Generalized Pruning Algorithm

Seong-Hwan Jun,[1] Hassan Nasif,[1,2] Frederick Matsen,[1,2,3]

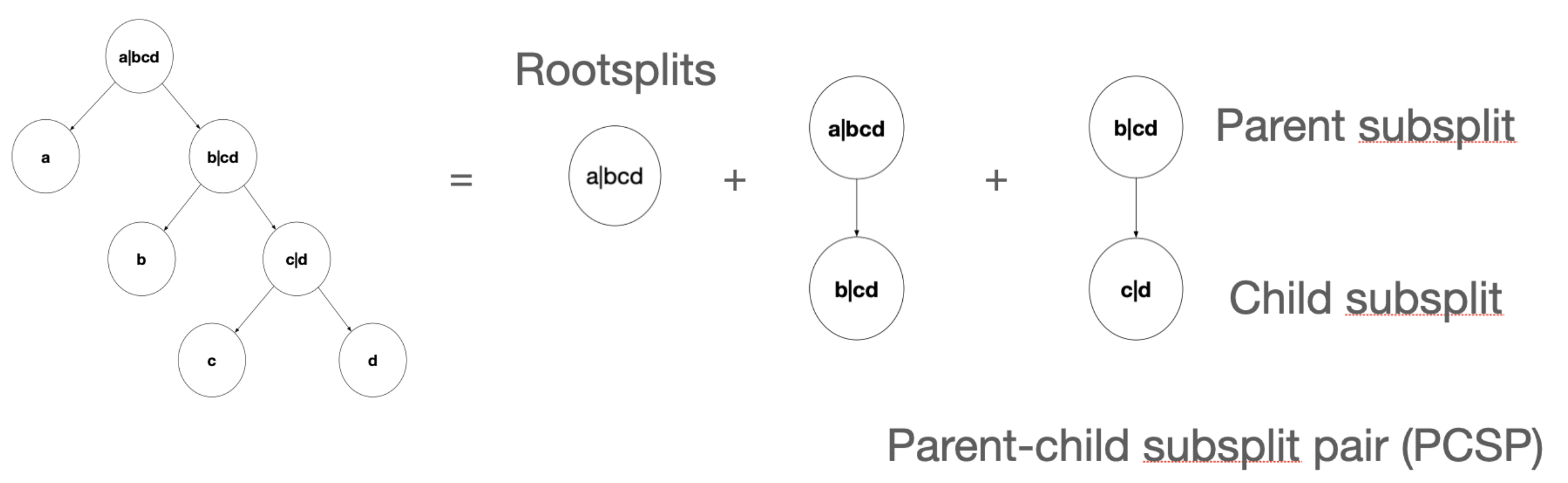
[1]Computational Biology, Public Health Science Division, Fred Hutchinson Cancer Research Center
[2]Department of Statistics, University of Washington
[3]Department of Genome Sciences, University of Washington

Overview

- Goal: Develop fast and accurate algorithm to estimate the distribution of rooted phylogenetic trees.
- Parameterize distribution of the trees using *subsplits* introduced in Zhang and Matsen (2018). This lets us consider marginalization over the trees.
- Perform dynamic programming on a directed acyclic graph (DAG) with subsplits as nodes.
- Messages are passed up and down the DAG to compute the *partial likelihood vectors*, which represent marginalization of the states at the internal nodes and the trees.
- This algorithm generalizes Felsenstein pruning algorithm and it is inspired by two-pass Felsenstein algorithm for efficient gradient computation proposed in Ji *et al.* (2020).

Subsplits

Subsplit representation of a tree:

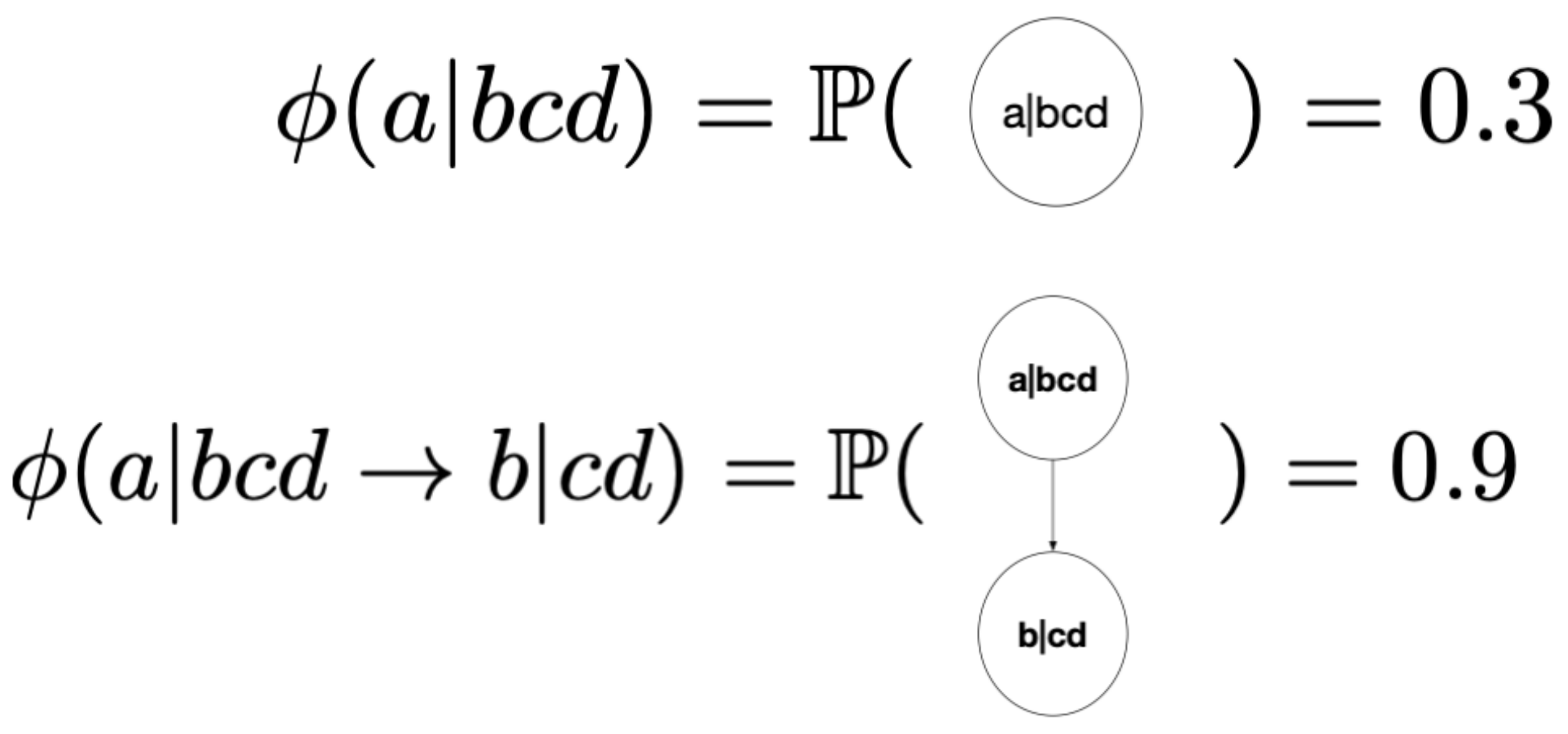


Rootsplits split all of the taxa set. Child subsplits one of the two clades of the parent.

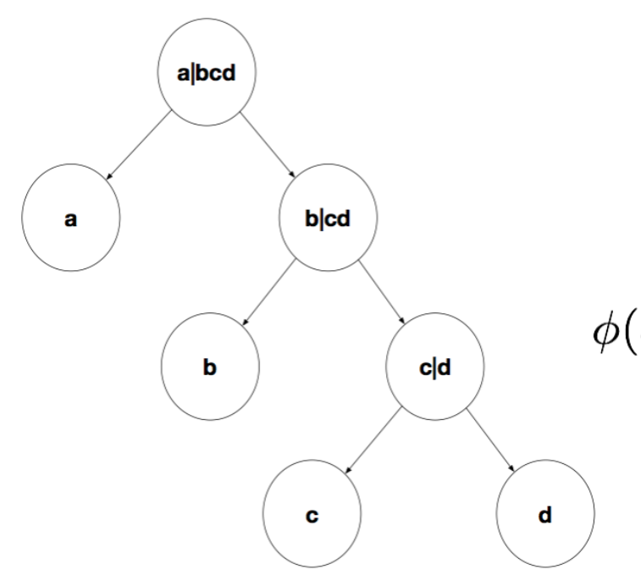
Subsplit support can be specified via conditional probability table. We use ϕ to denote the entire collection of subsplit parameters and $\phi(s)$ or $\phi(t \to s)$ to denote the probability for a rootsplit or a PCSP.

Child / Parent	$a\|bcd$	$ab\|cd$
$abcd$	0.3	0.7

Child / Parent	$b\|cd$	$bc\|d$
$a\|bcd$	0.9	0.1


$\phi(a|bcd) = \mathbb{P}($ a|bcd $) = 0.3$
$\phi(a|bcd \to b|cd) = \mathbb{P}($ a|bcd b|cd $) = 0.9$

Note: the vertical bar denotes split of clades and not conditioning.



