Beyond level-1: Identifying phylogenetic network features

ICERM
Oct. 23, 2024

John A. Rhodes
UNIVERSITY OF ALASKA FAIRBANKS

Collaborators:

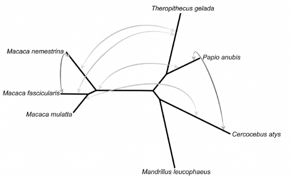
Cecile Anè, U Wisconsin Madison
Hector Baños, CSU San Bernadino
Jingcheng Xu, U Wisconsin Madison

Significant Non-Collaborator:

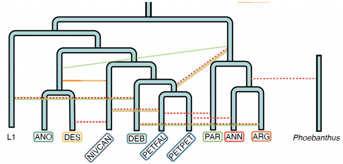
Elizabeth Allman, UAF

DMS-2023239, DMS-2051760, DMS-2331660

Q: If a collection of species are related by a phylogenetic network — describing hybridization, introgression, or some form of lateral gene flow —



(Vanderpool et al., PLOS Biol.,2020)



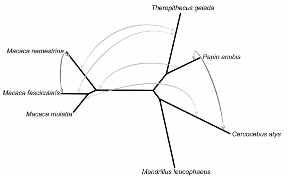
(Owens et al., MBE, 2023)

(1) What features of the network are **identifiable**?

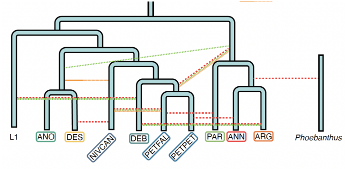
identifiable = is determined by the probability distribution for data under some model
≈ is theoretically inferable from data

"data" ≈ something a biologist can collect or infer well,
e.g., alignable sequences, unrooted gene tree topologies, allele frequencies, . . .

Q: If a collection of species are related by a phylogenetic network — describing hybridization, introgression, or some form of lateral gene flow —

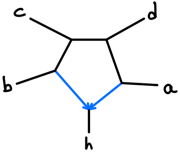


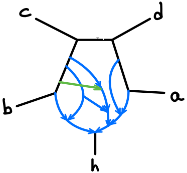
(Vanderpool et al., PLOS Biol.,2020)



(Owens et al., MBE, 2023)

(2) What is **not** identifiable?





Identifiability results can help in development/use of practical inference methods:

- What data type makes inference of certain features possible?
- What assumptions are needed to justify an inference method?
- What is un-inferable?

and sometimes suggests useful inference approaches.

E.g., Under the NMSC model:

- Unrooted species tree topology is identifiable from quartets on gene trees $\rightsquigarrow$ ASTRAL
- Network's Tree of Blobs is identifiable from quartets on gene trees $\rightsquigarrow$ TINNiK

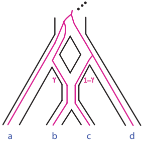
Network identifiability is actually **many** questions, depending on...

- **model:**
  - gene tree formation: NMSC, NMSC w/ common inheritance, Displayed trees (no ILS)
    gene duplication/loss,. . .
  - network structure: level-k, tree-child, outer-labelled planar, . . .
  - sequence evolution on gts: rate matrices, scalar rate variation, conversion of coal. units to subs. units,. . .
- **features sought:** full structure, tree of blobs, something in-between
- **data type:** site patterns for non-recomb. loci, genomic site patterns, gene trees (metric/topo, rooted/unrooted), quartet CFs, allele frequencies,. . .

Quartet CFs are a currently popular data type:

Inferred trees for different genes on the same taxa have many different topologies,

Likely causes: incomplete lineage sorting and/or
network reticulations
(+ dup/loss, gt inference error,. . .).



For any fixed model of gene tree formation on a species network,

$$CF_{ab|cd} = \text{probability that } ab|cd \text{ is displayed on a gene tree}$$

$$CF_{abcd} = (CF_{ab|cd}, CF_{ac|bd}, CF_{ab|bc})$$

Quartet CFs

- have nice properties under standard models
- easy to estimate well from sequences (infer unrooted gts, ignore edge lengths, count)
- allow for faster inference schemes than full gene trees (possibly at some cost...)

