# BIOINFORMATICS AND STATISTICAL GENETICS

GABRIEL VALIENTE

ALGORITHMS, BIOINFORMATICS, COMPLEXITY AND FORMAL METHODS RESEARCH GROUP,
TECHNICAL UNIVERSITY OF CATALONIA

2023–2024

Phylogenetic reconstruction III

- Phylogenetic networks
- Galled trees
- Tree-child networks
- Tree-sibling networks
- Time consistency of phylogenetic networks
- A hierarchy of phylogenetic networks

- The evolutionary relationships among a group of organisms are often illustrated by means of a phylogenetic tree, whose nodes represent taxonomic units (which can be species or taxa, higher or nested taxa, populations, individuals, or genes) and whose branches define the evolutionary relationships among the taxonomic units (where children nodes descend from their parents by mutation).
- However, there are evolutionary processes acting at the population level, such as recombination between genes, hybridization between lineages, and lateral gene transfer, that lead to reticulate relationships that can no longer be modeled by a phylogenetic tree.
- The modeling and explicit representation of reticulate evolutionary events turn phylogenetic trees into a particular form of directed acyclic graphs called phylogenetic networks.

- A phylogenetic network is a directed acyclic graph whose terminal nodes are labeled by taxa names and whose internal nodes are either tree nodes (if they have only one parent) or hybrid nodes (if they have two or more parents).
- As in the case of phylogenetic trees, a phylogenetic network is fully resolved if every internal tree node in the network has two children and every hybrid node has two parents and a single child, which is often a tree node.

## Example

The alcohol dehydrogenase enzyme is one of the most abundant proteins in *Drosophila melanogaster*, and it is encoded by a single gene.

The alcohol dehydrogenase gene was studied on a sample of eleven species from five natural populations of Drosophila melanogaster, and the sampled sequences contain 44 polymorphic (segregating) sites.

```
CCGCAATAATGGCGCTACTCTCACAATAACCCACTAGACAGCCT
CCCCAATATGGGCGCTACTTTCACAATAACCCACTAGACAGCCT
CCGCAATATGGGCGCTACCCCCCGGAATCTCCACTAAACAGTCA
CCGCAATATGGGCGCTGTCCCCCGGAATCTCCACTAAACTACCT
CCGAGATAAGTCCGAGGTCCCCCGGAATCTCCACTAGCCAGCCT
CCCCAATATGGGCGCGACCCCCCGGAATCTCTATTCACCAGCTT
CCCCAATATGGGCGCGACCCCCCGGAATCTGTCTCCGCCAGCCT
TGCAGATAAGTCGGCGACCCCCCGGAATCTGTCTCCGCGAGCCT
TGCAGATAAGTCGGCGACCCCCCGGAATCTGTCTCCGCGAGCCT
TGCAGATAAGTCGGCGACCCCCCGGAATCTGTCTCCGCGAGCCT
TGCAGGGGAGGGCTCGACCCCACGGGATCTGTCTCCGCCAGCCT
```

## Example

Under the infinite sites assumption, by which mutations are rare enough to discard the possibility of more than one mutation to occur at the same site in a sample of sequences, no site of a sample can contain more than two different nucleotides.

The most frequent nucleotide along a site is often taken as the base, with the least frequent nucleotide being taken as the mutant.

The base and mutant nucleotides for each site of the previous sequences are as follows.

```
CCCCAATAAGGGCGCGACCCCCCGGAATCTCTATTCGCCAGCCT
TGGAGGGGTTTCGTATGTTTTAACAGTAACGCCCCAAAGTATTA
```

