How to use the DENIM package for tests

Graham Jones

2017-03-06, March 7, 2017

Introduction

DENIM is aimed at species tree estimation using the multispecies coalescent, in the presence of some migration. It is based on an approximation which will break down if there is a lot of migration. Leaché et al. (2014) showed migration causes problems for species tree inference using the multispecies coalescent when migration is present but ignored.

'Migration' is used to refer to gene flow between species (usually introgression but not restricted to that). A migration event occurs when an allele comes from a parent in another species. An 'embedding' of a gene tree specifies which species tree branch each coalescence belongs to, together with migration events, which specify the times along gene tree branches at which an allele moved between species tree branches, and which species tree branch is the destination. We always describe events going back in time from the present, so alleles have parents to which they 'go', and a destination branch refers to an earlier time than a source branch. This is because coalesences are easier to model this way, and is the same convention as IMa2 (Hey and Nielsen, 2004, 2007; Hey, 2010).

There is no upper limit on the number of possible migration events, and it is difficult to model the situation in full. IMa2 requires that the true population phylogeny (equivalent to species tree here) is known. The method of Dalquen et al. (2016) can estimate the species tree, but is currently restricted to at most 3 species and 3 sequences per locus, although it can handle very large numbers of loci.

Here we make two simplifying assumptions. Firstly, we use a simpler model for migration than IMa2 or Dalquen et al. (2016). The aim is to have a prior density for a gene tree, given the species tree, the population parameters, and the way in which the gene tree is embedded, which is straightforward to compute.

We also approximate the posterior by ignoring most of the 'unlikely' embeddings. If the migration rate is high, some of the ignored embeddings will be quite likely and the approximation will break down.

Testing on prior only

To make XML files for testing prior, can use template `DENIMfixed-lam-sig`. This has fixed growth rate for the species tree, and fixed population size parameter (which controls amount of ILS). It also has collapse wight fixed at zero (ie fixed species delimitation).

For using with data, use template `DENIM`. This is very like STACEY, and allows the species delimitation to be estimated (in theory - experiments should start simpler!).

1. `make-nex-empty-seqs.R` makes nex files with empty "?" sequences. Eg s8i3locus5.nex has 8 species, three individuals per species, and is the fifth locus like this.

2. Run Beauti2. Choose DENIMfixed-lam-sig.xml template. Load (eg) 8i3locus1.nex, ... 8i3locus9.nex Link all clocks (or put a prior on all but first.)
3. Edit XML. Set bdcGrowthRate.t:Species and migrationIntensity values.
4. Need to use Beauti2 for each basic configuration (number of species, individuals, loci). Then copy and edit XML to make different ones to run.

Testing on data made by MCcoal

1. Run MCcoal. Example `run-mccoal.BAT`, `ctrl.txt`, `imap.txt`, make `trees.txt`, `seq.txt`
2. `mccoal-data-to-nex.R` converts `seq.txt` into a bunch of NEX files which can be loaded into Beauti2.
3. `mccoal-data-tobeastxml.R` is an alternative which make a fragment of a BEAST2 XML file containing the data. This could be pasted in if the rest of the XML is set up for the same species, individuals, and loci.
4. In Beauti2, use DENIM template. In general, will need to edit:
   (a) Taxon Sets: guess, everything after first `^`
   (b) sensible priors on growth rate, relative death rate (or don't estimate)
   (c) sensible prior popPriorScale
   (d) fix collapse weight at zero with [-.0.5,0.5] prior.
   (e) Prior on migration rate. To be experimented with. Perhaps exponential with mean around 1. The results on priors use a value of 5.