From quartet CFs and NMSC model, current understanding of network identifiability is roughly...

For a level-1 network we can identify the semidirected (unrooted) structure*.

* mostly...

.

.

.

dark abyss[1]

.

.

.

For a general network we can identify the tree of blobs.

[1] Milton, *Paradise Lost*

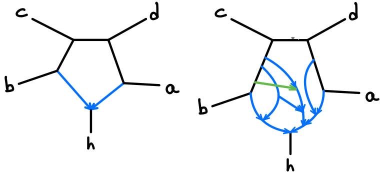
New result

To begin filling this gap...

When it makes sense to have an "order of taxa around a blob", the order is identifiable for several models & data types, even though the full blob structure may not be.

and some non-identifiability examples

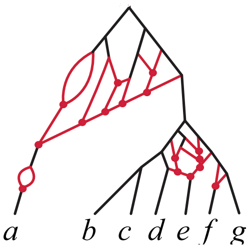
Networks may be indistinguishable using common data types.



Part I (combinatorics)

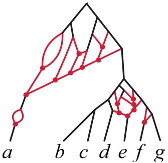
Consider binary networks with **outer-labelled planar (OLP)** blobs.

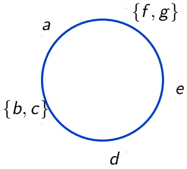
Ex:



Outer-labelled planar = embeddable in plane with
(1) no edge crossings,
(2) taxa on "outside"

An OLP blob embedded in the plane defines a circular order of taxon groups:

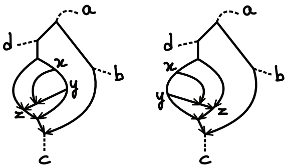




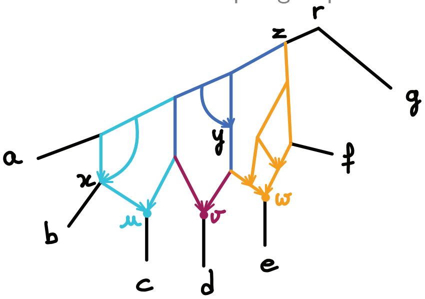
This order gives a weak notion of relatedness of taxa.

Theorem: Outer-labelled planar blobs have a unique circular order of taxon groups, independent of planar embedding.

Two embeddings, same order:



Consider an OLP blob in N restricted to 1 taxon per group around it.

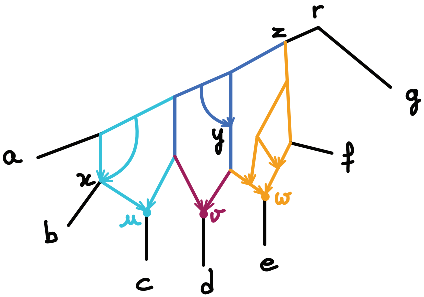


For a subset of four taxa $Q = \{a, b, c, d\}$, the induced subnetwork has 1 or 2 compatible circular orders

$\{a, b, c, e\}$ has one order: $(a, b, c, e) = (b, c, e, a) = \cdots = (e, c, b, a)$.

$\{a, b, e, f\}$ has two orders: $(a, b, e, f) \neq (a, b, f, e)$.

Quartet order information for the blob is the set of all such circular orders.



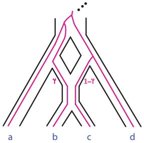
Theorem: For an OLP blob, quartet order information determines the blob's circular order.

For **identifiability** of the circular order, we still need to show quartet order information can be identified from data under some model.

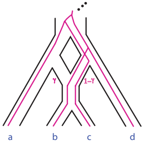
Part II (Data/Model)

Consider **3 models** of gene tree formation within a network:

- ILS modeled by NMSC w/ independent inheritance
  + 'no anomaly' condition



- ILS modeled by NMSC w/ common inheritance



- Displayed trees (no ILS), limit of coalescent models as pop. size → 0 and **3 summary data types**:

Data types:

1. average distances from gene sequences

   compute distances between taxa for each gene, then average

2. quartet concordance factors (CFs)

   infer gene trees, count their displayed quartets

3. logDet distances from genomic sequences

   concatenate gene sequences, compute logDet distances between taxa

(1 and 3 require specifying sequence evolution model)

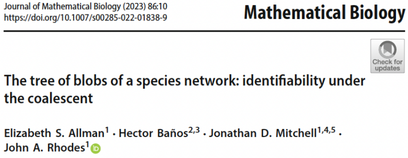
Theorem: For an OLP blob in a binary network, quartet order information is identifiable for any of the 3 gene tree models and 3 data types.

