Phylogenetic Network Inference with PhyloNet

Zhen Cao, Huw A. Ogilvie, and Zhi Yan
Department of Computer Science
Rice University

The 2020 Phylogenomics Symposium and Software School
Gainesville, Florida
3 January 2020

What is PhyloNet?

Syst. Biol. 67(4):735–740, 2018

DOI:10.1093/sysbio/syy015
Advance Access publication March 5, 2018

Inferring Phylogenetic Networks Using PhyloNet

DINGQIAO WEN[1], YUN YU[1], JIAFAN ZHU[1], AND LUAY NAKHLEH[1,2,*]

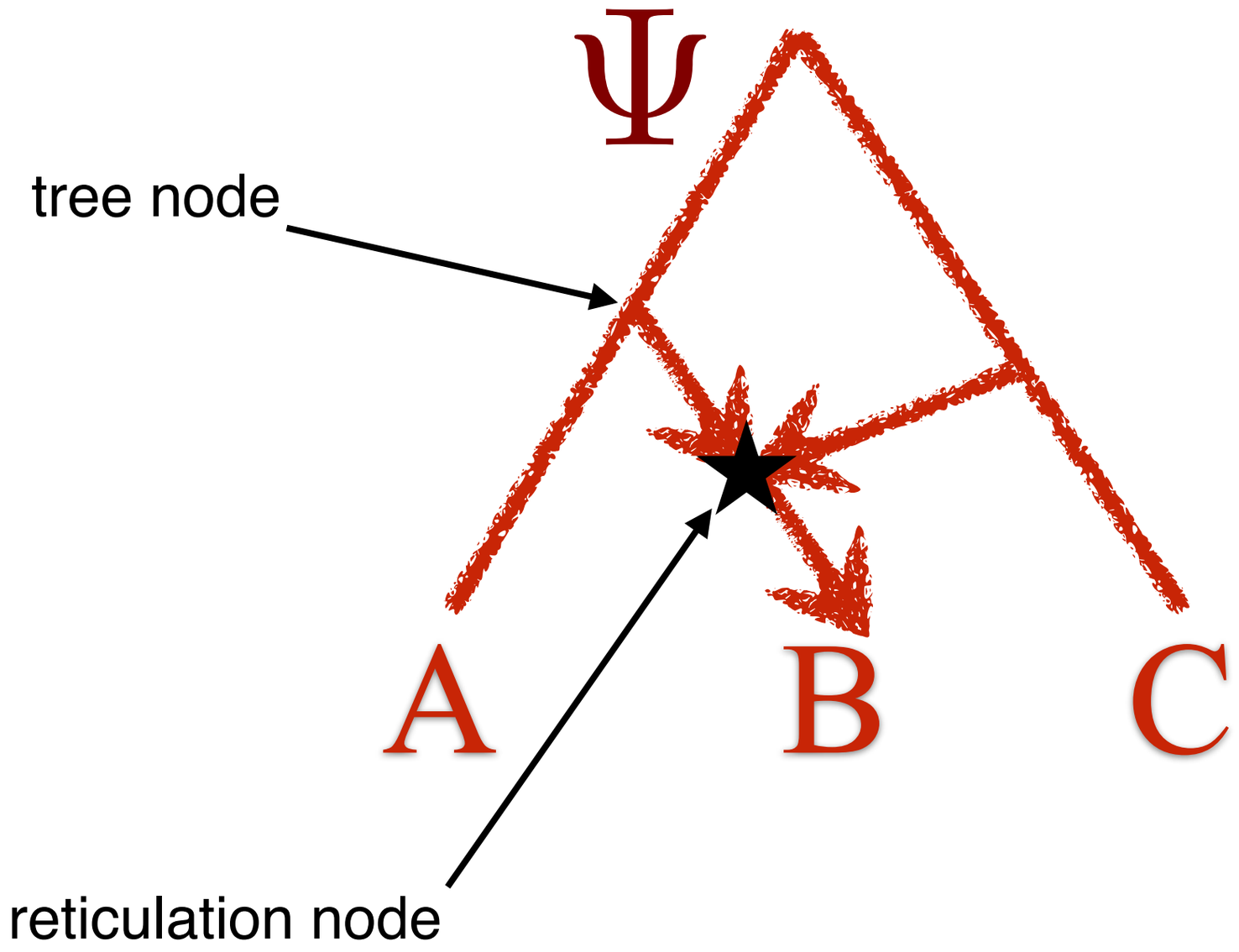
[1]*Computer Science and* [2]*BioSciences, Rice University, 6100 Main Street, Houston, TX 77005, USA;*
[*]*Correspondence to be sent to: Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA;*
E-mail: nakhleh@rice.edu.

