Lille – 24/06/2009 – CPM'09

The Structure of Level-k Phylogenetic Networks

Philippe Gambette

in collaboration with

Vincent Berry, Christophe Paul

Outline

- **Phylogenetic networks**
- **Decomposition of level-*k* networks**
- **Construction of level-*k* generators**
- **Number of level-*k* generators**
- **Simulated level-*k* networks**

Outline

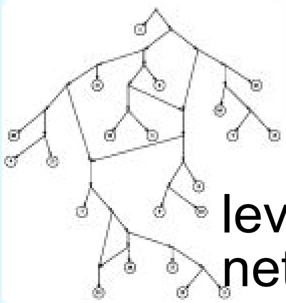
- **Phylogenetic networks**
- **Decomposition of level-*k* networks**
- **Construction of level-*k* generators**
- **Number of level-*k* generators**
- **Simulated level-k networks**

Phylogenetic networks

Phylogenetic network

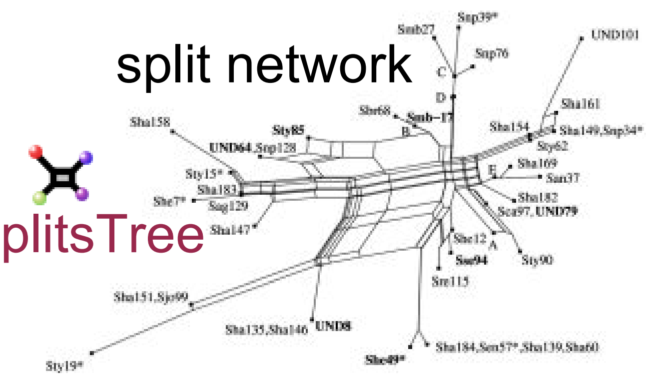
From Wikipedia, the free encyclopedia

A **phylogenetic network** is *any* graph used to visualize evolutionary relationships between species or organisms. It is employed when reticulate events such as hybridization, horizontal gene transfer, recombination, or gene duplication and loss are believed to be involved. Phylogenetic trees are a subset of phylogenetic networks.



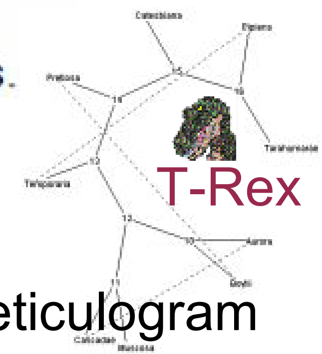
level-2 network

Level-2



split network

SplitsTree

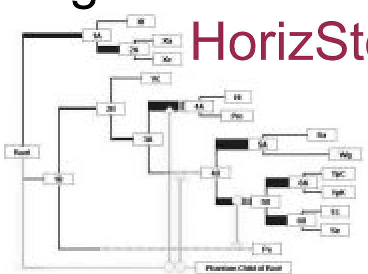


T-Rex

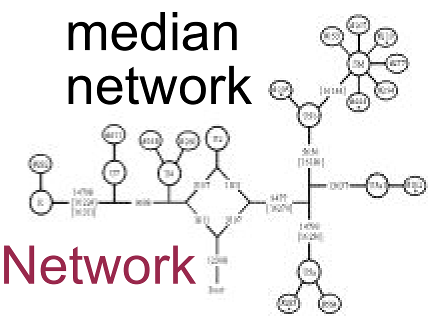
reticulogram

synthesis diagram

HorizStory

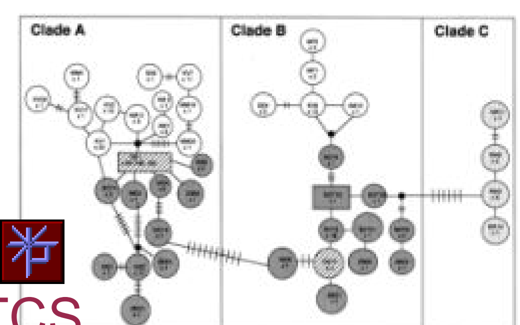


median network



Network

minimum spanning network



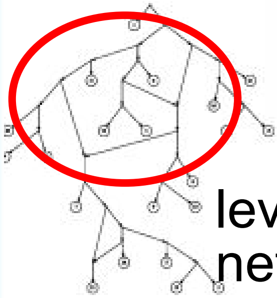
TCS

Phylogenetic networks

Phylogenetic network

From Wikipedia, the free encyclopedia

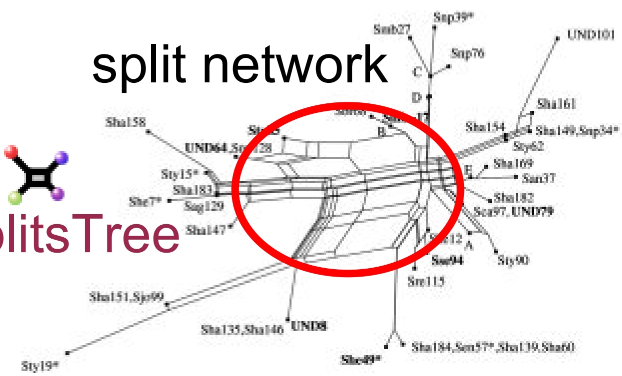
A **phylogenetic network** is *any* graph used to visualize evolutionary relationships between species or organisms. It is employed when reticulate events such as hybridization, horizontal gene transfer, recombination, or gene duplication and loss are believed to be involved. Phylogenetic trees are a subset of phylogenetic networks.



