Whole genome alignments

Genome 559: Introduction to Statistical and Computational Genomics

Prof. James H. Thomas

On Problem Set 2

(this is a slide from a lecture)

Traceback (local)

		A	A	G
	0	0	0	0
G	0	0	0	2
A	0	2	2	0
A	0	2	4	0
G	0	0	0	6
G	0	0	0	2
C	0	0	0	0

AAG

AAG

for local alignment only the aligned residues are shown

On Problem Set 2

- Most common programming mistake was not to be sure your program works on any input.

- I provide an example to clarify what the output should look like, but your program needs to work for ANY input.

- Think of what possible inputs might give a problem - ideally your program will always do something sensible regardless of what you throw at it.

- Later in the course we'll cover ways to manage problem input much more systematically.

Review

- What a score matrix is and how to calculate one.
- Why an affine gap penalty is desirable.
- How to align sequences using dynamic programming.
- How to calculate and interpret p-values and E-values for pair alignments and database searches.

Whole genome alignments

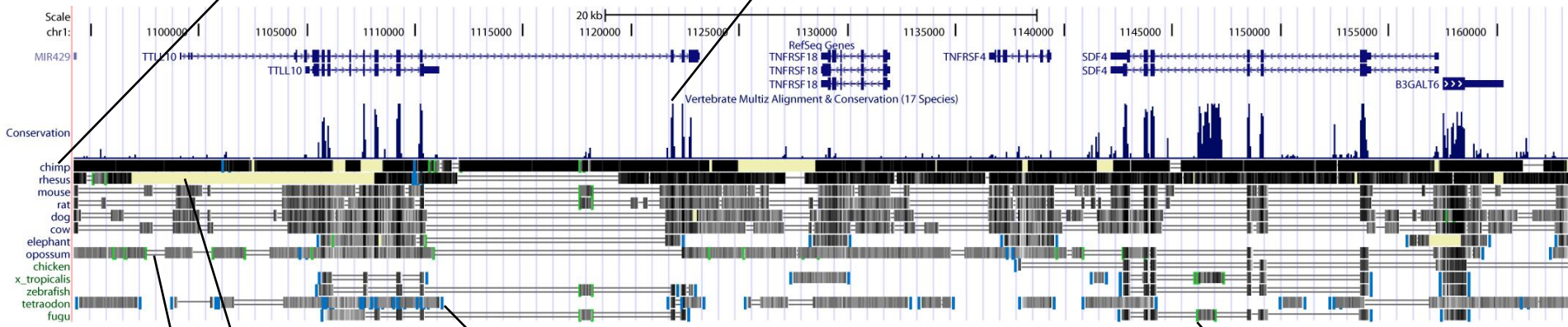
Why?

- genome-wide alignment data (efficient)
- inference of shared genes across many species
- genome evolution