Received 21 December 2017; reviews returned 20 February 2018; accepted 23 February 2018
Associate Editor: David Posada

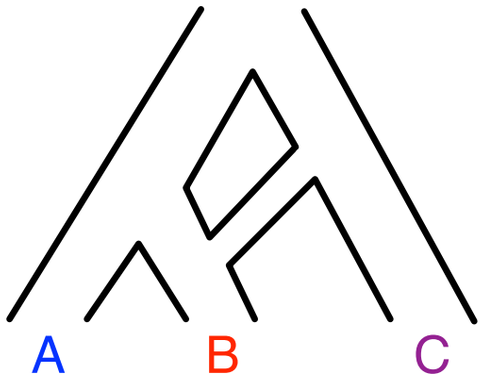
Abstract.—PhyloNet was released in 2008 as a software package for representing and analyzing phylogenetic networks. At the time of its release, the main functionalities in PhyloNet consisted of measures for comparing network topologies and a single heuristic for reconciling gene trees with a species tree. Since then, PhyloNet has grown significantly. The software package now includes a wide array of methods for inferring phylogenetic networks from data sets of unlinked loci while accounting for both reticulation (e.g., hybridization) and incomplete lineage sorting. In particular, PhyloNet now allows for maximum parsimony, maximum likelihood, and Bayesian inference of phylogenetic networks from gene tree estimates. Furthermore, Bayesian inference directly from sequence data (sequence alignments or biallelic markers) is implemented. Maximum parsimony is based on an extension of the "minimizing deep coalescences" criterion to phylogenetic networks, whereas maximum likelihood and Bayesian inference are based on the multispecies network coalescent. All methods allow for multiple individuals per species. As computing the likelihood of a phylogenetic network is computationally hard, PhyloNet allows for evaluation and inference of networks using a pseudolikelihood measure. PhyloNet summarizes the results of the various analyzes and generates phylogenetic networks in the extended Newick format that is readily viewable by existing visualization software. [Bayesian inference; incomplete lineage sorting; maximum likelihood; maximum parsimony; multispecies network coalescent; phylogenetic networks; reticulation.]

What is a Phylogenetic Network?

