9 June 2011

A Step-by-Step Tutorial: Divergence Time Estimation with Approximate Likelihood Calculation Using MCMCTREE in PAML

by Jun Inoue, Mario dos Reis, and Ziheng Yang

In this tutorial we will analyze the data sets of Inoue et al. (2010). It is assumed that you have some basic knowledge of using the command line in Windows or Unix systems (e.g. Linux and MacOS). You need to download and install the PAML package from http://abacus.gene.ucl.ac.uk/software/paml.html. If you have previously downloaded the package make sure you have the latest version. There are executables for Windows (*.exe) but Unix users may need to compile the programs. Please follow the instructions in the PAML website to modify your operating system path variable. This is necessary so that you can call the programs from the command line without having to type their full folder (path) location.

DNA dataset

Rough estimation of the substitution rate

The first data set that we will analyze is the 12tr data set in the 12tr folder. File 12tr_Chi23.phy contains a nucleotide alignment divided into four partitions: the 1st and 2nd partitions are the 1st and 2nd codon sites for 12 mitochondrial proteins; and the 3rd and 4th partitions are the mitochondrial tRNA and RNAs respectively. There are 23 fish species for which we would like to estimate their divergence times.

The first step is to get an idea of the overall mutation rate in this data set. Go into the 12tr folder. There you will see a baseml.ctl file. This is the control file for the BASEML program in the PAML package. You can open this file with your favorite text editor (Notepad, TexEdit, Vim, etc.). This file contains the instructions for the program. Look at the PAML documentation for the meaning of each variable in this file. BASEML will estimate the branch lengths and substitution parameters by maximum likelihood (ML) under the strict clock, general reversible substitution model and discrete gamma rates. The phylogeny for the 23 species is in the Chi23PE.tree file. Please note that the root node has a '@4.5' tag. This is a simple fossil calibration of 450Ma ago for the root.

Make sure you are in the 12tr folder. At the command line type (don't type the '>' that just indicates the command line prompt):

```
> baseml
```

This will start the BASEML program. The program will read the baseml.ctl file and it will follow its instructions. The results will be written to the mlb file. Open this file with a text editor and look for the following

```
Substitution rate is per time unit
    0.062931 +- 0.002345
```

This is an estimate of the per site substitution rate for the nucleotide data set. We will use this

value to set the prior for the mean substitution rate in the Bayesian analysis.

Estimation of the gradient and Hessian

The second step is the estimation of the gradient and Hessian of the branch lengths for the 23 species tree topology. The gradient (*g*) and Hessian (*H*) are the vector and matrix of first and second derivatives of the log-likelihood function evaluated at the maximum likelihood estimates (MLE) of the branch lengths. The gradient and Hessian describe the shape of the log-likelihood surface around the MLEs, and they can be used to approximate the log-likelihood curve using Taylor expansion (for details see dos Reis and Yang 2011). The *g* and *H* must be calculated on the unrooted tree without the clock.

Open the mcmctree.ctl file in the 12tr folder with a text editor. This is the control file for the MCMCTREE program that will perform Bayesian inference of divergence times. The sequence file to be analyzed will be 12tr_Chi23.phy and the tree topology is in Chi23.tree. The tree file must be rooted and must not contain any branch lengths. If you open Chi23.tree with a text editor you will be able to see the fossil calibrations. They are indicated within single quotes, for example 'B(4.22,4.63)' for the root. The MCMCTREE documentation describes all the possible fossil calibrations in detail. The program will not use the fossil calibrations at this stage.

In the mcmctree.ctl file look for and set the line

usedata = 3

This tells MCMCTREE that we will not perform divergence time estimation yet. We only want to calculate *g*, *H* and the MLEs of the branch lengths. Making sure you are in 12tr type at the terminal

> mcmctree

This will start the MCMCTREE program. MCMCTREE will unroot the tree and write it to a temporary file. Then it will call BASEML to perform ML estimation of the branch lengths, *g* and *H* without the clock. Because we have four partitions, BASEML will estimate four sets of branches, four *g* vectors and four *H* matrices. The results will be written to a file called out.BV. Open this file and examine it in a text editor. The first line has the number of species for the first partition. Then there is the tree with branch lengths, the branch lengths on their own in another line, then the gradient (if there are no zero length branches, the gradient should be zero, or close to zero for all branches) and finally the Hessian matrix. If you browse down the file you will see the estimates for the other three partitions. Rename the out.BV file as in.BV.

