PhyloNet

Yun Yu

Department of Computer Science
Bioinformatics Group
Rice University
yy9@rice.edu

Installation

* System requirement:

  Java 1.8 or later

* Download executable of PhyloNet (latest version 3.6.0):

  http://bioinfo.cs.rice.edu/PhyloNet

Basic Usage

* Command:

  > java -jar PhyloNet_3.6.0.jar script.nex

* Input NEXUS file:

```
#NEXUS

BEGIN TREES;

Tree g1 = ((((a,b),c),d),e);
Tree g2 = ((a,b),((c,e),d));
Tree g3 = ((a,c),((b,e),d));

END;


BEGIN PHYLONET;

Infer_ST_MDC (g1, g2, g3);

END;
```

Estimating species phylogenies from gene trees

* Maximum Likelihood: *InferNetwork_ML*

* Maximum Pseudo-Likelihood: *InferNetwork_MPL*

* Bayesian: *MCMC_GT*

* Maximum Parsimony:

  * Trees: *Infer_ST_MDC, Infer_ST_MDC_UR*

  * Networks: *InferNetwork_MP*

ILS

Maximum Parsimony

* ## Command Usage

`infer_ST_MDC/infer_ST_MDC_UR (geneTree1 [, geneTree2...]) [-e proportion] [-x] [-b threshold] [-a taxaMap] [-ur] [-t time] [resultOutputFile]`

geneTree1 [, geneTree2...]	Comma delimited collection of gene trees.	mandatory
-a *taxaMap*	Gene trees / species tree taxa association.	optional
-b *threshold*	Gene trees bootstrap threshold.	optional
-e *proportion*	Get optimal and sub-optimal species trees whose scores are less than *proportion*% worse than the optimal score.	optional
-x	Use all clusters instead of clusters in gene trees.	optional
-ur	Allow returning unresolved species tree.	optional
-t *timeLimit*	Limit search time to *timeLimit* minutes.	optional
resultOutputFile	Optional file destination for command output.	optional

Maximum Parsimony

✤ Input NEXUS File

```
#NEXUS

BEGIN TREES;
Tree gt0 = (((((((Scer,Spar),Smik),Skud),Sbay),Scas),Sklu),Calb);
…
Tree gt105 = ((((((Scer,Spar),Smik),Skud),Sbay),(Scas,Sklu)),Calb);
END;

BEGIN PHYLONET;
Infer_ST_MDC (gt0-gt105);
END;
```
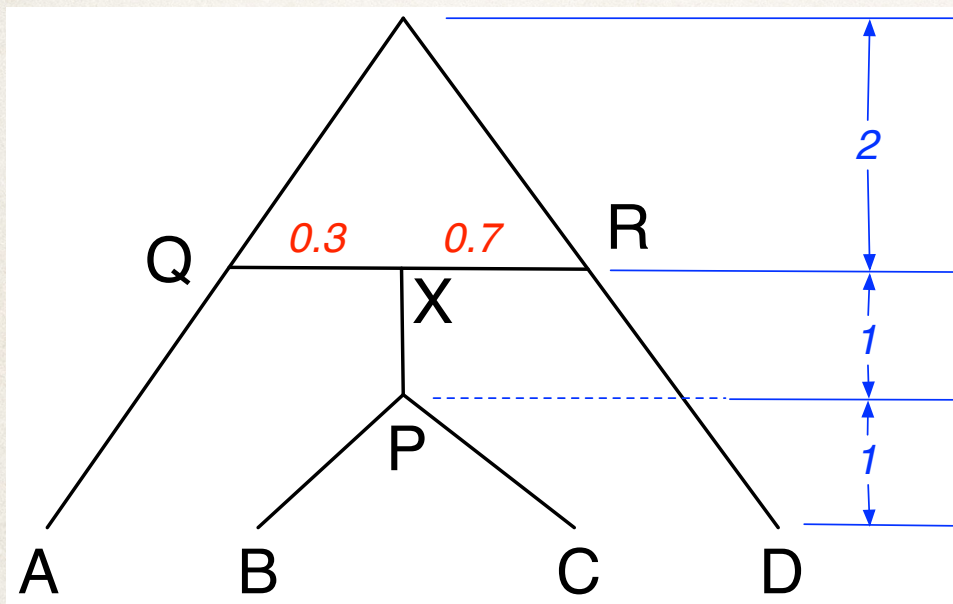
✤ Output

```
(Calb:0,(Sklu:0,(Scas:0,(Sbay:0,(Skud:0,(Smik:0,(Scer:0,Spar:0):2):21):54):0):35):0):0;
Total number of extra lineages:112
```