UCSC Browser track

individual genome alignments, darker = higher scoring

averaged conservation for 17 genomes

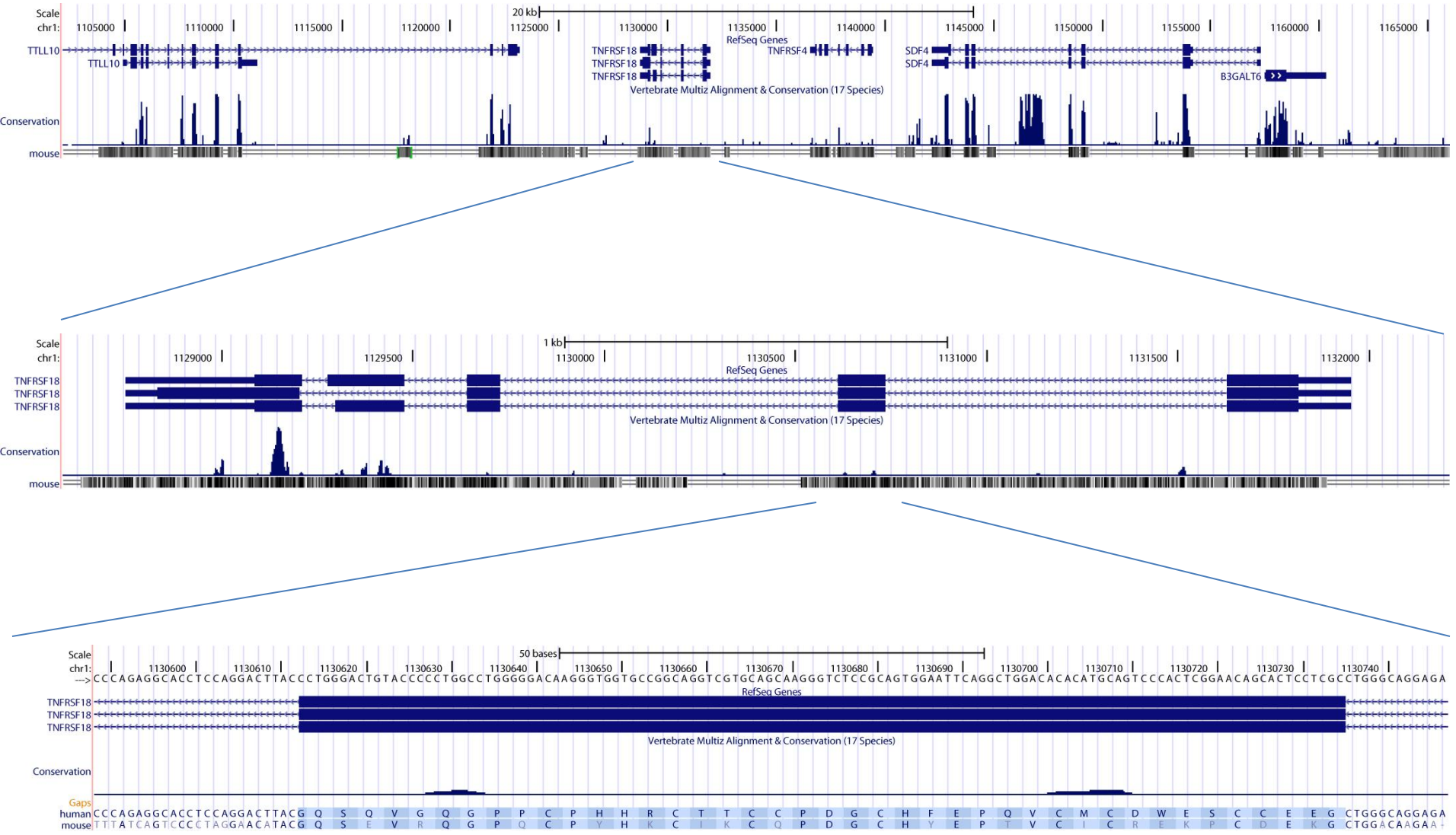


known gap in assembly

alignment discontinuity (e.g. translocation break point)

questionable alignment segment

= sequence present but unalignable



GQSQVGQGPPCPHHRCTTCCPDGCHFEPQVCMCDWESCCEEG

GQSEVRQGPQCPYHKCIKCQPDGCHYEPTVCICREKPCDEKG

How are genome-wide alignments made?

- mouse and human genomes are each about 3×10^9 nucleotides.

- how many calculations would a dynamic programming alignment have to make?

- at a minimum - 3 integer additions and 3 inequality tests for each matrix position

- matrix size is 3×10^9 by 3×10^9

- about $6 \times (3 \times 3 \times 10^{18}) = 5.4 \times 10^{19}$ calculations!