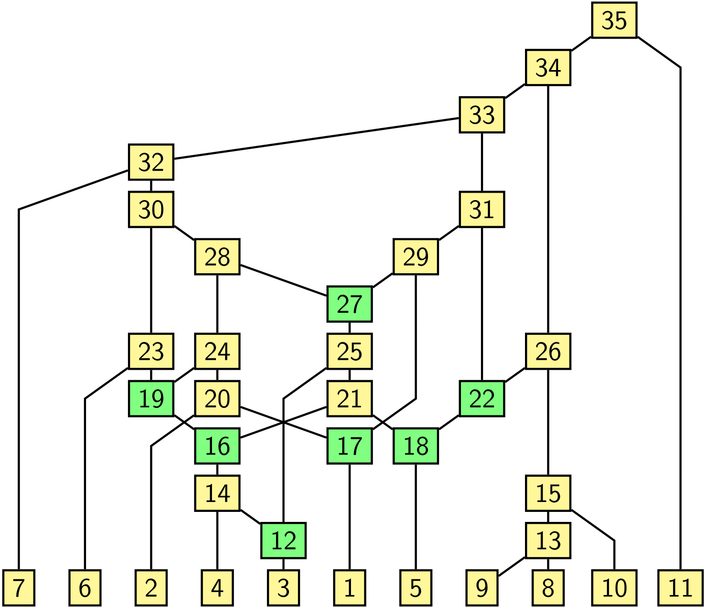
## Example

This allows for a binary representation of a sample of sequences, where the base nucleotide in each site is encoded as 0 and the mutant nucleotide is encoded as 1.

```
00100000010000010010101110111101010101000000
00000000100000010011101110111101010101000000
00100000100000010000000000000001010111000101
00100000100000011100000000000001010111011000
00111000001100101100000000000001010100000000
00000000100000000000000000000000000010000010
00000000100000000000000000000010101000000000
11011000001110000000000000000010101000100000
11011000001110000000000000000010101000100000
11011000001110000000000000000010101000100000
11011111000001000000010001000010101000000000
```

- The evolutionary relationships among these eleven cloned genes cannot be modeled by a phylogenetic tree.
- In fact, any phylogenetic network with hybrid nodes representing recombination events explaining these evolutionary relationships must include at least 7 recombination events.
- One such possible explanation is the following fully resolved phylogenetic network, which has 11 leaves and exactly 7 hybrid nodes.
- The edges are directed from top to bottom, and some of them represent mutation events.
- The edges going into a hybrid node, on the other hand, represent recombination events.

- R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–164, 1985

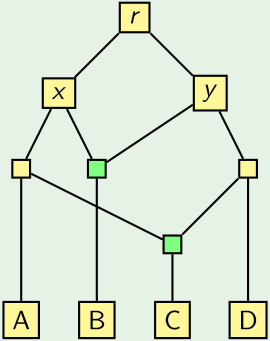

35
34
33
32
30
31
28
29
27
23
24
25
26
19
20
21
22
16
17
18
14
15
12
13
7
6
2
4
3
1
5
9
8
10
11

- A phylogenetic network explaining the evolutionary relationships among a given set of taxonomic units can be very large indeed, as there is no upper bound on the number of hybrid nodes.
- There is, however, a lower bound on the number of recombination events needed to explain the evolutionary relationships among a set of organisms given their DNA or RNA sequences, and this lower bound is 7 for the phylogenetic network of the previous example.
- While most phylogenetic reconstruction algorithms attempt to achieve the lower bound on the number of hybridization, recombination, or lateral gene transfer events, the resulting phylogenetic networks often lack topological properties that are essential to their further analysis.
- This is especially relevant to the comparative analysis of phylogenetic networks, and most distances and alignment algorithms impose some condition on phylogenetic network topology.

- Recombination or hybridization cycles be pairwise disjoint
- Internal nodes have some child that is a tree node
- Hybrid nodes have some sibling that is a tree node

- The path between any two given terminal nodes of a rooted phylogenetic tree traverses the LCA of the nodes but, in a phylogenetic network, the path between any two terminal nodes need not be unique, because of the existence of hybrid nodes, which make it necessary to distinguish between strict and non-strict descendants of a node.
- For a strict descendant of a node, every path from the root to the descendant must contain the node, while for a non-strict descendant, there is at least one path containing the node and at least one path not containing it.
- A common semi-strict ancestor (CSA) of two nodes in a phylogenetic network is a common ancestor of the nodes which is also a strict ancestor of at least one of them.
- The path between any two given terminal nodes of a phylogenetic network is the shortest path that traverses the lowest common semi-strict ancestor (LCSA) of the nodes in the network, which always exists.

## Example



- The LCSA of A and B is $x$, because it is a strict ancestor of A and a (non-strict) ancestor of B, and none of its descendants is an ancestor of both.
- The LCSA of B and C is $r$, because it is a strict ancestor of both, and none of its descendants is also a strict ancestor of any of them.

