Identifying features of phylogenetic networks from various data types

IMSI workshop - Aug 11-14, 2025

Cécile Ané

- phylogenetic networks
- known non-identifiable features
- known identifiable features

review & joint work with

Hector Baños , John Rhodes, Elizabeth Allman, Jingcheng Xu

phylogenetic networks

and admixture graphs

examples

Fig. 4a in **Salem et al (2025, Nature)** using ancient human DNA, from f_4 statistics: site patterns across 4 taxa

Fig. 3 in **Lagou et al. 2024** on slipper orchids, from 913 gene trees, quartets

data types:

- **quartet concordance factors**:
  % genes with $ab|cd$, $ac|bd$ and $ad|bc$, for all subsets a,b,c,d
- average genetic **distances**, log-det distances
- f_4 statistics: linear combinations of $f_2 \simeq$ distances
- frequencies of full gene trees, or full site patterns

models:

- coalescent: **common** or **independent** inheritance at hybrids
- without the coalescent: gene trees **displayed** in the network

the network coalescent model

edge lengths in coalescent units: # generations / N_e
hybrid edges: inheritance probability γ

---

species network, or admixture graph

---

gene tree that can occur

(C,D) sister in some genes: from gene flow *or* the coalescent

the network coalescent model

model parental ancestry of lineages at a hybrid node

- independent inheritance: independent parents
- common inheritance: identical parents

Fogg, Allman & Ané **2023** PhyloCoalSimulations

---

multiple lineages at the hybrid node: their parents could be correlated, e.g. due to selection

with the coalescent

Solís-Lemus & Ané **2016**
Baños **2019**
Allman, Baños & Rhodes **2022**: log-det
Allman, Baños, Mitchell & Rhodes **2023**
Allman, Baños, Garrote-Lopez & Rhodes **2024**
Rhodes, Baños, Xu & Ané **2025**
Allman, Ané, Baños & Rhodes **2025**
Holtgrefe et al. **2025**

without the coalescent
(less gene tree variation)

Gross et al. **2021**
Xu & Ané **2023**
Englander, Frohn et al. **2025**

known non-identifiable features

from most data types
under most models (allowing for rate variation)

the root position is *not* identifiable

---

infer the **semidirected** network:

- no root
- hybrid edges: directed
- tree edges: not directed

small blobs are *not quite* identifiable

- blob: not disconnected by removing an edge, maximal
- *m*-blob: *m* attachment nodes
  disconnects network into *m* blocks of taxa

2-blobs are not identifiable

- average distances
- quartet CFs (but perhaps from *quintet* CFs, Cummings et al.)

3-blobs are not identifiable

- average distances
- quartet CFs if 2 blocks have only 1 taxon

The **hybrid position** is not identifiable

- in a 3-cycle
- in a 4-cycle: distances, quartet CFs if 4 blocks of 1 taxon

known identifiable features

from most data types

the reduced tree-of-blobs is identifiable

- shrink each blob
- suppress degree-2 nodes

level-1 networks are (mostly) identifiable

the circular order is identifiable

in **outer-labeled planar** networks

\+ extra conditions depending on data & model

outer-labeled planar blobs

planar: no crossing edges
outer-labeled: taxa (or taxon blocks) on the outside

In an outer-labeled planar blob,
the **circular order** of taxa is **well defined**.

---

different planar embedding must have *a*,*b*,*c*,*d* in the same order along the outer face

> For a binary outer-labeled planar blob, the **full** circular order is **identifiable** from the order on **4-taxon subsets**.

