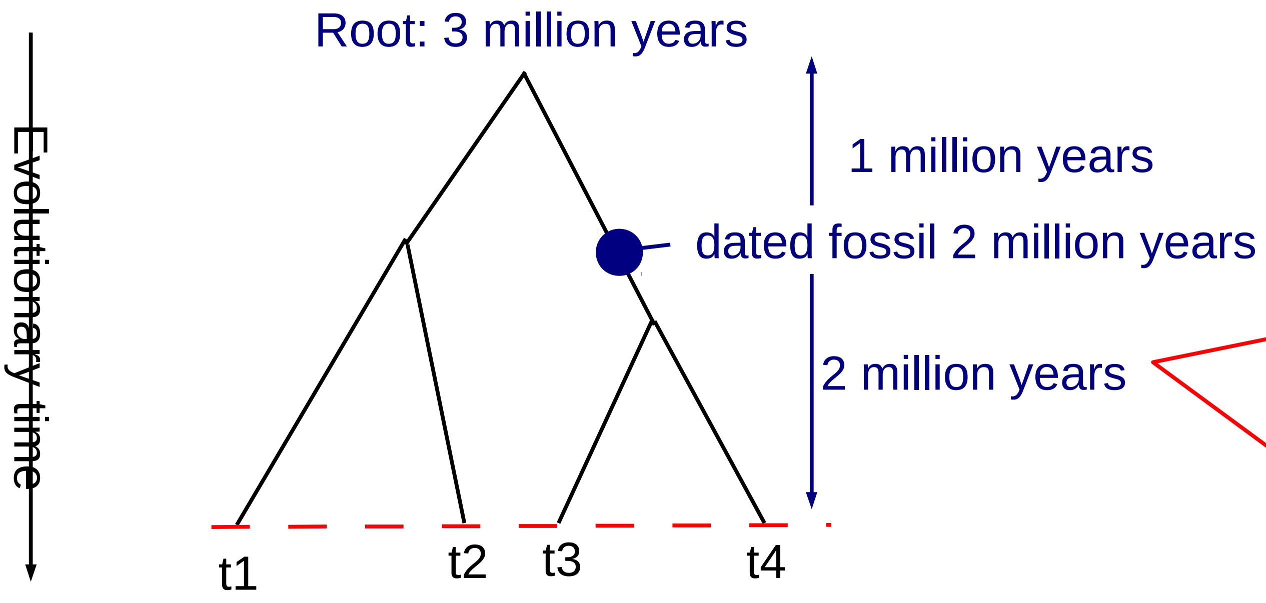
Dating Trees



Dating Trees



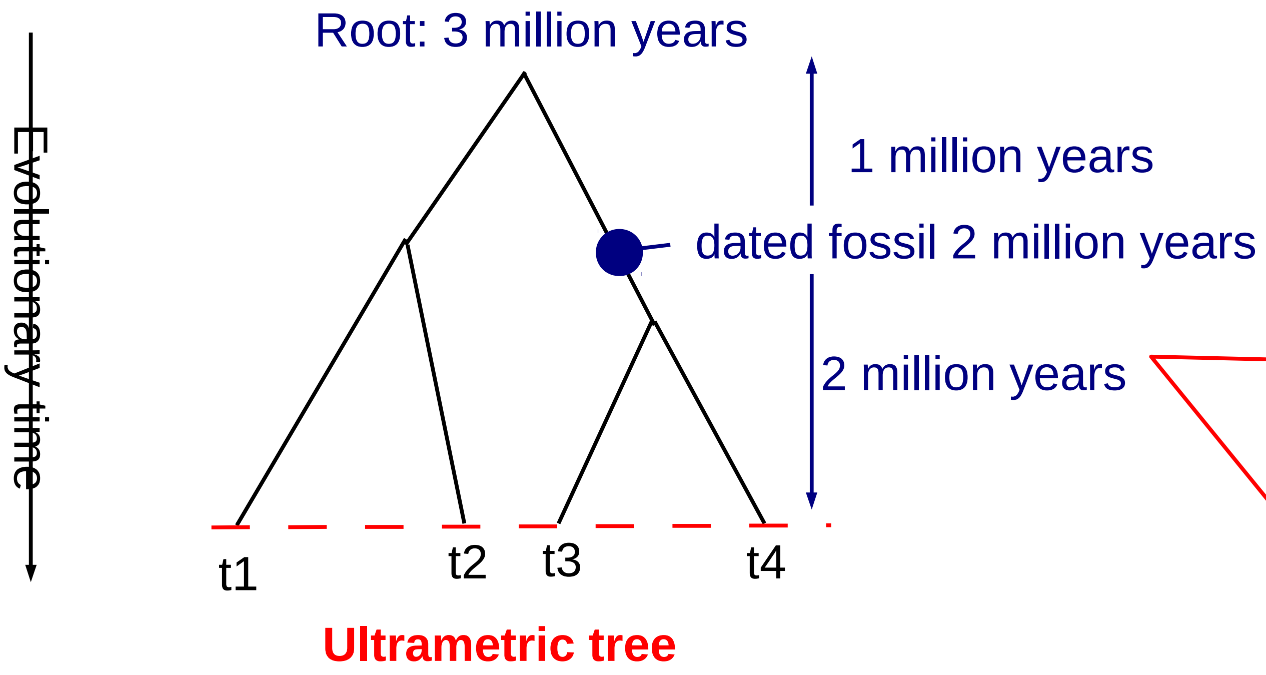
Dating Trees



Ultrametric tree

We need a rooted & ultrametric tree!
→ rooting with outgroups
→ ultrametricity with programs for *divergence time estimation*
→ active research area
→ most codes rely on the phylogenetic likelihood function and Bayesian Statistics (MCMC methods)

Dating Trees



Ultrametric tree

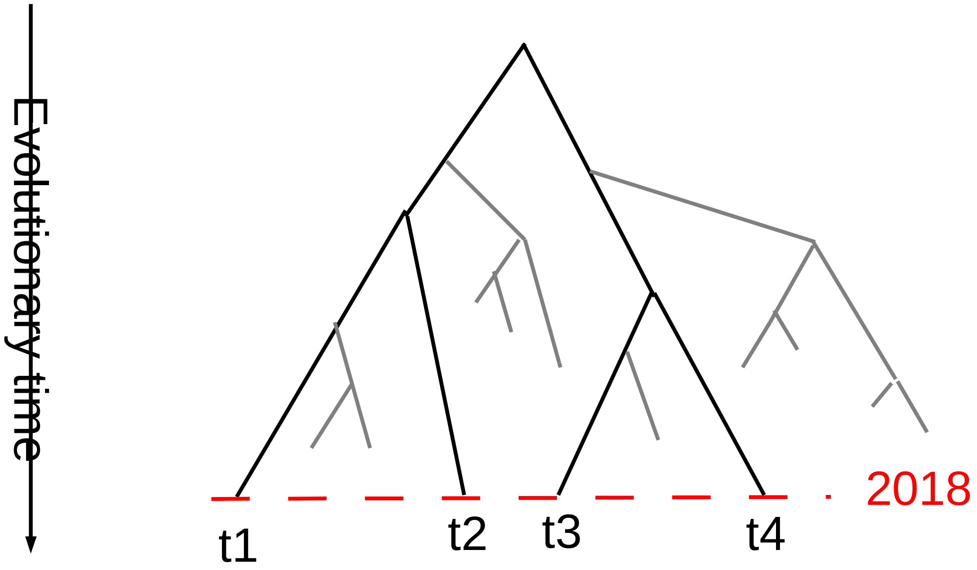
But how do we place the fossil?
→ typically no DNA data available

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”)!

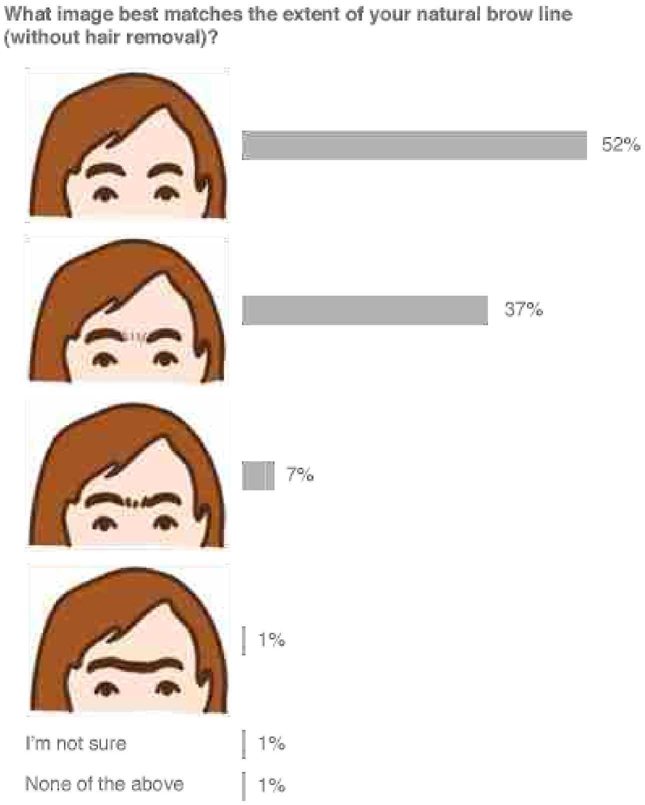
Remember that we deal with extant species!



Ultrametric tree

Morphological Traits

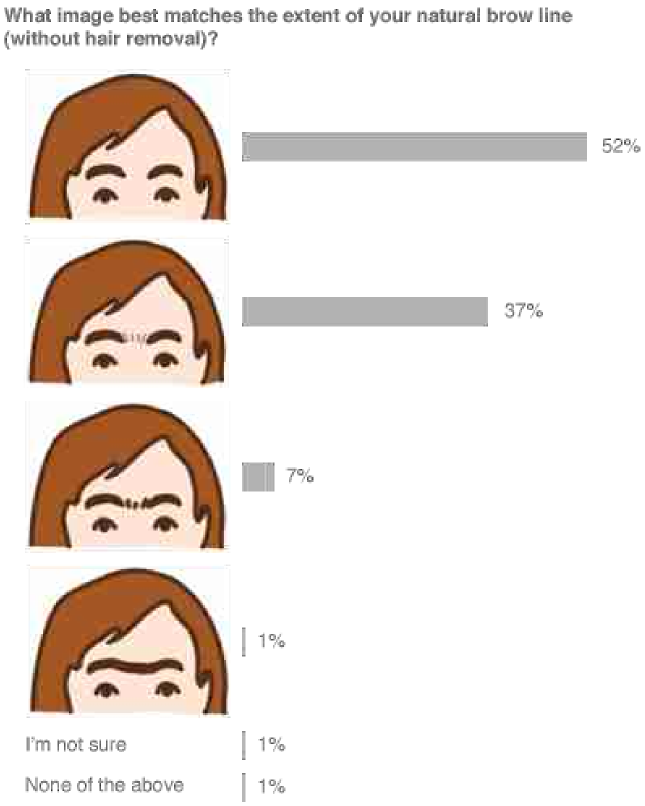
```
t1: 1000
t2: 0100
t3: 0010
T4: 0001

or:

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



What image best matches the extent of your natural brow line (without hair removal)?

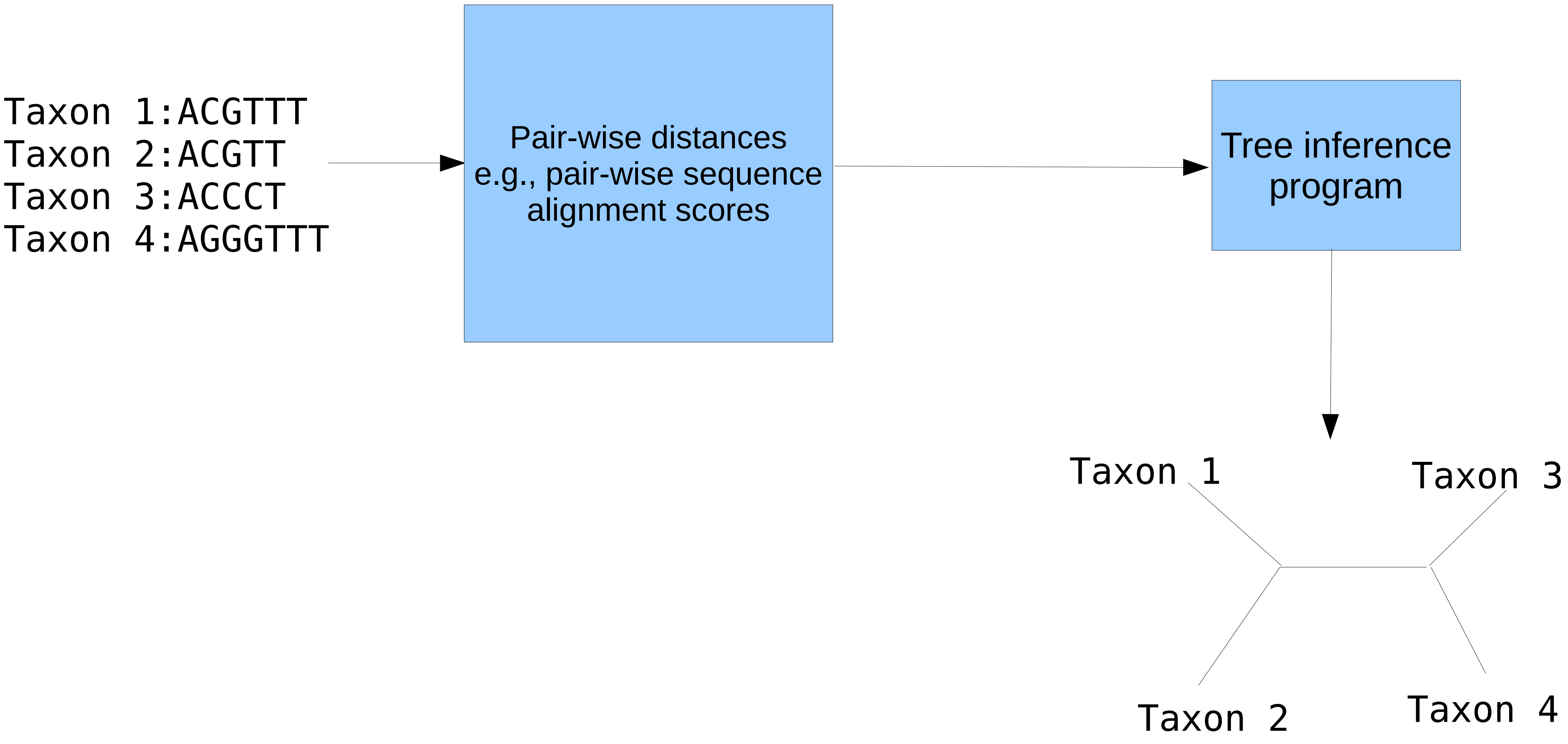
Morphological Traits

