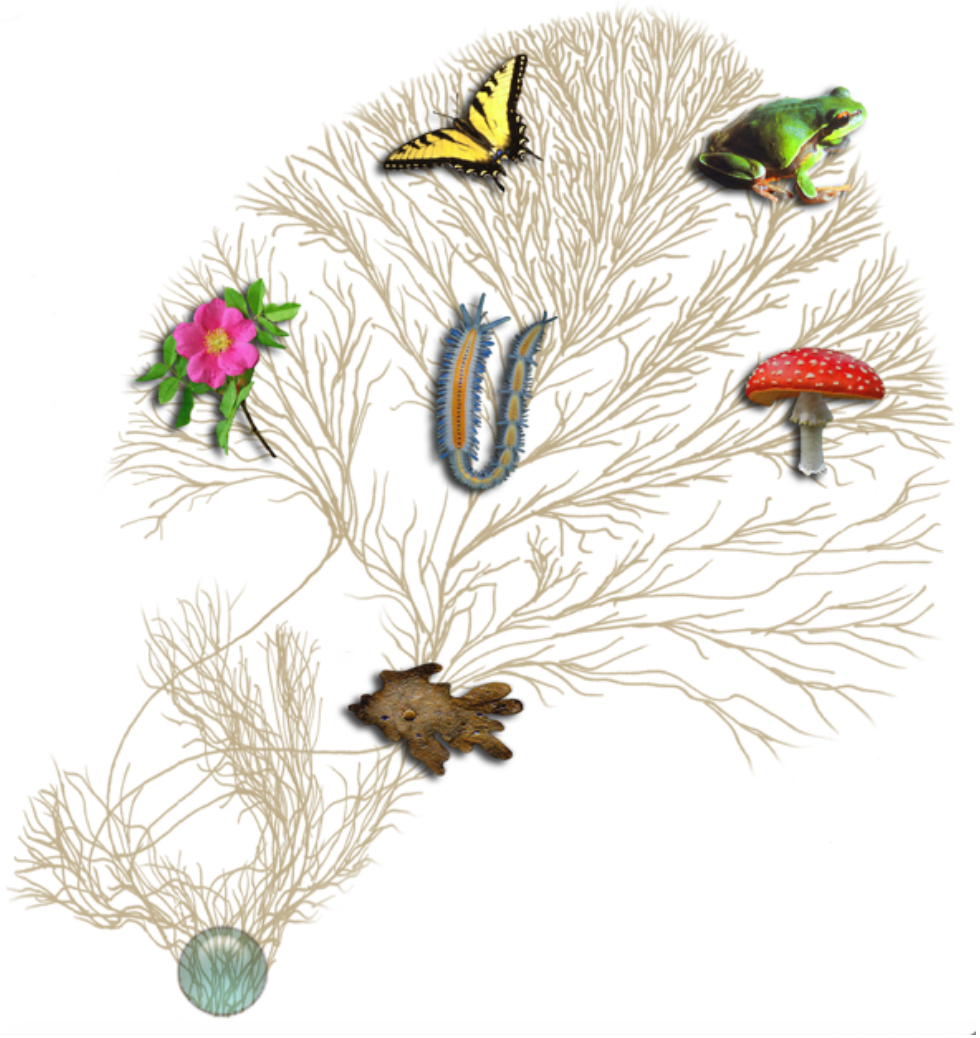
Phylogenetic networks with multiple roots

Katharina Huber

September 18, 2024

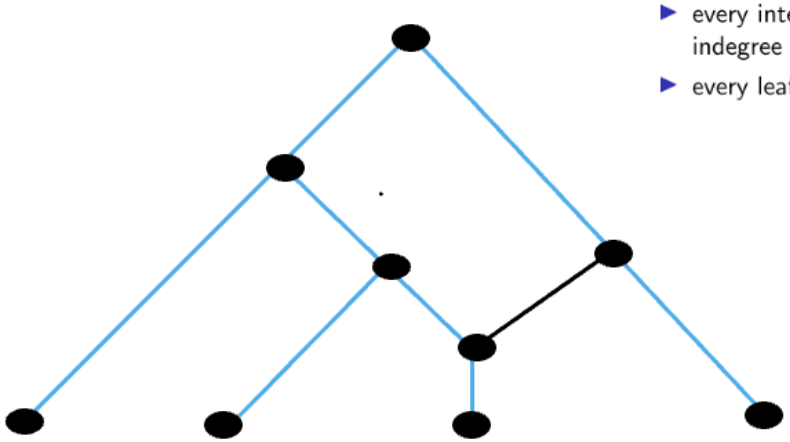
Many of us believe ...



Tree of Life image © 2007 Tree of Life Web Project. Image of rose © 1999 Nick Kurzenko. Image of annelid worm © 2001 Greg W. Rouse.

But if we look closer ..

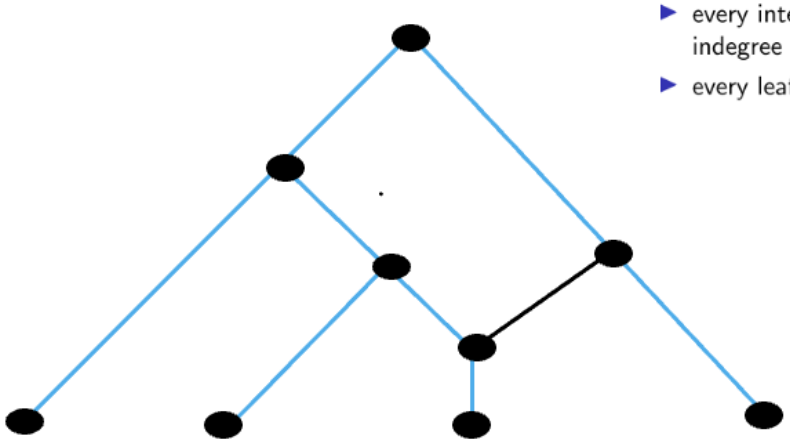
One might think that ...



A (rooted) phylogenetic network for a set X of species is a rooted DAG with leaf set X such that

- the root has indegree 0 and outdegree 2;
- every internal vertex either has indegree 1 and outdegree 2 or indegree 2 and outdegree 1;
- every leaf has indegree 1.

One might think that ...



A (rooted) phylogenetic network for a set X of species is a rooted DAG with leaf set X such that

- the root has indegree 0 and outdegree 2;
- every internal vertex either has indegree 1 and outdegree 2 or indegree 2 and outdegree 1;
- every leaf has indegree 1.

A phylogenetic network with vertex set V and leaf set X is called *tree-based* if there exists a subset A of arcs of N so that (V, A) is a tree with leaf set X (Francis and Steel, Systematic Biology, 2015).

but...