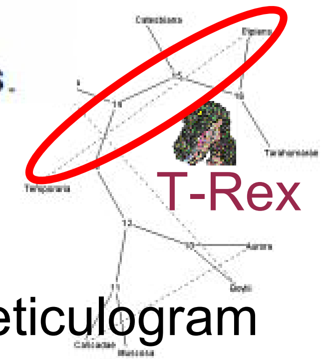
level-2 network

Level-2

split network



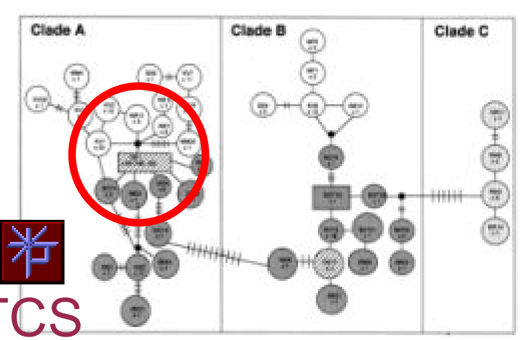
SplitsTree



T-Rex

reticulogram

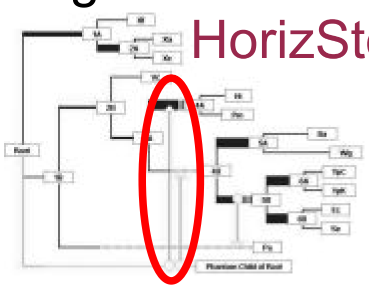
minimum spanning network



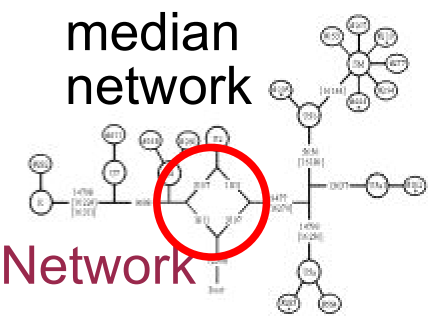
TCS

synthesis diagram

HorizStory



median network

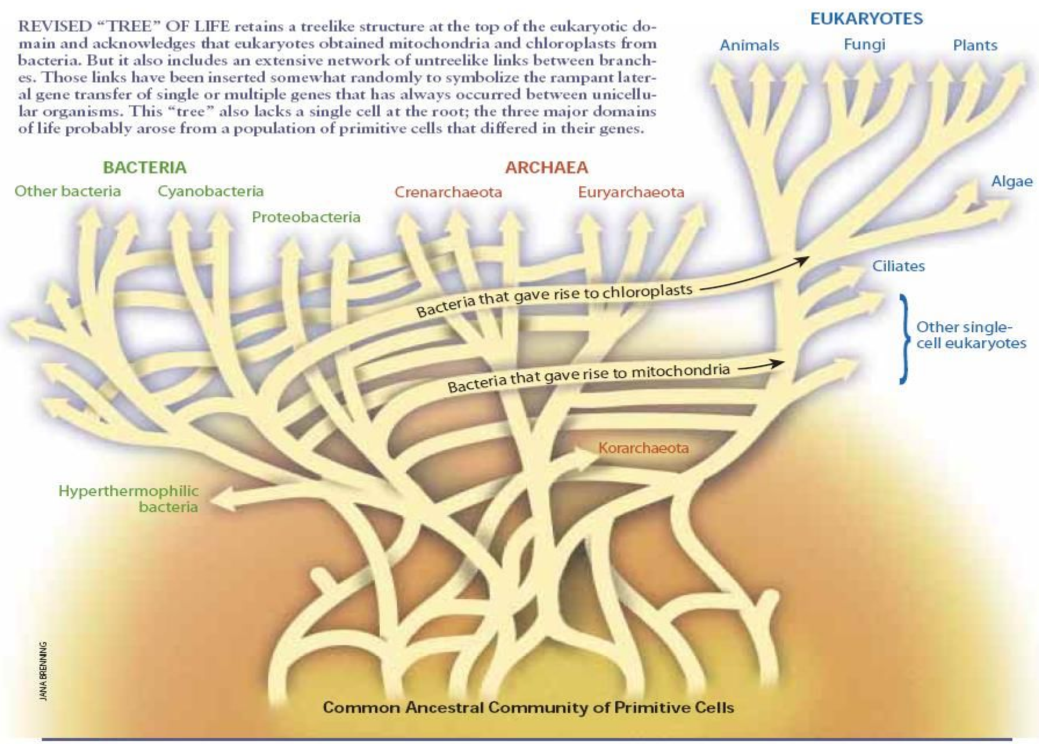


Network

Abstract or explicit networks

An **explicit phylogenetic network** is a phyogenetic network where all reticulations can be interpreted as precise biological events.

An **abstract network** reflects some phylogenetic signals rather than explicitly displaying biological reticulation events.



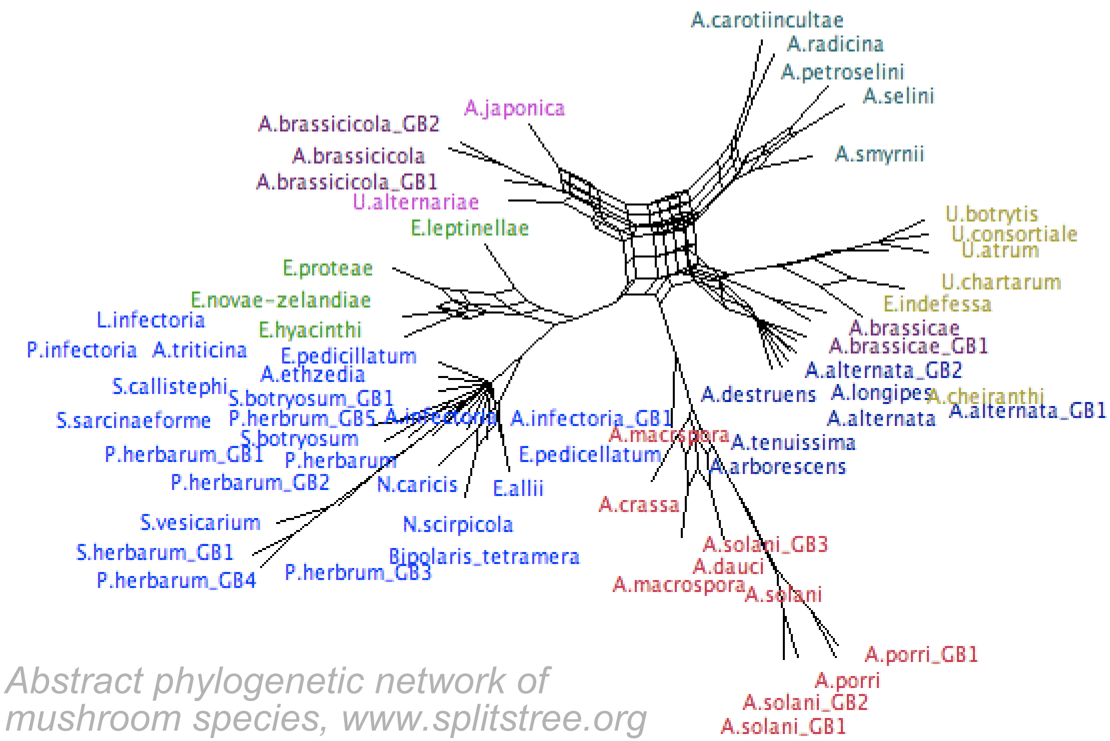
REVISED "TREE" OF LIFE retains a treelike structure at the top of the eukaryotic domain and acknowledges that eukaryotes obtained mitochondria and chloroplasts from bacteria. But it also includes an extensive network of untreelike links between branches. Those links have been inserted somewhat randomly to symbolize the rampant lateral gene transfer of single or multiple genes that has always occurred between unicellular organisms. This "tree" also lacks a single cell at the root; the three major domains of life probably arose from a population of primitive cells that differed in their genes.

Doolittle : Uprooting the Tree of Life, Scientific American (Feb..2000)

Abstract or explicit networks

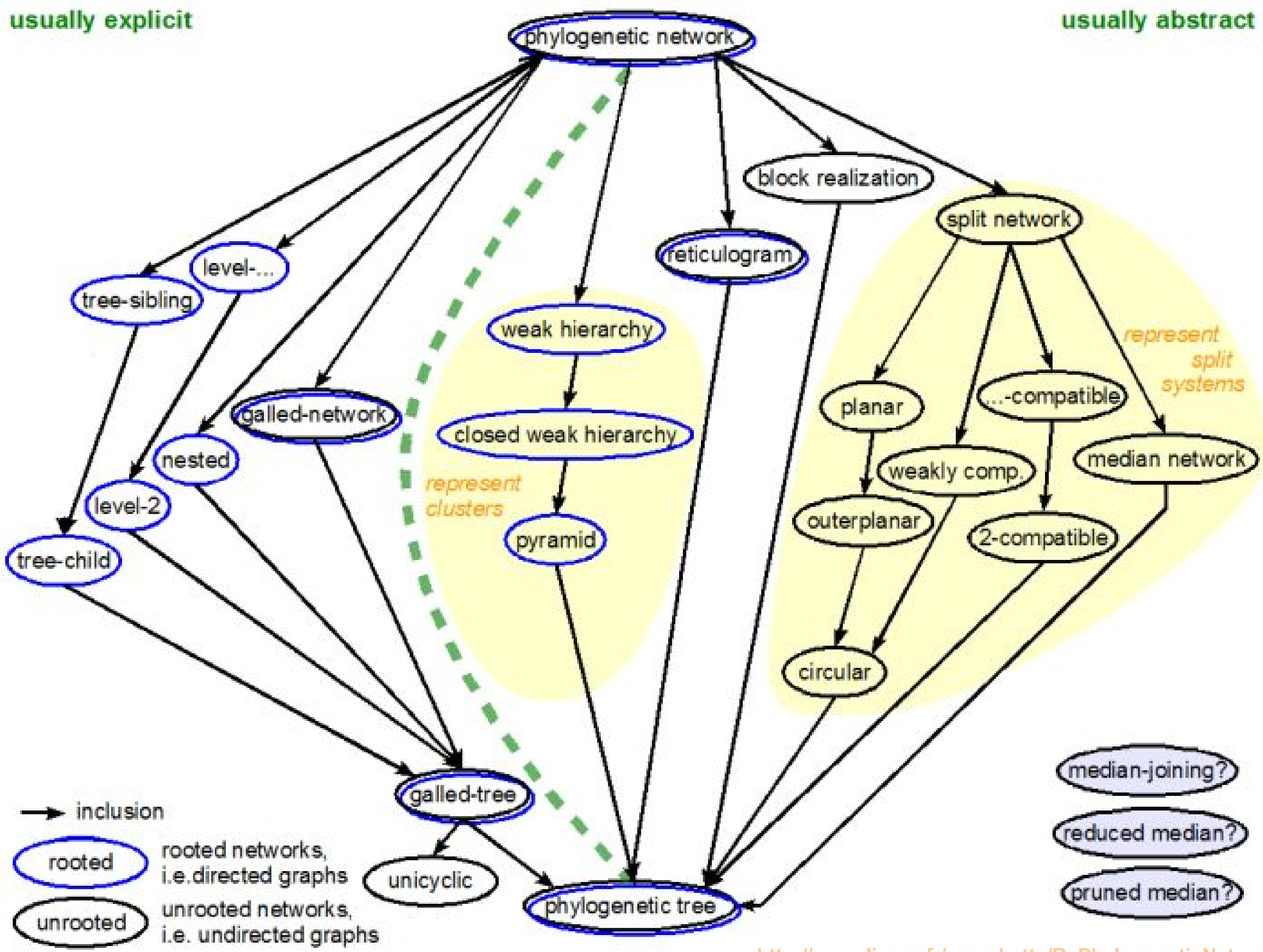
An **explicit phylogenetic network** is a phyogenetic network where all reticulations can be interpreted as precise biological events.