  Age of the universe is about 4.3×10^{17} seconds

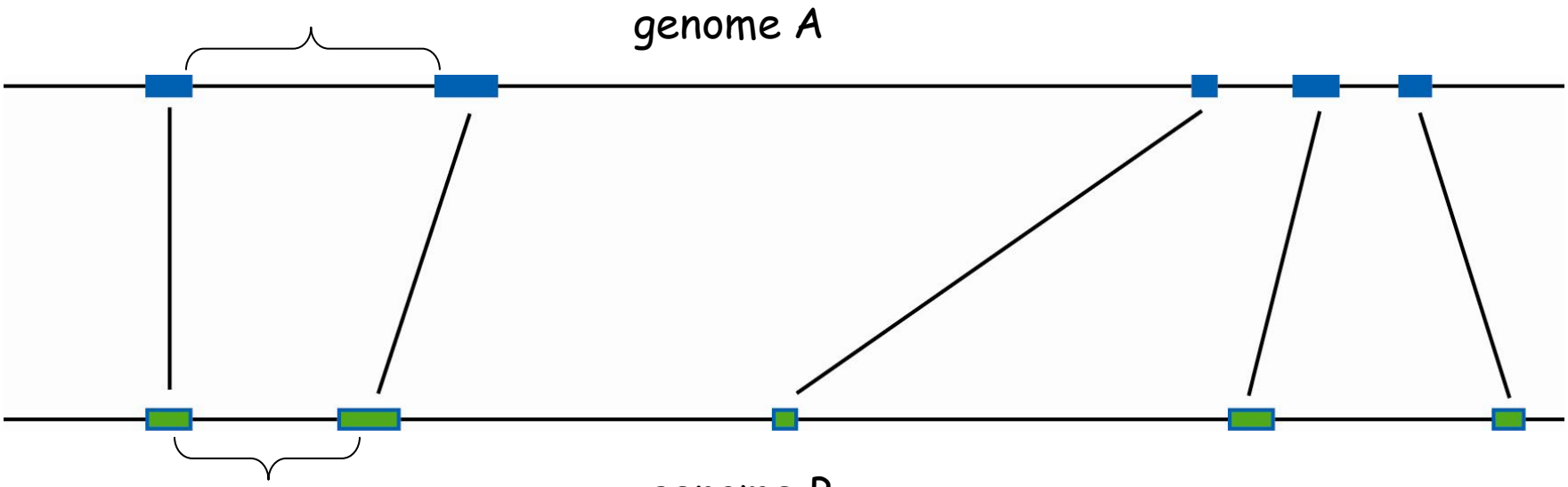
(by the way, there are other problems too, including assuming colinearity)

BLAST-like alignment seeding

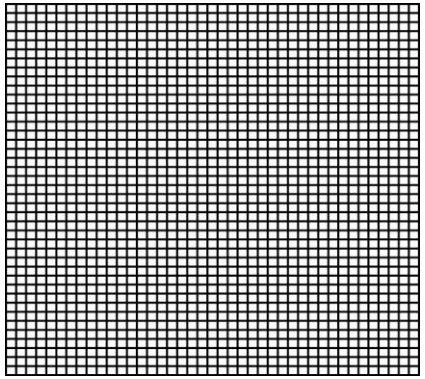
- later in the course you'll learn how BLAST works, for now it is sufficient that it is several orders of magnitude faster than DP.