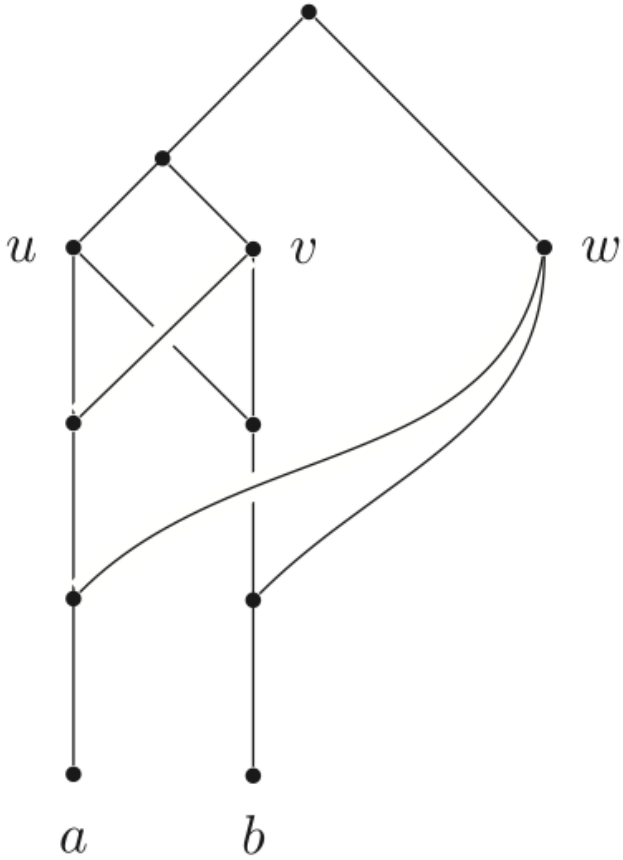
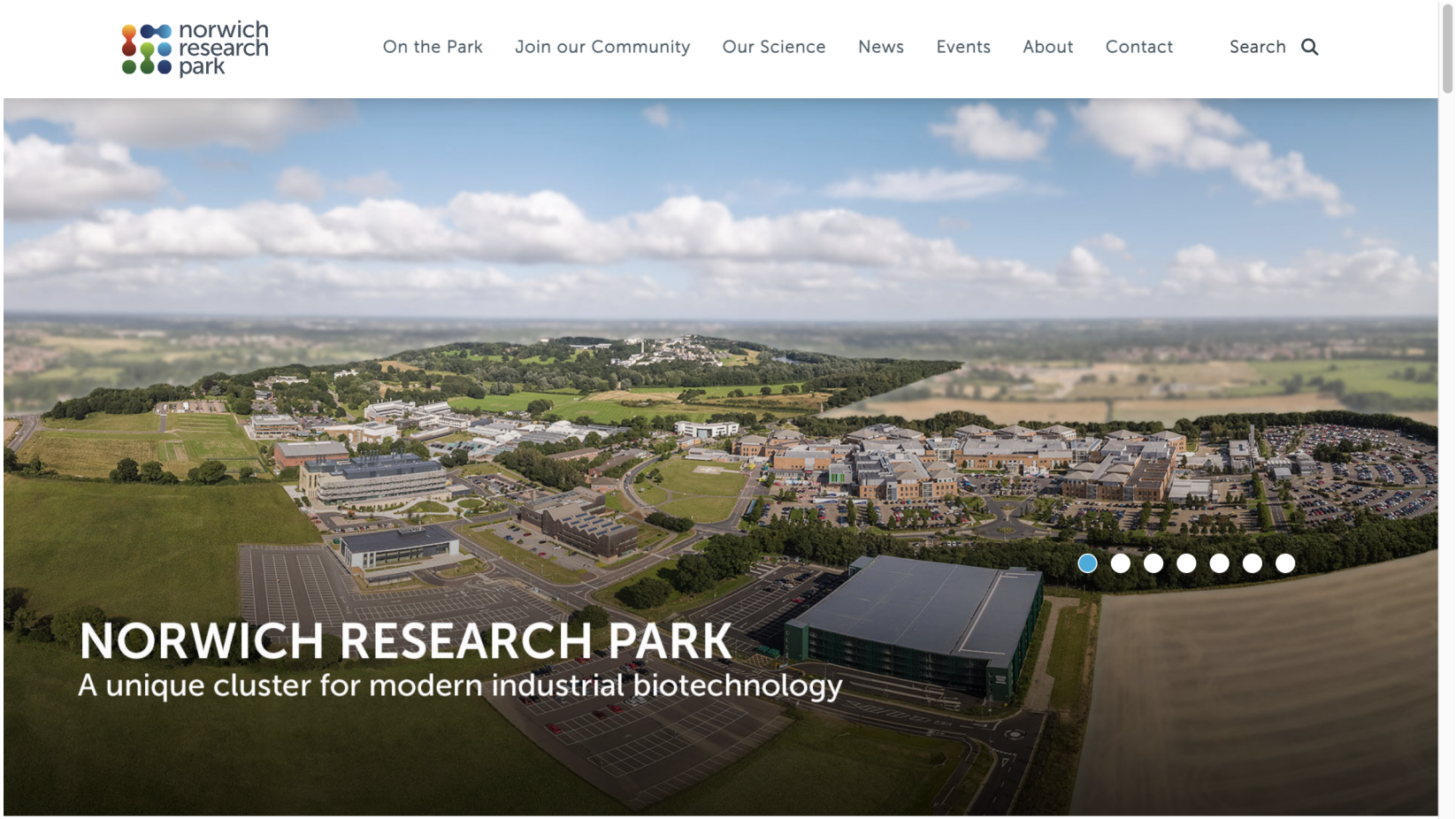


Figure: Francis and Steel, Systematic Biology, 2015 (see also van Iersel, 2013).

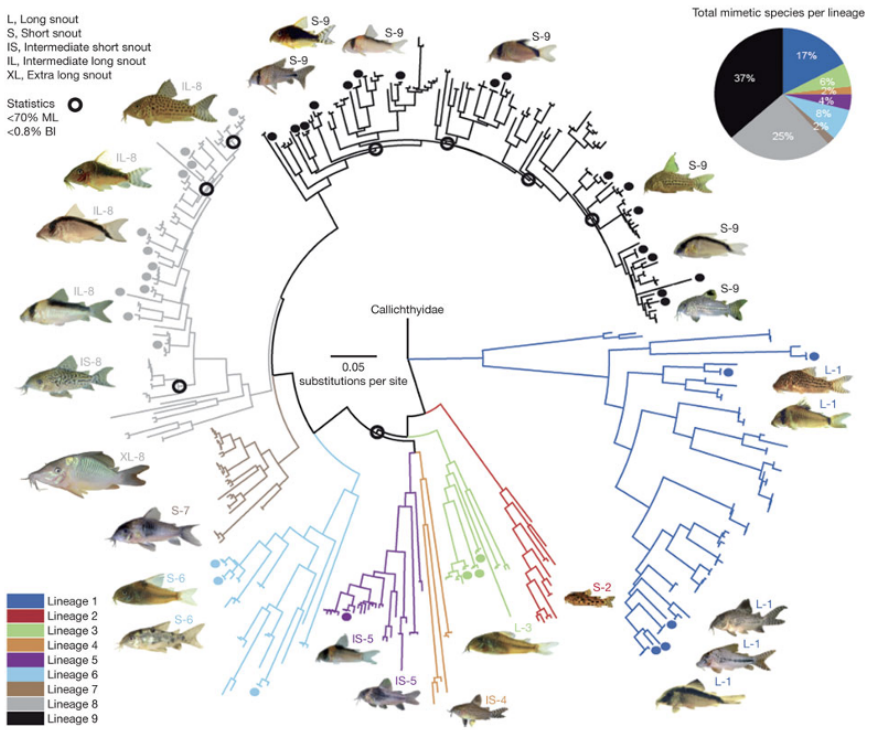
However this is not the whole story![1]



