BRIEF DOCUMENTATION FOR BEST, VERSION 1.6

Patrícia H. Brito, Scott V. Edwards, and Liang Liu and Dennis Pearl
September 2007

1. INTRODUCTION

AIM: To estimate the posterior distribution of species trees using multilocus, multiple-allele DNA sequence data accounting for deep coalescence of alleles. It is intended to implement the Bayesian hierarchical model proposed by Liang Liu and Dennis Pearl (2007; Syst. Biol. 56, 504-14) and further developed in collaboration with Scott Edwards (Edwards, S. V., Liu, L. & Pearl, D. K. (2007) Proc. Natl. Acad. Sci. (USA) 104, 5936-41). Details of the Bayesian hierarchical model are also available in Liang's Thesis (available at: http://www.stat.osu.edu/~dkp/BEST/Thesis.pdf).

CITATION: Liu, L., and D. K. Pearl. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504-514.

The current version (1.6) is based on the following work:

Liang Liu, Dennis K. Pearl, Robb T. Brumfield, Scott V. Edwards. 2007. Estimating species trees using multiple allele DNA sequence data. Evolution, submitted.

BACKGROUND: Version 1.6 of BEST can take single-allele or multiple allele sequence data from an arbitrary number of individuals and species and estimate a posterior distribution of trees containing only individual speccies as tips. Estimating the posterior distribution of species trees involves two consecutive Markov Chain Monte Carlo (MCMC) procedures. The first MCMC is performed in a revised MrBayes (BEST part 1) in which a new function is added to approximate the joint probability of gene trees. The output consists of a control file and a vector of gene trees consisting of gene trees for each locus. These two files constitute the inputs for the second MCMC program, BEST part 2. BEST part 2 also calculates the importance sampling weights and creates a posterior distribution of species trees based on these weights. This posterior distribution can be summarized in a program such as MrBayes or PAUP.

2. COMILING THE SOURCE CODES

BEST parts 1 and 2 use the same readline libraries and makefile as MrBayes. If these libraries are in place, by simply typing **make** in the folder BEST_1_v1.6**,** the compiler can generate the executable program, called mbbest (for MrBayes, best). If you are having trouble compiling the program it is likely that you do not have the appropriate readline libraries on your computer, or they are not in the appropriate directory. Many useful libraries for compiling sourcecodes can be found at:

http://ftp.gnu.org/gnu/readline/

If you encounter problems with compiling, refer to the Mrbayes manual for suggestions. **We strongly suggest working with someone familiar with compiling sourcecodes on the appropriate platform, or using the pre-compiled executables we have provided on the web site.**
To compile the parallel version of BEST part 1, you need to install the MPICC compiler on the computer and see details in the mrbayes manual about the compiling. The executable program BEST part 2 can also be compiled by typing **make** in the folder BEST_2_v1.6.

3. Runing BEST part 1 (mbbest), the modified version of MrBayes

To run best part 1, simply type: mbbest -i filename.nex or ./mbbest -i filename.nex if the executables are in the same directory.

FILES NEEDED: nexus file with MrBayes block and BEST block

Indicate the outgroup in the Mrbayes block, otherwise BEST will use by default the first sequence of the dataset as outgroup. The outgroup taxa should have NO MORE THAN ONE sequence. You must have an outgroup, since BEST works mostly with rooted gene and species trees.

The nexus file should present the DNA sequence data for each OTU as a concatenated gene sequence, with separate character sets indicating each locus. BEST does not use concatenation for estimation, it only uses the format for the input file. Several complete example input files are available on the BEST website.

