STACEY package documentation: species delimitation with BEAST2

Graham Jones, www.indriid.com

2014-10-08, June 14, 2015

STACEY version 1.0.4

Contents

1 Introduction

The BEAST2 package STACEY can be used for species delimitation and species tree estimation. It incorporates a new model for the per-branch population parameters which is simpler than the one in *BEAST, but more flexible than one with equal populations in every branch. There are also new MCMC operators for the multispecies coalescent model. STACEY incorporates the 'birth-death-collapse' model from DISSECT. The models are described in two papers: Jones et al. (2014) (the 'DISSECT paper') and Jones (2015) (the 'STACEY paper').

For species delimitation, the main reason for using STACEY instead of DISSECT in BEAST1 is that it converges faster. It can be a good alternative to *BEAST for species tree estimation for the same reason, especially when there are many loci. It *may* be that the new operators work well with *BEAST (ignoring the rest of STACEY). However STACEY has so far mainly been tested on species delimitation problems. A couple of reasons for *not* using STACEY are that it does not allow you to use tip dates or calibrations, and it does not allow you to estimate individual per-branch population sizes.

STACEY stands for Species Tree And Classification Estimation, Yarely.

1.1 Model

When used for delimitation the species tree, the species tree has tips which represent minimal clusters of individuals (DISSECT paper). These minimal clusters may be merged but not split to form potential species. Thus 'species tree' is not a good name for this tree, and instead I will refer to it a the 'SMC-tree', as a shorthand for 'species or minimal clusters tree'.

In the multispecies coalescent model implemented in STACEY, it is assumed that each branch in the SMC-tree has a population parameter which is constant along the branch, and that these parameters are independent and identically distributed over branches. Instead of sampling these parameters, they are integrated out. The method caters for variation among branches, but does not allow individual populations to be estimated. The method is analogous to the common one for modeling site rate heterogeneity where it is assumed 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 birth-death-collapse model used by both STACEY and DISSECT allows species delimitations to be estimated.

1.2 Operators

STACEY has three new operators `NodesNudge`, `FocusedNodeHeightScaler`, and `CoordinatedPruneRegraft`. `NodesNudge` is described in detail in the STACEY paper.

`NodesNudge` changes height of a node in the SMC-tree and some nodes in the gene trees in such a way that all topologies are preserved and the gene trees always (in v1.0.0) remain compatible with the SMC-tree. The same (typically small) value is added or subtracted from all node heights affected.

`FocusedNodeHeightScaler` changes the height of many nodes in the species and gene trees. The scaling is 'focused' on a node in the SMC-tree. This node is scaled by the largest amount. The further away a node is from the focus (in a certain sense), the less it is affected by the move. All topologies are preserved and the gene trees always (in v1.0.0) remain compatible with the SMC-tree.

`CoordinatedPruneRegraft` makes coordinated topological changes to the species tree and gene trees. It makes a 'Fixed Nodeheight Prune and Regraft' (FNPR) move (Höhna et al., 2008), then a set of FNPR moves on each gene tree to maintion compatibility. It can be seen as an extension of the NNI move described in Yang and Rannala (2014).

1.3 Output

The output is similar to that of StarBEAST. There will be a tree file for each gene tree and the SMC-tree, and a log file for the numerical parameters. Since STACEY does not estimate individual populations in branches, there will be no demographic information in the SMC-tree file.

There are some new items that are available to log (and which will be logged by default). See 2.5 for more details.

- The logarithm of the probability of the STACEY multispecies coalescent model.
- The logarithm of the probability of the birth-death-collapse model.
- The overall population scaling factor.
- The birth-death-collapse parameters.
- The number of clusters.
- Samples from the per-branch population sizes.

2 Usage

2.1 Installation

STACEY can be installed using BEAUTi. Choose `File`→`Manage packages` to install, upgrade and uninstall packages. It seems necessary to restart BEAUTi before STACEY will work with BEAUTi.