[1] see (Fischer, Hamann, Wicke, DAM, 2023) for an alternative twist of the story.

Corydoradinae evolution

Phylogenetic relationships of Corydoradinae including co-mimics.



MA Alexandrou *et al. Nature* **469**, 84-88 (2011) doi:10.1038/nature09660

Butterfly wing pattern evolution

Fig 5. Hypothesis for the origins and introgression of the dennis and ray regions inferred from dated trees.

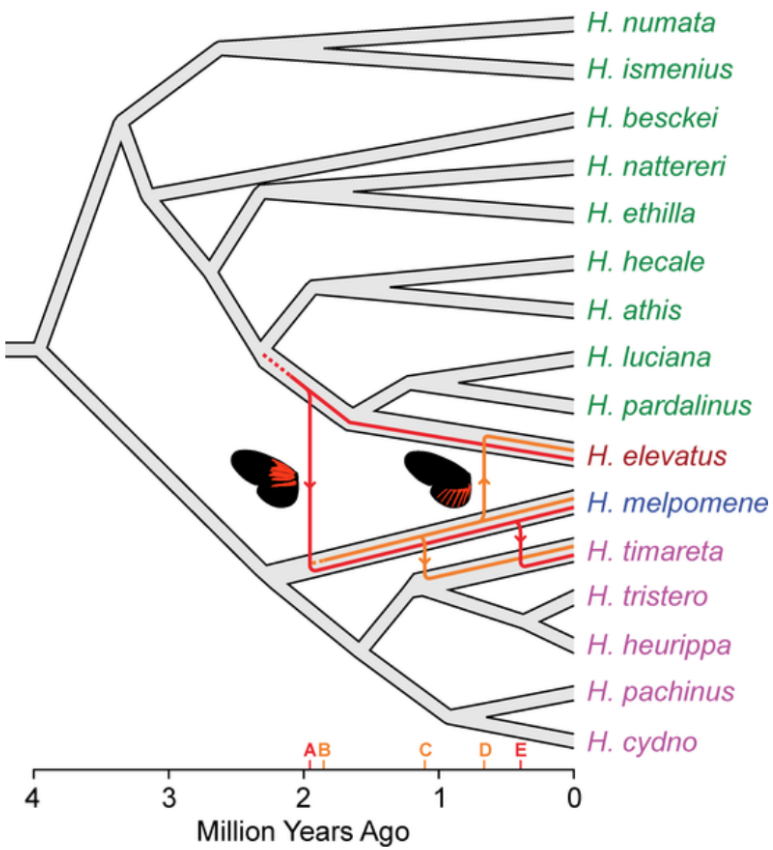


Figure: Wallbank et al, PLOS Biology, 2016

Not a phylogenetic network in the usual sense!

