DENIM package documentstion

Graham Jones

2017-06-14, March 21, 2019

DENIM version 1.0.0

Contents

1 Introduction

The BEAST2 package DENIM can be used for species tree estimation in the presence of small amounts of migration. DENIM stands for Divergence Estimation Notwithstanding ILS and Migration. It is a development from the STACEY package, but note that the operators which were designed for STACEY do not work when there is migration.

Migration is a difficult problem to deal with, especially when the species tree is unknown. There is no upper limit on the number of migrations that might have occured. DENIM uses an approximation which limits this number. If the migration rates are high, the approximation will break down. Looking at this the other way, you could say that all phylogenetic analyses which ignore migration when it is present involve a worse approximation.

The paper describing DENIM is at
`https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy041/5003138`
The paper on STACEY (Jones, 2016) is also very relevant. Also see my website `http://indriid.com/` for other information, especially recent working notes, and the software page `http://indriid.com/software.html`

1.1 Model

The underlying evolutionary model is similar to the of IMa2 (Hey, 2010), except that the species tree follows a birth-death model instead of being assumed. The time periods between the species tree node heights are 'epochs'. Within an epoch the number of branches is constant. If it is n, there is a n-island model between any pair of branches for migration. A pair of branches b and d can both exist for one or more epochs, and if they do, there are two parameters m_{bd} and m_{db} for the migration rates in both directions. They are assumed constant while b and d exist, as in IMa2.

The implementation of DENIM differs from IMa2 in that both population size parameters and migration rate parameters are integrated out, and DENIM uses the approximation mentioned above. DENIM is on the phylogenetics side, and IMa2 is on the population genetics side, of the boundary between the fields.

1.2 Output

As well as the usual gene trees, species tree, and parameters for the site and clock models, and so on, DENIM produces some extra information about migrations. The simplest is a count of the number of migrations within each gene tree. This may identify which loci have migrated. There is more detail about individual migrations in the species tree, as annotations. These can be analysed using a command-line tool called `MigrationAnalyser`.

2 Usage

2.1 Installation

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

2.2 Setting up an analysis with BEAUTi

Start BEAUTi and choose `DENIM` from the `File->Template` menu. Do this before loading any alignments, or you'll have to reload the alignments. You should see a two tabs called `Ploidy` (with nothing in it until you load an alignment) and `Coalescence+Migration` appear.

2.2.1 Data

Use `File->Import Alignments` to load the data as usual. Apart from the `Coalescence+Migration` panel, the other panels work the same way as STACEY, and are similar to *BEAST and StarBEAST2. Use the `Taxon sets` tab to divide the taxa into species (the `Species/Population` column) as you would for StarBEAST.

Missing data. DENIM should work with missing data, but this has not been tested much, so use with caution. In any case, you need to have at least one sequence in each species for each locus. Some sequences can consist of explicitly missing data, such as `-----` or `??????`. If there is a problem with missing data, this provides a fallback: you can add explicitly missing data to fill the gaps.

2.2.2 Site Model and Clock Model

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

2.2.3 Ploidy panel

This allows you to set the ploidy for each locus. This is 2.0 for autosomal nuclear DNA from diploid species, but different for plastids and sex chromosomes. Ignore the rest of the settings in this panel. Tip dates cannot be used with DENIM. The Migration points and Destination choices are parameters for how the gene tree is embedded, but there is nothing useful to be done with them here.

2.2.4 Coalescence+Migration panel

`InvGammaComponent.1` is the prior which models the branch-to-branch variation in the population size parameters, as in STACEY. Click the pencil icon to edit the hyperparameters for this.

`popPriorScale`. The multispecies coalescent model has population size parameters for each branch in the species tree. These are values for $N_e\mu$ for each branch, where N_e is the effective population size and μ is the mutation rate in substitutions per site per generation. These cannot be estimated individually in DENIM, but still require a prior. The prior for the population size parameters is made of two parts. Firstly, there is `popPriorScale` which is the overall scaling factor. It is denoted as σ in the STACEY paper. The actual population size parameters for the species tree branches are `popPriorScale` multiplied by values drawn independently for each branch from a second distribution. This second distribution represents the variation between branches.

`GammaComponent.1` is the prior for the migration rates. Click the pencil icon to edit the hyperparameters for this. The `Weight` is ignored in DENIM 1.0.0. The values for `Shape` and `Scale` control the amount of migration.

DENIM has two models for migration rates, which I call 'simple' and 'flexible' (see the DENIM paper).

In the simple one, there is the same rate m for every pair of contemporaneous branches. To use the simple migration model, set both `Relatedness Factor` and `Migration Decay Scale` to a negative value. In this case, m

is drawn from the gamma prior `GammaComponent.1`. Because the approximation used by DENIM breaks down with a lot of migration, normally the mean of this prior should be very small, perhaps around 0.001. This is a matter for experimentation. (In the Leaché data, values were 0.001, 0.01, 0.1, 1.0, and applied to one or a few of the contemporaneous pairs, not all.)