If you can't see STACEY listed in the `Package Manager` window, you can download it from `http://www.indriid.com/software.html`, and install it manually. This means extracting the contents of the zip file (which has a name like `STACEY.addon.v1.0.4.zip`) into a directory called `STACEY` where BEAST can find it. This might be `<Your home directory>\BEAST\2.3`.

2.2 BEAUTi

The template supplied with v1.0.4 allows you to set up a STACEY analysis in BEAUTi. There are some rough edges, and a few things that cannot be edited in BEAUTi.

Start BEAUTi and choose `STACEY` from the `File->Template` menu. Do this before loading any alignments, or you'll have to reload the alignments. You should see a tab called `STACEY Coalescent` appear (with nothing in it).

Use `File->Import Alignments` to load the data as usual. Use the `Taxon sets` tab to divide the taxa into minimal clusters (the `Species/Population` column) as you would for StarBEAST.

Missing data. Currently (1.0.4) STACEY does not handle missing data as well as it should. You must ensure that for each minimal cluster, there is a sequence for every taxon in the minimal cluster. The sequence can consist of missing data, such as `-----` or `??????` but a sequence must be present. For example, if you have two sequences for some genes from an individual, and only one for others, you need to add sequences (of missing data, or a copy if the reason for the single sequence is homozygosity), so that there two sequences per individual for each gene. In XML, this situation would look like:

```
<taxon id="mincluster6" spec="TaxonSet">
  <taxon id="individual6_sequence1" spec="Taxon"/>
  <taxon id="individual6_sequence2" spec="Taxon"/>
</taxon>
```

Here, there must be a sequence in each alignment with `taxon="individual6_sequence1"` and another with `taxon="individual6_sequence2"`. A future version of STACEY is expected to relax this requirement.

Use the `Site Model` and `Clock Model` to set up the gene tree models as you would for StarBEAST.

In the `STACEY Coalescent` tab you should see a list of loci, labeled like `gTreeCF.t:locus_name`. You can set the ploidy values here for individual genes. The normal value is 2.0. Genes from sex chromosomes and organelles are different. You can't use tip dates with STACEY.
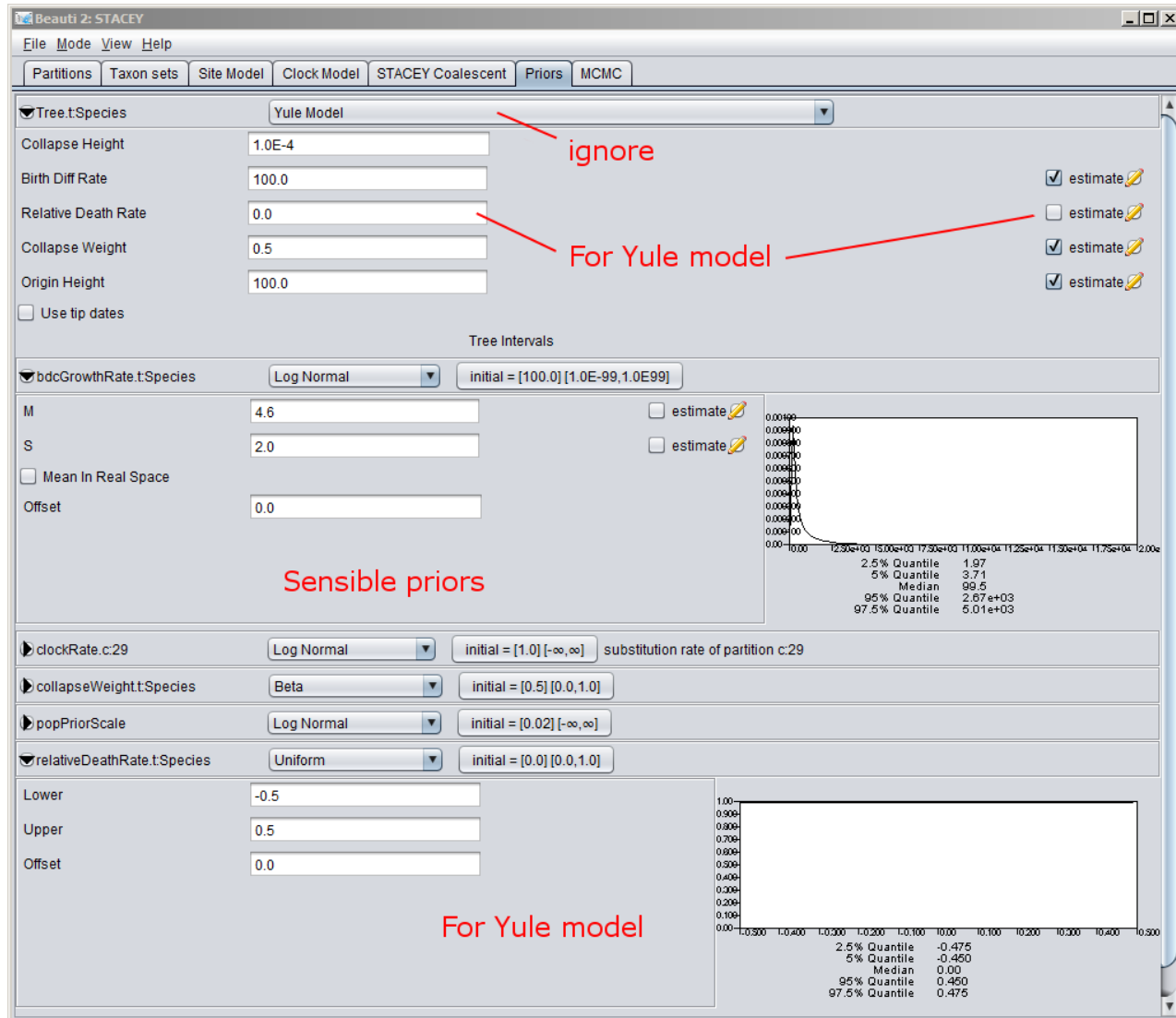

