How to Cite:
Sivaraman, R., & Sengothai, R. (2022). Mathematics of coalescent gene trees. *International Journal of Health Sciences*, *6*(S1), 5785–5789. https://doi.org/10.53730/ijhs.v6nS1.6146

Mathematics of coalescent gene trees

Dr. R. Sivaraman
Associate Professor, Department of Mathematics, Dwaraka Doss Goverdhan Doss Vaishnav College, Chennai, India and National Awardee for Popularizing Mathematics among masses
Email: rsivaraman1729@yahoo.co.in

Dr. R. Sengothai
Mathematics Educator, Pie Mathematics Association, Choolaimedu, Chennai, India
Email: kothai1729@gmail.com

***Abstract*---**Using the coalescent process continuously until all the lineages have become one, we get different possible gene trees. In this paper I had computed the probabilities of such different gene trees that occur depending on the way lineages can form. In particular I had shown the probability of any 3 sample gene tree is $\frac{1}{3}$ and for four samples resembling caterpillar gene trees the probability is $\frac{1}{18}$ and probability for any balanced gene tree with 4 samples is $\frac{1}{9}$. I have also proved a theorem providing the number of ranked gene trees. This estimate will help us to determine the probability of any gene tree.

***Keywords*---**Gene Tree, Coalescence, Lineages, Balanced Gene Tree, Ranked Gene Trees.

Introduction

Suppose we sample the same gene in several individuals within a population and assume the coalescent model describes the probabilistic way their lineages coalesce. If the coalescent process is continued until all the lineages have become one, then a rooted gene tree is formed. Each possible gene tree can arise with some probability, which can be computed. In this paper, I demonstrate how this can be done in the simplest situation – the gene samples are taken from a single population, which persists back in time `forever'. (In essence, we assume a species tree with only a single taxon.) Gene tree probabilities can be computed for either topological gene trees or metric gene trees. In the following sections, I will present examples of both cases.

Modeling for 3-sample trees

First we consider sampling 3 extant lineages, which we will denote by A_1, A_2, A_3 at $u = 0$. If we are only concerned with gene tree topologies, then we observe that the rooted gene tree which relates them is determined by the first pair of lineages to coalesce. Since the coalescence of pairs is identically and independently distributed (i.i.d.), this first coalescence involves A_1, A_2 or A_2, A_3 or A_1, A_3 with equal probability $\frac{1}{3}$. Thus each of the 3 gene trees ((A_1, A_2), A_3), ((A_2, A_3), A_1), and ((A_1, A_3), A_2) occur with probability $\frac{1}{3}$.

For a metric gene tree probability note that distances will be measured in coalescent units, and the tree must be ultrametric. Remember that a gene tree is said to be ultrametric if the distance of any leaves remain same from the root which is considered to be common ancestor. For knowing more about metric and ultrametric trees see [1].

Consider the gene tree ((A_1:u_1, A_2:u_1):u_2, A_3:u_3), where $u_3 = u_1 + u_2$. This is formed by:

- 3 lineages and no coalescent events for a time interval of length u_1,
- a coalescent event of 3 lineages to 2 at time u_1, with the specific lineages A_1, A_2 coalescing,
- 2 lineages and no coalescent event for a time interval of length u_2,
- a coalescent event of 2 lineages to 1 at time $u_3 = u_2 + u_1$.

Now we know from coalescence theory that (a) has probability e^{-3u_1} and (c) has probability e^{-u_2}. Both (b) and (d) require probabilities of coalescence at an instant, which is simply the rate of coalescence times du. Thus (d) has probability $du_3 = du_2$, while (b) involves both the probability of a coalescence, $3du_1$, and an additional factor of $\frac{1}{3}$ that the lineages involved in the coalescence are A_1, A_2. The total probability is thus given by $P(u_1,u_2) = e^{-(3u_1+u_2)}\,du_1du_2$ so the probability density function is $f(u_1,u_2) = e^{-(3u_1+u_2)}$ (2.1)

Integrating (2.1) between the limits $0 \le u_1, u_2 < \infty$ we get

$$\int_{u_2=0}^{\infty}\int_{u_1=0}^{\infty} e^{-(3u_1+u_2)}\,du_1du_2 = \left(\int_{u_2=0}^{\infty} e^{-u_2}du_2\right)\times\left(\int_{u_1=0}^{\infty} e^{-3u_1}du_1\right) = 1\times\frac{1}{3} = \frac{1}{3} \quad (2.2)$$

Thus from (2.2), we notice that the probability of the topological tree ((A_1, A_2), A_3) is $\frac{1}{3}$ as mentioned above.