$$\mathbb{P}_{\phi}(\tau) = \prod_{t \to s \in \tau} \phi(t \to s)$$

$$\phi(a|bcd) \times \phi(a|bcd \to b|cd) = 0.3 \times 0.9 = 0.27$$

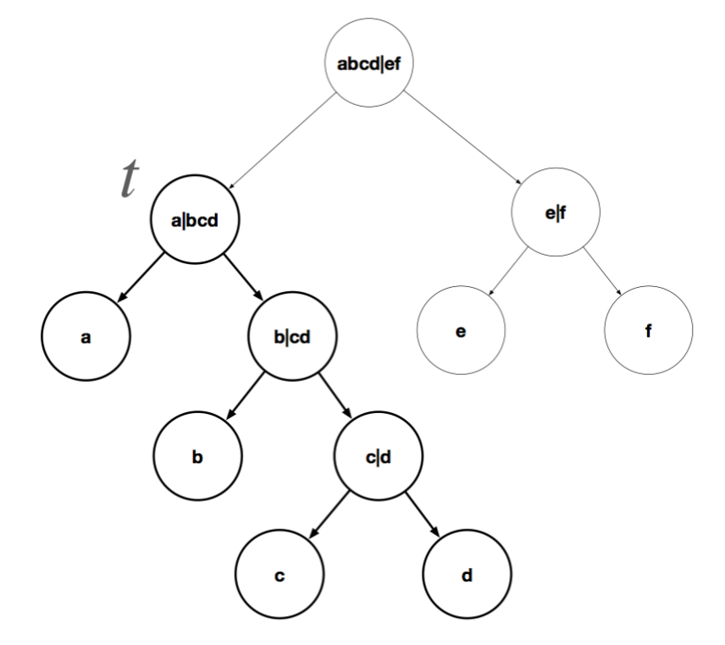
Child / Parent	$a\|bcd$	$ab\|cd$
$abcd$	0.3	0.7

Child / Parent	$b\|cd$	$bc\|d$
$a\|bcd$	0.9	0.1

Subsplit probabilities are used for evaluating the probability of a tree. For a given tree, multiply the probabilities of each rootsplit and parent-child subsplits that appear in the tree.

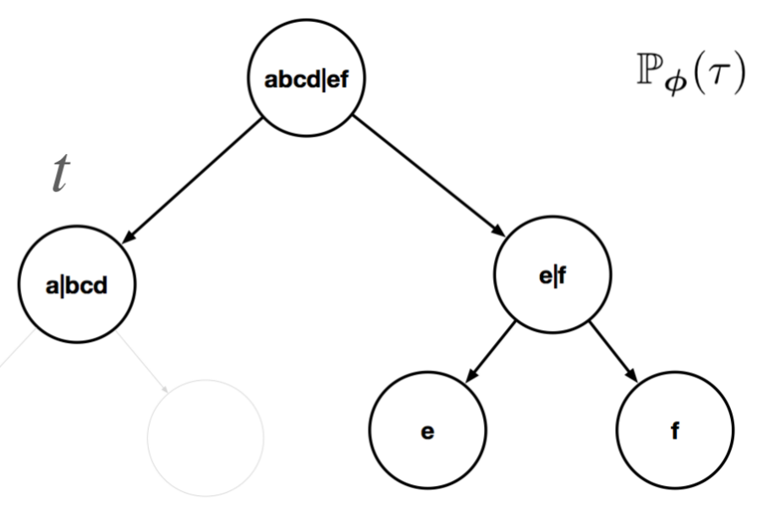
Justification for Subsplits

Dynamic evaluation of tree probability makes it well suited for dynamic programming.



$$\mathbb{P}_{\phi}(\tau_t|t) = \prod_{s \to u \in \tau_t} \phi(s \to u)$$

Child / Parent	$b\|cd$	$bc\|d$
$a\|bcd$	0.9	0.1



$$\mathbb{P}_{\phi}(\tau) = \phi(abcd|ef) \times \phi(abcd|ef \to a|bcd) = 0.8 \times 0.3 = 0.24$$

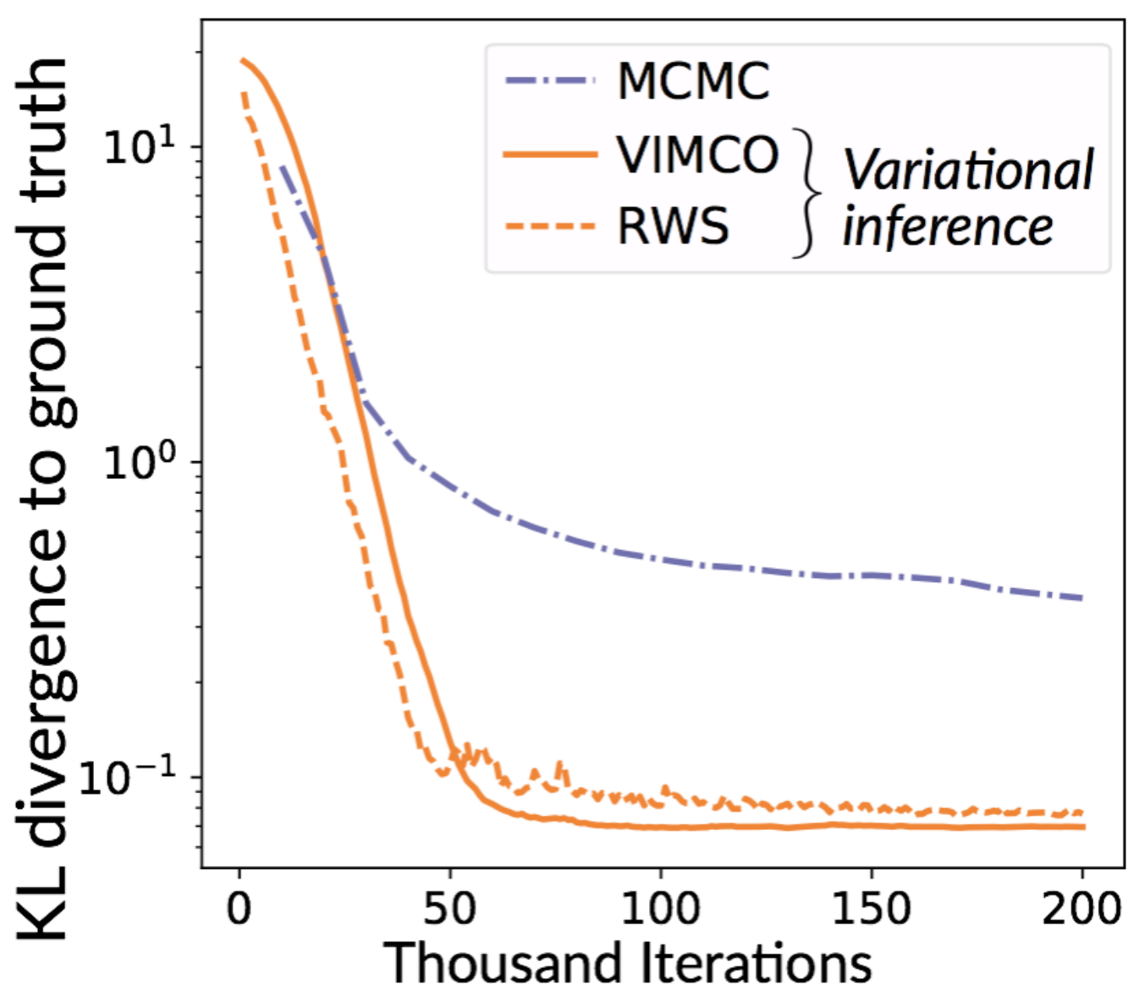
$$\mathbb{P}_{\phi}(\tau) = \prod_{t \to s \in \tau} \phi(t \to s)$$

