Simplified manual for multi-taxa multi-locus msbayes (MTML-msBayes)

TABLE OF CONTENTS

1. Overview

Multi-taxa multi-locus msbayes implements a comparative phylogeographic analysis of multiple co-distributed taxon-pairs using a hierarchical approximate Bayesian computation (HABC) model. It is assumed that ancestral populations are split into pairs of populations. The sizes of these ancestral and descendent populations are free to vary and therefore isolation by colonization is allowed within the general model. More generally, ancestral and current effective population sizes, divergence times and migration rates are allowed to vary across taxon-pairs as are size-change parameters in all of the descendant populations. Optionally, one can run a ABC on a single taxon-pair.

Here is what you can do:

1. test if the taxon-pairs formed at the same time (tau) using estimates of the dispersion index of divergence times, Omega = (Var(tau)/E(tau))

2. estimate when the simultaneous divergence or colonization occurred

3. estimate how many divergence/colonization events occurred across multiple taxon-pairs.

4. estimate when these events occurred and how many taxon-pairs split at each of divergence event (using a constrained prior by fixing the number of divergence times to the estimate under the unconstrained prior)

5. calculate a number of summary statistics from each of the taxon-pairs

6. conduct model choice by running msbayes under different constrained priors (scenarios) and subsequently using ABC to estimate a model indicator parameter.

7. estimate divergence time under a single taxon-pair model with or without migration.

8. Allow for migration rates to vary between descendent sister populations.

9. allow for multiple unlinked loci with rate heterogeneity across loci

10. allows for locus specific ploidy and mutation rate scalars

For full descriptions of the methods that can be implemented, see

2006 Hickerson, M.J., E, Stahl, and H.A. Lessios. Test for simultaneous divergence using approximate Bayesian computation. Evolution 60: 2435-2453

and

2007 Hickerson, M.J., E. Stahl, and N. Takebayashi. msBayes: A flexible pipeline for comparative phylogeographic inference using approximate Bayesian computation (ABC). *BMC Bioinformatics*. 8:268

This program runs on the command line using a unix/linux or Mac OS-X Terminal.app. In this document, we use "$" to indicate the command line prompt (your actual command line prompt may differ). The line which starts from "$" is the line which you should type in the command line, but do NOT type in "$" character, and start type the commands following the "$" character.

2. Installation

a. Installation by system admin

If you are using this program on a computer which you do not have root/superuser/administrator access (e.g. departmental unix server), ask the administrator to install all needed components for you.

After the administrator installed the program, open a command line (e.g., Terminal.app in Mac OS-X), and type:

```
$ msbayes.pl -h
```

If it complains that "command not found", you need to check your PATH. Check the section "SETTING UP EXECUTION PATH" of file named "INSTALL". If it prints out the brief description of usage, you are ready to analyze your data.

b. general installation instructions

See installation instructions

make flow chart cartoon

3. Running it

Below is a simplified 4-step description for running multi-locus msbayes. In general, typing '**-h**' at each of the command lines will give you a set of options

Here are the four command line steps (ABC steps):

Brief explanations of the four steps

a. Making the Input files

```
$convertIM.pl infile.list
```

b. Observed summary statistics vector

```
$obsSumStats.pl -s 7 -T obsSS.table batch.masterIn.fromIM > obsSS.txt
```

c. Simulating the prior

```
$msbayes.pl -c batch.masterIn.fromIM -s 7 -r 1000000 -o priorfile
```

d. Getting the posterior

```
$acceptRej.pl -t 0.001 -p outfig.pdf obsSS.txt priorfile >
modeEstimatesOut.txt
```

a. Making the input files

The easiest way to construct your infiles for msbayes is to first convert the DNA sequence data of each taxon-pair into IMa2 infile format. Go HERE for information about IMa2

You will need one IM infile for every taxon-pair and **the name of each IM infile should end with a suffix of '.im'**

To convert these IM files into msbayes infiles, first, you need to make a text file " **infile.list** " with the names of your IM files like this (where one IM infile per taxon-pair):

```
Fairywren.im
Treecreeper.im
Grassfinch.im
```

In this example there are three taxon pairs.

Then, on the command line type:

```
$convertIM.pl infile.list
```

infile.list = text file list of IM infile names. Each taxon-pair's data is contained in an IM infile **(must have '.im' at the ends of each file name).** See IM documentation for IM infile format.

