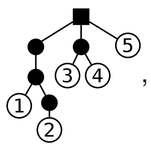
How many ways can we explain the same data?

Mathias C. Cronjäger (cronjager@stats.ox.ac.uk),
D. Khurana (Oxford), P. Jenkins (Warwick), J. Hein (Oxford)

Motivation: how complex are probability recursions?

Given a set of aligned sequences, e.g.

Sequence 1: … T … G … A … A …
Sequence 2: … T … G … A … G …
Sequence 3: … A … A … T … A …
Sequence 4: … A … A … T … A …
Sequence 5: … A … A … A … A …

represented as  ,

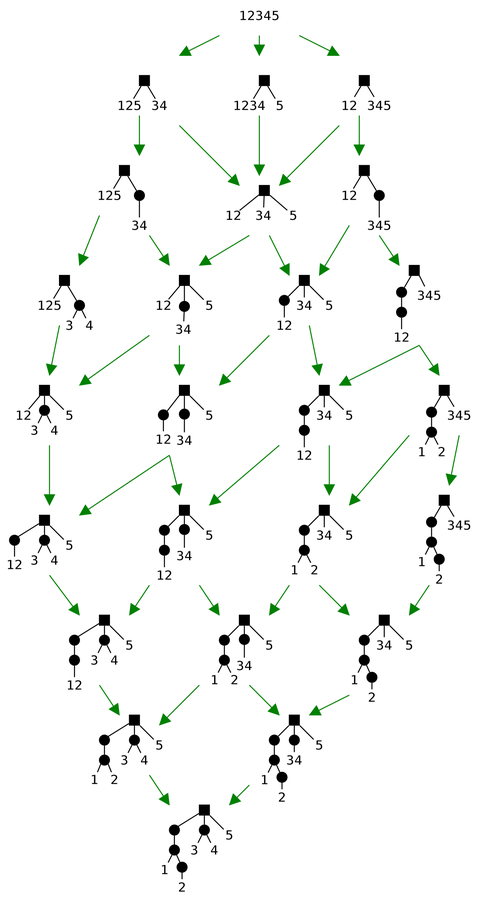
the probability of having evolved from some initial sequence–here $\ldots A \ldots A \ldots A \ldots A \ldots$–may be expressed:

$$\mathbb{P}\left(\text{tree } 12345\right) = \sum_{x \in \text{Histories}\left(\text{tree } 12345\right)} \mathbb{P}(x)$$

which in turn may be expressed by conditioning on the most recent event:

$$\mathbb{P}\left(\text{tree } 12345\right) = \frac{1}{\theta+4}\mathbb{P}\left(\text{tree } 12345\right) + \frac{\theta}{\theta+4}\frac{1}{5}\mathbb{P}\left(\text{tree } 12345\right).$$

If we apply this conditioning-trick recursively, computing $\mathbb{P}\left(\text{tree } 12345\right)$ reduces to computing a weighted sum over all paths from "12345" to "tree 12345" in the graph below:



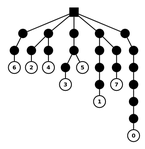
We may note that

- number of nodes = number of distinct terms in $\mathbb{P}$-recursion,
- number of paths "$12345 \to \ldots \to$ tree 12345" = number of execution paths when evaluating $\mathbb{P}\left(\text{tree } 12345\right)$ via tail-recursion (without memoization).

We refer to the nodes in the above graph as **ancestral states**, and the paths "$12345 \to \ldots \to$ tree 12345" as **ancestral histories**.

Central problem: counting quickly

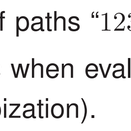
The number of ancestral states and histories grows very quickly. Whereas tree 12345 has 71 histories and 25 ancestral states (including itself), the dataset



for example already has 89,325,103,544,240,200 histories and 12,356 ancestral states. Since *we do not know the graph a-priori* our main aim is to **count the number of ancestral states and/or histories just by considering the initial data** (i.e. without having to determine the entire ancestral graph).