An **abstract network** reflects some phylogenetic signals rather than explicitly displaying biological reticulation events.



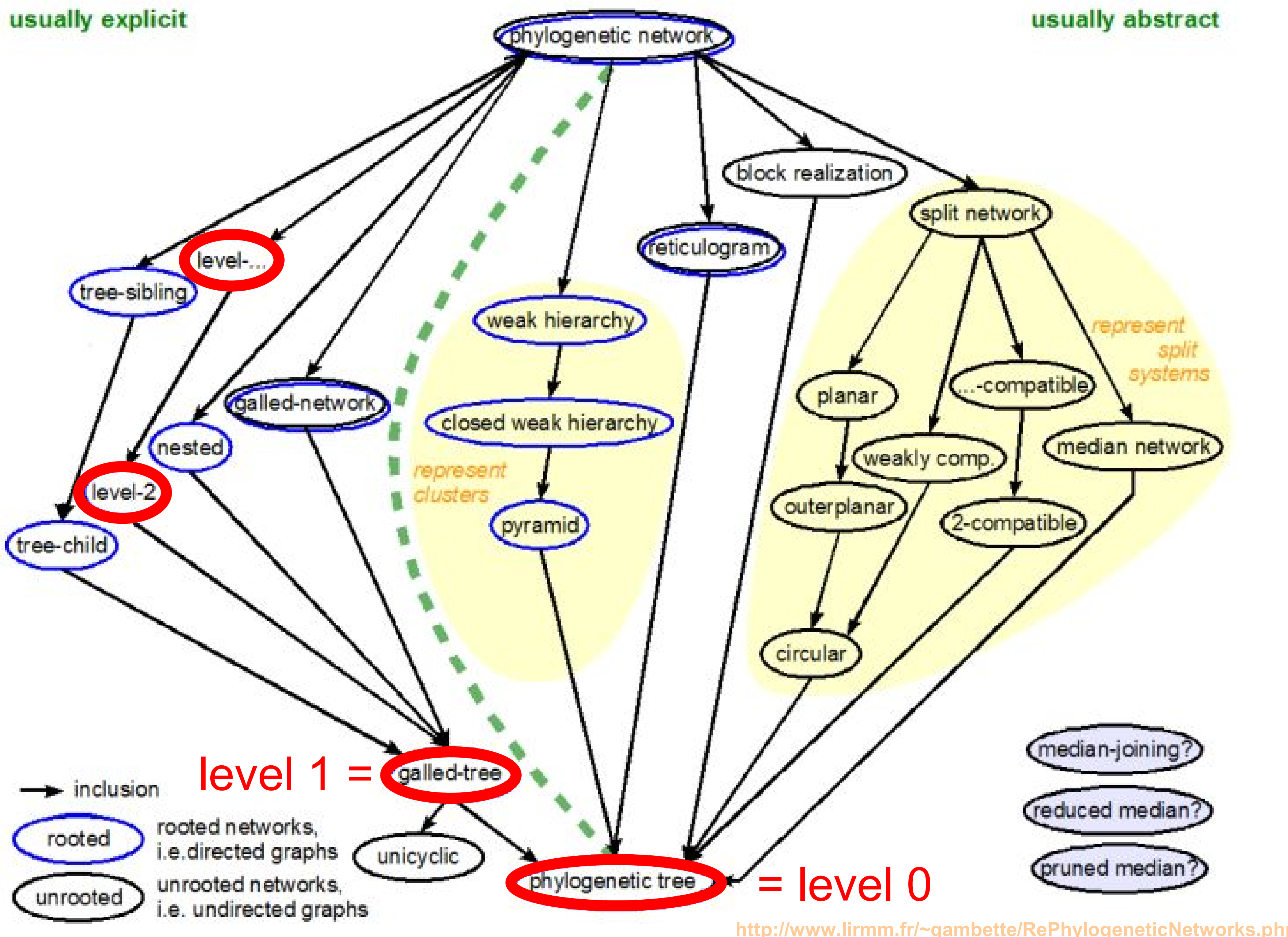
Abstract phylogenetic network of mushroom species, www.splitstree.org

Hierarchy of network subclasses



http://www.lirmm.fr/~gambette/RePhylogeneticNetworks.php

Level-*k* phylogenetic networks



Level-*k* phylogenetic networks

Motivation to generalize "galled trees" (= level-1) :

Table 1: Number of simulated networks falling in each class as a function of the recombination rate $\rho = 0, 1, 2, 4, 8, 16, 32$, for sample size $n = 10$.

Network class	Recombination rate 0	1	2	4	8	16	32
Regular	1,000	200	58	5	0	0	0
Tree-sibling	1,000	832	514	151	14	0	0
Tree-child	1,000	560	205	39	1	0	0
Galled-trees	1,000	440	137	21	1	0	0
Trees	1,000	139	27	1	0	0	0

Table 2: Number of simulated networks falling in each class as a function of the recombination rate $\rho = 0, 1, 2, 4, 8, 16, 32$, for sample size $n = 50$.

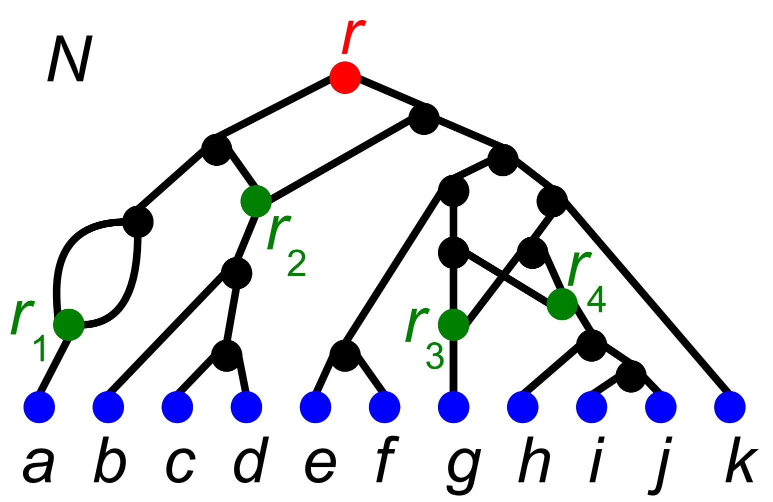
Network class	Recombination rate 0	1	2	4	8	16	32
Regular	1,000	57	1	0	0	0	0
Tree-sibling	1,000	784	469	101	2	0	0
Tree-child	1,000	463	126	9	0	0	0
Galled-trees	1,000	161	5	0	0	0	0
Trees	1,000	34	0	0	0	0	0

Arenas, Valiente, Posada : Characterization of Phylogenetic Reticulate Networks based on the Coalescent with Recombination, Molecular Biology and Evolution, to appear.

