STACEY package documentation: species delimitation and species tree estimation with BEAST2

Graham Jones, www.indriid.com

2014-10-08, February 4, 2016

STACEY version 1.1.1

Contents

1 Introduction

Brief overview. The BEAST2 package STACEY can be used for species delimitation and species tree estimation, based on the multispecies coalescent model. It incorporates a model for the per-branch population parameters which is simpler than the one in StarBEAST, but allows these parameters to be integrated out. 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').

What it is good for. 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 StarBEAST for species tree estimation for the same reason, especially when there are many loci. It *may* be that the new operators work well with StarBEAST (ignoring the rest of STACEY). 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.

This manual covers the basics. There is more advanced information in the appendices. In particular, if you are looking for advice on choosing priors, or if you are using STACEY on larger data sets (say, the total number of sequences is a 1000 or more), or are having trouble obtaining convergence in a reasonable time, read the appendices too.

STACEY stands for Species Tree And Classification Estimation, Yarely.

1.1 Model

The basic model of the evolutionary process used by STACEY is the same as StarBEAST (Heled and Drummond, 2010), namely the multispecies coalescent. The birth-death-collapse model used by both STACEY and DISSECT is an extension of this which allows species delimitations to be estimated.

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.

1.2 Operators

STACEY has five 'big' operators `StaceyNodeReheight`, `NodesNudge`, `FocusedNodeHeightScaler`, `CoordinatedPruneRegraft`, and `ThreeBranchAdjuster` which change the shape of the SMC-tree. Apart from `StaceyNodeReheight` they also change some or all gene trees. The other operators will be familiar to StarBEAST users.

`StaceyNodeReheight` is a replacement for NodeReheight as found in StarBEAST. The implementation is faster, especially for large numbers of minimal clusters, and it samples a new height for the SMC-tree node from a density which is skewed towards the maximum compatible height. This improves the acceptance ratio when there are large numbers of loci. See the appendices for more details.

`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 v1.1.1) remain compatible with the SMC-tree. The same

(typically small) value is added or subtracted from all node heights affected. See the STACEY paper for more details.

`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 v1.1.1) remain compatible with the SMC-tree. See the STACEY paper for more details.

`CoordinatedPruneRegraft` makes coordinated topological changes to the SMC-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). See the STACEY paper for more details.

`ThreeBranchAdjuster` is an implementation of Rannala and Yang (2003). It changes the height of one SMC-tree node, and the heights of all gene tree nodes within the the three adjacent branches. See the appendices for more details.

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.v1.1.1.zip`) into a directory called `STACEY` where BEAST can find it. This might be `<Your home directory>\BEAST\2.3`.

2.2 BEAUTi

There are two templates supplied with v1.1.1. The one called `StarBeastWithSTACEYops` is for using StarBEAST, not STACEY, but it uses the five 'big' operators from STACEY instead of the `NodeReheight` operator in StarBEAST. In theory, `StarBeastWithSTACEYops` should work just like StarBEAST, but faster, especially when there are lots of loci. It has not been much tested.

The rest of this section is about the `STACEY` template which 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. In 1.1.1 it is no longer necessary to have a sequence for every taxon. However, this is a new feature which has not been tested much, so use with caution. You still need to have at least one sequence in each minimal cluster for each locus. Some sequences can consist of explicitly missing data, such as `-----` or `??????`. For reference, the situation in 1.0.5 is below.

> Currently (1.0.5) 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"`.

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.

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 $\lambda - \mu$.

`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 ω 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 StarBEAST 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.