- All the CSA of two terminal nodes $v$ and $w$ in a phylogenetic network $N$ are collected in an (initially empty) queue $Q$ of nodes.

```
function CSA($N, v, w$)
    for all nodes $u$ of $N$ do
        if $v$ and $w$ are reachable from $u$ in $N$ then
            if $strict\_ancestor(N, u, v)$ or $strict\_ancestor(N, u, w)$ then
                $enqueue(Q, u)$
    return $Q$
```

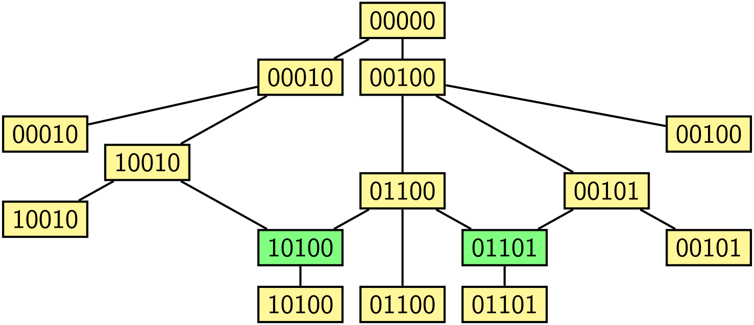
- The test for a node $u$ being a strict ancestor of a terminal node $v$ involves removing $u$ from a copy $N'$ of the phylogenetic network $N$.

```
**function** strict_ancestor($N, u, v$)
    $N' \leftarrow N$
    remove node $u$ from $N'$
    $r \leftarrow root(N')$
    **if** $v$ is reachable from $r$ in $N'$ **then**
        **return** *false*
    **else**
        **return** *true*
```

- The LCSA of two terminal nodes is just the CSA of the two nodes that has the smallest height in the DAG representation of the phylogenetic network.

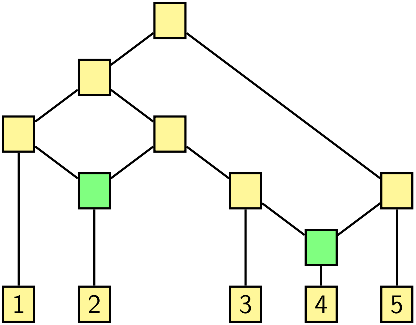
```
**function** LCSA($N, v, w$)
    $Q \leftarrow CSA(N, v, w)$
    $u \leftarrow dequeue(Q)$
    **while** $Q$ is not empty **do**
        $x \leftarrow dequeue(Q)$
        **if** $height(x) < height(u)$ **then**
            $u \leftarrow x$
    **return** $u$
```

- A recombination or hybridization cycle consists of two paths from some common ancestor to the two parents of a hybrid node.
- A phylogenetic network is called a galled-tree if all recombination or hybridization cycles are pairwise disjoint.



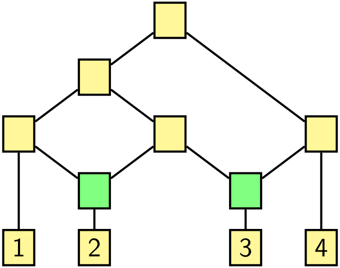
- D. Gusfield. *ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*. The MIT Press, Cambridge, MA, 2014

- A phylogenetic network is called tree-child if every internal node has at least one child that is a tree node.
- The biological meaning of the tree-child condition is that every non-extant species has some descendant through mutation.



- G. Cardona, F. Rosselló, and G. Valiente. Comparison of tree-child phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):552–569, 2009

- A phylogenetic network is called tree-sibling if every hybrid node has at least one sibling that is a tree node.
- The biological meaning of the tree-sibling condition is that in each of the recombination or hybridization processes, at least one of the species involved in them also has some descendant through mutation.

