MutAnce Manual

Anna Nagel

May 30, 2025

1 Overview

`MutAnce` is a method to estimate mutation ages, populations of origin, and ancestral and derived states with a structured population model. This program relies on the output of `bpp` as input. Currently, users must use the mutanceDev branch of `bpp`. Directions on how to use the branch are provided in the README for `MutAnce`. This manual focuses on how to use `MutAnce`. Explanation of the details of `bpp` are available in the `bpp` manual. This manual follows the updated (version v4.8.0 and later) of `bpp`, which changes some of the control file options. The only `bpp` settings described here are the settings that are specific to or are required for `MutAnce`.

2 Control Files

Here we describe new control files options for `bpp` and the control file options for `MutAnce`.

2.1 BPP control file variables

1

`printlocus = +d +d*`

DESCRIPTION

Print the migration histories to file for the specified loci. Without this option, the migration histories will not be printed.

VALUES

`+d`, a positive integer specifying the number of loci for which the gene trees and migration histories will be printed to file. `+d*`, a list of `+d` positive integers specifying the index of each locus based on the order of the loci in the sequence file. The values can range from 1 to the number of loci in the analysis.

DEPENDENCIES

The gene trees must be printed (The fourth value in `print` must be 1; `print = 1 0 0 1 0`).

COMMENTS

If `printlocus` is used, the gene tree files for the same loci will be created. The gene tree files for the other loci will not be created. The same applies to the locus-specific substitution-rate parameters if they are printed (the fifth value in `print` is 1). Without `printlocus`, the gene tree files and locus-specific substitution-rate parameters are printed for all loci is the corresponding `print` option is used.

Printing hundreds of loci is not advised as it will slow down the program. There is also a limit to the number open files allowed, which may cause an error if too many (several hundred), loci are printed.

EXAMPLES

```
printlocus = 2 1 2
```

```
printlocus = 5 3 6 9 10 11
```

2.2 Simulator Control file for `bpp`

The simulator will not be needed for users who only wish to analyze empirical data. If users wish to simulate data, they should use the `--simulate` in `bpp`, described in more depth in the `bpp` manual. `printlocus` is specified in the same way as in the inference program. This determines which loci the mutational histories will be printed to screen.

2.3 `MutAnce` Control Files

1

`seed = (-d,+d)`

DESCRIPTION

Specifies the seed used for the random number generator.

VALUES

`-d`, use the wall clock a seed (store seed in file `SeedUsed`).

`+d`, use +d as the seed.

DEFAULT

`-1`

COMMENTS

Using the same positive integer seed will produce identical results in different runs, which is useful for debugging, and using different positive integers in different runs produces different results. Using `-d`, for example, -1, the computer's time is used as a source for the seed, and different runs will produce different results. The positive integer seed automatically generated when using option `-d` is stored in a file named SeedUsed and can be used explicitly to replicate a result.

EXAMPLES

```
seed = -1
```

```
seed = 278
```

2

`seqfile = s`

DESCRIPTION

Sets the name/path of the file containing the sequence data to be the string `s`

VALUES

`s`, a string of characters specifying the directory path and name of file that contains the sequence data. This file should be the same as the file used in the `bpp` to create the other input files for the program (`migfile`, `gtreefile`, and possibly `HKYfile`)

EXAMPLES

```
seqfile = /home/mickey/mouse_seq.txt
```

```
seqfile = sequences.txt
```

3

`gtreefile = s`

DESCRIPTION

Sets the path/name of the input file containing the gene trees to be the string `s`. This file is produced by running `bpp`.

VALUES

`s`, a string of characters specifying the directory path and name of file that contains the MCMC samples of the gene trees for `locus`.

EXAMPLES

```
gtreefile = out.gtree.L3
```

4

`migfile = s`

DESCRIPTION

Sets the path/name of the migration input file to be the string `s`. This file is produced by running `bpp`.

VALUES

`s`, a string of characters specifying the directory path and name of file that contains the migration history and speciation times for each sampled gene tree for `locus`.

EXAMPLES

```
outfile = out.mig.L3
```

5

HKYfile = s

DESCRIPTION

Sets the path/name of the file containing the MCMC samples of the substitution model parameters to be the string s. This file is produced by running bpp.