How to count (abstractly)

We encode non-plane trees as nested systems of sets, e.g.

 $= \{\{\{1,\{2\}\}\},\{3,4\},5\}.$

Given a tree $T = \{T_1, \ldots, T_r\}$, we may determine the number of ancestral histories and ancestral states–denoted $h(T)$ and $a(T)$ respectively– as follows:

- If T is a leaf (i.e. has no children), $a(T) = h(T) = 1$ holds.
- If $T = \{T'\}$ (i.e. T has root degree 1), $a(T) = 1 + a(T')$ and $h(T) = h(T')$ holds.
- If $T = \{T_1, \ldots T_r\}$ holds for $r > 1$, then:

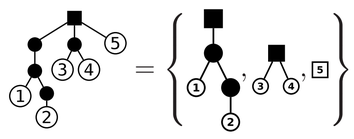
$$a'(T) = 1 + \sum_{S \subseteq [k]} \left(\sum_{\substack{\lambda \vdash |S| \\ \alpha_1 = 0}} \frac{|S|!}{\prod_i \lambda_i! \prod_i \alpha_i!} \right) \prod_{i \notin S} a'(T_i)$$

$$h(T) = \sum_{\substack{S \subsetneq [r] \\ 1 \in S}} h(\{T_i \mid i \in S\}) h(\{T_i \mid i \notin S\}) \binom{(\sum_{i=1}^{r} k_i) - 2}{(\sum_{i \in S} k_i) - 1}$$

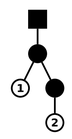
Whereby $a'(T') = a(T') + \mathbb{1}(\deg(\text{root}(T')) > 1)$ and k_i denotes the number of nodes in sub-tree T_i.

How to count – a simple example

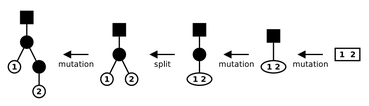
1. We decompose our gene-tree into sub-trees:



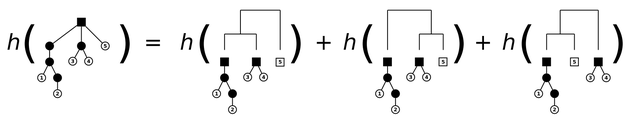
2. We consider the the problem for each sub-tree separately:

T_{sub}	$h(T_{\text{sub}})$	$a'(T_{\text{sub}})$
	1	5
	1	3
5	1	1

Ancestors of T_{sub} (verification)



Computing $h\left(\text{tree } 12345\right)$: we consider all three ways of ordering mergers immediately below the root:



Each of the right hand terms can then be computed with relative ease:

$$h(T_1)\left(h(T_2)h(T_3)\binom{3+1-2}{3-1}\right)\binom{5+3+1-2}{5-1} = 35$$

$$h(T_2)\left(h(T_1)h(T_3)\binom{5+1-2}{5-1}\right)\binom{5+3+1-2}{3-1} = 21$$

$$h(T_3)\left(h(T_1)h(T_2)\binom{5+3-2}{5-1}\right)\binom{5+3+1-2}{1-1} = 15$$

Which implies that $h\left(\text{tree } 12345\right) = 35 + 21 + 15 = 71$ holds.

Computing $a\left(\text{tree } 12345\right)$: Here, we first compute $a'\left(\text{tree } 12345\right)$ using our recursion:

$$a'\left(\text{tree } 12345\right) = 1 + \underbrace{5 \cdot 3 \cdot 1}_{S=\emptyset} + \underbrace{5}_{S=\{2,3\}} + \underbrace{3}_{S=\{1,3\}} + \underbrace{1}_{S=\{1,2\}} + \underbrace{1}_{S=\{1,2,3\}}$$