ILS + Introgression

Phylogenetic Network

* ## Rich Newick Format

  * Newick for phylogenetic tree & Extended Newick for phylogenetic network (Cardona et al., 2008)

e/display/PHYLONET/Rich

((A:2,((B:1,C:1)P:1)X#H1:0::0.3)Q:2,
(D:2,X#H1:0::0.7)R:2);

:branch length:support:inheritance probability

ILS + Introgression

Parsimony

Maximum Parsimony

❖ Command Usage

```
InferNetwork_MP (geneTree1 [, geneTree2...]) numReticulations [-a taxa
map] [-b threshold] [-s startingNetwork] [-n numNetReturned]  [-m
maxNetExamined] [-d maxDiameter] [-h {s1 [,s2...]}] [-w (w1,w2,w3,w4)] [-f
maxFailure] [-x numRuns] [-pl numProcessors] [-di] [result output file]
```

geneTree1 [, geneTree2...]	Comma delimited collection of gene trees.	mandatory
numReticulations	Maximum number of reticulation nodes to added.	mandatory
-s *startingNetwork*	Specify the network to start search. Default value is the optimal MDC tree.	optional
-n *numNetReturned*	Number of optimal networks to return. Default value is 1.	optional
-h *{s1 [, s2...]}*	A set of specified hybrid species. The size of this set equals the number of reticulations in the inferred network	optional
-pl *numProcessors*	Number of processors if you want the computation to be done in parallel. Default value is 1.	optional
-di	Output the Rich Newick string of the inferred network that can be read by Dendroscope.	optional

See https://wiki.rice.edu/confluence/display/PHYLONET/InferNetwork_MP for details.

Maximum Parsimony

✤ Input NEXUS File

```
#NEXUS

BEGIN TREES;
Tree gt0 = (((((((Scer,Spar),Smik),Skud),Sbay),Scas),Sklu),Calb);
…
Tree gt105 = ((((((Scer,Spar),Smik),Skud),Sbay),(Scas,Sklu)),Calb);
END;

BEGIN PHYLONET;
InferNetwork_MP (gt0-gt105) 1;
END;
```

✤ Output

```
(((((((Sbay)#H1:::0.3608,Skud),((Spar,Scer),Smik)), #H1:::0.6392),Scas),Sklu),Calb);
Total number of extra lineages: 60
```

ILS + Introgression

Likelihood

Maximum Likelihood

* ## Command Usage

```
InferNetwork_ML (geneTree1 [, geneTree2...])
numReticulations [-a taxa map] [-bl] [-b threshold] [-s
startingNetwork] [-n numNetReturned] [-h {s1 [,s2...]}] [-w
(w1,w2,w3,w4)] [-x numRuns] [-m maxNetExamined] [-md
moveDiameter] [-rd reticulationDiameter] [-f maxFailure] [-
o] [-po] [-p (rel,abs)] [-r maxRounds] [-t maxTryPerBr] [-i
improveThreshold] [-l maxBL] [-pl numProcessors] [-di]
[result output file]
```

-po	Branch lengths are sampled during the search. This option allows users to optimize branch lengths and inheritance probabilities of the inferred networks as a post-process	optional
-bl	The branch lengths of the input gene trees need to be considered in the computation.	

See https://wiki.rice.edu/confluence/display/PHYLONET/InferNetwork_ML+Command for details.

Maximum Likelihood

✤ Input NEXUS File

```
#NEXUS

BEGIN TREES;
Tree gt0 = ((((Scer,Spar),Smik),Skud),Sbay);
…
Tree gt105 = ((Scer,Spar),Smik,Skud,Sbay);
END;

BEGIN PHYLONET;
InferNetwork_ML (gt0-gt105) 1;
END;
```

✤ Output

```
Inferred Network #1:
((Sbay:1.0)#H1:1.0::0.6130,((Smik:1.0,(Scer:1.0,Spar:1.0):3.5436):1.0585,
(#H1:1.0::0.3869,Skud:1.0):2.1717):5.9272);
Total log probability: -151.57753843275103
```

