Lecture 4: PDA and Yule models

Cuong Van Than

November 07, 2013

The PDA (short for proportional to distinguishable arrangements) and Yule models for generating random trees are commonly used in phylogenetics and population genetics. We study the Yule model in detail in this lecture. You will be asked to prove some results for the PDA model.

Ranked trees. Let T be a rooted phylogenetic tree on X. Recall that the internal vertices of T have no labels. Let $\mathring{V}(T)$ be the set of the internal vertices of T, and let $h\colon \mathring{V}(T) \to \{1, \ldots, |\mathring{V}(T)|\}$ be a bijective map with the property that

$$h(u) \leq h(v) \quad \text{if } u \text{ is the parent of } v.$$

The pair (T, h) is called a ranked phylogenetic tree.

Theorem 1. *Let T be a rooted phylogenetic tree. The number of ranked phylogenetic trees that can be obtained from T is*

$$\frac{|\mathring{V}(T)|!}{\prod_{v \in \mathring{V}(T)} \lambda_v},$$

whre λ_v is the number of internal vertices of T that are also descendants of v (including v itself).

Proof. We prove by induction on $|\mathring{V}(T)|$. The claim cleary holds for trees with zero internal vertices. Assume $|\mathring{V}(T)| = n$ and that the claim holds for trees with up to $n-1$ internal vertices. Let $v_1, \ldots, v_k$ be the children of the root of T, and let $n_1, \ldots, n_k$ be the number of internal vertices of $T(v_1), \ldots, T(v_k)$. Clearly $n_1 + \cdots + n_k = n - 1$, and hence $n_1, \ldots, n_k \leq n-1$. In any rank function for T, the root of T must be assigned rank 1. Therefore, to rank the internal vertices of $T(v_1), \ldots, T(v_k)$, we need to partition $\{2, \ldots, |\mathring{V}(T)|\}$ into (disjoint) subsets of sizes $n_1, \ldots, n_k$. Applying the inductive hypothesis, the number of possible ranked trees that can be obtained from T is

$$\binom{n-1}{n_1, \ldots, n_k} \times \frac{n_1!}{\prod_{v \in \mathring{V}(T_1)} \lambda_v} \cdots \frac{n_k!}{\prod_{v \in \mathring{V}(T_k)} \lambda_v} = \frac{n!}{\prod_{v \in \mathring{V}(T)} \lambda_v},$$

where the last step follows from the fact that $\mathring{V}(T_1), \ldots, \mathring{V}(T_k)$ consist of all the non-root internal vertices of T, and that the root of T has exactly n descendant internal vertices. □

Clearly by ignoring the labels of a ranked tree, one have a "ranked tree shape." On the other hand, one can simply take a tree shape t and then rank its internal vertices as above. We now describe a procedure for associating a ranked tree shape with a rank string s consisting of the ranks of its internal vertices.

Assume that t is binary (and rooted). If t has no internal vertex, then set s as an empty string. If it has exactly one internal vertex, then set s as the string consisting of the rank of the root of t. Otherwise, let s_1 and s_2 be the rank strings for $T(v_1)$ and $T(v_2)$, where v_1 and v_2 are children of the root of t. Then set s as the concatenation of s_1, the rank of the root of t, and s_2 (in this order).

Proposition 2. *Let (t, h) be a ranked tree shape, where t is binary and rooted. Then there are exactly 2^{n-1-c} rank strings for (t, h), where n and c are the numbers of leaves and cherries of t, respectively.*

Proposition 3. *The number of leaf-labelings of a ranked tree shape (t, h), where t is binary and rooted, is $n!/2^c$.*

Yule model. A ranked tree shape in this model is generated as follows. We start with an one-vertex tree shape t, and integer $i = 1$. Choose a random leaf of t with uniform probability (i.e. each leaf of t has equal chance of being selected). The chosen leaf is assigned rank i, and is then split into two new leaves. We increase i by one, and repeat the splitting procedure until t has the desired number of leaves.

Proposition 4. *The probability of a ranked tree shape in the Yule model is*

$$\frac{2^{n-1-c}}{(n-1)!},$$

where n and c are the numbers of leaves and cherries of the tree shape, respectively

We now label the leaves of a ranked tree shape by elements of X. Here, we consider each of $1/n!$ is equally probable.

Theorem 5. *The probability of a ranked tree generated under the Yule model is*

$$\frac{2^{n-1}}{n!(n-1)!}.$$

Coalescent process. A coalescent process on X is a sequence $\pi_0, \ldots, \pi_{n-1}$ of partitions of X, where $|X| = n$, such that π_0 consists of the singleton subsets of X, $\pi_{n-1} = \{X\}$, and for each $1 \leq i \leq n-1$,

$$\pi_i = (\pi_{i-1} \setminus \{A_{i-1}, B_{i-1}\}) \cup \{A_{i-1} \cup B_{i-1}\},$$

for some distinct blocks A_{i-1} and B_{i-1} of π_{i-1}.

It is easy to see that each ranked binary phylogenetic tree on X corresponds to one and exactly one coalescent sequence. Further, the number of coalescent sequences is $n!(n-1)!/2^{n-1}$.

Theorem 6. *The uniform probability distribution on the ranked binary phylogenetic trees is exactly the probability distribution that the Yule model produces.*