The flexible model for migration rates has two migration rate parameters for each pair of contemporaneous branches. In a species tree with n tips, there are $2(n-1)^2$ such parameters. These are integrated out by assuming they are drawn independently from the gamma prior `GammaComponent.1`. Ahgain the mean should normally be very small. In the flexible model, you can use different priors for different pairs of contemporaneous branches. There two ways to control how the migration rates vary over the tree.

`Relatedness Factor`. This is a number x between 0 and 1. (Values above 1.0 make no biological sense, but DENIM will not stop you doing this.) The migration rate between sister branches will be drawn from the gamma distribution `GammaComponent.1`. For nonsisters, if the number if branches separating the rootward ends of two branches is i, the migration rate between then will be drawn from a gamma distribution with a scale that is x^i smaller than `GammaComponent.1`. So sister branches have $i = 0$ and $x^i = 1$, an uncle-nephew pair of branches have $i = 1$ and are scaled by x, and so on. See Figure 1. A reasonable value for `Relatedness Factor` might be 0.5.

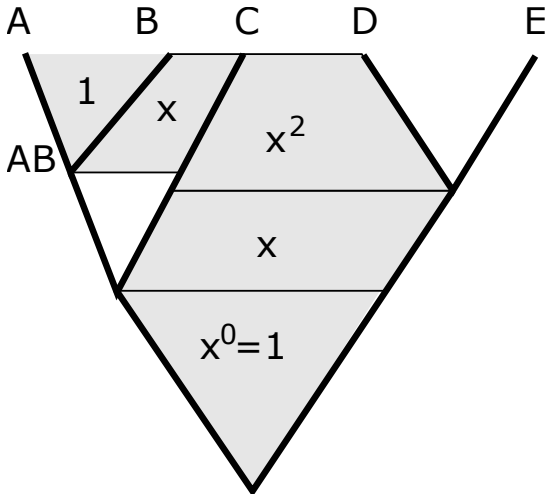

Figure 1: Relatedness Factor x. The diagram shows examples of how x affects migration rates between AB and C, and between C and D.

`Migration Decay Scale`. This is a value x in substitution units. Here, the idea is that migration rates decay exponentially with time, and x is the scale at which this happens. After a time t, the rate is e^{-xt} what it was at divergence. Since DENIM uses constant migration rates between pairs of contemporaneous branches, this idea is implemented by using the time between the most recent common ancestor of the two branches and the midpoint of the time during which both branches exist. See Figure 2. A reasonable value for `Migration Decay Scale` might be 0.01.

You can use both `Relatedness Factor` and `Migration Decay Scale` together.

If you want use the more flexible model, but to treat all the pairs of contemporaneous branches in the same way, you can set `Migration Decay Scale` to a negative value, and `Relatedness Factor` to 1.0.

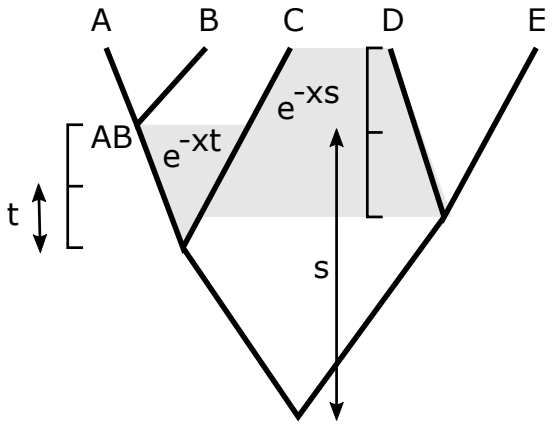

Figure 2: Migration Decay Scale x. The diagram shows examples of how x affects migration rates between A and B; B and C; C and D; C and DE; and ABC and DE.

2.2.5 Priors panel

Ignore the drop-down menu showing `Yule Model`, and click the arrow by `Tree.t:Species`.

For reasons given in subsection 2.4 you should normally make the delimitation fixed. To do this, set the initial `Collapse Weight` to 0, do not estimate it, and change the default prior to have an offset of -0.5. The latter is so that the values of 0 has a nonzero probability in the prior: if you don't do this, you will get a 'Failed to initialise...' error. The `Collapse Height` is irrelevant for fixed delimitations.

Other priors can be set in the usual way. It is strongly advised to change improper priors, such as `Uniform` on $[0, \infty]$ and `1/X` to reasonable ones. For example you might use a `Log Normal` with `M=5, S=2` for `bdcGrowthRate.t:Species`, and a `Log Normal` with `M=-7, S=2` for `popPriorScale`.

