HIVtree Manual

Anna Nagel and Bruce Rannala

October 11, 2022

HIVtree is a Bayesian phylogenetic inference program that estimates HIV latent integration times and node ages on a fixed phylogenetic tree. This program was originally modified from PAML version 4.9. The program takes as input (1) a tree in newick format without branch lengths, (2) an alignment with tip dates, (3) a file with the list of latent sequences, and (4) a control file. The control file contains specification of the model and MCMC. The control file is very similar to that of mcmctree, a program in PAML. Here, only new or changed options to the control file will be detailed. We refer readers to the mcmctree manual for a full description.

1 Control File

clock: 1 must be used for the clock model, which specifies a strict clock. Other clock models are available in PAML but not in HIVtree.

latentFile: This is a the name of text file that provides the names of all of the latent sequences. The sequence names should match the names in the alignment and tree file with one name per line.

latentBound: This provides a hard upper bound on all of the latent integration times in the analysis. This is specified in backward time in the time units specified by the TipDate option. For example, consider the options "latentBound = 3", "tipDate = 1 1000", and time specified in days in the sequence names. This means no latent ages can be more than 3000 days older than the time of the last sample.

RootAge: This specifies the prior on the root age. There are two options, either a shifted gamma prior, $G(\alpha, \beta)$, or a uniform prior, U(a, b). The gamma distribution is shifted by adding the first sample time to the distribution. This ensures there is no density after sequences are sampled. The uniform prior has hard bound, so there is no density outside of the range between a and b. The parameters for both the uniform and gamma distributions must also be chosen with the time unit transformation going backward in time. For example, with option "tipDate = 1 1000" and the dates for the sequences specified in days, U(3,4) would be a uniform root age prior between 3000 and 4000 days prior to the *last* sample time. G(1, 1) would be a gamma prior with mean 1000 days prior to the *first* sample time with variance 1000 days. Note that the user specified prior will not match the induced prior when running without data (option usedata = 0) because of the constraints imposed by the tip ages and rank order of the node. The user should run without data to see what the induced prior will be.

2 Running the Program

The program is run in the same way as mcmctree. All path names in the control file must be relative to the current directory.

```
./HIVtree control.ctl
```

Example files are provided in the `examples` directory.

3 Combining results across genomic regions

For the combined analysis, the use must run the program on a unix computer using the bash shell and R must first be installed. The R packages "GoFKernel" and "kdensity" must also be installed. To combine the results from different regions of the genome from a single latent provirus, the user must create a file that specifies which latent sequences are from the same latent provirus.

For example, let's say there are two latent sequences, W14 and W19. These were split into two regions, C1C2 and C2C3 (both part of ENV). The sequences are named "C1C2_W14_QVOA_3921", "C1C2_W19_QVOA_3921", "C2C3_W14_QVOA_3921", and "C2C3_W19_QVOA_3921" in the fasta files. The user ran HIVtree on a phylogeny with C1C2 sequences and a phylogeny with C2C3 sequences. The user also ran HIVtree under the prior (usedata = 0) for both of those. For the analysis under the prior, "prior" was added to the latent sequence names. This is optional. This was done in the example so that sequence names are unique in the csv file for illustration purposes. The output of the MCMCs are in directories named "ENV2", "ENV3", ""ENV2_Prior", and "ENV2_Prior". To combine the estimates for the latency times for W14 in C1C2 and C2C3 and W19 for C1C2 and C2C3, the user should create a csv file as shown below.

```
ENV2,ENV3,ENV2_Prior,ENV3_Prior
C1C2_W14_QVOA_3921,C2C3_W14_QVOA_3921,C1C2_W14_QVOA_prior_3921,C2C3_W14_QVOA_prior_3921
C1C2_W19_QVOA_3921,C2C3_W19_QVOA_3921,C2C3_W19_QVOA_prior_3921,C2C3_W19_QVOA_prior_3921
```

The first row includes the directory names for all of the MCMC runs. The control file should contain the line"mcmcfile = mcmc.txt", which gives sets the name of the output file of the MCMC. The screen output of HIVtree should be saved to "output" in the same directory as the MCMC run, as is shown in the command line example below.

```
./HIVtree control.ctl &> output &
```

The runs with the data should be first in the csv file and then the runs without data (priors). The genes should be in the same order for the priors and posteriors. The following rows give the name of the latent sequence in each of the MCMC runs. These should match the names given in the fasta file for run. An improperly formatted csv file will result in the script not running correctly. If the script does not appear to give the correct output, check the to make sure the input file is in a csv file format with the same number of commas on each row and names that match the names in the fasta files. This program allows for missing data. The entry in the csv file should be blank if there is no sequence for a particular gene for a given latent provirus. This file is given as an input to parseMCMC.sh, which is run as follows.

```
./parseMCMC.sh sequences.csv
```

This will generate the files needed to run the analysis in R. There will be one less file generated than there are rows in the sequences.csv file. This is a csv file where each column is the estimate of a single latent integration time from a MCMC. If there are n different genomic regions, the first n columns will be the posterior distributions for the latent time and the $n+1$ to $2n$ columns will be the prior distributions. This will be in the same order as the input csv file. The names of the output files will match the name of the first sequence in each row of the csv file. For example, the above csv would have files names "C1C2_W14_QVOA_3921.txt" and "C1C2_W19_QVOA_3921.txt". If there is missing data, the first non-missing sequence name will be used.

There are 5 arguments needed by the R script. The first is the file created by parseMCMC.sh. The second argument is the sample time of the latent sequence in the same time units as HIVtree. The third argument is the same the "latentBound" used in the MCMCs. This is used as an integration bound and should be the same for all MCMCs. The fourth argument is the time unit. This should match the second argument of TipDate used in the control file for HIVtree. If "tipDate = 1 1000", the fourth argument should be 1000. The assumes the sample tipDate was used for all of the runs of HIVtree. The fifth argument is the last sample date in the whole phylogeny. This assumes that the last sample date is the same for all of the phylogenies. This will be the same as the highest number at the end of the sequence names in the fasta file. The sixth argument is the number of genomic regions to be used. This should be half the number of columns in the csv file and is used for error checking. This cannot be greater than 10. This argument assumes no missing data. It correspond to the dimensions of the csv file. If there is missing data for some genes for a provirus, the number of genomic regions will be greater than the number genes used for analysis. The analysis will still work correctly with missing data.

```
Rscript combineEstimates.R C1C2_W14_QVOA_3921.txt 0 3.695 1000 3650 2
```

The RScript will print out the mean and 95% credible set of the posterior distribution for the latent integration time. The prior distributions are divided out for all genes, resulting in a prior for the latency time that is uniform between the lower and upper integration bounds. If the user wishes the view the posterior distribution, the plot command in the Rscript can be uncommented. Note that in some rare cases, the numerical integration of the kernel density estimate can fail in R. One reason this may occur is if the posterior densities for the genes analyzed do not overlap. Examples of this are given in the `examples/combineGenes` directory.