Child / Parent	$abcd\|ef$	$aef\|bcd$
$abcdef$	0.8	0.2

Child / Parent	$ab\|cd$	$a\|bcd$	$ad\|bc$
$abcd\|ef$	0.5	0.3	0.2

Left: Evaluating probability of a subtree, **Right**: Evaluating probability of a partially specified tree where the subtree below node $a|bcd$ is unspecified.

Subsplits are used to parameterize variational distribution over trees in Zhang and Matsen (2019). VIMCO and RWS are inference algorithms to minimize the KL divergence between variational distribution and the posterior distribution. Variational approximation achieved lower KL divergence compared to 2 million iterations of MCMC run using Mr. Bayes. This suggests that subsplit representation sufficiently captures the trees with high posterior mass.



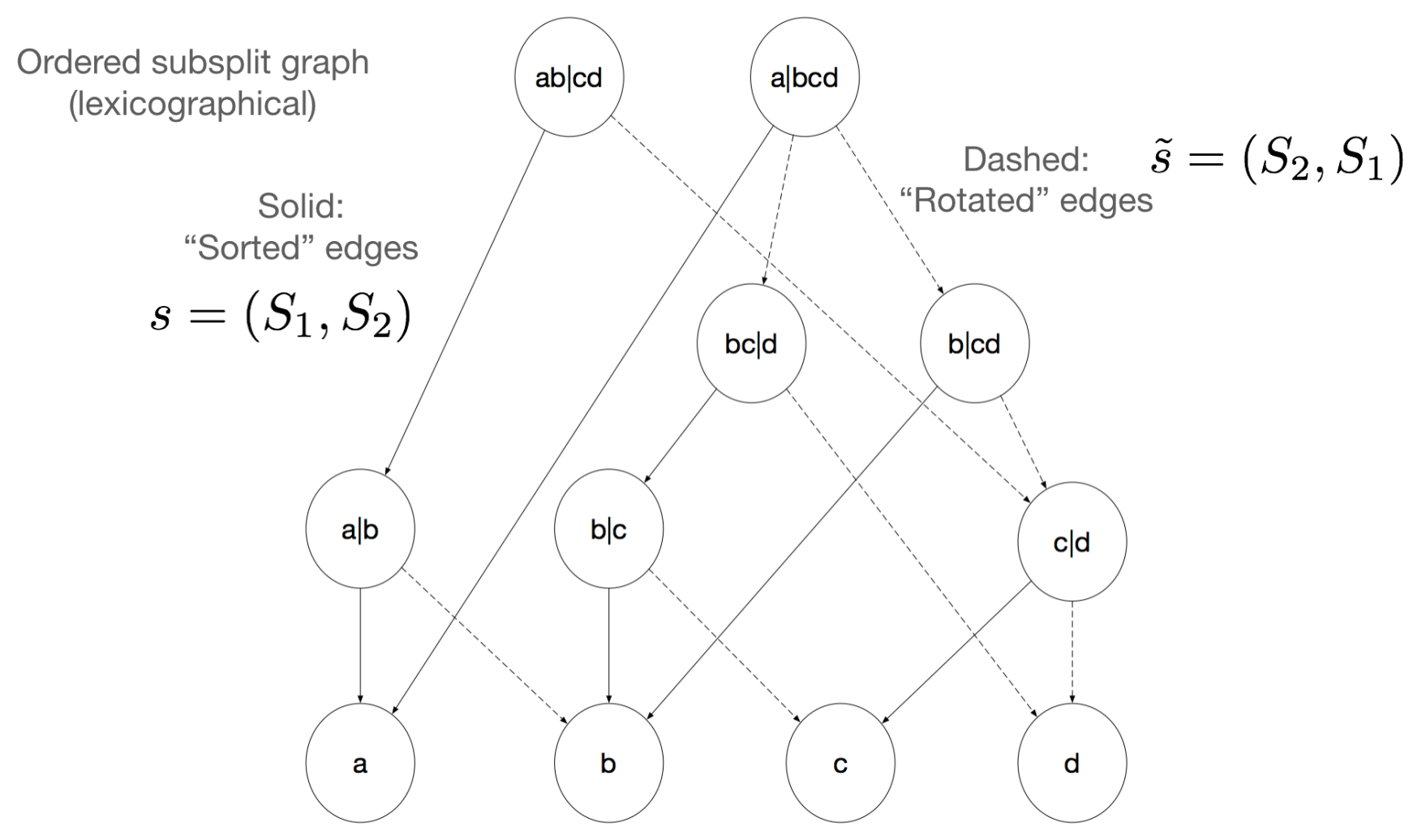
Notation and Review of Felsenstein Pruning Algorithm