VALUES

s, a string of characters specifying the directory path and name of file that contains the MCMC samples of the substitution model parameters in the HKY model.

COMMENTS

If HKYfile is not given, a Jukes-Cantor model is assumed.

EXAMPLES

```
outfile = out.locus_3_params_sample.txt
```

6

locus = +d

DESCRIPTION

The index of the locus used in inference. The index matches the ordering in the sequence file.

VALUES

+d, a positive integer specifying the index of the locus. Acceptable values range between 1 and the number of loci in the sequence file.

COMMENTS

If files are used directly from bpp, +d will be in the files names for gtreefile, migfile, and HKYfile.

EXAMPLES

```
locus = 3
```

7

site = +d

DESCRIPTION

The site used

VALUES

+d, a positive integer specifying the site in the alignment. Acceptable values range between 1 and the number of sites in the alignment of the locus specified with `locus`.

COMMENTS

By default, inference will be performed on all polymorphic sites for the specified `locus`. The run time and memory usage are approximately linear with the number of polymorphic sites. If only one or a few sites are of interest, it is recommended to use the `site` option. If a few sites are of interest, `MutAnce` should be run separately for each site.

EXAMPLES

```
site = 105
```

3 Example Control File

3.1 `bpp` Control File

`bpp` must be run as a precursor to running `MutAnce`. Below is an example control file for `bpp`. For more details on `bpp`, refer to the `bpp` manual at https://bpp.github.io/bpp-manual/bpp-4-manual/.

```
seed = 1

seqfile =  simulate_IM.txt
Imapfile = simple.Imap.txt
jobname = out

# fixed species tree
species&tree = 3 A B C
5 5 5
((A, B), C);


# phased data for population
phase =   0 0 0

# use sequence likelihood
usedata = 1

nloci = 5
clock = 1
model = HKY

# invgamma(a, b) for root tau & Dirichlet(a) for other tau's
tauprior = gamma 50 100000
thetaprior = gamma 50 100000

# MCMC samples, locusrate, heredityscalars, Genetrees
print = 1 0 0 1 1
burnin = 2000
```

```
sampfreq = 4
nsample = 10000
printlocus = 2 1 3

wprior = 15 .01
migration = 2
A B
B A
```

The options that a required for `MutAnce` are discussed individually.

```
# fixed species tree
species&tree = 3 A B C
5 5 5
((A, B), C);
```

The species tree is assumed to be known and is specified with the `species&tree` option. Character used in newick format, such as parenthesis and semicolons should not be used in species names. Additionally, dollar signs and asterisks should not be used in species names, as this will interfere with the format for the migration file used in `MutAnce`.

```
model = HKY
```

Model determines the substitution model for the analysis. Either an HKY or Jukes-Cantor model must be used. Jukes-Cantor is the default model, so this line can be omitted when using an Jukes-Cantor model.

```
phase =   0 0 0
```

All of the data must be phased.

```
print = 1 0 0 1 1
```

The first value creates an MCMC file for the run. This must always be 1. If the fourth value, is a 1, `bpp` saves the gene trees sampled in the MCMC to files, which is required for `MutAnce`. If the last values is a 1, `bpp` saves the locus specific substitution model parameters to files, which is also required for `MutAnce` when using a HKY model. It is not required when using a Jukes-Cantor model.

```
printlocus = 2 1 3
```

The first value specifies the number of loci that have their gene trees and substitution model parameters written to file and the next numbers specify which loci in order of appearance in the DNA sequence file. One based indexing is used, so the first locus in the sequence file is locus 1. In the example, the first and third loci in the sequence file will have their samples of gene trees and substitution model parameters from the MCMC written to file. Every locus for which `MutAnce` will be run must be specified in this line. Migration files will be created with the `printlocus` option even if there are no migration connections present in the model. No migration files will be created if this option is not used, and `MutAnce` will not run.

```
wprior = 15 .01
migration = 2
A B
B A
```

As of version v4.8.0, bpp uses the mutation scaled migration rate, w, instead of M, which is used in older versions. $w_{xy} = \frac{4M_{xy}}{\theta_y}$ where w_{xy} is the migration rate from population x to population y in forward time, M_{xy} is the expected number of migrants per generation from x to y in forward time, and θ_y is the mutation scaled effective population size for y. Note that the original publication of MutAnce used M not w.