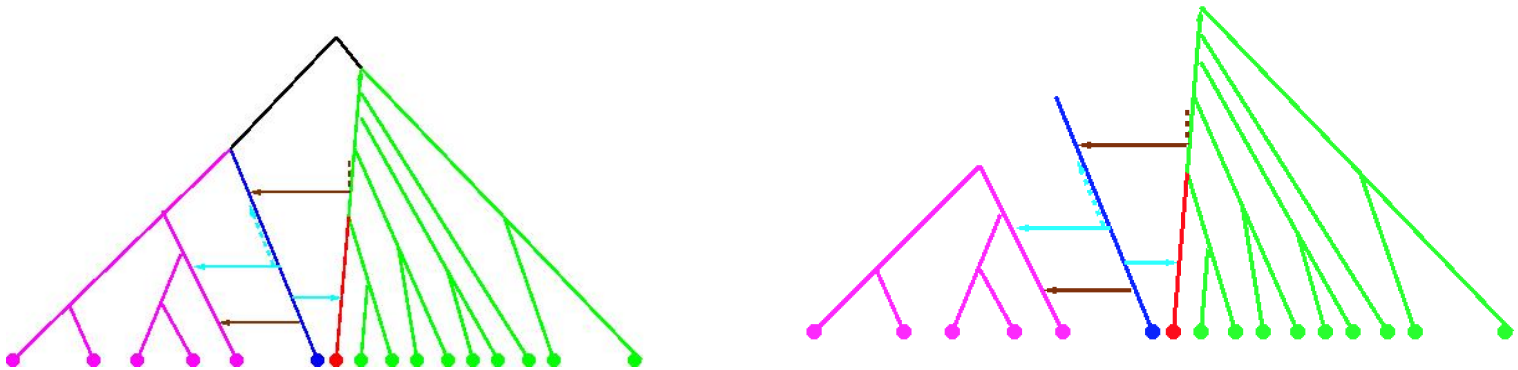


Figure: arrows: introgression events; colors: lineages

Multiply rooted networks

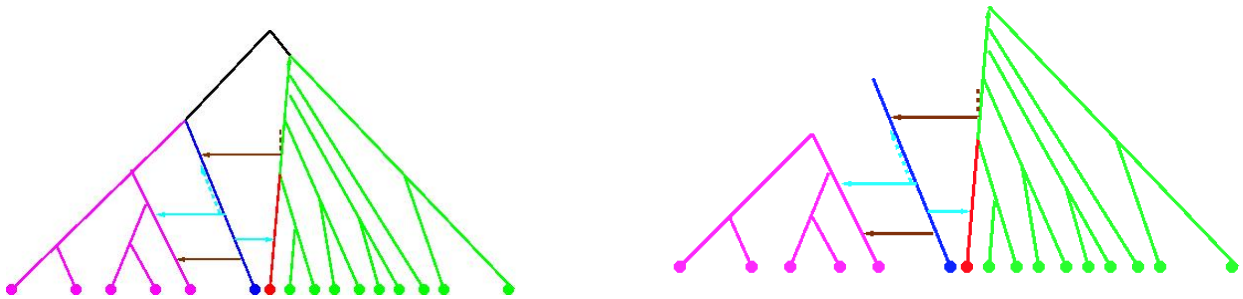


Figure: Left: 1-rooted network; Right: 3-rooted network

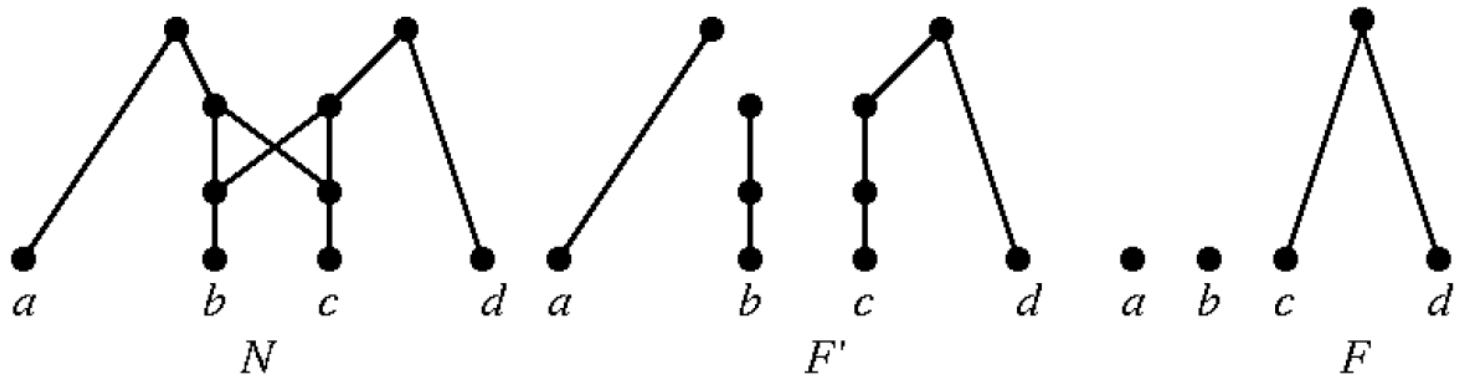
A *m-(rooted) network (on a set X of taxa)* is either an isolated vertex or a connected DAG with leaf set X and $m \geq 1$ roots, such that

- a root has indegree 0 and outdegree 2;
- an internal vertex has either indegree 2 and outdegree 1 or indegree 1 and outdegree 2;
- every leaf has indegree 1.

Overlaid Species Forests (OSF)

OSF-Builder (Scholz, Popescu, Taylor, Moulton, Huber, Systematic Biology, 2019) for the Heliconius butterfly data set from Wallbank et al. (2016) and Kozak et al. (2015),

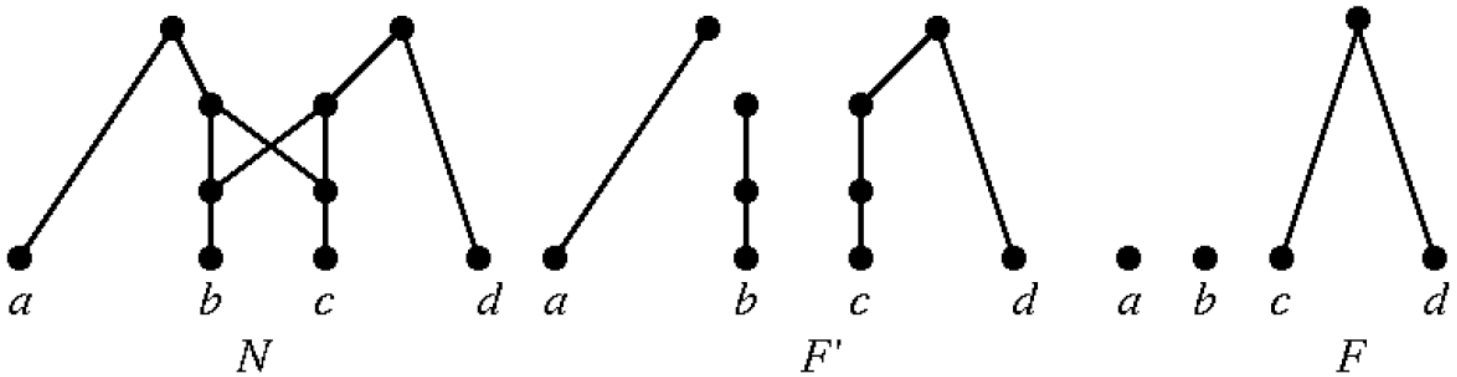
From tree-based to forest-based



- A m-network $N = (V, A)$ is called *forest-based* if there exists a subset $A' \subseteq A$ such that $F' = (V, A')$ is a forest with the same leaf set as N and every arc in $A - A'$ has end vertices contained in different trees of F'.
- The *base forest* F is the forest obtained from F' by repeatedly suppressing vertices of indegree and outdegree 1 and also roots of outdegree 1.

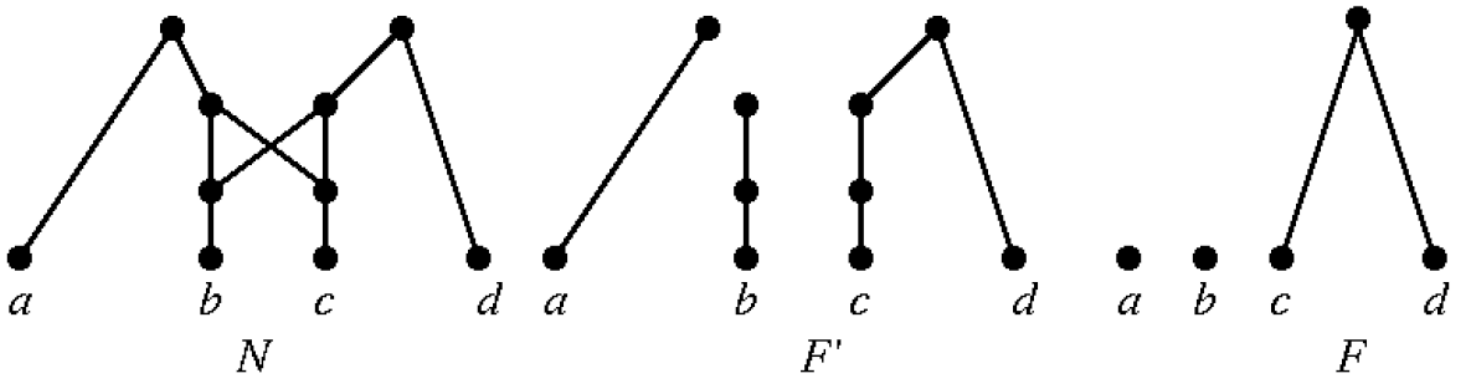
Some remarks, ..