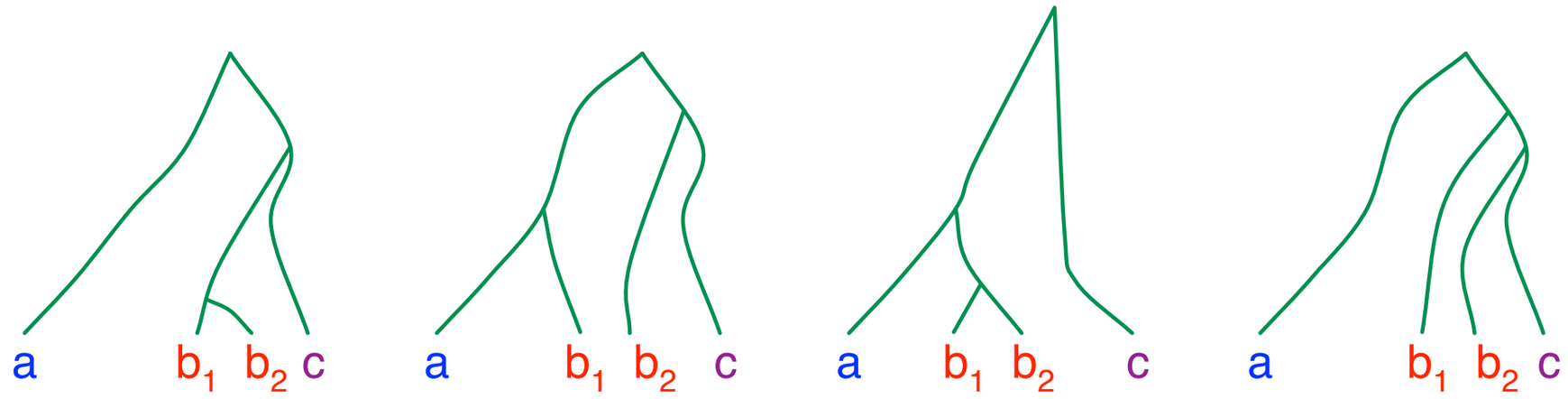
A leaf-labeled, rooted, directed, acyclic graph (rDAG)



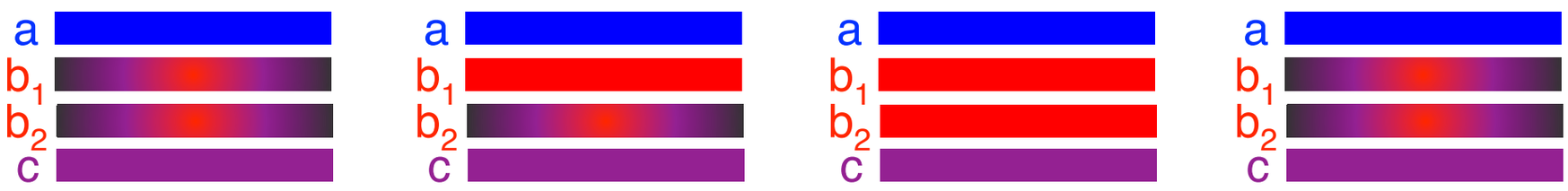
Phylogenomics: ILS + Reticulation



Multispecies Network Coalescent



Model of Sequence Evolution



locus 1 **locus 2** **locus 3** **locus 4**

Inference of Phylogenetic Networks

Input: Sequence alignments for m loci

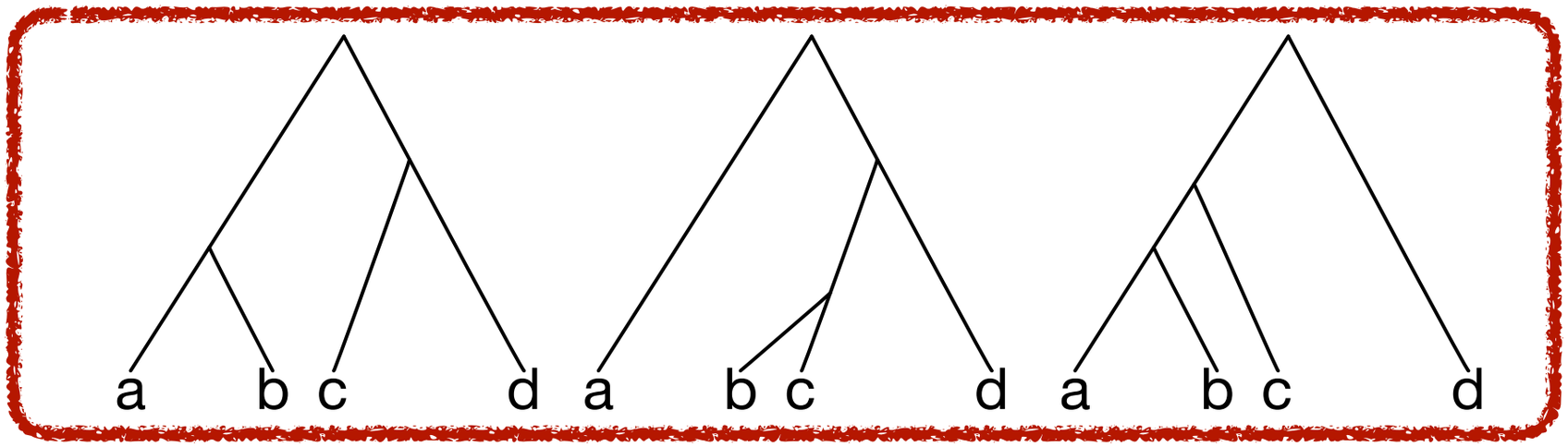
$$\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$$