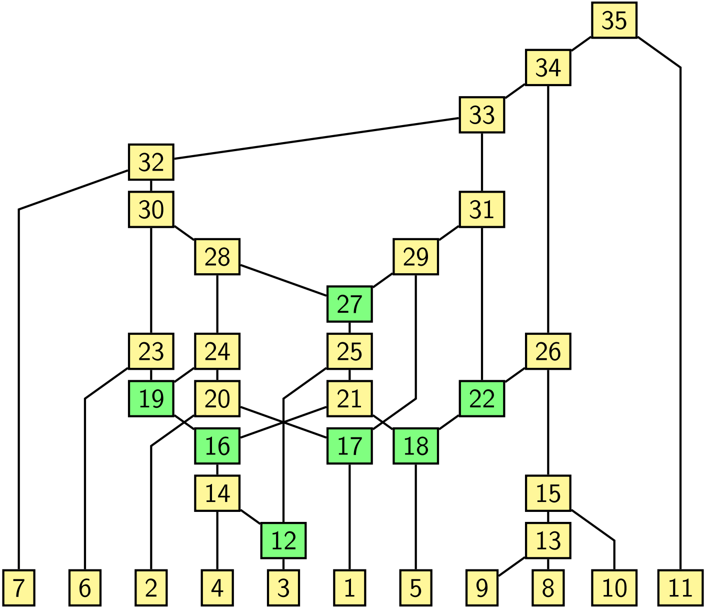


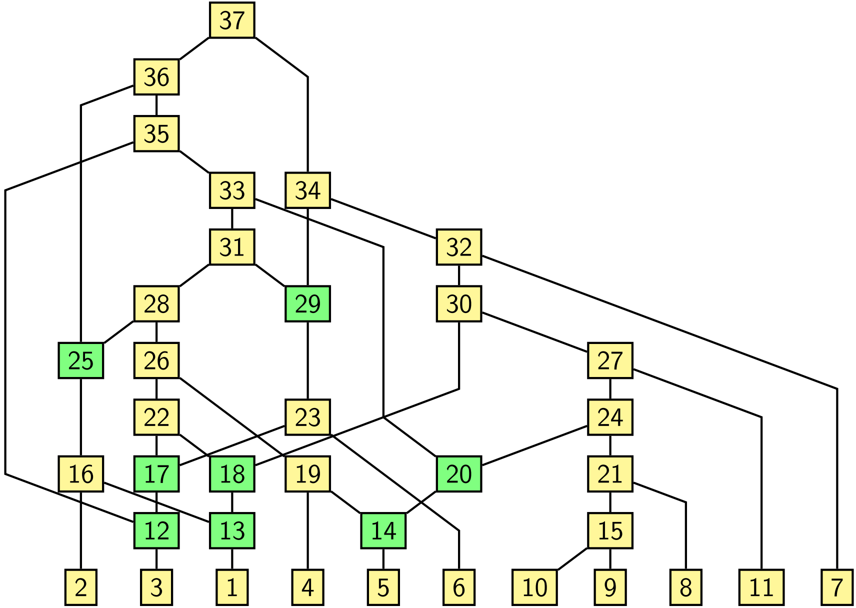
- G. Cardona, F. Rosselló, and G. Valiente. A distance metric for a class of tree-sibling phylogenetic networks. *Bioinformatics*, 24(13):1481–1488, 2008

- A temporal representation of a phylogenetic network is an assignment of times to the nodes of the network that strictly increases on tree edges (those edges whose head is a tree node) and remains the same on hybrid edges (whose head is a hybrid node).
- A phylogenetic network is time-consistent if it has a temporal representation.
- The biological meaning of a temporal assignment is the time when certain species exist or when certain hybridization processes occur, because for these processes to take place, the species involved must coexist in time.

- G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente. Path lengths in tree-child time consistent hybridization networks. *Information Sciences*, 180(3):366–383, 2010

- A phylogenetic network with the least possible number of hybrid nodes does not necessarily exhibit any of the topological properties of being a galled-tree, tree-child, tree-sibling, or time-consistent.
- However, under the hypothesis of more recombination events, the evolutionary history of the sample from the previous example can be explained by a tree-sibling phylogenetic network.
- Another possible explanation of the evolutionary relationships among the eleven alcohol dehydrogenase genes from the previous example is the following fully resolved tree-sibling phylogenetic network, which has 8 hybrid nodes.

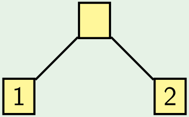

35
34
33
32
30
31
28
29
27
23
24
25
26
19
20
21
22
16
17
18
14
15
12
13
7
6
2
4
3
1
5
9
8
10
11


37
36
35
33
34
31
32
28
29
30
25
26
27
22
23
24
16
17
18
19
20
21
12
13
14
15
2
3
1
4
5
6
10
9
8
11
7

- A (phylogenetic) network is a rooted DAG with leaves bijectively labeled in a given set, such that no tree node has out-degree 1, and every hybrid node has out-degree 1, and its single child is a tree node.
- A node in a DAG is a tree node if it has in-degree at most 1, and it is a hybrid node if it has in-degree at least 2.
- A network is 2-hybrid if all hybrid nodes have in-degree 2, and it is hybrid-1 if all hybrid nodes have out-degree 1.
- A network is semi-binary if all hybrid nodes have in-degree 2 and out-degree 1.
- A network is binary if it is semi-binary and all internal tree nodes have out-degree 2.