- There are forest-based networks that are not OSFs.



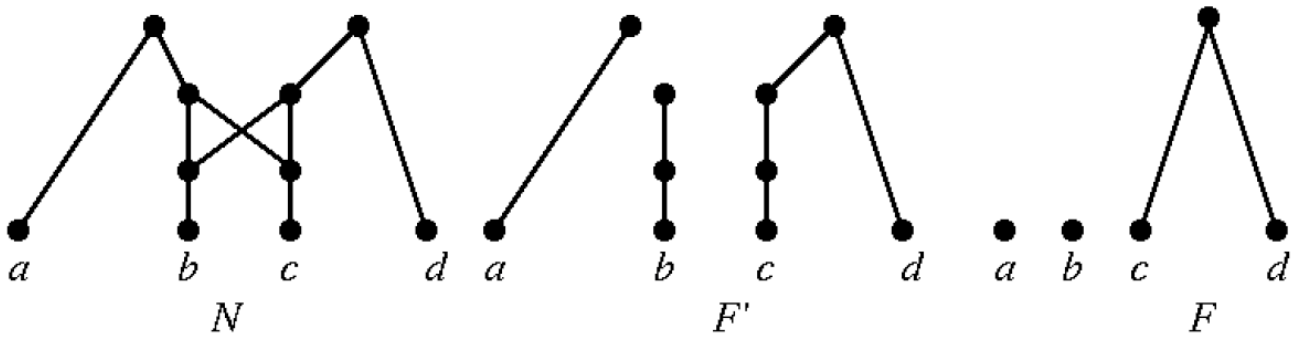
Some remarks, ..

- There are forest-based networks that are not OSFs.

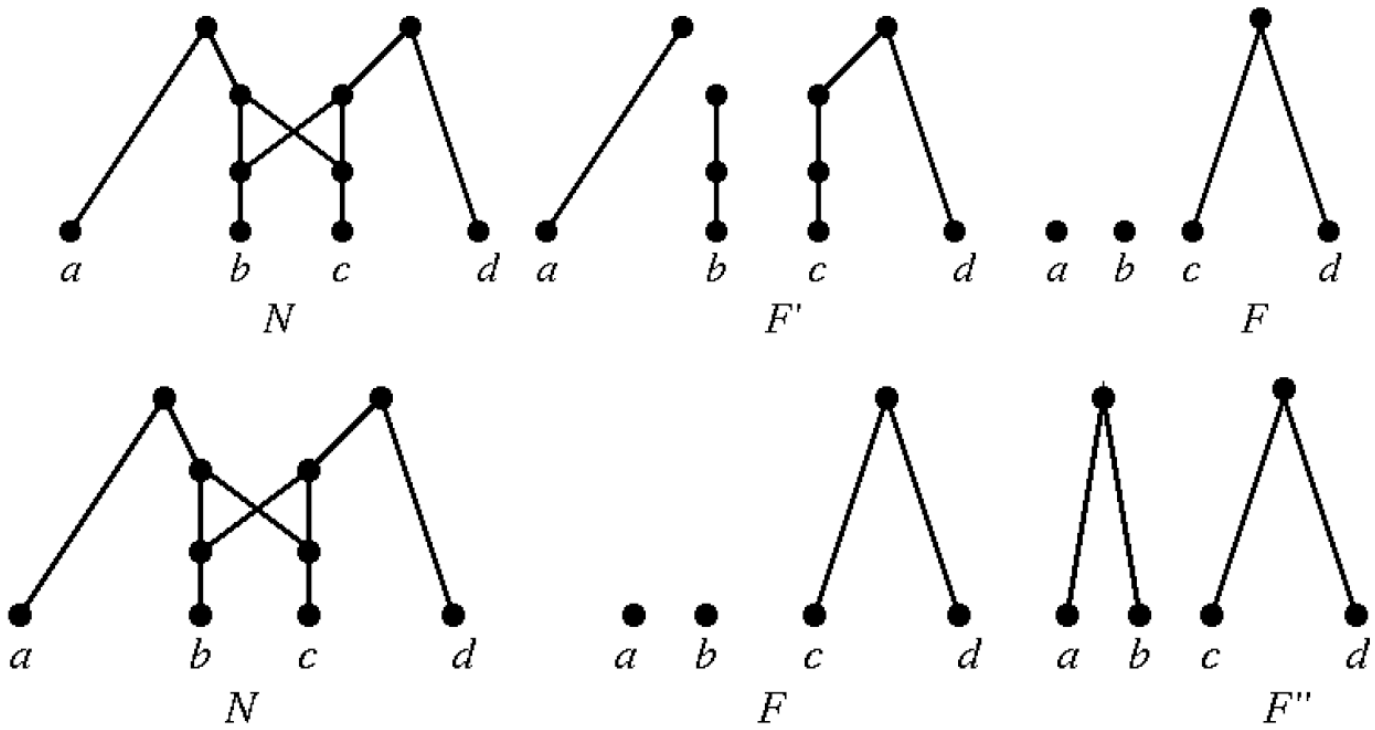


- Forest-based networks might also lend themselves as a tool for modelling how organisms in different environments have swapped genetic material.

Number of roots need not be equal to the size of a base forest

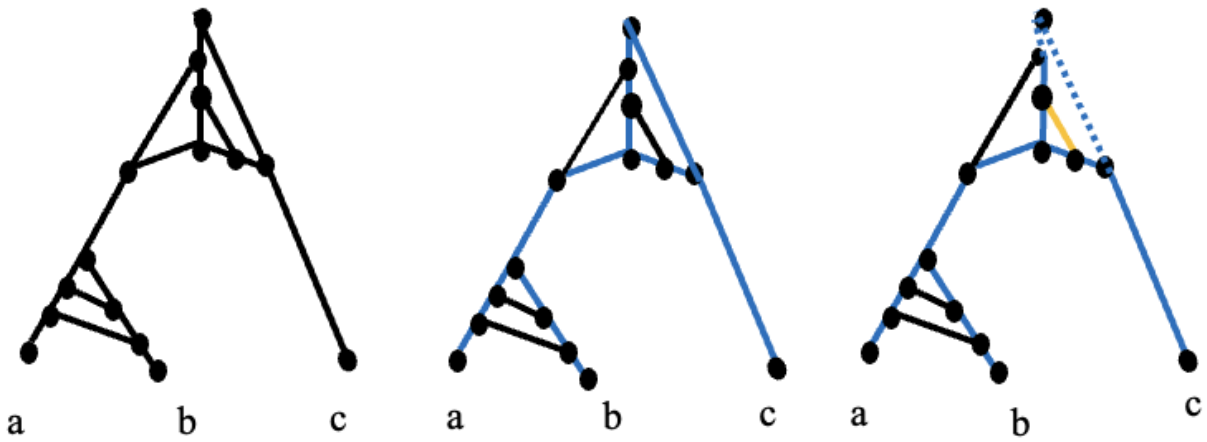


Number of roots need not be equal to the size of a base forest



A forest-based m-network, $m \geq 2$, that contains a base forest of size m is called a *proper* forest-based m-network.