Output: Phylogenetic network and inheritance probabilities

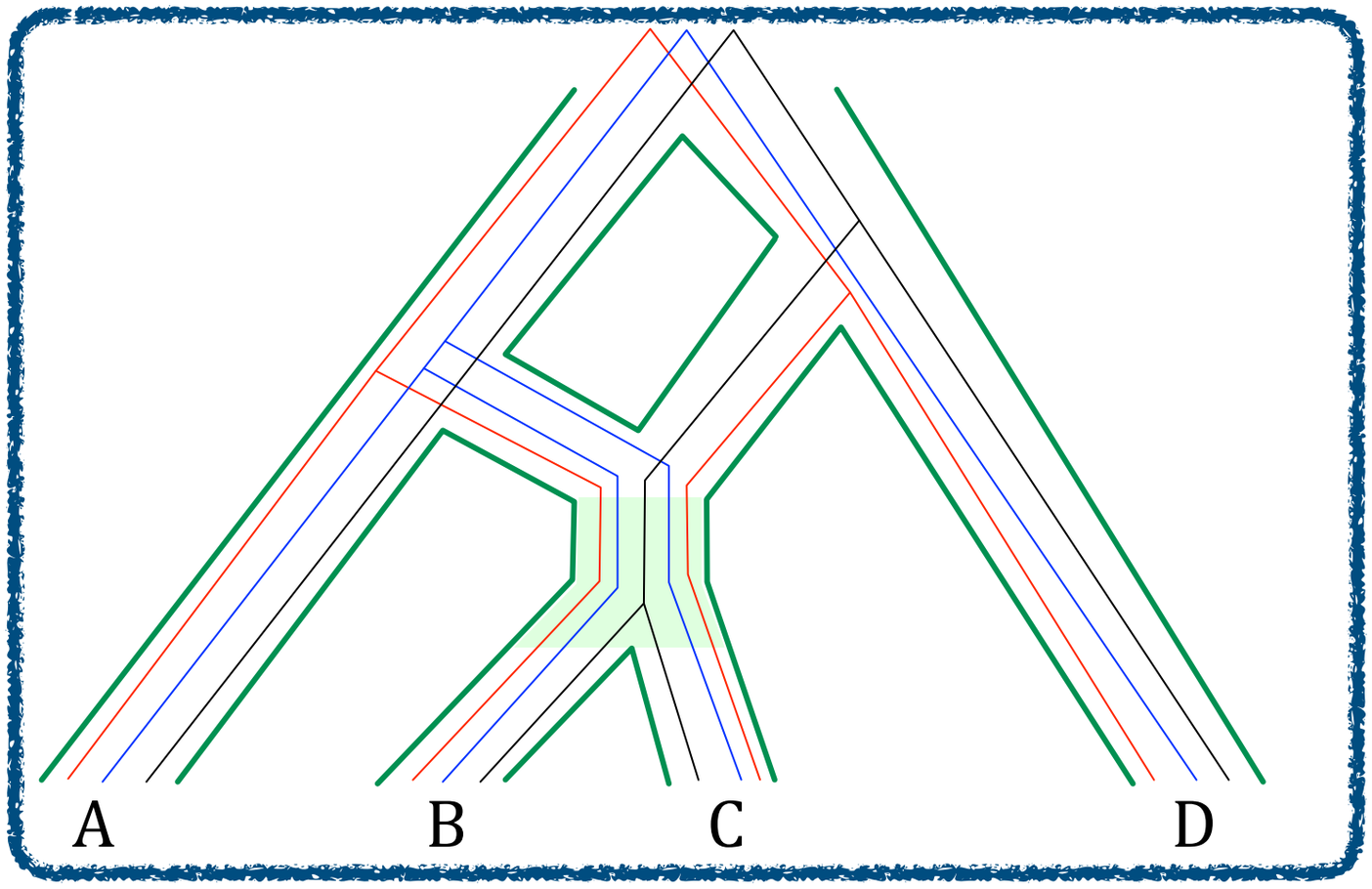
$$(\Psi, \Gamma)$$