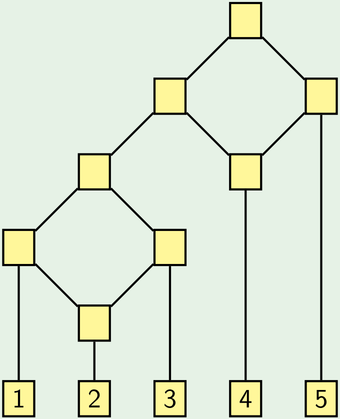
- A network is a tree if it has no hybrid nodes.

## Example (Tree but not path)



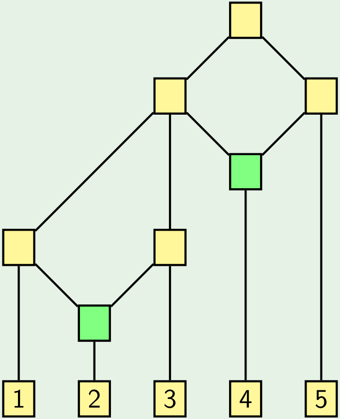
- A network is a galled tree if every pair of reticulation cycles have disjoint sets of nodes.

## Example (Galled tree but not tree)



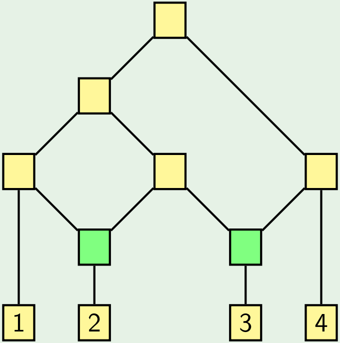
- A network is a weakly galled tree if every pair of reticulation cycles have disjoint sets of arcs.

## Example (Weakly galled tree but not galled tree)



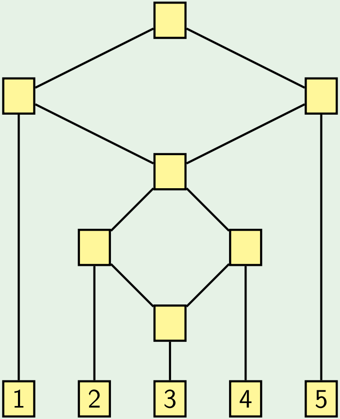
- A network is level-1 if no (maximal) biconnected subgraph contains more than one hybrid node.

## Example (Level-1 network but not weakly galled tree)



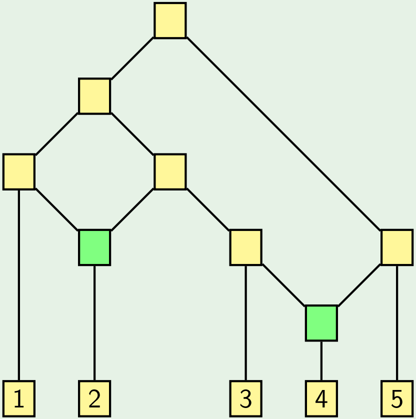
- A network is 1-nested if every pair of reticulation cycles for two different hybrid nodes has disjoint sets of intermediate nodes.

### Example (1-nested but not level-1)



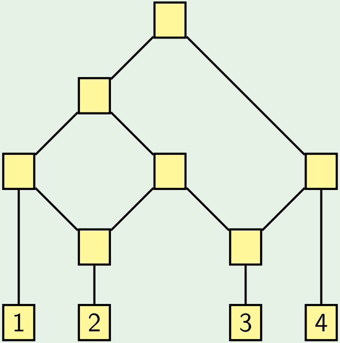
- A network is tree-child if every internal node has at least one child that is a tree node.