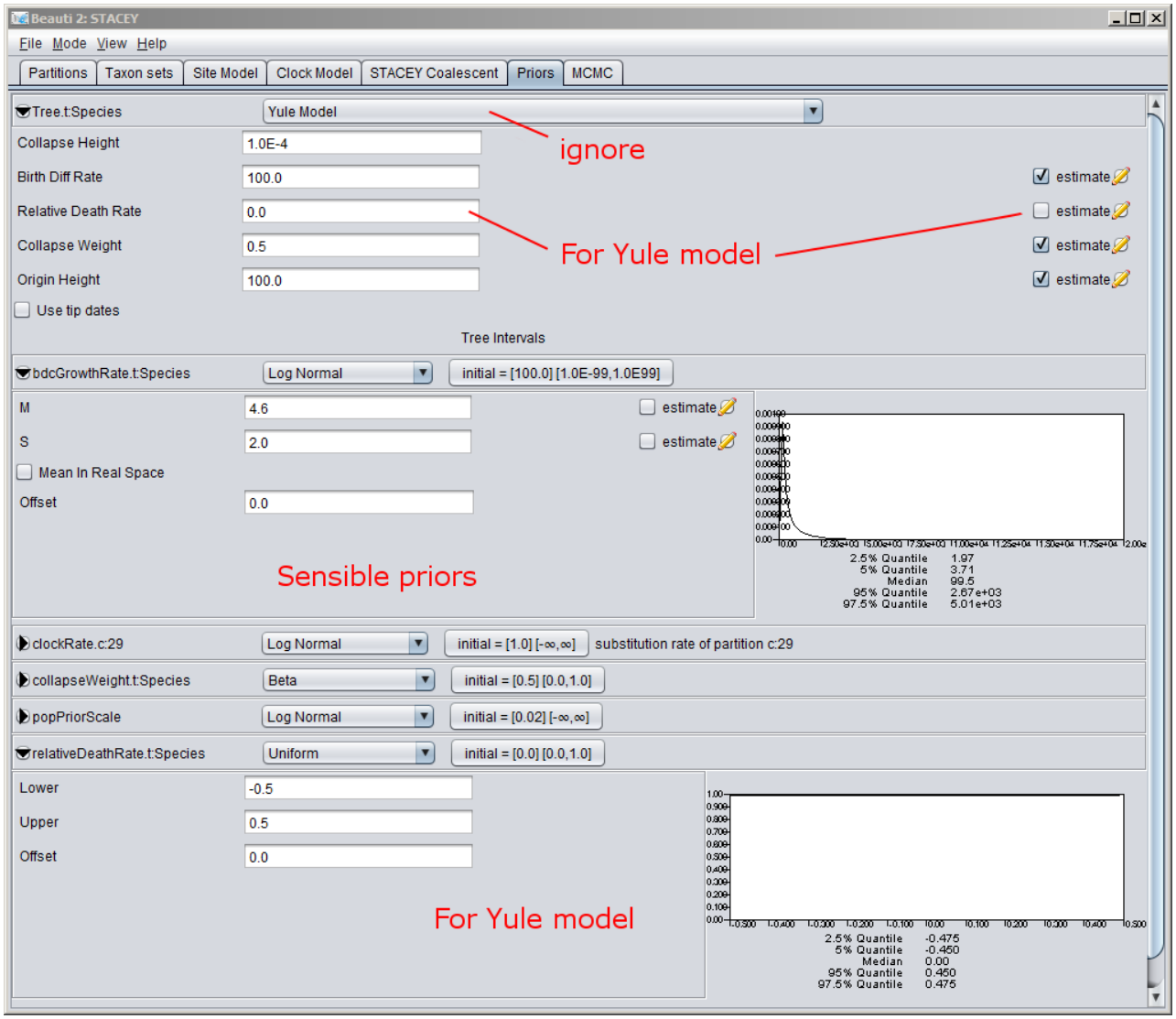

Figure 1: Screenshot of BEAUTi priors panel.

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. See the appendices for advice on choosing priors.

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. See the appendicies for more discussion on convergence and operator weights.

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

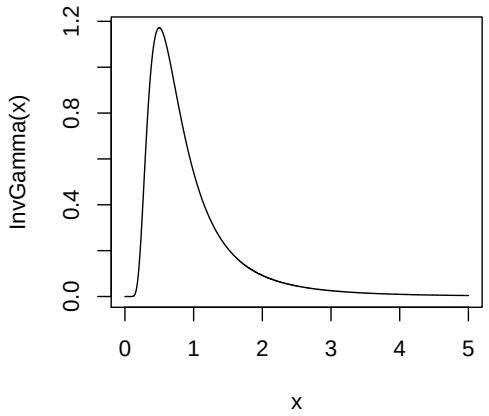

Figure 2: Inverse gamma distribution with shape parameter 3.0 and rate parameter 2.0.

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. Each component of the mixture has a density like

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

Changing α changes the shape of the density. Increasing α makes the density more concentrated, and decreasing α makes it more diffuse. For large x, the $\exp(-\beta x^{-1})$ term becomes very close to 1, so the tail of the distribution is proportional to $x^{-\alpha-1}$. This means that for $\alpha \leq 1$, the density has infinite mean, and for $\alpha \leq 2$ it has infinite variance. The β parameter is a scale parameter, so increasing β 'stretches' the density away from zero. For $\alpha > 1$, the mean is $\beta/(\alpha - 1)$.

2.4 Examples

The examples folder contains five `.nex` files `U-g5t10-1.nex` to `U-g5t10-5.nex` and a BEAST XML file `Ug5t10-FromBeauti.xml` made from this data. The nex files contain simulated data, where the truth is known. 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.

These files can be used to practice with setting up a STACEY analysis in BEAUTi. Here is a specific example. In the `Taxon sets` tab use `Guess` with 'use everything before first _' to make a minimal cluster for each individual. In the `Site Model` tab select the first locus `U-g5t10-1` and change model from JC69 to HKY. Then select all the loci (with shift-click) and clone from `U-g5t10-1`. In the `Priors sets` tab, change the growth rate prior to a lognormal with M=5 and S=2. (M and S are the mean and sd in log space.) Change the four clock rate priors to lognormals with M=0 and S=1. Change the PopPriorScale prior to a lognormal with M=-7 and S=2. Save and run the BEAST XML file.

2.5 Analyzing the results

The sampled SMC-trees that are generated by STACEY can be analyzed using `SpeciesDelimitationAnalyser` (DISSECT paper) which can be downloaded from `http://www.indriid.com/software.html`. When the results from `Ug5t10-FromBeauti.xml` were processed using this command

```
java -jar speciesDA.jar -burnin 200 -collapseheight .0001 -simcutoff 1
```

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 v1.1.1

This is a major update.

- The implementation has been speeded up. This affects the calculation of the multispecies coalescent, and the operators. It is especially beneficial for large numbers of loci.
- `StaceyNodeReheight` replaces `NodeReheight`. It has a faster implementation, especially for large numbers of minimal clusters. It has a non-uniform density for new node heights which has a better acceptance ratefor large numbers of loci.
- New `ThreeBranchAdjuster` operator.
- Better handling of missing data.
- XML change: Log file names no longer contain the random seed. The inclusion of the seeds was aimed at avoiding overwriting old log files if two analyses were done in the same directory. However it complicates resuming. It is recommended to use different directories for different analyses, to avoid overwriting and other confusion.
- Substantial changes to manual.

3.2 Changes in v1.0.5

- Bug fixed: If there were exactly 4 minimal clusters, the `FocusedNodeHeightScaler` could (and usually did) go into an infinite loop.
- Made some improvements to the code suggested by Findbugs (see `http://findbugs.sourceforge.net/`).
- Manual updated: operator weights, v1.1.x.

3.3 Changes in v1.0.4

- Manual updated, especially about missing data.

3.4 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.

References

J Heled and A Drummond. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27:570–580, 2010.

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.

B Rannala and Z Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164:1645–1656, 2003.

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.