BEST 2.3 Manual

Liang Liu[1], Dennis K. Pearl[2], and Scott V. Edwards[1]

[1]lliu@oeb.harvard.edu
[2]dkp@stat.osu.edu
[1]sedwards@fas.harvard.edu

[1]Department of Organismic and Evolutionary Biology
Harvard University
Cambridge, MA, 02138

[2]Department of Statistics
The Ohio State University
Columbus, OH, 43210

1 Introduction

BEST is a Bayesian program for estimating species phylogenies from multilocus and multiple-allele sequences. BEST 1.6 and 1.7 implement a two-step MCMC algorithm to estmate the posterior distribution of species trees [1-3]. From version 2.0 [4-5], BEST has integrated the two-step MCMC into a single MCMC algorithm built upon the popular phylogenetic program MrBayes. Castillo et al [6] wrote a book chapter for BEST 2.3 that contains recommendations on issues of chain convergence, model violation, relative mutation rates of genes, non-clocklike evolution. This manual is written to explain the commands used by BEST 2.3 for estimating species trees. For the information about commands used in the regular MrBayes, that may be also used in BEST 2.3, users may refer to the MrBayes manual which can be downloaded at `http://mrbayes.csit.fsu.edu/mb3.1_manual.pdf`. We strongly recommend that users read the MrBayes manual and become familiar with the commands used in MrBayes before reading this manual.

2 Installing and compiling BEST 2.3

The windows and Mac executables of BEST 2.3 are free for download at `http://www.stat.osu.edu/~dkp/BEST`. For those who want to run the program on Unix/Linux, the source code is provided at the same website. The program can be compiled by simply typing ”make” under the folder where the source code is contained and you will see

```
    gcc -DUNIX_VERSION -DUSE_READLINE -O3 -Wall -c -o best.o best.c
    gcc -DUNIX_VERSION -DUSE_READLINE -O3 -Wall -c -o mcmc.o mcmc.c
    gcc -DUNIX_VERSION -DUSE_READLINE -O3 -Wall -c -o bayes.o bayes.c
    gcc -DUNIX_VERSION -DUSE_READLINE -O3 -Wall -c -o command.o command.c
    gcc -DUNIX_VERSION -DUSE_READLINE -O3 -Wall -c -o mbmath.o mbmath.c
    gcc -DUNIX_VERSION -DUSE_READLINE -O3 -Wall -c -o model.o model.c
    gcc -DUNIX_VERSION -DUSE_READLINE -O3 -Wall -c -o plot.o plot.c
    gcc -DUNIX_VERSION -DUSE_READLINE -O3 -Wall -c -o sump.o sump.c
    gcc -DUNIX_VERSION -DUSE_READLINE -O3 -Wall -c -o sumt.o sumt.c
    gcc -DUNIX_VERSION -DUSE_READLINE -O3 -Wall best.o bayes.o command.o tool.o best.o
mbmath.o mcmc.o model.o plot.o sump.o sumt.o -lncurses -lreadline -lm -o best
```

When compiling BEST 2.3, please first make sure that Architecture in the file Makefile is correctly set to the platform you are using. The default is Architecture = unix which should be changed to Architecture = windows if you want to compile the program on Windows. For other problems such as compiling parallel version of the program, please refer to section 7 in the Mrbayes manual.

If you are encounting problems regarding compiling the program and unable to find solutions in the Mrbayes manual or find a bug in BEST 2.3, you can report the problem or bug to lliu@oeb.harvard.edu. Your report can help us greatly improve the program and we are always grateful for it. We have created a user mailing list so that registered users can get the latest information about the ongoing projects and updates of the BEST program.

There are two ways to run BEST 2.3. Simply typing "./best" will prompt a working environment where you can type in commands to manipulate and analyze aligned sequences. Or you can use the command line "./best -i data.nex" to analyze pre-prepared input file data.nex in which the commands in the mrbayes block will be executed by the program.

