Lab meeting 10.10.16

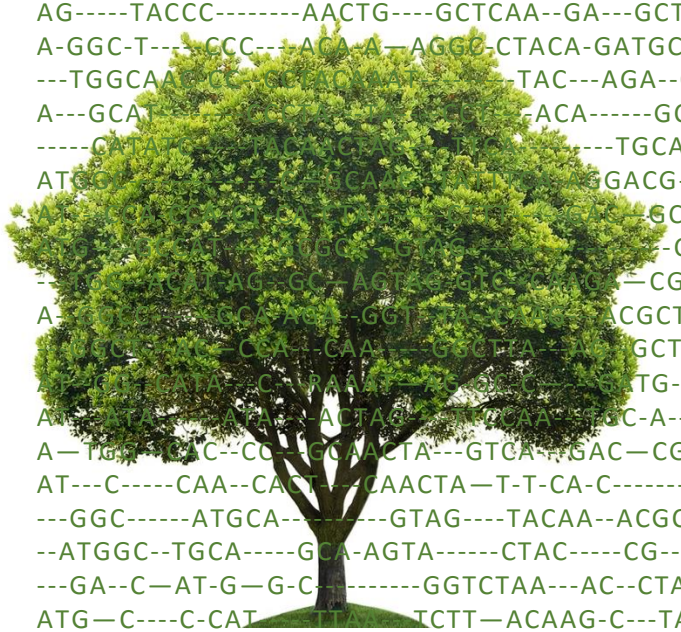
What's the problem with big trees?

Highly diverged taxa + rapidly evolving loci ->

poor alignments ->

inaccurate trees.

Species1: ATGGCATATCC[illegible]GGATTCCAAGATGCAACC
Species2: ATGGCACA[illegible]GGTTTCAAGGACGCGGC
Species3: ATGGCCAA[illegible]GGCTTTCAAGACGCCTCC
Species4: ATGGCACA[illegible]GGTCTACAAGACGCTAC
Species5: ATGGCACA[illegible]GGTCTACAAGACGCTAC
Species6: ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTAC
Species7: ATGGCTTACCCA[illegible]GCTTGGCTTACAAGACGCTACC
Species8: ATGGCATACCCCCTA[illegible]AATAGGCCTACAAGATGCAAC



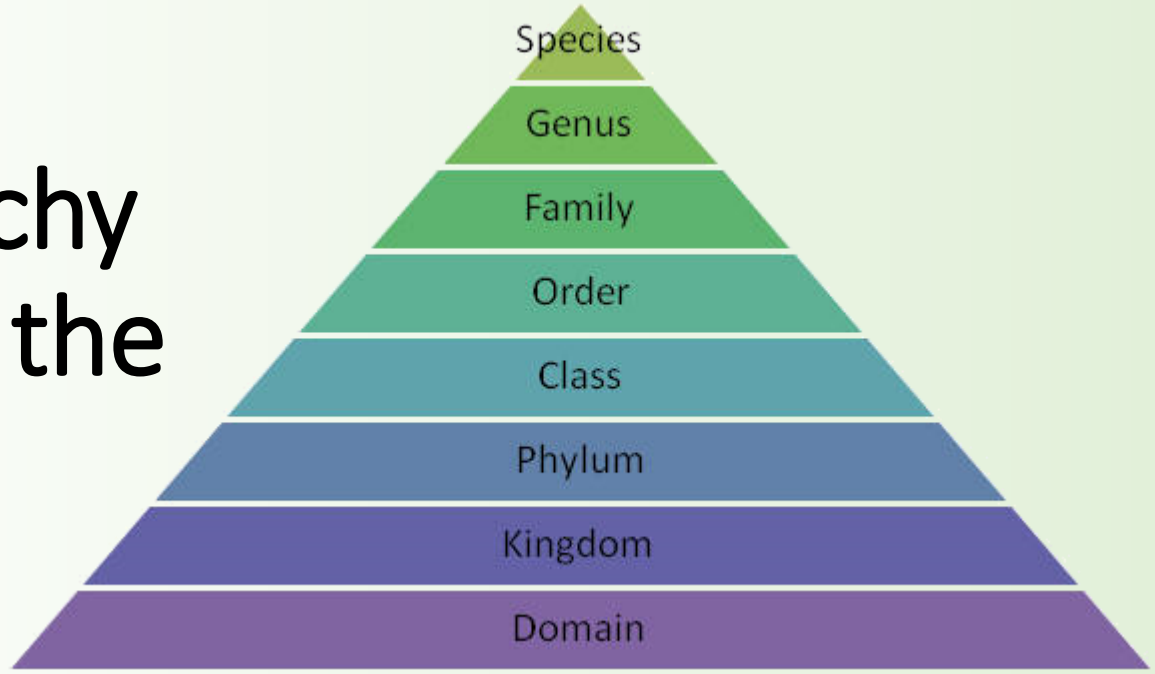
Species1: AT-----CATA-CCATA-----A—GATTC-AG-AT--GC-AA-CC
Species2: ATGG-CAC—CCCAAC------CTAGGTT----GGA-CGC-GGC-
Species3: ATGG-----AA—CCATC—CAA---AGGC--------------G------
Species4: --ATG—--AT—GC--GCG—AGTA—GG-----A—G----CTAC
Species5: ATGGC-------A----GC00-----TAG--TCTACAAGAC-----TAC
Species6: AT—G—CACATG-CA----CA-AAG-------TCCA--GACGTAC-
Species7: AG-----TACCC--------AACTG----GCTCAA--GA---GCTACC
Species8: A-GGC-T----CCC---ACA-A—AGGC-CTACA-GATGCAAC-
Species9: ---TGGCAA[illegible]TAC---AGA--CAAC
Species10: A---GCA[illegible]ACA------GC---C
Species11: -----CA[illegible]TGCA--CC
Species12: AT[illegible]GACG---GC
Species13: A[illegible]GCC-CC
Species14: [illegible]-CTAC
Species15: [illegible]—CGC--C
Species16: A[illegible]CGCTAC--
Species17: [illegible]GCTA-CC
Species18: [illegible]TG--AC-
Species19: [illegible]C-A----C--
Species20: A—TG[illegible]GCAACTA---GTCA--GAC—CG-GC-
Species21: AT---C-----CAA--CAC[illegible]CAACTA—T-T-CA-C-------TCC-
Species22: ---GGC------ATGCA[illegible]GTAG----TACAA--ACGCTAC
Species23: --ATGGC--TGCA-----GCA-AGTA------CTAC-----CG--TAC-
Species24: ---GA--C—AT-G—G-C[illegible]GGTCTAA---AC--CTA—C-
Species25: ATG—C----C-CAT[illegible]TCTT—ACAAG-C---TAC-C
Species26: --TGG--CA-----CCC------A-AT—AGG-CCTACA-----TGCAC
Species27: -------CATA--TCCC---------------TTCCA----ATG-—AACC-
Species28: A--TGGC-ACAC---AA--GC---CTGGT---C-AGGAC—G-GGC
Species29: ---G---CC--AC----CTCCC—AAC-A----CTTC-----ACGC-TCC

