Supplemental Material: Tumor Phylogeny Inference Using Tree-Constrained Importance Sampling

Gryte Satas[1] and Benjamin J. Raphael[2,†]

[1]Department of Computer Science, Brown University, Providence, Rhode Island
[2]Department of Computer Science, Princeton University, Princeton, New Jersey
[†]Correspondence: braphael@princeton.edu

Contents

A Supplementary Methods

A.1 Sequential Importance Sampling

Importance sampling functions well when the number of vertices in the tree k and the number of samples m are low. However, the number of dimensions we are sampling from for $\mathbf{F}$ increases with $k \cdot m$. Thus, in the presence of noise, sometimes, for all samples $\bar{F}$, $\mathcal{T}_{\bar{F}}$ is empty, and thus the calculated posterior probability of all trees is 0. This occurs when the largest tree that $\bar{F}$ admits, $T_{\bar{F}}$ has less than k nodes, $|V(T_{\bar{F}})| < k$. However, some of these samples may in fact be near another value $\bar{F}' \approx \bar{F}$ such that there exists a (T, π) for which $\bar{F}'$ respects the Sum Condition. As the number of nodes in the largest tree admitted by samples increases, we would expect we are nearing such a value.

We generalize importance sampling to use multiple proposal distributions, $\mathrm{Q}_1, \ldots, \mathrm{Q}_N$, where Q_i is the distribution used at step i. As the estimate obtained from each distribution Q_i is an unbiased estimator of $\int \mathrm{P}(\mathbf{X}) d\mathbf{X}$, the mean is of these values is also an unbiased estimator.

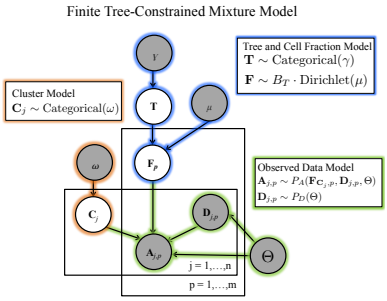
$$\int \mathrm{P}(\mathbf{X}) dX \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\mathrm{P}(\bar{X}_i)}{\mathrm{Q}_i(\bar{X}_i)}. \tag{1}$$

As such, we propose an MCMC inspired multiple-importance sampling approach. Q_0 is the proposal distribution supplied by SciClone in Section 2.3.1. Let $\mathrm{Beta}(\bar{F}, \Sigma)$ be the beta distribution with the values $\bar{F}$ as the means, and matrix Σ as the variances. At step $i+1$,

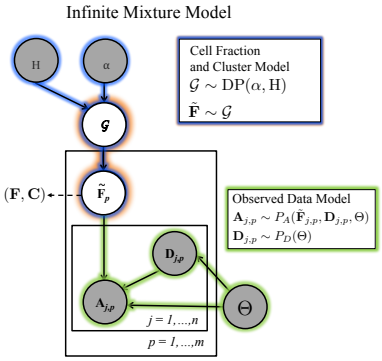
1. let $\mathrm{Q}_{i+1} = \mathrm{Beta}(\bar{F}_i)$ if $|V(T_{\bar{F}_i})| > |V(T_{\bar{F}_{i-1}})|$ or with probability
$$p = \min\left(\frac{\Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = \bar{F}^{(i)})}{\Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = \bar{F}^{(i)})}, 1\right).$$
2. Let $\mathrm{Q}_{i+1} = \mathrm{Q}_0$ with probability ϵ.
3. Else $\mathrm{Q}_{i+1} = \mathrm{Q}_i$

This approach uses information from previous samples to find samples which may have finite probability. Unlike Markov Chain Monte Carlo (MCMC), this approach is able to use information outside of model likelihood, i.e. the unconstrained clustering information and the tree size, to guide sampling.

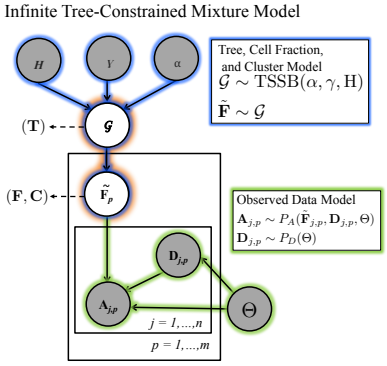
A.2 Comparison to Other Models



(a) The model used by PASTRI .



(b) A mixture model for binomial observations with a variable number of clusters (e.g. PyClone). G is a discrete distribution generated by a Dirichlet Process. $\tilde{F}$ is a $n \times m$ matrix, corresponding to an assignment of vector of frequencies (one per sample) to each mutation. All mutations with the same frequencies belong to the same cluster. Thus $\tilde{F}$ corresponds to both F and C from parts (a) and (b).



