# MCMC_GT

## Description

Bayesian estimation of the posterior distribution of phylogenetic networks given a list of gene tree topologies.

## Usage

### Bayesian Inference

```
MCMC_GT geneTreeList [-cl chainLength] [-bl burnInLength] [-sf sampleFrequency] [-sd seed] [-pp
poissonParameter] [-mr maximumReticulation] [-pl parallelThreads] [-tp temperatureList] [-sn
startingNetworkList] [-tm taxonMap] [-pseudo]
```

| | | |
|---|---|---|
| *geneTreeList* | Comma delimited list of gene tree identifiers or comma delimited list of sets of gene tree identifiers. See details. | mandatory |
| -cl *chainLength* | The length of the MCMC chain. The default value is 1,100,000. | optional |
| -bl *burnInLength* | The number of iterations in burn-in period. The default value is 100,000. | optional |
| -sf *sampleFrequency* | The sample frequency. The default value is 1,000. | optional |
| -sd *seed* | The random seed. The default seed is 12345678 | optional |
| -pp *poissonParameter* | The Poisson parameter in the prior on the number of reticulation nodes. The default value is 1.0 | optional |
| -mr *maximumReticulation* | The maximum number of reticulation nodes in the sampled phylogenetic networks. The default value is infinity. | optional |
| -pl *parallelThreads* | The number of threads running in parallel. The default value is 1. | optional |
| -tp *temperatureList* | The list of temperatures for the Metropolis-coupled MCMC chains. For example, a list (1.0, 2.0, 3.0) indicates a cold chain with temperature 1.0 and two hot chains with temperatures 2.0 and 3.0 respectively will be run. The first value in the list should always be 1.0. The default list is (1.0). | optional |
| -sn *startingNetworkList* | Comma delimited list of network identifiers. See details. If the list contains only one network, the network will be used as the starting network for all the chains. If the length of the list equals to the the length of the temperature list, each chain will have a corresponding starting network. The default list is empty and the MDC tree will be used as the starting network. | 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. However, this option allows multiple alleles to be sampled. | optional |
| -pseudo | Use pseudo likelihood instead of full likelihood to reduce runtime, see details. | optional |

### Summarization

```
MCMC_GT -sum fileList
```

| | | |
|---|---|---|
| *fileList* | The list of output files from **Bayesian Inference**. | mandatory |

# Examples

### Bayesian Inference

```
#NEXUS

BEGIN TREES;

Tree gt1 = (((DF,MZ),(DG,MK)),MC);
Tree gt2 = (((DF,(MK,MZ)),DG),MC);
Tree gt3 = (((DF,MK),(MC,MZ)),DG);
Tree gt4 = ((DF,(MC,(MK,MZ))),DG);
Tree gt5 = (DF,((DG,(MC,MZ)),MK));
Tree gt6 = (((DF,DG),MK),(MC,MZ));
Tree gt7 = ((((DF,DG),MC),MZ),MK);
Tree gt8 = (((DF,MK),MZ),(DG,MC));
Tree gt9 = ((DF,(DG,MC)),(MK,MZ));
Tree gt10 = (((DF,MC),(DG,MK)),MZ);

END;

BEGIN PHYLONET;
MCMC_GT (gt1-gt10) -cl 10000 -bl 5000 -sf 1000;
END;
```

```
#NEXUS


BEGIN NETWORKS;
Network net1 = (((F:1,K:1):1,G:1):1,C:1):1;
Network net2 = ((F:1,(C:1,K:1):1):1,G:1):1;
END;

BEGIN TREES;

Tree gt1 = (((DF,MZ),(DG,MK)),MC);
Tree gt2 = (((DF,(MK,MZ)),DG),MC);
Tree gt3 = (((DF,MK),(MC,MZ)),DG);
Tree gt4 = ((DF,(MC,(MK,MZ))),DG);
Tree gt5 = (DF,((DG,(MC,MZ)),MK));
Tree gt6 = (((DF,DG),MK),(MC,MZ));
Tree gt7 = ((((DF,DG),MC),MZ),MK);
Tree gt8 = (((DF,MK),MZ),(DG,MC));
Tree gt9 = ((DF,(DG,MC)),(MK,MZ));
Tree gt10 = (((DF,MC),(DG,MK)),MZ);

END;

BEGIN PHYLONET;
MCMC_GT ({gt1-gt2}, {gt3-gt4}, {gt5-gt6}, {gt7-gt8}, {gt9-gt10}) -cl 10000 -bl 5000 -sf 1000 -sd 12345 -pp 0.1 -mr 3 -pl 4 -tp (1.0,2.0) -sn (net1,net2) -tm
<F:DF;G:DG;K:MK;C:MC,MZ>;
END;
```