Level-*k* phylogenetic networks

A **level-*k* phylogenetic network *N*** on a set *X* of *n* taxa is a multidigraph in which:

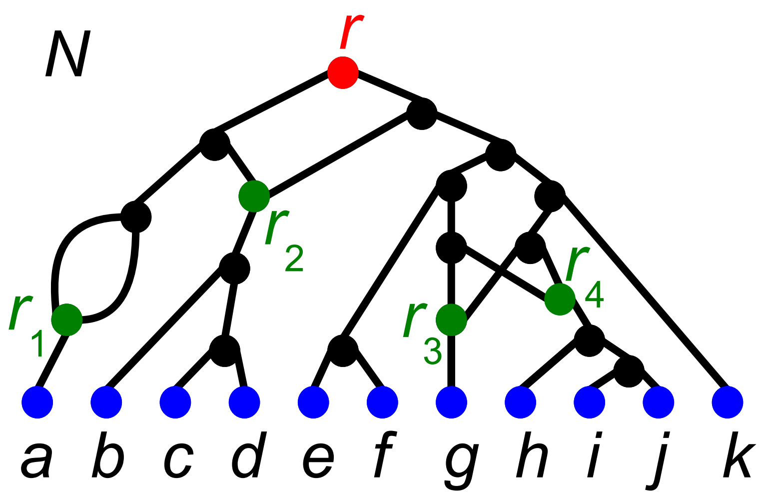
- exactly one vertex has indegree 0 and outdegree 2: the **root**,
- all other vertices have either:
    - indegree 1 and outdegree 2: **split vertices**,
    - indegree 2 and outdegree ≤ 1: **reticulation vertices**,
    - or indegree 1 and outdegree 0: **leaves** labeled by *X*,
- any **blob** has at most *k* reticulation vertices.



All arcs are oriented downwards

Level-*k* phylogenetic networks

A **level-*k* phylogenetic network *N*** on a set *X* of *n* taxa is a multidigraph in which:
- exactly one vertex has indegree 0 and outdegree 2: the **root**,
- all other vertices have either:
    - indegree 1 and outdegree 2: **split vertices**,
    - indegree 2 and outdegree ≤ 1: **reticulation vertices**,
    - or indegree 1 and outdegree 0: **leaves** labeled by *X*,
- any **blob** has at most *k* reticulation vertices.

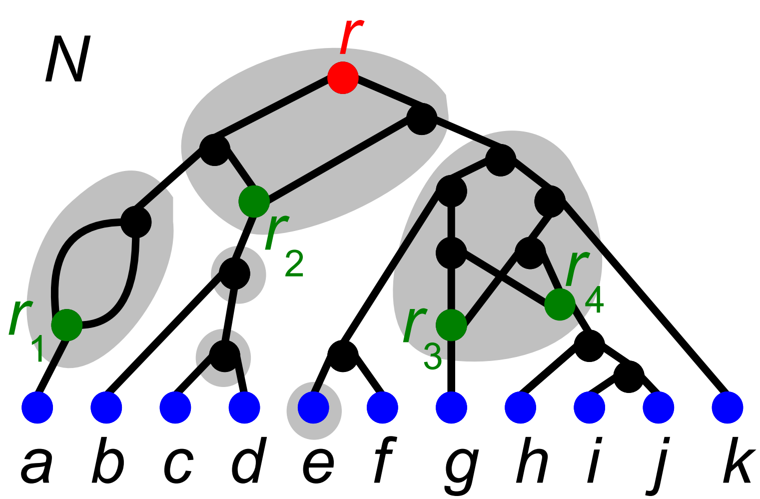


A **blob** is a maximal induced connected subgraph with no cut arc.

A **cut arc** is an arc which disconnects the graph.

Level-*k* phylogenetic networks

A **level-*k* phylogenetic network *N*** on a set *X* of *n* taxa is a multidigraph in which:
- exactly one vertex has indegree 0 and outdegree 2: the **root**,
- all other vertices have either:
    - indegree 1 and outdegree 2: **split vertices**,
    - indegree 2 and outdegree ≤ 1: **reticulation vertices**,
    - or indegree 1 and outdegree 0: **leaves** labeled by *X*,
- any **blob** has at most *k* reticulation vertices.



N has level 2.

A **blob** is a maximal induced connected subgraph with no cut arc.

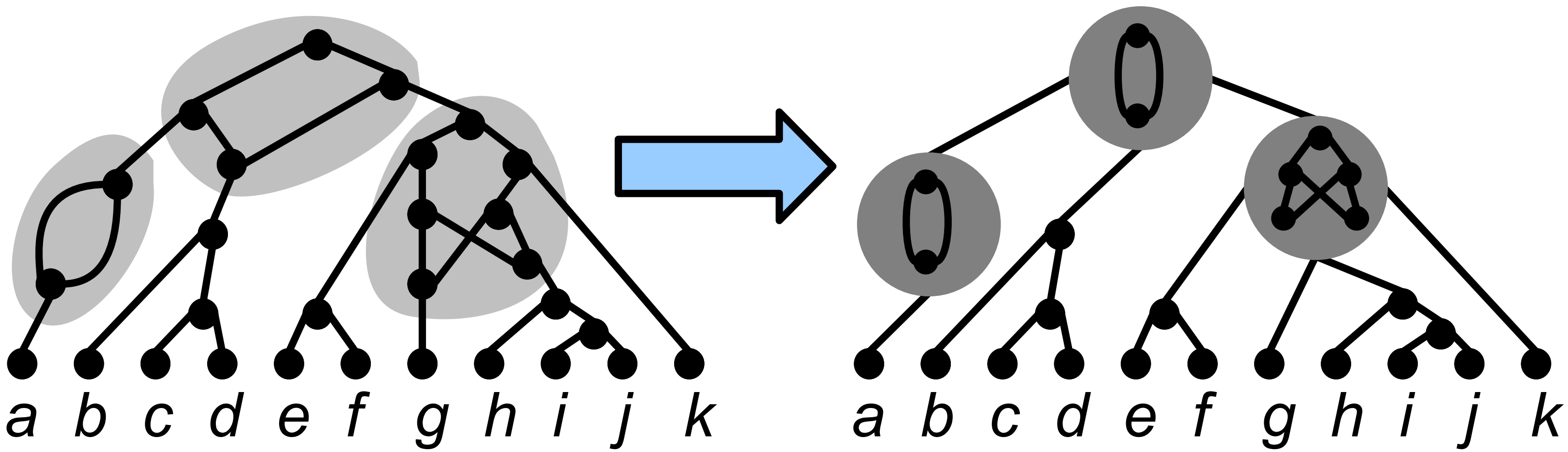
A **cut arc** is an arc which disconnects the graph.

Outline

- Phylogenetic networks
- **Decomposition of level-*k* networks**
- Construction of level-*k* generators
- Number of level-*k* generators
- Simulated level-k networks

Decomposition of level-*k* networks

We formalize the decomposition into blobs:



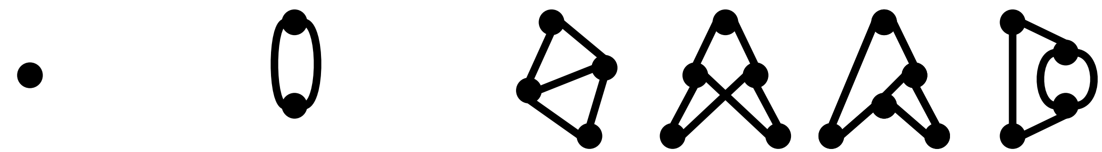
N, a level-*k* network.

N decomposed as a **tree** of simple graph patterns: **generators**.