The following bpp models and functionalities may not be used: among site rate variation (alphaprior), among locus rate variation (locusrate), MSC-I, species tree inference (speciestree = 1), or species delimitation (speciesdelimitation = 1). MutAnce does not accommodate ambiguity codes, but missing data is allowed.

3.2 MutAnce Control File

Here is an example control file for MutAnce.

```
seed = 1

locus = 1
seqfile = simulate_IM.txt

gtreefile = out.gtree.L1
migfile = out.mig.L1
HKYfile = out.locus_1_params_sample.txt

site = 310
```

seed sets the seed and allows for reproducible results.

```
seed = 1
```

locus specifies which locus in the sequence data file should be used. This should match the number specified in printlocus in the bpp control file. If more than one locus was specified in bpp, the user should create a control file for each locus. Then, MutAnce should be run independently for each locus.

```
locus = 1
```

seqfile specifies the DNA sequence file. This should match the seqfile line in the bpp control file.

```
seqfile = simulate_IM.txt
```

These lines specify the output files from bpp. The prefix out matches the bpp control file job option. The number 1 should match the locus specified in the locus option in the MutAnce control file.

```
gtreefile = out.gtree.L1
migfile = out.mig.L1
HKYfile = out.locus_1_params_sample.txt
```

This specified that inference of mutation history should only be done for site 310 in the alignment.

```
site = 310
```

4 Running the Program

4.1 Running `bpp`

Before running `MutAnce`, `bpp` must be run with the `printlocus` option. The `bpp` documentation is extensive, so we refer the user to https://bpp.github.io/bpp-manual/bpp-4-manual/ for directions. The program in run on the command line with a single option, `--cfile`, which gives the name of the control file.

```
bpp --cfile inference.ctl
```

4.2 Running `MutAnce`

`MutAnce` is run on the command line the control file argument.

```
MutAnce controlFile.ctl
```

This will show the progress as a percentage as the program is running and then print summaries of the output after it finishes. To save the screen output, direct stdout and stderr to a file.

```
MutAnce controlFile.ctl &> output
```

Depending on how this file is viewed, the first several lines may appear blank due to the use of ˆM (carriage-return character) to update the progress. This analysis should be considerably faster than running `bpp`. For datasets with 60 samples and 500,000 MCMC samples, the analyses typically took 2-3 hours on our university HPC. The run time is linear in the number of MCMC iterations in `bpp`.