2.2.6 Other tabs

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

2.3 Analysing results

The species tree and other parameters can be analysed using TreeAnnotator and Tracer as usual. There are statistics giving the posterior mean of the numbers of migrations for each locus called `MigCount-gTreeWE.t:xxx` where `xxx` is the ID of the alignment.

DENIM puts some annotations in the `species.trees` file, which give more detail about the migrations. These can be analysed using the command line program `MigrationAnalyser.jar` which can be downloaded from `http://indriid.com/software.html`. There are two arguments you can use. The first is `-burnin <percentage>` The default is no burnin. A simple use would be:

```
java -jar MigrationAnalyser.jar -burnin 10 species.trees output.txt
```

The second argument is `-outputtype [raw | clades]`. The default is `clades`.

2.3.1 MigrationAnalyser outputtype clades

This produces a summarised output like this:

```
tipLabels 4
0 a
1 b
2 c
3 d
alignmentIDs 10
0 locus8
1 locus5
2 locus2
3 locus3
4 locus7
5 locus6
6 locus1
7 locus4
8 locus10
9 locus9
count locus outClade inClade meanHeight
1999 6 {0, , , } { ,1, , } 4.154223341223918E-4
8 0 { , ,2, } { , , ,3} 0.03126845252011613
8 6 { , ,2, } {0,1, , } 0.03617963689366989
6 6 {0,1, , } { , ,2, } 0.03060420473613731
```

The `tipLabels` and the `alignmentIDs` are from the `species.trees` file. DENIM adds the `alignmentIDs` (which come from the BEAST XML file) to the `species.trees` file as a NEXUS comment, so that `MigrationAnalyser` can use them.

`1999 6 {0, , , } { ,1, , }` can be translated as
'1999 times, a migration occured in 6=locus1 from clade 0=a to clade 1=b'.

`8 6 { , ,2, } {0,1, , }` can be translated as
'8 times, a migration occured in 0=locus8 from clade 2=c to clade 0,1=(a,b)'.

The coalescence and migration process is modeled going backwards in time from present, so the outClade contains the more recent part of a gene tree branch, and the inClade contains the more ancient part.

2.3.2 MigrationAnalyser outputtype raw

It is more flexible to use `-outputtype raw`, but it does mean extra work afterwards. I used this method for the analyses in the paper (using R). `-outputtype raw` produces a larger output with every migration in every sampled species tree:

```
treeIndex locus outClade inClade height
1 0 {0, , , , , , , } { , , , , , , ,7} 0.0034784822840568962
1 1 {0, , , , , , , } { , , ,3, , , , } 0.001544563589387454
1 3 {0, , , , , , , } { , , , , , , ,6, } 0.003631403775616456
[...]
```

```
1855 5 { , , , ,4, , , } { , ,2, , , , , } 5.107440609744086E-4
1855 5 { , , , ,4, , , } {0, , , , , , , } 0.003983445162299825
1855 6 { , , , ,4, , , } { ,1, , , , , , } 1.8938260107405173E-4
1855 6 { , , , ,4, , , } { , ,2, , , , , } 0.003156275043390818
1855 0 { ,1, , , , , , } { , , , ,4, , , } 1.5194414859408747E-4
[...]
```

At `http://indriid.com/workingnotes2018.html` there are R scripts that I used for analysing this type of output from MigrationAnalyser.

2.4 Species delimitation

DENIM incorporates the birth-death-collapse model of Jones et al. (2015), which means that you can, in principle, estimate the species delimitation and species tree in the presence of migration. As mentioned in the introduction, the STACEY operators cannot be used so it is likely to be much slower than STACEY. With both unknown delimitation, and migration, there may be many ways of explaining the data, and the posterior could be extremely diffuse and difficult to explore. It is not yet known under what cicumstances DENIM is capable of producing useful results in this case. Tests with simulated data are needed before results could be trusted.

3 Changes from previous versions

Changes in v1.0.0

- Logging now works for more than 2 billion samples. v1.0.0 requires BEAST 2.5.
- More information about MigrationAnalyser in manual, section 2.3.

Changes in v0.3.1

- Citation updated.
- Manual updated.
- No "(beta)" in version number (to stop Beauti warnings).

References

J Hey. Isolation with migration models for more than two populations. *Mol Biol Evol*, 27:905–920, 2010.

Graham Jones. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of Mathematical Biology*, 74:447–467, 2016. doi: 10.1007/s00285-016-1034-0. URL `http://link.springer.com/article/10.1007/s00285-016-1034-0`.

Graham Jones, Zeynep Aydin, and Bengt Oxelman. Dissect: an assignment-free bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics*, 31(7):991–998, 2015. doi: 10.1093/bioinformatics/btu770. URL + `http://dx.doi.org/10.1093/bioinformatics/btu770`.