You could for example, prepare a baseml.ctl file with an arbitrary substitution model. Then you could run BASEML directly for each partition without using MCMCTREE. If you decide to do this, you must make sure to use the right unrooted tree(s). MCMCTREE will unroot the tree in a particular way and the tree used for BASEML has to be unrooted in the same manner. It may be a good idea to run MCMCTREE with usedata=3 and kill the program, just so that you can use the tree(s) generated in the temporary file(s) (say, tmp1.trees). The gradient and Hessian will be written to a file called rst2. Combine those output files for the partitions and rename the resulting file to in.BV.

Estimation of divergence times with the approximate likelihood method

Using a text editor, open the file mcmctree.ctl again. Modify the usedata line to

usedata = 2

This now tells the program that we will perform Bayesian estimation of divergence times using the approximate likelihood method. You may also set

outfile = out_usedata2

Now look for a line starting with rgene_gamma. This sets the gamma prior for the overall substitution rate. The gamma distribution is described by the shape parameter (*a*) and by the scale parameter (*b*). Let's write *m* for the mean and *s* for the standard deviation of the gamma distribution, then *a*, *b*, *m* and *s* are related by the following equations:

$a = (m/s)^2$ and $b = m/s^2$.

We will use $m = 0.06$ (close to the rate estimated by BASEML in the first step) and we will set $s = 0.06$ as well. The shape parameter is

$a = (0.06/0.06)^2 = 1$,

and the scale parameter is

$b = 0.06/0.06^2 = 16.6666$.

So in the control file we set

rgene_gamma = 1 16.7

Because the shape parameter is equal to one, the gamma density is in fact reduced to an exponential density. This is quite a diffuse prior. R users may plot the resulting prior in an R console with

> curve(dgamma(x, 1, 16.7), from=0, to=0.2)

(If you're not an R user do not worry about the line above!). We will now specify the prior for the variance of the log-normally distributed substitution rate. Find and set the following line in the mcmctree.ctl file

sigma2_gamma = 1 4.5

(Later on, you may want to try changing the shape and scale parameters to some other values to see the effect on the posterior time estimates).

We now set the burnin and the number of samples

burnin = 50000
nsample = 10000

We can now run the program

> mcmctree

The program will then read the alignment for checking purposes, the tree file with fossil calibrations, and the in.BV file to be used in the approximate likelihood calculation. As the MCMC analysis progresses, look at the mixing proportions. They should be close to 30% (between 20-40% is good, and 15-70% is acceptable). If the mixing proportions are too low or too high, kill the program (Ctl + C), modify the finetune parameters in the mcmctree.ctl file and start the program again. The MCMCTREE manual describes in detail how to set up the fine-tuning.

Version 4.4e of the program implements an option of automatic adjustment of those fine-tune step lengths, using the burn-in to calculate the jump probability or the proportion of proposals that are accepted. Please look at the monitor and confirm that the final acceptance proportions are reasonable.

A summary of the results will be written to the out_usedata2 file. There you will see the estimated times for each ancestral node in the tree, as well as 95% credibility intervals (CI). The raw output from the MCMCTREE is written to the mcmc.out file. This file can be analyzed with the Tracer program from the Beast family, or with the CODA package in R.

Amino Acid Dataset

The other data set that we will analyze is the single partition amino acid data alignment in the AA_Chi23.phy file in the AA folder. They are simply the corresponding amino acid sequences of partitions 1 and 2 in the nucleotide data set. We will basically repeat every step as for the nucleotide data: estimate rough substitution rate; estimate branch lengths, *g* and *H*; and estimate divergence times.

