Lecture 3: Counting trees

Cuong Van Than

October 31, 2013

In this lecture, we are concerned with the problem of counting phylogenetic trees. Unless otherwise stated, we assume in this lecture trees are phylogenetic trees on a fixed label set X. Recall that the sets of unrooted and rooted phylogenetic trees are denoted by $P(X)$ and $R(X)$, respectively; the counterparts for binary trees are $U(X)$ and $B(X)$. For brevity, denote $n = |X|$.

Counting phylogenetic trees.

Theorem 1. *The number of rooted binary trees on X is $|B(X)| = (2n-3)!!$.*

Proof. The theorem can be proved by induction on $|X|$. We give here another proof based on generating functions.

Clearly every tree $T \in B(X)$ gives rise to two subtrees $T(\ell)$ and $T(r)$, where ℓ and r are the children of the root of T, the union of whose leaf-label sets is X. Conversely, given nonempty subsets X_ℓ and X_r with $X_\ell \cup X_r = X$, one can join a rooted binary tree on X_ℓ with a rooted binary tree on X_r to obtain a rooted binary tree on X. It follows that

$$|B(X)| = \frac{1}{2} \sum_{X_\ell, X_r} |B(X_\ell)|\,|B(X_r)|,$$

where the sum is over all partitions $\{X_\ell, X_r\}$ of X; the factor $1/2$ accounts for the fact that there are two ways to designate ℓ and r as children of the root of T. Since $|B(X)|$ depends only on $|X|$, and *not* on particular elements of X, let $b(n) = |B(X)|$. We then have

$$b(n) = \frac{1}{2} \sum_{i=1}^{n-1} \binom{n}{i} b(i)\, b(n-i).$$

Let $b(x) = \sum_{n=1}^{\infty} b(n)x^n/n!$. The series $b(x)$ is called the exponetial generating function for $b(n)$. Then from the recursion above, we have

$$b(x) = x + b(x)^2/2.$$

Solving this equation, we have $b(x) = 1 \pm \sqrt{1-2x}$; however, note that $\lim_{x\to 0} b(x) = 0$, and hence $b(x) = 1 - \sqrt{1-2x}$. We now expand $\sqrt{1-2x}$ using the binomial theorem

$$\sqrt{1-2x} = (1-2x)^{1/2} = \sum_{n=0}^{\infty} \binom{1/2}{n} (-2x)^n.$$

Therefore,

$$\begin{aligned} b(n)/n! = [x^n]\, b(x) &= [x^n](1 - \sqrt{1-2x}) \\ &= [x^n] \left(1 - \sum_{n=0}^{\infty} \binom{1/2}{n} (-2x)^n \right) \\ &= - \binom{1/2}{n} (-2)^n = (2n-3)!!/n!. \end{aligned}$$

The theorem now follows. □

Using the bijective mapping ψ_x from the last lecture, we can easily prove the following corollary.

Corollary 2. *The number of unrooted binary trees on X is $|U(X)| = (2n-5)!!$.*

We next describe a correspondence between rooted binary trees and perfect matchings. For simplicity, we assume that $X = \{1, \dots, n\}$.

A perfect matching on $2n-2$ points is a set of $n-1$ unordered pairs of these points. Let T be rooted binary trees with its leaves labeled by numbers $1, \dots, n$, and let $I = \{n+1, \dots, 2n-2\}$. We extend the labeling to the internal vertices of T as follows.

1. Compute the set U of unlabeled internal vertices of T both of whose children have already been labeled.

2. Find among the children of the elements of U vertex w that has the smallest label.

3. Label the parent of w with the smallest integer of I, and update I by discarding that integer from it.

Note that since I initially has $n-2$ elements and T has $n-1$ internal vertices, the root of T is not labeled. For each internal vertex of T, we form an unordered pair of the labels of the vertex's children, and take the set of these pairs as a perfect matching on points $1, \dots, 2n-2$.

Theorem 3. *Each rooted binary tree corresponds to one and exactly one perfect matching as described above.*

The formular in Theorem 1 is a direct consequence of Theorem 3, for the number of perfect matchings on $2n-2$ points is:

$$\frac{1}{(n-1)!}\binom{2n-2}{2}\binom{2n-4}{2}\cdots\binom{2}{2} = (2n-3)!!,$$

where factor $(n-1)!$ accounts for the fact that the relative order of pairs of points in a perfect matching is disregarded.

Application: Catalan numbers. Recall that we can represent a rooted phylogenetic tree in the textual form using the Newick format. For example, $((a,b),c)$ represents a rooted caterpillar tree with three leaves a, b, and c. Clearly, a tree can have more than one Newick presentation. In fact, a rooted binary tree with n leaves has exactly 2^{n-1} Newick representations (why?).

Let us fix a particular string consisting of the elements of X, and let C_{n-1} be the number of ways to fully parenthesize the string (e.g. string *abcd* can be parenthesized as $(((ab)c)d)$, $(a((bc)d))$, $((ab)(cd))$, etc). It is easy to see that each parenthesization of the string is in fact the Newick representation of a rooted binary tree on X (to be precise, they are identical, except that commas do not appear in the former). Therefore, we can have C_{n-1} rooted binary trees given a fixed string of the elements of X. But we have $|X|! = n!$ such strings, and each rooted binary tree has, as noted in the preceding paragraph, 2^{n-1} paranthesizations of these strings. It follows that

$$n!\, C_{n-1} = (2n-3)!!\, 2^{n-1},$$

and hence $C_{n-1} = \frac{1}{n}\binom{2n-2}{n-1}$.

Counting tree shapes. We say that two phylogenetic trees have the same shape if they are isomorphic. We denote by $s(n)$ the number of binary rooted tree shapes on n vertices. The first few values of $s(n)$ are: $s(1) = s(2) = s(3) = 1$, $s(4) = 2$, $s(5) = 3$, $s(6) = 6$.

Theorem 4. *We have*

$$s(n) = \begin{cases} 1 & \text{if } n = 1, \\ \sum_{i=1}^{k} s(i)s(n-i) & \text{if } n = 2k+1,\ n > 1, \\ \frac{s(k)(s(k)+1)}{2} + \sum_{i=1}^{k-1} s(i)s(n-i) & \text{if } n = 2k,\ n > 1. \end{cases}$$

Proof. Assume $n > 1$. Consider a binary rooted tree T, and let T_ℓ and T_r be the two pendant subtrees rooted at the children of the root of T. If we impose that T_r has at least as many leaves as T_ℓ does, then each T is uniquely determined by an *ordered* pair (T_ℓ, T_r). The theorem is now a straighforward consequence of this (and a bit of care for the case $n = 2k$ and both T_ℓ and T_r have exactly k leaves). $\square$

Counting leaf-labelings. We next consider the problem of different leaf-labelings of a binary rooted tree shape. Let T a be tree shape with n leaves, and let v be an internal vertex of T. We say that v is a symmetry vertex if the two pendant subtrees rooted at children of v has the same shape.

Theorem 5. *Let $\sigma(T)$ be the number of symmetry vertices in a rooted binary tree shape T. Then the number of leaf-labelings of T is $n!/2^{\sigma(T)}$.*

We have so far considered trees with its leaves labeled. What about trees with all the vertices labeled? The following result is due to Cayley.

Theorem 6. *The number of trees with n vertices, all of which are labeled, is n^{n-2}.*

Proof. The theorem can be proved by using Prüfer code. $\square$