First approach:

Use the taxonomic hierarchy to estimate the quality of the alignment.

First approach:

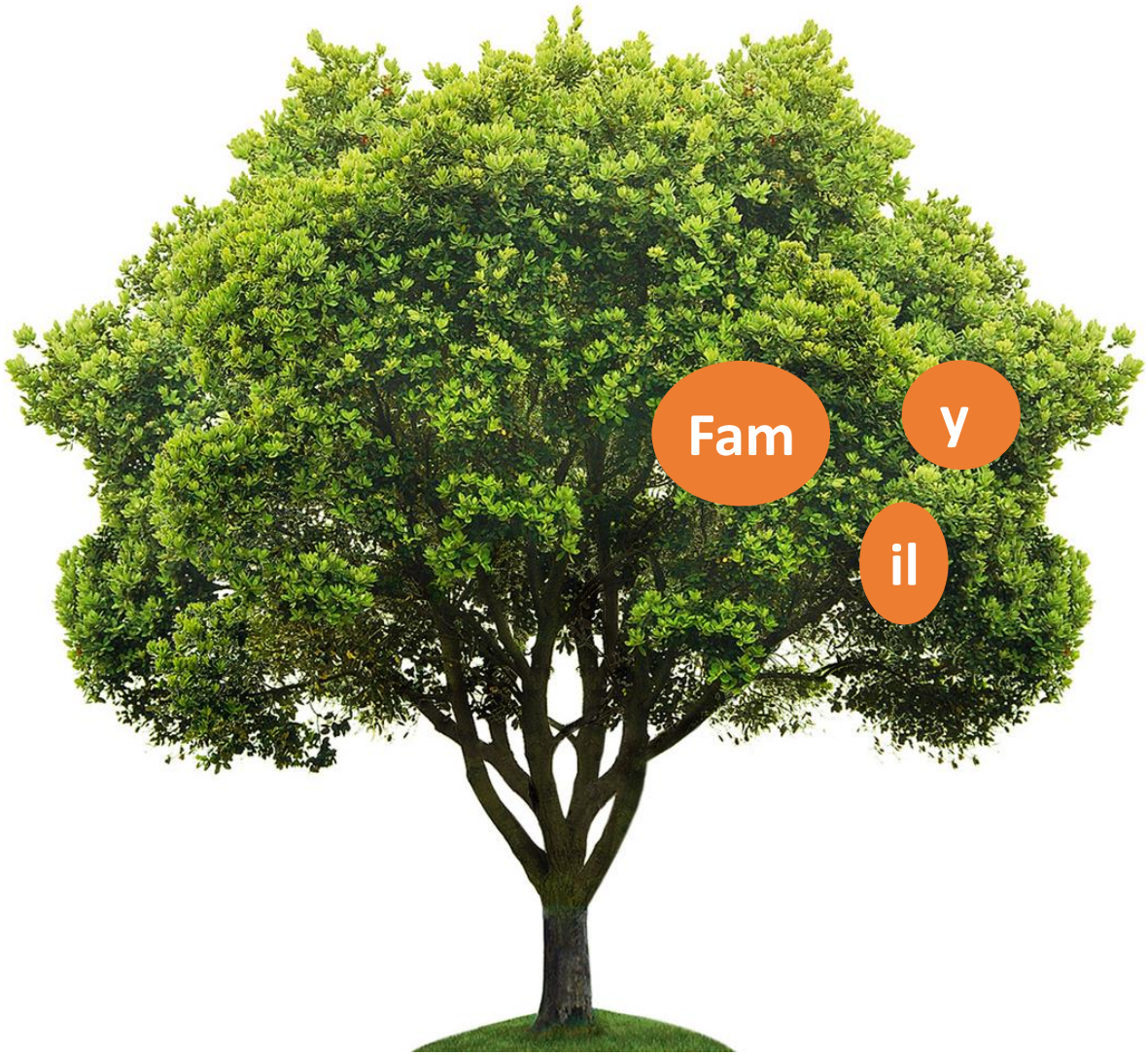
Use the taxonomic hierarchy to estimate the quality of the alignment.