- We will focus on nucleotide bases $\Sigma = \{A, C, G, T\}$.
- $\mathbf{P}(t \to s)$ denotes transition probability matrix for branch corresponding to $t \to s$.
- $\mathbf{Y}$: Observed sequences over N taxa with alignment length of M. $\mathbf{Y}^m \in \Sigma^N$ represents single-site observation.
- For a tree rooted at node s, we denote the observed sequences at the tips of the tree by $\mathbf{Y}_{\lfloor s \rfloor}$.
- We denote the observed sequences "above" s by $\mathbf{Y}_{\lceil s \rceil} = \mathbf{Y} \setminus \mathbf{Y}_{\lfloor s \rfloor}$.
- $\mathcal{T}_{leaf}(s)$: set of all trees with s as the root.
- $\mathcal{T}_{leaf}(s)$: for $s = (S_1, S_2)$, set of all partially-specified trees where S_2 is unspecified.
- $(\mathbf{p}(\tau_s))_i^m = \mathbb{P}(\mathbf{Y}^m_{\lfloor s \rfloor}|Y_s^m = i)$: likelihood of the sequences "below" node s, conditioned on the latent sequence at node s.
- $(\mathbf{r}(\tau_s))_i^m = \mathbb{P}(\mathbf{Y}^m_{\lceil s \rceil}, Y_t^m = i)$: joint likelihood of the sequences "above" node s and the latent state at the parent node t.
- To compute the likelihood over a tree: $\mathbb{P}(\mathbf{Y}^m|\tau) = \pi'\mathbf{p}$ where π denotes the stationary distribution.
- $\mathbf{p}, \mathbf{r}$ are referred to as partial likelihood vectors (PLV).