4-taxon information:
(abcd)
(hbcd)
(acdh)
(abdh)
bc — ah : (bcah) and (bcha)

> For binary outer-labeled planar networks, the **tree of blobs** and each blob **circular order** is identifiable.

from many data types:

- quartet concordance factors
- average distances
- logDet distances (assuming ultrametric networks)

> For binary outer-labeled planar networks, the **tree of blobs** and each blob **circular order** is identifiable.

and under various models:

- displayed-tree model (no coalescent)
- coalescent model with common inheritance
- coalescent model, independent inheritance *if no anomaly*

anomaly example

anomalous CFs if
$\%ca|bd = \%cb|ad > \%ab|cd$
anomalous distances if

$$\begin{aligned} D(c,a) + D(b,d) &= D(c,b) + D(a,d) \\ &< D(a,b) + D(c,d) \end{aligned}$$

but…

not distinguishable, from distances or quartet CFs

---

model with or without the coalescent

galled tree-child networks

are **identifiable**, if they have **large cycles**
from, e.g., quartet concordance factors

tree-child: each node has at least one tree child
galled: each hybrid in only 1 cycle

1. Assume: we can identify the tree of blobs

2. to identify each blob: sample one taxon from each block

network assumption: the bloblet is $\mathfrak{C}_k$, $k = 4$ or 5

- galled, tree-child, and
- for every taxon x of hybrid origin, the subnetwork on the skeleton taxa $Y \cup \{x\}$ has an m-cycle with $m \geq k$.

.

> **general data/model assumptions**
>
> 1. the **tree of blobs** is identifiable
> 2. for level-1 blobs on **4** taxa, the **circular order** is identifiable
> 3. for networks that reduce to a level-1 blob on **4** taxa, the **length** of internal tree edgs are identifiable
> 4. for networks on **3** taxa, we can identify whether the internal blob is **trivial or not**.

prove 2-4 on small networks, then

1-3: $\mathfrak{C}_5$ blobs are identifiable
1-4: $\mathfrak{C}_4$ blobs are identifiable

.

> **general data/model assumptions**
>
> 1. the **tree of blobs** is identifiable
> 2. for level-1 blobs on **4** taxa, the **circular order** is identifiable
> 3. for networks that reduce to a level-1 blob on **4** taxa, the **length** of internal tree edgs are identifiable
> 4. for networks on **3** taxa, we can identify whether the internal blob is **trivial or not**.

prove 2-4 on small networks, then

1-3: $\mathfrak{C}_5$ blobs are identifiable
1-4: $\mathfrak{C}_4$ blobs are identifiable

🥴 identify taxa Y not below a hybrid

🙂 skeleton subtree on Y

🙂 for x of hybrid origin: level-1 subnetwork on $Y \cup \{x\}$

- circular order → topology
- relative edge lengths along skeleton → combine

- $\mathfrak{C}_4$ blobs: identifiable from quartet CFs, 2 samples/taxon, coalescent model
- $\mathfrak{C}_5$ blobs: identifiable from quartet CFs, ≥ 1 sample/taxon; or average distances; coalescent or displayed tree model

full network example

$\mathfrak{C}_5$ blobs: identifiable from only 1 sample/taxon
3-cycle: its presence can be identified

but…

non-$\mathfrak{C}_5$ networks could be non-distinguishable

level-2 networks

need extra constraints

level-2, tree-child, without 3-cycles

are identifiable: Englander et al. **2025**
from frequencies of full site patterns, no coalescent

level-2, galled & outer-labeled planar

their ‘canonical’ form is identifiable: Holtgrefe et al. **2025**
from data & models that identify displayed quartets
e.g. quartet CFs under the coalescent

a canonical graph:

- has no 2-blob, no 3-blob, no 3-cycle
- 4-blobs are undirected cycles, 5-blobs are cycles
- is identified by the splits of its displayed trees

lots of open questions!

- to go beyond level-2+ or (tree-child & galled)
- for more models & data types

joint work with
Hector Baños , John Rhodes, Elizabeth Allman, Jingcheng Xu

circular order: **Rhodes et al. 2025**
galled tree-child: **Allman et al. 2025**

thanks to the National Science Foundation and to the
Wisconsin Alumni Research Foundation

Speaker notes