MCMC_SEQ

Description

Download PhyloNet

- Co-estimation of reticulate phylogenies (ILS & hybridization), gene trees, divergence times and population sizes on sequences from multiple independent loci.
    - For species phylogeny or phylogenetic network, we infer network topology, divergence times in units of expected number of mutations per site, population sizes in units of population mutation rate per site, and inheritance probabilities.
    - For gene trees, we infer gene tree topology and coalescent times in units of expected number of mutations per site.
    - To convert the divergence times/coalescent times to units of years, or to coalescent units, see our paper for details (Page 3, Lines 36-43).
- 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 error, try command: *java -Djava.library.path="/usr/local/lib" -jar PhyloNet_X.X.X.jar script.nex*

Usage

```
MCMC_SEQ -loci locusList [-cl chainLength] [-bl burnInLength] [-sf sampleFrequency] [-sd seed] [-pl parallelThreads] [-dir outDirectory] [-mc3 temperatureList] [-mr maxReticulation] [-tm taxonMap] [-fixps popSize] [-varyps] [-pp poissonParameter] [-dd] [-ee] [-sgt startingGeneTrees] [-snet startingNetwork] [-sps startingPopSize] [-pre preBurnIn] [-gtr paramList] [-diploid diploidSpeciesList] [-murate] [-mupi paramList] [-muweight paramList]
```

MCMC Settings		
-loci *locusList*	The list of loci used in the inference. For example, *-loci (YNR008W,YNL313C)* indicates the inference is performed on two loci YNR008W and YNL313C. See the format of multilocus data here. Note that our method is able to handle missing data, see the example below.	optional
-cl *chainLength*	The length of the MCMC chain. The default value is 10,000,000.	optional
-bl *burnInLength*	The number of iterations in burn-in period. The default value is 2,000,000.	optional
-sf *sampleFrequency*	The sample frequency. The default value is 5,000.	optional
-sd *seed*	The random seed. The default seed is 12345678.	optional

-pl *parallelThreads*	The number of threads running in parallel. The default value is the number of threads in your machine.	optional
-dir *outDirectory*	The absolute path to store the output files. The default path is your home directory.	optional
MC3 Settings		
-mc3 *temperatureList*	The list of temperatures for the Metropolis-coupled MCMC chains. For example, *-mc3 (2.0, 3.0)* indicates two hot chains with temperatures 2.0 and 3.0 respectively will be run along with the cold chain with temperature 1.0. By default only the cold chain will be run. Note that • The temperatures should be DIFFERENT! For example, *-mc3 (2.0, 2.0, 3.0)* is invalid. • The temperature of the cold chain should NOT be included. For example, *-mc3 (1.0, 2.0, 3.0)* is incorrect. • Metropolis-coupled MCMC leads to faster convergence and better mixing, however, the running time increases linearly with the number of chains. We suggest you first run a standard MCMC chain (cold chain) without this command. If the trace plot indicates the chain is not mixed well (jagged, stuck in local maxima for a long time), then try this command.	optional
Inference Settings		
-mr *maxReticulation*	The maximum number of reticulation nodes in the sampled phylogenetic networks. The default value is 4.	optional
-tm *taxonMap*	Gene tree / species tree taxa association. By default, it is assumed that only one individual is sampled per species in gene trees. This option allows multiple alleles to be sampled. For example, the gene tree is *(((a1,a2),(b1,b2)),c);* and the species tree is *((a,b),c);*, the command is *-tm <a:a1,a2; b:b1,b2;c:c>*. Note that the taxa association should cover all species, e.g. *-tm <a:a1,a2; b:b1,b2>* is incorrect because *c:c* is dropped out.	optional
-fixps *popSize*	Fix the population sizes associated with all branches of the phylogenetic network to this given value. By default, we estimate a constant population size across all branches.	optional
-varyps	Vary the population sizes across all branches. By default, we estimate a constant population size across all branches.	optional
-murate	Enabling the delta exchange operator for modeling varying substitution rates across loci.	optional
Prior Settings		
-pp *poissonParam*	The Poisson parameter in the prior on the number of reticulation nodes. The default value is 1.0.	optional
-dd	Disable the prior on the diameters of hybridizations. By default this prior on is exp(10).	optional