- use BLAST to find local high-quality alignments

- extend from these alignments using DP



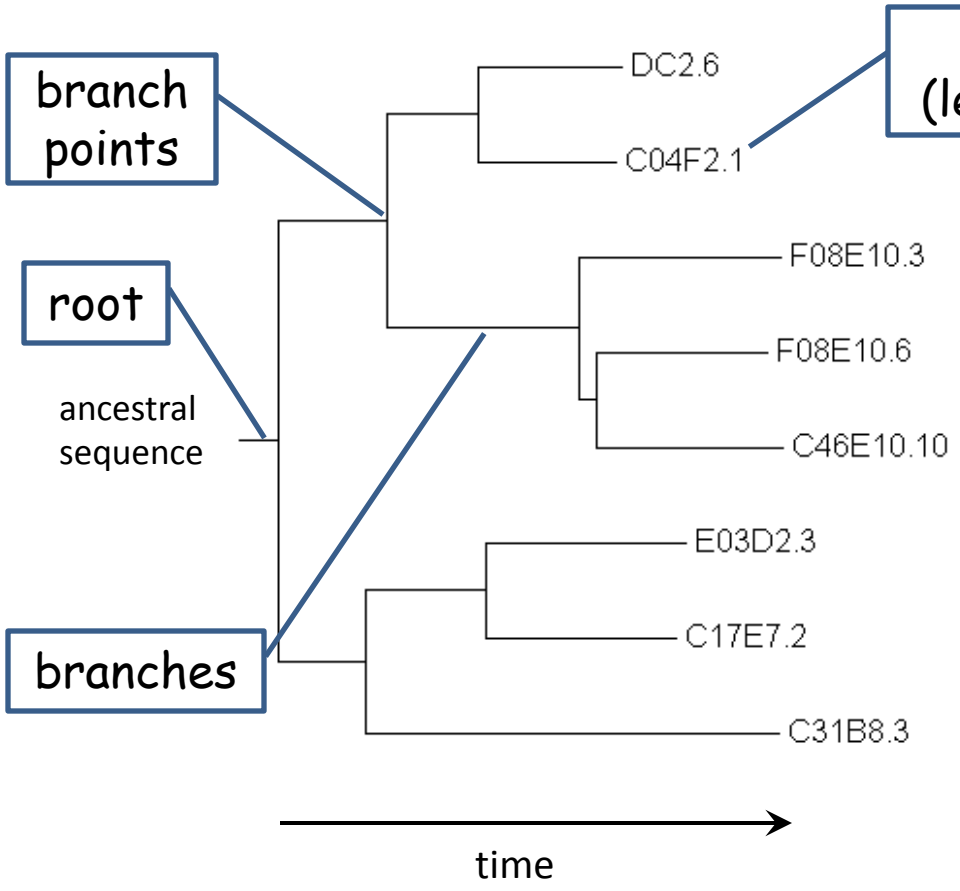
DP alignment region



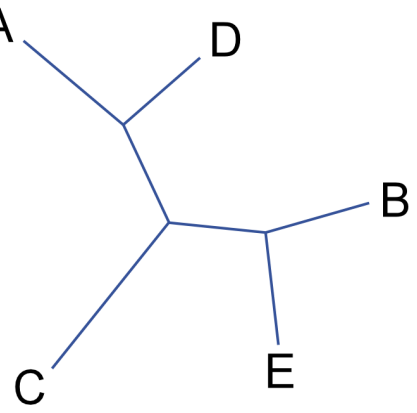
M x N manageable

Defining what a "tree" means

rooted tree (all real trees are rooted):



unrooted tree (used when the root isn't known):