Specifically, for taxa a, b, c, d, the quartet order can be found using

1. average distances across genes — by relative magnitudes of sums of distances like $d_{avg}(a, b) + d_{avg}(c, d)$
2. quartet CFs — by relative magnitudes of entries
3. logDet distance — by relative magnitudes of 3 distances among triples of taxa

To apply this, we next need to know taxon groups around the blobs.

Theorem: For a binary network, the tree of blobs is identifiable for these 3 gene tree models and 3 data types.

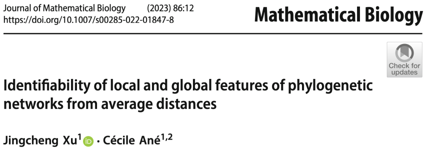


Journal of Mathematical Biology (2023) 86:10
https://doi.org/10.1007/s00285-022-01838-9

Mathematical Biology

The tree of blobs of a species network: identifiability under the coalescent

Elizabeth S. Allman[1] · Hector Baños[2,3] · Jonathan D. Mitchell[1,4,5] · John A. Rhodes[1]



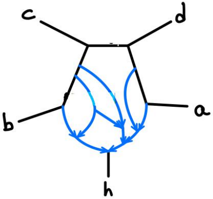
Journal of Mathematical Biology (2023) 86:12
https://doi.org/10.1007/s00285-022-01847-8

Mathematical Biology

Identifiability of local and global features of phylogenetic networks from average distances

Jingcheng Xu[1] · Cécile Ané[1,2]

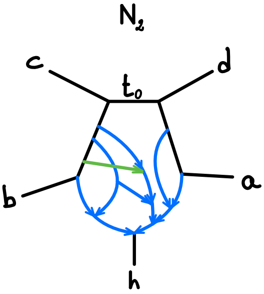
Corollary: For these 3 gene tree models and 3 data types, the tree of blobs with circular orders for OLP blobs are identifiable.

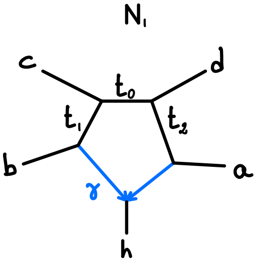


Part III: Examples of non-identifiability

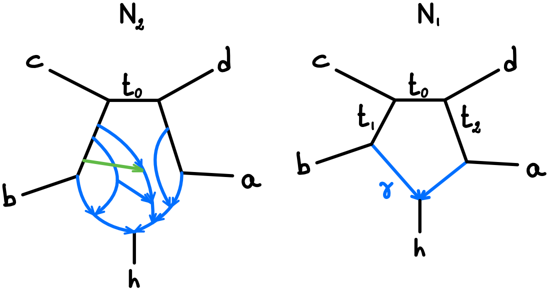
But these same models/data types can lead to non-identifiability of network topology:

Ex: Under any of the 3 models of gene tree formation, for arbitrary parameters on N_2 there exits t_1, t_2, γ on N_1 leading to identical quartet CFs.





Similar examples can be given for other two data types.



This means there are networks which from quartet CFs we cannot determine

1) anything about the level of the network
2) whether the network is tree-child
3) whether the network is OLP

Open question

Even when a blob's full structure is not identifiable, what other features of it might be identifiable, and what model/data type might allow this?

E.g.,
1) Which lineages exiting a blob have hybrid origin (descended from a hybird node in the blob)?

2) If a lineage is of hybrid origin, what are its hybrid siblings (lineages exiting the blob whose ancestors in the blob contributed to the hybid)?

Others?

Reference

“Identifying circular orders for blobs in phylogenetic networks,” Rhodes, Baños, Xu, Anè , 2024, `arXiv:2402.11693`