(c) An infinite mixture model with a tree constraint (e.g. PhyloSub). Instead of a Dirichlet Process, G is generated from a tree-structured stick breaking prior. This process generates frequencies $\tilde{F}$ that are consistent with a tree constraint.

Figure 1: **Comparison of models for mixtures with binomial observations.**

B Supplementary Results

B.1 Simulated Data Generation

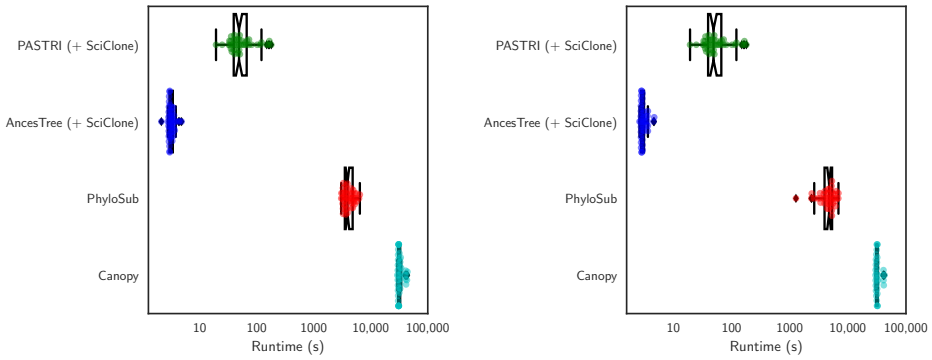
In Section 3.1, we evaluate PASTRI, AncesTree, PhyloSub and Canopy on simulated trees. Here, we generated trees with $k = 3, 4, 5$, $n = 20$, $m = 5$, and sequencing read depth $r = 200$. For each set of parameters, (k, m, n, r), 50 simulated instances were generated. Model parameters T, F, and C are generated according to the model presented in Section 2.1, with hyperparameters ω, γ, μ set such that the generating distributions are uniform. The observed data is generated as $d_{i,p} \sim \text{Poisson}(\lambda = r)$ and $a_{i,p} \sim \text{Binomial}(d_{i,p}, f_{c_j})$.

AncesTree, PhyloSub and Canopy were run with default parameters. Input for AncesTree was created using the clustering produced by SciClone. The observed read depths $\hat{A}$ and $\hat{D}$ were assigned as followed for cluster i in sample p,

$$a_{i,p} = \sum_{\ell; c_\ell = i} a_{\ell_p} \qquad\qquad d_{i,p} = \sum_{\ell; c_\ell = i} d_{\ell_p}. \tag{2}$$

PASTRI was run with $10,000$ iterations, using sequential importance sampling, with $\epsilon = 10^{-3}$.

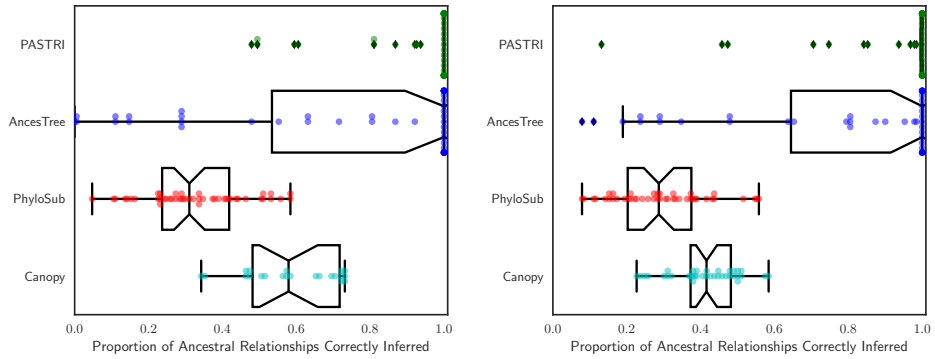
B.2 Additional Simulation Results from Section 3.1



(a) 3 nodes

(b) 4 nodes

Figure 2: **Runtime.**



(a) 3 nodes

(b) 4 nodes

Figure 3: **Accuracy in recovering correct ancestral relationships.**

B.3 Real Data Results

PASTRI was run as described in Supplementary Section B.1. The data log-likelihood was calculated as the logarithm of Equation 7. The PhyloSub model log-likelihood additionally allows for reads to contain sequencing error. Thus, they have a probability μ^r of observing a reference allele from a variant population and probability μ^v of observing a reference allele from the reference population. Thus, instead of the Binomial term, we have