Figure 1: Screenshot of BEAUTi priors panel.

In the `Priors` tab (Figure 1) you will see a label `Tree.t:Species` and a drop down menu. Do *not* choose anything from this menu, or you'll have to start all over again. Click the arrow by `Tree.t:Species` and you will see text fields for `Collapse Height`, `Birth Diff Rate`, `Relative Death Rate`, `Collapse Weight`, and `Origin Height`. You can specify initial values in the text fields.

`Collapse Height` is denoted ϵ in the DISSECT paper. This is a computational approximation to zero. It has no biological meaning, and its value is simply a trade-off between speed and accuracy. Smaller is more accurate, but slower. I suggest using a value which is about 1/100 or 1/1000 of a typical species tree branch length. The value is not critical, in that a wide range of values (such as from 0.000001 to 0.0001) usually produce similar results in similar times.

`Birth Diff Rate` is the growth rate of the SMC-tree, often denoted $\lambda - \mu$.

4

`Relative Death Rate` is the ratio of the extinction rate to the speciation rate, often denoted μ/λ. If you want to use a Yule model, set the initial value to zero, and untick `estimate`. It is also necessary to set a prior that is nonzero at 0, for example, a uniform distribution on [-0.5,0.5].

`Collapse Weight` is denoted ω in the DISSECT paper. It can be estimated or fixed. It is between 0 and 1, and supplies prior information about the likely number of species. Values near 1 mean fewer species (more merging of minimal clusters) are expected. If w is estimated and given a uniform prior on [0,1], every number of species between 1 and the number of minimal clusters is regarded as equally likely a priori. If you want to use fixed species assignments, set the initial value to zero, and untick `estimate`, which will prevent any merging, and make the analysis very similar to a *BEAST analysis. Again, it is necessary to set a prior that is nonzero at 0.

`Origin Height` is the height of the origin of the SMC-tree, that is, the height of the parent of the root. It must be estimated, and the initial value is ignored.

