Split times in multispecies coalescent

Graham Jones

2013-11-18, August 11, 2014

1 Introduction

Suppose that n individuals have been sampled from an unknown number of species. If the individuals are assigned to species by some method, it is possible to make various kinds of errors. Here, I concentrate on the case of erroneously splitting a single species into two. Suppose that m gene trees (topologies and node times) $T_1, \ldots, T_m$ are available for the n individuals, and that these are independent (the m genes are unlinked) and each follows the Kingman coalescence model with the same population parameter θ. In practice the gene trees would need to be estimated, but I assume they are known exactly.

It is impossible to rule out the possibility that there has been a very recent speciation, and it is interesting to find out about the distribution of the most ancient time that can split the individuals into two sets and is compatible with all the gene trees. Let $S = \{1, 2, \ldots, n\}$ label the individuals. Then I would like the distribution of

$$\max_{\emptyset \subset A \subset S} \left(\min_{1 \le j \le m} h(A, T_j) \right) \tag{1}$$

where $h(A, T)$ is the time of the first coalescence between a member of A and a member of $S \setminus A$ in tree T. This appears to be difficult, but I do have a first step, namely the distribution of $h(A, T)$ when $|A| = 1$. Since the trees are assumed independent, the distribution of $\min_{1 \le j \le m} h(A, T_m)$ follows for this case. I think extending this to a general but fixed A is feasible. I think that the result will be a weighted sum of exponentials, and I think a recursive definition for the weights is straightforward (but at the moment it looks messy). However, taking the maximum over all A looks hard.

2 Time to first coalescence for one individual

Theorem 1 *Assume $|A| = 1$. Let T be a tree sampled from the Kingman coalescence model with n tips and population parameter θ. Then the cumulative distribution of $h(A, T)$ is given by*

$$F_n(t) = \sum_{i=2}^{n} H(n, i)(1 - e^{-(i(i-1)/2)\theta t}) \tag{2}$$

where

$$H(n, i) = i(i-1)(2i-1)\frac{(n-1)!(n-2)!}{(n-i)!(n+i-1)!}. \tag{3}$$

First I prove a lemma. I found the form of $D(n, i)$ using the ideas in [2], especially those of Gosper [1], and used the software Reduce (http://sourceforge.net/projects/reduce-algebra/). The $H(n, i)$ turn out to be 'Gosper-summable'.

Lemma 2 *Let $H(n, i)$ be defined as in equation 26. Then $\sum_{i=1}^{n} H(n, i) = 1$.*

Proof. Let

$$D(n, i) = 1 - \frac{(n - 1 + i^2)(n-1)!(n-2)!}{(n-i-1)!(n+i-1)!}$$

for $0 \le i \le n-1$, and let $D(n,n) = 1$. I show that the $D(n,i)$ are partial sums of the $H(n,i)$. First

$$D(n,0) = 1 - \frac{(n-1)(n-1)!(n-2)!}{(n-1)!(n-1)!} = 0$$

and for $1 \le i \le n-1$,

$$\begin{aligned}
\frac{D(n,i) - D(n,i-1)}{(n-1)!(n-2)!} &= \left(\frac{(n-1+(i-1)^2)}{(n-i)!(n+i-2)!} - \frac{(n-1+i^2)}{(n-i-1)!(n+i-1)!} \right) \\
&= \frac{(n-1+(i-1)^2)(n+i-1) - (n-1+i^2)(n-i)}{(n-i)!(n+i-1)!} \\
&= \frac{(i^2-2i+n)(i+n-1) + (i^2+n-1)(i-n)}{(n-i)!(n+i-1)!} \\
&= \frac{2i^3 - 3i^2 + i}{(n-i)!(n+i-1)!} \\
&= \frac{i(i-1)(2i-1)}{(n-i)!(n+i-1)!}
\end{aligned}$$

so $D(n,i) - D(n,i-1) = H(n,i)$ and

$$D(n,n) - D(n,n-1) = \frac{(n-1+(n-1)^2)(n-1)!(n-2)!}{0!(2n-2)!} = H(n,n).$$

Thus

$$\sum_{i=1}^{n} H(n,i) = D(n,n) = 1. \tag{4}$$

□

Proof of theorem. Define $B(i) = i(i-1)/2$. The time for the next coalescence among i gene copies is given by the density $g_i(t) = B(i)\theta e^{-B(i)\theta t}$ for $i \ge 2$. Let X_i be a random variable with this distribution. Let Y_n be a random variable with the distribution we are looking for, that is, the first time A merges with some other lineage. Denote the density of Y_n by $f_n(t)$. Clearly $f_2 = g_2$. For $n \ge 3$, the probability that A merges at the first coalescence is $(n-1)/B(n) = 2/n$. If it is not in the first coalescence, there will be $n-1$ other gene copies left and the time for A to coalesce is $X_n + Y_{n-1}$. Thus