Maximum Parsimony Inference

Input gene trees



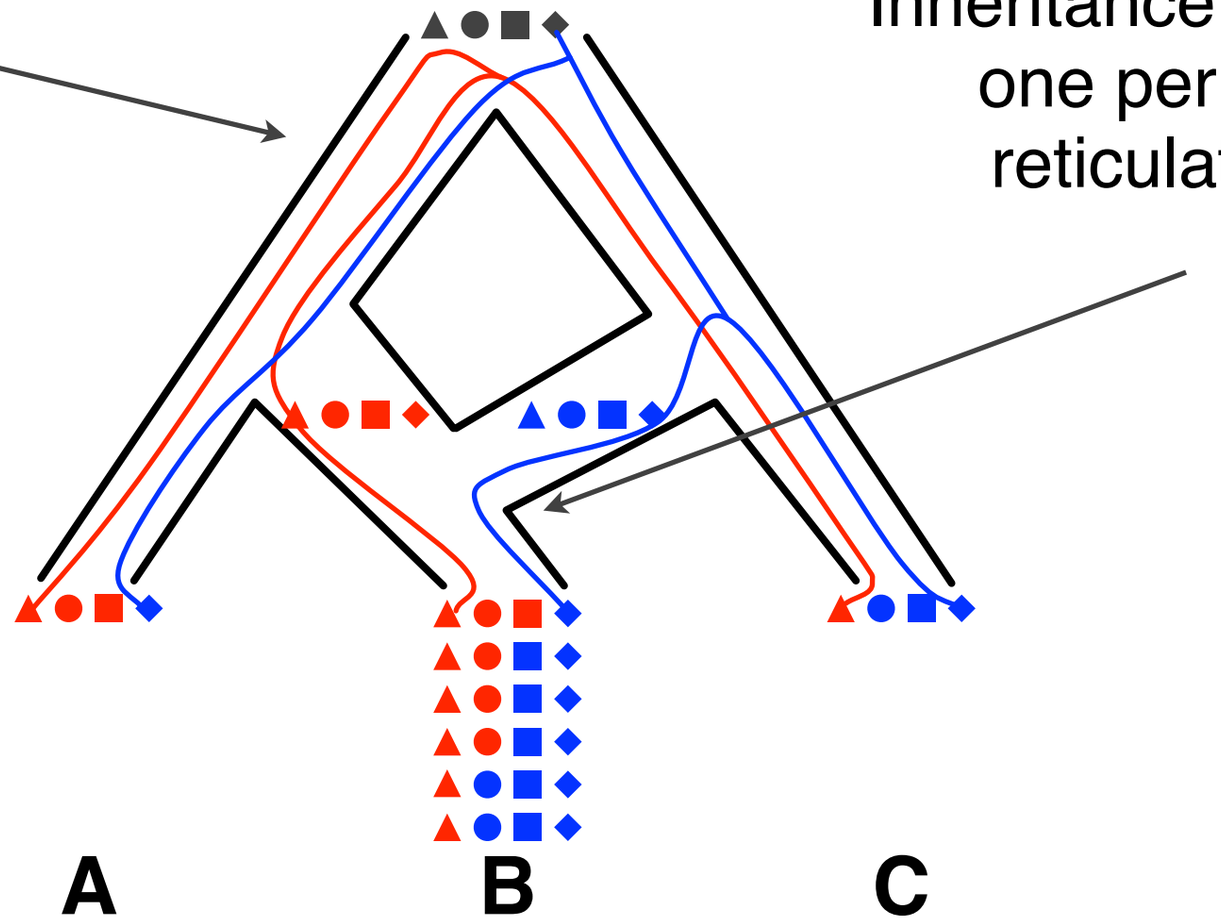
Network with 1 reticulation and 2 extra lineages