-ee	Enable the Exponential(10) prior on the divergence times of nodes in the phylogenetic network. By default we use Uniform prior.	optional
Starting State Settings		
-sgt	Specify the starting gene trees for each locus. Comma delimited list of gene tree identifiers. See details. The gene trees should be ultrametric trees with coalescent times in units of expected number of mutations per site. See example below. The default starting gene trees are UPGMA trees.	optional
-snet	Specify the starting network. The input network should be ultrametric with divergence times in units of expected number of mutations per site, inheritance probabilities and population sizes in units of population mutation rate (optional). See example below. The default starting network is the MDC trees given starting gene trees.	optional
-sps	Specify the starting population size. The default value is 0.02. See example below.	optional
-pre	Specify the number of iterations for pre burn-in, e.g. "-pre 20" means 20x *sampleFrequency* iterations will be run before the MCMC chain starts. By default, we run 10x *sampleFrequency* iterations for pre burn-in.	optional
Substitution Model		
-gtr *paramList*	Set GTR (general time-reversible) as the substitution model. The first four parameters in the list represent base frequencies for A, C, G, T. The rest six parameters represent transition probabilities for A>C, A>G, A>T, C>G, C>T and G>T. The default substitution model is JC69 model.	optional
Phasing		
-diploid diploidSpeciesList	Integrates over all possible phasings of heterozygous genotypes when computing likelihoods [2] given diploid species list. For example, a list of (Scer, Spar) indicates species Scer and Spar will be treated as diploid species in likelihood computation. See Section S4 in G-PhoCS manual for full details. By default we assume the sequences come from haploid species, or the sequences are randomly phased. Note that the substitution model is set to JC69 (fixed).	optional
Substitution rate sampling		
-mupiparamList	Specify the substitution rates when sampling locus-specific substitution rates, which are a list of double values with the order of loci in the nexus file. The default value for all loci is 1.0.	
-muweightparamList	Specify the weights for substitution rates when sampling locus-specific substitution rates, which are a list of integer values with the order of loci in the nexus file. The default value for all loci is 1.	

Simple Example

Download: MCMCseq_example0.nex

Please don't copy and paste, since some illegal characters might be copied.

```
#NEXUS
Begin data;
        Dimensions ntax=5 nchar=80;
        Format datatype=dna symbols="ACTG" missing=? gap=-;
        Matrix
[YAL053W, 25, ...]
Scer TCTTTATTGACGTGTATGGACAATT
Spar TCTTTGTTAACGTGCATGGACAATT
Smik TCCTTGCTAACATGCATGGACAATT
Skud TCTTTGCTAACGTGCATGGATAATT
Sbay TCTTTACTAACGTGCATGGATAACT
[YAR007C, 30, ...]
Scer ATGAGCAGTGTTCAACTTTCGAGGGGCGAT
Spar ATGAGCAGCGTTCAACTTTCGAAGGGCGAC
Smik ATGAGCAGCGTGCAACTATCAAAGGGCGAC
Skud ATGAGCAGTGTTCAACTTTCGAAGGGCGAC
Sbay ATGAGCAGCGTTCAACTTTCGAAGGGCGAC
[YBL015W, 25, ...]
Scer TCTAATTTGTTAAAGCAGAGAGTTA
Spar TCTAATTTGTTAAAGCAGAGAGTTA
Smik TCTAATTTGTTAAAACAGAGAGTTC
Skud TCTAATCTGTTGAAGCAGAGAGTTA
Sbay TCTAATCTGTTGAAGCAAAAAGTCA
;End;
BEGIN PHYLONET;
MCMC_SEQ -cl 250000 -bl 50000 -sf 5000;
END;
```