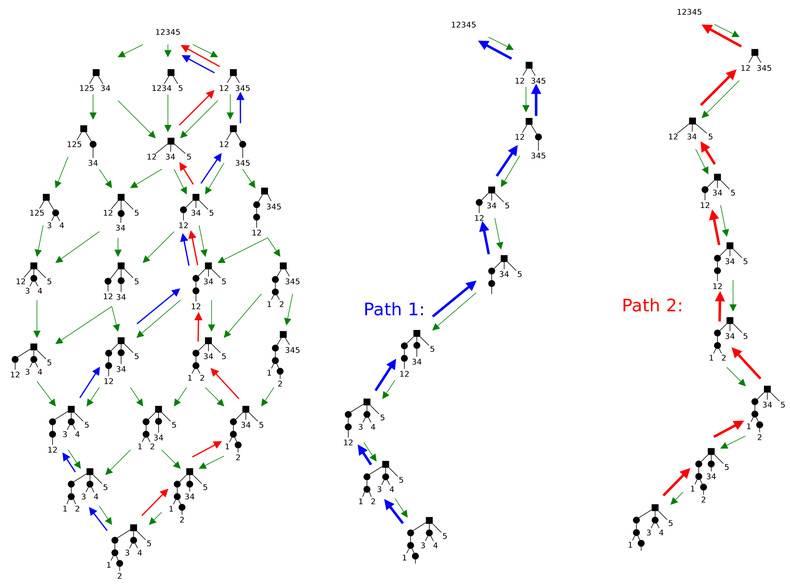
$$= 26$$

It therefore follows that $a\left(\text{tree } 12345\right) = a'\left(\text{tree } 12345\right) - 1 = 25$

Application: combinatorial importance sampler

We can approximate probabilities of aligned sequences–e.g. $\mathbb{P}\left(\text{tree } 12345\right)$–by sampling ancestral histories $X_1, \ldots, X_N \overset{iid}{\sim} \mathbb{Q} \ll \mathbb{P}$ and relying on the following approximation:

$$\mathbb{P}\left(\text{tree } 12345\right) = \sum_{x \in \text{H}\left(\text{tree } 12345\right)} \frac{\mathbb{P}(x)}{\mathbb{Q}(x)}\mathbb{Q}(x) = \mathbb{E}_{X \sim \mathbb{Q}}\left[\frac{\mathbb{P}(X)}{\mathbb{Q}(X)}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{P}(X_i)}{\mathbb{Q}(X_i)}$$

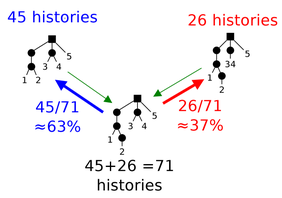


e.g. $\mathbb{P}\left(\text{tree } 12345\right) \approx \frac{1}{2}\left(\frac{\mathbb{P}(\text{Path 1})}{\mathbb{Q}(\text{Path 1})} + \frac{\mathbb{P}(\text{Path 2})}{\mathbb{Q}(\text{Path 2})}\right)$

For this approach to work effectively, $\mathbb{Q}$ should satisfy:

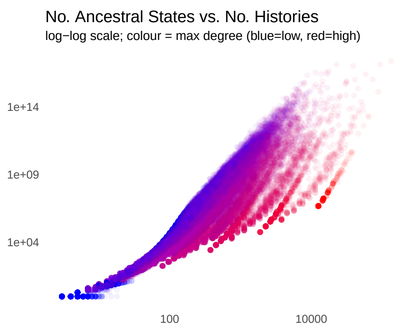
1. $\mathbb{Q}$ must approximate $\mathbb{P}$ well on the space of histories;
2. sampling $X_i \sim \mathbb{Q}$ should be fast;
3. computing $\mathbb{Q}(X_i)$ should be fast.