Some results with prior only

Made using `lookat-prior-logfiles.R`.

```
      nspp nindivs nloci   ils nseqs sHeight ratio    ESS mean.migs
 [1,]    2       1     3 5e-02     6 0.00505  1.01 5794.7   0.00302
 [2,]    2       1     3 5e-03     6 0.00535  1.07 5951.5   0.02400
 [3,]    2       1     3 5e-04     6 0.00576  1.15 2900.7   0.09294
 [4,]    2       1     9 5e-02    18 0.00512  1.02 4151.0   0.00420
 [5,]    2       1     9 5e-03    18 0.00569  1.14 2240.7   0.02841
 [6,]    2       1     9 5e-04    18 0.00677  1.35 1171.7   0.11382
 [7,]    2       3     3 5e-02    18 0.00535  1.07 5191.7   0.02858
 [8,]    2       3     3 5e-03    18 0.00594  1.19 3682.2   0.13756
 [9,]    2       3     3 5e-04    18 0.00579  1.16 1840.8   0.27038
[10,]    2       3     9 5e-02    54 0.00668  1.34  443.5   0.28892
[11,]    2       3     9 5e-03    54 0.00842  1.68  834.9   0.17198 *
[12,]    2       3     9 5e-04    54 0.00668  1.34  443.5   0.28892
[13,]    2       9     3 5e-02    54 0.00698  1.40 1001.7   0.20993
[14,]    2       9     3 5e-03    54 0.00711  1.42  976.8   0.53270
[15,]    2       9     3 5e-04    54 0.00586  1.17  417.7   0.66453
[16,]    2       9     9 5e-02   162 0.01243  2.49  692.8   0.33545 **
[17,]    2       9     9 5e-03   162 0.01023  2.05  911.1   0.58526 **
[18,]    2       9     9 5e-04   162 0.00762  1.52  309.8   0.68423
```

```
     nspp nindivs nloci   ils nseqs sHeight ratio    ESS mean.migs
[19,]    6       1     3 5e-02    18 0.01527  1.05 5522.3   0.10028
[20,]    6       1     3 5e-03    18 0.01638  1.13 4378.2   0.47834
[21,]    6       1     3 5e-04    18 0.01666  1.15 2682.4   0.96636
[22,]    6       1     9 5e-02    54 0.01691  1.17 2040.7   0.11134
[23,]    6       1     9 5e-03    54 0.02143  1.48 1146.4   0.56828
[24,]    6       1     9 5e-04    54 0.02020  1.39  284.1   1.06812
[25,]    6       3     3 5e-02    54 0.01773  1.22 1859.8   0.60257
[26,]    6       3     3 5e-03    54 0.01750  1.21 1462.1   1.38509
[27,]    6       3     3 5e-04    54 0.01673  1.15  796.3   1.90112
[28,]    6       3     9 5e-02   162 0.02592  1.79 1103.3   0.80550 *
[29,]    6       3     9 5e-03   162 0.02338  1.61 1063.2   1.60438 *
[30,]    6       3     9 5e-04   162 0.01891  1.30  472.3   1.99292
[31,]    6       9     3 5e-02   162 0.02169  1.50  925.8   2.12065 *
[32,]    6       9     3 5e-03   162 0.01835  1.27 1245.8   3.04164
[33,]    6       9     3 5e-04   162 0.01665  1.15  502.6   3.47127
[34,]    6       9     9 5e-02   486 0.03141  2.17  192.0   2.59987 **
[35,]    6       9     9 5e-03   486 0.02256  1.56  201.2   3.20446 *
[36,]    6       9     9 5e-04   486 0.02079  1.43   59.4   3.54862
```

- nspp = number of species

- nindivs = number of individuals per species
- nloci = number of loci
- ils = popPriorScale, wdhich controls amount of ILS
- nseqs = total number of sequence (nspp x nindivs x nloci)
- sHeight = estimated species tree height. Should be 0.01*(1/2) = 0.005 for 2 species, 0.01*(1/2+1/3+1/4+1/5+1/6) = 0.0145 for 6 species
- ratio = (estimated species tree height) / (true species tree height)
- ESS = ESS for posterior
- mean.migs = average number of migration per locus.

Although ESSs are low for larger data sets, the results are similar for other runs. The species tree gets stretched in the prior by the interaction between migration and the multispecies coalescent model. The worst cases are starred.

References

Daniel A. Dalquen, Tianqi Zhu, and Ziheng Yang. Maximum likelihood implementation of an isolation-with-migration model for three species. *Systematic Biology*, 00:00–00, 2016.

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

J Hey and R Nielsen. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of drosophila pseudoobscura and d. persimilis. *Genetics*, 167:747–760, 2004.

J Hey and R Nielsen. Integration within the felsenstein equation for improved markov chain monte carlo methods in population genetics. *PNAS*, 104:2785–2790, 2007.

A D Leaché, R B Harris, B Rannala, and Z Yang. The influence of gene flow on species tree estimation: A simulation study. *Systematic Biology*, 63(1):17–30, 2014.