

# CAFE v2.0: Software for Computational Analysis of gene Family Evolution

Sang-Gook Han, Jeffery P. Demuth, and Matthew W. Hahn

November 26, 2007

The purpose of CAFE is to analyze changes in gene family size in a way that accounts for phylogenetic history and provides a statistical foundation for evolutionary inferences. The program uses a random birth and death process to model gene gain and loss across a user-specified phylogenetic tree. The distribution of family sizes generated under the random model can provide a basis for assessing the significance of the observed family size differences among taxa.

CAFE v2.0 is an updated, command line only, version of the original CAFE software (De Bie et al. 2006). This document describes how to download and use CAFE v2.0. Major updates in version 2.0 include: 1) the ability to estimate separate values of the birth-death parameter ( $\lambda$ ) for each branch, or a subset of branches in the phylogenetic tree; 2) an improved optimization algorithm for estimating the maximum likelihood value of  $\lambda$ ; 3) improved overall computational efficiency; and 4) the ability to carry out extensive simulations.

The necessary inputs for CAFE v2.0 are:

- 1) a **data file** containing gene family sizes for the taxa included in the phylogenetic tree
- 2) a **Newick formatted phylogenetic tree**, including branch lengths

From the inputs above, CAFE v2.0 will compute:

- 1) the **maximum likelihood value of the birth & death parameter,  $\lambda$** , over the whole tree or for user-specified subsets of branches in the tree
- 2) **ancestral states** of family sizes for each node in the phylogenetic tree
- 3) **p-values** for each gene family describing the likelihood of the observed sizes given random gain and loss.
- 4) **average gene family expansion** along each branch in the tree
- 5) numbers of **gene families with expansions, contractions, or no change** along each branch in the tree

## CITING CAFE

An appropriate citation for use of CAFE in published research is:

De Bie T, Cristianini N, Demuth JP, and Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22:1269-1271.

Or the first application of CAFE 2.0:

Hahn MW, Demuth JP, and Han S-G (2007) Accelerated rate of gene gain and loss in primates. *Genetics*. 177:1941-1949.

Original development of the statistical framework and algorithms implemented in CAFE are published in:

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research* 15(8):1153-1160.

## DOWNLOADING

CAFE v2.0 is available from:

**<http://www.bio.indiana.edu/~hahnlab/Software.html>**

CAFE v2.0 is available in compiled versions for Mac OSX (PowerPC and Intel). Alternatively you may download and compile the source code yourself.

## LAUNCHING

CAFE v2.0 is implemented as a shell. The program can be run interactively by simply launching the shell, or the user may execute a shell script that lists a series of commands saved as a separate text file.

**ENTER CAFE SHELL:** To run CAFE interactively, launch the shell by typing `shell` at the unix prompt. If all is well the prompt should change to `#`. You may now begin inputting **commands**. To exit the shell, type `exit`.

**EXECUTING CAFE SHELL SCRIPTS:** Because many analyses will require a similar series of inputs, you may also run CAFE using a shell script. CAFE scripts should be saved as text files with unix line endings. Scripts may be executed from the unix prompt or from within the CAFE shell.

Example CAFE script :

```
#!/shell
#version
#date
load -i data/example2.tab -t 10 -l logfile.txt
tree (((chimp:6, human:6):81, (mouse:17, rat:17):70):6, dog:93)
lambda -s -t (((1,1)1, (2,2)2)2,2)
report lambda
```

In this example, the first line indicates the location of the CAFE shell program just like general shells, such as `bash` or `perl`. Subsequently, lines beginning with “`#`” are regarded as remarks. Thus, the example above only executes lines 4, 5, 6, and 7. Note that to run a script, you must make the file executable from the unix prompt, by `chmod +x filename`. CAFE will automatically exit after the last command in the script is completed, so it is not necessary to specify `exit`.

An example script is downloadable from the CAFE website.

## COMMANDS

Index links to commands:

<b>Command</b>	<b>Brief Description</b>
<code>source</code>	run shell script
<code>version</code>	version info
<code>date</code>	date/time
<code>exit</code>	exit CAFE shell
<code>log</code>	log output
<code>load</code>	data file & run parameters
<code>tree</code>	phylogenetic tree
<code>lambda</code>	find/specify birth-death parameter
<code>pvalue</code>	identify rapidly evolving families/branches
<code>report</code>	report values
<code>genefamily</code>	generate simulated data
<code>lhtest</code>	compare likelihoods of lambda models

# `source` filename

Load shell script file.

