Statistical models on phylogenetic networks

Claudia Solís-Lemus, PhD
Emory University

August 6, 2019
Phylogenetic Networks

Part I

What?

Why?

How?

When?

Part II

PCM
What?
Phylogenetic network
What?

Phylogenetic network

Explicit

Implicit

Hahn et al (2016)
Why?
Phylogenetic network
Why?
Phylogenetic network

Main tree
Why?
Phylogenetic network

Ignore gene flow
=> Wrong tree!

(S.-L., Yang, Ané, 2016, Syst Bio)
Why?
Phylogenetic network

Coalescent tree methods not robust to gene flow

Number of genes
(S.-L., Yang, Ané, 2016, Syst Bio)

ASTRAL (Mirarab et al, 2014)
NJst (Liu&Yu, 2011)
Anomaly zone with gene flow

Why?
Phylogenetic network

(S.-L., Yang, Ané, 2016, Syst Bio)
Why?
Phylogenetic network

Coalescent tree methods not robust to gene flow

Mean RF distance

Number of genes

(S.-L., Yang, Ané, 2016, Syst Bio)

ASTRAL (Mirarab et al, 2014)
NJst (Liu&Yu, 2011)
PhyloNet (Yu et al 2012, 2014)
Anomalous unrooted gene trees with gene flow

Quartet $\gamma = 0.0$ $\gamma = 0.1$ $\gamma = 0.3$

<table>
<thead>
<tr>
<th>Quartet</th>
<th>$\gamma = 0.0$</th>
<th>$\gamma = 0.1$</th>
<th>$\gamma = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AB</td>
<td>CD$</td>
<td>0.347</td>
<td>0.298</td>
</tr>
<tr>
<td>$CA</td>
<td>BD$</td>
<td>0.327</td>
<td>0.351</td>
</tr>
<tr>
<td>$CB</td>
<td>AD$</td>
<td>0.327</td>
<td>0.351</td>
</tr>
</tbody>
</table>

$t_1 = t_2 = 0.01, t_3 = t_4 = t_5 = 1$

Why?

Phylogenetic network

- **ILS**: no AUGT on 4 taxa (Degnan, 2013)
- **ILS+HGT**: AUGT on 4 taxa (S.-L., Yang, Ané, 2016, Syst Bio)

See also Long & Kubatko (2018) for AUGT under continuous gene flow between sister species
So far...

- Networks are good
- Explicit networks are better
- If you ignore gene flow, you can estimate the wrong tree
How?
Phylogenetic network

BEAST2 (Zhang et al, 2017)
PhyloNet (Wen et al, 2016)

MrBayes (Huelsenbeck, Ronquist, 2001)
RAxML (Stamatakis, 2014)
PhyML (Guindon et al, 2010)

SNaQ (S.-L., Ane, 2016)
PhyloNet (Yu et al, 2014)
How?
Phylogenetic network

Data

Maximum pseudo-likelihood

(Solís-Lemus, Ané, 2016, PLoS Genetics)
(Solís-Lemus et al, 2017, MBE)

www.github.com/CRSL4/PhyloNetworks
How?
Phylogenetic network

Maximum pseudo-likelihood

Data

Unrooted gene trees
No rooting error
No branch lengths
No molecular clock assumption
Concordance factors
Account for tree estimation error

(Solís-Lemus, Ané, 2016, PLoS Genetics)
(Solís-Lemus et al, 2017, MBE)
www.github.com/CRSL4/PhyloNetworks
Quartet-based inference

Concordance factors (CF):
% of genes having the quartet in their tree

1. \( \frac{3}{5} \) for tree A
2. \( \frac{1}{5} \) for tree B
3. \( \frac{1}{5} \) for tree C
Quartet-based inference

**Observed quartet CFs:**

<table>
<thead>
<tr>
<th>4 taxon set</th>
<th>$CF_1$</th>
<th>$CF_2$</th>
<th>$CF_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D</td>
<td>.80</td>
<td>.10</td>
<td>.10</td>
</tr>
<tr>
<td>A B C E</td>
<td>.40</td>
<td>.40</td>
<td>.20</td>
</tr>
<tr>
<td>A B D E</td>
<td>.40</td>
<td>.40</td>
<td>.20</td>
</tr>
<tr>
<td>A C D E</td>
<td>.84</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>B C D E</td>
<td>.82</td>
<td>.10</td>
<td>.08</td>
</tr>
</tbody>
</table>