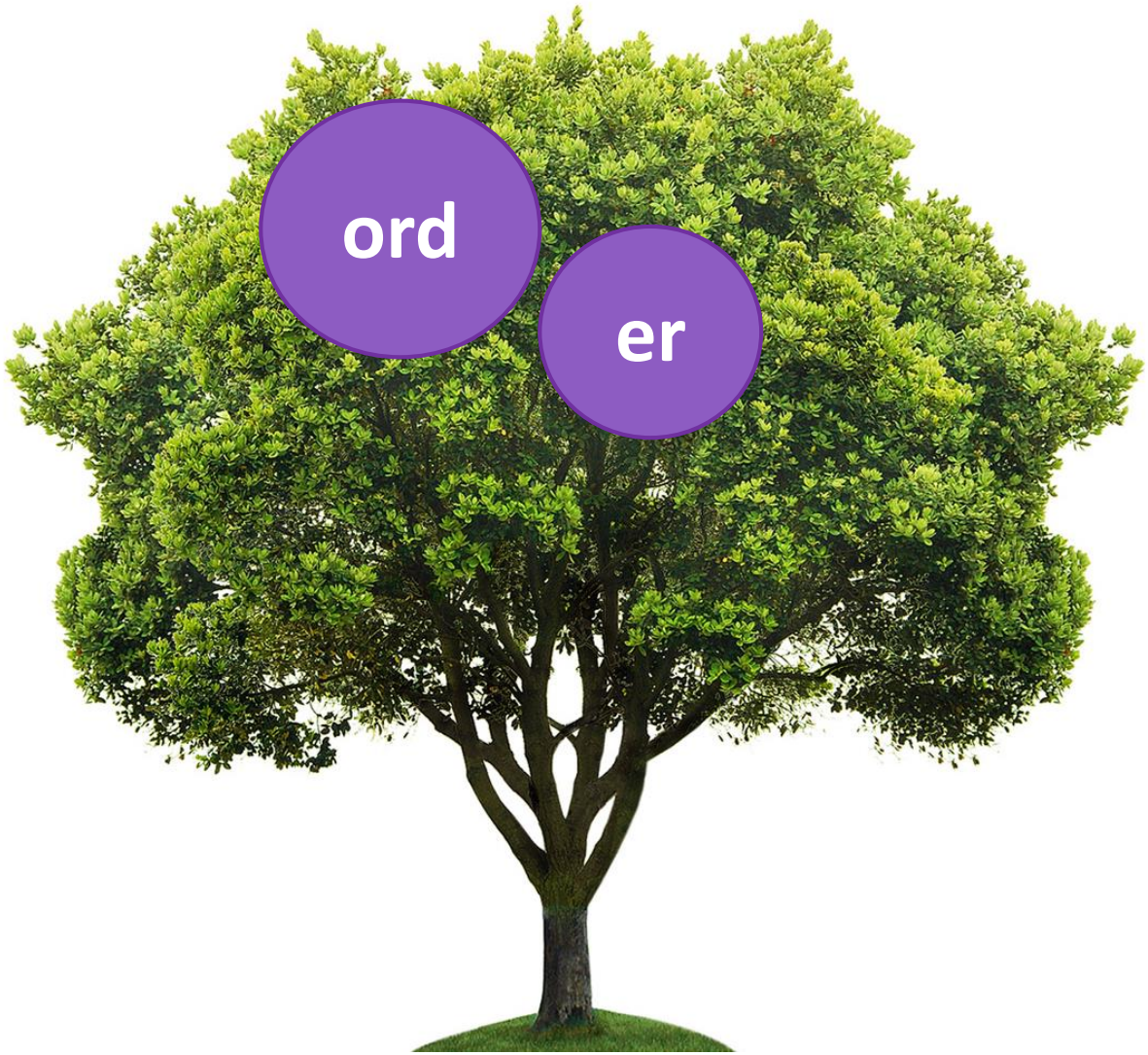
Family


Family


Family

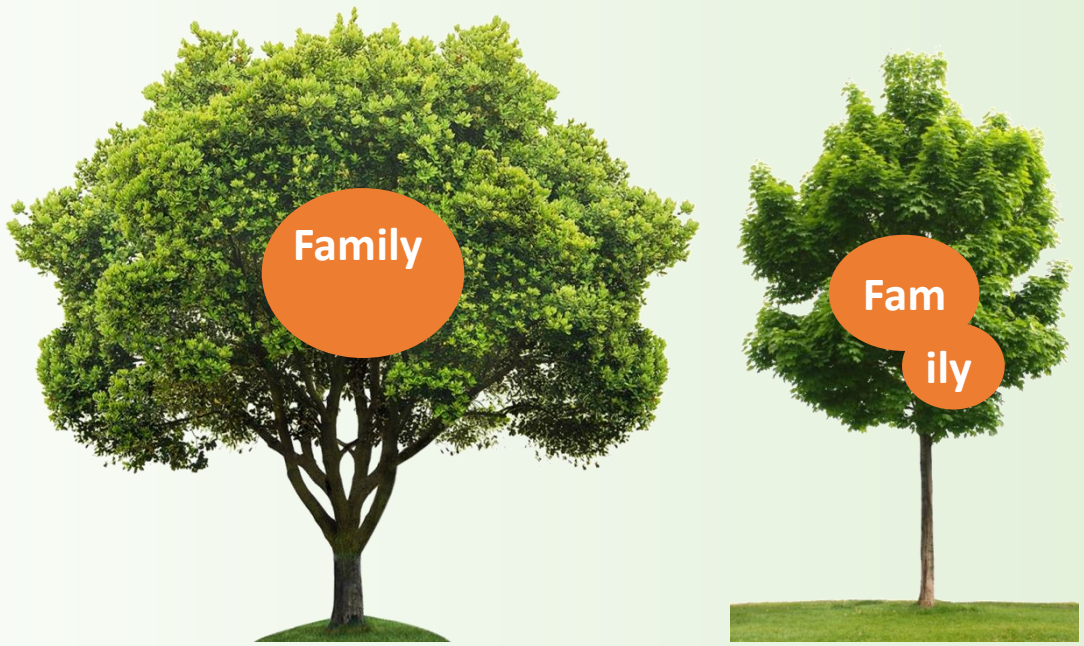

Fam
y
il


Family

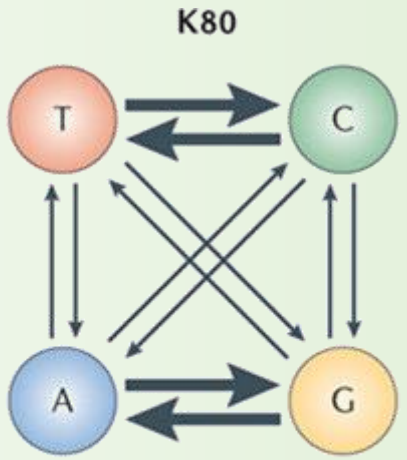

ord
er


order

Results were a little unexpected
Go to plan B…



Second approach:
Simulations.