# `version`

Display the CAFE version number

# `date`

Display the current date and time

# `exit`

Exit the CAFE shell. `quit` will perform the same action.

# `log` [filename]

If no filename is given, this command displays the current log file. If a filename is specified, CAFE will create (or overwrite) the log file. Default: stdout (output to screen only). The log file may also be specified in the `load` command using the `-l` option.

# `load` -i filename [-t integer | -l filename]

-i: DATA FILE: Enter the path to the file containing gene family data. The data file format must be tab delimited with Unix line endings. Family description may contain spaces (but not tabs). The first line must contain labels in the order: Description, ID, and then the names of each taxon separated by tabs. *If you do not have a Description or ID, CAFE still requires two tabs at the beginning of each line.* The taxon names must be spelled exactly as they are in the **tree structure**. Subsequent lines each correspond to a single gene family. If the data file contains taxa that do not appear in the tree structure, they are not considered in the analysis.

**Example Data File:**

DESCRIPTION	ID	Chimp	Human	Mouse	Rat	Dog
EF 1 ALPHA	ENSF000000000004	5	8	6	12	40
HLA CLASS II	ENSF000000000007	4	4	3	3	3
HLA CLASS I	ENSF000000000014	5	3	5	6	3
RAG 1	ENSF000000000015	1	1	1	1	1
IG HEAVY CHAIN	ENSF000000000020	32	42	51	60	18
ACTIN	ENSF000000000027	27	30	22	28	25
OPSIN	ENSF000000000029	2	2	2	2	2
HEAVY CHAIN	ENSF000000000030	25	25	23	24	18

If the file is loaded correctly, CAFE will output summary information about the current data file to the logfile.

An example data set is available for download from the CAFE website.

-t: The maximum number of CPU threads to be used. Default: 8

-l: LOG FILE: Enter the path and filename where CAFE will write the **main output**. This file will contain a summary of input parameters as well as details of  $\lambda$  searches, including likelihood scores and maximum likelihood values of  $\lambda$ . If the file does not exist, CAFE will create it for you; if the file already exists, CAFE will append results to the previous file. Default: output to screen (no log file will be created).

# **tree** tree\_structure

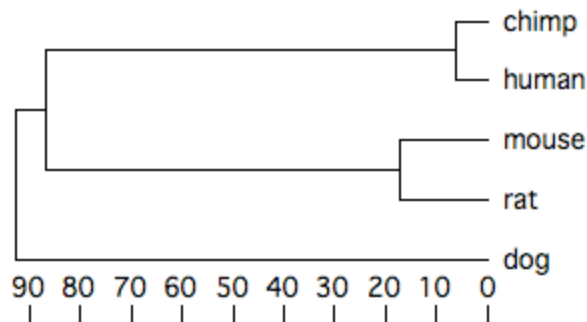
TREE STRUCTURE: A Newick formatted tree containing branch lengths and taxon names as they are specified in the input file. Branch lengths should be integer units of time and ultrametric (i.e. the sum of lengths from root to tip should be the same for all paths). For instructions on converting trees to Newick format visit: <http://evolution.genetics.washington.edu/phylip/newicktree.html>.

*Warning:* There should be no spaces in the tree string and no semicolon at the end of the line.

Example:

In Newick format, the tree diagram below is represented as:

(( (chimp:6, human:6) :81, (mouse:17, rat:17) :70) :6, dog:93)



# **lambda** [-l values | -s | -r start:step:end ]-t lambda\_structure [-e | -o filename]

*Warning:* The **load** and **tree** commands should be run prior to the **lambda** command.

-l: SPECIFY  $\lambda$  VALUES. This option allows the user to specify the value(s) of  $\lambda$  rather than performing a search. If more than one  $\lambda$  is specified, then the **-t** option must also be used.  $\lambda$  values are specified in the order 1 2 3 ... which correspond to the integers specified in the **-t** option.  $\lambda$  values should be separated by spaces.

*Warning:* 1) The product of  $\lambda$  and the depth of the tree structure should not exceed one (i.e.  $\lambda * t < 1$  must be true; where  $t$  is the time from the tips to the root). See Known Limitations for details to diagnose this problem.  
2) CAFE will use the last  $\lambda$  value(s) estimated (or user specified) to compute ancestral gene family sizes and to run Monte Carlo simulations

-s: SEARCH FOR  $\lambda$  VALUE(S). CAFE v2.0 will search using an optimization algorithm to find the value(s) of  $\lambda$  that maximize the log likelihood of the data for all families. CAFE starts with an intermediate value and then searches iteratively for the best value for  $\lambda$  (or set of  $\lambda$  values if used in conjunction with the **-t** option). Subsequent analyses will automatically use the results from the lambda search.