```
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

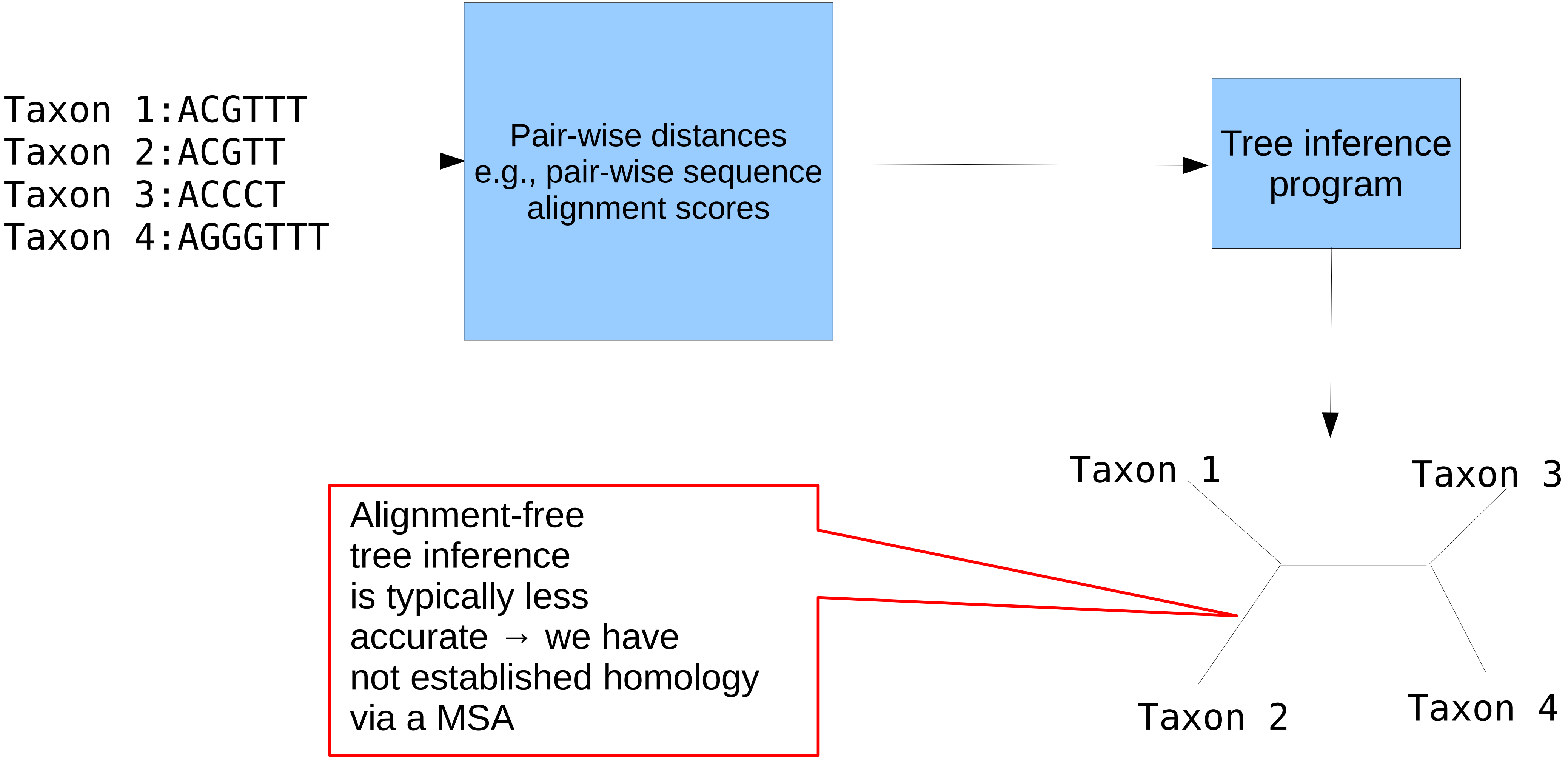


```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

Pair-wise distances
e.g., pair-wise sequence
alignment scores

Tree inference
program

Alignment-Free Tree Inference



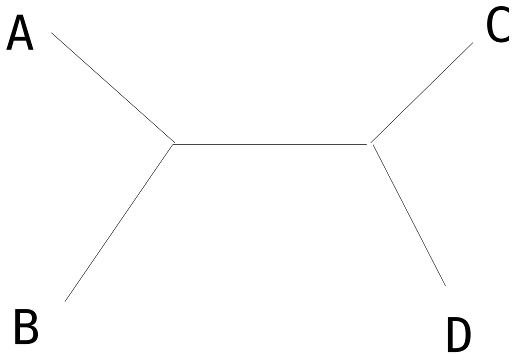
```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

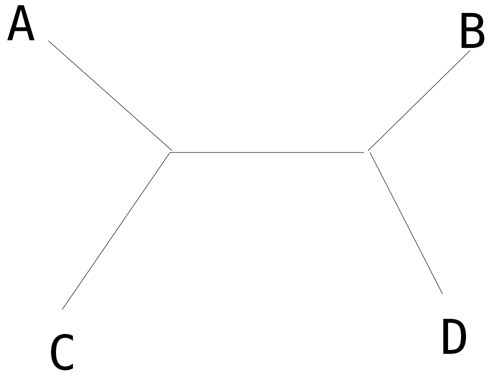
Pair-wise distances
e.g., pair-wise sequence
alignment scores

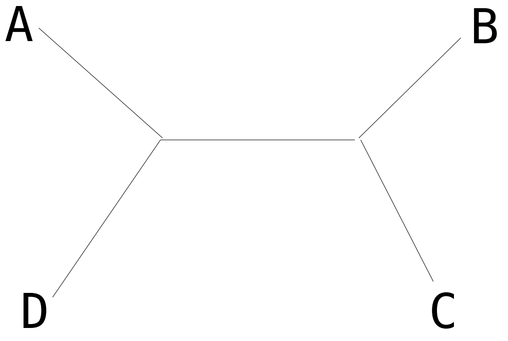
Tree inference
program

Alignment-free tree inference is typically less accurate → we have not established homology via a MSA

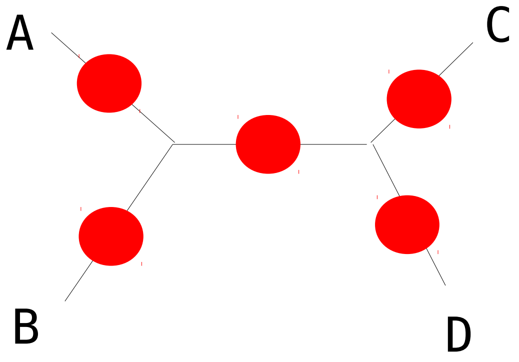
How many unrooted 4-taxon trees exist?

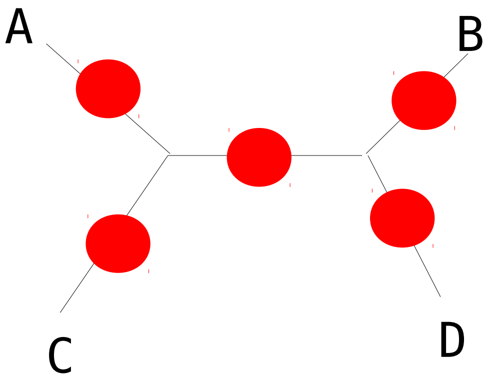


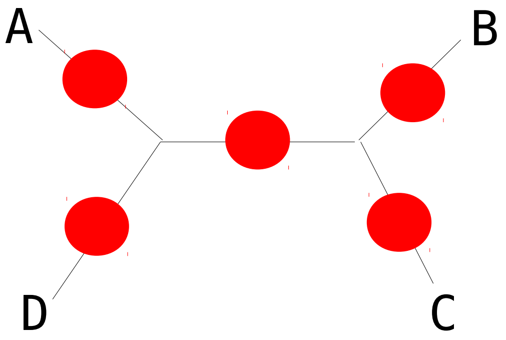




How many rooted 4-taxon trees exist?



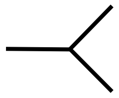




Tree Counts

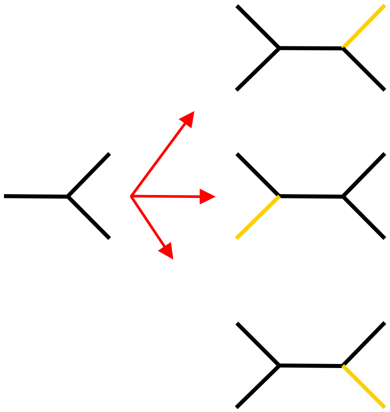
- Unrooted binary trees
  - *4* taxa → *3* distinct trees
  - A tree with *n* taxa has *n-2* inner nodes
  - And *2n-3* branches
- Rooted binary trees
  - *4* taxa → *3* unrooted trees * *5* branches each (rooting points) = *15* trees
  - *n-1* inner nodes
  - *2n-2* branches