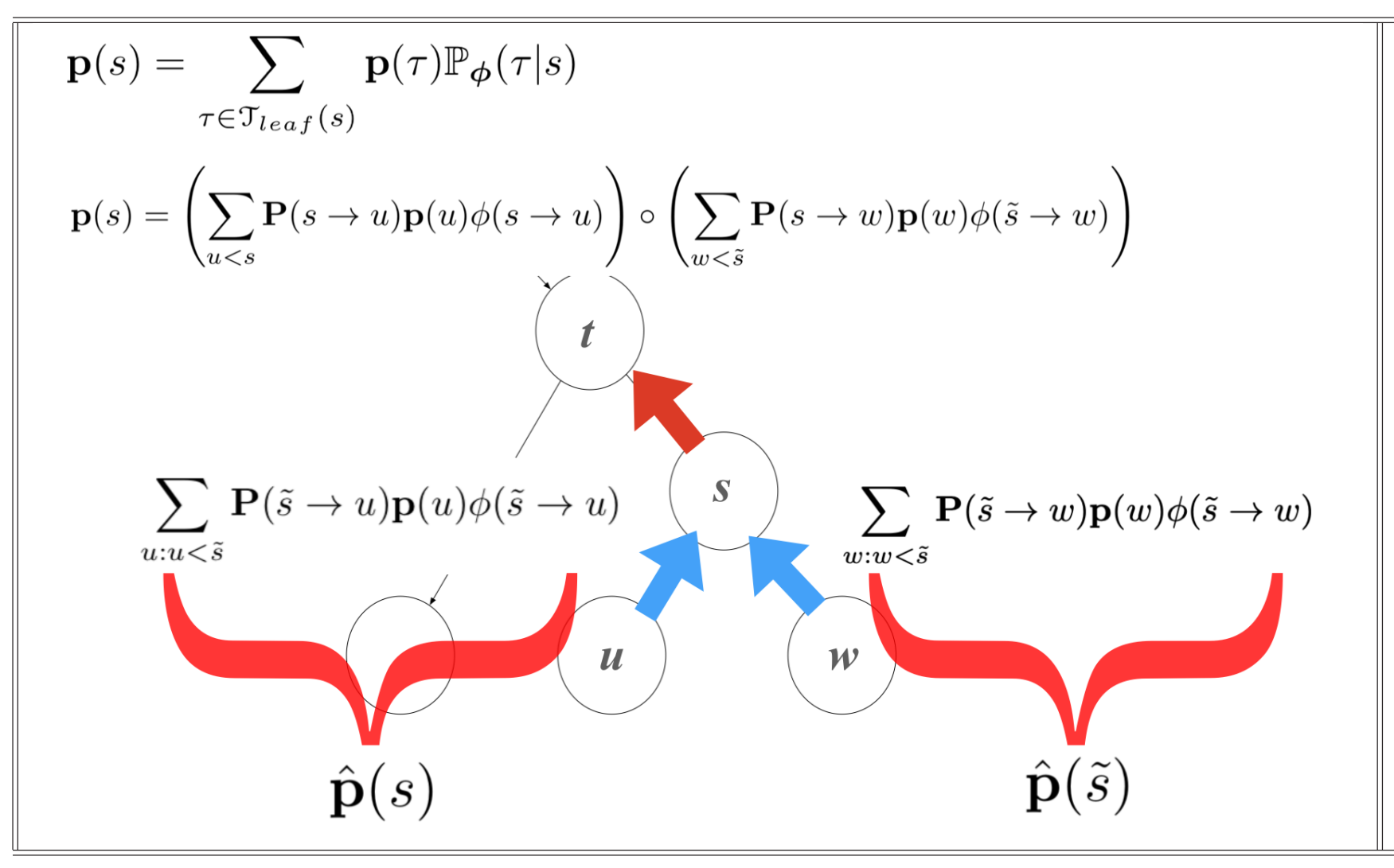
Generalized Pruning Algorithm

Ordered subsplit graph:

- The nodes are ordered subsplits, $s = (S_1, S_2)$ based on lexicographical ordering on S_1, S_2.
- Rotating a subsplit yields $s = (S_2, S_1)$.
- A *sorted* edge connects a parent s to any child u where u is a subsplit of S_1.
- A *rotated* edge connects a parent s to any child u that splits S_2.

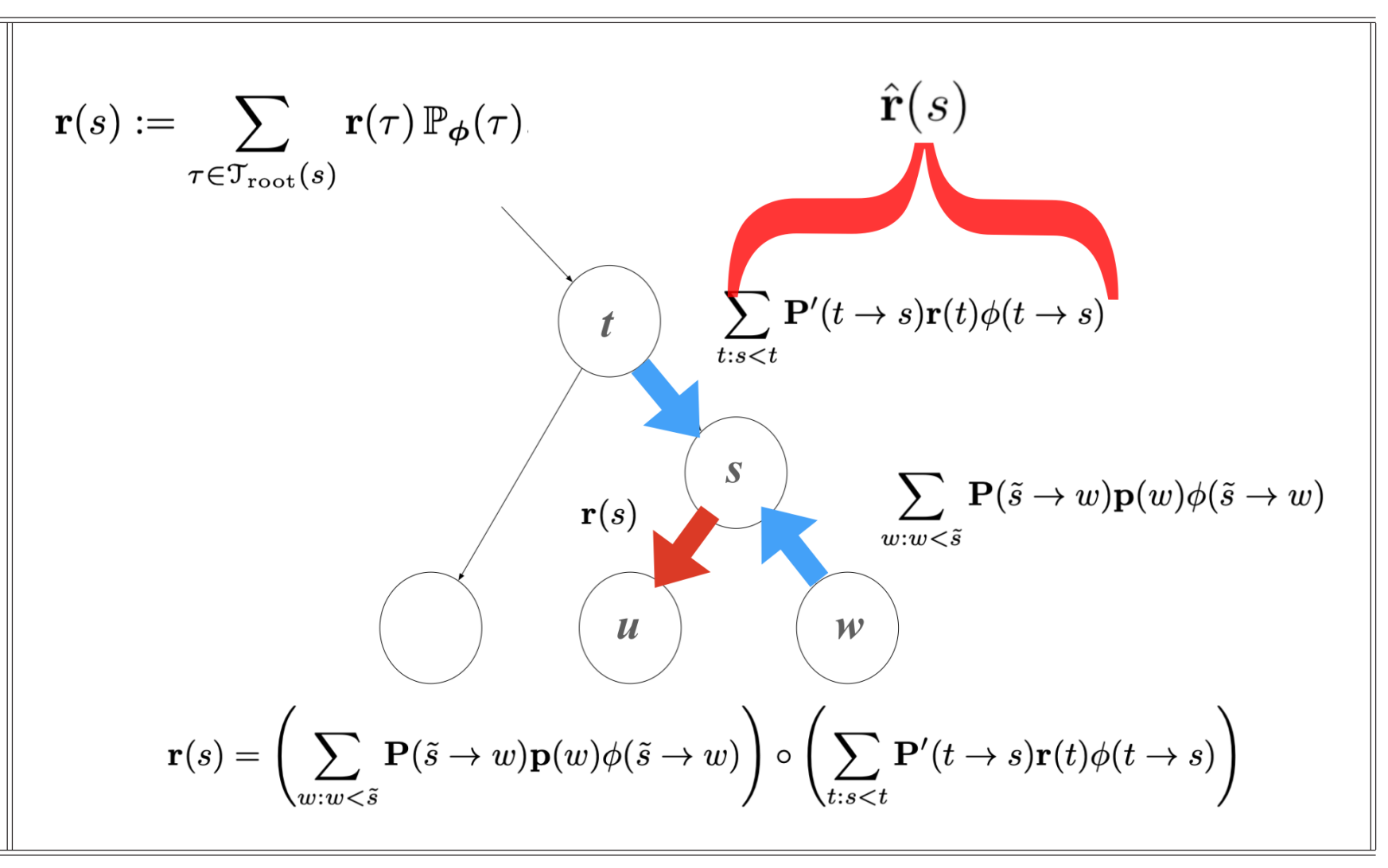


This graph encapsulates all of the trees that are in the tree support determined by the conditional probability table.