Example with Starting State

Download:

MCMCseq_example1.nex

Please don't copy and paste, since some illegal characters might be copied.

```
#NEXUS

Begin data;
        Dimensions ntax=3 nchar=500;
        Format datatype=dna symbols="ACTG" missing=? gap=-;
        Matrix
[0, 500]
A
TCGCGCTAACGTCGTTTATAAGTGATCAAAGATAAAAGGAAATCTAAGCTGCCTTCATGTTCCTCATCGGACCTGCACAAGGATGGGCGTGGAGATTCTGGCATGGATACTG
TACTTTTACGCGATCGCCCCAGCTACCGACCTCTATAATCACAGGGAATCTCGGGGAACGAATTGCTTCACTAGGTCACACCCGGTTTATAGCCCGTAGAAGTTAGAGCCCG
CGAATAAAGGACTAACAACTCTTATCAAGCTAAGGGACATCCTAGAGGGACCTCTGCGGGAGCAGCATGTTGTGTGACTCATCACGGTAAGAACTTGGCAAGCGCGACAGCG
GCTAAGCCAGCATGCTAGGCGTCGTCGGATAGTCGCCGTCACGGAATCGGATGAGATCCCTTGAGGGATTGATGATGTTCACATCACTACATGGTTGTTCTGAGTGTTGGTG
ATCAGGTGCAGCAATTGTGCTTGACGGAAATGGGCTCTCATAACCGAACCCA
C
GCGCACCTCCCTCGGATATAAGTGACCGAAGAGAAAAGGGAATCTAATTGGCCCTATCATCACTCATCGTACCTGATCACGTATGGCTGTGGAGATTGCGGCATGGATACTG
TACTTTTGAGCGATCATCCCAGTTACCGACCTTCTTAATAAGAGGGAACCTAGGGTAAAGGAATGCTCCACTCCGTCACACGGGGTATATATCCGGAATATGTTAGGCCCCC
CGAATGAAGGAGTAAAAACTCTTAACAAGCTCCGACAGATCCTAGGGTATCGTCTGCGGGGCCGGCAGGTCGTGGGACGCATCACGCTAAACACTTGGCAAGCGTGACAGCG
GCTGGGTCAGAATGCTCGGCCACGCCGTTTAGTCGCCGGCACCGAATCGAATGTGATCCCTTGAGGAAATGATGAAGTTAACATCATTACATGGGTGCTCTGAGTGATGGTG
ATAAGGTGGAGGACTTGTGTTTGACGGAAATGGGCTCTGAAAACCGAACTCT
B
GCGCACCTACTGCGGATATAAGTGACCGAAGAGTAATGGGAATCTATGCGGCCCTCGCGTCTCTCATCGTACCTGATCAAGTATGGGCGTGGAGATTGTGGCATGGATACTG
TACTTTTGAGCCATCATCCCAGTTACCGACCTTCGTAATAAGAGCGAGCCTAGGGGAAAGAAATGCTCCACTCCATCACACCGGGTATATATCCGGAATATGTTCGAGCCCC
CGAATAAAGGAGTAAAAACTCTTAACAAGCTCCGAAACATCCTAGGGTATCCTCTGCAGGGACGGCATGTTGTGGGGCCCATCACCCTAAGACCTTTGCAAGCATGAAAGCG
GCTCAGCCAGCATGCTCGATCCCGCCGTACAGTCGCCGGCACGGAATCGAGTGTGATCCCCTGAGGAATTGATGAAGTTAACATCACTACTTGGCTGCTCTGAGTGCTGGTG
ATCAGGTGCAGCACATATGTGTGACGGAAATGGGCACTGACAACCGAACTAT
[1, 500]
A