This makes a text file named **batch.masterIn.fromIM** which contains the user defined priors and sample configuration of your observed DNA sequence data. This contains things like sample sizes, number of base pairs, and also a bunch of priors of parameters that you can modify. Keep in mind that Upper Theta is automatically set as 2x the largest value of pi-within (scaled for mtDNA/nDNA rate differences). You might want to change the upper limit of ancestral population size depending on the biogeographic context.

other options

-o specify name of **batch.masterIn.fromIM**

-m specify mutation rate scalar for average mutation rate difference between an autosomal uni-paraentally inherited haploid locus (mtDNA). For example, if out-group analysis suggests that mtDNA rate is 10x faster than the average nDNA, then one can set this scalar by **$convertIM.pl -m 10 infile.list**. This value will appear in the 4th column of **batch.masterIn.fromIM** for loci that have an inheritance scalar of 0.25 (mtDNA and cDNA). The default for this is 20. If cDNA is estimated to evolve at half the speed as average nDNA used, then one can set this as **-m 0.5**. If one is only using single locus mtDNA data, one could set this mutation scalar as 1 (-m 1).

b. Observed summary statistics vector

Make your observed summary statistic vector AND a human-consumable summary statistics table.

```
$obsSumStats.pl -s 7 -T obsSS.table batch.masterIn.fromIM > obsSS.txt
```

obsSS.table = human-readable summary statistics table (open in your favorite spreadsheet program)

obsSS.txt = observed summary statistic vector (computer readable file)

batch.masterIn.fromIM = user defined priors and sample configuration of your observed DNA sequence data

-s followed by an integer value specifies the sorting pattern for the summary statistic vector. For testing simultaneous divergence we recommend using sorting pattern **-s 7** (sorting the order of taxon-pairs within the summary statistic vector by ascending values of average Pi.b across loci). See manuscript for details

Other integer value can be used to specify different sorting of rows of delineating the order of taxon-pairs.
Arguments:
-s 0: no sorting at all
-s 7: sorting rows by the average pi_b (first moment)
-s 6: sorting rows by the average pi_b (use first two moments across loci of each summary statistic)
-s 5: sorting rows by the average pi_b (use first three moments across loci of each summary statistic)
-s 4: sorting rows by the average pi_b (use first four moments across loci of each summary statistic)

c. Simulating the prior

```
$msbayes.pl -c batch.masterIn.fromIM -s 7 -r 1000000 -o priorfile
```

-o priorfile = file containing the simulated prior (randomly drawn hyper-parameters and associated summary statistics)
-r 1000000 = number of random draws from the hyper-prior (number of simulations)
-s = sorting algorithm for summary statistic vector
-c batch.masterIn.fromIM Sample configuration file with user defined priors and sample configuration of your observed DNA sequence data

Each row of the prior file contains a random draw from hyper-parameters (first set of columns) and their associated summary statistic vector (the subsequent columns). This file starts with a header file describing each column.

other option

-S = set the initial seed (UPPERcase “S”; this is optional)

Here is what prints out on the terminal when it is running:

```
INFO: using ./msprior
INFO: using ./msDQH
INFO: using ./sumstatsvector
```

If you run two or more prior files in parallel, then you can concatenate them into one big file after they are made (before acceptance/rejection). A good way to do this is the UNIX 'cat' utility.
For example

```
$cat priorfile1 priorfile2 priorfile3 ..... > BIGpriorfile
```

d. Getting the posterior

After obtaining a large number of prior draws, one must then obtain a subset of the simulations that contain the closest matches between the summary statistics calculated from the observed and simulated data in order to sample from the the posterior distribution.

```
$acceptRej.pl -t 0.0005 -s "pi.b" -p outfig.pdf obsSS.txt priorfile > modeEstimatesOut.txt
```

These command line arguments produce summaries (**modeEstimatesOut.txt**) and figures (**outfig.pdf**) of the approximate posterior distributions of the hyper-parameters. Additionally, a file ("posteriortable") is produced containing the raw and transformed (local linear regression) hyper-parameter values and their associated summary statistics. After loading this file into R or Excel, one can calculate Bayes factors and posterior probabilities for certain parameter values or ranges of parameter values.

Alternatively one can use various software tools for ABC available from M. Beaumont, DIY-ABC, K. Thornton, ABCtoolbox, popABC, and REJECTOR), to sample from the posterior using various ABC techniques.

obsSS.txt = observed summary statistic vector

modeEstimatesOut.txt = text file with summaries of the posteriors of the various hyper-parameters that are estimated (including modes and 95% quantiles)

outfig.pdf = graphical depictions of the posterior for various hyper-parameters;

-t = the proportion of accepted draws from the prior. For testing simultaneous divergence we recommend 500 - 1000 accepted draws from at least 1,000,000 simulations

other options
-n = print out the names of all available summary stats
-r = Simple rejection method without local linear regression will be used
-s = statString (e.g. -s "pi','wattTheta','pi.net','tajD.denom", <=default). The summary statistics listed here will be used

We recommend using **-s 'pi.b'** for data sets with low sample sizes (number of individuals per taxon-pair)

In depth explanations of various files

A. batch.masterIn.fromIM

This contains the user defined priors and sample configuration of your observed DNA sequence data. This file is divided into two sections (priors and sample configuration). The easiest way to construct this file is to convert each taxon-pair's data into an IM infile **(must have '.im' at the ends of each file name)**.

