Simple population model for multispecies coalescent

Graham Jones

2013-09-14, September 24, 2013

(This started two years ago as 2011-09-23-simple-pop-model.pdf. It seems a nice idea, but was half-baked. In particular the calculations were wrong: I had $N^{-(k-1)}$ not N^{-k}. Worse, there was confusion about multiplying the per branch formula.)

1 Description

Assume that each branch has a constant population. Assume the branch populations are independent and identically distributed. Then, instead of adding a parameter for each branch and sampling 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.

The probability density for the coalescence times of one gene in a species tree is given (p572, (3) of [1]) by the product over all branches of terms like this:

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

where L is the lineage history of a gene tree within a single branch, and $N = N(t)$ is the effective population for this branch, which is assumed constant in this paper. Here L consists of the number n of lineages at the tipward end of the branch, the number k of coalescences within the branch, plus the coalescence 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. The effective population N in equation (2) can be interpreted as a time, measured in generations: it is the expected time, for a single pair of lineages in a gene tree to coalesce. Since N is constant, equation (1) becomes

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

The complete multi-species coalescent probability density is the double product, over genes and over branches, of terms like this. To write down the full expression, some more notation is needed.

- The branches in the species tree are indexed by b. A sum or product over b should be understood as being over all branches. Note that this includes the root, so that all gene lineages eventually coalesce.
- The number of branches is B.
- Assume the mutation rate is μ_b substitutions per site per generation in branch b. Define $\theta_b = N_b \mu_b$ to be the population size parameter for branch b. Thus the θ_b are now in substitutions units.
- The genes are indexed by j. A sum or product over j should be understood as being over all genes.
- The number of coalescences of gene j within branch b is denoted by k_{jb}.
- The number of lineages in gene tree j at the tipward end of branch b is denoted by n_{jb}. Thus the number of lineages in gene tree j at the rootward end of branch b is $n_{jb} - k_{jb}$.
- The time interval between the tipward and rootward branch b is divided into $k_{jb} + 1$ intervals by the coalescent times of gene j. These $k_{jb} + 1$ intervals are denoted by c_{jbi} $(0 \leq i \leq k_{jb})$. There are $n_{jb} - i$ lineages in gene tree j, branch b during the time interval c_{jbi}.
- Let $q_b = \sum_j k_{jb}$ be the total number of coalescences of all genes in branch b.
- Let

$$\gamma_b = \sum_j \sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jb} - i}{2}$$

  which can be seen as a weighted sum of time intervals in branch b.
- The density π_σ is a user-chosen hyperprior, defined later.
- Terms c, λ_i, α_i, β_i are user-chosen values, defined later.

Let G denote all the lineage histories of all the genes in all the branches. The complete multi-species coalescent probability density is

$$\begin{aligned} f_G(G|\theta) &= \prod_j \prod_b \theta_b^{-k_{jb}} \exp\Big(- \Big[\sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jb} - i}{2} \Big] \theta_b^{-1} \Big) \\ &= \prod_b \theta_b^{-q_b} \exp\big(- \gamma_b \theta_b^{-1} \big) \qquad (3) \end{aligned}$$

For each b this has the form of an unnormalised inverse gamma density for θ_b. If, a priori, the θ_b are assumed independent and are assumed to have an inverse gamma density, it will possible to integrate out θ_b analytically. Suppose the common prior density is

$$g(x; \alpha, \beta) = \beta^\alpha \Gamma(\alpha)^{-1} x^{-\alpha-1} \exp(-\beta x^{-1}) \mathbf{1}_{[0,\infty)}$$

Then the joint density is

$$g(\theta; \alpha, \beta) = \prod_b \beta^\alpha \Gamma(\alpha)^{-1} \theta_b^{-\alpha-1} \exp(-\beta \theta_b^{-1}) \mathbf{1}_X$$

where X is the positive orthant in $\mathbf{R}^B$.

Then

$$\begin{aligned}
\int g(\theta) f_G(G|\theta)\mathrm{d}\theta &= \int_X \prod_b \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_b^{-\alpha-1} \exp(-\beta\theta_b^{-1})\theta_b^{-q_b} \exp\big(-\gamma_b\theta_b^{-1}\big)\mathrm{d}\theta \\
&= \prod_b \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_b^{-(\alpha+q_b)-1} \exp\big(-(\beta+\gamma_b)\theta_b^{-1}\big)\mathrm{d}\theta_b \\
&= \prod_b \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+q_b)}{(\beta+\gamma_b)^{\alpha+q_b}} \times \\
&\quad \int_0^\infty \frac{(\beta+\gamma_b)^{\alpha+q_b}}{\Gamma(\alpha+q_b)} \theta_b^{-(\alpha+q_b)-1} \exp\big(-(\beta+\gamma_b)\theta_b^{-1}\big)\mathrm{d}\theta_b \\
&= \prod_b \frac{\Gamma(\alpha+q_b)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta+\gamma_b)^{\alpha+q_b}} \qquad (4)
\end{aligned}$$

since the integrand in the penultimate line is an inverse gamma density.