The number of trees



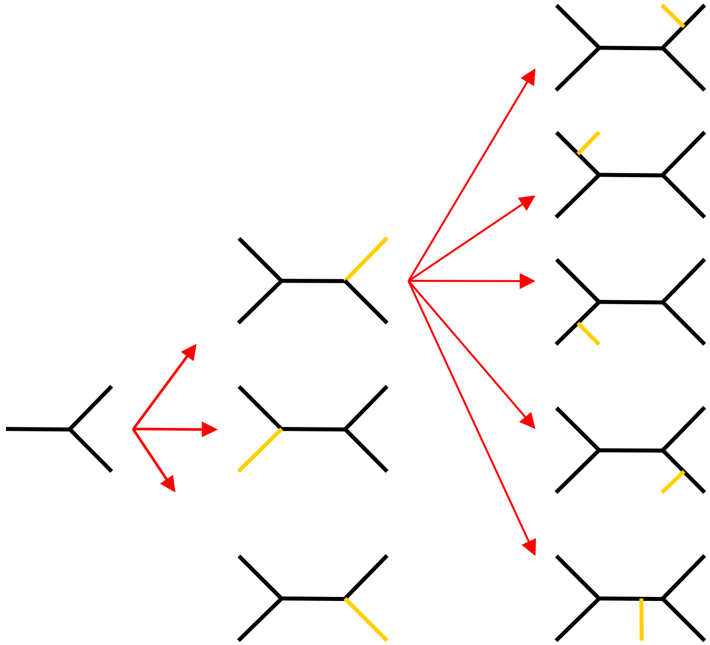
3 taxa = 1 tree

The number of trees



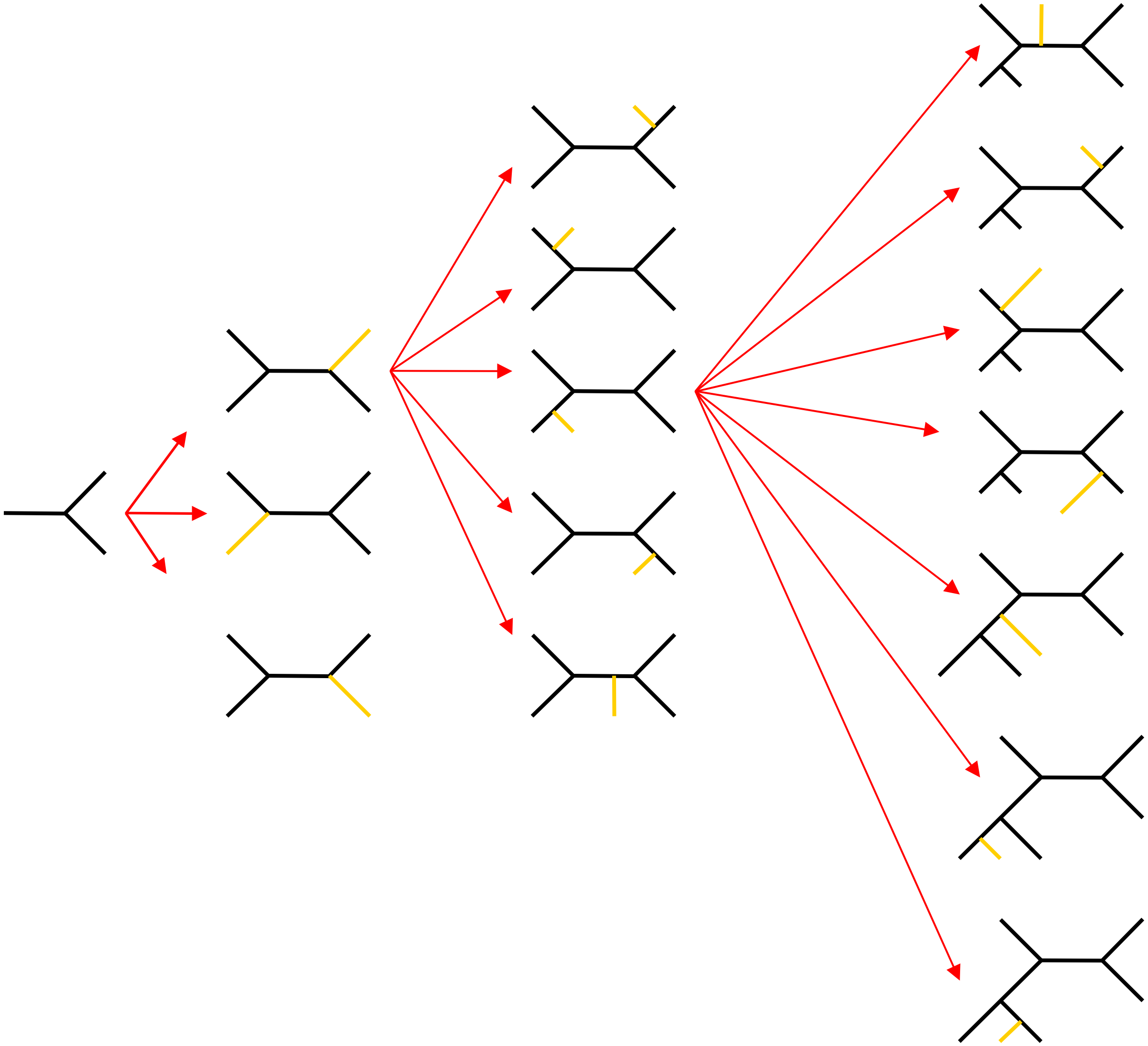
4 taxa: 3 trees
u: # trees of size 4-1 := 1
v: # branches in a tree of size 4-1 := 3
Number of unrooted binary trees with 4 taxa: u * v = 3

The number of trees



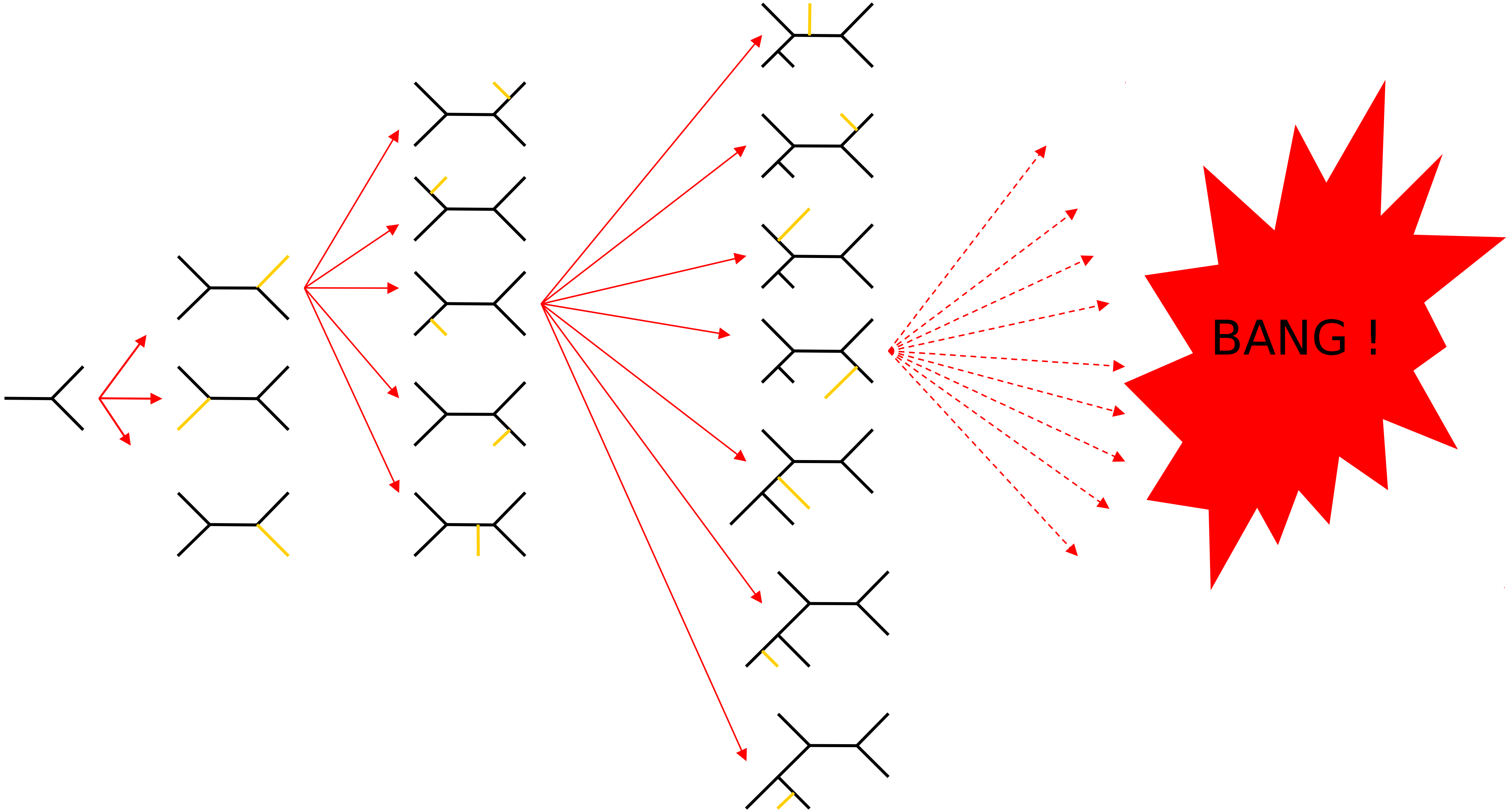
5 taxa: 15 trees
u = 3
v = 5
Number of unrooted trees with 5 taxa: 3 * 5 = 15

The number of trees

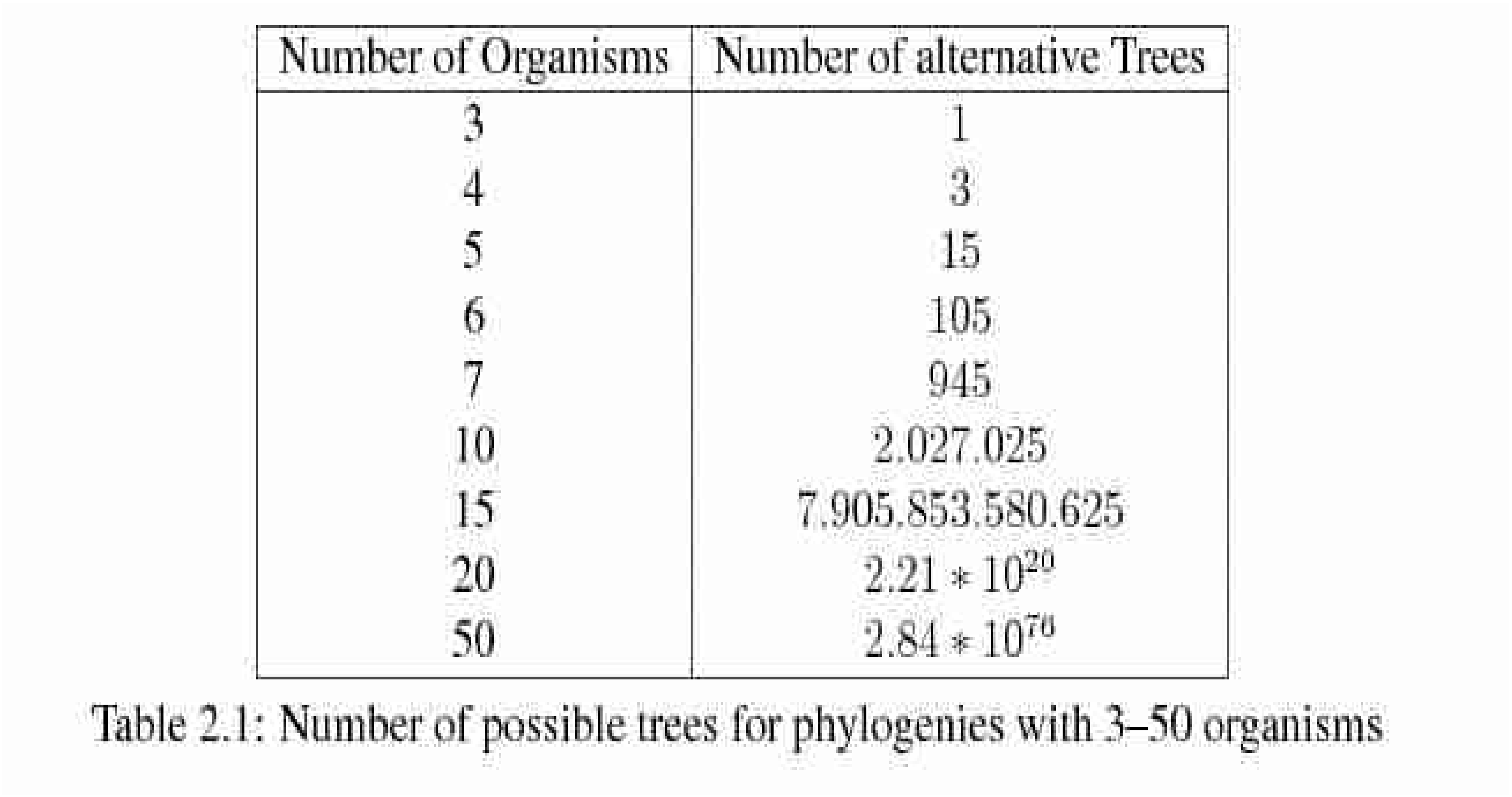


6 taxa: 105 trees
u = 15
v = 7
u * v = 105

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

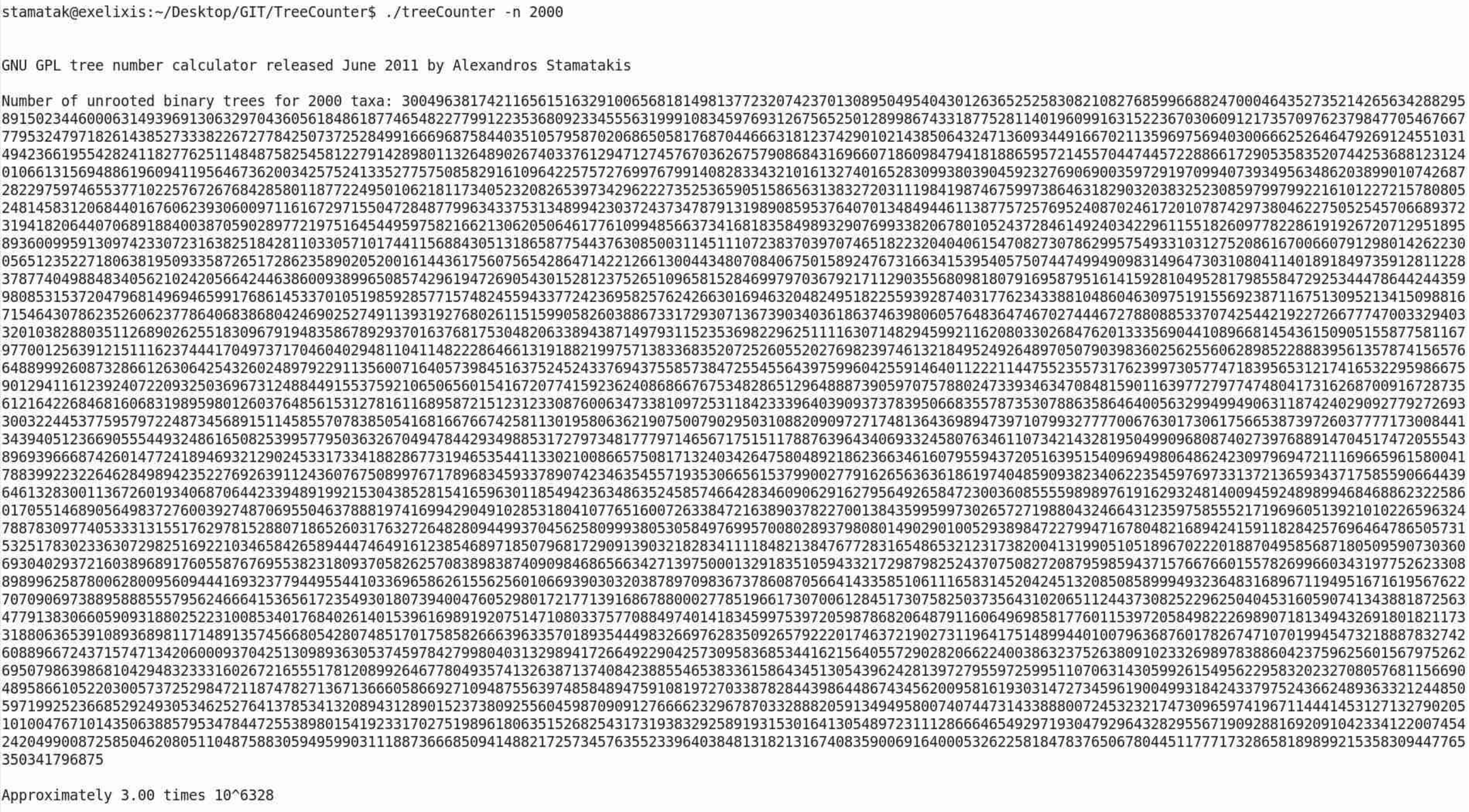
Equation for the number of unrooted trees