Generators were introduced by van Iersel & al (Recomb 2008) for the restricted class of simple level-*k* networks.

Level-*k* generators

A **level-*k* generator** is a level-k network with no cut arc.



$\mathcal{G}^0$ $\mathcal{G}^1$ 2a 2b 2c 2d

The **sides** of the generator are:
- its arcs
- its reticulation vertices of outdegree 0

Decomposition theorem of level-*k* networks

N is a level-k network

iff

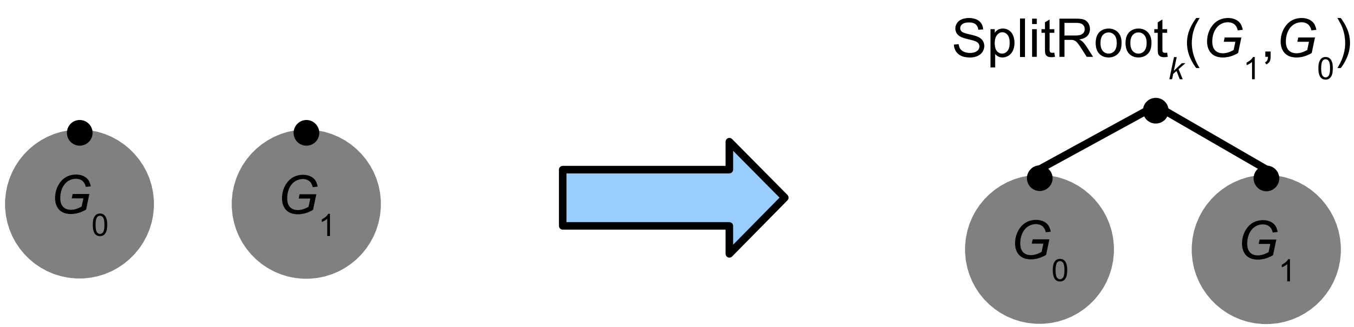
there exists a sequence $(l_j)_{j\in[1,r]}$ of r locations
(arcs or reticulation vertices of outdegree 0)
and a sequence $(G_j)_{j\in[0,r]}$ of generators of level at most k, such that:
- $N = \text{Attach}_k(l_r, G_r, \text{Attach}_k(\ldots \text{Attach}_k(l_2, G_2, \text{Attach}_k(l_1, G_1, G_0))\ldots))$,
- or $N = \text{Attach}_k(l_r, G_r, \text{Attach}_k(\ldots \text{Attach}_k(l_2, G_2, \text{SplitRoot}_k(G_1, G_0))\ldots))$.

Decomposition theorem of level-*k* networks

N is a level-*k* network

iff

there exists a sequence $(l_j)_{j \in [1,r]}$ of *r* locations
(arcs or reticulation vertices of outdegree 0)
and a sequence $(G_j)_{j \in [0,r]}$ of generators of level at most *k*, such that:
- $N = Attach_k(l_r, G_r, Attach_k(\ldots Attach_k(l_2, G_2, Attach_k(l_1, G_1, G_0))\ldots))$,
- or $N = Attach_k(l_r, G_r, Attach_k(\ldots Attach_k(l_2, G_2, SplitRoot_k(G_1, G_0))\ldots))$.



Decomposition theorem of level-*k* networks

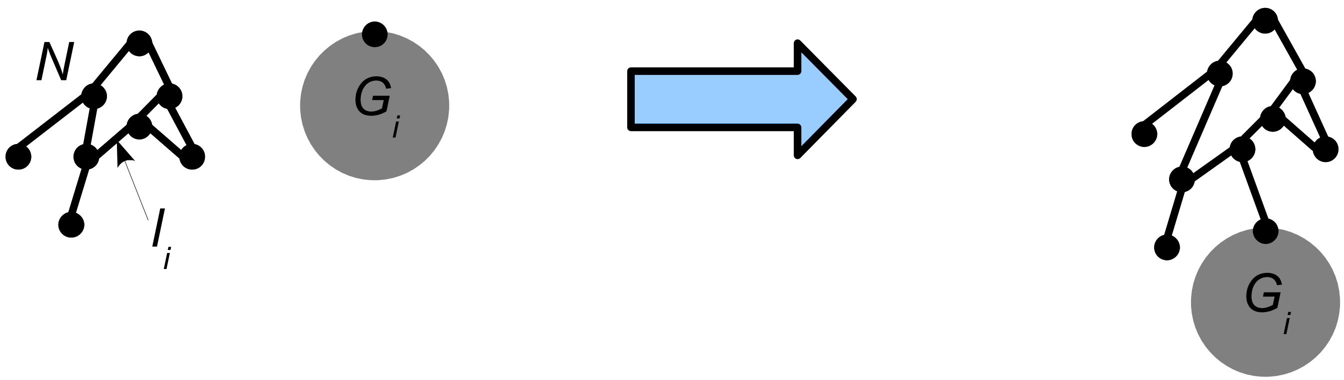
N is a level-*k* network

iff

there exists a sequence $(l_j)_{j\in[1,r]}$ of *r* locations
(arcs or reticulation vertices of outdegree 0)
and a sequence $(G_j)_{j\in[0,r]}$ of generators of level at most *k*, such that:

- N = $\text{Attach}_k(l_r,G_r,\text{Attach}_k(\ldots\ \text{Attach}_k(l_2,G_2,\text{Attach}_k(l_1,G_1,G_0))\ldots))$,
- or N = $\text{Attach}_k(l_r,G_r,\text{Attach}_k(\ldots\ \text{Attach}_k(l_2,G_2,\text{SplitRoot}_k(G_1,G_0))\ldots))$.

l_i is an arc of *N* $\qquad\qquad$ $\text{Attach}_k(l_i,G_i,N)$



Decomposition theorem of level-k networks

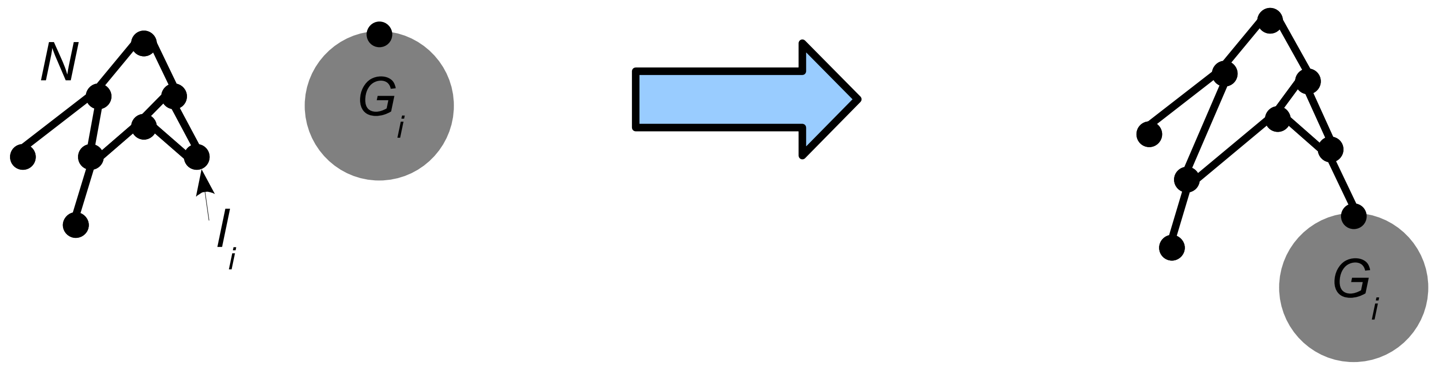
N is a level-k network

iff

there exists a sequence $(l_j)_{j\in[1,r]}$ of r locations
(arcs or reticulation vertices of outdegree 0)
and a sequence $(G_j)_{j\in[0,r]}$ of generators of level at most k, such that:
- $N = \text{Attach}_k(l_r,G_r,\text{Attach}_k(\ldots \text{Attach}_k(l_2,G_2,\text{Attach}_k(l_1,G_1,G_0))\ldots))$,
- or $N = \text{Attach}_k(l_r,G_r,\text{Attach}_k(\ldots \text{Attach}_k(l_2,G_2,\text{SplitRoot}_k(G_1,G_0))\ldots))$.

