Lecture 20

- Last time we ~~saw~~ that if we have all quartets induced by a phylogenetic tree over $[n]$, we can reconstruct the tree.

- This time we examine other combinatorial structures that are sufficient for this purpose based on distance information.

- **Def (Path metric):** Let $T=(V,E)$ be a tree w/ positive edge weights $w=\{w_e\}_{e\in E}$. For all pairs of vertices $u,v\in V$ let $\text{Path}(u,v)$ be the set of edges in the unique path between u and v. The path metric induced by (T,w) is:

$$d_{T,w}(u,v)=\sum_{e\in \text{Path}(u,v)} w_e .$$

- **Def (Dissimilarity map):** A dissimilarity map on a set X is a function $\delta: X\times X \to \mathbb{R}$ s.t. $\delta(x,x)=0$ and $\delta(x,y)=\delta(y,x)$ for all $x,y\in X$.

- **Def (Tree metric):** A dissimilarity map δ on $[n]$ is a tree metric if there exists a phylogenetic tree T over $[n]$ and a collection of pos. weights w on its edges so that the path metric induced by the tree and the edge-weights agrees w/ δ on all pairs of leaves.

i.e. $d_{T,w}(x,y)=\delta(x,y)$, $\forall\, x,y\in[n]$

In this case (T, w) is called a tree representation of the tree metric δ.

Theorem: Let δ be a tree metric on $[n]$. Up to isomorphism, there exists a unique tree representation of δ, which can be constructed in polynomial time.

Proof:
- Recall from last lecture that, up to isomorphisms, a tree is determined by its quartets (on all subsets of 4 leaves)
- Let (T, w) be a tree metric representation of δ. We'll show that, using δ, we can obtain all quartets induced by T and therefore obtain T; then finding w will be easy.
- Pick any subset $\{x, y, z, w\} \subseteq [n]$ of leaves and look at the expression:

$$\frac{1}{2}\left(\delta(x,w) + \delta(y,z) - \delta(x,y) - \delta(w,z)\right) \qquad (*)$$

this expression is equal to:

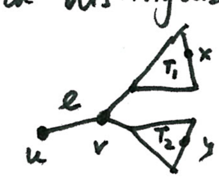
- the weight of the interior path of the quartet formed by x, y, z, w, if $xy|zw$ is that quartet
- minus the above if $xw|yz$ is the quartet formed by x, y, z, w
- 0 o.w. (i.e. if $xz|wy$ is the quartet)

- So using δ can find all quartets and obtain T (this also shows that the tree must be unique, as determining the quartets only depends on δ; in particular all tree representations of δ should have the same tree up to isomorphisms)

- Computing the weights:

pick an edge $e = \{u, v\}$ of tree and distinguish two cases:

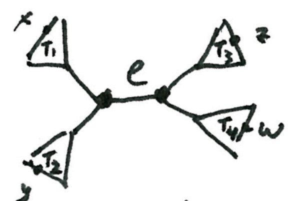
i> e is pendant



then choose leaf x in T_1 and y in T_2 and set

$$w_e = \frac{1}{2}\left(\delta(u,x) + \delta(u,y) - \delta(x,y)\right).$$

ii> e is interior



then choose leaves x in T_1, y in T_2, z in T_3, w in T_4 and set

$$w_e = \frac{1}{2}\left(\delta(x,w) + \delta(y,z) - \delta(x,y) - \delta(z,w)\right).$$

$\boxtimes$

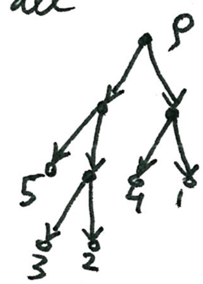
Remark:

Let δ be a tree metric and (T, w) its tree representation. Let also $w^* = \min_e \{w_e\}$, and suppose that we are given $\hat{\delta}$ s.t.

$$|\delta(x,y) - \hat{\delta}(x,y)| < \frac{1}{4} w^*, \text{ for all } x, y \in [n].$$

Using $\hat{\delta}$ we can compute T in polynomial time. See part (*) in the proof of the theorem.

- **Def (Markov Chain on a Tree):** Consider a rooted binary tree T, whose root is labeled ρ and whose leaves are labeled $1, \dots, n$. Suppose that all edges of T are directed away from the root. In particular, the root has out-degree 2 and in-degree 0, the leaves have in-degree 1 and out-degree 0, and all internal nodes have in-degree 1 and out-degree 2.