5 Interpreting the Output

This is representative output of `MutAnce` starting after the progress report at the beginning. After the entire output, the output will be explained section by section.

```
310 GGGGGGGGGGGGGAG
site 310

Posterior probability multiple mutations 0.0040

Posterior probability of three or more mutations 0.0000

population of mutation
1_mut_pop_C        0.972791
1_mut_pop_A,B,C        0.027209
1_mut_pop_A        0.000000
1_mut_pop_B        0.000000
1_mut_pop_A,B        0.000000

mean_time        0.0001181040
1_mut_02.5_HPD        0.0000000098
1_mut_97.5_HPD        0.0003991872

root state
1_mut_root_A 0.005823
1_mut_root_C 0.000000
1_mut_root_G 0.994177
1_mut_root_T 0.000000

mutation
1_mut_mutation_A 0.994177
1_mut_mutation_C 0.000000
1_mut_mutation_G 0.005823
1_mut_mutation_T 0.000000

population of mutation 1
2_mut_pop_1_C        0.325000
2_mut_pop_1_A,B,C        0.125000
2_mut_pop_1_A        0.200000
2_mut_pop_1_B        0.200000
2_mut_pop_1_A,B        0.150000

population of mutation 2
2_mut_pop_2C        0.900000
2_mut_pop_2A,B,C        0.000000
2_mut_pop_2A        0.075000
2_mut_pop_2B        0.025000
2_mut_pop_2A,B        0.000000

2_mut_time_1        0.0002925405
2_mut_02.5_HPD_1          0.0000239671
2_mut_97.5_HPD_1        0.0009675987

2_mut_time_2        0.0000911371
2_mut_02.5_HPD_2        0.0000044721
2_mut_97.5_HPD_2        0.0002740323

root state
2_mut_root_A        0.000000
2_mut_root_C        0.000000
2_mut_root_G        1.000000
```

```
2_mut_root_T        0.000000

First mutation
2_mut_mutation_1_A        0.225000
2_mut_mutation_1_C        0.000000
2_mut_mutation_1_G        0.775000
2_mut_mutation_1_T        0.000000

Second mutation
2_mut_mutation_2_A        0.775000
2_mut_mutation_2_C        0.000000
2_mut_mutation_2_G        0.225000
2_mut_mutation_2_T        0.000000

Joint probability distribution of mutations to bases conditional on two mutations
Row: first mutation, Column: second mutation
site 310
A       C       G        T
A       0.000000        0.000000        0.225000        0.000000
C       0.000000        0.000000        0.000000        0.000000
G       0.775000        0.000000        0.000000        0.000000
T       0.000000        0.000000        0.000000        0.000000


Joint probability distribution of populatons conditional on two mutations
Row: first populaton, Column: second population
site 310
C       A,B,C       A        B        A,B
C       0.225000        0.000000        0.075000        0.025000        0.000000
A,B,C       0.125000        0.000000        0.000000        0.000000        0.000000
A       0.200000        0.000000        0.000000        0.000000        0.000000
B       0.200000        0.000000        0.000000        0.000000        0.000000
A,B       0.150000        0.000000        0.000000        0.000000        0.000000
```

The site pattern(s) for either the specified site (if the `site` option was used) or all polymorphic sites are displayed. In this example, only a single site is displayed. Next the site number is displayed. If there are multiple sites, they will be listed on the same line. All of the reports will follow the same order of sites.

```
310 GGGGGGGGGGGGGAG
site 310
```

The posterior probability of multiple mutations is the probability that the observed nucleotide states arose by two or more mutations. This could be a results of multiple mutations to the same state or mutations to two or more different states. The posterior probability of three or more mutations is expected to be small in most cases. The posterior probability that the nucleotide states arose by exactly one mutation can be found by subtracting the probability of multiple mutations from 1, or 1 - 0.0040 in this case. Similarly, the posterior probability of exactly two mutations can be found subtracting the probability of three or more mutation from the probability of multiple mutations, or 0.0040 - 0.0000 in this case.

```
Posterior probability multiple mutations 0.0040

Posterior probability of three or more mutations 0.0000
```

The first set of probabilities are conditional on there being a single mutation. These lines start with `1_`. First are the probabilities for the populations of origin. If a population has posterior probability 0 of being the population of origin, it will not be displayed in list. The order of the populations may vary between runs with different seeds or input files. The population labels of the tips match the labels in the species (population) tree in the `bpp` control file. The internal nodes are the names of their daughter nodes separated by commas.

```
population of mutation
1_mut_pop_C        0.972791
1_mut_pop_A,B,C        0.027209
1_mut_pop_A        0.000000
1_mut_pop_B        0.000000
1_mut_pop_A,B        0.000000
```

Next is the mean of the posterior distribution for the mutation time and the 95% highest posterior density interval. Time present is zero and times increase going backward into the past. The units are in expected number of substitutions per site. To estimate a time in years, you must assume a mutation rate per year. Time in years equals time in expected number of substitutions per site divided by the mutation rate per year.

```
mean\_time        0.0001181040
1\_mut\_02.5\_HPD       0.0000000098
1\_mut\_97.5\_HPD       0.0003991872
```

Next are the posterior probabilities of the root state and the mutation. If there are two states and a single mutation, this information is redundant as 1 − 1_mut_root_A = 1_mut_mutation_A.

```
root state
1_mut_root_A 0.005823
1_mut_root_C 0.000000
1_mut_root_G 0.994177
1_mut_root_T 0.000000

mutation
1_mut_mutation_A 0.994177
1_mut_mutation_C 0.000000
1_mut_mutation_G 0.005823
1_mut_mutation_T 0.000000
```

After the summaries conditional on a single mutation, the same summaries are given conditional on exactly two mutations. These summaries start with `2_`. However, the times, populations of origin, and mutations are separated for each mutation. Mutation `1`, which is reported with the suffix _1 or _1_ before the population label, is the older mutation, occurred first in forward time. Mutation `2` is reported with the suffix _2 or _2_ before the population label.