- 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 all 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: 300496381742116561516329100656818149813772320742370130895049540430126365252583082108276859966882470004643527352142656342882 95
891502344600063149396913063297043605618486187746548227799122353680923345556319991083459769312675652501289986743318775281140196099163152236703060912173570976237984770546766 7
779532479718261438527333822672778425073725284991666968758440351057958702068650581768704466631812374290102143850643247136093449166702113596975694030066625264647926912455103 1
494236619554282411827762511484875825458122791428980113264890267403376129471274576703626757908684316966071860984794181886595721455704474457228866172905358352074425368812312 4
010661315694886196094119564673620034257524133527757508582916109642257572769976799140828334321016132740165283099380390459232769069003597291970994073934956348620389901074268 7
282297597465537710225767267684285801187722495010621811734052320826539734296222735253659051586563138327203111984198746759973864631829032038325230859799799221610122721578080 5
248145831206844016760623930600971161672971550472848779963433753134899423037243734787913198908595376407013484944611387757257695240870246172010787429738046227505254570668937 2
319418206440706891884003870590289772197516454495975821662130620506461776109948566373416818358498932907699338206780105243728461492403422961155182609778228619192672071295189 5
893600995913097423307231638251842811033057101744115688430513186587754437630850031145111072383703970746518223204040615470827307862995754933103127520861670066079129801426223 0
056512352271806381950933587265172862358902052001614436175607565428647142212661300443480708406750158924767316634153954057507447499490983149647303108041140189184973591281122 8
378774049884834056210242056642446386009389965085742961947269054301528123752651096581528469979703679217112903556809818079169587951614159281049528179855847292534447864424435 9
980853153720479681496946599176861453370105198592857715748245594337724236958257624266301694632048249518225593928740317762343388104860463097519155692387116751309521341509881 6
715464307862352606237786406838680424690252749113931927680261151599058260388673317293071367390340361863746398060576483647467027444672788088533707425442192272667774700332940 3
320103828803511268902625518309679194835867892937016376817530482063389438714979311523536982296251111630714829459921162080330268476201333569044108966814543615090515587758116 7
977001256391215111623744417049737170460402948110411482228646613191882199757138336835207252605520276982397461321849524926489705079039836025625560628985228883956135787415657 6
648899926087328661263064254326024897922911356007164057398451637524524337694375585738472554556439759960425591464011222114475523557317623997305774718395653121741653229598667 5
901294116123924072209325036967312488449155375921065065601541672077415923624086866767534828651296488873905970757880247339346347084815901163977279774748041731626870091672873 5
612164226846816068319895980126037648561531278161168958721512312330876006347338109725311842333964039093737839506683557873530788635864640056329949949063118742402909277927269 3
300322445377595797224873456891511458557078385054168166766742581130195806362190750079029503108820909727174813643698947397107993277770067630173061756653873972603777717300844 1
343940512366905554493248616508253995779503632670494784429349885317279734817779714656717515117887639643406933245807634611073421432819504990968087402739768891470451747205554 3
896939666874260147724189469321290245331733418828677319465354411330210086657508171324034264758048921862366346160795594372051639515409694980648624230979694721116966596158004 1
788399223226462849894235227692639112436076750899767178968345933789074234635455719353066561537990027791626563636186197404859093823406223545976973313721365934371758559066443 9
646132830011367260193406870644233948919921530438528154165963011854942363486352458574664283460906291627956492658472300360855559898976191629324814009459248989946846886232258 6
017055146890564983727600392748706955046378881974169942904910285318041077651600726338472163890378227001384359959973026572719880432466431235975855521719696051392101022659632 4
788783097740533313155176297815288071865260317632726482809449937045625809993805305849769957008028937980801490290100529389847227994716780482168942415911828425769646478650573 1
532517830233630729825169221034658426589444746491612385468971850796817290913903218283411118482138476772831654865321231738200413199051051896702220188704958568718050959073036 0
693040293721603896891760558767695538231809370582625708389838740909846865663427139750001329183510594332172987982524370750827208795985943715766766015578269966034319775262330 8
898996258780062800956094441693237794495544103369658626155625601066939030320387897098367378608705664143358510611165831452042451320850858999949323648316896711949516716195676 22
707090697388958885557956246664153656172354930180739400476052980172177139168678800027785196617307006128451730758250373564310206511244373082522962504045316059074134388187256 3
477913830660590931880252231008534017684026140153961698919207514710803375770884974014183459975397205987868206487911606496985817760115397205849822269890718134943269180182117 3
318806365391089368981171489135745668054280748517017585826663963357018935444983266976283509265792220174637219027311964175148994401007963687601782674710701994547321888783274 2
608896672437157471342060000937042513098936305374597842799804031329894172664922904257309583685344162156405572902820662240038632375263809102332698978388604237596256015679752 62
695079863986810429483233316026721655517812089926467780493574132638713740842388554653833615864345130543962428139727955972599511070631430599261549562295832023270805768115669 0
489586610522030057372529847211874782713671366605866927109487556397485848947591081972703387828443986448674345620095816193031472734596190049931842433797524366248936332124485 0
597199252366852924930534625276413785341320894312890152373809255604598709091276666232967870332888205913494958007407447314338880072453232174730965974196711444145312713279020 5
101004767101435063885795347844725538980154192331702751989618063515268254317319383292589193153016413054897231112866664654929719304792964328295567190928816920910423341220074 54
242049900872585046208051104875883059495990311188736668509414882172573457635523396403848131821316740835900691640005326225818478376506780445117771732865818989921535830944776 5
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

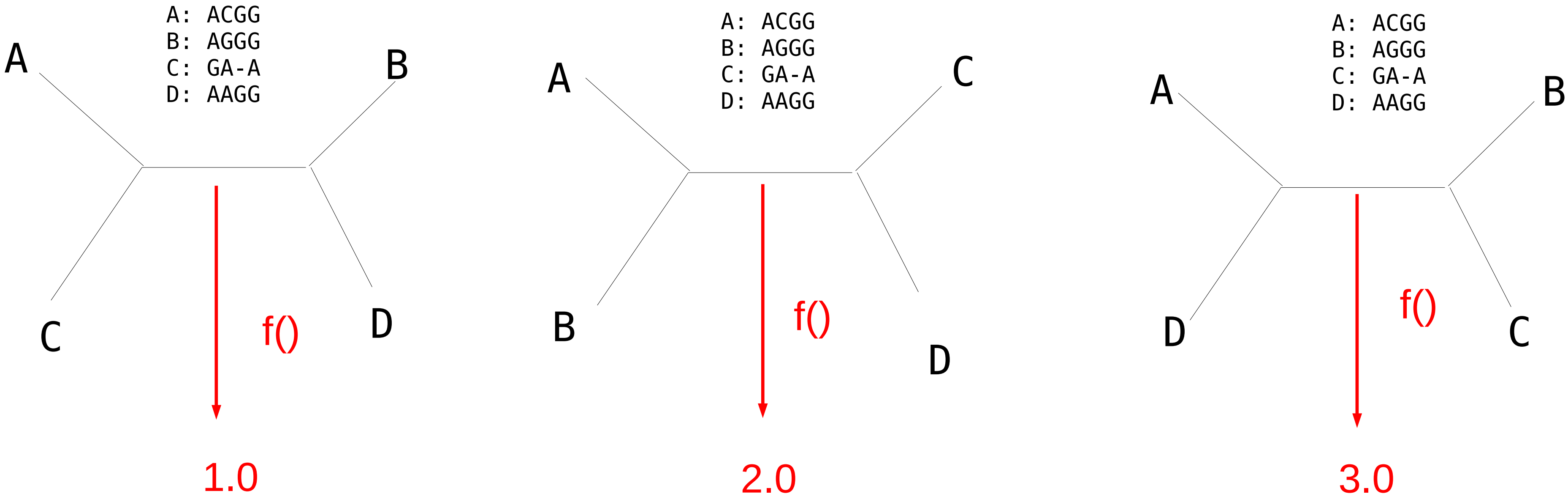
→ Has anybody already used GNU GMP?

Scoring Trees

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

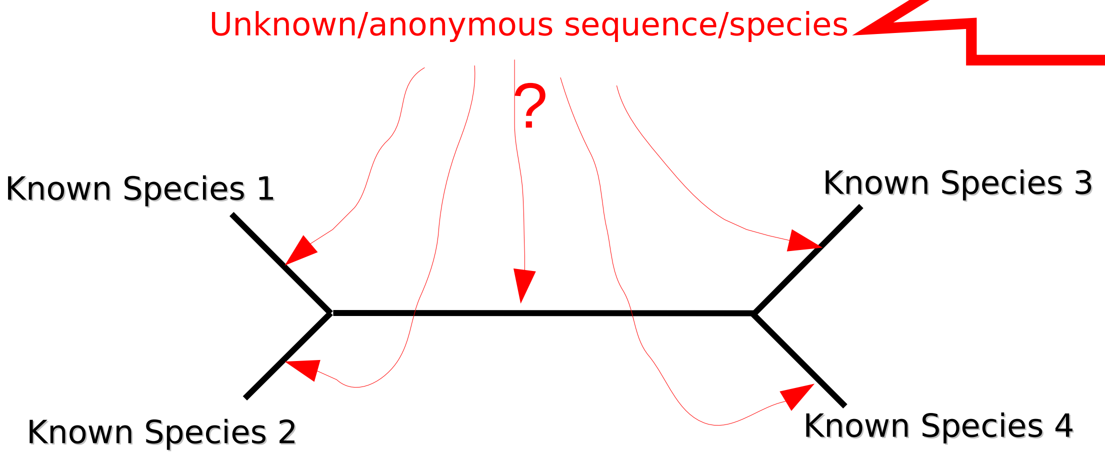
→ we need some scoring criterion f() to evaluate them

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



What can we do with Phylogenies?

Phylogenetic placement for identifying anonymous sequences



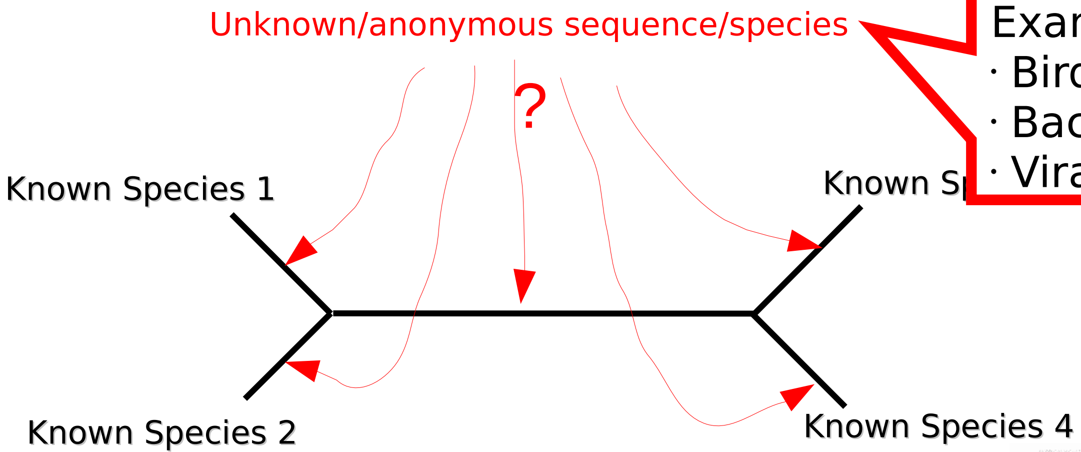
reference phylogeny

What can we do with Phylogenies?

