First, given a phylogenetic tree (T, ℓ) where T is a rooted binary tree and $\ell : E(T) \to \mathbb{R}^+$ are edge lengths (exp. # mutations).

Assume that foreach length ℓ we can obtain CPD $\theta_\ell = (\theta_\ell^A, \theta_\ell^C, \theta_\ell^G, \theta_\ell^T)$, where θ_ℓ^A gives the prob. of mutating from A to any other nucleotide (when ℓ mut. are expected).

Ex. Jukes-Cantor

m is Poisson dist. with exp. ℓ and

$$\theta_\ell \text{ is } \begin{array}{c|cccc} & A & C & G & T \\ \hline A & 0 & & & \\ C & & 0 & & 1/3 \\ G & & 1/3 & 0 & \\ T & & & & 0 \end{array}^{\,m}$$

This induce a DGM on any binary tree T with edge lengths ℓ and r.v.s $S_u = s_u^1, \dots, s_u^m$ $\forall u \in V(T)$ through

$P(S_u \mid S_{pa(u)}) = \prod_{j=1}^{m} P(s_u^j \mid s_{pa(u)}^j, \theta_{\ell((u, pa(u)))})$.

For observed sequences $\{S_e\}_{e \in L(T)}$, we can comp. the density

$$\gamma(T, \ell) = p(\{S_e\}_{e \in L(T)} \mid T, \{\theta_{\ell(e)}^e\})$$

$$= \sum_{\substack{s_v \text{ for} \\ v \in V(T)\setminus L(T)}} \prod_{i, u \in V(T)} p(s_u^i \mid s_{pa(u)}^i, \theta_{\ell(u,pa(u))}^{(u,pa(u))})$$

Phylogeny

Input: sequences $\{s_u\}_{u\in L}$ where $s_\ell = s_\ell^1, \dots, s_\ell^m$

Output: phylotree (T,ℓ) s/t $L(T)=L$ and $\ell(u)=s_u$ for $u \in L$ maximizing $\gamma(T,\ell)$

Alternative.

Output: the corresponding posterior

one less comp.

Idea: grow forrests rather than sequences ($F_i \to F_{i+1}$ ins. of $z_{1:i} \to z_{1:i+1}$).

We consider the ultratric case, i.e., all root to leaf paths have the same length.

We will use partially ordered sets (posets).

A poset is a pair $(S, \preceq)$ s/t $\preceq$ is a binary rel. on S^2 sat.

- reflexivity
- antisymetry
- transitivity.

Here S = set of ultrametric forrests with leaves $L(T)$ with edge lengths.

Extension γ_* of γ. Let A be a heuristic that join the most similar components and add lengths. Let $\gamma_*(F,\ell) = \gamma(A(F,\ell))$.

q proposal, q^n n steps with if

Assumption 1: $\exists n$ s/t $q^n(s \to s') > 0$ iff $s < s'$

Here $q^n(F,\ell \to F',\ell') > 0$ iff $(F,\ell) < (F',\ell')$

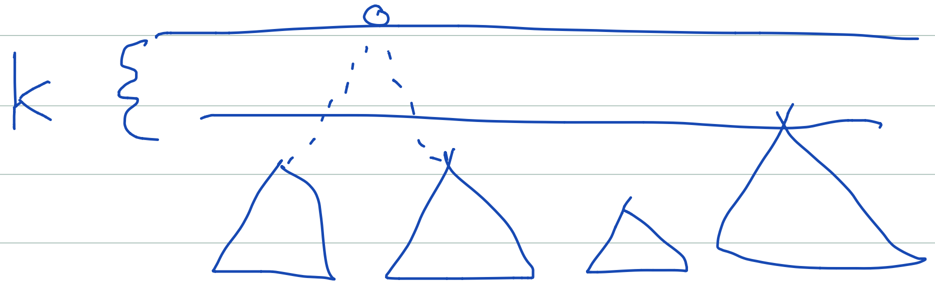
Assumption 2: each ~~element~~ Forrest has a unique predecessor (largest smaller) Can be relaxed! $\exists$ a path to each maximal element. I.e. tree.

Assumption 3: $\gamma_*(a) > 0$ $\forall$ elements a.

Under these assumptions SMC on the poset gives the dist. γ.

The goal is the dist. over the maximal elements.

We propose to merge s/t the height (i.e. longest root leaf path) is increased by α where α is exp. dist with rate $\binom{|F|}{2}$ (inspired by the coalescent process).



The roots to merge are picked uniformly. Call this distribution $q(F \to F')$.

Notice, every forrest F of non-zero height has a "highest" root removing it givs the unique predecessor of F.

Algorithm:

```
    Start with $F_1^k = L$,   $\forall k \in [K]$
    For $i := 2$ to $|L|$
        Sample $F_i^k \sim q(F_{i-1}^k \to F)$,  $\forall k \in [K]$
        let $w_{i,k} := \dfrac{\gamma_*(F_i^k)}{\gamma_*(F_{i-1}^k)\, q(F_{i-1}^k \to F_i^k)}$
```

Add resampling to this alg.

Theo. (here and in general)

Let $\pi_{r,K} = \sum_{k=1}^{K} w_{r,k}\, \delta_{F_r^k}(\cdot)$ and π the dist. ass. with γ.

Then

$$\pi_{L,K} \xrightarrow{K \to \infty} \pi.$$