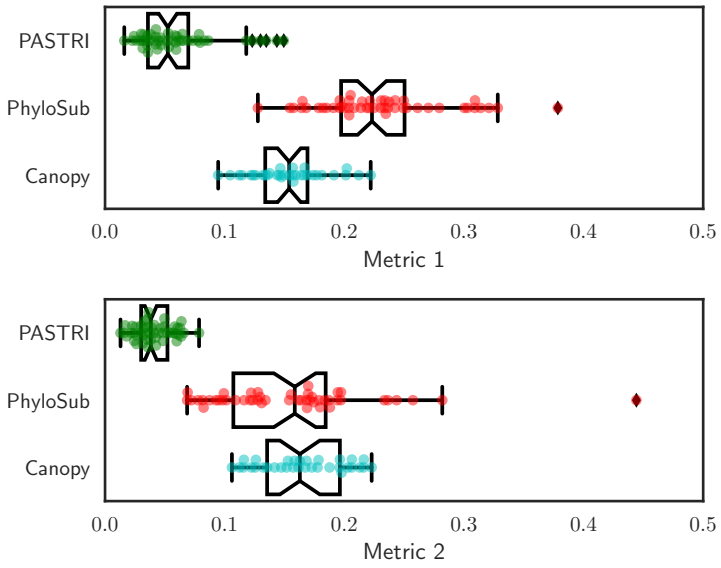
$$\text{Binomial}\left(a_{j,p} \mid \frac{1}{2} f_{c_j,p}, d_{j,p}\right),$$

we have

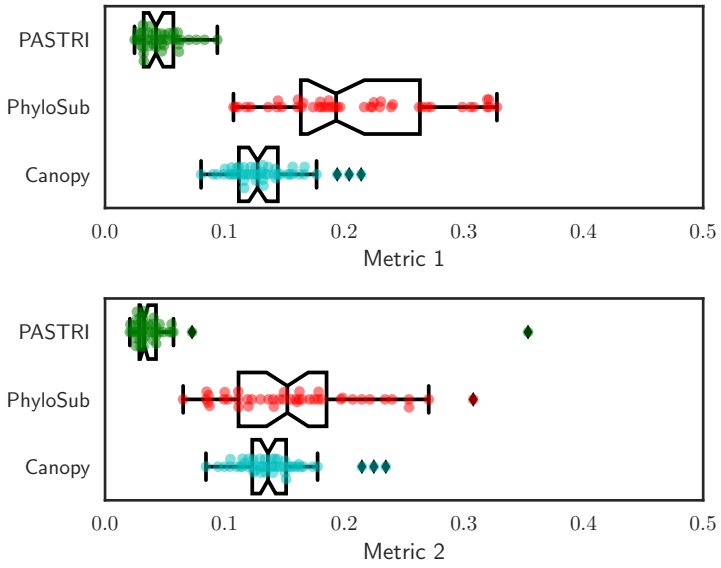
$$\text{Binomial}\left(a_{j,p} \mid f_{c_j,p}(1-\mu^v) + (1-f_{c_j,p})(1-\mu^r), d_{j,p}\right).$$

Given the 0.001 sequencing error for illumina data, we used $\mu_r = 0.999$ and $\mu_v = 0.499$.

Figure 5 contains the results of running AncesTree on the same data. SciClone found 8 clusters of mutations. AncesTree was not able to construct a tree containing all 8 clusters. The largest tree it was able to reconstruct contained 6 clusters, and 18/20 mutations.



(a) 3 nodes



(b) 4 nodes

Figure 4: **Accuracy in recovering true cluster frequencies.**

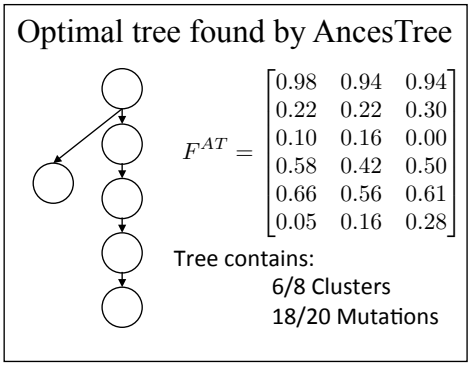


Figure 5: **AncesTree results on CLL patient 5.** AncesTree was not able to construct a tree containing all 8 clusters. The largest tree it was able to reconstruct contained 6 clusters, and 18/20 mutations.