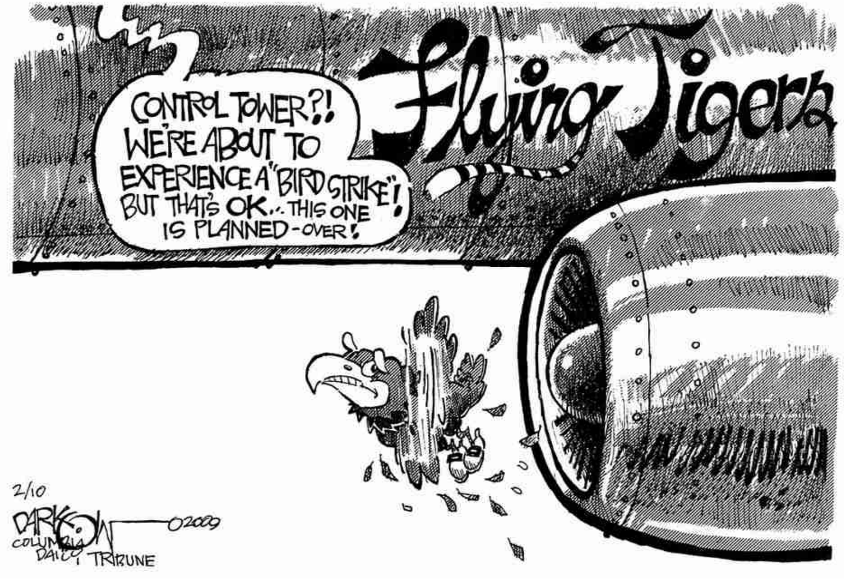
Phylogenetic placement for identifying anonymous sequences

Examples:

- Bird strike
- Bacteria
- Viral strains

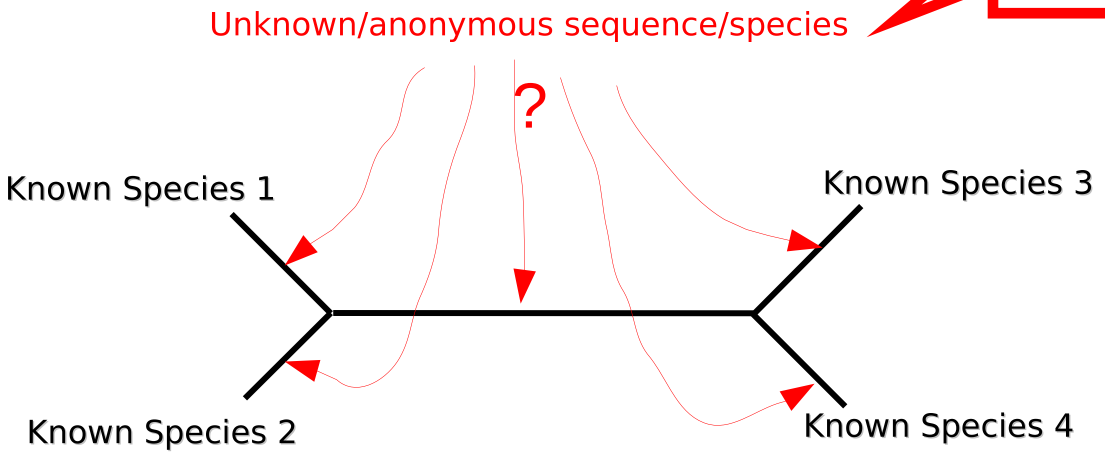


reference phylogeny


CONTROL TOWER?! WE'RE ABOUT TO EXPERIENCE A "BIRD STRIKE"! BUT THAT'S OK... THIS ONE IS PLANNED-OVER!
Flying Tigers
2/10
DARKOW ©2009
COLUMBIA DAILY TRIBUNE

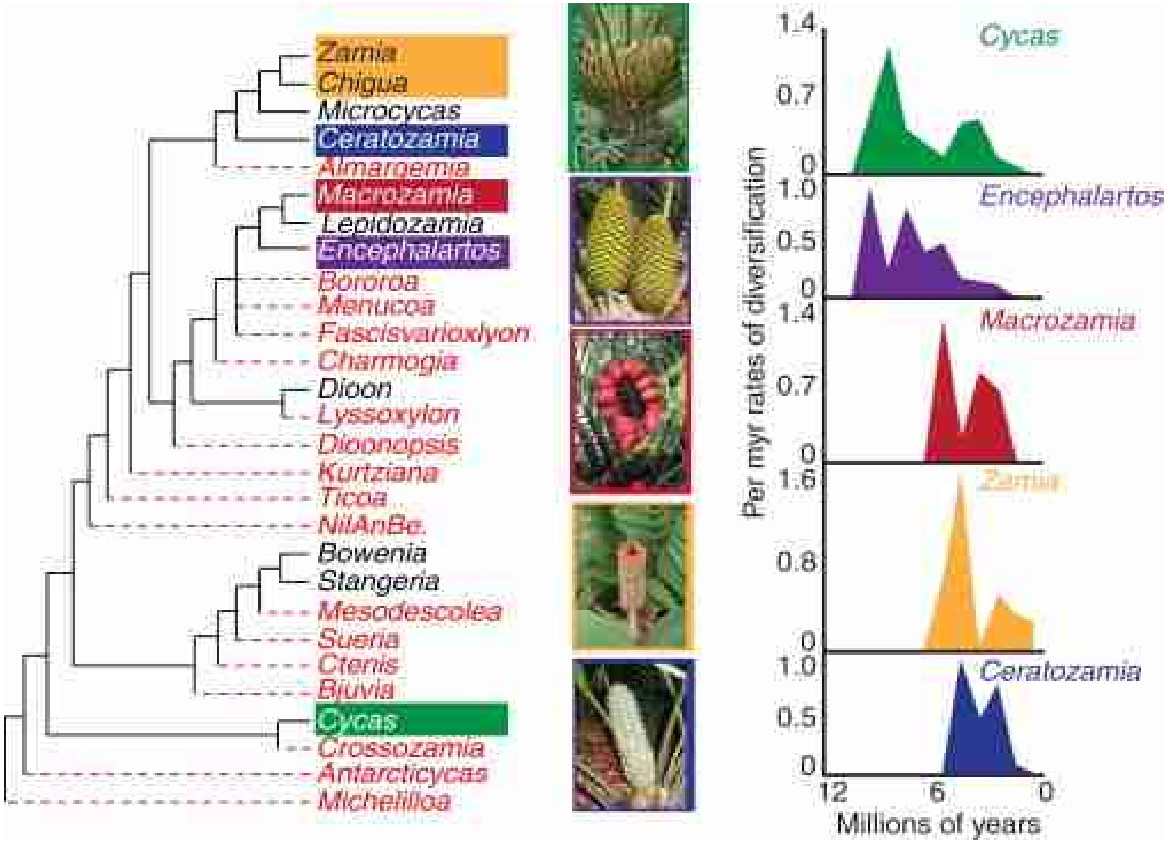
What can we do with Phylogenies?

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



reference phylogeny

Diversification Rates

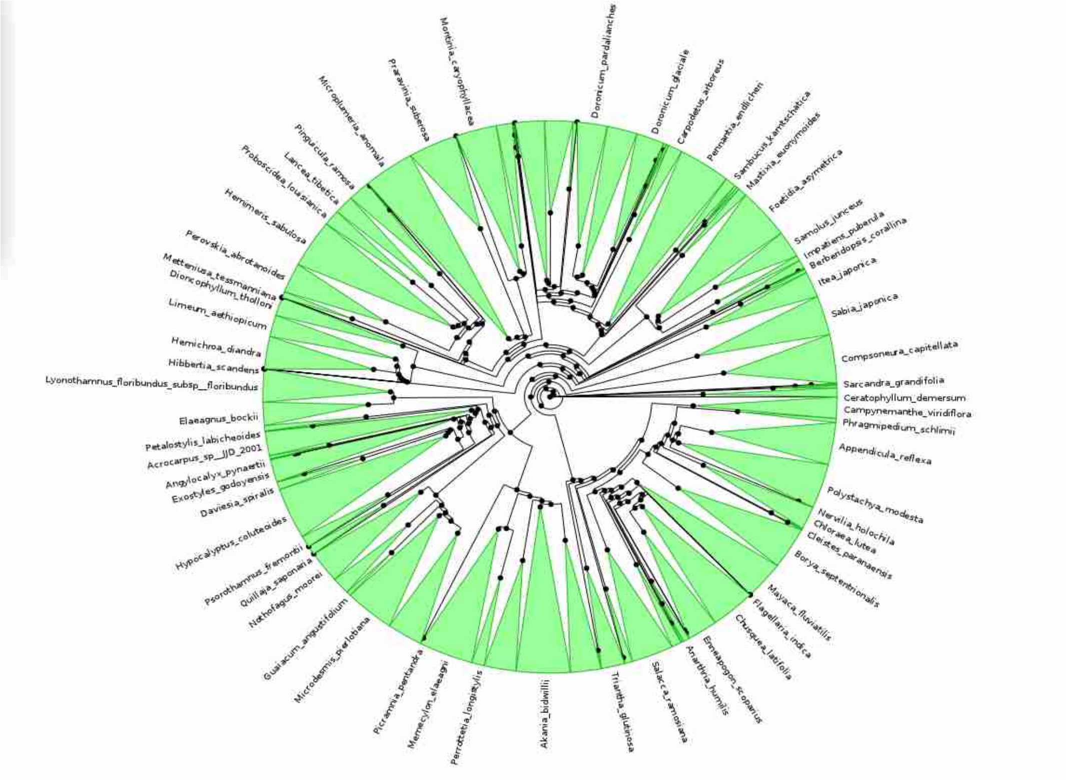


Time

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

Diversification Rates

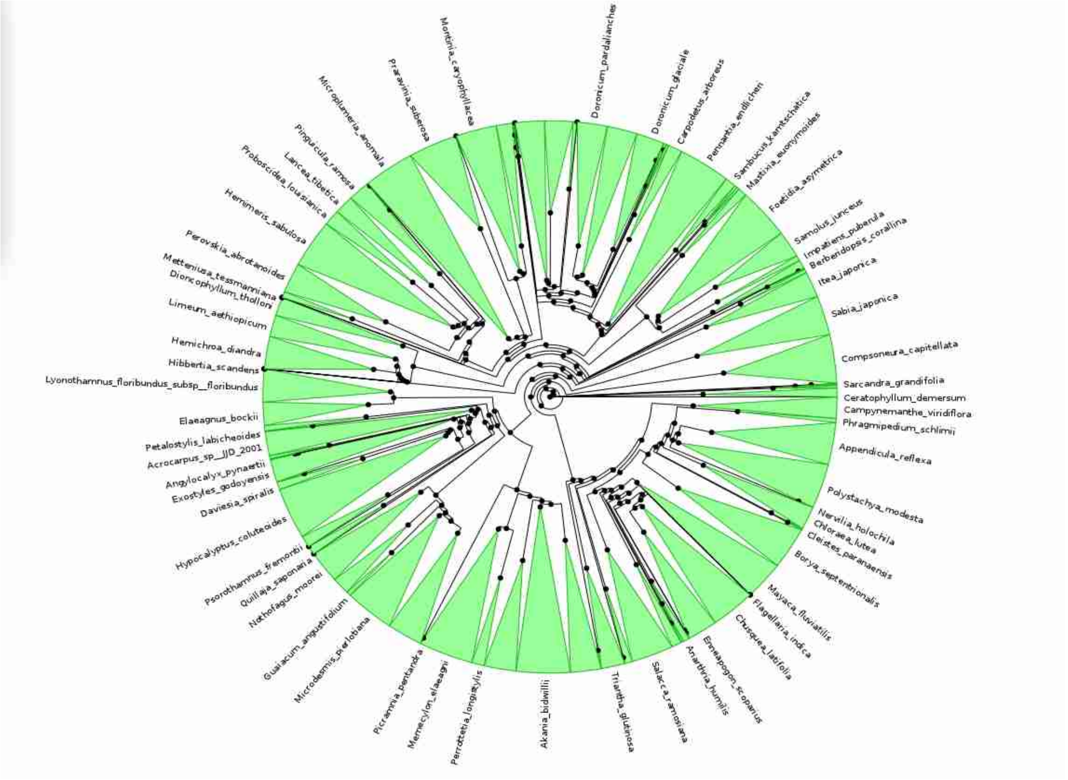
- With former PostDoc Stephen Smith: “Understanding angiosperm diversification using small and large phylogenetic trees”, *American Journal of Botany* 98 (3), 404-414, 2011.
- Largest tree of angiosperms computed to date
- 55,000 taxa