Statistical Inference

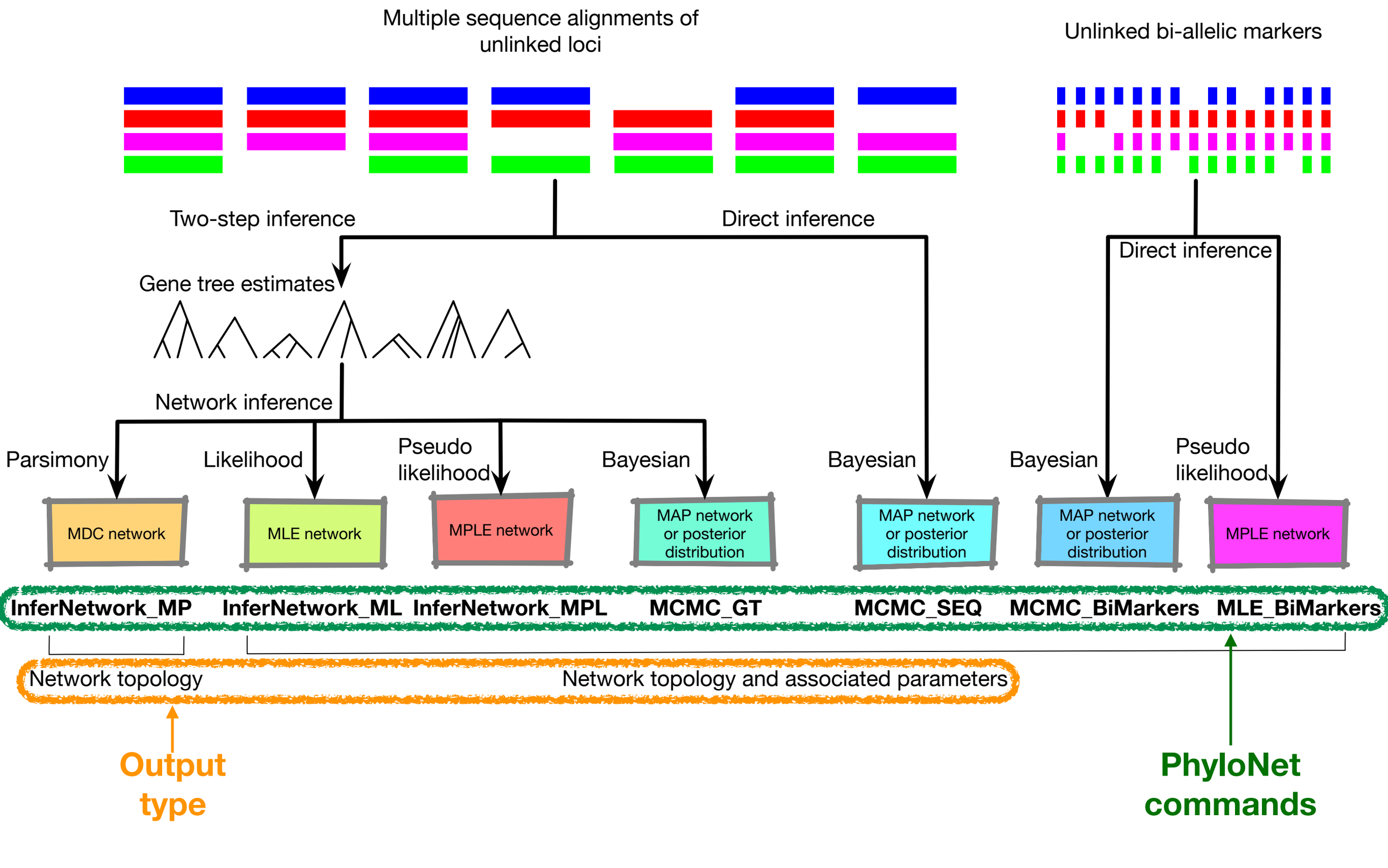
Population sizes, divergence times, …

Inheritance probabilities, one per locus per reticulation node



Assuming independent loci: $p(S|\Psi) = \prod_i \int_G p(S_i|g)p(g|\Psi)dg$

Inference Process and Approaches



List of Inference Commands

Name	Description	Ref	Since
Methods for Species Network (and Tree) Inference (all account for ILS)			
MCMC_SEQ	Bayesian MCMC posterior estimation of phylogenetic networks and gene trees on sequences from multiple independent loci.	here	3.6.1
MCMC_BiMarkers	Bayesian estimation of the posterior distribution of phylogenetic networks given bi-allelic genetic markers (SNPs, AFLPs, etc).	here	3.6.1
MCMC_GT	Bayesian MCMC posterior estimation of phylogenetic networks given a list of gene tree topologies.	here	3.6.0
MLE_BiMarkers	Maximum (pseudo-)likelihood estimation of phylogenetic networks given bi-allelic genetic markers (SNPs, AFLPs, etc).	here	3.6.4
InferNetwork_MPL	Infers a phylogenetic network from gene trees under maximum pseudo-likelihood.	here	3.5.5
InferNetwork_ML_Bootstrap	Infers a phylogenetic network from gene trees under maximum likelihood with parametric bootstrap.	here	3.5.2
InferNetwork_ML_CV	Infers a phylogenetic network from gene trees under maximum likelihood with cross-validation.	here	3.5.2
InferNetwork_ML	Infers a phylogenetic network from gene trees under maximum likelihood.	here	3.4.0
InferNetwork_MP	Infers a phylogenetic network from gene trees under the MDC criterion.	here	3.4.0
NetMerger	Merge subnetworks inferred by MCMC_SEQ or MCMC_BiMarkers to a full network.		

Can we Infer a Tree?