t
s
u
w
$\sum_{u:u<\tilde{s}} \mathbf{P}(\tilde{s} \to u)\mathbf{p}(u)\phi(\tilde{s} \to u)$
$\sum_{w:w<\tilde{s}} \mathbf{P}(\tilde{s} \to w)\mathbf{p}(w)\phi(\tilde{s} \to w)$
$\hat{\mathbf{p}}(s)$
$\hat{\mathbf{p}}(\tilde{s})$

$$\mathbf{p}(s) = \sum_{\tau \in \mathcal{T}_{leaf}(s)} \mathbf{p}(\tau)\mathbb{P}_{\phi}(\tau|s)$$

$$\mathbf{p}(s) = \left(\sum_{u<s} \mathbf{P}(s \to u)\mathbf{p}(u)\phi(s \to u)\right) \circ \left(\sum_{w<\tilde{s}} \mathbf{P}(s \to w)\mathbf{p}(w)\phi(\tilde{s} \to w)\right)$$


$\hat{\mathbf{r}}(s)$
$\sum_{t:s<t} \mathbf{P}'(t \to s)\mathbf{r}(t)\phi(t \to s)$
t
s
$\mathbf{r}(s)$
$\sum_{w:w<\tilde{s}} \mathbf{P}(\tilde{s} \to w)\mathbf{p}(w)\phi(\tilde{s} \to w)$
u
w

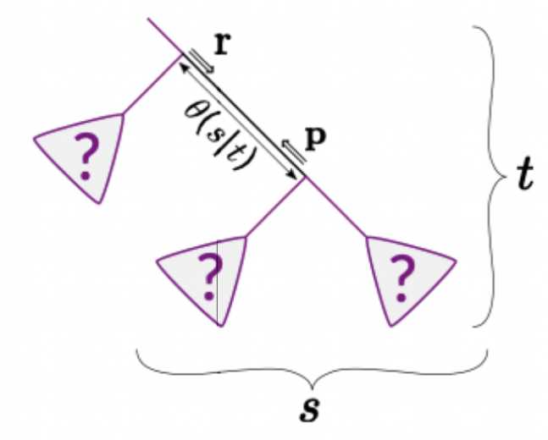
$$\mathbf{r}(s) := \sum_{\tau \in \mathcal{T}_{\text{root}}(s)} \mathbf{r}(\tau)\,\mathbb{P}_{\phi}(\tau)$$

$$\mathbf{r}(s) = \left(\sum_{w:w<\tilde{s}} \mathbf{P}(\tilde{s} \to w)\mathbf{p}(w)\phi(\tilde{s} \to w)\right) \circ \left(\sum_{t:s<t} \mathbf{P}'(t \to s)\mathbf{r}(t)\phi(t \to s)\right)$$

Left: p-PLV defined in terms of likelihood of subtrees rooted at s, weighted by probability of such trees conditioned on the value of subsplit at s. The definition for p-PLV can be expressed recursively in terms of children subsplits $u < s$ and $w < \tilde{s}$. **Right**: r-PLV is defined in terms of partially unspecified trees. It is expressed recursively in terms of p-PLV of one of its child and r-PLV of its parents. Note that $\mathbf{r}(s) \neq \mathbf{r}(\tilde{s})$. We perform and store intermediate computations $\hat{\mathbf{p}}(s), \hat{\mathbf{p}}(\tilde{s}), \hat{\mathbf{r}}(s)$.

Estimation

Our algorithm computes exact marginal for a single site m. We can combine the single-site marginals to form composite-likelihood function, which we optimize with respect to the branch length parameters θ for each PCSP $t \to s$ using the PLVs. This is fast since PLVs are already computed and admits efficient function evaluation in $O(M)$ time.



$$\ell^m(t \to s) = \phi(t \to s)\mathbf{r}'(t)\mathbf{P}(\theta(t \to s))\mathbf{p}(s) = \sum_{\tau : t \to s \in \tau} \mathbb{P}(\mathbf{Y}^m|\tau, \boldsymbol{\theta})\mathbb{P}_{\phi}(\tau)$$