ILS + Introgression

Pseudo-Likelihood

Maximum Pseudo-Likelihood

* ## Command Usage

```
InferNetwork_MPL (geneTree1 [, geneTree2...])
numReticulations [-a taxa map] [-b threshold] [-s
startingNetwork] [-n numNetReturned] [-h {s1 [,s2...]}]
[-w (w1,w2,w3,w4)] [-x numRuns] [-m maxNetExamined] [-md
moveDiameter] [-rd reticulationDiameter] [-f maxFailure]
[-o] [-po] [-p (rel,abs)] [-r maxRounds] [-t maxTryPerBr]
[-i improveThreshold] [-l maxBL] [-pl numProcessors] [-
di] [result output file]
```

-po	Branch lengths are sampled during the search. This option allows users to optimize branch lengths and inheritance probabilities of the inferred networks under FULL likelihood as a post-process	optional

See https://wiki.rice.edu/confluence/display/PHYLONET/InferNetwork_MPL for details.

Maximum Pseudo-Likelihood

* Input NEXUS File

```
#NEXUS

BEGIN TREES;
Tree gt0 = (((((Scer,Spar),Smik),Skud),Sbay);
…
Tree gt105 = ((Scer,Spar),Smik,Skud,Sbay);
END;

BEGIN PHYLONET;
InferNetwork_MPL (gt0-gt105) 1 -x 10 -po;
END;
```

Maximum Pseudo-Likelihood

* Output

```
Inferred Network #1:
((Sbay:1.0)#H1:1.0::0.6130,((Smik:1.0,(Scer:1.0,Spar:1.0):3.5436):1.0585,
(#H1:1.0::0.3869,Skud:1.0):2.1717):5.9272);
Total log probability: -151.57753843275103
Inferred Network #2:
(((((Smik:1.0,(Scer:1.0,Spar:1.0):3.5517):0.8718)#H1:1.5844::0.6690,Skud:1.0):1.6469,Sbay:
1.0):5.914823983639853,#H1:0.0012::0.3309);
Total log probability: -168.28650850921707
Inferred Network #3:
(((Smik:1.0,(Scer:1.0,Spar:1.0):3.5379):1.0444,(Skud:1.0)#H1:1.0::0.7007):0.8788,
(#H1:1.0::0.2993,Sbay:1.0):5.9324);
Total log probability: -177.6791272776054
Inferred Network #4:
((Sbay:1.0,(Skud:1.0)#H1:1.0::0.9288):0.2035,((#H1:1.0::0.0712,Smik:1.0):0.0012,(Spar:
1.0,Scer:1.0):3.5535):1.3471);
Total log probability: -199.98142731241273
Inferred Network #5:
(((Skud:1.0)#H1:1.0::0.9326,Sbay:1.0):0.1974,((#H1:1.0::0.0674,(Spar:1.0,Scer:1.0):3.5517):
0.0012,Smik:1.0):1.3412);
Total log probability: -199.98564708955553
```

ILS + Introgression

Bayesian

Bayesian

* ## Command Usage

```
MCMC_GT (geneTree1 [, geneTree2...]) [-cl chainLength]
[-bl burnInLength] [-sf sampleFrequency] [-sd seed] [-pp
poissonParameter] [-mr maximumReticulation] [-pl
parallelThreads] [-tp temperatureList] [-sn
startingNetworkList] [-tm taxonMap]
```

-mr	The maximum number of reticulation nodes in the sampled phylogenetic networks. The default value is infinity.	optional

See https://wiki.rice.edu/confluence/display/PHYLONET/MCMC_GT+Command for details.

Bayesian

✤ Input NEXUS File

```
#NEXUS

BEGIN TREES;
Tree gt0 = ((((Scer,Spar),Smik),Skud),Sbay);
…
Tree gt105 = ((Scer,Spar),Smik,Skud,Sbay);
END;

BEGIN PHYLONET;
MCMC_GT (gt0-gt105);
END;
```

