BayArea v1.0.2
User Manual

July 9, 2013

Contents

1 Overview

BayArea infers the biogeographic histories of taxa sharing a phylogeny, using Markov chain Monte Carlo (MCMC) to approximate the joint posterior of biogeographic histories and range evolution parameters given presence-absence (binary) data and a time-calibrated phylogeny. Model details are available in our forthcoming paper. BayArea is written in C++ and is open source.

Please contact mlandis (at) gmail (dot) com with any comments or questions.

2 Updates

v1.0.2 (July 9, 2013)

- File handling now works for all end line symbols.
- Initialization output block now shows the random seed when left uninitialized.
- Added setting to constrain the distance power parameter to be positive (set "`-geoDistancePowerPositive=T`", default False). This is useful for global and sparse distributions with high rates of dispersal, which may send the distance power to very large negative values.
- Added setting to approximate distances computations to speed up analyses for large numbers of areas (set "`-geoDistanceTruncate=T`", default False).

v1.0.1 (June 28, 2013)

- File handling now deals with excess whitespace more smoothly.
- Fixed issue where analyses with small numbers of areas (e.g. $N < 10$) would randomly freeze during analysis.

v1.0.0 (May 7, 2013)

- Manual now available
- Default settings set to be conservative for general analysis
- Added settings for prior and proposal densities
- Added lnL scores to area_probs.txt
- Example command-line script added
- Example Vireya files added

3 Install

3.1 Executable

The zip file contains the executable file, `bayarea`. This file was compiled on Mac OS X 10.8.3 with Intel Xeon processors using gcc version 4.2.1.

3.2 Compile source code

An executable may not run properly on unsupported operating systems. If you suspect an incompatibility issue, compile the source code to generate an executable for your computing environment. To create the executable `bayarea` while in the source code directory, issue the terminal command:

```
> g++ -O3 *.cpp -o bayarea
```

Other compilers are likely to work, but are untested.

4 Run

BayArea is a command-line program. After installing the program, execute the program:

```
> ./bayarea -areaFileName=my_areas.txt -geoFileName=my_geo.txt -treeFileName=my_tree.txt
```

(The input files are defined in Section 5.)

4.1 Settings

Arguments are appended to the command string, separated by spaces. Each argument follows the format `-argumentName=argumentValue`. Below is a list of currently available arguments and their default values when unspecified (second column in italics).

4.1.1 Input/Output settings

Argument name	Default value	Description
`areaFileName`	*none*	The data file containing presence-absence data for all taxa
`geoFileName`	*none*	The geography file containing latitudes and longitudes
`treeFileName`	*none*	The tree file containing the Newick string
`inputFilePath`	*local directory*	The directory for input files
`outputPrefix`	*none*	Analysis prefix name for all output files
`outputTimestamp`	*T*	Append timestamp string to outputPrefix
`outputFilePath`	*local directory*	The directory for output files
`parameterSampleFrequency`	*1000*	The MCMC sample frequency to populate the `.parameters.txt` file
`historySampleFrequency`	*1000*	The MCMC sample frequency to populate the `.area_states.txt`, `.area_probs.txt`, and `.nhx` files
`printFrequency`	*1000*	Print frequency to the screen (stdout)

4.1.2 Analysis settings

Argument name	Default value	Description
`seed`	*random*	Random number generator seed
`chainLength`	*10000000*	Number of MCMC cycles
`chainBurnIn`	*0*	First MCMC cycle sampling point for `.parameters.txt` and `.area_states.txt`
`probBurnIn`	*2500000*	First MCMC cycle sampling point for `.area_probs.txt` and `.nhx` (Note: For these files to be reflective of the posterior, the chain must first burn-in – i.e. a small `probBurnIn` will give inaccurate results.)
`modelType`	*3*	Indicates which likelihood model to use. Currently available models: 1=INDEPENDENCE 3=DISTANCE_NORM
`gainPrior`	*1.0*	scale parameter for half-Cauchy with location 0
`lossPrior`	*1.0*	scale parameter for half-Cauchy with location 0
`distancePowerPrior`	*1.0*	scale parameter for Cauchy with location 0
`areaProposalTuner`	*0.2*	The number of areas sampled for a history update is Poisson-distributed with rate equal to this value multiplied by the number of areas in the analysis. The value must be between 0 and 1. Adjust this value if you find your MCMC mixes slowly.
`guessInitialRates`	*T*	Use heuristic to initialize MCMC area loss/gain rates
`geoDistancePowerPositive`	*F*	Constrains the distance power parameter to be positive when True. This may lead to more sensible parameter inferences when range data are sparse and global.
`geoDistanceTruncate`	*F*	When True, the pairwise distance calculator approximates extremely improbable dispersal events as having a very small positive rate (effectively zero). This introduces negligible error into the analysis, but greatly speeds up computations when the number of areas is large.

4.2 Batch file

Issuing command line settings is a typo prone process. To minimize this risk, the BayArea archive contains the batch file, `my_run.sh`, which simply executes BayArea with all settings assigned to their default values. If you would like to execute a run with custom settings, copy this file, edit the copy, then execute the copy.

