

# VI\_coalHMM

## Description

- Black box variational inference of evolutionary parameters (node heights and population sizes of each internal node and internal branch) on a species tree under the coalescent with recombination. There is also an option to infer branch lengths instead of node heights. The data is a sequence alignment of recombinant DNA. We estimate node heights (or branch lengths) in the unit of generations and population sizes in the unit of individuals. We estimate the mean and standard deviation of the posterior of each parameter.
- When inferring the branch lengths of the species tree instead of node heights, we only infer the lengths for some branches because the tree is assumed to be ultrametric. For details of which branch lengths are inferred, please see section Notes on Branch Lengths.
- We use `mspms`, an `ms`-compatible command-line interface to the `msprime` library. Details of installation can be found [here](#). Please make sure `mspms` (usually in `~/local/bin`) is on your system search path.
- We use BEAGLE, a high-performance library to calculate the "Felsenstein Likelihood". Full details of installation instructions can be found [here](#), always follow "Installing from source".
  - [Instructions for installing BEAGLE on Mac OS X](#)
  - [Instructions for installing BEAGLE on Windows](#)
  - [Instructions for installing BEAGLE on Linux](#)
  - If there is an error, try the command: `java -Djava.library.path="/usr/local/lib" -jar PhyloNet.X.X.X.jar script nex`

## Usage

```
VI_coalHMM [-bl] -st startingTree -mu mutationRate -rho recombinationRate -r crossoverRate -nb numSubBranch [-len simulationShortRegionLength] [-nhsigma nodeHeightInitialSigma] [-pssigma popSizeInitialSigma] [-blsigma branchLengthInitialSigma] [-psp popSizePrior] [-n0 NOForMS] [-ns samplePerIter] [-niter numIter] [-nhmeanlr nodeHeightMeanLearningRate] [-psmeanlr popSizeMeanLearningRate] [-blmeanlr branchLengthMeanLearningRate] [-nhsigmalr nodeHeightSigmaLearningRate] [-pssigmalr popSizeSigmaLearningRate] [-blsigmalr branchLengthSigmaLearningRate] [-nhsigmamin nodeHeightSigmaMinimum] [-pssigmamin popSizeSigmaMinimum] [-blsigmamin branchLengthSigmaMinimum]
```

Parametrization setting		
-bl	Infer branch length of the branches in the species tree, instead of node heights of internal nodes. This option can reduce sampling illegal tree configurations when estimating the gradient of the ELBO by Monte Carlo samples.	optional
Starting State Settings		
-st <i>startingTree</i>	Specify the starting tree topology and node heights. The input tree should be ultrametric with branch lengths in units of generations. For example, ((H:160000, C:160000):60000, G:220000); species a three-taxon tree with an internal node height of 160,000 generations and root node height of 220,000 generations. See the example below.	mandatory
-mu <i>mutationRate</i>	The mutation rate in unit of expected number of mutations per site per generation. For example, 2.5e-8.	mandatory
-rho <i>recombinationRate</i>	The recombination rate in unit of expected number of recombinations per site per generation. For example, 1.5e-8.	mandatory
-nhsigma <i>nodeHeightInitialSigma</i>	The starting standard deviation of the variational posterior of each node height. The default value is 20,000. (Only used when -bl is not set.)	optional

-pssigma <i>popSizeInitialSigma</i>	The starting standard deviation of the variational posterior of each population size. The default value is 10,000.	optional
-blsigma <i>branchLengthInitialSigma</i>	The starting standard deviation of the variational posterior of each branch length. The default value is 20,000. (Only used when -bl is set.)	optional
<b>Prior Settings</b>		
-psp <i>popSizePrior</i>	Mean value of the prior of population sizes. The default value is 50,000.	optional
<b>Likelihood Simulator Settings</b>		
-n0 <i>N0ForMS</i>	N0 for <code>ms</code> . The default value is 10,000. For details see <a href="#">ms documentation</a> (subsection "Two species with population size differences" in section "Some examples") and our paper.	optional
-r <i>crossoverRate</i>	The cross-over rate that determines the length of simulation for building coalHMM. For details see <a href="#">ms documentation</a> ("Crossing over") and our paper. Can use 1,000 as a starting point.	mandatory
-nb <i>numSubBranch</i>	The number of sub-branches on each internal branch of the species tree for refining coalHMM state space. For details see our paper. Can use 2 as a starting point.	mandatory
-len <i>simulationShortRegionLength</i>	Simulating multiple independent short regions when building the HMM could save time compared to simulating a long region. This parameter is the length of each independent short region simulation. The default value is 5,000.	optional
<b>BBVI Settings</b>		
-ns <i>samplePerIter</i>	The number of samples per iteration of BBVI for estimating gradient. The default value is 50.	optional
-niter <i>numIter</i>	The number of iterations of BBVI. The default value is 200.	optional
-nhmeanlr <i>nodeHeightMeanLearningRate</i>	Learning rate for the mean parameter of the variational posterior of node heights. The default value is 20,000. (Only used when -bl is not set.)	optional
-psmeanlr <i>popSizeMeanLearningRate</i>	Learning rate for the mean parameter of the variational posterior of population sizes. The default value is 10,000.	optional
-blmeanlr <i>branchLengthMeanLearningRate</i>	Learning rate for the mean parameter of the variational posterior of branch lengths. The default value is 20,000. (Only used when -bl is set.)	optional