Bayesian

* Output

```
Iteration;    Posterior;  ESS;     Likelihood;    Prior;  ESS;    #Reticulation
0;     -252.23994;    0.00000;    -244.23994;   -8.00000;    0.00000;    0;
(Sbay:1.0,(Skud:1.0,(Smik:1.0,(Scer:1.0,Spar:1.0):1.0):1.0):1.0);
.........
1100;    -182.48421;    6.31204;    -154.15857;   -28.32565;    8.02957;    1;
((Sbay:0.8364)I0#H1:0.5373::0.5949,((Smik:1.0,(Scer:1.0,Spar:1.0)I5:3.3052)I4:0.8173,(Skud:
2.7098,I0#H1:0.6856::0.4051)I2:2.9664)I3:2.2888)I1;


---------------------- Summarization: -----------------------
Burn-in = 100000, Chain length = 1100000, Sample size = 1000 Acceptance rate = 0.55319


        -------------- Operations ---------------
Operation:Move-Head; Used:143520; Accepted:1 ACrate:6.967670011148272E-6
.........
Operation:Delete-Reticulation; Used:5358; Accepted:2626 ACrate:0.49010824934677116

Overall MAP = -179.89656989770972
((((Sbay:0.8364)I0#H1:0.6856::0.3776,Skud:2.7098)I2:1.7771,(Smik:1.0,(Scer:1.0,Spar:
1.0)I5:2.9942)I4:1.0385)I3:2.4721,I0#H1:0.5373::0.6224)I1;
```

Bayesian

* Output

```
-------------- Top Topologies: --------------
Rank = 0; Size = 818; Percent = 0.818;
MAP = -179.89656989770972: ((((Sbay:0.8364)I0#H1:0.6856::0.3776,Skud:2.7098)I2:1.7771,(Smik:1.0,(Scer:
1.0,Spar:1.0)I5:2.9942)I4:1.0385)I3:2.4721,I0#H1:0.5373::0.6224)I1;
Ave = -182.497829758399; ((Sbay:0.8364)I0#H1:0.5373::0.6231,(((Spar:1.0,Scer:1.0)I5:3.3546,Smik:
1.0)I4:1.0773,(Skud:2.7098,I0#H1:0.6857::0.3769)I2:1.9924)I3:2.6931)I1;

Rank = 1; Size = 180; Percent = 0.18;
MAP = -191.61104349697848:((((((Spar:1.0,Scer:1.0)I5:3.3889,Smik:
1.0)I4:0.8307)I3#H1:0.9523::0.6741,Skud:0.1783)I2:1.2530,Sbay:0.3632)I0:1.6455,I3#H1:0.0210::0.3258)I1;
Ave = -194.12662772729175; (((Smik:1.0,(Scer:1.0,Spar:1.0)I5:3.3053)I4:0.9721)I3#H1:0.0479::0.3634,
(Sbay:0.3632,(Skud:0.1919,I3#H1:1.2856::0.6365)I2:1.4111)I0:1.5633)I1;

Rank = 2; Size = 2; Percent = 0.002;
MAP = -189.56637204871708:(((((Skud:2.7098,(Sbay:
0.8363)I0#H2:0.6856::0.3355)I2:3.8612)I6#H1:0.0667::0.3223,((Spar:1.0,Scer:1.0)I5:2.6678,Smik:
1.0)I4:0.4006)I7:1.0679,I6#H1:0.2275::0.6776)I3:2.1196,I0#H2:0.5372::0.6644)I1;
Ave = -190.47255907248774; ((Sbay:0.8363)I0#H1:0.5372::0.8354,(((I0#H1:0.6856::0.1645,Skud:
2.7098)I2:3.7692)I6#H2:0.2659::0.7540,((Smik:1.0,(Scer:1.0,Spar:
1.0)I5:2.4576)I4:0.4274,I6#H2:0.0704::0.2459)I7:1.2316)I3:2.0443)I1;
```

Scalability

* Maximum Parsimony:

    ~30 taxa, 3 or 4 reticulations

* Maximum Likelihood:

    ~10 taxa, 2 or 3 reticulations

* Maximum Pseudo-Likelihood:

    ~30 taxa

* Bayesian:

    ~10 taxa, 2 or 3 reticulations

Testing Your Hypothesis

* Parsimony:
    * Trees: *DeepCoalCount_tree*
    * Networks: *DeepCoalCount_network*

* Likelihood: *CalGTProb*

Relevant Publications

* Likelihood: Yu et. al, *PloS Genetics*, vol. 8, no. 4, p. e1002660, 2012

* Parsimony: Yu et. al, *Systematic Biology*, vol. 62, no. 5, pp. 738-751, 2013

* Likelihood: Yu et. al, *PNAS*, vol. 111, no. 46, pp. 16448-16453, 2014

* Pseudo-likelihood: Yu et. al, *BMC Genomics*, vol. 16, no. Suppl 10, p. S10, 2015

* Application: Wen et. al, *Molecular Ecology*, vol. 25, pp. 2361-2372, 2016

* Bayesian: Wen et. al, *PloS Genetics*, vol. 12, no. 5, p. e1006006, 2016

Try it yourself!