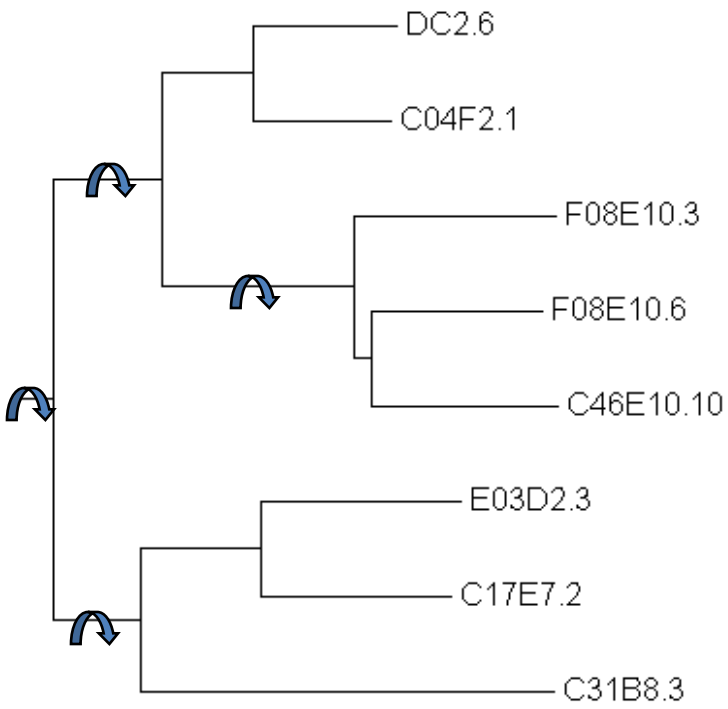


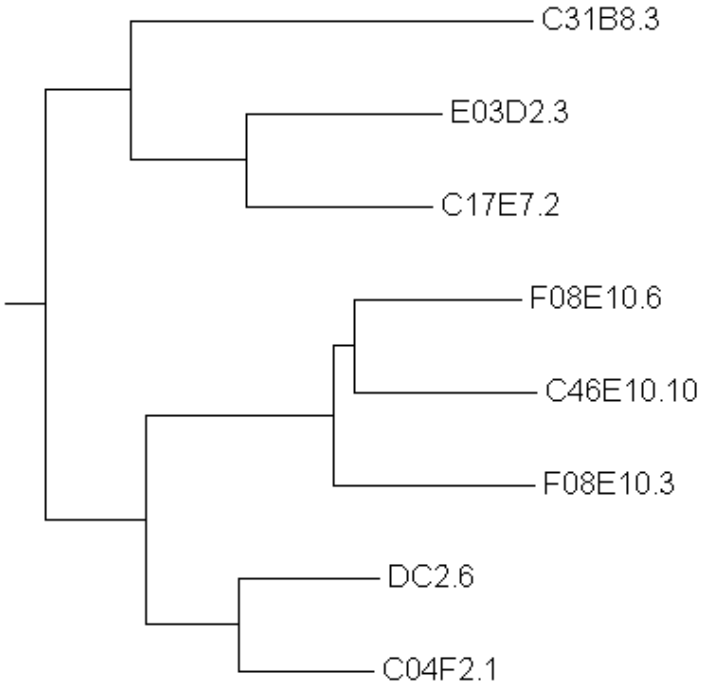
time vaguely radiates out from somewhere near the center

…divergence time is the sum of (horizontal) branch lengths

A tree has topology and distances

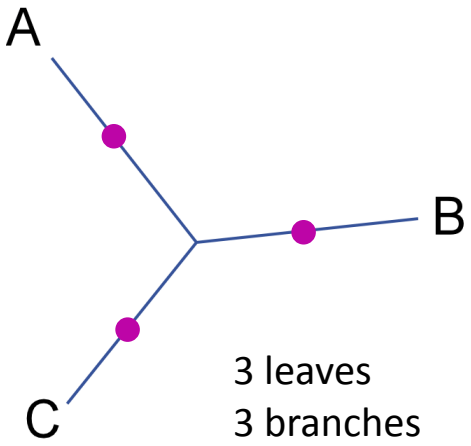
Are these different trees?



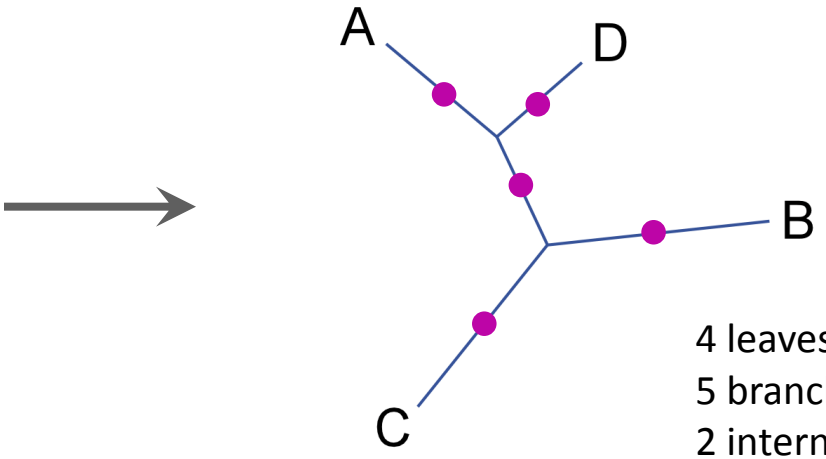


Topologically, these are the SAME tree. In general, two trees are the same if they can be inter-converted by branch rotations.