Go into the AA folder. Because we are using amino acid data, we need to use CODEML (instead of BASEML) to estimate the rough substitution rate. Later on, MCMCTREE will also use CODEML to estimate the MLE of branch lengths, *g* and *H*. Copy the mtREV24.data file into the AA directory. This file is located in the dat folder from the paml package. From the terminal window, making sure you are at the AA folder, type

> codeml

This will run the ML analysis with the clock on the amino acid data set. Examine the output file and you should find the substitution rate

0.047131

Now edit the mcmctree.ctl file in the AA folder as to perform estimation of *g* and *H* without the clock. Make sure you set the line

seqtype = 2

This tells MCMCTREE that we will be analyzing amino acid sequences. Examine the substitution

model variables to work out what they mean. Now run the program from the terminal

> mcmctree

MCMCTREE will prepare the appropriate files and then will call CODEML to do the analysis. As previously, an out.BV file will be generated with the branch lengths, *g* and *H*. Rename it as in.BV and proceed with divergence time estimation. Edit the mcmctree.ctl file again as to set the rate prior according to rate estimated under the clock. Run the divergence time analysis. Are the divergence times estimated with the amino acid data similar to the nucleotide analysis?

Analysis of a Mixed Dataset

An advantage of the approximate likelihood method is that arbitrary data sets could be combined and analyzed together. For example, one partition could be amino acid data and another partition could be nucleotide data. Once the branch lengths, *g* and *H* have been calculated, there is no way MCMCTREE can tell what type of data was used to generate the in.BV file.

We will now analyse the amino acid data set for the 23 fish species together with the tRNA and RNA genes (the last two partitions in the nucleotide data set). Copy the in.BV file from 12tr into AAtr. Open it in a text editor. Delete the first two partitions (1st and 2nd codon sites), this means deleting two trees, two gradient vectors, two branch length vectors, and two matrices. Open the in.BV file from AA and copy everything from this file into the newly created AAtr/in.BV file. You now have a new in.BV file with tree partitions: amino acid sequences, tRNA and RNA genes.

Please note that you also need to create a new alignment file with three partitions (12tr_Chi23_P3.phy). The sequences themselves in the file are read but not used. However, each partition must have the same number of species present in the corresponding in.BV partitions. For example, you could be analyzing tRNA genes from 20 species and amino acid sequences from 30, so the alignment file would need to have 20 and 30 species in each partition respectively. The master tree (the tree with fossil calibrations) must have all 30 species. This is necessary because MCMCTREE will read the alignment file to check the number of species. If they don't match the in.BV file you may get an error!

Note that the sequence data file has to have only one type of data (either nucleotides or amino acids for example) and you should never mix DNA and amino acid sequences in one sequence data file. Instead you use baseml or codeml to analyze the data of the right type and then combine the branch lengths and Hessian matrices for the different partitions. Those branch lengths and Hessian matrices may be from DNA sequences for some partitions, from protein sequences for some other partitions, and from codon sequences for yet some other partitions, for example. Effectively we are cheating MCMCTREE into believing that the data are of the same data type, but in fact the in.BV file is generated from analyzing different types of sequence alignments.

Edit the mcmctree.ctl file in AAtr. Adjust the rate prior to a sensible number. You may also want to adjust the prior for sigma2. Run the analysis and compare the results with the other two MCMC analyses above. Are the divergence times similar?

Remember that each MCMC analysis should be run at least twice to check for convergence to

the posterior distribution. In this tutorial it means that you would have run MCMCTREE six times with usedata=2. If results from different MCMC runs (from the same control file) look too different, convergence was not achieved. You may try running the MCMC for a greater number of generations (i.e. increase the value of the sampfreq variable in the control file). You may also want to increase burnin or change finetune parameters to get as close to 30% of acceptance proportions as possible.

Good luck and enjoy divergence time estimation with MCMCTREE!

References

Inoue, J.G., Miya, M., Lam, K., Tay B.H., Danks, J.A., Bell, J., Walker, T.I., Venkatesh, B. 2010. Evolutionary origin and phylogeny of the modern holocephalans (Chondrichthyes: Chimaeriformes): a mitogenomic perspective. *Molecular Biology and Evolution*, **27:** 2576-2586.

dos Reis, M. and Yang Z. (2011) Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Molecular Biology and Evolution* (*in press*).