l_i is a reticulation vertex of N

$\text{Attach}_k(l_i,G_i,N)$



Decomposition theorem of level-k networks

N is a level-k network

iff

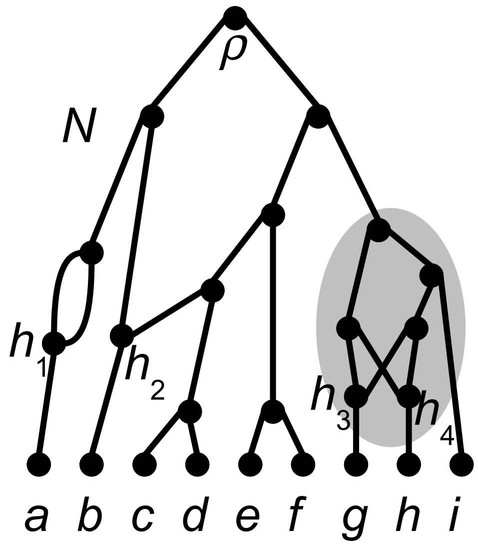
there exists a sequence $(l_j)_{j \in [1,r]}$ of r locations
(arcs or reticulation vertices of outdegree 0)
and a sequence $(G_j)_{j \in [0,r]}$ of generators of level at most k, such that:

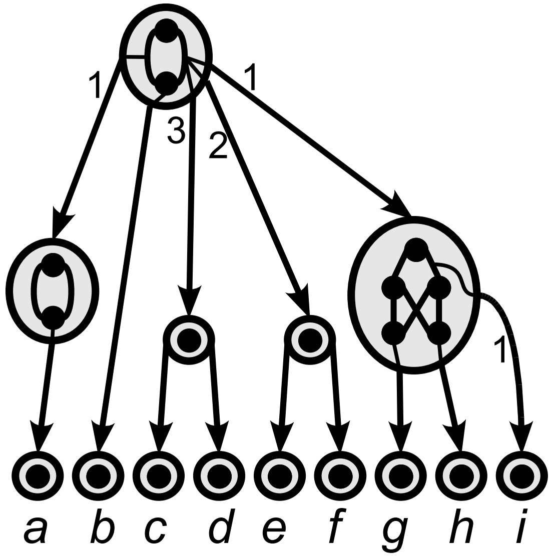
- N = $\text{Attach}_k(l_r, G_r, \text{Attach}_k(\ldots \text{Attach}_k(l_2, G_2, \text{Attach}_k(l_1, G_1, G_0))\ldots))$,
- or N = $\text{Attach}_k(l_r, G_r, \text{Attach}_k(\ldots \text{Attach}_k(l_2, G_2, \text{SplitRoot}_k(G_1, G_0))\ldots))$.

This decomposition is **not unique!**

Decomposition theorem of level-*k* networks

Unique **"graph-labeled tree" decomposition**:





Possible applications:
- exhaustive generation of level-*k* networks
- counting of level-*k* networks

Outline

- Phylogenetic networks
- Decomposition of level-*k* networks
- **Construction of level-*k* generators**
- Number of level-*k* generators
- Simulated level-*k* networks

Construction of the generators

Van Iersel & al build the 4 level-2 generators by a case analysis, generalized by Steven Kelk into an exponential algorithm to find all 65 level-3 generators.

AT&T Integer Sequences RESEARCH

Greetings from The On-Line Encyclopedia of Integer Sequences!

1,4,65 | Search | Hints

Search: **1, 4, 65**

Displaying 1-2 of 2 results found. page 1

Format: long | short | internal | text Sort: relevance | references | number Highlight: on | off

A041119 Denominators of continued fraction convergents to sqrt(68). +20 2

1, 4, 65, 264, 4289, 17420, 283009, 1149456, 18674305, 75846676, 1232221121, 5004731160, 81307919681, 330236409884, 5365090477825, 21790598321184, 354014663616769, 1437849252788260, 23359602708228929 (list; graph; listen)

OFFSET 0,2

CROSSREFS Cf. A041118.
Sequence in context: A138835 A119601 A058438 this_sequence A015475 A025585 A048828
Adjacent sequences: A041116 A041117 A041118 this_sequence A041120 A041121 A041122

KEYWORD nonn,cofr,easy

AUTHOR njas

A015475 q-Fibonacci numbers for q=4. +20 1

0, 1, 4, 65, 4164, 1066049, 1091638340, 4471351706689, 73258627454030916, 4801077413298721817665, 1258573637505038759624004676, 1319710110525284599824799048959041 (list; graph; listen)

OFFSET 0,3

FORMULA a(n) = 4^(n-1) a(n-1) + a(n-2).

CROSSREFS Sequence in context: A119601 A058438 A041119 this_sequence A025585 A048828

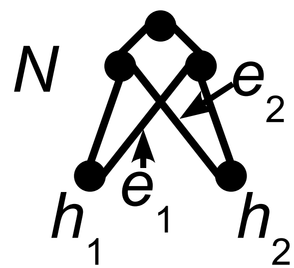
Construction of the generators

Van Iersel & al give a simple case analysis for level-2.

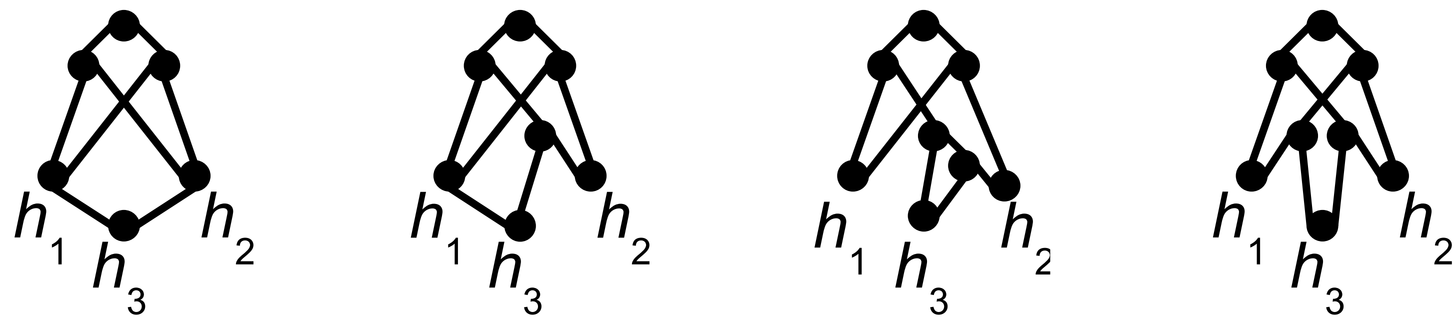
We give **rules to build level-(*k*+1) from level-*k* generators**.