EXAMPLE INPUT FILE:

```
#NEXUS
[Zosterops anonymous loci - PHB MAY 2007]

BEGIN DATA;
      DIMENSIONS  NTAX=32 NCHAR=2458;
      FORMAT DATATYPE=DNA  MISSING=? GAP=- ;
MATRIX
MKL131_a
AAGGATTGGCACAATAGTAATGAAACTCATCCACATCTTCAGATGAACTTAGTAGAGTTAAGAATTGAGCTCACTTTGGTAGA
AACCAGGCTGACCATTTGCTTGCAAATGAGTATTTTAGTGAAAAAGATTATTAAAT------------
CCTTTGGCCCCTTCTCTGCACTCTGTCAATATCCTTTCCTCTCTTTAGAATGGCAGGTAAATATTTTATTTGTTGTTCTCTCA
CTGCCAGTGTATTTATAGGGGTTTGTGTTCATATTAATCTTGTTGCTTTCTTATTTGTATTCCATAATGTGTGTTGGTTTTTC
TTATTTTTCACTGTGTGCTTTTGTGCTTGTCATACTTTATGTTTGTTGAAGATCTGTGAAATATCTGATTTTGTGCAATTCCC
GTGTTTGTCTCCACTGGAAAAAAATTACTCAGAACTCCCCAAAAAAAAATAACCCAAACCCATCCTGAGAGTTTCCCTCACCTT
CCTGGAGCTCTCACAGGTCAGAGCTTCCAATTCTGCCACTAAATCACAATCCTCCAGAGCCACTTTGGTGATGTAACGTTTCC
CTGGAGACAAAATTTCAGGAGAACACAAGAGGGTGAGGAATTAATTCTTCAAGGCTTTCTCCTCTTTTCTCCCACTAAAATCC
CTGCTCTGCCAGCAAAACCTGACTCACAACTGAATAAATGGAACAAATCTGGCAACAGGAGACTTTAAAATAATGATTGTAAC
AAATCTGGGCTTCAAAATTCTTTTACACAGGTCAGGGCAAGTGTCTCTCTCTTATTAAGTCTGCCAGCTCCCAGTTTATAGTC
TAAATTGAATAGAACTATGCTTCTCT-------
... etc ...
;
END;

Begin mrbayes;
set autoclose=yes nowarn=yes    # this allows you to run BEST in batch mode ;
outgroup ML131_a                # indicate the outgroup here ;
```

```
charset 2B = 1 - 430;
charset 3A = 431 - 752;
charset 5A = 753 - 1204;
charset 10A = 1205 - 1629;
charset 11B = 1630 - 2071;
charset 14B = 2072 - 2458;
partition bylocus = 6: 2B, 3A, 5A, 10A, 11B, 14B;
set partition = bylocus;
lset applyto=(1, 3, 4, 6) nst=1 rates=equal;
lset applyto=(2, 5) nst=2 rates=equal;

prset treeheight=exponential(0.000001) GeneMuPr=uniform(0.2,2) JointtreePr=1;

# Prior settings – the program will use all the default settings or and plus the
ones specified here
# treeheigh=exponential(0.000001) – this is basically an uninformative prior for
the treeheight, assuming a molecular clock the default in MrBayes is
exponential(1). We want this setting because we already dealt with the tree height
prior when we specified the joint prior based on the coalescent.

# GeneMuPr=uniform(0.2,2) – this setting is unique to the modified version of
mrBayes. This indicates the prior for the mutation rate across loci. Here we
specify that the relative rates across loci can vary between 0.2 and 2, with the
average as 1. This setting should cover all biologically realistic situations, and
allows for rate variation across loci.

# JointtreePr=1 this setting is unique to the modified version of mrBayes.
JointtreePr=1 indicates that the the joint prior is desired, as opposed to the
independent prior (the usual approach when an ‘uninformative’ prior is used).
JointtreePr=0 is used to indicate the independent prior, essentially the default
version of mrBayes.


unlink topology=(all) brlens=(all) genemu=(all) shape=(all) Statefreq=(all)
tratio=(all) revmat=(all) pinvar=(all);

# genemu=(all) is unique to the modified version.  It indicates that the locus-
specific mutation rates can vary across loci. With these settings we are saying
that we want to allow all parameters to be different across loci.

mcmcp ngen=5000000 nruns=1 printfreq=100 samplefreq=100 nchains=1 savebrlens=yes;
mcmc;
end;

the mcmc and end commands are standard in MrBayes.

#below is the best block used in the first step

Begin best                 # use the same capitalization as here;
-1                         # positive number: seed defined by user; negative number:
random seed generated by program;

1000 900 10                # maximum number of proposals per MCMC cycle for species
tree, burnin, printfreq;
```

```
(The larger number for maximum proposals the better, but if it is larger than
around 5000 it will considerably slow down the mcmc run;
1                     # pilot run  - this has a shortcut used to speed the run,
but needs more testing – setting can be 0=no or 1=yes;

1 3 1                 # 1:indicates inverse gamma prior of theta, alpha,beta;

(or, for example 0 1 200 means, using a gamma prior for theta, with associated
alpha and beta. Using the inverse gamma allows BEST to integrate out theta if this
parameter is not of interest.  BEST will still deliver estimates of theta when
using the inverse gamma but the estimates of the species tree will have theta
integrated out. When using the gamma prior, theta is estimated for each node as a
regular paremeter.)

6 32 32 32 32 32 32     #no_ of loci, no_ of individuals per locus ;
MKL131_a MKL131_b CEF382_a CEF382_b CEF846_a CEF846_b PRS2684_a PRS2684_b CEF811_a
CEF811_b CEF825B_a CEF825B_b CEF823_a CEF823_b PRS2615_a PRS2615_b VGR666_a
VGR666_b MKL60_a MKL60_b MKL28_a MKL28_b CES716_a CES716_b CEF470_a CEF470_b
... etc ...

#above are names of the terminals for each sequence; these are the OTUs ;

1.000000 1.000000 1.000000 1.000000 1.000000 1.000000   #prior for the mutation
rate ratio – 1.000000 means same mutation rate for all loci;

0 0 0 0 0 0                     # 0 = diploid, 1=haploid ;

9 2 2 2 4 4 4 4 4 6             # no_ of species, and no. of terminals (alleles)
per species; this must be in the same order as sequences and labels ;

MKL131_a MKL131_b CEF382_a CEF382_b CEF846_a CEF846_b PRS2684_a PRS2684_b CEF811_a
CEF811_b CEF825B_a CEF825B_b CEF823_a CEF823_b PRS2615_a PRS2615_b VGR666_a
VGR666_b MKL60_a MKL60_b MKL28_a MKL28_b CES716_a CES716_b CEF470_a CEF470_b
CEF873_a CEF873_b CEF871_a CEF871_b VGR631_a VGR631_b
#above are names of terminals in the species tree. These are usually just a repeat
of the names used to designate sequence or OTUs as above. ;
end ;
```