-nhsigmalr <i>nodeHeightSigmaLearningRate</i>	Learning rate for the standard deviation parameter of the variational posterior of node heights. The default value is 500. (Only used when -bl is not set.)	optional
-pssigmalr <i>popSizeSigmaLearningRate</i>	Learning rate for the standard deviation parameter of the variational posterior of population sizes. The default value is 500.	optional
-blsigmalr <i>branchLengthSigmaLearningRate</i>	Learning rate for the standard deviation parameter of the variational posterior of branch lengths. The default value is 500. (Only used when -bl is set.)	optional
-nhsigmamin <i>nodeHeightSigmaMinimum</i>	The minimum value of the standard deviation of node heights variational posterior. Since BBVI is possible to reach a negative standard deviation if the learning rate is not set carefully, a minimum value is required so that the standard deviation would not drop below the specified value during BBVI searches. The default value is 10,000. (Only used when -bl is not set.)	optional
-pssigmamin <i>popSizeSigmaMinimum</i>	The minimum value of the standard deviation of population sizes variational posterior. The default value is 3,000.	optional
-blsigmamin <i>branchLengthSigmaMinimum</i>	The minimum value of the standard deviation of branch length variational posterior. The default value is 10,000. (Only used when -bl is set.)	optional

## Example

Download: [test.nex](#)

```
#NEXUS
Begin data;
Dimensions ntax=3 nchar=500000;
Format datatype=dna symbols="ACTG" missing=? gap=-;
Matrix

H TCGCTGTCTCATACTATATGGAGAGTCAAGGGGGTTGAGATAATTGTCGCATTGTCTAAGTGAATGGCGTAAAGCGAAC.....
C CCGCTGTCTCATACTATATGGAGAGTCAAGGGGGTTGAGATAATTGTCGCATTGTCTAAGTGTATGGCGTAAAGCGAAC.....
G TCGCTGTCTCATACTATATGGAGAGTCAAGTGGGTTGAGATAATTGTCGCATTGTCTAAGTGAATGGCGTAAAGCGAAC.....
;End;

BEGIN TREES;
Tree t0 = ((H:150000,C:150000):150000, G:300000);
END;

BEGIN PHYLONET;
VI_coalHMM -st (t0) -mu 2.5e-8 -rho 1.5e-8 -r 1000 -nb 2 -psp 50000 -nhsigma 20000 -pssigma 10000 -n0 10000 -ns 50 -niter 200 -nhmeanlr
20000 -psmeanlr 10000 -nhsigmalr 500 -pssigmalr 500 -nhsigmamin 10000 -pssigmamin 3000;
END;
```

This command will run `VI_coalHMM` for the data given. It will infer the divergence times of HC ancestor and HCG ancestor, as well as the population sizes of HC ancestor and HCG ancestor. The starting tree is `((H:150000,C:150000):150000, G:300000);`. That is, we start the search with HC ancestor divergence time of 150,000 generations and HCG ancestor divergence time of 300,000 generations. Note that the Newick string must be given in the TREES section and referenced in the PHYLONET section. The mutation rate is set to 2.5e-8 mutations per site per generation. The recombination rate is set to 1.5e-8 recombinations per site per generation. The cross-over rate `-r` is set to 1000 and the number of sub-branches `-nb` is set to 2. For details of `-r` and `-nb` see our paper. All other parameters are set as default.

## Note on Branch Lengths

When inferring the branch lengths of a species tree, since the tree is assumed to be ultrametric, we only need to infer the length of some branches in order to fully characterize the tree. During each iteration of the inference and at the end of the inference, our program will print the lengths of the branches inferred in the order described in the next paragraph.

For each node of the species tree, we infer branch lengths according to the following policy. Leaf nodes do not incur branch lengths. If an internal node has two leaf children, incur a branch length of the branch connecting the node to its left child. If an internal node has one leaf child and one non-leaf child, incur a branch length of the branch connecting the node to its non-leaf child. If an internal node has two non-leaf children, incur a branch length of the branch connecting the node to its left child. We walk the species tree nodes in postorder and add branch lengths to infer according to this policy. The branch lengths printed at each iteration of the inference and at the end of the inference are ordered this way.

## Note on Learning Rates

Users can set separate learning rates for four kinds of free parameters of variational posterior: mean of node heights, standard deviation of node heights, mean of population sizes, and standard deviation of population sizes. These learning rates need to be set very carefully so that BBVI can converge quickly. During the BBVI search, `VI_coalHMM` will print the gradient of each parameter (mean and standard deviation of each demographic parameter) to the console. It is recommended that the user set the four learning rates according to the scale of the gradient of each parameter so that the step size of each parameter in each iteration is reasonable.

If learning rates or starting states are not set properly, you may often see the warning "Illegal value sampled this iteration." This happens when an illegal configuration is sampled during BBVI gradient estimation (For example, child node has a higher node height than the parent node). If you see this message a lot, please change starting states and learning rates so that the variational posterior of node heights do not overlap and the variational posterior of population sizes do not cover negative values.

## Command References

1. Xinhao Liu, Huw A. Ogilvie, and Luay Nakhleh. Variational Inference Using Approximate Likelihood Under the Coalescent With Recombination.