Simple population model for *BEAST or allopolyploids

Graham Jones

2011-09-23

1 Assumptions

Assume that each branch has a constant population. Assume a prior for this where the branch populations are independent and have the same distribution. Then, instead of adding a parameter for each branch and samplong from it, these parameters will be integrated out. The method is similar to the usual one for modelling site rate heterogeneity where you assume that each site independently chooses a rate from a gamma (or other) distribution. Unlike the site heterogeneity case, there is no need to approximate the integral.

2 Calculation

The *BEAST equation [1], p572, (3) for a constant population is

$$\Pr(L|N) = N^{-(k-1)} \prod_{i=0}^{k} \exp\left(- \int_{t_i}^{t_{i+1}} \binom{n-i}{2} N^{-1} \right)$$

where L is the lineage history of a gene tree within a single branch, and N is the effective population for this branch. L consists of the number n of lineages at the tipward end of the branch, plus the times $(t_0, t_1, ...t_k, t_{k+1})$ where t_0 is the node time at the tipward end, t_{k+1} is the node time at the rootward end, and $(t_1, ...t_k)$ are the coalescent times. Between t_i and t_{i+1} there are $n-i$ lineages.

$$\Pr(L|N) = N^{-(k-1)} \prod_{i=0}^{k} \exp\left(- (t_{i+1} - t_i) \binom{n-i}{2} N^{-1} \right)$$

$$\Pr(L|N) = N^{-(k-1)} \exp\left(- \left[\sum_{i=0}^{k} (t_{i+1} - t_i) \binom{n-i}{2} \right] N^{-1} \right)$$

$$\Pr(L|N) = N^{-(k-1)} \exp(-\gamma N^{-1})$$

where $\gamma = \sum_{i=0}^{k} (t_{i+1} - t_i)\binom{n-i}{2}$. This has the form of an (unnormalised) inverse gamma density for N. If the prior for N is assumed to also have an inverse gamma density, it will possible to integrate out N analytically. Suppose the prior is

$$g(N) = \beta^\alpha \Gamma(\alpha)^{-1} N^{-\alpha-1} \exp(-\beta N^{-1})$$

Then

$$\int_0^\infty g(N) \Pr(L|N) \mathrm{d}N = \int_0^\infty \beta^\alpha \Gamma(\alpha)^{-1} N^{-\alpha-k} \exp(-(\beta+\gamma)N^{-1})$$

$$= \frac{\beta^\alpha}{(\beta+\gamma)^{\alpha+k+1}} \frac{\Gamma(\alpha+k+1)}{\Gamma(\alpha)} \int_0^\infty (\beta+\gamma)^{\alpha+k+1} \Gamma(\alpha+k+1)^{-1} N^{-\alpha-k} \exp(-(\beta+\gamma)N^{-1})$$

Since the integrand is an inverse gamma density it integrates to 1, so

$$\int_0^\infty g(N) \Pr(L|N) \mathrm{d}N = \frac{\beta^\alpha}{(\beta+\gamma)^{\alpha+k+1}} \frac{\Gamma(\alpha+k+1)}{\Gamma(\alpha)}$$

The inverse gamma density is not very suitable for a prior for populaton sizes. If its shape parameter α is chosen to be small in order to give a large variance, the density is extremely small for small N but has a very long tail for large N. It would either rule out moderately small N or allow absurdly large N with too high a probability. However, by taking a mixture of inverse gamma densities, a reasonable prior can be formed. If enough components are used in the mixture, any reasonable prior could be approximated. I think about 10 components would be sufficient for nearly all purposes. If

$$h(N) = \sum_{i=1}^{c} \lambda_i g_i(N; \alpha_i, \beta_i)$$

where all $\lambda_i \geq 0$ and $\sum_{i=1}^{c} \lambda_i = 1$ and the g_i are inverse gamma densities with parameters α_i, β_i then

$$\int_0^\infty h(N) \Pr(L|N) \mathrm{d}N = \sum_{i=1}^{c} \lambda_i \frac{\beta_i^{\alpha_i}}{(\beta_i+\gamma)^{\alpha_i+k+1}} \frac{\Gamma(\alpha_i+k+1)}{\Gamma(\alpha_i)}$$

This expression is then multpilied over all branches. γ and k vary between branches.

3 Example distribution from R

```
scales <- 40 * 3^(1:10)
ms <- c(0.001, 0.009, 0.066, 0.132, 0.132, 0.132, 0.132, 0.132, 0.132, 0.132)
```

This means

$$\beta_i = 1/(40 \times 3^i)$$
$$\{\lambda_i\} = \{0.001, 0.009, 0.066, 0.132, 0.132, 0.132, 0.132, 0.132, 0.132, 0.132\}$$

All $\alpha_i = 1$.

Cumulative distribution

```
 N       cdf     1-cdf
10    6.144e-09 1
100   0.0005484 0.999451
1000  0.0347569 0.965243
10000 0.22149   0.778510
1e+05 0.4860904 0.513910
1e+06 0.7614095 0.238590
1e+07 0.957105  0.042895
1e+08 0.9953658 0.004634
1e+09 0.9995329 0.000467
```

References

[1] Heled and Drummond