$$f_n = (2/n)g_n + ((n-2)/n)g_n \circ f_{n-1} \tag{5}$$

where $\circ$ is convolution. A straightforward calculation shows that

$$g_i \circ g_j = \frac{B(j)g_i - B(i)g_j}{B(j) - B(i)} \tag{6}$$

so by induction, f_n is a linear combination of the g_i $(2 \le i \le n)$. Write

$$f_n = \sum_{i=2}^{n} C(n,i) g_i$$

where the $C(n,i)$ are constants. We have $C(2,2) = 1$. For $n \ge 3$,

$$\begin{aligned}
f_n &= \frac{2}{n} g_n + \frac{n-2}{n} g_n \circ \sum_{i=2}^{n-1} C(n-1,i) g_i \\
&= \frac{2}{n} g_n + \frac{n-2}{n} \sum_{i=2}^{n-1} C(n-1,i) g_n \circ g_i \\
&= \frac{2}{n} g_n + \frac{n-2}{n} \sum_{i=2}^{n-1} C(n-1,i) \frac{B(n)g_i - B(i)g_n}{B(n) - B(i)} \\
&= \frac{2}{n} g_n + \frac{n-2}{n} \sum_{i=2}^{n-1} \frac{B(n)C(n-1,i)g_i}{B(n) - B(i)} - \frac{n-2}{n} \sum_{i=2}^{n-1} \frac{B(i)C(n-1,i)g_n}{B(n) - B(i)} \\
&= \left(\frac{2}{n} - \frac{n-2}{n} \sum_{i=2}^{n-1} \frac{B(i)C(n-1,i)}{B(n) - B(i)} \right) g_n + \frac{n-2}{n} \sum_{i=2}^{n-1} \frac{B(n)C(n-1,i)}{B(n) - B(i)} g_i.
\end{aligned}$$

So

$$C(n,i) = \frac{n-2}{n}\frac{B(n)}{B(n)-B(i)}C(n-1,i) \tag{7}$$

for $i < n$ and

$$C(n,n) = \frac{2}{n} - \frac{n-2}{n}\sum_{i=2}^{n-1}\frac{B(i)}{B(n)-B(i)}C(n-1,i). \tag{8}$$

Note that since the $C(n,i)$ are the mixture weights of a normalised density, we must have

$$\sum_{i=1}^{n} C(n,i) = 1. \tag{9}$$

We use induction on n to show that $C(n,i) = H(n,i)$, starting with $H(2,2) = 1$. Suppose that $n \geq 3$ and that $C(n-1,i) = H(n-1,i)$ for $1 \leq i \leq n-1$.

$$\begin{aligned}
\frac{H(n,i)}{H(n-1,i)} &= \frac{(n-1)!(n-2)!}{(n-i)!(n+i-1)!}\frac{(n-1-i)!(n+i-2)!}{(n-2)!(n-3)!} \\
&= \frac{(n-1)(n-2)}{(n-i)(n+i-1)} \\
&= \frac{(n-1)(n-2)}{n(n-1)-i(i-1)} \\
&= \frac{n-2}{n}\frac{n(n-1)}{n(n-1)-i(i-1)} \\
&= \frac{n-2}{n}\frac{B(n)}{B(n)-B(i)}.
\end{aligned}$$

which shows that the $H(n,i)$ satisfy equation (7) for $i < n$, and so $C(n,i) = H(n,i)$ for $i < n$. From equations (9) and Lemma 2 we must also have $C(n,n) = H(n,n)$. □

A few observations.

- It is not at all obvious from equation (8) that $C(n,n) = H(n,n)$ (or even that $C(n,n)$ is positive). That's why my proof requires the lemma. I don't have an intuitive understanding of equation (8).
- A calculation shows that
$$f_n(0) = \sum_{i=2}^{n} C(n,i)B(i) = n-1$$
- $C(n,n) = n\binom{2(n-1)}{n-1}^{-1}$ which is the reciprocal of the $(n-1)$st Catalan number.
- $i(i-1)(2i-1)/6$ is the sum of the first i squares.

3 Time to first coalescence for more individuals

Let $f_{r,n}$ be the density of the first coalescence of A and $S \setminus A$ when $|A| = r$. Then

$$f_{r,n} = \frac{2r(n-r)g_n + (n-r)(n-r-1)f_{r,n-1} \circ g_n + r(r-1)f_{r-1,n-1} \circ g_n}{n(n-1)} \tag{10}$$

so we can write

$$f_{r,n} = \sum_{i=2}^{n} C_r(n,i)g_i \tag{11}$$

and using (6) it follows that for $i < n$,

$$C_r(n,i) = \frac{(n-r)(n-r-1)C_r(n-1,i) + r(r-1)C_{r-1}(n-1,i)}{(n-i)(n+i-1)} \tag{12}$$