Construction of the generators

Construction rules of level-k+1 generators
from level k-generators:



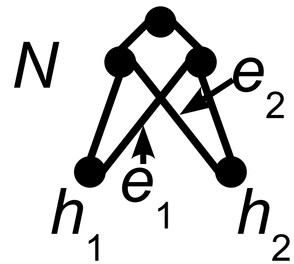
Rule R_1 :



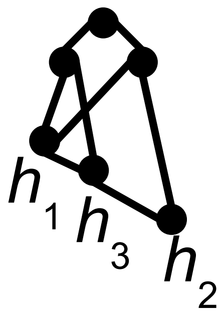
$R_1(N,h_1,h_2)$ $R_1(N,h_1,e_2)$ $R_1(N,e_2,e_2)$ $R_1(N,e_1,e_2)$

Construction of the generators

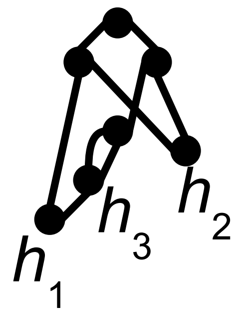
Construction rules of level-k+1 generators
from level k-generators:



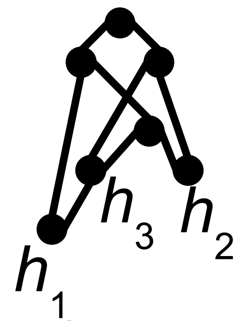
Rule R_2 :



$R_2(N,h_1,e_2)$



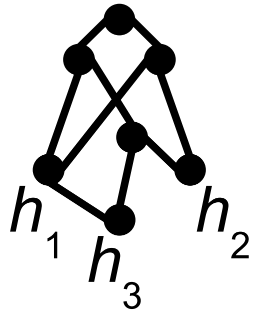
$R_2(N,e_1,e_1)$



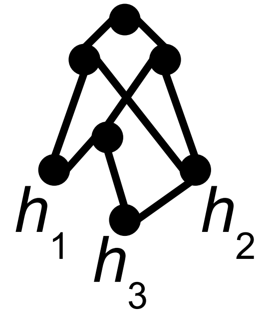
$R_2(N,e_2,e_1)$

Construction of the generators

Problem!
Some of the level-k+1 generators obtained from level-k generators are isomorphic!



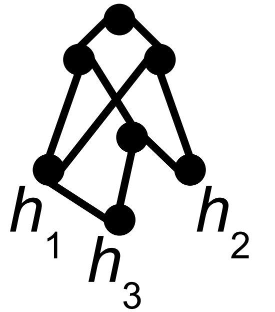
$R_1(2\text{b},h_1,e_2)$



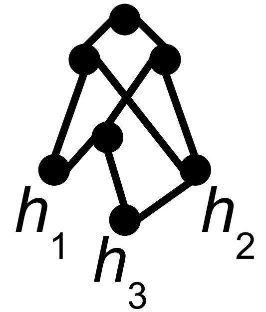
$R_1(2\text{b},h_2,e_1)$

Construction of the generators

Problem!
Some of the level-*k*+1 generators obtained from level-*k* generators are isomorphic!



$R_1(N,h_1,e_2)$



$R_1(N,h_2,e_1)$

→ difficult to count!

Outline

- Phylogenetic networks
- Decomposition of level-*k* networks
- Construction of level-*k* generators
- **Number of level-*k* generators**
- Simulated level-*k* networks

Upper bound

R_1 and R_2 can be applied at most on all pairs of sides
A level-k generator has at most $5k$ slides:

$$g_{k+1} < 50\, k^2\, g_k$$

Upper bound:

$$g_k < k!^2\, 50^k$$

Theoretical corollary:
There is a polynomial algorithm to build the set of level-k+1 generators from the set of level-k generators.

Practical corollary:

$$g_4 < 28350$$

→ it is possible to enumerate all level-4 generators.

Number of level-*k* generators

It is possible to enumerate all level-4 generators.

Isomorphism of graphs of bounded valence:
polynomial

(Luks, FOCS 1980)

Practical algorithm?
Simple backtracking exponential algorithm sufficient for level 4 :

go through both graphs from their root in parallel and identify their vertices: $O(n2^{n-h})$

→ $g_4 = 1993$
→ $g_5 > 71000$

Number of level-*k* generators

Greetings from The On-Line Encyclopedia of Integer Sequences!

1,4,65,1993 Search Hints

Search: **1, 4, 65, 1993**

I am sorry, but the terms do not match anything in the table.

Lower bound

Lower bound:

$$g_k \geq 2^{k-1}$$

There is an **exponential number** of generators!

Idea:
Code every number between 0 and 2^{k-1}-1 by a level-k generator.

Lower bound

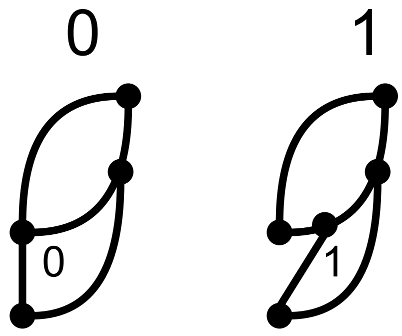
Lower bound:

$$g_k \geq 2^{k-1}$$

There is an **exponential number** of generators!

Idea:
Code every number between 0 and 2^{k-1}-1 by a level-k generator.

k = 1



Lower bound

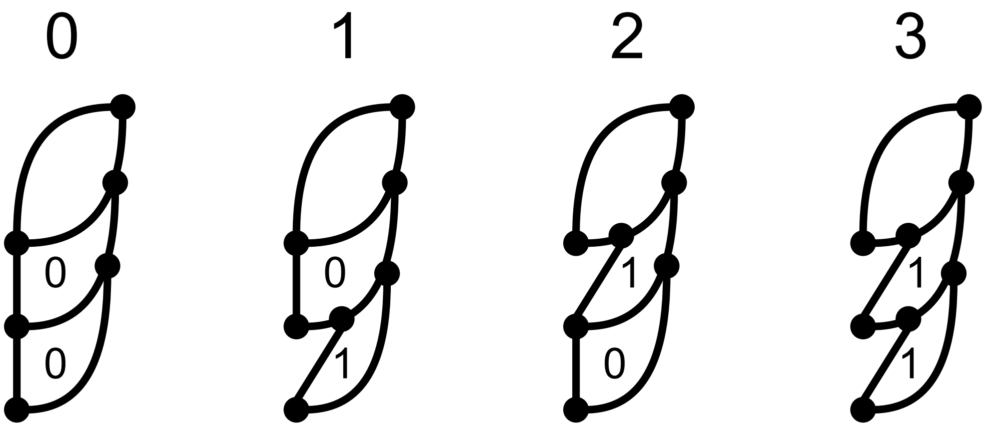
Lower bound:

$$g_k \geq 2^{k\text{-}1}$$

There is an **exponential number** of generators!

Idea:
Code every number between 0 and 2^{k-1}-1 by a level-k generator.

$k = 2$



Outline

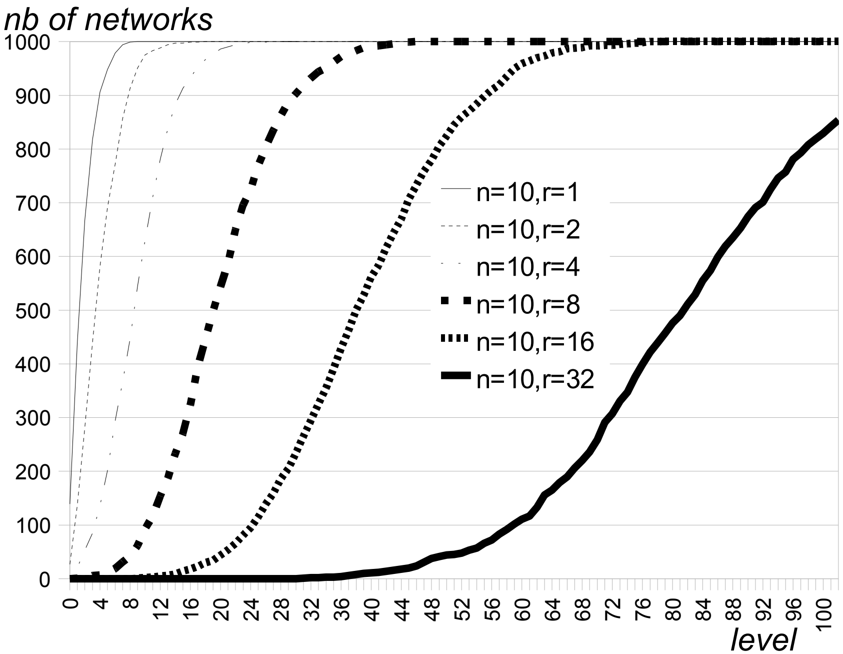
- Phylogenetic networks
- Decomposition of level-*k* networks
- Construction of level-*k* generators
- Number of level-*k* generators
- **Simulated level-*k* networks**