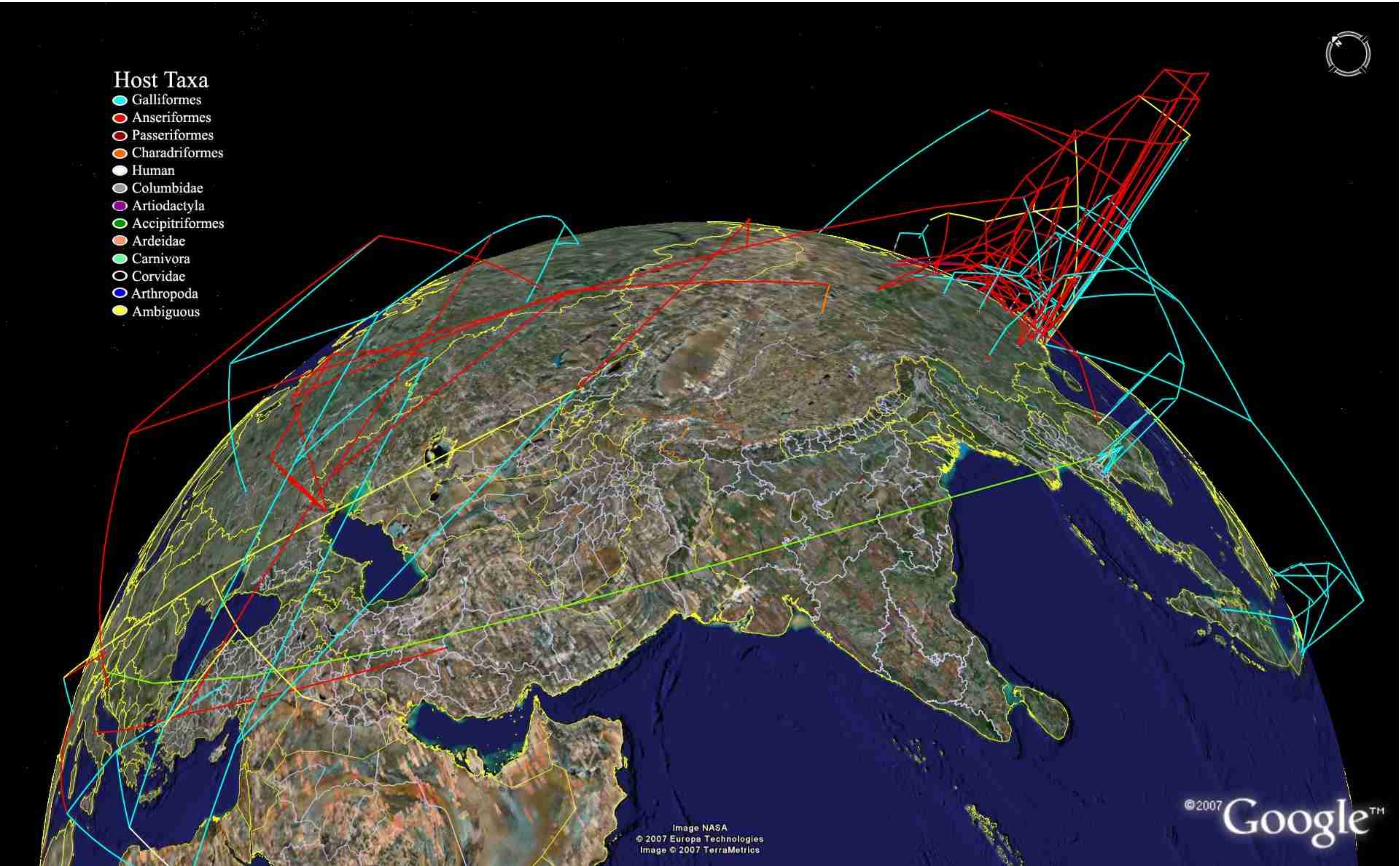
Diversification Rates

- 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

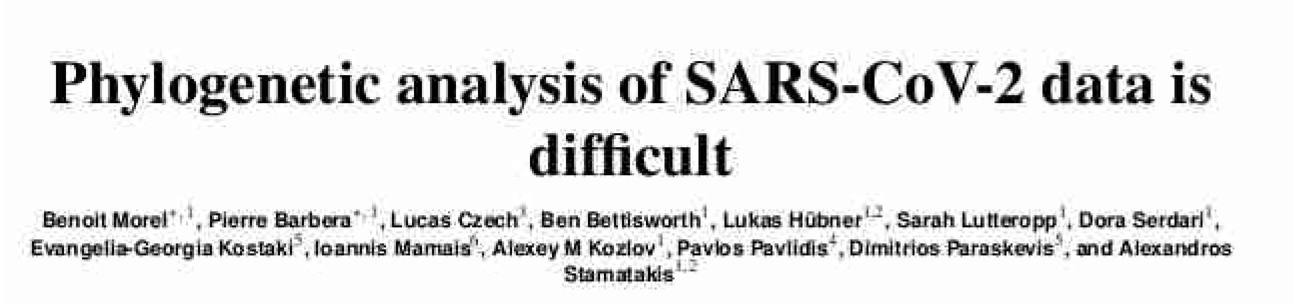
Visualizing big trees also represents a challenge → graph drawing & layout algorithms.



Influenza Outbreaks

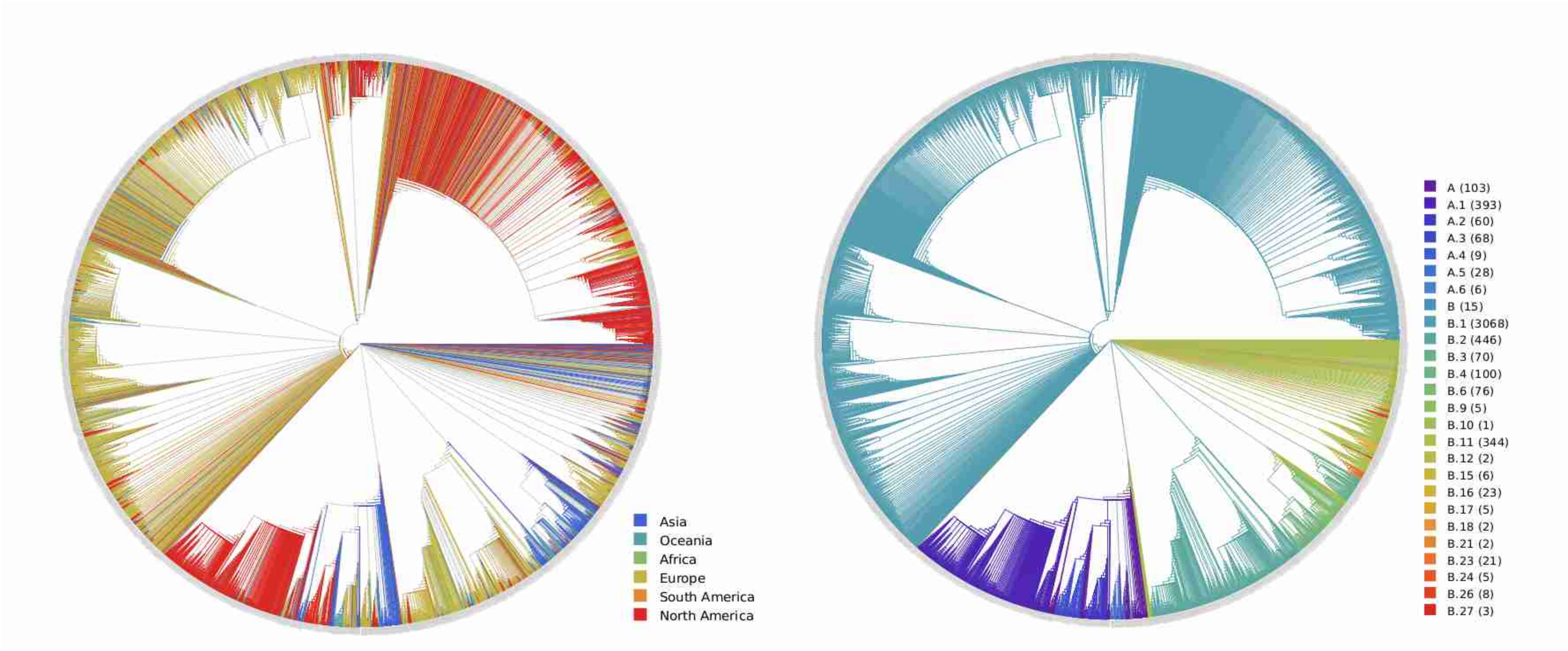


And of course SARS-CoV-2



Phylogenetic analysis of SARS-CoV-2 data is difficult

Benoit Morel*,1, Pierre Barbera*,1, Lucas Czech3, Ben Bettisworth1, Lukas Hübner1,2, Sarah Lutteropp1, Dora Serdari1, Evangelia-Georgia Kostaki5, Ioannis Mamais6, Alexey M Kozlov1, Pavlos Pavlidis4, Dimitrios Paraskevis5, and Alexandros Stamatakis1,2



Snakebites

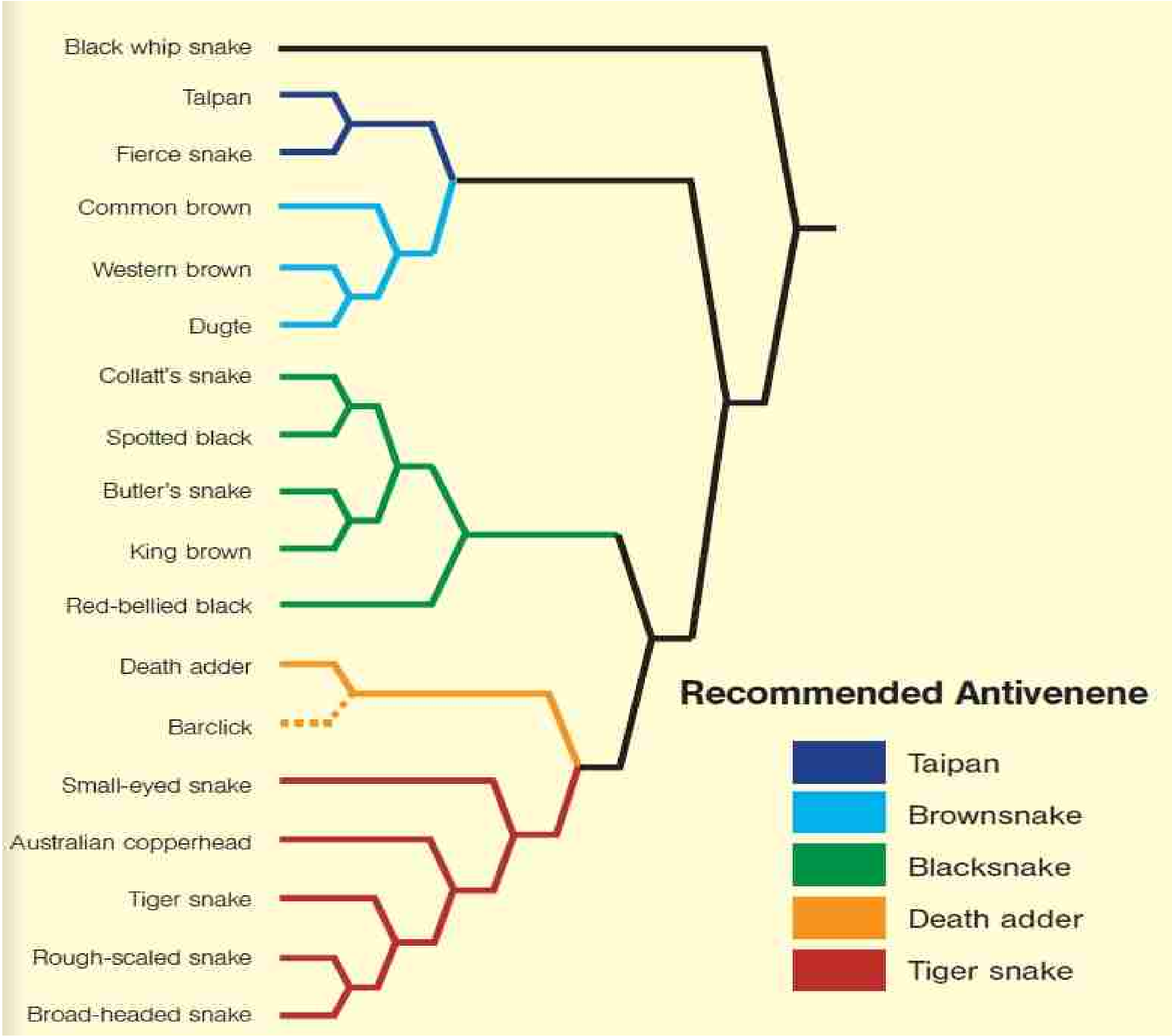
Australia has more poisonous snakes than any other continent, and many people die from snakebites each year. Developing effective antivenins is thus a high priority, but little is known about the venins of most species.