First section of **batch.masterIn.fromIM** (user-defined priors)

1. **upperTheta (q_{MAX}):** Upper and lower bounds of the uniform prior distribution for each taxon-pair's theta (q=4Nm). The upper value of Theta q is calculated as 2x the largest value of p-within (scaled for mtDNA/nDNA rate differences). Theta is in units of **per site per generation**. The mutation rate (m) for each locus is allowed to vary by drawing a m scalar from a gamma distribution (where the mean is 1.0). In this case, if the taxon-specific q = 0.01, and the locus specific m-scalar = 1.2, then the q for this locus and this taxon-pair is 0.012.

2. **upperTau:** the upper bound of the uniform distribution of tau (t), which is in units if 4N_{AVE} generations. The lower bound is 0.0.

Converting *t* into units if real time (generations). Use the following equations. If estimate of m_{mtDNA} is known, then

$m_{nuc} = m_{mtDNA}$/rate scalar. (1)

Therefore if (2)
and $4N_{AVE} = q_{AVE}/m_{nuc}$, (3)
then divergence time in generational time (t) is
t=4$N_{AVE}t$. (4)

3. **number of tau classes (Psi):** Y is the hyper-parameter for the number of different divergence times across Y taxon-pairs. If it is set to 0, then this discrete hyper-parameter will be drawn from integer values between 1 and Y, the number of taxon-pairs (discrete uniform prior). If another integer value is chosen, then this value will be the number of different divergence times across Y taxon-pairs (constrained hyper-prior).

4. **upperMig and upperRec:** These are the upper bounds of the migration (number if individuals per generation) and intragenic recombination rates.

5. **Ancestral theta multiplier:** Upper bound for the uniform prior of ancestral q relative to q_{MAX}.

Second section of **batch.masterIn.fromIM** (sample size configuration)

This section is divided into columns that correspond to various values describing the data configuration. The rows correspond to loci of particular taxon-pairs.

1. Column 1: label for taxon-pair (taken from IM file name)
2. Column 2: locus label
3. Column 3: ploidy of locus (this column can also specify relative differences in generation times; for example, if one pair has a generation time of 1 and the rest have 2 year generation times, then it can be 0.5 for the first and 1 for the rest at at that 3rd column.)
4. Column 4: mutation rate scalar of locus
5. Column 5: sample size of daughter population 1 from corresponding locus and taxon-pair
6. Column 6: sample size of daughter population 2 from corresponding locus and taxon-pair
7. Column 7:Transition/Transversion ratio
8. Column 8: number of base-pairs at corresponding locus and taxon-pair
9. Columns 9, 10, and 11: the A, C, G base frequencies at corresponding locus and taxon-pair
10. Column 10: FASTA file corresponding to each locus and taxon-pair (this is automatically generated using **convertIM.pl** and **infile.list**)

B. obsSS.txt

The summary statistic vector for the entire multi-taxon-pair/multi-locus data set. This file is not really meant from human consumption. Instead, the HABC acceptance/rejection step uses this file to pick the best matches between the observed data and simulated data. This file only consists of two rows. The first row is a header file that describes the hyper-parameters and corresponding summary statistics in the second row. The hyper-parameter values within **obsSS.txt** are really dummy values that are never used since we are estimating these values and do not really know what these true values really are. This file is identical to the prior file in format (the prior file has the header and subsequent number rows corresponding to the number of simulationed draws from the prior).

C. obsSS.table

This file summarizes file **obsSS.txt** for human consumption. The file contains a series of tables corresponding to every summary statistic for every taxon-pair.

The table will display all of the summary statistics like this :

```
pi.b
            locus 1   locus 2   locus 3
taxon 1
taxon 2
.
.
taxon 3

pi.w
            locus 1   locus 2   locus 3
taxon 1
taxon 2
.
.
taxon 3

.
.
```

D. modeEstimatesOut.txt

This text file gives summaries of the posteriors of the various estimated hyper-parameters. This includes modes and 95% quantiles. Sometimes the mode can not be determined if the posterior distribution is too extreme (i.e. L-shaped). If this happens, then a warning will be posted. When this situation occurs, the mode is often the lower bound of the prior (i.e. 0 or 1).

E. outfig.pdf

This pdf file contains graphical depictions of the posterior for various hyper-parameters that are estimated.

F. posteriortable

This file contains the raw and transformed (local linear regression) hyper-parameter values and their associated summary statistics. After loading this file into R or Excel, one can calculate Bayes factors and posterior probabilities for certain parameter values or ranges of parameter values. It is produced after one runs the acceptance/rejection step (**acceptRej.pl**)