3 Prepare the input file

The input file for BEST 2.3 is in NEXUS format and consists of two major blocks: data block and mrbayes block. There is no difference between the input files of BEST 2.3 and MrBayes in terms of the data block. Thus, create the data block in the same way as that in the regular MrBayes. For example, the data block in the example file test.nex (available at `http://www.stat.osu.edu/~dkp/BEST/examples/test.nex`) is

```
begin data;
    dimensions ntax=4 nchar=3000;
    format datatype=DNA interleave missing=? gap=-;
    matrix
    H TTTCGGGTAT GATTGAACCG .....
    C TTTCGGGTAT GATTGAACCG .....
    G TTTCGGGTAT GATTGAACCG .....
    O TTTCGGGTAT GATTGAACCG .....
    .......
;
end;
```

The multilocus sequences should be concatenated across loci in the data block. Missing nucleotides or sequences are replaced by question marks. Although BEST 2.3 can still estimate the species tree when the whole sequences are missing for some genes, users should remain cautious about the result because the placement of the taxa in the tree is purely derived from the prior distribution. If the dataset contains both diploid genes (nucleotide DNA) and haploid genes (e.g. mtDNA), users may duplicate the haploid genes to make it compatible with the diploid genes or randomly choose one of the two sequences for the diploid genes to make it compatible with the haploid genes.

BEST 2.3 is the extension of the regular MrBayes revised for the purpose of estimating species phylogenies. The commands for specifying substitution models and prior distributions at gene tree level in the regular MrBayes are still valid for being used in BEST 2.3. There are some new commands being added in BEST 2.3 in order to set the prior distribution for the species tree, population sizes, and variable mutation rates across genes. A typical mrbayes block for BEST 2.3 specifies

1. outgroup: only a single outgroup sequence is allowed.
2. the locations of genes along sequences.
3. the sequence-species relationship, i.e., which sequences belong to which species.
4. substitution models for genes.
5. priors for the parameters in the substitution model.
6. priors for the species tree, mutation rates across genes, and population sizes.
7. haploid genes.

which will be explained item by item using the example file test.nex as shown below.

```
begin mrbayes;
    set autoclose=yes nowarn=yes;
    outgroup 4;
    taxset H = 1;
    taxset C = 2;
    taxset G = 3;
    taxset O =4;
    CHARSET gene1 = 1 - 500;
    CHARSET gene2 = 501 - 1000;
    CHARSET gene3 = 1001 - 1500;
    CHARSET gene4 = 1501 - 2000;
    CHARSET gene5 = 2001 - 2500;
    CHARSET gene6 = 2501 - 3000;
    partition Genes = 6: gene1, gene2, gene3, gene4, gene5, gene6;
    set partition=Genes;
    prset thetapr=invgamma(3,0.003) GeneMuPr=uniform(0.5,1.5) BEST=1;
    unlink topology=(all) brlens=(all) genemu=(all);
    mcmc ngen=1000 nrun = 2 nchain = 2 samplefreq=100;
    sumt nrun=2 filename=test.nex.sptree;
end;
```

3.1 set genes

The location of each gene is defined by CHARSET. For example, CHARSET gene1 = 1 - 500; says that the first 500 nucleotides belong to the gene gene1. The command "partition" divides sequences into genes specified by CHARSET. The partition must be activated by the command"set partition=gene".

It is quite common in gene tree estimation to use codon models to group the nucleotides in triplets and assume different evolutionary models for the three groups of nucleotides. While users are still allowed to use a codon model to group nucleotides in triplets, we do not recommend dividing data by triplets and treating each of the three groups of nucleotides as a "gene". It would be more appropriate to define a codon model within each gene specified by CHARSET, but the current version of BEST is unable to support this possibility.

3.2 set the sequence-species relationship

The command "TAXSET" tells the program which sequences belong to which species. In the example file, the list of TAXSETs implies that there are four species and each species has only one sequence (single allele data). Although the species names (H, C, G, O) coincide with the sequences' names (H, C, G, O) in the example file, it is totally valid to use other names for species, for instance,

```
taxset s1 = 1;
taxset s2 = 2;
taxset s3 = 3;
taxset s4 =4;
```

For multiple allele data, multiple sequences may belong to the same species. For example,

```
taxset s1 = 1-4;
taxset s2 = 5,7,9;
taxset s3 = 6,8;
taxset s4 =10;
```

indicates that sequences 1 to 4 belong to species s1, sequences 5, 7, 9 to species s2, etc.

3.3 set substitution models for genes

The substitution model for each partition is specified by the command "lset". Users may refer to the mrbayes manual for the information about the command "lset".