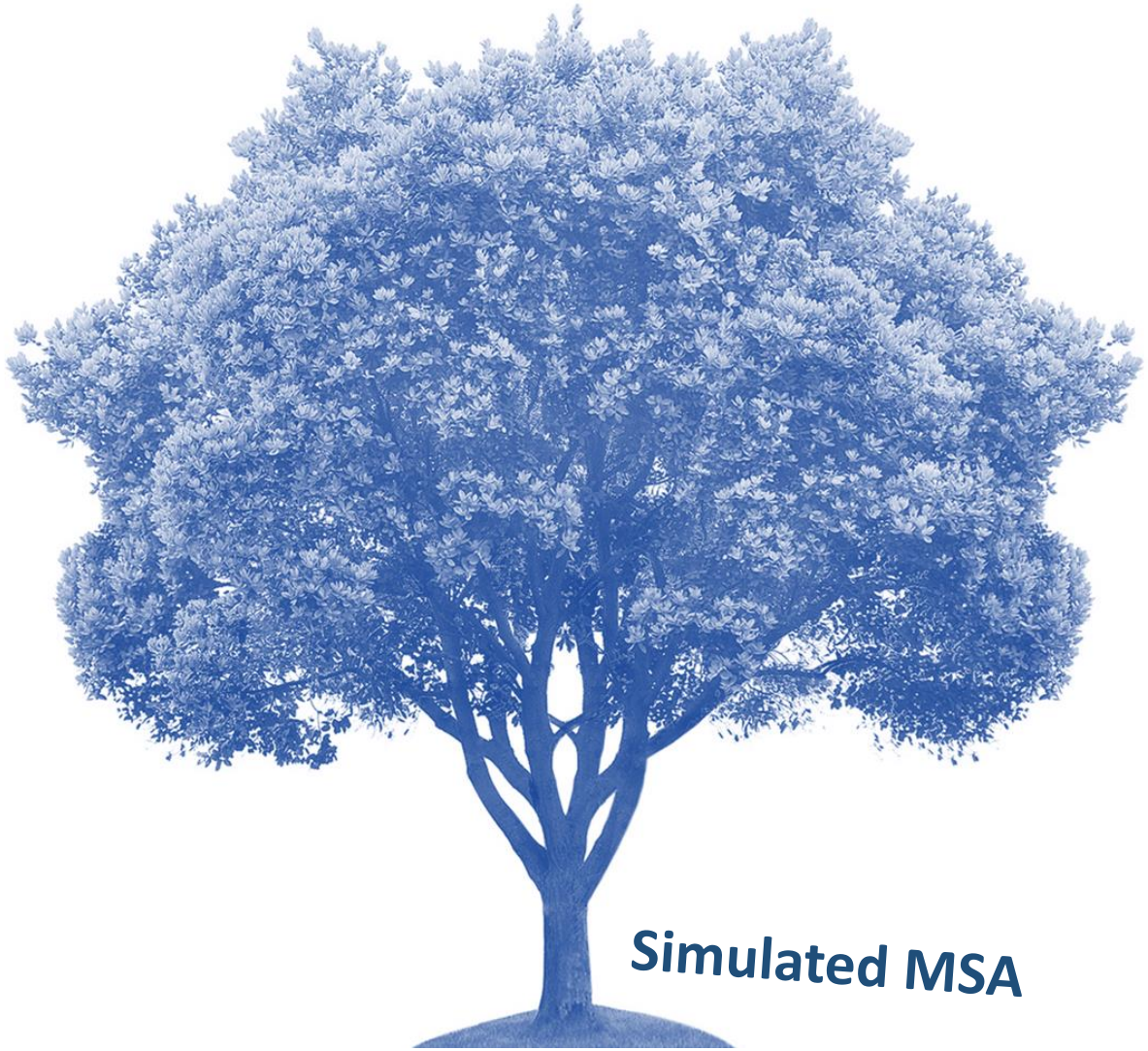
Second approach:
Simulations.

Characters of alignments:

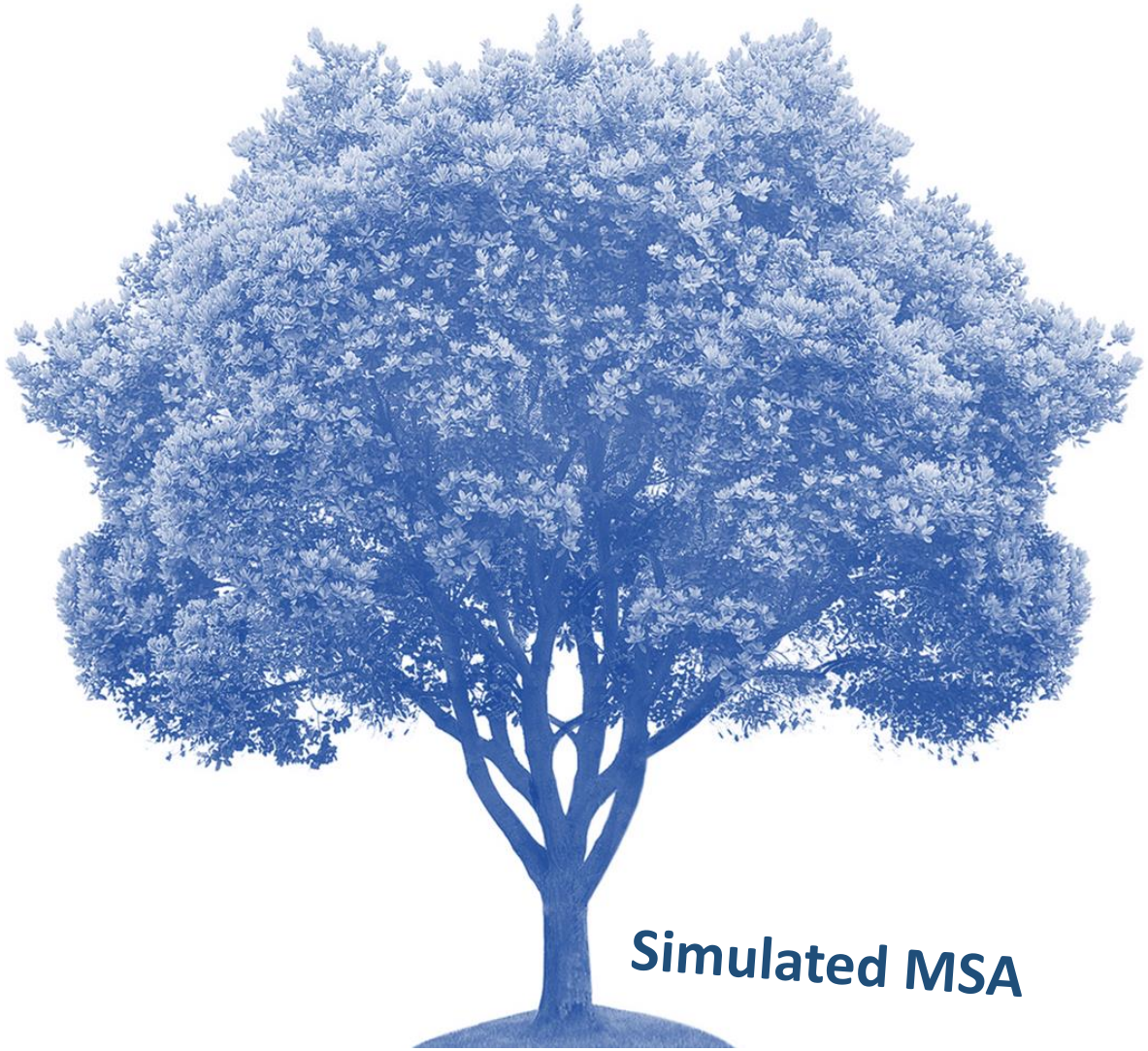
- nucleotides frequencies
- substitutions model
- insertion-deletion parameters (rate, length) - SpartaABC!