Maximize per PCSP composite likelihood

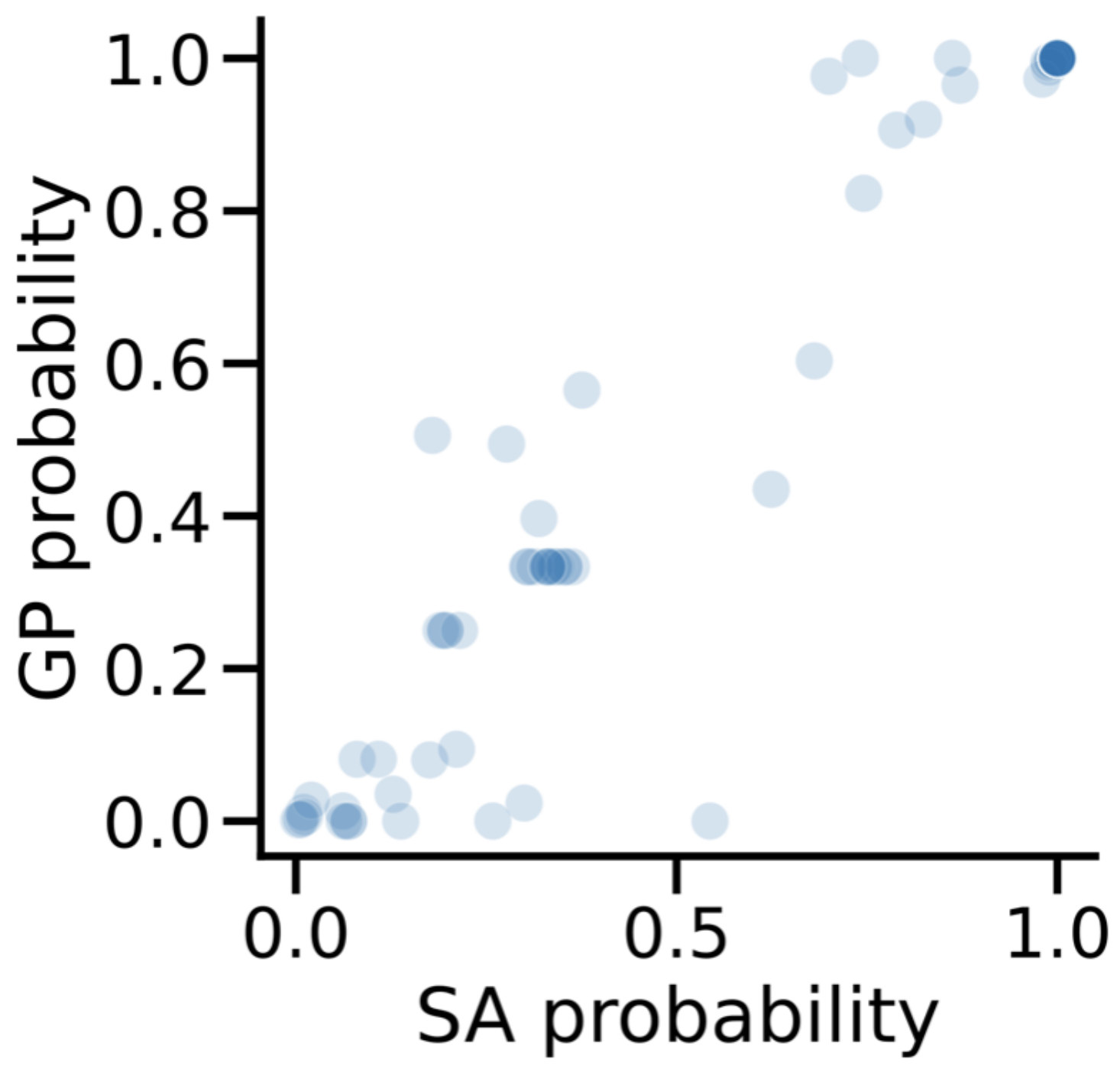
$$\ell(t \to s) = \prod_{m=1}^{M} \ell^m(t \to s)$$

For estimating subsplit parameters, we initialize ϕ such that it defines $\mathbb{P}_{\phi}$ to be a uniform prior over the rooted phylogenetic trees. Then, we fix the branch lengths and update subsplit parameters using Bayes rule.

$$\phi_{post}(t \to s) = \mathbb{P}(t \to s|\mathbf{Y}) = \frac{\ell(t \to s)\phi_{prior}(t \to s)}{\sum_{s' \in \mathcal{S}_t} \ell(t \to s')\phi_{prior}(t \to s')}$$

Preliminary Results

We performed experiments on DS7 data from Yang and Yoder (2003). The subsplit parameters estimated using GP algorithm is compared against the method proposed in Zhang and Matsen (2018). The estimated values of two methods show strong concordance.



References

Xiang Ji, Zhenyu Zhang, Andrew Holbrook, Akihiko Nishimura, Guy Baele, Andrew Rambaut, Philippe Lemey, and Marc A Suchard. Gradients do grow on trees: a linear-time o(n)-dimensional gradient for statistical phylogenetics. *Mol. Biol. Evol.*, May 2020.

Ziheng Yang and Anne D Yoder. Comparison of likelihood and bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic biology*, 52(5):705–716, 2003.

Cheng Zhang and Frederick A Matsen, IV. Generalizing tree probability estimation via bayesian networks. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1449–1458. Curran Associates, Inc., 2018.

Cheng Zhang and Frederick A Matsen, IV. Variational bayesian phylogenetic inference. In *International Conference on Learning Representations (ICLR)*, 2019.