Probabilistic modeling of gene family evolution

Gene family content evolution

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

$$P(\theta|X, \mathcal{S}) \propto P(X|\theta, \mathcal{S})P(\theta|S)$$

The main challenge is to specify a model of gene family evolution that results in a well defined probability measure $P(X|\theta, \mathcal{S})$ that can be efficiently computed for any θ and $\mathcal{S}$. An additional point of interest is to specify reasonable priors $P(\theta|\mathcal{S})$, encoding our assumptions on the evolutionary process. If both are available, standard techniques from Bayesian statistics can be applied to approximate the posterior density.

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

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