Below are priors for these parameters (except `Origin Height` which has the prior described in the DISSECT paper). I recommend changing the improper `1/X` ones to something sensible.

There is also a prior for `popPriorScale`. It is denoted as σ in the STACEY paper. It should be possible to set the initial value in BEAUTi but currently it isn't. It is set to 0.02 (in the STACEY template), which should be fine for most analyses. If you want to change it, you will have to edit the XML produced by BEAUTi. Again, I recommend changing the improper `1/X` prior to something sensible. See subsection 2.5 for more on `popPriorScale`.

The `MCMC` tab and other tabs you can see via `View` work as usual.

2.3 Editing the XML

There are a few things which cannot be done in BEAUTi.

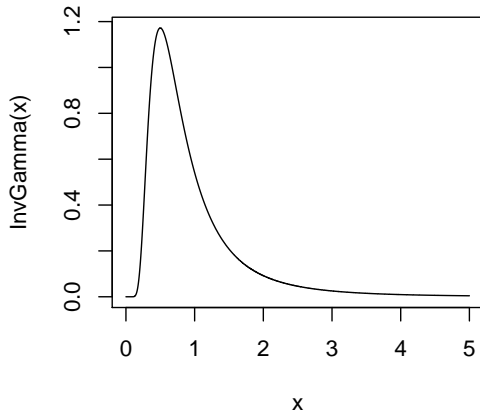### 2.3.1 Prior for per-branch variation of population

Figure 2: Inverse gamma distribution with $\alpha = 3.0$ and $\beta = 2.0$.

The multispecies coalescent model described in the STACEY paper, called the 'STACEY coalescent model' from now on, represents the per-branch variation of population parameters by a mixture of inverse gamma distributions. This mixture is a very flexible distribution and you can choose more or less what you want by choosing appropriate parameters (though choosing them is not very easy). The default is a single inverse gamma distribution which has

mean and variance 1 and looks like Figure 2. If you want a different distribution you will have to edit the XML. The relevant bit of XML looks like this:

```
<popPriorInvGamma id="InvGammaComponent.1" spec="stacey.InverseGammaComponent">
  <parameter name="weight" id="InvGammaComponentWeight.1" estimate="false">1.0</parameter>
  <parameter name="alpha" id="InvGammaComponentAlpha.1" estimate="false">3.0</parameter>
  <parameter name="beta" id="InvGammaComponentBeta.1" estimate="false">2.0</parameter>
</popPriorInvGamma>
```

You can change the `alpha` and `beta` values, and add other components `InvGammaComponent.2` etc. The `weight` values will be normalized to sum to 1.

2.3.2 Operator weights

BEAUTi 2.1.3 automatically assigns 20% of the total weight to operators which affect the SMC-tree, and the rest to gene tree operators. This seems reasonable for operators which *only* affect the SMC-tree, but the three new operators affect all trees, so a larger total percentage seems appropriate. By default, NodeReheight and the three new ones will get 5% each. I suggest increasing the weight for NodeReheight by a factor of about 3, and leaving the others as they are. In the code below, this means changing the 8.74817518248175 to about 25. This cannot be done in BEAUTi, since it rescales everything so the total is still 20%.

```
    <operator id="Reheight.t:Species" spec="NodeReheight" taxonset="@taxonsuperset"
                                 tree="@Tree.t:Species" weight="8.74817518248175">
        <tree idref="Tree.t:29" name="genetree"/>
        ...
    </operator>
```

For large data sets, it may be worth experimenting with operator weights (these and others) in order to improve the rate at which ESSs/hour can be obtained.

2.4 Examples

The examples folder contains three BEAST2 XML files `A3BXg2t10r1.xml`, `3BXr21.xml`, `Ug5t10-FromBeauti.xml`, and five `.nex` files `U2SQg5t10r1-1.nex` to `U2SQg5t10r1-5.nex`. All these contain simulated data, where the truth is known.