Similar to the results conditional on a single mutation, the posterior probabilities of the population of origin are shown for each mutation. If the posterior probability a mutation occurred in a population is exactly zero, it will not be listed. As before, the internal nodes in the species (population) tree are named by the names of their daughters separated with commas, as before.

```
population of mutation 1
2_mut_pop_1_C        0.325000
2_mut_pop_1_A,B,C        0.125000
2_mut_pop_1_A        0.200000
2_mut_pop_1_B        0.200000
2_mut_pop_1_A,B        0.150000

population of mutation 2
2_mut_pop_2C        0.900000
2_mut_pop_2A,B,C        0.000000
2_mut_pop_2A        0.075000
2_mut_pop_2B        0.025000
2_mut_pop_2A,B        0.000000
```

Next the times of the first and second mutation are shown. The first mutation time is larger than the second mutation time since the values are expected number of mutations before time present. Similar to the results conditional on a single mutation, the 95% highest posterior density intervals are reported for both mutations.

```
2_mut_time_1        0.0002925405
2_mut_02.5_HPD_1          0.0000239671
2_mut_97.5_HPD_1        0.0009675987

2_mut_time_2        0.0000911371
2_mut_02.5_HPD_2        0.0000044721
2_mut_97.5_HPD_2        0.0002740323
```

Next the posterior probabilities for the root state and the mutations are shown. These are the marginal probabilities and are not conditional the other mutation. The _1 and _2 match the numbering described for the populations.

```
root state
2_mut_root_A        0.000000
2_mut_root_C        0.000000
2_mut_root_G        1.000000
2_mut_root_T        0.000000

First mutation
2_mut_mutation_1_A        0.225000
2_mut_mutation_1_C        0.000000
2_mut_mutation_1_G        0.775000
2_mut_mutation_1_T        0.000000

Second mutation
```

```
2_mut_mutation_2_A        0.775000
2_mut_mutation_2_C        0.000000
2_mut_mutation_2_G        0.225000
2_mut_mutation_2_T        0.000000
```

The final section has the joint distribution of the mutations conditional on exactly two mutations at the site and the joint distribution of population of origin conditional on exactly two mutations at the site. The rows in the table are the first (older) mutation and the columns are the second (younger) mutation. For example, the probability that the first mutation was a A and the second mutation was a G conditional on two mutation is 0.225000. The probability of recurrent mutation (two mutations to the same state) can be found by summing the diagonal elements of the table.

```
Joint probability distribution of mutations to bases conditional on two mutations
Row: first mutation, Column: second mutation
site 310
A        C        G        T
A        0.000000        0.000000        0.225000        0.000000
C        0.000000        0.000000        0.000000        0.000000
G        0.775000        0.000000        0.000000        0.000000
T        0.000000        0.000000        0.000000        0.000000
```

The final table shows the joint distribution of site of origins conditional on two mutations. This table is in the same format as the previous table, with the older mutation on the rows and the younger mutation on the columns. For example, the probability the first mutation occurred in population A,B and the second mutation occurred in population C is 0.150000.

```
Joint probability distribution of populatons conditional on two mutations
Row: first populaton, Column: second population
site 310
C        A,B,C        A        B        A,B
C        0.225000        0.000000        0.075000        0.025000        0.000000
A,B,C        0.125000        0.000000        0.000000        0.000000        0.000000
A        0.200000        0.000000        0.000000        0.000000        0.000000
B        0.200000        0.000000        0.000000        0.000000        0.000000
A,B        0.150000        0.000000        0.000000        0.000000        0.000000
```

If inference is done for all polymorphic sites (omitting `site` in the control file), all of the summary statistics until the tables will have multiple columns with each column corresponding to a different site. The sites are in increasing order and listed at the top of the output. The tables for the joint distributions conditional on two mutations will have the site printed on the line prior to the table, as is shown in the example.