**Inferred network:**

Maximum Pseudo-Likelihood:

$$\log \tilde{L} = \sum_{q \in Q(N)} CF_{in,1} \log(CF_{net,1}) + CF_{in,2} \log(CF_{net,2}) + CF_{in,3} \log(CF_{net,3})$$
Coalescent model within 1 population

Past

Present

\[ g = \text{generations} \]

\[ N = \text{population size} \]

Probability of no coalescence in \( g \) generations:

\[ \left( 1 - \frac{1}{N} \right)^g \]

\[ t = \frac{g}{N} \Rightarrow \left( 1 - \frac{t}{Nt} \right)^{Nt} \xrightarrow{N \to \infty} e^{-t} \]
Multispecies coalescent on a tree

\[ P(T > t) = e^{-t} \]

\[ T = \frac{g}{N} \text{ coalescent units } \sim Exp(1) \]
Multispecies coalescent on a tree

\[ P(T > t) = e^{-t} \]
Multispecies coalescent on a tree

\[ P(T > t) = e^{-t} \]

\[ P(\bigwedge_{A,B,C}) = 1 - e^{-t} \]
Multispecies coalescent on a tree

\[
P(T > t) = e^{-t}
\]

\[
P\left(\bigwedge_{\text{ABC}}\right) = 1 - e^{-t}
\]
Multispecies coalescent on a tree

\[
P(T > t) = e^{-t}
\]

\[
P(\bigwedge_{A B C}) = 1 - e^{-t} + e^{-t} \times 1/3
\]
Multispecies coalescent on a tree

\[ P(T > t) = e^{-t} \]

\[
P(\bigcap_{A,B,C}) = 1 - e^{-t} + e^{-t} \times \frac{1}{3} = 1 - \frac{2}{3}e^{-t}
\]
Multispecies coalescent on a tree

1 - \(\frac{2}{3}e^{-t}\)

\(\frac{1}{3}e^{-t}\)

\(\frac{1}{3}e^{-t}\)
Multispecies coalescent on a network

(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)
Multispecies coalescent on a network

$P(\quad)$

(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)
Multispecies coalescent on a network

\[ P( 1 \leq t \leq 2) = (1 - \gamma) \frac{1}{3} e^{-t} + \gamma \left( 1 - \frac{2}{3} e^{-t_2} \right) \]

(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)
Multispecies coalescent on a network

\[ P(\quad ) \]

\[ CF_{BC|AD}(t, t_2, \gamma) = (1 - \gamma) \frac{1}{3} e^{-t} + \gamma(1 - \frac{2}{3} e^{-t_2}) \]

(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)
1183 genes, 24 swordtails and platyfish

Xiphophorus fish data

(Solís-Lemus, Ané, 2016, PLoS Genetics)
How?
Phylogenetic network

MrBayes
(Huelsenbeck, Ronquist, 2001)
RAXML
(Stamatakis, 2014)
PhyML
(Guindon et al, 2010)

BEAST2
(Zhang et al, 2017)
PhyloNet
(Wen et al, 2016)

SNaQ
(S.-L., Ane, 2016)
PhyloNet
(Yu et al, 2014)
\[ P(N, G, \theta|D) \propto \pi(N)\pi(\theta) \prod_{i=1}^{L} P(D_i|G_i)P(G_i|N) \]

Prior
Network

Birth-hybridization process
BEAST2
(Zhang et al, 2017)

# reticulations, cycle diameter
PhyloNet
Bayesian
(Wen et al, 2016)
The prior distribution for the network is given by:

$$P(N, G, \theta | D) \propto \pi(N) \pi(\theta) \prod_{i=1}^{L} P(D_i | G_i) P(G_i | N)$$

The likelihood of the network is:

$$P(N, \theta | G) \propto \pi(N) \pi(\theta) \prod_{i=1}^{L} P(G_i | N, \theta)$$