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 θ but has a very long tail for large θ. It would either rule out moderately small θ or allow absurdly large θ with too high a probability. By using a mixture of inverse gamma densities, and introducing an overall scaling parameter σ, much more flexible priors can be formed. So let

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

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 and $\sigma\beta_i$, and where $\pi_\sigma(\sigma)$ is a hyperprior. Then, using (4),

$$\begin{aligned}
\int_0^\infty h(\theta,\sigma) f_G(G|\theta)\mathrm{d}\theta &= \pi_\sigma(\sigma) \sum_{i=1}^c \lambda_i \int_0^\infty g_i(\theta) f_G(G|\theta)\mathrm{d}\theta \\
&= \pi_\sigma(\sigma) \sum_{i=1}^c \lambda_i \prod_b \frac{\Gamma(\alpha_i+q_b)}{\Gamma(\alpha_i)} \frac{(\sigma\beta_i)^{\alpha_i}}{(\sigma\beta_i+\gamma_b)^{\alpha_i+q_b}} \qquad (5)
\end{aligned}$$

Here, c, the λ_i, α_i, β_i, are user-chosen values, which are constant for the analysis, and π_σ is a user-chosen density.

2 Usage

Element pioSpeciesTree replaces usual species tree
This provides a global prior on the population size parameters.

```
<pioSpeciesTree id="pio.species.tree">
    <pioSpeciesBindings idref="pio.species.bindings"/>
    <pioPopPriorScale>
        <parameter id="pio.pop.scale" value="1.0" lower="0.0" upper="Infinity"/>
    </pioPopPriorScale>
    <pioPopPriorInvGammas>
        <pioPopPriorComponent weight="1.0" alpha="4.0" beta=".003"/>
        <pioPopPriorComponent weight="1.0" alpha="4.0" beta=".001"/>
        <pioPopPriorComponent weight="1.0" alpha="4.0" beta=".0003"/>
    </pioPopPriorInvGammas>
</pioSpeciesTree>
```

Prior for species tree
`birthDeathModel` or others can be used.

```
<PopsIOSpeciesTreePrior id="PopsIOSpeciesTreePrior">
    <model>
        <birthDeathModel idref="birthDeath"/>
    </model>
    <pioTree>
        <pioSpeciesTree idref="pio.species.tree"/>
    </pioTree>
</PopsIOSpeciesTreePrior>
```

Coalescent likelihood for gene trees under species tree

```
<PopsIOMSCoalescent id="popsIO.MScoalescent">
    <pioSpeciesBindings idref="pio.species.bindings"/>
    <pioSpeciesTree idref="pio.species.tree"/>
</PopsIOMSCoalescent>
```

Operator to stretch/squash all gene trees and species tree
The pio.pop.scale is added to usual ones.

```
<upDownOperator scaleFactor="0.75" weight="30">
    <up>
        <parameter idref="29.clock.rate"/>
        <parameter idref="47.clock.rate"/>
        <parameter idref="53.clock.rate"/>
        <parameter idref="59.clock.rate"/>
        <parameter idref="64.clock.rate"/>
        <parameter idref="72.clock.rate"/>
        <parameter idref="species.birthDeath.meanGrowthRate"/>
    </up>
    <down>
        <pioSpeciesTree idref="pio.species.tree"/>
        <pioSpeciesTree idref="pio.pop.scale"/>
        <parameter idref="26.treeModel.allInternalNodeHeights"/>
        <parameter idref="29.treeModel.allInternalNodeHeights"/>
        <parameter idref="47.treeModel.allInternalNodeHeights"/>
        <parameter idref="53.treeModel.allInternalNodeHeights"/>
        <parameter idref="59.treeModel.allInternalNodeHeights"/>
        <parameter idref="64.treeModel.allInternalNodeHeights"/>
        <parameter idref="72.treeModel.allInternalNodeHeights"/>
    </down>
</upDownOperator>
```

Operators for species tree
Operator for pio.pop.scale added to usual; pioTreeNodeSlide replaces treeNodeSlide.

```
<scaleOperator scaleFactor="0.75" weight="3">
    <parameter idref="pio.pop.scale"/>
</scaleOperator>
<scaleOperator scaleFactor="0.75" weight="3">
    <parameter idref="species.birthDeath.meanGrowthRate"/>
</scaleOperator>
<scaleOperator scaleFactor="0.75" weight="3">
    <parameter idref="species.birthDeath.relativeDeathRate"/>
</scaleOperator>

<pioTreeNodeSlide weight="45">
    <pioSpeciesBindings idref="pio.species.bindings"/>
    <pioSpeciesTree idref="pio.species.tree"/>
</pioTreeNodeSlide>
```

Hyperprior for overall scaling of populations
This is in `prior` element in `posterior` element in `mcmc` element.

```
<logNormalPrior mean="0" stdev="3" offset="0" meanInRealSpace="false">
    <parameter idref="pio.pop.scale"/>
</logNormalPrior>
```

Log population scaling parameter
This is in main logger for parameters.

```
<parameter idref="pio.pop.scale"/>
```

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