`A3BXg2t10r1.xml`. This is small test file which should take less than a minute to run. There are three true species `a,b,c`, two individuals per species, one sequence from each individual. The minimal clusters are the six individuals.

`3BXr21.xml` is one of the files used in the STACEY paper. It contains data from Olave et al. (2014). Here there are 8 true species, 5 individuals per species, two sequences from each individual and 14 loci (total 1120 sequences) for the 'hard' N=0.4 case with an asymmetric tree. It should take less than an hour to run.

The five `.nex` files can be used to practice with setting up a STACEY analysis in BEAUTi. In the `Taxon sets` tab you can use `Guess` with 'use everything before first _' to make a minimal cluster for each individual. `Ug5t10-FromBeauti.xml` is what I made with my choice of priors, etc.

The `.nex` files are 5 alignments simulated from the following scenario. The true SMC-tree is `((a:.001,b:.001):0.001,c:0.002);`. There are 5 individuals from each species, one sequence from each individual of length 500. The mutation rate in substitutions per site per generation is 1e-8 and the populations vary linearly along each branch from 69300 at tip to 34650 at root. This is very similar to a constant population of 50000.

2.5 Analyzing the results

The sampled SMC-trees that are generated by STACEY can be analyzed using `SpeciesDelimitationAnalyser` (DISSECT paper). When the results from `Ug5t10-FromBeauti.xml` were processed they looked like this

(abbreviated from `SpeciesDelimitationAnalyser` output):

```
fraction nclusters  b01  b05  b04  b03  b02  a01  a02  a03  a04  a05  c05  c04  c03  c02  c01
0.53     3          1    1    1    1    1    2    2    2    2    2    3    3    3    3    3
0.055    4          2    2    2    1    2    3    3    3    3    3    4    4    4    4    4
0.023    4          2    2    2    2    2    3    3    3    3    3    4    4    4    1    4
0.018    4          1    2    1    2    1    3    3    3    3    3    4    4    4    4    4
0.017    2          1    1    1    1    1    1    1    1    1    1    2    2    2    2    2
0.015    4          2    2    2    2    2    3    1    3    3    3    4    4    4    4    4
0.011    4          1    1    1    2    2    3    3    3    3    3    4    4    4    4    4
...
```

This shows a posterior probability of 0.53 for the true clustering, then posterior probability of 0.055 for splitting `b03` from the other `b`s and so on.

There are some new parameters and statistics which are written to the `tracelog`. These are `BirthDeathCollapseNClustersStatistic`, `BirthDeathCollapseModel.t:Species`, `smcCoalescent`, `popPriorScale`, `PopSampleStatistic`, and `originHeight.t:Species`.

`BirthDeathCollapseNClustersStatistic` is the most interesting one for species delimitation. It represents the number of species for this sample from the posterior. It is also written to screen.

`BirthDeathCollapseModel.t:Species` and `smcCoalescent` are log probability densities which contribute to the posterior. The first is for the birth-death-collapse prior for the SMC-Tree prior, and `smcCoalescent` is for the STACEY coalescent mode.

`popPriorScale` is the overall scaling factor for the population sizes. It is denoted as σ in the STACEY paper.

`PopSampleStatistic`. In the STACEY coalescent model, the population sizes are assumed to be `popPriorScale` multiplied by a value drawn from a mixture of inverse gammas. At each generation in the MCMC where a value is required, this statistic draws a random value from the mixture and multiplies this by `popPriorScale`. If the mean of the mixture of inverse gammas is 1, and the distribution of this mixture is not very skewed (both of which are true in the default mixture), this will usually be a similar value to `popPriorScale`. In the general case, `PopSampleStatistic` is the most easily understood value. It samples from the posterior distribution of $N_b\mu_b$, over all branches b. Here N_b is the effective population for branch b and μ_b is the mutation rate in substitutions per site per generation along branch b.

`originHeight.t:Species` is not usually estimated in phylogenetics. It is here because it forms part of the birth-death-collapse model (DISSECT paper). Somebody might find it useful.