GAAACGGATCTAAGTGTACGGTTTCTCTCGAAGGGGGCACCTTTGCTATGCCCACCCCCATCTTGGAAGTGCGAGACCATACTCGCGCGTGCGTCAGGTTCTTACTTGATTT
CGGCGGGGGTGGCTAAATTTTAGCTAGGGATCTAGAAATCCGTCATAGTCCTACAGGGCCATTCTGCCGCTTGCTAGCGTTGGTGATACGAGGGCAACTTTGAACTTTACGC
GGAACTCCCCACCTCAGAGACTGTTACGACGTAGGCTAAATGTGCCGTGATTTCTGAGGGCAAAAGCCGTGCAAGGATGGACGGGGGTGCTCAAACAACTGCATCAGCCTCG
GCATTATCTTGCATGAGCGCCTTCGATCGGTCACCAGTCGGCTAGATTACAAGCAAGCTCTTCGGAGGAGATGAGCTCGCATGGATCACGCGTCTACGTAACTTTCAGGGTC
CATCCAAATGTCAATCATTCACCGAATGGCGATCGTCAGGTACGCGATTCCA
C
CGCTCGGATCTAAGTGTACGGTTTCTCTCGAAGGTGGAACCATTGCTATACCCACCCCCATCTTGGAAGTGCCAAACCATTCTCCCAAGAGCGTCGGGTTCTTACTCGATTT
CGGCGGGGGTGGCTACAATTTAGGTAGGGATCTAGAAATCGGTTATAATCCTACAAAGCCATTCTGGCGCTTGCTAGTGTTGGTGATACGAGGGCAGCTTTGAACTTTACCG
GGAACTGGGCACCTAAGGGACTGTGTCGACGTAGGCTAAATGTGCCGTGATTTCAGCGAGCAAAAGCCATGCAAGATTGGACGGGGGGCCTCAAACAACTGCATCAGCCTCG
ATATTATCTTGCATGAGCTCCTTCGATCGGTTCCCAGTCGGCTATATTATAAGCAAGCTCTTCGGAGGATATGAGCACGCACGGATTCCGCGTCTACGTAACTTTGAGGGCC
CAGCCAGCAGTCAATCATTCAACGAATGGCGATCATAACGAACGCGATTCCA
B
CGCTCGGATCTAAGTGTACGGTTTCTCTGGAAGGTGGAACCATTGCTATACCCATCCCCATCTTGGAAGTGCCAGACCATTCTCCCAAGAGCGTCTGGTTCTTACTCGATTT
CGGCGGGGGTGGCTACAATTTAGGTAGGGATCTAGAAATCGGTGATAATCGTACAAAGCCATTCTGGCGCTTGCTAGTGTCGGTGATACGAGAGCAGCTTTGAACTTTACCC
GGAACTGCGCACCTAAGGGACTGTGTCGACGTAGGCTAAATGTGCCGTGATTTCAGCGAGCAAAAGCCATGCAAGATTGGACGGGCGGCCTCAAACAACTGCATCAGCCTCG
ATATTATCTTGCATGAGCTCCTTCGATCGGTTCCCAGTCGGCTATCTTATAAGCAAGCTCTTCGGAGGATATGAGCACGCACGGATTCCGCGTCTACGTAACTTTGAGGGCC
CAGCCAGCAGTCAATCATTGACCGAATGGCGATCATAACGAACGCGATTCCA
;End;

BEGIN TREES;
Tree gt0 = (A:0.119900443,(C:0.058838639,B:0.058838639):0.061061803);
Tree gt1 = (A:0.068766378,(C:0.016229589,B:0.016229589):0.052536789);
END;

BEGIN NETWORKS;
Network net1 = (((B:0.0)I3#H1:0.05::0.8,(C:2.0E-8,I3#H1:2.0E-8::0.2)I2:0.04999998)I1:0.01,A:0.06)I0;
END;

BEGIN PHYLONET;
MCMC_SEQ -cl 50000 -bl 10000 -sgt (gt0,gt1) -snet (net1) -sps 0.04 -pre 20;
END;
```