### Summarization

```
#NEXUS
BEGIN PHYLONET;
MCMC_GT -sum (mcmc1.txt, mcmc2.txt, mcmc3.txt, mcmc4.txt);
END;
```

# Understanding the Output

### Bayesian Inference

- Logger: each time a sample is collected, the program prints out the Posterior value, current ESS (Effective Sample Size) based on the posterior values, likelihood value, prior value, current ESS based on the prior values, and the sampled phylogenetic network sampled.
- 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.
- The 95% credible set: For each unique topology in the 95% credible set, the network with the maximum posterior value and the averaged (branch lengths and inheritance probabilities) network are printed out. The topologies are ranked on their posterior probabilities.
- Run time: the elapsed time.

Output of the second example

```
---------- Logger ----------
Iteration;  Posterior;  ESS;  Likelihood;  Prior;  ESS;  #Reticulation

0;  -32.31804;  0.00000;  -26.31804;  -6.00000;  0.00000;  0;

(C:1.0,(F:1.0,(G:1.0,K:1.0):1.0):1.0);

1;  -30.67945;  0.00000;  -24.36941;  -6.31004;  0.00000;  0;

(((F:1.100296825380626,G:3.091462635493304):0.0672071335563177,C:1.2631834608546109):0.022439620222693815,K:0.7654486703017418);

.....

9;  -39.23243;  5.00000;  -24.51469;  -14.71774;  5.00000;  0;

(C:1.303229359865713,(G:1.184262389767834,(F:9.355326885355721,K:1.9913949707859038)I0:0.8785013943118355)I3:0.005027219859507291)I2;

10;  -47.01271;  6.00000;  -23.75691;  -23.25580;  6.00000;  1;

((((G:0.6775006643112376)I1#H1:0.10254403746151251::0.6222138281679854,(K:0.40305290671104776,F:11.774054935191478)I3:1.0127444936379693)I4:
0.11119651081773953,C:0.6245266587065907)I2:0.24534630112752273,I1#H1:0.6332472675472998::0.3777861718320146)I0;

---------- Summarization: ----------

Burn-in = 5000, Chain length = 10000, Sample size = 5 Acceptance rate = 0.62080

        ---------- Operations ----------

Operation:Move-Head; Used:196; Accepted:35 ACrate:0.17857142857142858

Operation:Change-Length; Used:5782; Accepted:5628 ACrate:0.9733656174334141

Operation:Add-Reticulation; Used:92; Accepted:1 ACrate:0.010869565217391304

Operation:Flip-Reticulation; Used:191; Accepted:44 ACrate:0.23036649214659685

Operation:Change-Probability; Used:172; Accepted:165 ACrate:0.9593023255813954

Operation:Move-Tail; Used:3564; Accepted:332 ACrate:0.0931537598204265

Operation:Delete-Reticulation; Used:3; Accepted:3 ACrate:1.0



Overall MAP = -39.13178296704008

(((G:1.8467353971326255,F:3.669429912707025)I4:0.72972613976057,C:0.7672463997220461)I0:0.14233512944191162,K:7.34631035390478)I3;

        ---------- 95% credible set of topologies: ----------

Rank = 0; Size = 2; Percent = 40.00; MAP = -39.23243337560144:(C:1.303229359865713,(G:1.184262389767834,(F:9.355326885355721,K:1.9913949707859038)I0:
0.8785013943118355)I3:0.005027219859507291)I2; Ave=-39.25394715434493; (C:1.20916762172346,(G:2.057607023787008,(K:4.61838854181666,F:6.605842763421295)I0:
0.5911795016873735)I4:0.1657841343330578)I3;

Rank = 1; Size = 1; Percent = 20.00; MAP = -47.01271365298103:((((G:0.6775006643112376)I1#H1:0.10254403746151251::0.6222138281679854,(K:0.40305290671104776,F:
11.774054935191478)I3:1.0127444936379693)I4:0.11119651081773953,C:0.6245266587065907)I2:0.24534630112752273,I1#H1:0.6332472675472998::0.3777861718320146)I0;
Ave=-47.01271365298103; ((G:0.6775006643112376)I1#H1:0.6332472675472998::0.3777861718320146,(C:0.6245266587065907,((F:11.774054935191478,K:
0.40305290671104776)I3:1.0127444936379693,I1#H1:0.10254403746151251::0.6222138281679854)I4:0.11119651081773953)I2:0.24534630112752273)I0;

Rank = 2; Size = 1; Percent = 20.00; MAP = -39.13178296704008:(((G:1.8467353971326255,F:3.669429912707025)I4:0.72972613976057,C:0.7672463997220461)I0:
0.14233512944191162,K:7.34631035390478)I3; Ave=-39.13178296704008; (K:7.34631035390478,(C:0.7672463997220461,(F:3.669429912707025,G:1.8467353971326255)I4:
0.72972613976057)I0:0.14233512944191162)I3;

Rank = 3; Size = 1; Percent = 20.00; MAP = -40.2382981193348:(((F:7.823288510084807,K:6.50888387662072)I0:0.32861454580215427,C:0.8862007905628368)I4:
0.1556168196177095,G:1.218111399301718)I3; Ave=-40.2382981193348; (G:1.218111399301718,(C:0.8862007905628368,(K:6.50888387662072,F:7.823288510084807)I0:
0.32861454580215427)I4:0.1556168196177095)I3;

Total elapsed time : 31.53400 s
```