The log-likelihood of the network is:

$$L(N, \theta) = \prod_{i=1}^{L} P(G_i | N, \theta)$$

The log-likelihood for quartet sets is:

$$\tilde{L}(N, \theta) \propto \prod_{q} L(q | N, \theta)$$

The substitution model is:

$$\pi(N) \pi(\theta)$$

The multispecies coalescent is:

$$P(G_i | N, \theta)$$

The birth-hybridization process is:

$$P(D_i | G_i)$$

The 4-taxon sets are:

$$CFS$$

The quartet sets are:

$$12|34 \ 13|24 \ 14|23$$
<table>
<thead>
<tr>
<th>Method</th>
<th>Quadruplet Sets</th>
<th>Birth-hybridization process</th>
<th>Rooted/Unrooted</th>
<th>Most accurate, not scalable</th>
<th>MCMC: Network moves, mixing</th>
<th>Heuristic search: Network moves</th>
<th>More accurate, Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAST2 (Zhang et al, 2017)</td>
<td></td>
<td></td>
<td>Rooted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyloNet Bayesian (Wen et al, 2016)</td>
<td></td>
<td># reticulations, cycle diameter</td>
<td>Rooted</td>
<td></td>
<td>MCMC: Network moves, mixing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyloNet Bayesian (Wen et al, 2016)</td>
<td></td>
<td></td>
<td>Rooted</td>
<td></td>
<td></td>
<td>Heuristic search: Network moves</td>
<td></td>
</tr>
<tr>
<td>PhyloNet Likelihood (Yu et al, 2014)</td>
<td></td>
<td></td>
<td>Rooted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNaQ (S.-L., Ane, 2016)</td>
<td></td>
<td></td>
<td>Unrooted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tool</td>
<td>Type</td>
<td>Method</td>
<td>Prior/Assumption</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------</td>
<td>----------------</td>
<td>-------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STEM-hy</td>
<td>gene trees rooted, BL</td>
<td>likelihood</td>
<td>hybridization b/w sister lineages</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyloNet</td>
<td>gene trees rooted</td>
<td>likelihood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyloNet</td>
<td>gene trees rooted</td>
<td>triplet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyloNetworks</td>
<td>gene trees or quartet CFs</td>
<td>quartet</td>
<td>level-1 network</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNaQ</td>
<td></td>
<td>likelihood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyloNet</td>
<td>gene trees rooted</td>
<td>Bayesian</td>
<td>compound prior</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyloNet</td>
<td>alignments</td>
<td>Bayesian</td>
<td>compound prior no rate variation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyloNet</td>
<td>alignments</td>
<td>Bayesian</td>
<td>birth-hyb prior</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SpeciesNetwork</td>
<td></td>
<td>Bayesian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyloNet</td>
<td>biallelic sites</td>
<td>likelihood</td>
<td>compound prior</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyloNet</td>
<td>biallelic sites</td>
<td>Bayesian</td>
<td>compound prior</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HyDe</td>
<td>sites</td>
<td>invariants</td>
<td>4 taxa, 1 hyb.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Challenges

- Network space
- Identifiability
- Network comparison
Identifiability

$L(\theta|X)$
Reconstructible Phylogenetic Networks: Do Not Distinguish the Indistinguishable

Fabio Pardi¹,³*, Celine Scornavacca²,³

1 Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM, UMR 5506) CNRS, Université de Montpellier, France, 2 Institut des Sciences de l'Évolution de Montpellier (ISE-M, UMR 5554) CNRS, IRD, Université de Montpellier, France, 3 Institut de Biologie Computationnelle, Montpellier, France

Undistinguishable with the “displayed trees” criterion

Solution: Canonical network (“unzipped”)
Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

Sha Zhu\textsuperscript{1}, James H. Degnan\textsuperscript{2}

Distinguishable under the MSC
Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

Sha Zhu\textsuperscript{1}, James H. Degnan\textsuperscript{2}

Decomposing network in \textit{parental} trees
Can we detect the presence of hybridization in level-1 networks?