INPUT FILE FOR BEST PART 1:

filename.nex

OUTPUT FILES FROM BEST PART 1

filename.nex.p	This is the table of posterior probability distributions, without the trees, just as in the regular MrBayes. You can import this file into Excel or TRACER to analyze posterior distributions of parameters
filename.nex.mcmc	Same as in MrBayes
filename.nex.tree1.t, filename.nex.tree2.t, etc.	Posterior distribution of gene trees, just as in MrBayes; one file for each gene. These gene trees are not ultrametric
filename.nex.tree1.newt, filename.nex.tree2.newt, etc.	Posterior distribution of gene trees, with each gene tree made ultrametric and normalized across

	genes.
filename.nex_treevector.t	The gene tree vector file that combines gene trees from all .newt files. If you have used nrun = 2 or more, this file is not produced in version 1.6.
filename.nex_bestcontrol	The control file for BEST part 2. This file can be modified to change the run parameters for BEST part 2
filename.nex_prob	This file contains the approximate joint probabilities of gene trees that will be used to calculate the importance sampling weights in BEST part 2.

4. Running BEST part 2

To run BEST part 2, simply type: best filename.nex_bestcontrol or
./best filename.nex_bestcontrol if the executables are in the same directory.

Files needed: To run BEST part 2:

```
filename.nex_bestcontrol    # file with the BEST block and MCMC sampling parameters
filename.nex_treevector.t   # file with gene trees.
filename.nex_prob
```