## Summarization

- Topologies: the samples (samples from burn-in period are excluded) from all the files are combined. The topologies are ranked on their posterior probabilities. Only the topologies in the 95% credible set are printed out. The first topology is MPP (maximum posterior probability) topology.
- PSRF (Potential Scale Reduction Factor): The PSRF values of posterior, likelihood and prior are calculated respectively. A PSRF value indicates good mixing when it approaches 1.0.
- Sojourns: a method to evaluate convergence/mixing. 'Sojourn' is a consecutive series of samples in which only the topology of interest was found. For each file, we summarize the frequency, the posterior probability, the number of sojourns, and the max and average length of sojourns for each topology in the 95% credible set. Ideally the number of sojourns is large while the length of sojourns is small -- the ability to leave and then return quickly and repeatedly to the same topology suggests good mixing with respect to topologies.

- SRQ (Scaled Regeneration Quantile) Plot: a method to evaluate convergence/mixing via topologies. Let Ti be the number of sampled MPP topology in the first i iteration of the stationary phase. For each file, we extract <i/n, Ti/Tn> for every i in 1...n into <x,y>. The slope of <x,y> in an SRQ plot should ideally be close to the posterior probability of MPP. Departures from this indicate that at some points the chain was on a trajectory that should have led to a different final posterior probability.
- Trace Plot: a method to evaluate convergence/mixing via posterior values. For each file, we extract a list of posterior values from the sampling phase. The format is compatible with Matlab and Python.

```
---------- Topologies ----------
PosteriorProbability Topology

0.4500 ((TaB:0.2658858620952823)I2#H1:0.7341141379047177::0.8708100650901477,(TaA:1.0,(TaD:0.32466899225557677,I2#H1:
1.3465736635156926::0.1291899349098523)I3:2.4002600314073748)I1:0.19747493612759992)I0;
0.2500 ((TaD:1.0,TaA:1.0):1.0,TaB:1.0);
0.1500 (((TaD:1.57737312713865)I3#H1:0.7217516903113591::0.42596342801648157,TaB:0.3702926554366514)I2:0.5822736350667996,(I3#H1:
0.8251033741688848::0.5740365719835184,TaA:0.5942441726893598)I1:0.6915888106669655)I0;
0.1500 ((((TaA:0.2655477426293258)I1#H1:0.11325339702702007::0.4241273916272569,TaD:1.935302259486147)I3:0.786549353980848,TaB:
0.40541038525824113)I2:0.2974313193957629,I1#H1:1.0824385015100584::0.5758726083727431)I0;
---------- PSRF ----------
Posterior - 1.0498673507956626
Likelihood - 1.09497860895306455
Prior - 1.0576479101618842

---------- Sojourn ----------
Topology Frequency Posterior #sojourns max(sjs) ave(sjs)
1 & 4 & 0.4000 & 2 & 3 & 2.00 \\
2 & 4 & 0.4000 & 1 & 4 & 4.00 \\
3 & 1 & 0.1000 & 1 & 1 & 1.00 \\
4 & 1 & 0.1000 & 1 & 1 & 1.00 \\
Topology Frequency Posterior #sojourns max(sjs) ave(sjs)
1 & 8 & 0.8000 & 1 & 8 & 8.00 \\
2 & 2 & 0.2000 & 1 & 2 & 2.00 \\
3 & 0 & 0.0000 & 0 & 0 & NaN \\
4 & 0 & 0.0000 & 0 & 0 & NaN \\
Topology Frequency Posterior #sojourns max(sjs) ave(sjs)
1 & 4 & 0.4000 & 2 & 2 & 2.00 \\
2 & 2 & 0.2000 & 1 & 2 & 2.00 \\
3 & 1 & 0.1000 & 1 & 1 & 1.00 \\
4 & 3 & 0.3000 & 2 & 2 & 1.50 \\
Topology Frequency Posterior #sojourns max(sjs) ave(sjs)
1 & 2 & 0.2000 & 2 & 1 & 1.00 \\
2 & 2 & 0.2000 & 1 & 2 & 2.00 \\
3 & 4 & 0.4000 & 3 & 2 & 1.33 \\
4 & 2 & 0.2000 & 2 & 1 & 1.00 \\

---------- SRQ Plot ----------
y0 = [0.25,0.5,0.75,1.0];
x0 = [0.5,0.6,0.7,1.0];
y1 = [0.125,0.25,0.375,0.5,0.625,0.75,0.875,1.0];
x1 = [0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0];
y2 = [0.25,0.5,0.75,1.0];
x2 = [0.3,0.4,0.8,0.9];
y3 = [0.5,1.0];
x3 = [0.5,0.7];
---------- Trace Plot ----------

y0 = [-2736.2903,-2487.17354,-2464.78381,-2465.97101,-2452.65531,-2443.28357,-2439.46916,-2441.01877,-2439.93589,-2442.0554,];

y1 = [-2736.2903,-2469.68094,-2438.08212,-2438.52812,-2439.02494,-2438.4918,-2439.33072,-2438.70721,-2438.68125,-2438.68995,];
y2 = [-2736.2903,-2474.94649,-2446.39497,-2438.87233,-2440.56426,-2442.1411,-2438.8736,-2439.18374,-2438.65952,-2439.23713,];
y3 = [-2736.2903,-2466.11338,-2451.75728,-2439.36378,-2439.58731,-2437.75625,-2439.12275,-2438.99142,-2439.20819,-2440.1315,];
```

# Command References


- D.Wen, Y. Yu, and L. Nakhleh. Bayesian Inference of Reticulate Phylogenies Under the Multispecies Network Coalescent. PLoS Genet 12(5): e1006006, 2016.


# See Also

- List of PhyloNet Commands