Models of gene content evolution

Consider a species tree S. Denote by x the vector of gene counts in each species observed for a gene family. Furthermore, denote by x_v the part of x corresponding to gene counts in the leaves of the subtree of S rooted in v. Denote by X_v the number of gene copies that actually existed at vertex v (with X the number at the root of S). The main goal is to obtain the posterior density for parameters of some model of gene family evolution $\mathcal{M}$ given a known species tree S

$$p(\theta|x, \mathcal{M}, S) \propto p(x|\theta, \mathcal{M}, S)p(\theta|S)$$

If we can compute the right hand side, standard techniques from Bayesian inference allow us to approximate or sample from the posterior density. The tricky part is of course the likelihood term $P(x|\theta, \mathcal{M}, S)$.

In the following, we suppress the dependence on $\mathcal{M}$. The key idea is to condition on the number of *surviving lineages* Y_v, which is the number of lineages that existed at vertex v that have left descendants at the leaves of S (we denote this number at the root of S by Y). Using this survival probability, the likelihood can be expressed as

$$p(x|\theta, S) = \sum_{n=0}^{\infty} p(x|\theta, S, Y = n)p(Y = n)$$

The next key step is to similarly define the extinction probability ϵ_v which is the probability that a lineage that existed at vertex v went extinct such that it did not leave observed descendants. We can further rewrite the expression for the likelihood as

$$\begin{aligned} p(x|\theta, S) &= \sum_{n=0}^{\infty} p(x|\theta, S, Y = n) \sum_{i=0}^{\infty} p(Y = n|X = n + i)p(X = n + i) \quad (1) \\ &= \sum_{n=0}^{\infty} p(x|\theta, S, Y = n) \sum_{i=0}^{\infty} \binom{n+i}{i} \epsilon^i (1-\epsilon)^n p(X = n + i) \quad (2) \end{aligned}$$

Two important realizations help further computations. Firstly, the infinite sum over n (the number of surviving lineages) is not infinite in practice, as the maximum number of lineages extant at the root surviving up to the present is bounded by the number of presently observed lineages expressed in x. Secondly, for some distributions of the number of lineages at the root of S (e.g. geometric), the second infinite sum has a closed form. For instance, if we assume a geometric probability distribution with parameter η for X we get

$$
\begin{aligned}
\sum_{i=0}^{\infty} \binom{n+i}{i} \epsilon^i (1-\epsilon)^n p(X = n+i) &= \sum_{i=0}^{\infty} \binom{n+i}{i} \epsilon^i (1-\epsilon)^n \eta (1-\eta)^{n+i-1} \\
&= \eta (1-\eta)^{n-1} (1-\epsilon)^n \sum_{i=0}^{\infty} \binom{n+i}{i} \epsilon^i (1-\eta)^i \\
&= (1-\epsilon)^n \frac{\eta (1-\eta)^{n-1}}{(1-(1-\eta)\epsilon)^{n+1}} \qquad (3)
\end{aligned}
$$