- **Yes**. Since a tree is a special case of network (a network with zero reticulation nodes), all these methods can be used to **infer species trees.**

- Simply **set the maximum number of reticulations to 0** and the methods will search the tree (not network) space!

Can we Fix a Species Tree and Search for Reticulations?

- **NO**, <u>in general</u>, PhyloNet **does not** designate a species tree and search for reticulations to add to it!
- But, if the user wants to start with a species tree and search for “best” reticulations to add to it, these are implemented in InferNetwork_MPL and InferNetwork_MP using -fs.

Computational Difficulty

- Phylogenetic network inference is computationally very hard.
- All the methods in PhyloNet are **heuristics**.
  - (This answers the question “Why did different runs return different networks?”)

How to determine the Number of Reticulations

- It is a very hard problem (the same as the problem of determining the number of clusters in the clustering problem)

- The Bayesian approach performs best at determining the number, as the prior “naturally” accounts for model complexity.

- In general, we recommend incrementally increasing the number of reticulations allowed and comparing the results.

How to determine the Number of Reticulations

- We also recommend **limiting the number** of reticulations allowed in the analysis since it has a huge impact on the computational complexity.

- Inferences based on pseudo-likelihood can scale to larger data sets, though.

Individuals Per Species

- All methods allow data from multiple individuals per species (but that further adds to the computational complexity).
- Missing data (as in missing an entire sequence for a certain locus) is also handled.