Our scheme *uniformly samples* elements from the space of ancestral histories by generating them "from the bottom up". To sample an ancestral history of tree 1234, there are for example two "first steps" to pick from:

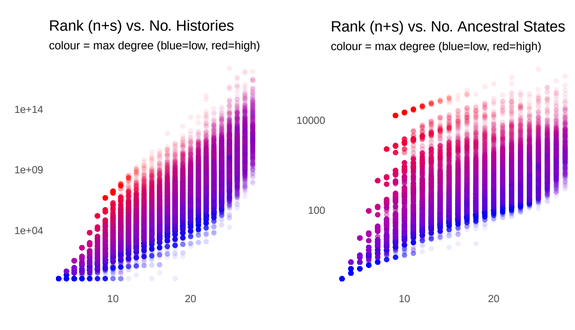


Since tree 12345 has 45 ancestral histories whereas tree 1234 has only 26. Our sampler will therefore generate paths that "go left" as a first step with probability $45/71 \approx 63\%$.

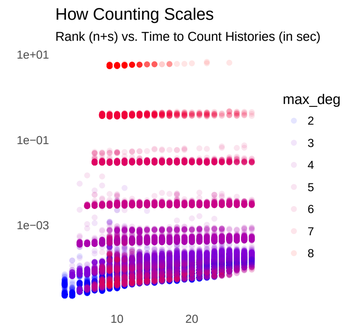
Computational experiments



Counts for 70,000 simulated datasets (generated using ms); 500 per choice of $n = 2, \ldots, 8$ sequences and $s = 1, \ldots, 20$ segregating sites



A natural measure of how complex a dataset is is the number of evolutionary events (n + s -1). We see that some degree of proportionality exists (on a log-scale) and note that the maximal degree of the gene tree seems very important as well.



The number of events in the histories of a dataset is of little importance relative to the maximal degree of the gene tree

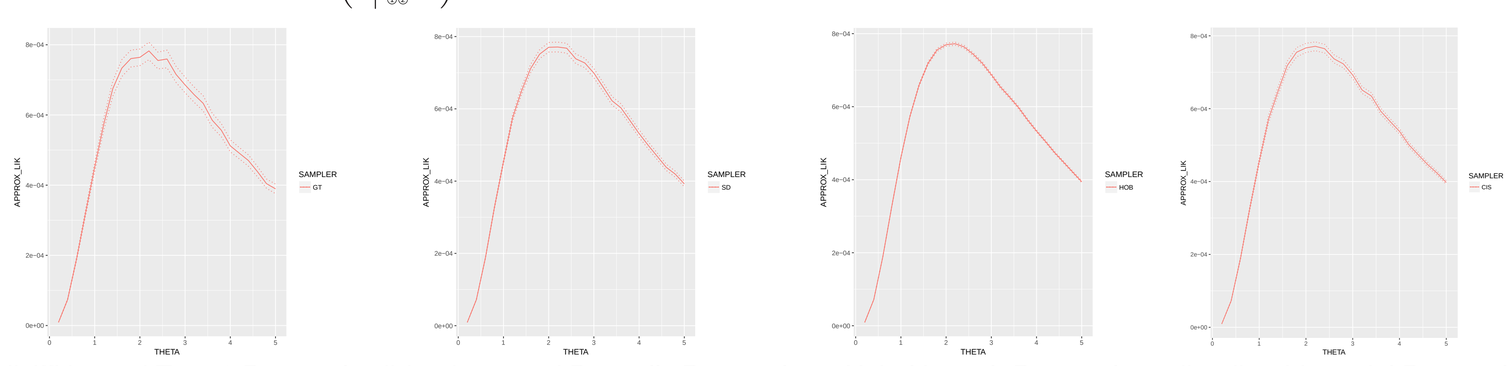
Median of $h(T)$ for sim. data w. n sequences and s seg. sites; $N = 500$