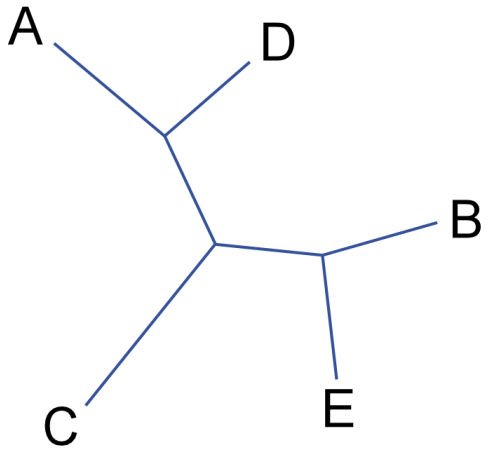
The number of tree topologies grows extremely fast



3 leaves
3 branches
1 internal node
1 topology
(3 insertions)



4 leaves
5 branches
2 internal nodes
3 topologies (x3)
(5 insertions)

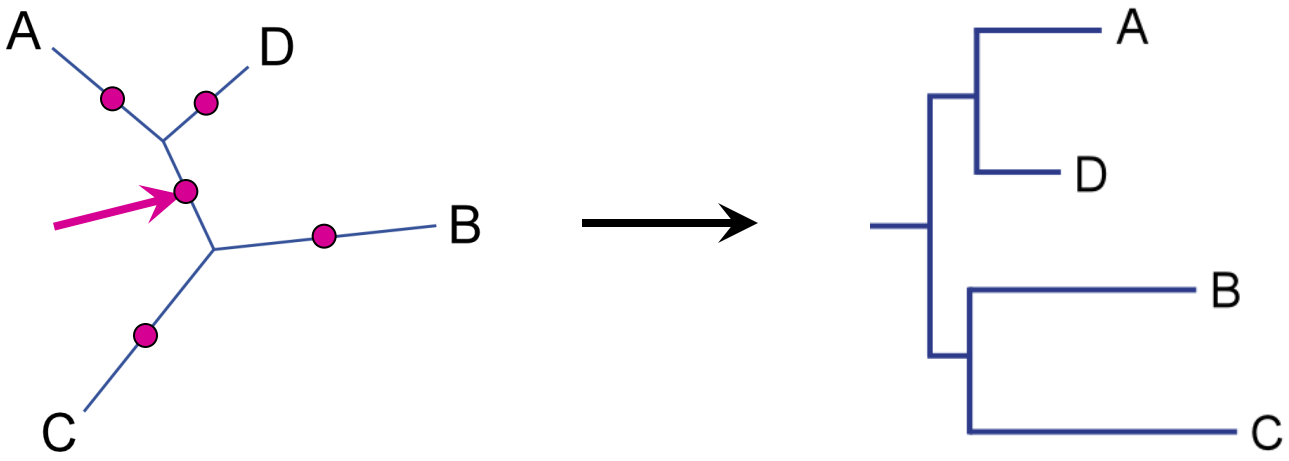


5 leaves
7 branches
3 internal nodes
15 topologies (x5)
(7 insertions)

In general, an unrooted tree with N leaves has:
2N – 3 branches
N – 2 internal nodes
~ O(N!) topologies $3*5*7*\ldots*\ 2N-5$

There are many rooted trees for each unrooted tree

For each unrooted tree, there are 2N - 3 times as many rooted trees, where N is the number of leaves (# internal branches = 2N – 3).



20 leaves - 564,480,989,588,730,591,336,960,000,000 topologies