Example of a control file:

```
filename.nex_treevector.t filename.nex_prob # Gene tree file – tells
the program where the trees are ;
-1                      # positive number: seed defined by user, in case, for
example, one wants to run identical MCMC runs. - negative number: random seed ;
1 5000 1                # start and stop of genetrees, number of chains - start
and stop indicate which vectors to include in the analysis. The burnin can be
regulated at this point – a 1 indicates no burnin. Otherwise start should be larger
to include burn-in. These settings indicate start at vector 1 and end at vector
5000.
;
1 1 0.2                  # swapFreq, numswap, temperature  - settings for heating
of MCMC chain;
100000 90000 1000 1000   # max gen for species tree, burnin, savefreq, printfreq
setting for the second mcmc run. Usually the length of the chain here does not need
to be as long as in the first mcmc if the priors for part 1 have been effective in
generating the appropriate distribution of gene trees
;
1 3 1                 # 1:using inverse gamma_ prior of theta, gamma(alpha,beta).
These settings should be the same as in part 1;
6 32 32 32 32 32 32           #no_ of loci, no_ of ind per locus ;
MKL131_a MKL131_b CEF382_a CEF382_b CEF846_a CEF846_b PRS2684_a PRS2684_b CEF811_a
CEF811_b CEF825B_a CEF825B_b CEF823_a CEF823_b PRS2615_a PRS2615_b VGR666_a
VGR666_b MKL60_a MKL60_b MKL28_a MKL28_b CES716_a CES716_b CEF470_a CEF470_b
CEF873_a CEF873_b CEF871_a CEF871_b VGR631_a VGR631_b
... etc ...
#above are names of the terminals for each gene tree in order ;
```

```
1.000000 1.000000 1.000000 1.000000 1.000000 1.000000   #mutation ratio –
indidcates the same mutation rate for all loci (treelengths have been normalized in
part 1;
0 0 0 0 0 0                                            # 0 = diploid, 1=haploid ;
9 2 2 2 4 4 4 4 4 6                                       # no_ of species, and no.
of terminals (alleles) per species, this must be in the same order as sequences and
labels ;
MKL131_a MKL131_b CEF382_a CEF382_b CEF846_a CEF846_b PRS2684_a PRS2684_b CEF811_a
CEF811_b CEF825B_a CEF825B_b CEF823_a CEF823_b PRS2615_a PRS2615_b VGR666_a
VGR666_b MKL60_a MKL60_b MKL28_a MKL28_b CES716_a CES716_b CEF470_a CEF470_b
CEF873_a CEF873_b CEF871_a CEF871_b VGR631_a VGR631_b
#above are names of terminals in the species tree ;
```

OUTFILES FROM BEST PART 2:

filename.nex_treevector.t.t	main output from BEST part 2. This file has the posterior distribution of species trees that can then be summarized via consensus or other methods. These trees will have posterior estimates of the ancestral population sizes (theta) for each tree in square brackets.
filename.nex_treevector.t.p	Similar to the .p file of MrBayes and BEST part 1. This file contains the likelihoods and the prior likelihoods of each species tree sampled. sR refers to the birth paremeter in generating the priors on species trees, and eR indicates the extinction parameter for these prior trees.

The file filename.nex can then be executed in MrBayes or PAUP and the species tree file (filename.nex_treevector.t.t) imported. To get the majority rule consensus do sumt (in Mrbayes with no burnin) or majrul (in PAUP). Sometimes an "end;" has to be inserted into the treefile to import into these programs. The BEST tree has information on both speciation times and ancestral population sizes.You will need to use a perlscript to extract the estimaates of the ancestral population sizes [in square brackets] from the treevector file.

5. NOTES ON VERSION 1.6:

A. When you conduct a MCMC run in part 1 with nrun greater than 2, no genetree vectors or control file will be produced. This has been fixed in version 1.7 (available from L. Liu, soon available on website). The program produceoutfile can be used to construct gene tree vectors from single gene posterior distributions for use in BEST part 2. This program is available from Liang Liu.

B. Currently there is no way to indicate to the program the ploidy of X or Z chromosomes, i.e., ones that are _ the effective size of autosomes. Only organelle, hemizygous sex chromosomes (Y or W) and autosomal genes can be indicated at this point.

C. Make sure to designate a single sequence as the outgroup in BEST part 1.

D. When there is a mixture of ploidy values for a data set, version 1. 6 does not export the ploidy designations correctly in the control file. These ploidy designations need to bc changed manually before running best part 2

E. In the best block of parts 1 and in the control file of part 2, you need at least one space between the end of each line and the semicolon. If you get an error reading:

```
"allocating problem for sptree.nodes
Bugs in ReadControlfile
Error in command "Mcmc"
Error in command "Execute"
```

It probably means you have not put a space before each semicolon.