Tree-based minus root is not forest-based!



In fact, we have ...

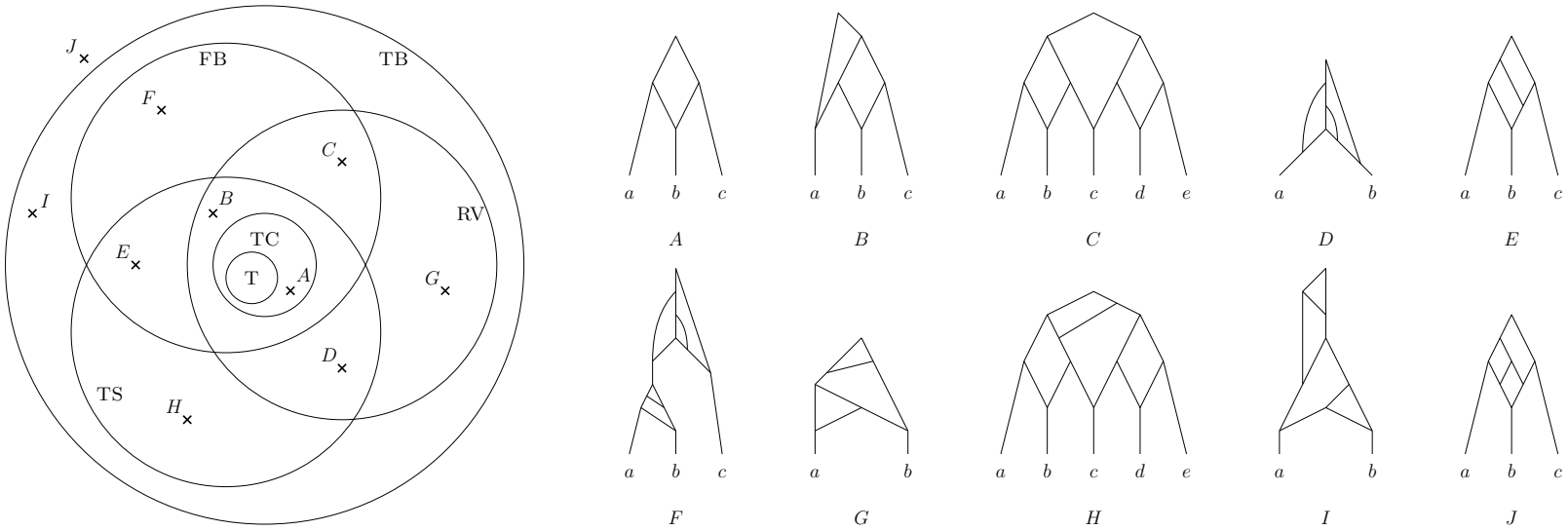


Figure: T: phylogenetic trees; TC: tree-child; TS: tree-sibling; RV: reticulation visible, TB: tree-based; FB: forest-based. – see Steel M., CBMS89, SIAM, 2016 for definitions.

What can we learn from ...

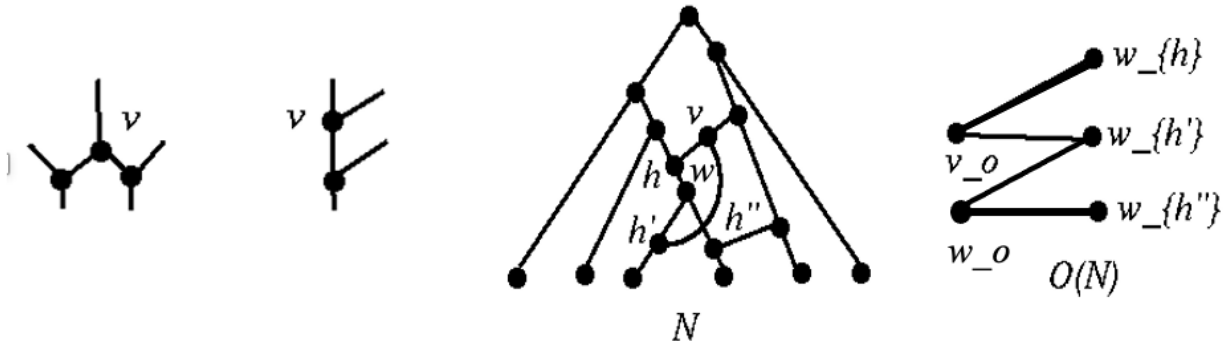


Figure: Left: v is an omnion (Jetten and van Iersel, 2018); Right: $O(N)$.

$O(N) = (U \cup H, E)$ is the bipartite graph given by

- for every omnion v of N there exists a vertex v_o in U.
- for every *hybrid* (i.e. indegree 2 vertex) w of N there exists a vertex w_h in H.
- $\{v_o, w_h\} \in E$ if there exists an omnion v and a hybrid w in N such that (v, w) is an arc in N.

In the case of a phylogenetic network, ...

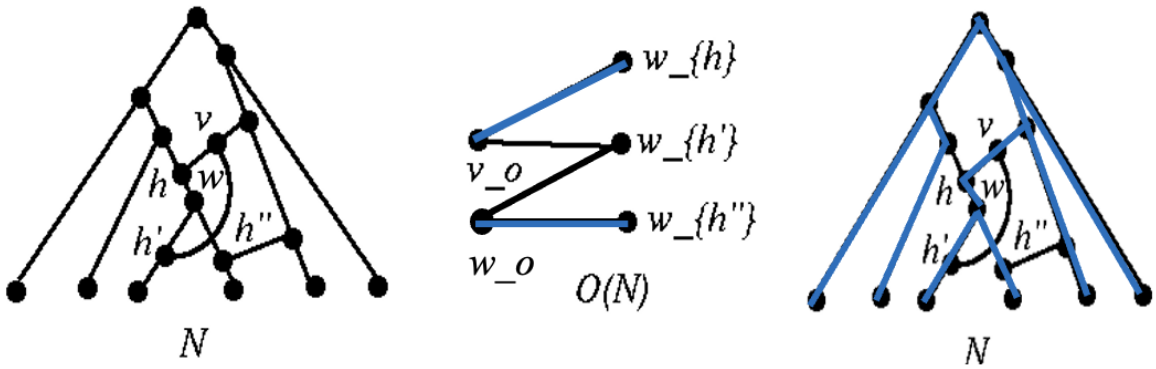
Theorem (Jetten and van Iersel, 2018)

A phylogenetic network is tree-based if and only if $O(N)$ contains a matching M and $|M| = |U|$.