TRUE


TRUE


Simulated MSA


TRUE

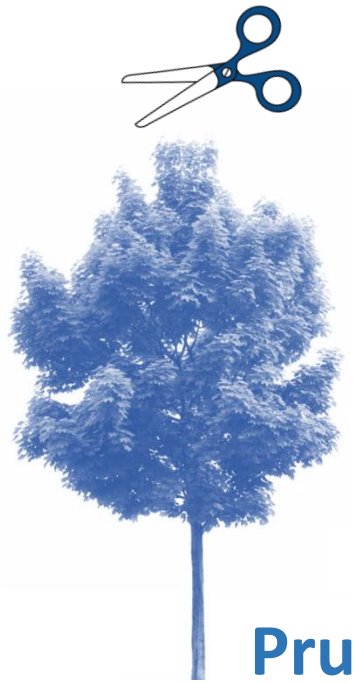

Simulated MSA


TRUE


Simulated MSA


TRUE

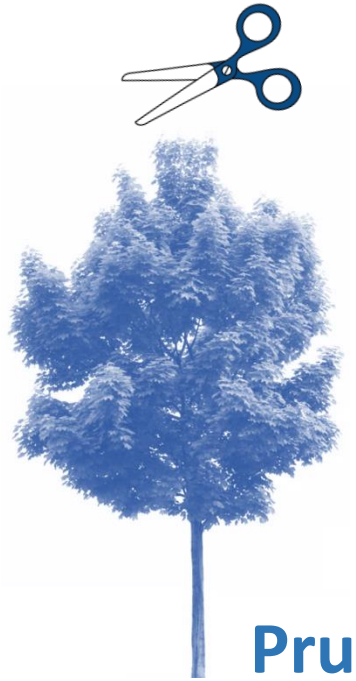

Simulated MSA


Pruned


TRUE

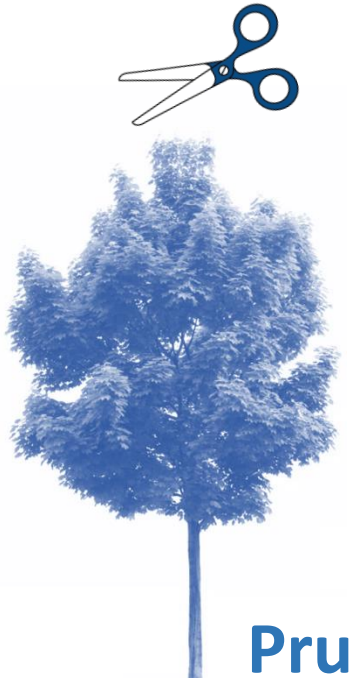

Simulated MSA


ACGGTCTGTGCG
ACGGTATGTGCG
Re-built


Pruned

Pruned