  - Let now C be a finite character set, e.g. $C = \{A, C, G, T\}$ or $C = \{0,1\}$ and M_C be the set of all transition matrices on C, i.e. stochastic matrices $|C| \times |C|$.
  - Suppose $P = \{P^e\}_{e \in E} \in M_C^E$ and μ_ρ a probability distri over C.
  - A Markov Chain on a tree (T, P, μ_ρ) is the following stochastic process $\Xi_V = \{\Xi_v\}_{v \in V}$:
    - pick a state Ξ_ρ for ρ according to μ_ρ
    - moving away from the root towards the leaves, apply to each edge e the transition P^e independently from everything else.
  - Denote by μ_V the distribution over C^V thus obtained.

- For $\xi_V = \{\xi_v\}_{v \in V} \in C^V$,

$$\mu_V(\xi_V) = \underbrace{\mu_\rho(\xi_\rho)}_{\Pr[\Xi_\rho = \xi_\rho]} \prod_{\substack{e = \{u,v\} \in E \\ u \leq_T v}} \underbrace{P^e_{\xi_u, \xi_v}}_{\Pr[\Xi_v = \xi_v \mid \Xi_u = \xi_u]} \qquad (**)$$

- e.g. CFN model:
  - $\mathcal{C} = \{0, 1\}$
  - $\mu_\rho = \left(\frac{1}{2}, \frac{1}{2}\right)$
  - $P^e = \begin{bmatrix} 1-p_e & p_e \\ p_e & 1-p_e \end{bmatrix}$ p_e: mutation probability on edge e

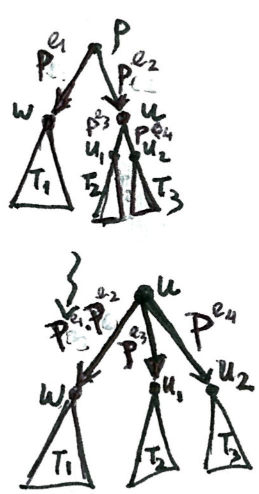
- **Def (Reconstruction Problem):** Let $\underset{\sim}{\Xi} = \{\Xi^1_{[n]}, \dots, \Xi^k_{[n]}\}$ be i.i.d. samples from a MCT (T, P, μ_ρ) projected on the leaf-set $[n]$. Given $\underset{\sim}{\Xi}$:
  - the reconstruction problem is to find a phylogenetic tree $\hat{T}$ over $[n]$ s.t. $\hat{T} = T^{-\rho}$

    $T^{-\rho}$ is obtained from T by undirecting all edges & removing ρ merging its adjacent edges

  - the full reconstruction problem: for a given $\varepsilon > 0$, find MCT $(\hat{T}, \hat{P}, \hat{\mu}_{\hat{\rho}})$ s.t. $\hat{T}^{-\hat{\rho}} = T^{-\rho}$ and the corresponding distn' $\hat{\mu}_{[n]}$ satisfies:

$$\|\mu_{[n]} - \hat{\mu}_{[n]}\|_{TV} \leq \varepsilon.$$

- **Remarks:** 1. cannot hope to get location of the root ρ



e.g. take CFN model w/ root ρ; easy to re-root at some child of ρ without changing distn' at the leaves; and hence root at any node w/out changing distn' at the leaves.

2. The reconstruction problem is identifiable under the conditions:

$$\mu_\rho > 0$$

$$\forall e \in E, \ \det(P^e) \neq 0, \pm 1$$

e.g. cannot distinguish quartets $ab|cd$ vs $ac|bd$

- if all edges of quartet have transition matrix

$$P = \begin{pmatrix} p & 1-p \\ p & 1-p \end{pmatrix} \quad (\det P = 0)$$

- if all edges of quartet have transition matrix

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (\det P = 1)$$

- Reconstruction in the CFN Model

$$u \xrightarrow{e} v \qquad P^e = \begin{pmatrix} 1-p_e & p_e \\ p_e & 1-p_e \end{pmatrix} = (1-2p_e)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + 2p_e \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

interpretation: w/ prob. p_e there is a mutation

new interpretation:
- w/ prob $1-2p_e$ v copies u
- w/ prob $2p_e$ v is independent of u

call $\theta_e = 1 - 2p_e$ the probability of copying

Next time we use θ_e's to define a tree-metric on $[n]$.