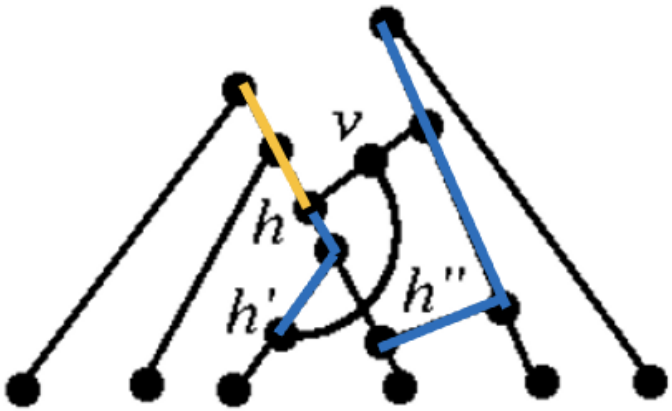
In the case of a phylogenetic network, ...

Theorem (Jetten and van Iersel, 2018)

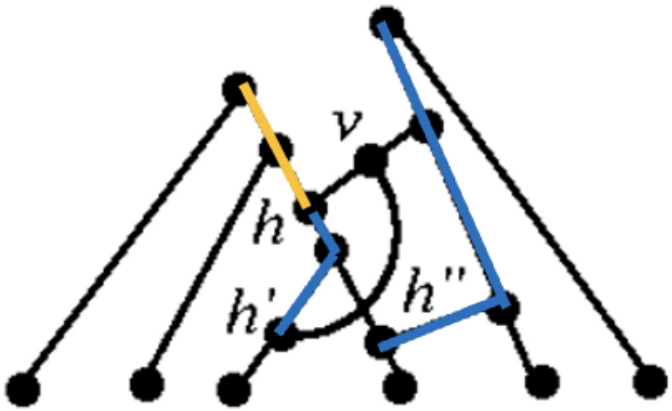
A phylogenetic network is tree-based if and only if $O(N)$ contains a matching M and $|M| = |U|$.



Clearly , ...

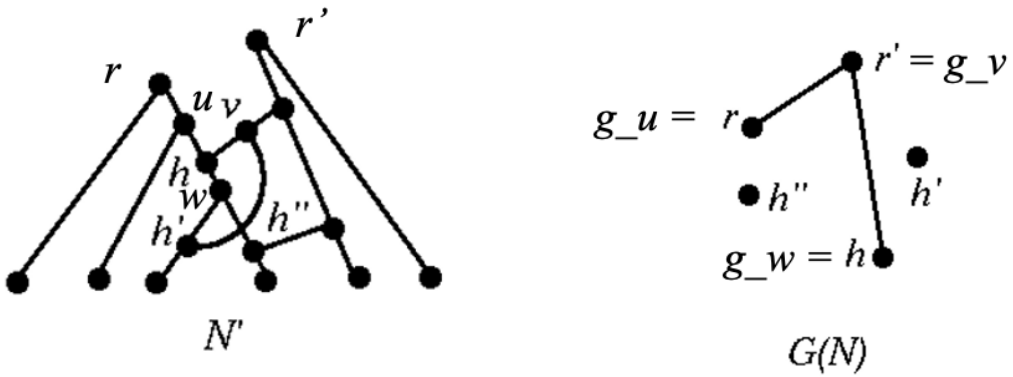


Clearly , ...



For v a vertex, g_v is the unique ancestor of v such that no root or hybrid is contained in the directed path from g_v to v.

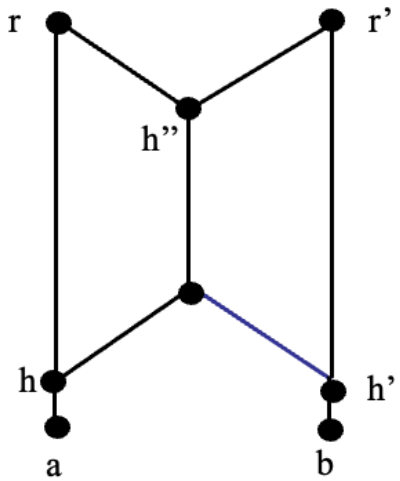
So it suffices to concentrate on the roots and hybrids ...



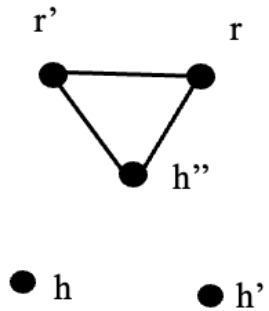
$G(N) = (V, E)$ is the graph (potentially with loops) given by

- V is the set of roots and hybrids of N.
- $\{u, v\} \in E$ if there exists a hybrid h with parents u' and v' such that $u = g_{u'}$ and $v = g_{v'}$.

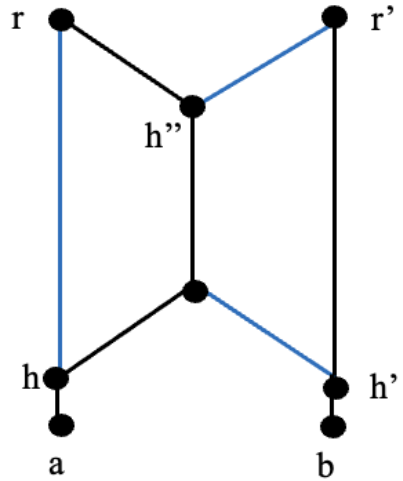
But $G(N)$ is not enough, ...



N

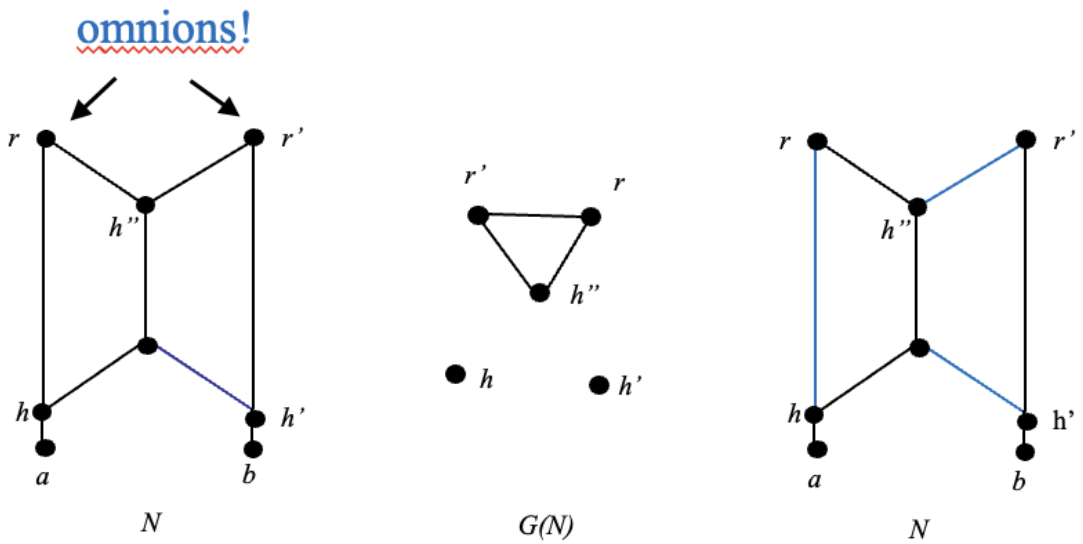


G(N)