Modeling for Larger trees

If 4 extant genes A_1, A_2, A_3, A_4 are sampled at $u = 0$, the computation of probabilities of topological trees is a little more complicated. A caterpillar gene tree like (((*A*1, *A*2), *A*3), *A*4) can only be formed by a specific sequence of coalescent events. The probability that the first lineages A_1, A_2 to merge is given by $\frac{1}{\binom{4}{2}} = \frac{1}{6}$.

Here $\binom{4}{2} = 6$ represent selecting any two among four available samples.

Once there are only 3 lineages, the probability that A_1A_2 lineage merges with the A_3 lineage is given by $\frac{1}{\binom{3}{2}} = \frac{1}{3}$. Finally, the probability that the $A_1A_2A_3$ lineage merges with A_4 is 1.
Thus the probability of this, or any of the other caterpillar gene trees is given by $\frac{1}{6} \times \frac{1}{3} \times 1 = \frac{1}{18}$.

A balanced tree like ((*A*1, *A*2), (*A*3, *A*4)) can be formed by either of two sequences of coalescent events: A_1A_2 may merge first, followed by A_3A_4, or vice versa. Each of these has probability $\frac{1}{18}$ since they are determined by a specific sequence of coalescent events. Thus the total probability of this tree or any other balanced gene tree is $2 \times \frac{1}{18} = \frac{1}{9}$.

This last calculation shows a significant feature of the coalescent model in a single population. Topological gene trees that show more `balance' tend to have higher probability than those that are less balanced, because they can be achieved by more distinct orderings of coalescent events. In fact, it is not hard to generalize the calculation of topological gene tree probabilities from 4-samples to more. We now define a ranked gene tree.

A ranked gene tree is a rooted binary leaf-labelled topological tree with an ordering to the internal nodes (from the leaves to the root) such that the ranking of any node is greater than all its descendants. Then under the coalescent all ranked gene trees are equally probable. I now compute the number of ranked gene trees through the following theorem.

Theorem 1

The number of ranked gene trees is $\frac{n!\times(n-1)!}{2^{n-1}}$ (4.1)

Proof: If $R(n)$ denotes the number of ranked gene trees then it is given by

$$R(n)=\prod_{k=2}^{n}\binom{k}{2}=\binom{2}{2}\times\binom{3}{2}\times\binom{4}{2}\times\cdots\times\binom{n-1}{2}\times\binom{n}{2}$$
$$=\frac{2\times1}{1\times2}\times\frac{3\times2}{1\times2}\times\frac{4\times3}{1\times2}\times\cdots\times\frac{(n-1)\times(n-2)}{1\times2}\times\frac{n\times(n-1)}{1\times2}$$
$$=\frac{n!\times(n-1)!}{2^{n-1}}$$

This proves (4.1) and completes the proof.

With the aid of theorem 1, we notice that the probability of any gene tree is simply the number of rankings it may be given divided by $R(n)^2$

Conclusion

Considering topological gene trees, I had created two models one for 3 sample trees and other for larger sample trees. In particular, I had proved that the probability of any 3 – sample gene tree is $\frac{1}{3}$ in section 2. Similarly in section 3, I had proved that the probability of any caterpillar type gene tree is $\frac{1}{18}$ and probability of any balanced gene tree is $\frac{1}{9}$.

After defining ranked gene tree in section 3, I had obtained a closed expression for number of ranked gene trees through $R(n)$ in equation (4.1). Using this expression, we see that the probability of any gene tree is simply the number of rankings it may be given divided by $R(n)^2$. Overall, these calculations provide a better understanding of how gene trees behave.

References

[1] R. Sivaraman, On Metric and Ultrametric Trees, European Journal of Molecular and Clinical Medicine, Volume 7, Issue 8, 2611 – 2615, 2020.
[2] John Wakeley, Coalescent Theory: An Introduction, Roberts and Company, Greenwood Village, Colorado, 2009.
[3] Michael S. Waterman, Introduction to Computational Biology: Maps, Sequences and Genomes, Chapman and Hall, London, 1995.

[4] Elizabeth S. Allman and John A. Rhodes, Mathematical Models in Biology: An Introduction, Cambridge University Press, Cambridge, 2004.
[5] L. Knowles and L. Kubatko, Estimating Species Trees: Practical and Theoretical Aspects, Wiley-Blackwell, College Station, Texas, 2010.