{Lecture 19}

- Introduction to Phylogenetic Reconstruction.
- See Slideshow for structured lecture; here we provide some formal definitions, as well as some omitted details.
- Notation: - present-day species are $1, 2, \dots, n$
  - denote $\{1, 2, \dots, n\}$ by $[n]$.
- **Def:** A phylogenetic tree over $[n]$ is a $[n]$-leaf labeled binary tree, i.e. an undirected tree $T = (V, E)$ w/ n leaves that are labeled $1, 2, \dots, n$ and all internal nodes w/ degree 3.

- **Lemma 1:** A phylogenetic tree over $[n]$ has exactly $2n-2$ nodes.

  Proof: - Let $T = (V, E)$ be a phylogenetic tree.

  - $2|E| = n + 3(|V| - n)$ (since leaves have degree 1, & internal nodes have degree 3)

  also $2 \cdot |E| = 2 \cdot (|V| - 1)$ (this is true for all trees)

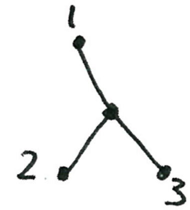
  $$\Rightarrow 2|V| - 2 = n + 3|V| - 3n \Rightarrow |V| = 2n - 2$$ $\boxtimes$

- We proceed to count the number of phylogenetic trees over $[n]$.

- **Lemma 2:** There are exactly $(2n-5)!! = (2n-5)(2n-7)(2n-9)\dots 3$ phylogenetic trees over $[n]$ (up to graph isomorphisms).

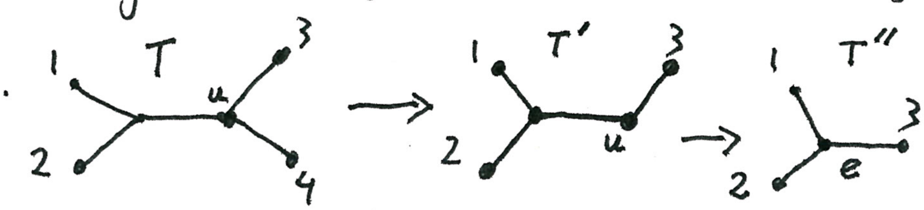
Proof: (by induction)

- $n=3$: there is clearly exactly one phylogenetic tree over $[3]$



- suppose claim is true for n
- inductive step: we define a bijective mapping from the set of phylogenetic trees over $[n+1]$ to the cartesian product of phylogenetic trees over $[n]$ and their edge sets.
  - let $T=(V,E)$ be a phylogenetic tree over $[n+1]$
  - remove leaf $n+1$ and its edge.
  - this creates a node u of degree 2 in the resulting graph T'
  - contract the edges adjacent to that node into one edge (hence eliminating the node); let T'' be the resulting tree

e.g.



  - record the edge e that was created in previous step

$$T \longmapsto (T'', e)$$

  - easy to see that the mapping is a bijection

- It follows that the number

$$\begin{pmatrix}\text{\# phylogenetic} \\ \text{trees over} \\ [n+1]\end{pmatrix} = \begin{pmatrix}\text{\# phylogenetic} \\ \text{trees over} \\ [n]\end{pmatrix} \times (2n-3)$$

since by Lemma 1 a phylogenetic tree over $[n]$ has $2n-3$ edges

$$\Rightarrow \quad = (2n-3)!!$$

$\boxtimes$

- **Lemma 3 (Information Theoretic Lower Bound on Sequence Length):**
  - Suppose the input to a phylogenetic reconstruction algorithm is a sequence of length k over $\{A,C,G,T\}$ for every leaf in $[n]$.
  - Suppose that, w prob $\geq \frac{3}{4}$, over its internal randomness the algorithm is correct
  - Suppose all phylogenetic trees over $[n]$ are possible correct answers, over the set of possible inputs to the algorithm.
  - Then $k = \Omega(\log n)$.