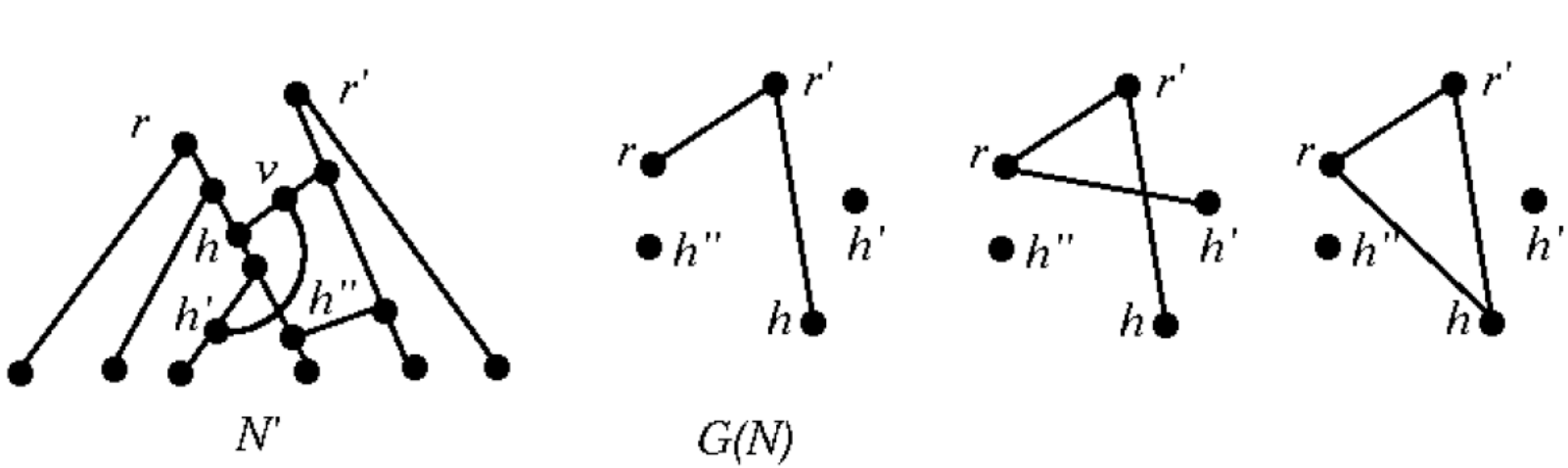


N

Omnions again,



No additional arc within ...



A supergraph $G'(N)$ is an *omni-extension* of $G(N)$ if

- the vertex set of $G'(N)$ is $G(N)$.
- for every omnion v of N there exists a child h of v such that $\{g_u, h\}$ is an edge of $G'(N)$ for u the second parent of h.

Theorem (Huber, Moulton, Scholz, (BMB, 2022))

A 2-rooted network is proper forest-based if and only if $G(N)$ has a bipartite omni-extension.

More generally, ...

Theorem (Huber, Moulton, Scholz, (BMB, 2022))

Let N be a m-rooted network on X, some $m \geq 2$, and let $\{s_1, \ldots, s_m\}$ be a set of m colors. Then N is proper forest based if and only if there exists an omni-extension $\Gamma'(N)$ of $\Gamma(N)$ and a proper vertex coloring $\sigma : R(N)^2 \cup H(N)^3 \to \{s_1, \ldots, s_m\}$ of $\Gamma'(N)$ satisfying:

- *The restriction of σ to $R(N)$ is a bijection.*
- *For all $u \in R(N)$ and all $v \in H(N)$ such that $\sigma(u) = \sigma(v)$ there must exist a directed path P in N from u to v such that $\sigma(w) = \sigma(u)$ holds for all vertices $w \in H(N)$ that lie on P.*

[2] set of roots of N

[3] set of hybrids of N

When is a given network proper forest based?

The good news ...

Theorem (Huber, van Iersel, Moulton, Scholz (IPL 2024))

Given a tree-child[4] *2-rooted network* N *with* $n \geq 2$ *leaves, it can be decided in* $O(n \times |H(N)|)$ *time if* N *is proper forest-based.*

[4] A phylogenetic network is called *tree-child* if every vertex that is not a leaf has a child that is not a hybrid.

The bad news ...

Theorem (Huber, van Iersel, Moulton, Scholz (IPL 2024))

Given a m-network N it is NP-complete to decide if N is forest-based even when N is restricted to be tree-child, for each fixed $m \geq 3$.

The bad news ...

Theorem (Huber, van Iersel, Moulton, Scholz (IPL 2024))

Given a m-network N it is NP-complete to decide if N is forest-based even when N is restricted to be tree-child, for each fixed $m \geq 3$.

Proof: Reduction from GRAPH m-COLORABILITY.

Relaxing the degree requirement ...

Theorem (Huber, van Iersel, Moulton, Scholz (IPL 2024))

Given a m-network N

- *it is NP-complete to decide if N is forest-based even when N is restricted to be tree-child, with maximum outdegree 2 and maximum indegree at most 3, for each fixed $m \geq 2$.*
- *an FPT algorithm can be used to decide if N is forest-based with parameters Δ[5], $|H(N)|$, and m, which is linear in $|V(N)|$, assuming the maximum indegree is 2.*

[5] maximum outdegree

Relaxing the degree requirement ...

Theorem (Huber, van Iersel, Moulton, Scholz (IPL 2024))

Given a m-network N

- *it is NP-complete to decide if N is forest-based even when N is restricted to be tree-child, with maximum outdegree 2 and maximum indegree at most 3, for each fixed $m \geq 2$.*
- *an FPT algorithm can be used to decide if N is forest-based with parameters Δ[5], $|H(N)|$, and m, which is linear in $|V(N)|$, assuming the maximum indegree is 2.*

Proof: Reduction from SET SPLITTING ($m = 2$) and GRAPH m-COLORABILITY ($m \geq 3$).

[5] maximum outdegree

Open Problems

- Pedigrees are closely related to networks with multiple roots. Given that they can be split into a female and a male line and so induce a forest, what is the relationship between forest-based networks and pedigrees?
- What can we say about forest-based networks if we also allow additional arcs within a tree in the forest?

Acknowledgements