16

```
#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;
```

Locus 1:
2 individuals from A
2 individuals from B
0 individuals from C

Locus 3:
1 individual from A
2 individuals from B
1 individual from C

Bayesian inference directly from the sequence data

Mapping individuals to species

Visualizing the Inferred Networks

- Dendroscope: dendroscope.org
- IcyTree: icytree.org

Try PhyloNet for yourself:

Bioinformatics Group

- About
- Add content
- Contact
- Login
- Members
- Projects
- Publications
- Software
- Student Awards
- Support
- Feed aggregator

PhyloNet

View Edit

Read Advances in Computational Methods for Phylogenetic Networks in the Presence of Hybridization.

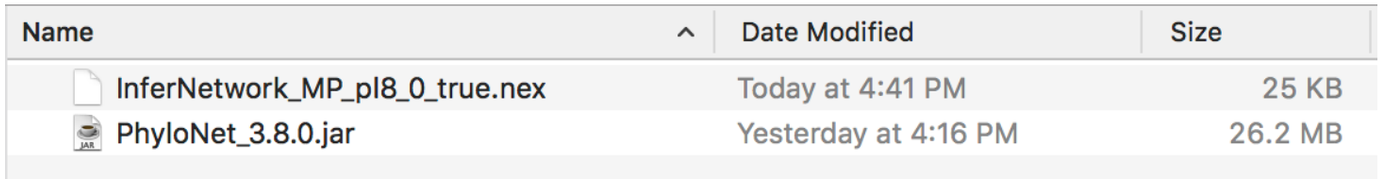
The current version of PhyloNet is 3.8.0.

- **Download**
  - Binary jar file
- **Usage**
  - Talk about PhyloNet (June 4, 2018, by Luay Nakhleh at the SSB Standalone Meeting in Ohio)
  - General overview
  - Tutorial (3 January 2020) at the 2020 Phylogenomics Software Symposium held in conjunction with the SSB Standalone Meeting in Gainesville, Florida.
  - Tutorial: Species phylogeny inference (2017)
  - List of PhyloNet commands (see the figure at the bottom of this page for a summary of the available inference methods)
  - The phylogenetic network format (the Rice Newick format) used in PhyloNet can be readily visualized by Dendroscope.

bioinfocs.rice.edu/phylonet

Warmup Example

1. Download PhyloNet and NEXUS file.



Name	Date Modified	Size
InferNetwork_MP_pl8_0_true.nex	Today at 4:41 PM	25 KB
PhyloNet_3.8.0.jar	Yesterday at 4:16 PM	26.2 MB

2. Open your terminal, change path

```
(base) sousmacbookpuro:ssb zhen$ pwd
/Users/zhen/Desktop/ssb
(base) sousmacbookpuro:ssb zhen$ ls
InferNetwork_MP_pl8_0_true.nex  PhyloNet_3.8.0.jar
```

3. Type the command

```
java -jar PhyloNet_3.8.0.jar InferNetwork_MP_pl8_0_true.nex
```

4. Enter! See your outputs!

```
Results after run #1
363.0: ((((K,P)I4,F)I3,(C,O)I2)I1,L)I0;
Running Time (min): 0.0061333333333333335
================================


Results after run #2
363.0: ((((K,P)I4,F)I3,(C,O)I2)I1,L)I0;
Running Time (min): 0.0029
================================


Results after run #3
363.0: ((((K,P)I4,F)I3,(C,O)I2)I1,L)I0;
Running Time (min): 0.0032166666666666667
================================


Results after run #4
363.0: ((((K,P)I4,F)I3,(C,O)I2)I1,L)I0;
Running Time (min): 0.0026
================================


Results after run #5
363.0: ((((K,P)I4,F)I3,(C,O)I2)I1,L)I0;
Running Time (min): 0.0018333333333333333
================================


Inferred Network #1:
((((K,P),F),(C,O)),L);
Total number of extra lineages: 363.0
```