Example given Missing Data

Download: MCMCseq_example2.nex
Please don't copy and paste, since some illegal characters might be copied.

```
#NEXUS

Begin data;
        Dimensions ntax=5 nchar=108;
        Format datatype=dna symbols="ACTG" missing=? gap=-;
        Matrix
[loci1, 53, ...]
a1        ATTGGAGACRAGCGARGACCGAGCTCACGAACCTGAGGAATGGAATCGATTAC
a2        ATTTGAGACRAGCGARGACCGAGCTCACGAACCTGAGGANTGGAATCGATTAC
b1        TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTAAGGATTGGAATCGATTGT
b2        TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTGAGGATTGGAATCGATTGT
[loci2, 58, ...]
a2        ACTTTGCAAGCCAAAAATGGTATGCGAGACAACGCCTGTCATGGATGATGAACCAGAT
b1        GCTTTGCAAGCCTAAGATGGTTTGCGAGACGACGATGGCAGTCGACGATGAATCAGAC
b2        GCTTTGCAAGCCTAAGATGGTTTGCGAGACGACGATGGCAGTCGACGATGAATCAGAC
c1        GCTTTGRAAGRCAAAAATGATATGCGAAACAACGCCCGTGATGGACGATGAACAGGAT
;End;
BEGIN PHYLONET;
MCMC_SEQ -loci (loci1,loci2) -cl 5000000 -bl 1000000 -tm <A:a1,a2; B:b1,b2; C:c1>;
END;
```

Understanding the Output

System Output

- Logger: each time a sample is collected, the program prints out
    1. first line: the Posterior value, current ESS (Effective Sample Size) based on the posterior values, likelihood value, prior value, current ESS based on the prior values.
    2. second line: the sampled phylogenetic network, with divergence times, population sizes and inheritance probabilities. Note that the value in the bracket is the population size of the root branch. If a constant population size across all branches is assumed, then the value represents the general population size.
- Summarization: the program prints out the chain length, burn-in length, sample frequency and the overall acceptance rate of proposals.
- Operations: the usage and the acceptance rate for each operation.
- 95% credible set of network topologies:
    - size: the number of times the topology being sampled
    - percent: the proportion of the topology being sampled
    - MAP (Maximum A Posterior): the maximum posterior value and the corresponding topology the MAP topology are given.
    - AVE: the average posterior value and the averaged (branch lengths and inheritance probabilities) network are printed out.
    - rank: the topologies are ranked on their proportion.
- Run time: the elapsed time.

```
MCMC_SEQ -cl 250000 -bl 50000 -sf 5000
---------- Logger: ----------
Iteration; Posterior; ESS; Likelihood; Prior; ESS; #Reticulation
0; -257.55069; 0.00000; -263.27861; 5.72791; 0.00000; 0;
[0.036](((Scer:0.00857375,Spar:0.00857375):4.5125000000000026E-4,Skud:0.009025):4.7499999999999973E-4,Sbay:0.0095):
0.11472678834065418,Smik:0.12422678834065418);
.....
50; -176.50732; 10.65831; -181.15553; 4.64822; 11.84238; 0;
[0.017160158027924775](((Sbay:0.01968119866828454,Skud:0.01968119866828454):0.042016035724419504,(Spar:0.04364900393745317,Scer:
0.04364900393745317):0.018048230455250877):0.01662669337541752,Smik:0.07832392776812157);
---------- Summarization: ----------
Burn-in = 50000, Chain length = 250000, Sample size = 40, Acceptance rate = 0.10274
---------- Operations ----------
Operation:NarrowNNI; Used:34781; Accepted:3750 ACrate:0.10781748655875334
Operation:Swap-Nodes; Used:5273; Accepted:155 ACrate:0.02939503129148492
Operation:SubtreeSlide; Used:34904; Accepted:3566 ACrate:0.10216594086637634
.....

Overall MAP = -139.6655535361708

(((Spar:0.054401097303896875,Scer:0.054401097303896875):0.02940261095452569,Smik:0.08380370825842257):0.015640290731517764,(Sbay:
0.038489186677349164,Skud:0.038489186677349164):0.060954812312591165);
---------- 95% credible set: ----------
Rank = 0; Size = 20; Percent = 48.78; MAP = -139.6655535361708:(((Spar:0.054401097303896875,Scer:0.054401097303896875):
0.02940261095452569,Smik:0.08380370825842257):0.015640290731517764,(Sbay:0.038489186677349164,Skud:0.038489186677349164):
0.060954812312591165); Ave=-159.81967227297005; ((Smik:0.07802648460532205,(Scer:0.04734369459293139,Spar:0.04734369459293139):
0.03068279001239066):0.012243968912293374,(Skud:0.0399365140411103,Sbay:0.0399365140411103):0.05033393947650512);
Rank = 1; Size = 16; Percent = 39.02; MAP = -150.1407504811838:(Smik:0.08832671142241318,((Sbay:0.0574884656789708,Skud:
0.0574884656789708):0.029947862652692656,(Spar:0.05271204611595535,Scer:0.05271204611595535):0.034724282215708106):
8.903830907497218E-4); Ave=-171.0422748749801; (Smik:0.09785346027572299,((Sbay:0.040857883347008524,Skud:0.040857883347008524):
0.03552943228704832,(Scer:0.055009891356695276,Spar:0.055009891356695276):0.02137742427736157):0.021466144641666143);
.....

Total elapsed time : 27.35100 s
```