- Proof:
  - We prove the lemma for deterministic algorithms using counting, and leave the generalization to randomized algorithms as an exercise (1pt).
  - $\#$ possible outputs $= (2n-5)!! \geq \sqrt{(2n-6)!!} = 4^{\Omega(n \cdot \log n)}$ (using lemma 2 & the fact that all phylogenetic trees over $[n]$ are possible outputs)
  - $\#$ possible inputs $= 4^{k \cdot n}$

- For all possible outputs to be output by the algorithm for some input we need:

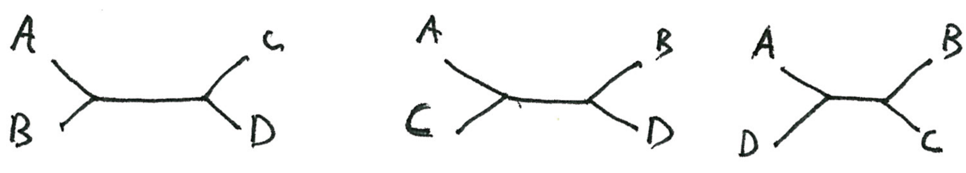
$$4^{kn} \geq (2n-5)!! = 4^{\Omega(n \cdot \log n)}$$

$$\Rightarrow k = \Omega(\log n).$$

$\boxtimes$

Quartet - Methods

- Observation: there are 3 possible trees on 4 species

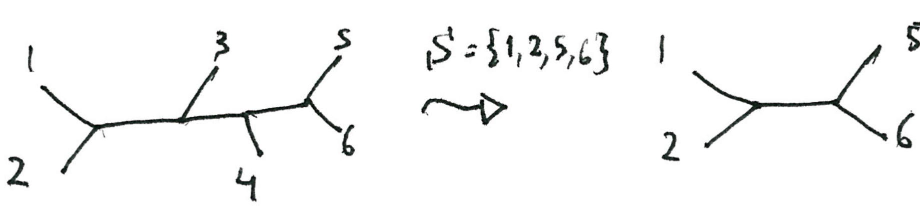


AB|CD AC|BD AD|BC

these are called "quartets"

- A phylogenetic tree induces a quartet on all subsets of 4 species by removing all other species and then contracting all paths comprised of nodes of degree 2 into a single edge

e.g.



- **Theorem**: Let T be a phylogenetic tree over $[n]$. Suppose we are given all quartets induced by T on all subsets of 4 leaves $S \in \binom{[n]}{4}$. Using the quartets, can reconstruct T.

Proof:

Claim 1: Every phylogenetic tree over $[n]$ has a cherry $\{i,j\} \subseteq [n]$, i.e. a pair of leaves at distance 2.

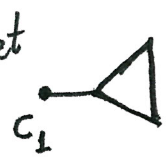
Proof: - Suppose not; then the tree should have at least $2 \cdot n$ nodes (since the "father" of a leaf belongs only to that leaf)

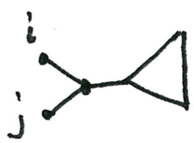
- But we've shown that a phylogenetic tree over $[n]$ has $2n-2$ nodes (contradiction). $\boxtimes$

Claim 2: If I have all quartets induced by a phylogenetic tree over $[n]$, I can identify all cherries.

Proof: If a pair of leaves $\{i,j\}$ is a cherry then i,j never appear on opposite sides of a quartet; and, vice versa, if a pair of leaves $\{i,j\}$ never appears on opposite sides of a quartet, then it's a cherry. $\boxtimes$

To conclude the proof of the theorem:

- look at quartets to identify a cherry $\{i,j\}$
- replace all occurrences of i and j by c_1 in all quartets and throw away all resulting quartets w/ two occurrences of c_1
- inductively reconstruct phylogenetic tree over $([n] \setminus \{i,j\}) \cup \{c_1\}$
- Let  be the tree

- return 

$\boxtimes$