5 Input

5.1 Range data

The first line contains two numbers separated by a space, indicating the number of taxa and number of areas in the data matrix. The following lines begin with the unique taxon label, followed by a space, then followed by that taxon's presence-absence data recorded as a bit string. Taxon presence in area i is marked with a 1, absence is marked with a 0. It is assumed the order of presence-absence data per area matches the order of geographic data (below). An example, where `true_mallard` is present in the first two areas only:

```
3 4
true_mallard 1100
false_mallard 0101
fuzzy_mallard 1110
```

5.2 Geographic data

The first line contains the time interval described by the geographic data in the file. Assign this line to "`# 0.0`" (temporally dynamic geographies will be supported in future versions of BayArea). Each following line reports the geographic coordinates of an area. These values are assumed to match the ordered presence-absence data in the range data file. Example file:

```
# 0.0
38 -122.5
38 -122.4437
38 -122.3875
38 -122.3312
```

5.3 Newick tree

The first and only line contains the Newick string, summarizing the rooted topology and branch lengths of the phylogeny. Taxon labels in the tree file must match those in the range file exactly. As described in the methods manuscript, we approximate the stationary frequencies of the model by letting the Markov chain run until stationarity. We do this by assigning the root an immediate ancestor connected long branch, then sample biogeographic histories at the root using that length. By default, the root branch length is twice the tree height, but may be supplied (e.g. 50) as demonstrated below:

```
((true_mallard:10.0,false_mallard:10.0):10.0,fuzzy_mallard:20.0):50.0;
```

6 Output

6.1 MCMC samples (.parameters.txt)

The .parameters.txt file contains samples from the MCMC, with first sample point and the sample frequency are specified through the analysis settings. The first line reports the column headers, which are (in order, as below) the MCMC cycle, the log likelihood, the full model gain and loss parameters, the sampling model gain and loss parameters, the distance power parameter, and the number of areas gained and lost over the entire tree. Every following line reports the MCMC state corresponding to the cycle indicated by the first element of that line. To visualize your MCMC results, we recommend using this file with Tracer (http://tree.bio.ed.ac.uk/software/tracer/).

```
n       lnL       gain        loss       gain-p      loss-p     distP    numGain  numLoss
200000  -13531.8  0.00525798  0.0535335  0.00539434  0.0512821  2.22786  1329     1399
201000  -13027.2  0.00503444  0.0501542  0.00520644  0.0517992  2.15599  1264     1341
...
```

6.2 Ancestral state reconstruction (.area_states.txt)

The .area_states.txt file contains geographic range samples from the MCMC for all internal nodes in the phylogeny. (Future versions of BayArea will also allow you to sample the range evolution histories between nodes.) The first column indicates the MCMC cycle. The second column reports the node name or node index. The third column reports the geographic range stored at that node at that MCMC cycle. For an analysis on a phylogeny with M tips, the first M rows are the observed presence-absence data. At each sampling interval, the file is updated with the sampled geographic range for each node, listed in postorder tree traversal in blocks of $M - 1$ rows.

```
1000 true_mallard 0010
1000 false_mallard 0101
1000 fuzzy_mallard 1110
2000 4 0010
2000 5 0111
3000 4 0011
3000 5 0110
...
```

6.3 Ancestral state probabilities (.area_probs.txt)

The .area_probs.txt file contains the marginal posterior probabilities of an area occupancy by the ancestral lineage an internal node. The contents of the file are computed and written only at the end of the analysis. Each row corresponds to a node in the tree, with the first M nodes being the tips of the tree, and the final $M - 1$ nodes being the internal nodes. The first column indicates the node name or index, and each following column gives the occupancy probability. More complex posterior summaries of area occupancy (e.g. per-configuration rather than per-area) are not available at this time, but can be computed from the .area_states.txt file as needed.

```
true_mallard 0 0 1 0
false_mallard 0 1 0 1
```

```
fuzzy_mallard 1 1 1 0
Taxon_4 0.2 0.8 0.2 0.1
Taxon_5 0.3 0.6 0.4 0.2
```

6.4 New Hampshire eXtended file (.nhx)

The .nhx file is offered for the user's convenience, containing the input data and the posterior state probabilities as metadata in the tree block. This file may be used as input for two ancestral range reconstruction visualization programs: Phylowood and BayArea-Fig.

```
#NEXUS

Begin taxa;
    Dimensions ntax=3;
    Taxlabels
        true_mallard
        false_mallard
        fuzzy_mallard
        ;
End;

Begin geo;
    Dimensions ngeo=4;
    Coords
        0 38 -122.5,
        1 38 -122.4437,
        2 38 -122.3875,
        3 38 -122.3312
        ;
End;

Begin trees;
    Translate
        0 true_mallard,
        1 false_mallard,
        2 fuzzy_mallard
        ;
tree TREE1 = ((0[&area_pp={0,0,1,0}]:10.0,1[&area_pp={0,1,0,1}]:10.0)[&area_pp={0.2,
    0.8,0.2,0.1}]:10.0,2[&area_pp={1,1,1,0}]:20.0)[&area_pp={0.3,0.6,0.4,0.2)]:50.0;
End;
```

7 Additional software

7.1 Phylowood

Phylowood is a Javascript web service that generates interactive animations for phylogeographic and biogeographic reconstructions. To use Phylowood, navigate to `http://mlandis.github.com/phylowood`. Simply copy and paste the contents of your analysis' `.nhx` file into the text area, then click Load. For further help, visit `https://github.com/mlandis/phylowood/wiki`.

7.2 BayArea-Fig

BayArea-Fig is a Javascript tool that plots ancestral range reconstructions onto a phylogeny, producing figures suitable for publication. To use BayArea-Fig, navigate to `https://mlandis.github.com/bayarea-fig`. Simply copy and paste the contents of your analysis' `.nhx` file into the text area, then click Load. This project is still under development, so *caveat indagator* and whatnot.