Sample Files

The phylogenetic network, gene trees and the hyper-parameter of the population size are logged into files under your home directory or the directory specified by "-dir outDirectory".

- Phylogenetic Network: ~/outDirectory/network.log
- Hyper-parameter of Population size: ~/outDirectory/popSizePrior.log
- Gene tree: ~/outDirectory/tree_locusName.log

Downloads

- example.zip
  - example.nexus: input file for PhyloNet
  - example.txt: system output
  - network.log, popSizePrior.log, tree_YAL053W.log, tree_YAR007C.log, tree_YBL015W.log: sample files
- The yeast data set (Rokas et al., 2003) sampled from seven Saccharomyces species *S. cerevisiae* (*Scer*), *S. paradoxus* (*Spar*), *S. mikatae* (*Smik*) , *S. kudriavzevii* (*Skud*), *S. bayanus* (*Sbay*), *S. castellii* (*Scas*) and *S. kluyveri* (*Sklu*)
  - 106-locus
  - 28-locus (with strong phylogenetic signals)
  - 106-locus restricted by five Saccharomyces species *Scer*, *Spar*, *Smik*, *Skud* and *Sbay*.
- The wheat data set (Marcussen et al., 2014) sampled from hexaploid bread wheat subgenomes *T. aestivum* TaA (A subgenome), TaB (B subgenome) and TaD (D subgenome), and five diploid relatives *T. monococcum* (Tm), *T. urartu* (Tu), *Ae. sharonensis* (Ash), *Ae. speltoides* (Asp) and *Ae. tauschii* (At)
  - 137-locus
  - 68-locus
- The mosquito data set (Fontaine et al., 2014) sampled from six Anopheles species *An. gambiae* (G), *An. coluzzii* (C), *An. arabiensis* (A), *An. quadriannulatus* (Q), *An. merus* (R) and *An. melas* (L)
  - 228-locus from X chromosome
  - 59-locus (with strong phylogenetic signals) from X chromosome
  - 382-locus (with strong phylogenetic signals) from autosomes

Visualization

If you want to analyze parameters in the sampled networks using Tracer, please use command SummarizeMCMCResults to generate Tracer readable log file.

Command References

1. D.Wen and L. Nakhleh. Co-estimating reticulate phylogenies and gene trees on sequences from multiple independent loci. *Submitted*.
2. Gronau, Ilan, et al. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics* 43.10 (2011): 1031-1034.

See Also

- List of PhyloNet Commands