-r: SEARCH  $\lambda$  IN SPECIFIED RANGE. Returns the likelihood scores for  $\lambda$  values in the user specified range. The format of a range is start:step:end. For example, to see score distribution with a lambda between 0.003 and 0.005, the range would be 0:003:0.001:0.005. In case of more than one lambda, ranges are separated by space.

-t:  $\lambda$  STRUCTURE. To investigate whether different parts of the tree are evolving at different rates, the user must specify which branches of the tree will take the same or different lambda values. Input the same Newick tree structure as in the **tree** command, but exclude branch lengths and substitute integer values from 1 up to  $n$  for taxon names ( $n$  = total number of branches in the tree; matching integer values will take the same value of  $\lambda$ ). Default: all branches have the same value for  $\lambda$ .

Example  $\lambda$  structure:

The following lambda structure specifies one  $\lambda$  value for the Human, Chimp, and Ape, and a second lambda value for all other branches base on the tree specified above.

(( (1, 1) 1, (2, 2) 2) 2, 2)

*Warning:* CAFE will not always converge to a single optimum with models that contain many parameters. See Known Limitations for details on assessing this problem.

**-e:** SEARCH  $\lambda$  FOR EACH FAMILY INDIVIDUALLY. Use this option to estimate  $\lambda$  for each family separately. The **-o** option can be used in concert with **-e** to write output for each family to summary files. Raw output is written to the log file specified in the **load** command.

*Warning:* If the **-e** option is used, CAFE will not estimate  $\lambda$  value(s) for all families combined and thus cannot complete ancestral node assignments or Monte Carlo simulations.

**-o:** OUTPUT FILE. Use in conjunction with **-e**. Enter the name of the output file. Do not assign an extension as it may interfere with the metapost file conversion (see `filename.mp`). CAFE will write the following 3 files:

`filename.lambda`: A tab delimited summary of  $\lambda$  values for each family in Newick notation. CAFE outputs the tree from the **load** command, with the number of genes and the  $\lambda$  estimate for each branch inserted (taxon<#of genes>\_lambda value:branch length)

`filename.mp`: This file contains the same information as `filename.lambda`, but in metapost format for conversion to pdf files. The user should create a directory called "pdf" and move the `filename.mp` file into the pdf directory before running the conversion (you will need to have a metapost interpreter installed to utilize this functionality). The end result will be an individual pdf file for each family illustrating numbers of genes and  $\lambda$  values on each branch of the tree.

`filename.html`: This is an index which allows the user to easily link to the pdf results for each family (output of the metapost file conversion).

**# pvalue** -r integer -p real\_number [branchcutting]

The **pvalue** command will generate the expected distribution of family sizes under the stochastic birth-death model for the tree specified in the **load** command. The last value of  $\lambda$  specified or found by search is used. Running the simulations uses the most machine resources and thus is the most time intensive step in CAFE.

**-r:** NUMBER OF RANDOM SAMPLES: To determine the probability of a gene family with the observed sizes among taxa, CAFE uses a Monte Carlo re-sampling procedure. Enter the number of samples CAFE should use to calculate p-values. The tradeoff is between precision and computation time; in most cases 1000 samples should provide reasonable balance. Default: 1000

**-p:** P-VALUE THRESHOLD: For each family in the data file, CAFE computes a probability (p-value) of observing the data given random gain and loss of genes. All else being equal, families with more variance in size are expected to have lower p-values. The p-value threshold allows the user to specify the cutoff for subsequent analyses. Families with p-values larger than the designated threshold will not be included in identification of the most unlikely branch. Default: 0.01

**branchcutting**: This option provides an additional method for identifying rapidly evolving branches. The branch-cutting method calculates whether the overall p-value associated with a gene family increases when the probabilistic coupling between the parent and child family sizes for a given branch is removed. A p-value is computed for the gene family given the tree with one branch removed as a model (and this is done for each branch separately). If the p-value increases considerably after cutting a branch, this branch may be held responsible for the overall low p-value of the complete model.

Default: The Viterbi method is the default option for identifying rapidly evolving branches. This method calculates exact p-values for transitions between the parent and child family sizes for all branches of the phylogenetic tree. A low p-value indicates a rapidly evolving branch.

Additional details germane to methods of branch identification are presented in Hahn et al. 2005.

```
# report [lambda | ancestor | pvalue] filename
```

The **report** command allows the user to choose what output to see. Although all analyses must be specified by their own commands to run, **report** specifies the output of CAFE.