Define

$$E_r(n,i) = \frac{(n-i)!(n+i-1)!}{(n-r)!(n-r-1)!} C_r(n,i) \tag{13}$$

Then can write the recursion as

$$E_r(n,i) = E_r(n-1,i) + r(r-1)E_{r-1}(n-1,i) \tag{14}$$

Applying this j times,

$$E_r(n,i) = E_r(n-j,i) + r(r-1)\sum_{k=1}^{j} E_{r-1}(n-k,i) \tag{15}$$

3.1 Two individuals

Theorem 3 *Assume $|A| = 2$. Let T be a tree sampled from the Kingman coalescence model with n tips and population parameter θ. Then the cumulative distribution of $h(A,T)$ is given by*

$$F_{2,n}(t) = \sum_{i=2}^{n} H_2(n,i)(1 - e^{-(i(i-1)/2)\theta t}) \tag{16}$$

where

$$H_2(n,i) = \frac{(n-2)!(n-3)!i(2i-1)(2i-2)}{(n-i)!(n+i-1)!} \left(\frac{(2i-3)!}{(i-2)!(i-3)!} \frac{i^2+i-6}{6} \frac{(i-1)!^2}{(2i-2)!} + (n-i) \right) \tag{17}$$

Proof.

$$\begin{aligned} C_2(n,i) &= \frac{(n-2)(n-3)C_2(n-1,i) + 2C_1(n-1,i)}{(n-i)(n+i-1)} && (18)\\ &= \frac{(n-2)(n-3)C_2(n-1,i) + 2i(i-1)(2i-1)\frac{(n-2)!(n-3)!}{(n-i-1)!(n+i-2)!}}{(n-i)(n+i-1)} && (19)\\ & && (20) \end{aligned}$$

So

$$\frac{(n-i)!(n+i-1)!}{(n-2)!(n-3)!} C_2(n,i) = \frac{(n-i-1)!(n+i-2)!}{(n-3)!(n-4)!} C_2(n-1,i) + 2i(i-1)(2i-1) \tag{21}$$

The RHS and first term on the LHS are the same except that n on the right becomes $n-1$ on the left. Applying this j times,

$$\frac{(n-i)!(n+i-1)!}{(n-2)!(n-3)!} C_2(n,i) = \frac{(n-i-j)!(n+i-j-1)!}{(n-j-2)!(n-j-3)!} C_2(n-j,i) + 2ji(i-1)(2i-1) \tag{22}$$

Setting $j = n-i$

$$\frac{(n-i)!(n+i-1)!}{(n-2)!(n-3)!} C_2(n,i) = \frac{0!(2i-1)!}{(i-2)!(i-3)!} C_2(i,i) + 2(n-i)i(i-1)(2i-1) \tag{23}$$

To be continued...

In `reduce`,

```
summand:=k*(k-1)*(2*k-1)*factorial(n-2)*factorial(n-3)*binomial(2*n-1,n-k)*
         ((k+3)*(k-2)*(k-2)*(k-1)+12*(n-k))/(6*factorial(2*n-1));
summand;
```

$$\frac{\binom{2n-1}{-k+n}(n-3)!\,(n-2)!k\left(2k^6-7k^5-7k^4+35k^3+24k^2n-55k^2-36kn+44k+12n-12\right)}{6\,(2n-1)!} \tag{24}$$

```
gosper(summand,k,n);
```

$$\frac{\left(k^6+6n^3-24n^2+30n-12+(3n-14)\,k^4+\left(6n^2-21n+25\right)k^2\right)(k-n)\binom{2n-1}{-(k-n)}(n-3)!\,(n-2)!}{6\,(2n-1)!} \tag{25}$$

Note that

$$\begin{aligned} H_2(n,n) &= \frac{(n-2)!(n-3)!n(n-1)(2n-1)}{0!(2n-1)!}\frac{(n+3)(n-2)^2(n-1)}{6} && (26)\\ &= \frac{(n-1)!(n-1)!n}{(2n-2)!}\frac{(n+3)(n-2)}{6} && (27)\\ &= n\binom{2n-2}{n-1}^{-1}\frac{(n+3)(n-2)}{6} && (28) \end{aligned}$$

3.2 Three individuals

It goes on. No proof of the following but the numerical evidence is strong.

Theorem 4 *Assume $|A| = 3$. Let T be a tree sampled from the Kingman coalescence model with n tips and population parameter θ. Then the cumulative distribution of $h(A,T)$ is given by*

$$F_{3,n}(t) = \sum_{i=2}^{n} H_3(n,i)(1-e^{-(i(i-1)/2)\theta t}) \tag{29}$$

where

$$H_3(n,n) = n\binom{2n-2}{n-1}^{-1}\frac{(n-3)((n-3)^3+16(n-3)^2+65(n-3)+38)}{120} \tag{30}$$

I think other $H(n,i)$ can be found from this. I have written the polynomial in $(n-3)$ beacuse it seems to follow a bit of a pattern. $H_2(n,n)$ multiplied by the $(n-1)$st Catalan number is

$$\frac{(n-2)((n-2)+5)}{3!}$$

$H_3(n,n)$ multiplied by the $(n-1)$st Catalan number is

$$\frac{(n-3)((n-3)^3+16(n-3)^2+65(n-3)+38)}{5!}$$

and note the coefficients are all positive and sum to 1.

References

[1] R. W. Gosper. Decision procedure for indefinite hypergeometric summation. *Proc. Natl. Acad. Sci. USA*, 75:40–42, 1978.

[2] Marko Petkovsek, Herbert Wilf, and Doron Zeilberger. *A=B*. ???, 1997.