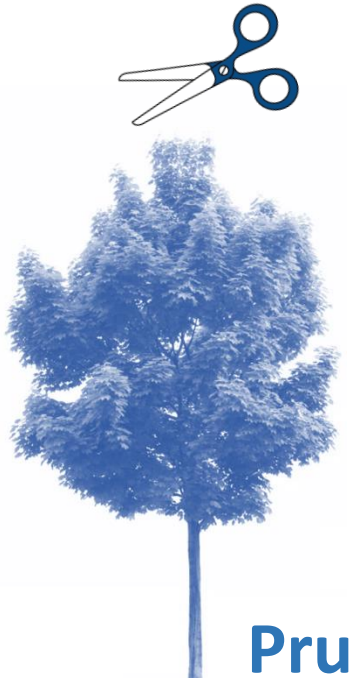
Re-built



Pruned

Pruned

Re-built



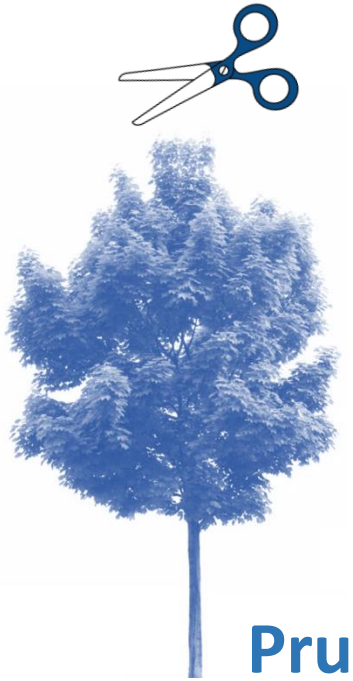
Pruned

Pruned

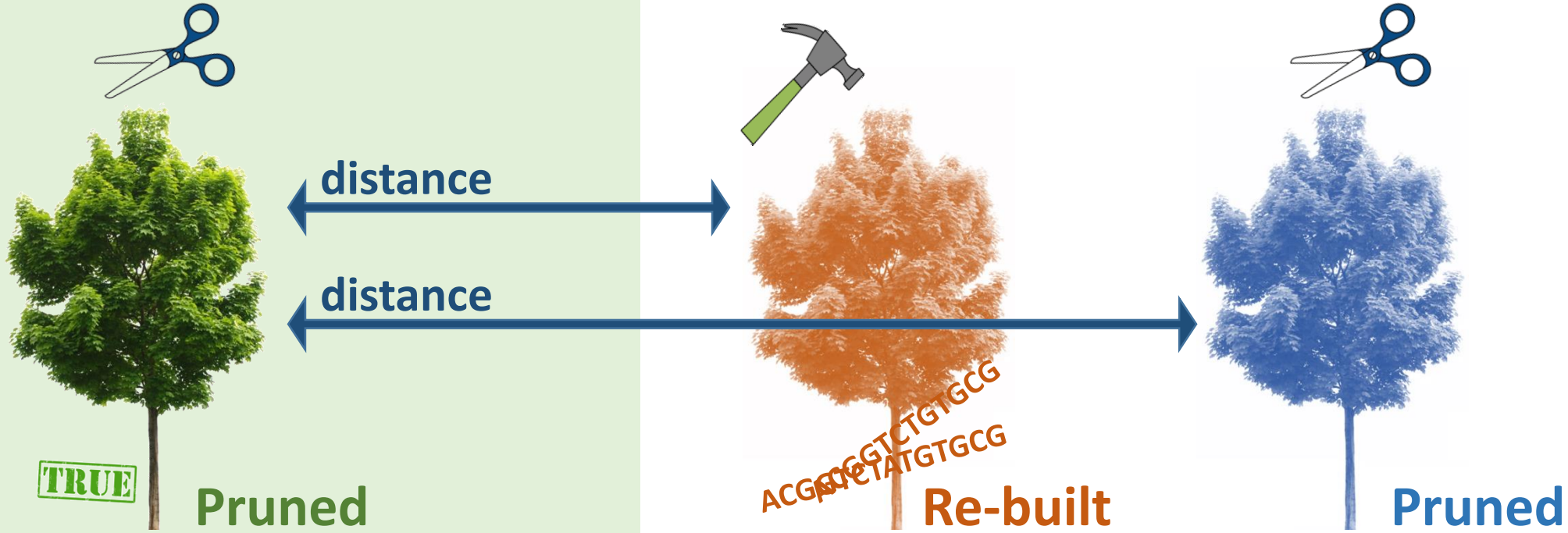
distance

distance

Re-built

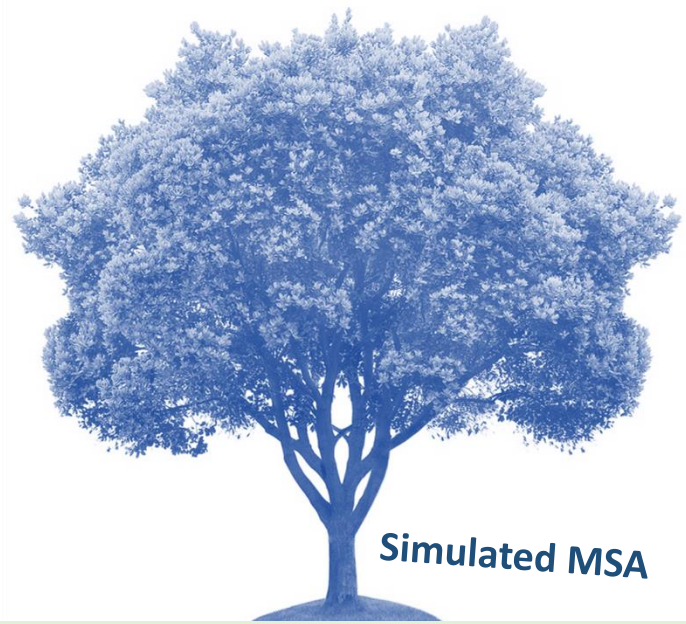


Pruned



Which of the distances we expect to be larger?