**lambda**: Reports the likelihood and parameter value(s) from the **lambda** command. If the **-l** option was used to specify  $\lambda$ , the likelihood of the data as well as the original  $\lambda$  value(s) are returned. If the **-s** option was used, the likelihood of the best  $\lambda$  parameters as well as their values are returned. The **lambda** command must be invoked separately, or no output is given.

**ancestor**: Reports the maximum likelihood values of all ancestral states for all gene families. The **lambda** command must be invoked separately, or no output is given.

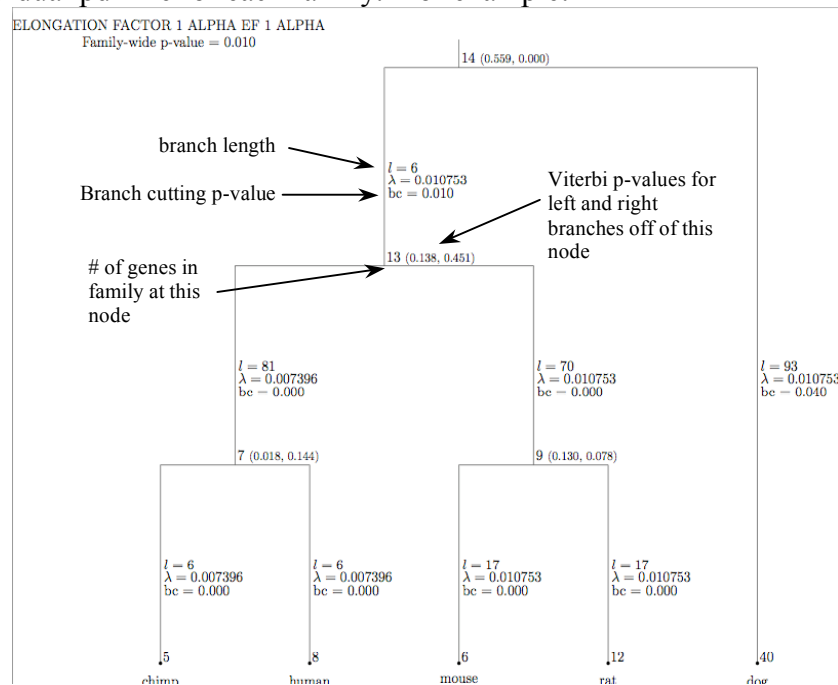
**pvalue**: Reports the overall p-values for each family, as well as the branch-specific p-values for families below the threshold specified in the **pvalue** command. The **pvalue** command must be invoked separately, or no output is given.

**filename**: The file where CAFE will write the main results of gene family analysis. Do not add an extension to the file name. The name of the report file should be different than the filename given for **lambda -o** output, otherwise output may be overwritten. As with the **lambda -o** option, 3 files will be created:

`filename.cafe`: A tab delimited summary of results.

CAFE outputs: 1) the current tree, 2) current lambda(s) and likelihoods, 3) lambda tree structure, 4) ID numbers of each node in tree format, 5) output formats according to node ID numbers, 5) average expansion (mean number of genes gained or lost per family, where “-“ expansion is a net contraction), 6) number of families with expansions, 7) contractions, or 8) no change. Finally, CAFE lists information for each family including the description, Newick format tree with ancestral numbers of genes at each node, the family-wide p-value, branch-specific p-values (from the Viterbi method). If the `branchcutting` option was specified, the string of p-values from this method is listed following the Viterbi results. A key to mapping the output to each node is also provided in report.

`filename.mp`: This file contains the same family specific information as `filename.cafe`, but in metapost format for conversion to pdf files. The user should create a directory called “pdf” and move the `filename.mp` file into the pdf directory before running the conversion (you will need to have a metapost interpreter installed to utilize this functionality). The end result will be an individual pdf file for each family. For example:



`filename.html`: This is an index which allows the user to easily link to the pdf results for each family (output of the metapost file conversion).

```
# genefamily directory/fileprefix -t integer
```

The `genefamily` command generates simulated data based on the properties of observed data. These simulated data can be used for many purposes, including generating null distributions of likelihood ratios to assess the significance of multi-parameter models (see `lhtest`). CAFE uses the estimated root sizes from the observed

data, the most recently specified global lambda (either by search or user input), and the tree specified in the `tree` command to generate new data sets. Each simulated data file will contain the same number of families and distribution of root sizes as the observed data. To specify the data file and the value of lambda (or lambda search), the `load` and `lambda` commands must precede `genefamily`. Note that because of the dependence of the `lhtest` command on `genefamily`, you are limited to simulating data using a single, global lambda. So if your script estimates a multiple parameter model, you will also need to precede `genefamily` with an additional `lambda` command to assign or search for the global parameter that you wish to be used for simulation (see example script on the website for an example).