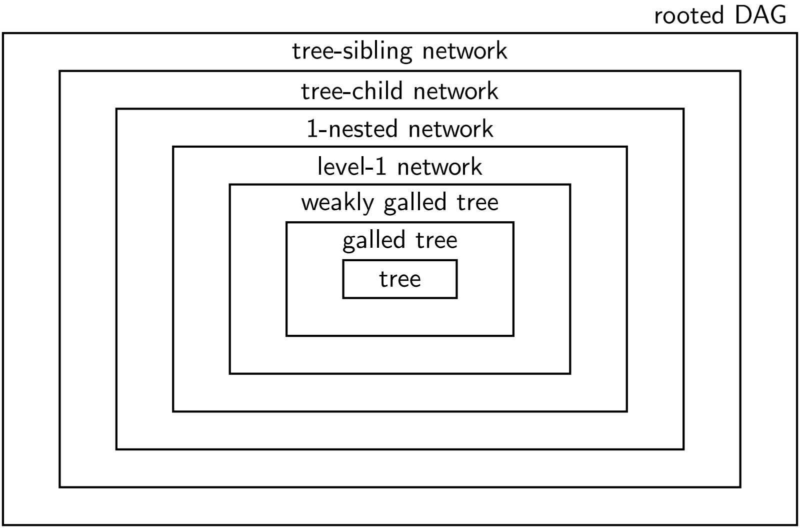
## Example (Tree-child but not 1-nested network)

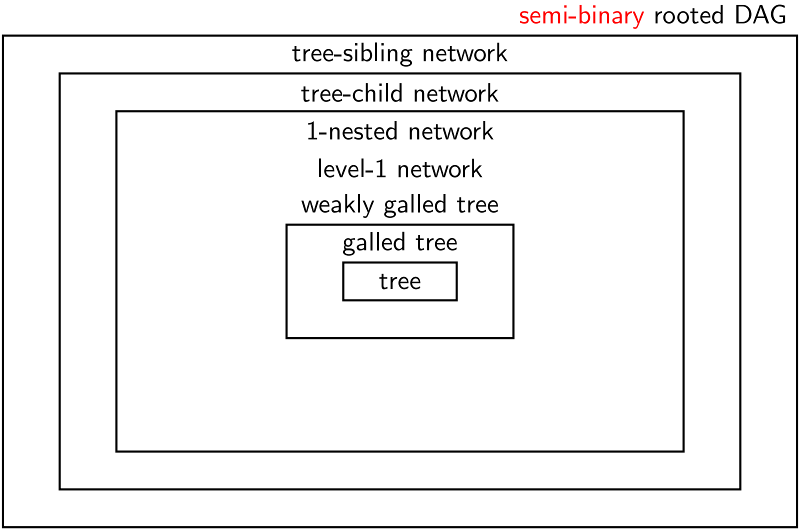


- A network is tree-sibling if every hybrid node has at least one sibling that is a tree node.

### Example (Tree-sibling but not tree-child network)




rooted DAG
tree-sibling network
tree-child network
1-nested network
level-1 network
weakly galled tree
galled tree
tree



semi-binary rooted DAG

tree-sibling network

tree-child network

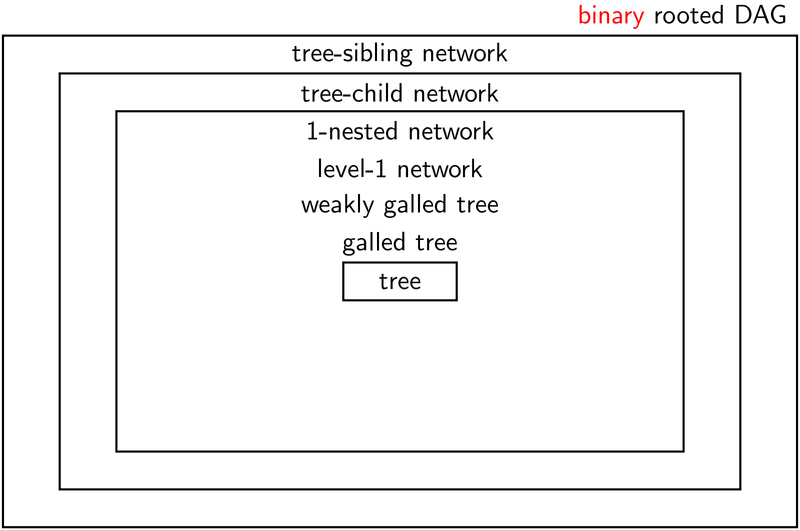
1-nested network

level-1 network

weakly galled tree

galled tree

tree



binary rooted DAG

tree-sibling network

tree-child network

1-nested network

level-1 network

weakly galled tree

galled tree

tree