No

Yes

$(n_i, n_j \geq 2)$

Yes

$(n_i \geq 2)$

Yes

Generic Identifiability $t_i \in (0, \infty), \gamma \in (0, 1)$
In practice:
flat pseudolikelihood

(S.-L., Ané, 2016, PLoS Genetics)
Can we estimate numerical parameters?

**Yes**

Good diamond

\( n_0, n_2 \geq 2 \)

**Yes**

Generic Identifiability

\( t_i \in (0, \infty), \gamma \in (0, 1) \)

**No**

Good triangle

\( t_{12} = 0 \)
Identifiability matters: SNaQ performance

Good diamond

Bad diamond

(S.-L., Ané, 2016, PLoS Genetics)
Challenges

- Network space
- Identifiability
- Network comparison

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
**Missing:** likelihood, level-k semi-directed
Challenges

- Network space
- Identifiability
- Network comparison

**Missing**: path through tree space, semi-directed

K. Huber, V. Moulton, C. Scornavacca,...

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
**Missing**: likelihood, level-k semi-directed
Challenges

• Network space
  - Missing: path through tree space, semi-directed

• Identifiability
  - Displayed vs Parental trees
  - Level-1 semi-directed networks
  - Hybridizations: case by case
  - Missing: likelihood, level-k semi-directed

• Network comparison
  - Missing: distance function
  - Hardwired-cluster distance only for rooted networks
  - Summary of networks: clades!
Network summary

(S.-L. et al, 2017, MBE)
Network summary

Hybrid clades

Minor sister clades
When?
Phylogenetic network

Goodness-of-fit test
Hypothesis test: Is a tree a good fit?

Data

TICR

GitHub

https://github.com/nstenz/TICR
(Stenz et al, 2015, Syst Bio)
Part II
I have the network, now what?

- Sword index
- Female preference

(Cui et al., 2013)
(Solís-Lemus, Ané, 2016, PLoS Genetics)
Trait models of evolution in networks

Brownian Motion + weighted average in hybrid

\[ X_h = \gamma_1 X_{p1} + \gamma_2 X_{p2} \]

(Bastide et al, 2018, Syst Bio)

\[ X \sim N(X_{root}, \sigma^2 V) \]

- Phylogenetic signal
- Ancestral reconstruction
- Phylogenetic regression
- Phylogenetic ANOVA

www.github.com/CRSL4/PhyloNetworks
- Sword index
- Female preference

- Ancestral reconstruction: common ancestor likely had sword
- Phylogenetic regression: positive association between sword index and female preference but not significant ($p = 0.106$)
Test for transgressive evolution

\[ X_h = \gamma_1 X_{p1} + \gamma_2 X_{p2} + \Delta_h \]

\[ \Delta_h = 0 \quad \text{No transgressive evolution} \]
\[ \Delta_h = \Delta \quad \text{Single-effect transgressive evolution} \]
\[ \Delta_h \quad \text{Multi-effect transgressive evolution} \]

Hybrid value: shift from parents range

F tests
Test for transgressive evolution

- Sword index: $p=0.55$
- Female preference: $p=0.0064$

Hybrid value: shift from parents range
PhyloNetworks: analysis for phylogenetic networks

Overview

PhyloNetworks is a Julia package with utilities to:

- read / write phylogenetic trees and networks, in (extended) Newick format. Networks are considered explicit: nodes represent ancestral species. They can be rooted or unrooted.
- manipulate networks: re-root, prune taxa, remove hybrid edges, extract the major tree from a network, extract displayed networks / trees
- compare networks / trees with dissimilarity measures (Robinson-Foulds distance on trees)
- summarize samples of bootstrap networks (or trees) with edge and node support
- estimate species networks from multilocus data (see below)
- phylogenetic comparative methods for continuous trait evolution on species networks / trees

http://crsl4.github.io/
(S.-L. et al, 2017, MBE)
Acknowledgements

Cécile Ané (UW)  Mengyao Yang
Paul Bastide (KU-Leuven)  John Malloy
Bret Larget (UW)  John Spaw
Douglas Bates (UW)  Noah Stenz
David Baum (UW)  Nan Ji
Sarah Friedrich (UW)  Jordan Vonderwell
Michael Epstein (Emory)  Josh McGrath

http://crsl4.github.io/