`directory/fileprefix`: Designates the directory where CAFE will write the simulated data sets. This should be specified relative to the working directory, and must be created prior to running CAFE (i.e. CAFE will not create the directory). Each simulated data set will have the name `fileprefix_#.tab`.

`-t integer`: Specifies the number of simulated data sets for CAFE to generate

Example: `genefamily rndtree/rnd -t 100` This command line will generate the simulated data sets: `rnd_1.tab`, `rnd_2.tab` ... `rnd_100.tab` and write them in the “`rndtree`” directory. These data sets will have the same format as typical CAFE input.

```
# lhtest -d directory -l lambda_seed_value -t lambda_structure -o filename
```

The `lhtest` command computes two likelihood scores for each simulated data set: one based on a model with a single, global lambda and one based on a model with multiple lambdas. Because the simulated data from `genefamily` have been generated under a single-parameter model, the results of `lhtest` can be used to assess the significance of models with more than one lambda parameter by comparing the likelihood ratio of the observed data to the distribution of ratios generated by analysis of simulated data sets [ $LR = 2^{(\text{score of global lambda model} - \text{score of multi-lambda model})}$ ]. A multi-parameter model is significantly better than a single-parameter model if the observed LR is greater than 95% of the distribution of simulated LRs.

`-d directory`: Specifies the directory where CAFE can find the simulated data sets generated by the `genefamily` command.

`-l lambda_seed_value`: For each simulated data set CAFE will begin the lambda search algorithm at the value specified here. Since CAFE re-estimates lambdas for each simulated data set, specifying a seed value close to the actual lambda used in the simulations may save considerable time.

`-t lambda_structure`: specify the lambda structure as in the `lambda` command above. This command specifies the multi-lambda model to be estimated.

`-o filename`: the file where CAFE will write output of `lhtest`. The file contains columns for each of the following values: likelihood score for global lambda |

estimated global lambda | likelihood score for multiple lambda model | estimated lambda 1 | estimated lambda 2 |...| estimated lambda n

## KNOWN LIMITATIONS

1. Because the random birth and death process assumes that each family has at least one gene at the root of the tree, CAFE will not provide accurate results if gene families are included that are not present in the most recent common ancestor (MRCA) of all taxa included in the tree. For example, even if all included taxa have gene family size = 0, CAFE will assign the MRCA a gene family size of 1, and include the family in estimation of the birth and death rate. This difficulty does not affect analyses containing families that go extinct subsequent to the root node.
2. If a change in gene family size is very large on a single branch, CAFE may fail to provide accurate lambda estimation and/or die during computation. To see if this is a problem, look at the likelihood scores computed during the  $\lambda$  search (reported in the logfile if the job finishes). If ALL scores = “-inf” then there is a problem with large size change giving CAFE a probability = 0. Removing the family with the largest difference in size among species and rerunning CAFE should allow  $\lambda$  to be estimated on the remaining data. If the problem persists remove the family with the next largest difference and proceed in a like manner until CAFE no longer finds families with zero probability. However, if rapidly evolving families are removed, care should be taken in interpretation of the estimated rate of evolution for the remaining data.
3. If the product of  $\lambda$  and the distance from the tips to the root is greater than 1, then CAFE will not return accurate results. If  $\lambda$  is specified by the user, this problem is seen as @@. If the  $\lambda$ -search option is used, then the value of  $\lambda$  output will be the maximum value possible for  $\lambda * t < 1$ . If this is a problem, CAFE will print a caution message and “@@” will appear before the Newick-formatted tree in the output. ***In our experience, this is the most common error encountered by users.***
4. In very large phylogenetic trees there can be many independent lambda parameters ( $2N-2$  in a rooted tree, where  $N$  is the number of taxa). CAFE does not always converge to a single global maximum with large numbers of  $\lambda$  parameters, and therefore can give misleading results. To check for this, you should always run the `lambda` search multiple times to ensure that the same estimated values are found. Also, the likelihood of models with more parameters should always be lower than models with fewer parameters, which may not be true if CAFE has failed to find a global maximum. If CAFE does not converge over multiple runs then one should reduce the number of parameters estimated and try again.