3 Changes from previous versions

3.1 Changes in v1.0.4

- Manual updated, especially about missing data.

3.2 Changes in v1.0.3

- Failure to make sense of taxon names now gives error message, instead of null pointer exception. (The usual cause was having a Taxon and a Species/Population with identical name in Beauti.)
- Removed checks in `BirthDeathCollapseModel.initAndValidate()` for collapse weight and relative death rate. (The checks failed if the parameters were not estimated, I don't know why.)
- Fixed bug that prevented resume from working.
- New template for BEAST 2.3.0 (thanks to Remco).
- Minor updates to manual.

4 StarBEAST XML vs STACEY XML

I hope most of you won't need to read this section. You may find it useful if you start editing XML directly. It describes the differences between the XML for *BEAST and STACEY, when STACEY is used for delimitation. When STACEY is used for delimitation, the species tree of *BEAST is replaced by a tree in which the tips are minimal clusters. It is called an 'SMC-tree' ('species or minimal clusters tree'), and is denoted `smcTree` in the example XML files below.

4.1 <state> element

The birth-death-collapse hyper-parameters (namely `bdcGrowthRate`, `bdcRelativeDeathRate`, `bdcCollapseWeight`, and `bdcOriginHeight`) for `smcTree` replace those for a species tree prior such as the Yule prior. `popPriorScale` replaces `popMean`.

The initial value of the growth rate may be used for initializing trees, and as a time scale in operators in future versions, so a sensible value is recommended. The initial value of origin height is ignored: it is set automatically.

```
<parameter id="bdcGrowthRate.smcTree" name="stateNode" lower="1.0E-99" upper="1.0E99" >100.0</parameter>
<parameter id="bdcRelativeDeathRate.smcTree" name="stateNode" lower="0.0" upper="1.0" >0.5</parameter>
<parameter id="bdcCollapseWeight.smcTree" name="stateNode" lower="0.0" upper="1.0" >0.5</parameter>
<parameter id="bdcOriginHeight.smcTree" name="stateNode" lower="1.0E-99" upper="1.0E99" >100.0</parameter>
<parameter id="popPriorScale" name="stateNode" lower="1.0E-9" upper="1.0E9" >0.02</parameter>
```

4.2 <init> element

The `init` element provides the initial value for `smcTree` and for all the gene trees. These are trees with random topologies, not the UPGMA trees used in the `SBI` element for *BEAST. The SMC-tree is scaled so its root has the height given by `rootHeight`, and all the gene trees have this value added to their internal node heights to make them compatible with the SMC-tree. Thus initially all coalescences occur in the root of the SMC-tree. It is important that `rootHeight` is given a fixed value so that the birth-death-collapse model can set a origin height that is bigger than this.

```
<init id="randomSpeciesTree" spec="RandomTree" estimate="false" taxonset="@taxonSetOfSets"
                                            initial="@smcTree" rootHeight="0.05" >
  <populationModel id="ConstantPopulation.smc" spec="ConstantPopulation" >
    <parameter id="randomPopSize.smc" name="popSize" >0.05</parameter>
  </populationModel>
</init>
<init id="randomGeneTree.1" spec="RandomGeneTree" estimate="false" taxonset="@geneTaxonSet.1"
                                          initial="@geneTree.1" speciesTree="@smcTree" >
  <populationModel id="RGTPopulationModel.1" spec="ConstantPopulation" >
    <parameter id="RGTPopSize.1" name="popSize" >0.05</parameter>
  </populationModel>
</init>
<init id="randomGeneTree.2" spec="RandomGeneTree" estimate="false" taxonset="@geneTaxonSet.2"
                                          initial="@geneTree.2" speciesTree="@smcTree" >
  <populationModel id="RGTPopulationModel.2" spec="ConstantPopulation" >
    <parameter id="RGTPopSize.2" name="popSize" >0.05</parameter>
  </populationModel>
</init>
```