	n = 2	n = 3	n = 4	n = 5	n = 6	n = 7	n = 8
s = 1	1	4	30	45	630	12,600	340,200
s = 2	1	5	13	90	1,008	10,692	151,200
s = 3	3	6	20	108	1,190	14,640	226,800
s = 4	4	10	45	215	1,840	21,426	328,320
s = 5	10	18	70	460	3,024	35,736	452,790
s = 6	15	35	161	770	6,213	56,240	870,396
s = 7	35	56	252	1,270	13,860	124,614	1,804,374
s = 8	56	126	588	3,927	30,800	308,880	4,616,892
s = 9	84	240	1,386	9,323	75,933	796,887	10,666,656
s = 10	210	462	3,465	19,552	181,818	2,001,285	24,094,581
s = 11	330	924	6,006	43,056	420,420	3,474,900	54,600,260
s = 12	792	1,716	12,012	102,668	1,345,344	11,401,968	138,116,160
s = 13	1,287	3,536	30,030	224,224	3,063,060	27,748,539	323,188,632
s = 14	3,003	9,783	80,080	556,920	5,250,960	76,246,688	905,123,520
s = 15	5,005	17,836	144,330	1,435,752	18,354,152	211,629,600	2,930,581,584
s = 16	11,440	48,620	403,920	4,576,910	36,555,525	553,655,592	8,713,086,690
s = 17	19,448	97,240	949,620	9,922,560	127,588,230	1,462,912,220	22,814,964,172
s = 18	43,758	235,467	2,771,340	24,690,120	298,217,920	3,088,978,200	72,231,873,336
s = 19	75,582	503,880	6,342,105	58,326,048	995,255,701	12,382,746,348	139,048,420,500
s = 20	125,970	1,175,720	10,893,498	137,287,920	2,759,218,000	32,668,284,600	758,910,438,000

Median of $a(T)$ for sim. data w. n sequences and s seg. sites; $N = 500$

	n = 2	n = 3	n = 4	n = 5	n = 6	n = 7	n = 8
s = 1	3	7	24	96	115	528	2,640
s = 2	5	7	13	38	134	238	1,126
s = 3	7	10	17	35	81	260	931
s = 4	9	13	21	37	92	258	777
s = 5	13	17	26	45	105	226	616
s = 6	16	21	31	51	111	226	572
s = 7	21	25	37	61	125	237	671
s = 8	25	31	49	74	139	267	673
s = 9	31	37	57	91	176	311	699
s = 10	36	43	69	101	181	322	733
s = 11	41	50	79	130	217	399	821
s = 12	49	61	97	145	244	445	906
s = 13	55	73	106	163	280	487	1,017
s = 14	64	81	126	193	337	589	1,105
s = 15	71	94	143	232	385	685	1,116
s = 16	81	105	161	261	421	726	1,384
s = 17	89	118	190	289	512	865	1,604
s = 18	100	130	221	328	546	959	1,678
s = 19	109	154	240	373	617	1,000	1,990
s = 20	121	169	261	409	689	1,133	2,175

Approximating $L\left(\theta \,\middle|\, \text{tree } 1234\right)$ with importance sampling (10^4 samples per θ value; plots by Drona Khurana)



Griffiths and Tavaré Proposal | Sthephens and Donnelly Proposal | Hobolth et al. Proposal | Our Combinatorial Proposal

References

Griffiths, R., 1989. Genealogical-tree probabilities in the infinitely-many-site model. J. Math. Biol. 27, 667–680. doi:10.1007/BF00276949; Griffiths, R., Tavaré, S., 1994. Ancestral Inference in Population Genetics. Stat. Sci. 9, 307–319. doi:10.1214/ss/1177010378; Stephens, M., Donnelly, P., 2000. Inference in molecular population genetics. J. R. Stat. Soc. Ser. B. 62, 605–635. doi:10.1111/1467-9868.00254 ; Hobolth, A., Uyenoyama, M., Wiuf, C., 2008. Importance Sampling for the Infinite Sites Model. Stat. Appl. Genet. Mol. Biol. 7, 32. doi:10.2202/1544-6115.1400