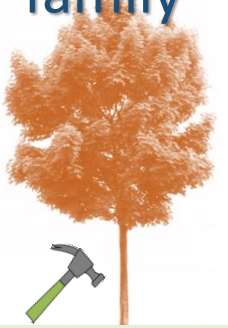
~ 15,000 species



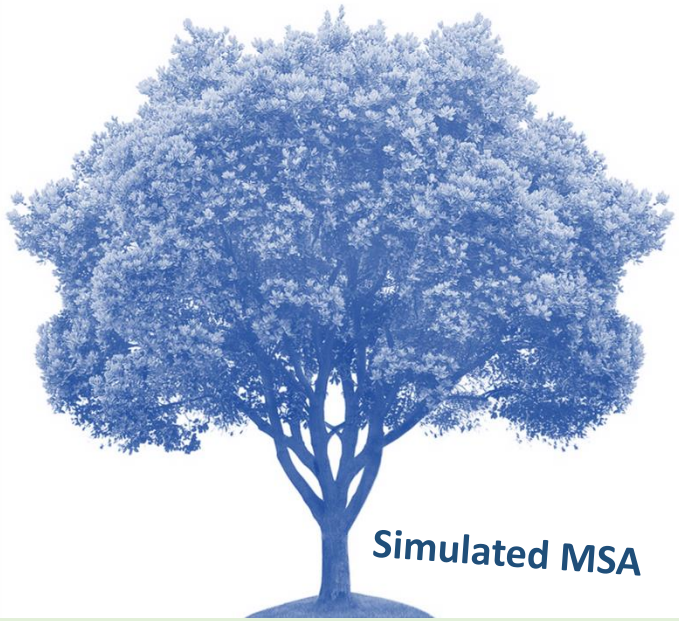
order

family

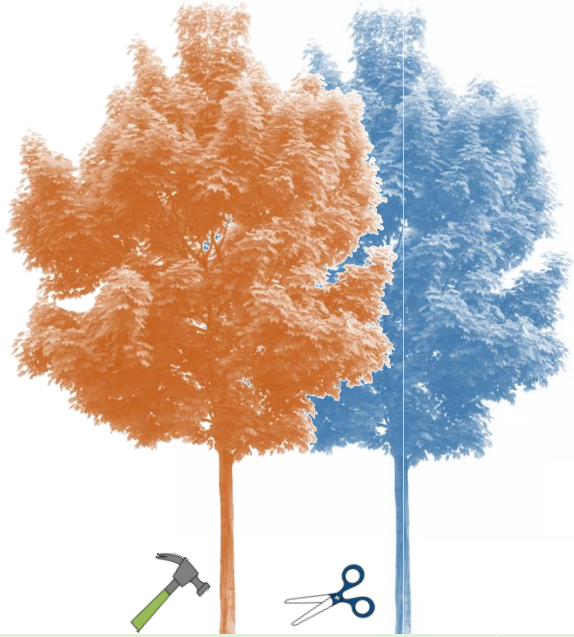
sub-
family



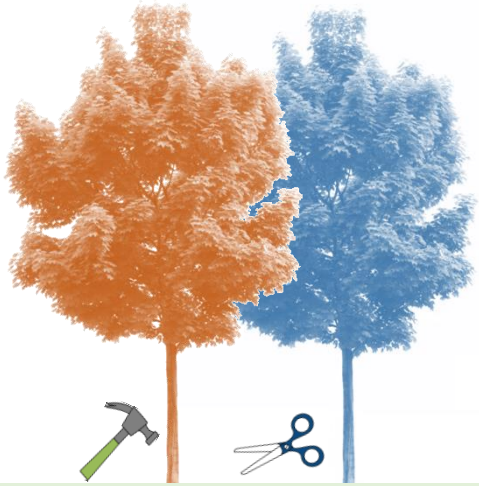
~ 15,000 species



order



family



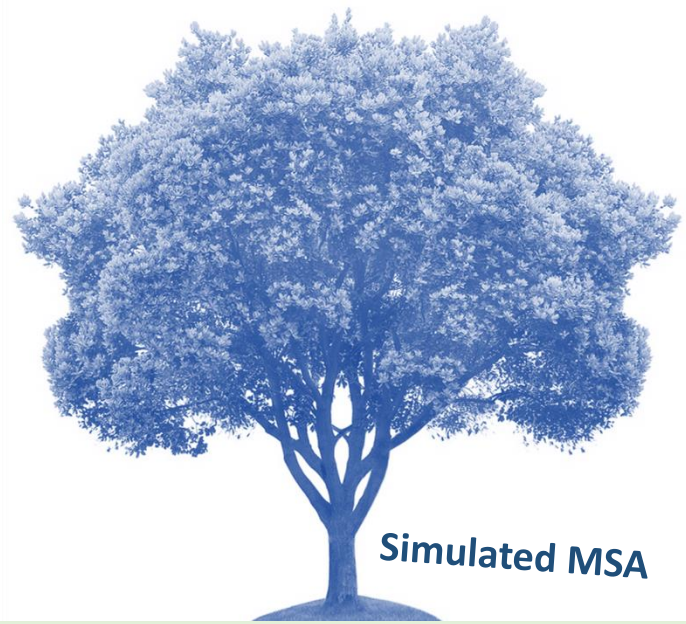
sub-family

426 440

330 342

210 212

~ 15,000 species

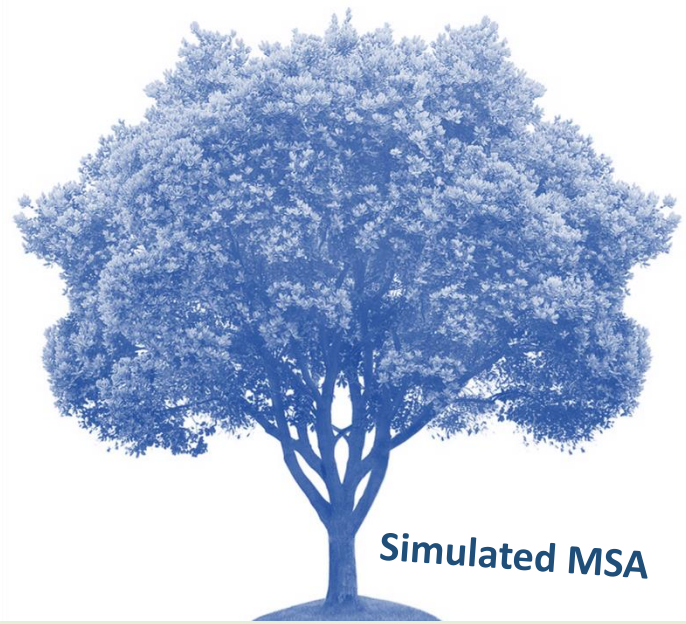


order

family

sub-
family

~ 15,000 species



order

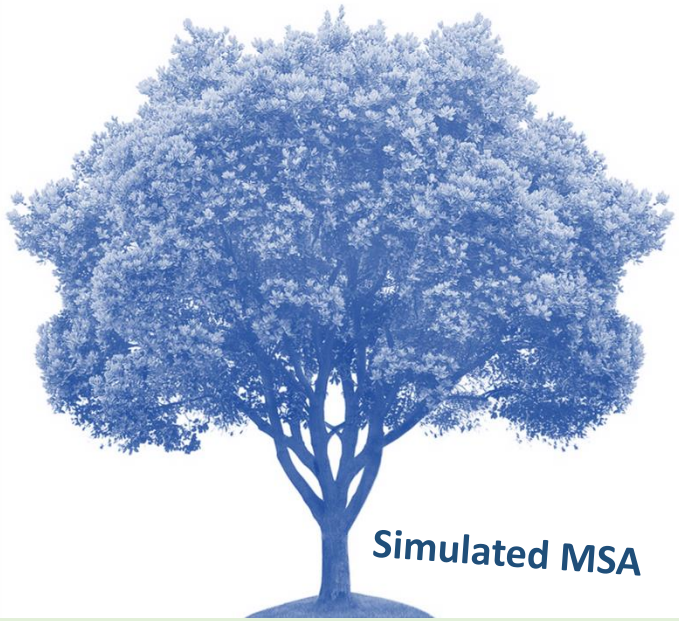
family

sub-
family

~ 15,000 species

sub-
family

~ 15,000 species



order

family

sub-
family

~ 15,000 species

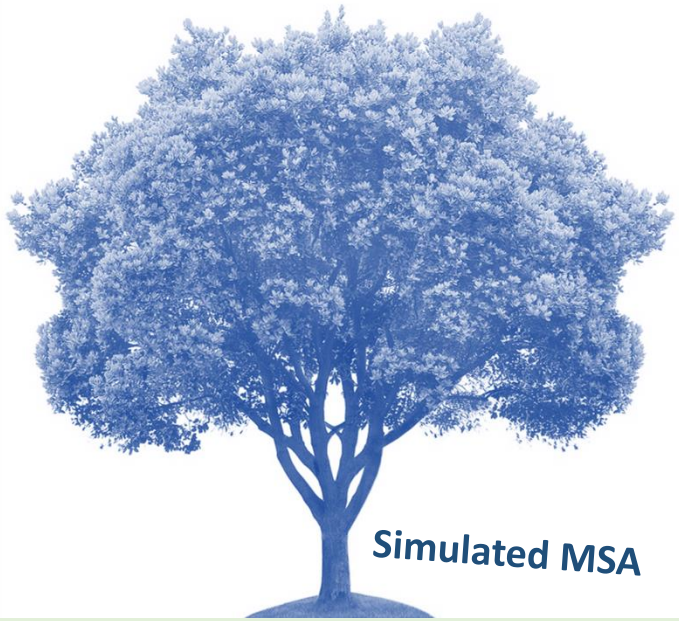
d1

d2

d3

sub-
family

~ 15,000 species



order

family

sub-
family

~ 15,000 species

d1

d2