4.3 <distribution> element

The `distribution` element contains the central component of the model in STACEY, namely the multispecies coalescent distribution in which population parameters integrated out. It contains references to `smcTree` and all the gene trees, the mixture of inverse gamma distributions used as a prior for the per-branch variability of population sizes, and a reference to the overall population scaling factor σ. It replaces the *BEAST multispecies coalescent distribution (the one with id `speciescoalescent` and a `GeneTreeForSpeciesTreeDistribution` for each gene tree).

```
<distribution id="smcCoalescent" spec="stacey.PIOMSCoalescentDistribution" tree="@smcTree" >
  <geneTree id="gTreeCF.1" spec="stacey.GtreeAndCoalFactor" tree="@geneTree.1" coalFactor="2.0" />
  <geneTree id="gTreeCF.2" spec="stacey.GtreeAndCoalFactor" tree="@geneTree.2" coalFactor="2.0" />
  <popPriorInvGamma id="InvGammaComponent.1" spec="stacey.InverseGammaComponent"
                                             weight="1" alpha="3" beta="2" />
  <popPriorScale idref="popPriorScale"/>
</distribution>
```

The `distribution` element also contains the `BirthDeathCollapseModel` which replaces a species tree prior such as the Yule model, with a prior for `smcTree` allowing delimitations to be inferred.

```
<distribution id="prior" spec="util.CompoundDistribution" >
<distribution id="BirthDeathCollapseModel" spec="stacey.BirthDeathCollapseModel" collapseHeight="1e-04" >
  <tree idref="smcTree"/>
  <parameter idref="bdcGrowthRate.smcTree" name="birthDiffRate"/>
  <parameter idref="bdcRelativeDeathRate.smcTree" name="relativeDeathRate"/>
  <parameter idref="bdcCollapseWeight.smcTree" name="collapseWeight"/>
  <parameter idref="bdcOriginHeight.smcTree" name="originHeight"/>
</distribution>
```

The `distribution` element also contains priors for the hyperparameters associated with the birth-death-collapse model, and for the overall population scaling factor.

```
<prior id="bdcGrowthRate.smcTreePrior" name="distribution" x="@bdcGrowthRate.smcTree" >
  <LogNormal name="distr" id="LogNormal.bdcGrowthRate.smcTreePrior" >
    <parameter name="M" id="mean.bdcGrowthRate.smcTreePrior" estimate="false" >4.6</parameter>
    <parameter name="S" id="sd.bdcGrowthRate.smcTreePrior" estimate="false" >2</parameter>
  </LogNormal>
</prior>
<prior id="bdcRelativeDeathRate.smcTreePrior" name="distribution" x="@bdcRelativeDeathRate.smcTree" >
  <Beta name="distr" id="Beta.bdcRelativeDeathRate.smcTreePrior" >
    <parameter name="alpha" id="alpha.bdcRelativeDeathRate.smcTreePrior" estimate="false" >3</parameter>
   <parameter name="beta" id="beta.bdcRelativeDeathRate.smcTreePrior" estimate="false" >1</parameter>
  </Beta>
</prior>
<prior id="bdcCollapseWeight.smcTreePrior" name="distribution" x="@bdcCollapseWeight.smcTree" >
  <Beta name="distr" id="Beta.bdcCollapseWeight.smcTreePrior" >
    <parameter name="alpha" id="alpha.bdcCollapseWeight.smcTreePrior" estimate="false" >8</parameter>
    <parameter name="beta" id="beta.bdcCollapseWeight.smcTreePrior" estimate="false" >2</parameter>
  </Beta>
</prior>
<prior id="popPriorScalePrior" name="distribution" x="@popPriorScale" >
  <LogNormal name="distr" id="LogNormal.popPriorScalePrior" >
    <parameter name="M" id="mean.popPriorScalePrior" estimate="false" >-7</parameter>
    <parameter name="S" id="sd.popPriorScalePrior" estimate="false" >2</parameter>
  </LogNormal>
</prior>
```