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 barclick 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

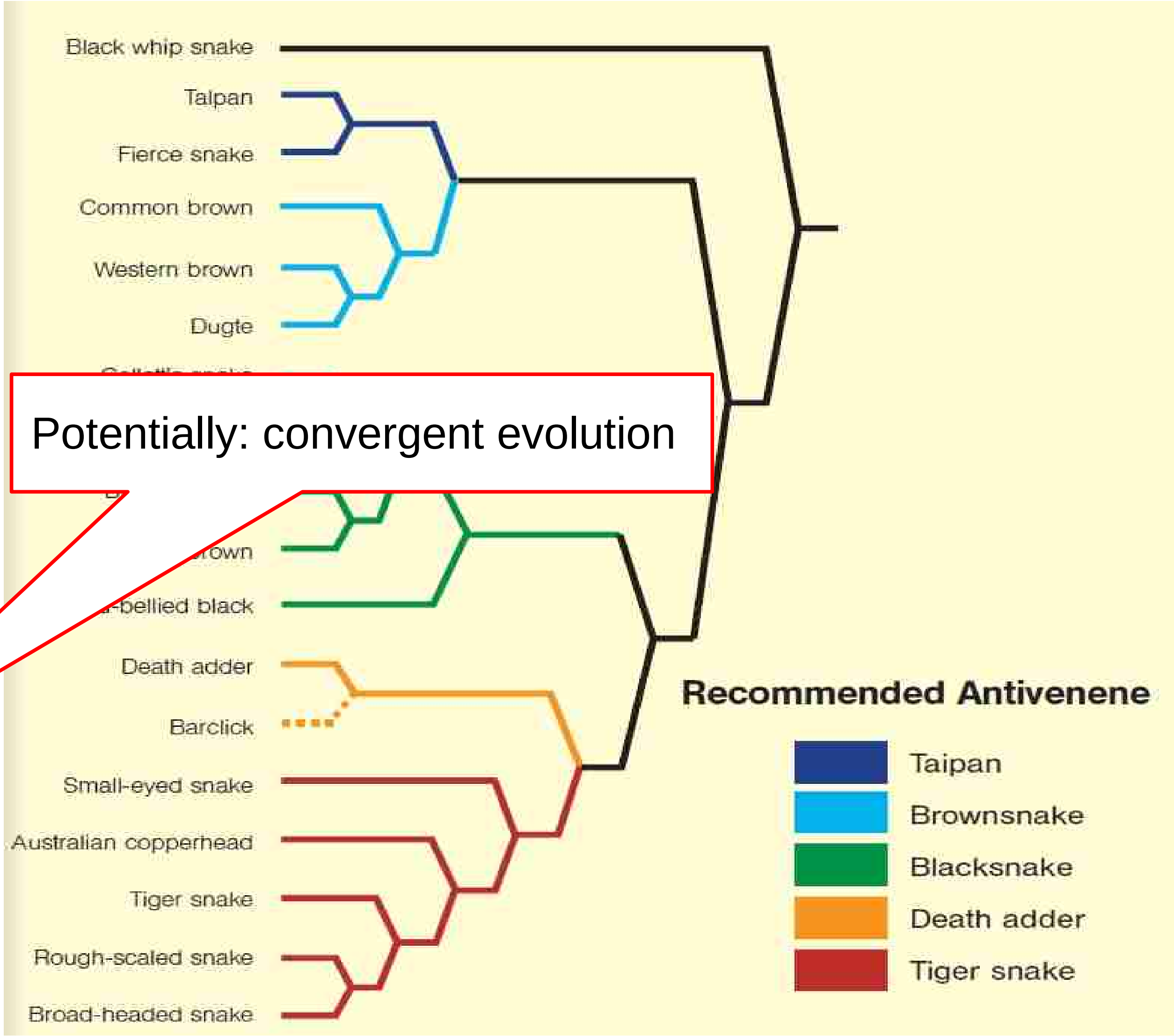
Australia has more poisonous snakes than any other continent, and many people die from snakebites each year. Developing effective antivenins is thus a high priority, but little is known about the venins of most species.

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 barclick 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.



Potentially: convergent evolution

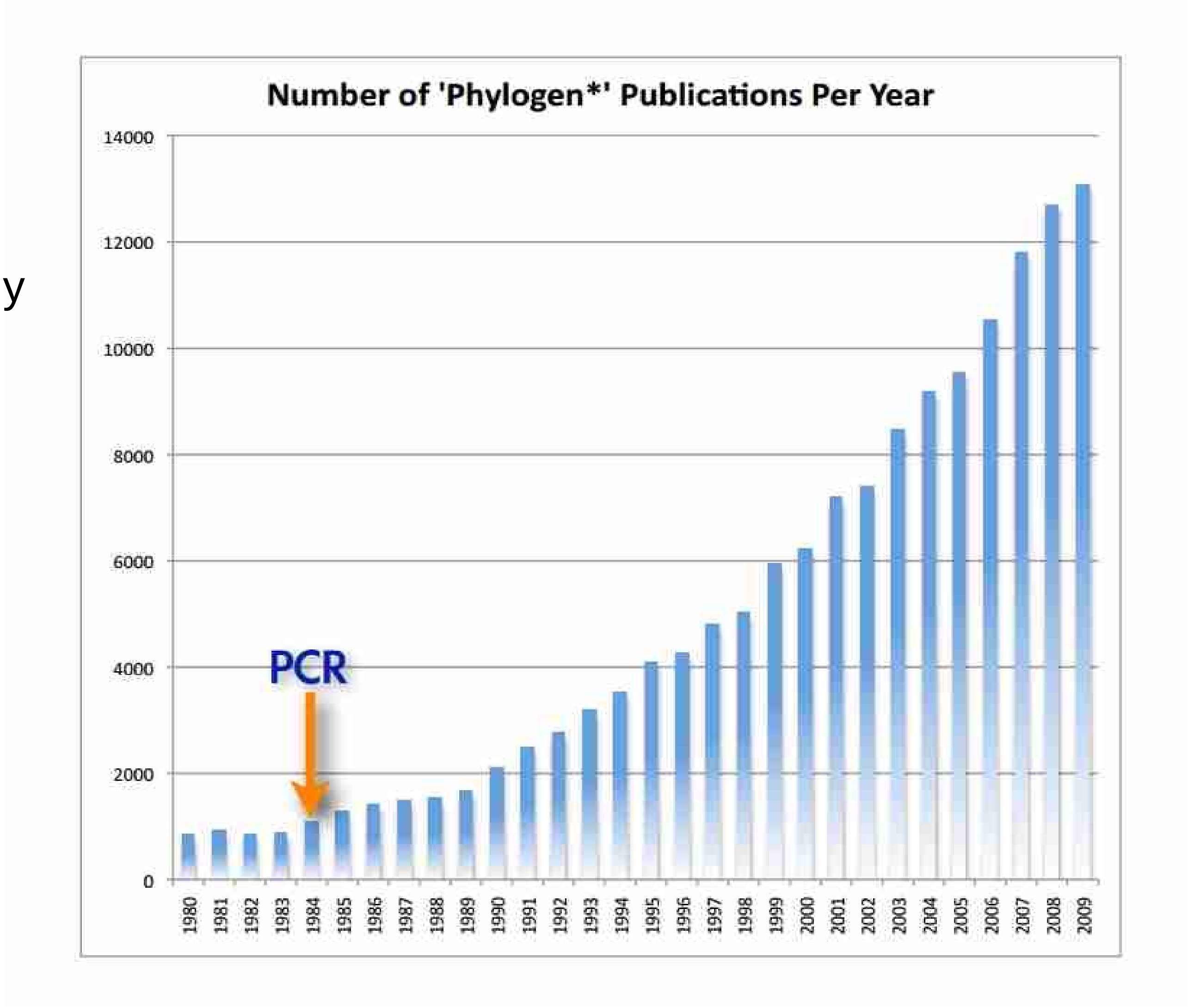
What can we do with phylogenetic trees?

- 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.

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

Why this increase in Phylogenetics papers?
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 & tree
      - → 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 & tree
      - → 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 Mean:* UPGMA
    - → least-squares method: explicit optimality
  - *Character-based methods*
    - → optimality criteria *f()* operate directly
      - → 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

What could be the computational limitation here?

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 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.

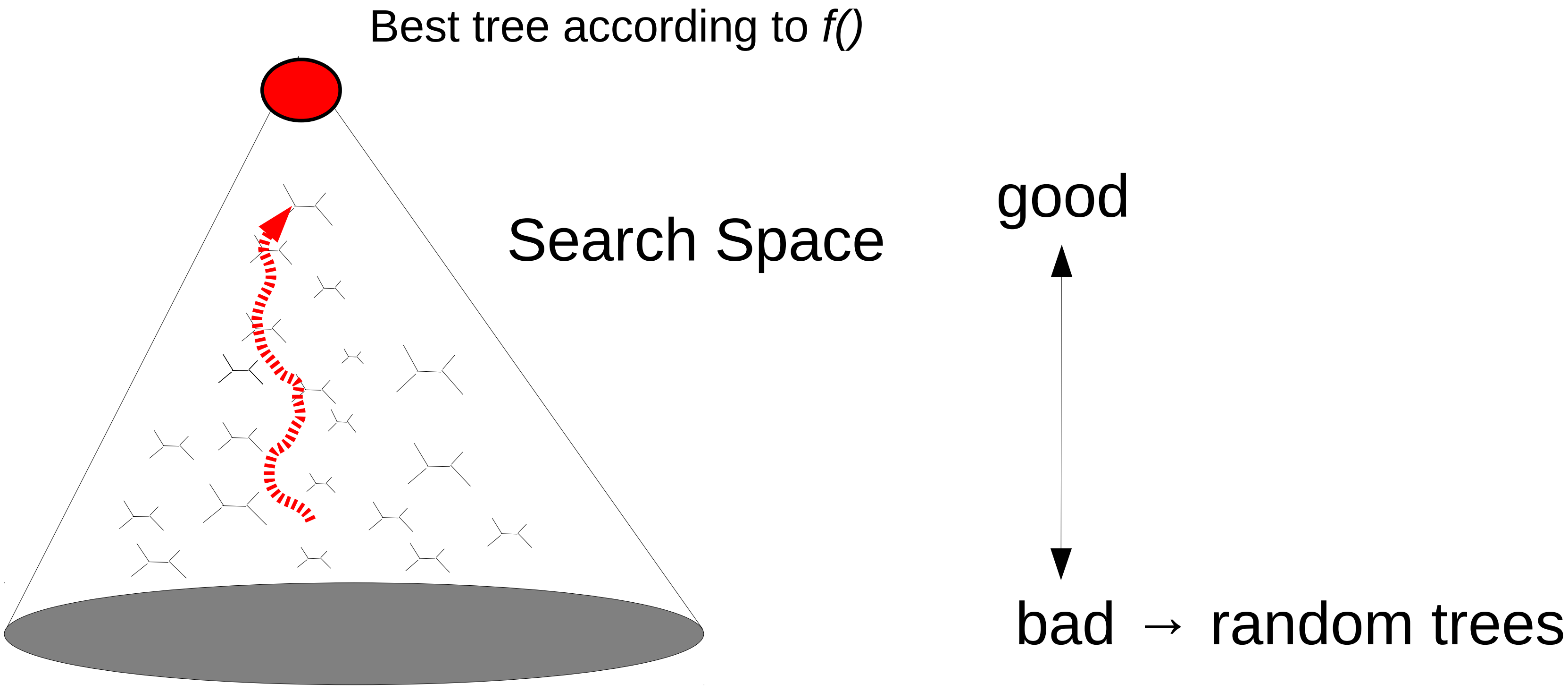
- We do the data transfer RAM ↔ disk explicitly from within the application code by using application-specific knowledge (e.g., about the data access patterns)
- This is to circumvent the paging procedure that would normally be initiated by the OS
- Out-of-core algorithms are typically much faster than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



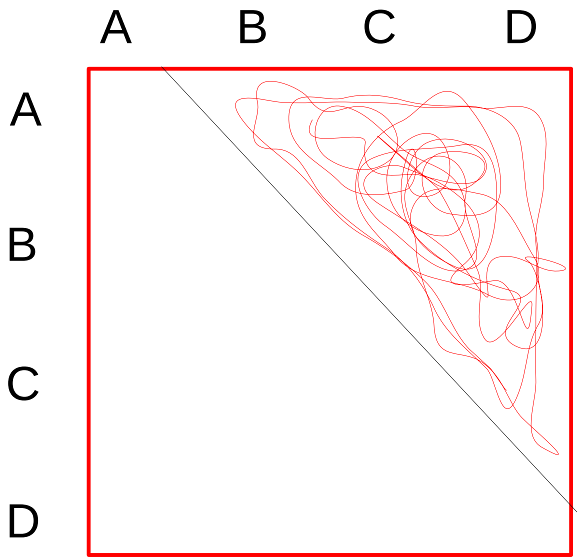
Best tree according to *f()*

Search Space

good

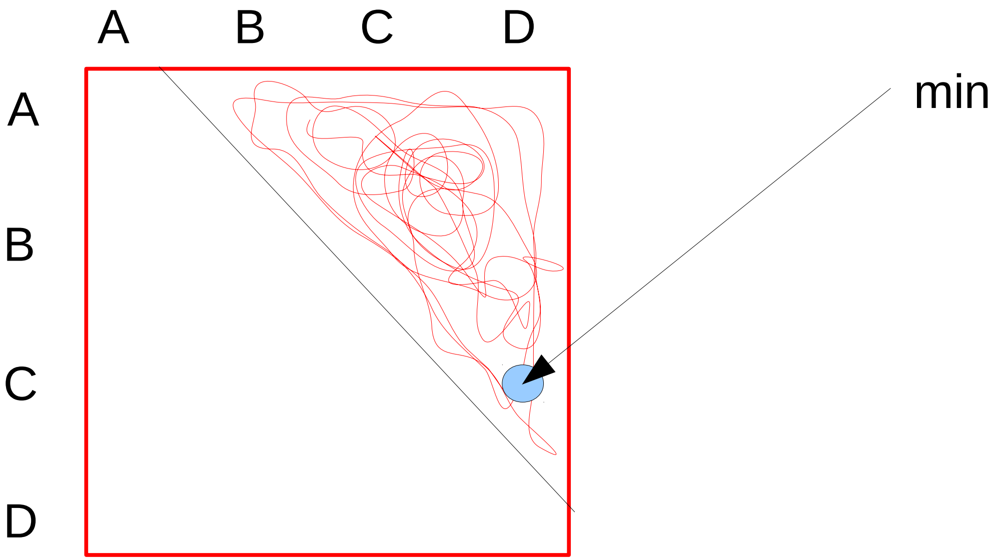
bad → random trees

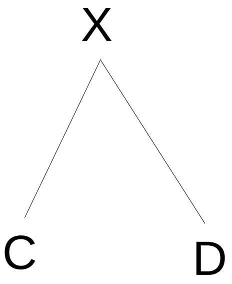
Neighbor Joining → Principle



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

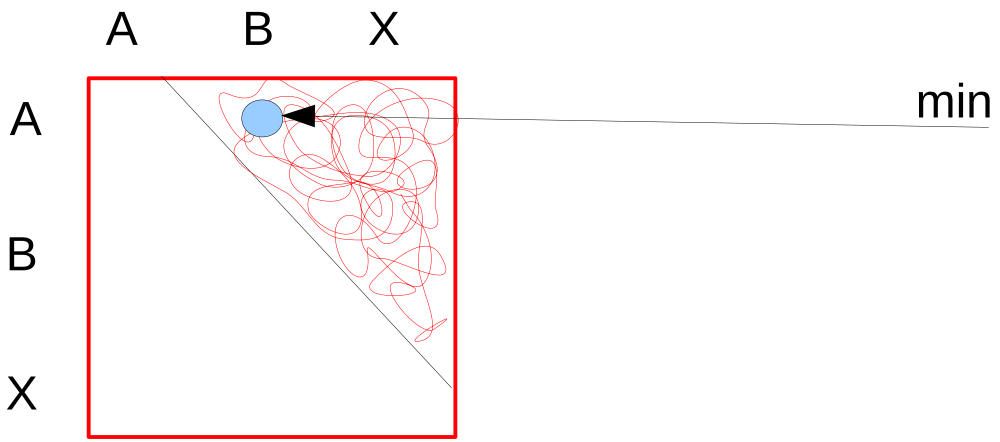
Neighbor Joining → Principle

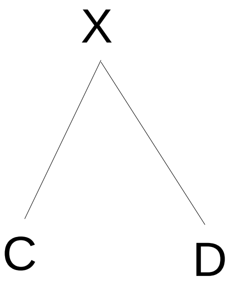




Given a kind of distance matrix $D_{i,j}$ where $i,j=1...4$
Find minimum and merge taxa

