

Beta Coalescent

In the last part of my dissertation research I have been examining an alternative to the standard coalescent (Kingman's) which occurs when offspring distributions have infinite variance is called the *Beta coalescent* (Schweinsberg [2003]). Some of the questions are 1) Is the Beta coalescent a better model for some marine or terrestrial species? 2) How do we build a statistical framework to estimate family size parameters?

The coalescent describes the connection between demographic history and genetic data. It describes the ancestral relationship (gene genealogy) of a sample of DNA sequences, and it is widely used to make predictions about patterns of genetic variation. Kingman's coalescent is described by gene genealogies that are characterized by rooted binary trees. The root represents the most recent common ancestor of the sample, and each tip corresponds to a sampled sequence at the present time. If one has a sample of size n , at most two genes can have the same common ancestor in a specific earlier generation. These binary trees occur because there is an underlying assumption: the variance of the number of offspring among individuals converges to a finite constant as the population size goes to infinity. The Wright-Fisher model is a forward model that considers a population of size $2N$, and the probability that an individual will have m children is binomial with parameters $2N$ and $p = 1/2N$ which is approximately Poisson with parameter 1 when N is large. Under the appropriate time scaling (1 unit of time is equal to $2N$ generations) and as $N \rightarrow \infty$, one obtains Kingman's coalescent. But what if the number of offspring of an individual is not binomial, and the variance of the offspring distribution is not small as a proportion of a large population? There appears to be evidence that in some species, such as marine animals, an individual can produce a very large number of offspring, comparable in fact to the size of the entire population (Hedgecock [1994]). Modeling these populations with Wright-Fisher dynamics is not appropriate. This can lead to an effective population size that is much smaller than under the finite variance model. However, if the variance of offspring goes to infinity, then Mohle and Sagitov [2001] have shown that the limiting process is qualitatively different, the resulting coalescent process they named the Λ -coalescent.

There are various types of the Λ -coalescent, and for the last part of my PhD thesis I have been studying the Beta coalescent (with Rick Durrett and Carlos Bustamante). The Beta coalescent is characterized by family size distributions that have a power law tail with exponent $-\alpha$, where $1 \leq \alpha < 2$. This results in gene genealogies that are characterized by *multiple gene mergers* (see Schweinsberg [2003]). α close to 1 implies bigger family sizes whereas α close to 2 implies smaller family sizes and a return to a binary tree.

As part of my PhD work, I am developing a maximum likelihood estimation procedure to test whether Beta coalescents are appropriate for marine species genealogies using published data sets from Arnason [2004] and Boom et al. [1994], by developing a maximum likelihood estimation procedure. I have been looking at estimates of α which is the most important parameter due to its drastic effect on the time scale. I am using results derived by Berestycki et al. [2007], which include formulas for the number of segregating sites and the site frequency spectrum (sites where i individuals in the sample of size n have the mutant allele) which show in the limit as the sample size $n \rightarrow \infty$. From these formulas one can estimate α and θ , but these estimates are biased for fixed sample size, n , which is possible to correct through simulation. Therefore, I have written a coalescent simulation that incorporates multiple mergers to compute the expected site frequency

spectrum for fixed n , which then can be directly used to define a multinomial likelihood function to estimate α . This model will be applied to dog and cow data currently being analyzed in the Bustamante lab. Dogs and cows, due to the fact that a small number of males produce most of the offspring, manifest a high degree of homozygosity. This is a signal that their genetics are not well explained by Kingman's coalescent, and will benefit from using an improved null model such as the beta coalescent.

References

- E. Arnason. Mitochondrial cytochrome b variation in the high-fecundity atlantic cod: trans-atlantic clines and shallow gene geneology. *Genetics*, 166:1871–1885, 2004.
- N. Berestycki, J. Berestycki, and J. Schweinsberg. Beta-coalescents and continuous stable random trees. *Ann. Probab.*, 35:1835–1887, 2007.
- J.D.G. Boom, E.G. Boulding, and A.T. Beckenback. Mitochondrial dna variation in introduced populations of pacific oyster, *Crassostrea gigas*, in british columbia. *Can. J. Fish. Aquat. Sci.*, 51:1608–1614, January 1994.
- D. Hedgecock. *Does variance in reproductive success limit effective population sizes of marine organisms*, pp. 1222-1344 in *Genetics and Evolution of Aquatic Organism*, edited by A. Beaumont. Chapman and Hall, London, 1994.
- M. Mohle and S. Sagitov. A clasification of coalescent processes for haploid exchangeable population models. *Ann. Prob.*, 27:1547–1562, 2001.
- Jason Schweinsberg. Coalescent processes obtained from supercritical galton-watson processes. *Stoch. Proc. Appl.*, 106, 2003.