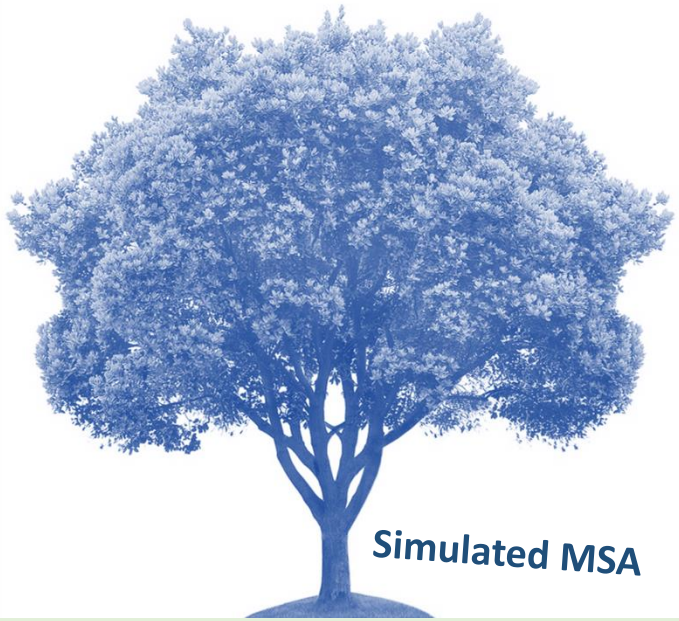
d3

?

d1	d2	d3

sub-
family

~ 15,000 species



order

family

sub-
family

~ 15,000 species

d1

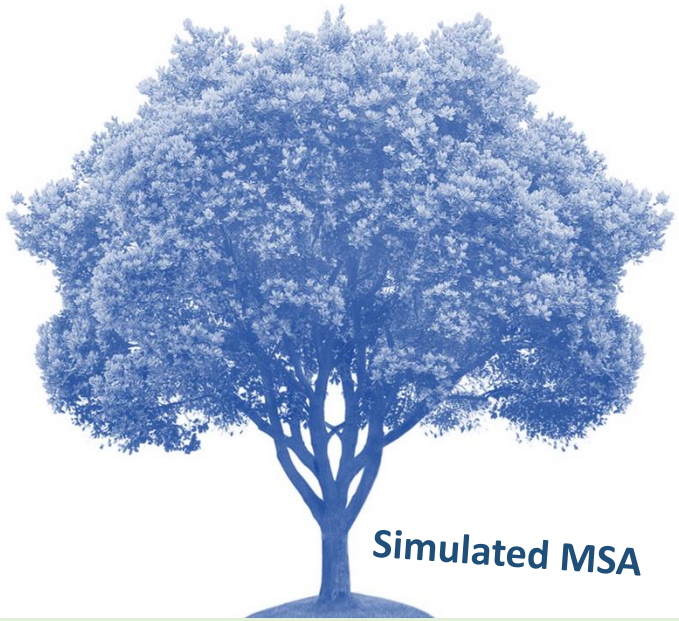
d2

d3

d1 > d2 > d3

sub-
family

~ 15,000 species



order

family

sub-
family

~ 15,000 species

210

204

210

d1 > d2 > d3

sub-
family

Two ways to create the sub-family tree:

family

family MSA

Leave rows of subfamily only

family tree

build

prune

sub-family tree

order

sub-family

family

sub-family

sub-family

sub-family

d1

d2

d3

sub-family

TRUE

SpartaABC:

GENE	ORIGINAL_MSA_LENGTH	INDELIBLE_MSA_LENGTH
18S	1818	340
26S	3490	872
atpB	1509	779
ITS	1221	1253
matK	1996	1119
rbcL	1423	905
trnLtrnF	1316	1408


Hope you could
see the forest...