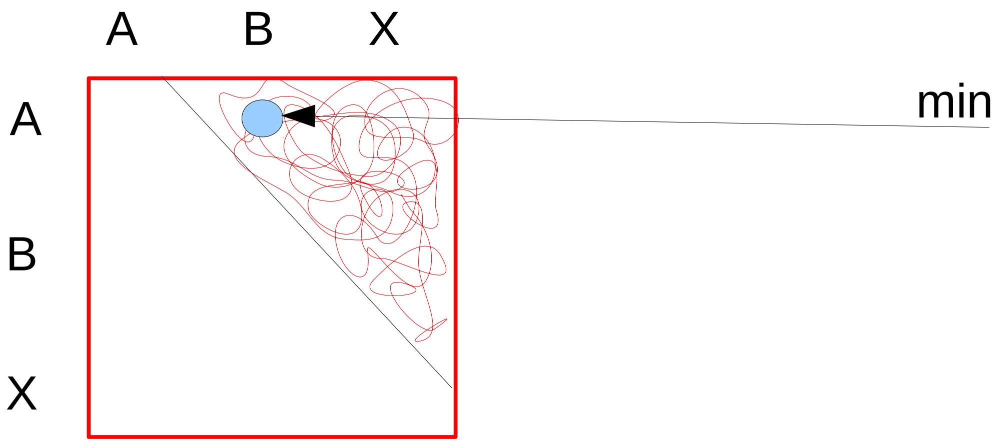
Neighbor Joining → Principle

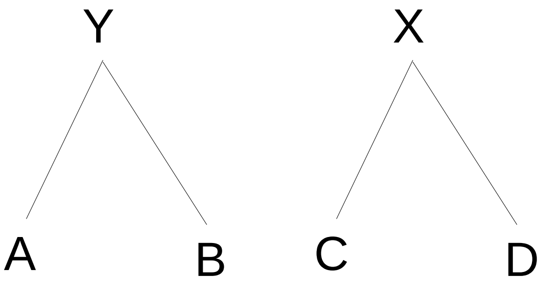




Given a kind of distance matrix $D_{i,j}$ where $i,j=1...4$
Find minimum and merge taxa
Compute a new distance matrix of size $n-1 = 3$
Find minimum

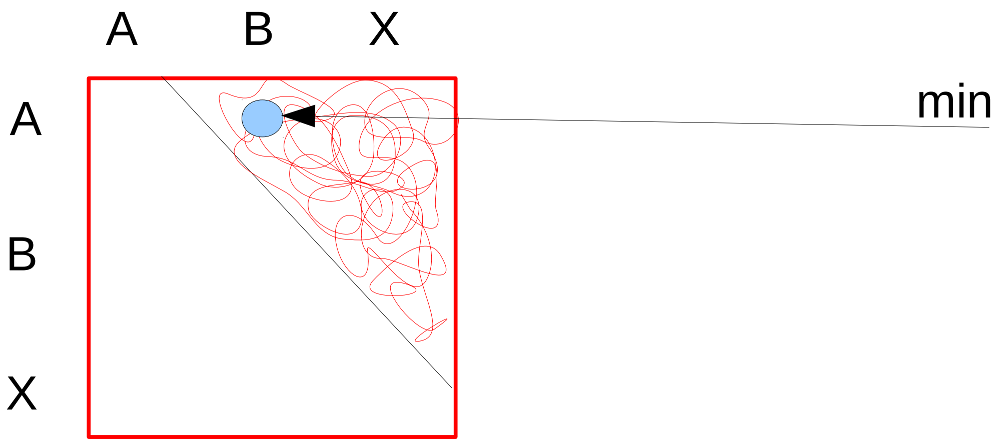
Neighbor Joining → Principle

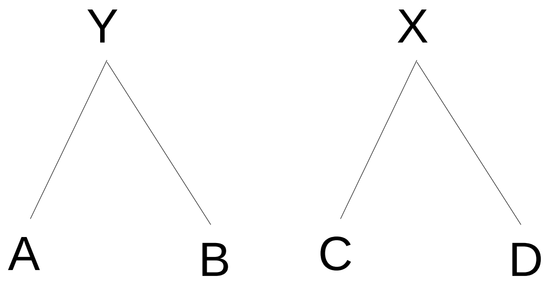




Given a kind of distance matrix $D_{i,j}$ where $i,j=1...4$
Find minimum and merge taxa
Compute a new distance matrix of size $n\text{-}1 = 3$
Find minimum and merge taxa

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,* see below why

- 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				-

$D_{ij}-u_i-u_j$

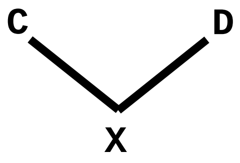
Neighbor Joining Algorithm

	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				-

$D_{ij}-u_i-u_j$



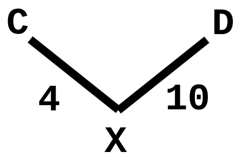
Neighbor Joining Algorithm

	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				-

$D_{ij}-u_i-u_j$

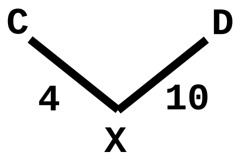


$b_C = 0.5 \times 14 + 0.5 \times (23.5-29.5) = 4$

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

Neighbor Joining Algorithm

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

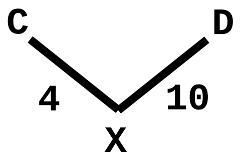


Neighbor Joining Algorithm

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

$$\begin{aligned} D_{XA} &= (D_{CA} + D_{DA} - D_{CD})/2 \\ &= (21 + 27 - 14)/2 \\ &= 17 \end{aligned}$$

$$\begin{aligned} D_{XB} &= (D_{CB} + D_{DB} - D_{CD})/2 \\ &= (12 + 18 - 14)/2 \\ &= 8 \end{aligned}$$

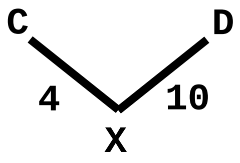


Neighbor Joining Algorithm

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

$$\begin{aligned} D_{XA} &= (D_{CA} + D_{DA} - D_{CD})/2 \\ &= (21 + 27 - 14)/2 \\ &= 17 \end{aligned}$$

$$\begin{aligned} D_{XB} &= (D_{CB} + D_{DB} - D_{CD})/2 \\ &= (12 + 18 - 14)/2 \\ &= 8 \end{aligned}$$

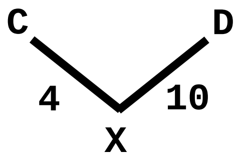


Neighbor Joining Algorithm

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

$$\begin{aligned} D_{XA} &= (D_{CA} + D_{DA} - D_{CD})/2 \\ &= (21 + 27 - 14)/2 \\ &= 17 \end{aligned}$$

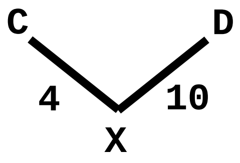
$$\begin{aligned} D_{XB} &= (D_{CB} + D_{DB} - D_{CD})/2 \\ &= (12 + 18 - 14)/2 \\ &= 8 \end{aligned}$$



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



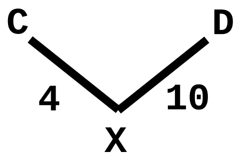
Neighbor Joining Algorithm

	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			-

$D_{ij}-u_i-u_j$

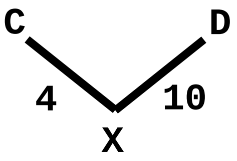


Neighbor Joining Algorithm

	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			-



$D_{ij}-u_i-u_j$

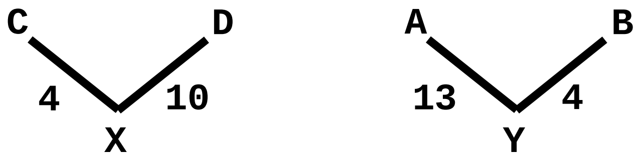
Neighbor Joining Algorithm

	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			-

$D_{ij}-u_i-u_j$

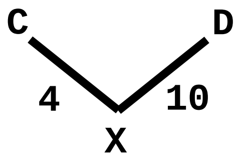


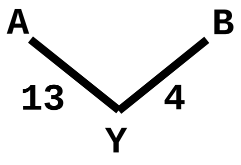
b_A = 0.5 x 17 + 0.5 x (34-25) = 13

b_D = 0.5 x 17 + 0.5 x (25-34) = 4

Neighbor Joining Algorithm

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

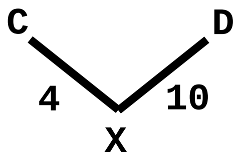


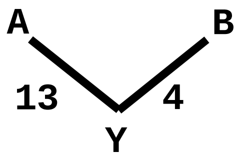


Neighbor Joining Algorithm

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

$$
\begin{aligned}
D_{YX} &= (D_{AX} + D_{BX} - D_{AB})/2 \\
&= (17 + 8 - 17)/2 \\
&= 4
\end{aligned}
$$

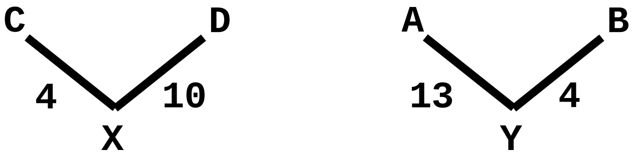




Neighbor Joining Algorithm

	X	Y
X	-	4
Y		-

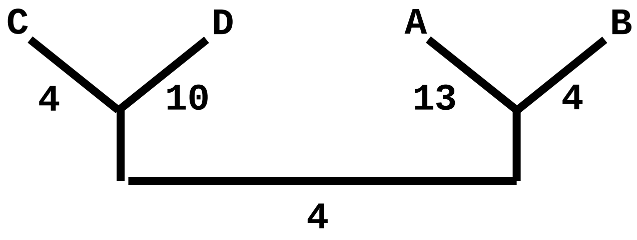
$$
\begin{aligned}
D_{YX} &= (D_{AX} + D_{BX} - D_{AB})/2 \\
&= (17 + 8 - 17)/2 \\
&= 4
\end{aligned}
$$



Neighbor Joining Algorithm

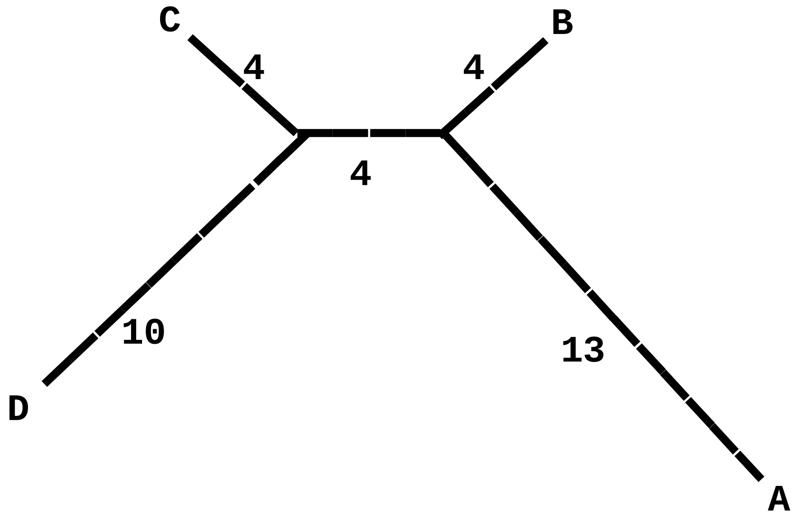
	X	Y
X	-	4
Y		-

$$\begin{aligned} D_{YX} &= (D_{AX} + D_{BX} - D_{AB})/2 \\ &= (17 + 8 - 17)/2 \\ &= 4 \end{aligned}$$



Neighbor Joining Algorithm

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


C
B
4
4
4
10
13
D
A