3.4 set priors for the parameters in the substitution model

Please refer to the MrBayes manual for the information about how to specify prior distributions for the parameters in the substitution model.

3.5 set priors for the species tree, mutation rates across genes, and population sizes.

The priors for the species tree, mutation rates, and population sizes are set in the command "prset".

Options:

BEST: this parameter initiates the Bayesian analysis for estimating species trees when setting BEST =1. If BEST=0, the regular MrBayes program is implemented.

thetapr: this parameter sets the prior distribution for population sizes. There is only one option, inverse gamma distribution with parameter α and β. The mean of the inverse gamma distribution is $\beta/(\alpha-1)$. Users should choose reasonable values for α and β such that the prior mean of the population size θ is in a reasonable range. In the example file, thetapr=invgamma(3,0.003) implies that the prior mean of the θ is 0.0015.

genemupr: this parameter sets the prior distribution for the mutation rates across genes. Two options: genemupr=uniform or genemupr=fixed(a).

Default model settings for BEST

Parameters	Options	Default settings
BEST	0/1	0
thetapr	invgamma	
genemupr	uniform/fixed	fixed(1.0)

In order to implement BEST 2.3, the command "unlink" must be used to unlink topology, branch lengths, and mutation rates across loci, i.e., unlink topology=(all) brlens=(all) genemupr=(all);. Users may unlink other parameters in the model if necessary, but they are optional.

3.6 set haploid genes

Use ”lset Ploidy=haploid” to define haploid genes. For example, ”lset applyto=(1,2) ploidy=haploid” implies that the first two genes are haploid while other genes are diploid (diploid is the default setting for ploidy).

4 Output files

Like the regular MrBayes, BEST 2.3 produces .p, .mcmc and .t files for which the description can be found in the MrBayes manual. The species trees generated from the posterior distribution is saved in the .sptree file. For multiple runs (for example, nruns=2), BEST 2.3 produces a .sptree file for each run. The names in the translation table in the .sptree file matches the species names specified by the command ”TAXSET” in the MrBayes block in the input file. In the tree block, the number right after the pound sign is the population size θ.

5 Summarizing the posterior distribution of the species tree

The estimated posterior distribution of the species tree can be summarized by the command ”sumt”. For single allele data such as the example file test.nex, if the species names match the sequences’ names, the command ”sumt” can be executed directly in the input data file. For multiple-allele data or when the species names are not identical to the sequences’ names, users must create a new input file for summarizing the species trees in the .sptree file. The sequences’ names in the new input file must match the names in the translation table in the .sptree file. BEST 2.3 will produce such an input file automatically after the MCMC run. This file has a name of the form xxx.sumt. To summarize the esitmated posterior distribution of the species tree, type ”execute xxx.sumt” at the BEST command line or ”./best -i xxx.sumt” outside BEST. The command ”sumt” produces .con, .trprobs, and .part files. The consensus tree as well as the estimates of population sizes (after ”#”) is output to the .con file. The population sizes are estimated by the posterior mean. The posterior mean and standard deviation of the population size and divergence time for each population are saved in the .part file.

6 References

[1] Liu, L. and D.K. Pearl. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Systematic Biology, 2007, 56:504-514.

[2] Edwards, S.V., L. Liu., and D.K. Pearl. High resolution species trees without concatenation. Proceedings of the National Academy of Sciences (USA), 2007, 104:5936-5941.

[3] Liu, L., D.K. Pearl, R.T. Brumfield, and S.V. Edwards. Estimating species trees using multiple-allele DNA sequence data. Evolution. 2008, 62(8):2080-2091.

[4] Liu, L. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 2008, 24(21):2542:2543.

[5] Liu, L., L. Yu, L. Kubatko, D.K. Pearl, and S.V. Edwards. Coalescent methods for estimating multilocus phylogenetic trees. Molecular Phylogenetics and Evolution, 2009, doi:10.1016/j.ympev.2009.05.033.

[6] Castillo, S., L. Liu, D.K. Pearl, S.V. Edwards, Bayesian estimation of species trees: a practical guide to optimal sampling and analysis, in book ”Estimating species trees” (edited by Laura Kubatko and Lacey Knowles), 2010.