4.4 <operators> element

The `operators` element contains the new `NodesNudge` operator.

```
<operator id="nodesNudge" spec="stacey.NodesNudge" weight="140" >
  <smcTree idref="smcTree"/>
  <tree idref="geneTree.1" name="geneTree"/>
  <tree idref="geneTree.2" name="geneTree"/>
</operator>
```

In the following `UpDownOperator`, the `popPriorScale` parameter is added to the usual ones.

```
<operator id="upGrowthClocks.downPopsHeights" spec="UpDownOperator" weight="40" scaleFactor="0.75" >
  <parameter idref="bdcGrowthRate.smcTree" name="up"/>
  <parameter idref="clockRate.2" name="up"/>
  <parameter idref="popPriorScale" name="down"/>
  <tree idref="smcTree" name="down"/>
  <tree idref="geneTree.1" name="down"/>
  <tree idref="geneTree.2" name="down"/>
</operator>
```

The `operators` element also contains operators for the various hyper-parameters.

```
<operator id="popSFScaler" spec="ScaleOperator" scaleFactor="0.75" weight="7" >
  <parameter idref="popPriorScale"/>
</operator>
<operator id="bdcGrowthScaler" spec="ScaleOperator" scaleFactor="0.75" weight="7" >
  <parameter idref="bdcGrowthRate.smcTree"/>
</operator>
<operator id="bdcReldeathScaler" spec="ScaleOperator" scaleFactor="0.75" weight="7" >
  <parameter idref="bdcRelativeDeathRate.smcTree"/>
</operator>
<operator id="bdcCollapseWtScaler" spec="ScaleOperator" scaleFactor="0.75" weight="7" >
  <parameter idref="bdcCollapseWeight.smcTree"/>
</operator>
<operator id="bdcOriginHtScaler" spec="ScaleOperator" scaleFactor="0.75" weight="7" >
  <parameter idref="bdcOriginHeight.smcTree"/>
</operator>
```

4.5 <loggers> element

New items available to log are the logarithm of the probabilities of the multispecies coalescent model and the birth-death-collapse model; the overall population scaling factor; and the birth-death-collapse parameters.

```
<log idref="smcCoalescent"/>
<log idref="popPriorScale"/>
<log idref="BirthDeathCollapseModel"/>
<parameter idref="bdcGrowthRate.smcTree" name="log"/>
<parameter idref="bdcRelativeDeathRate.smcTree" name="log"/>
<parameter idref="bdcCollapseWeight.smcTree" name="log"/>
<parameter idref="bdcOriginHeight.smcTree" name="log"/>
```

There are also two statistics. The first provides the number of clusters. The other statistic is from the distribution of per-branch population sizes.

```
<log spec="stacey.BirthDeathCollapseNClustersStatistic" bdcm="@BirthDeathCollapseModel"
                                                   spptree="@smcTree" />
<log spec="stacey.PopSampleStatistic" popPriorScale="@popPriorScale"
                            piomsCoalDist="@smcCoalescent" />
```

References

Sebastian Höhna, Michael Defoin-Platel, and Alexei J Drummond. Clock-constrained tree proposal operators in Bayesian phylogenetic inference. *8th IEEE International Conference on BioInformatics and BioEngineering, 8–10 October 2008, Athens, Greece*, pages 1–7, 2008.

G Jones. Species delimitation and phylogeny estimation under the multispecies coalescent. *bioRχiv*, 2015. doi: 10.1101/010199. URL http://biorxiv.org/content/early/2015/03/22/010199.

Graham Jones, Zeynep Aydin, and Bengt Oxelman. DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics*, 2014. doi: 10.1093/bioinformatics/btu770.

Melisa Olave, Eduard Solà, and L Lacey Knowles. Upstream analyses create problems with DNA-based species delimitation. *Syst Biol*, 63:263–271, 2014. doi: 10.1093/sysbio/syt106.

Z Yang and B Rannala. Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, 31(12):3125–3135, 2014. doi: 10.1093/molbev/msu279.