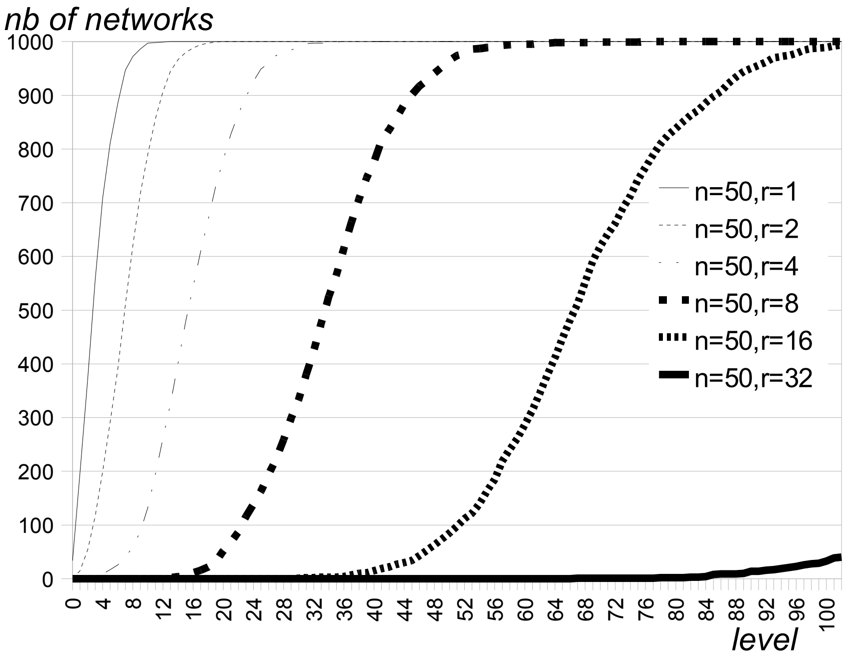
Simulated level-*k* networks

Simulate 1000 phylogenetic networks using the coalescent model with recombination.

Arenas, Valiente, Posada 2008
Program Recodon

How many are level-1,2,3... networks?

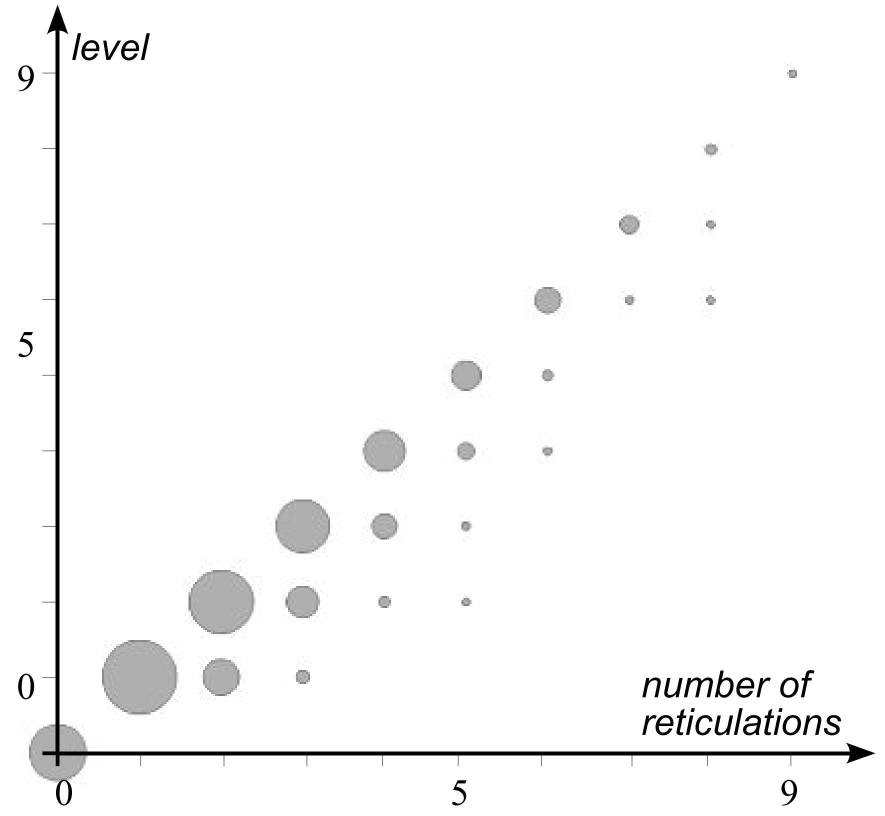




Simulated level-*k* networks

Simulate 1000 phylogenetic networks using the coalescent model with recombination.

Link between level and number of reticulations:



Summary on the level parameter

Advantages:

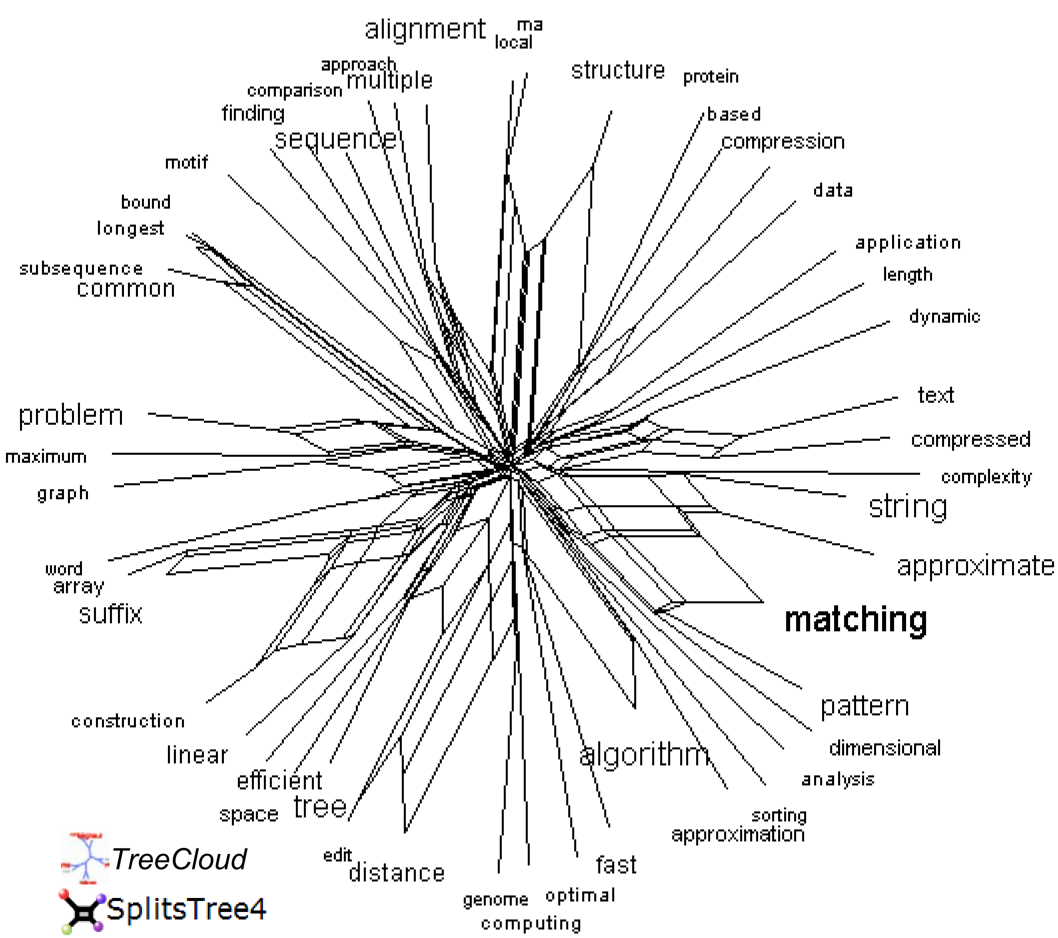
- natural structure for all explicit phylogenetic networks
- global tree-structure used algorithmically
- finite graph patterns to represent blobs: generators

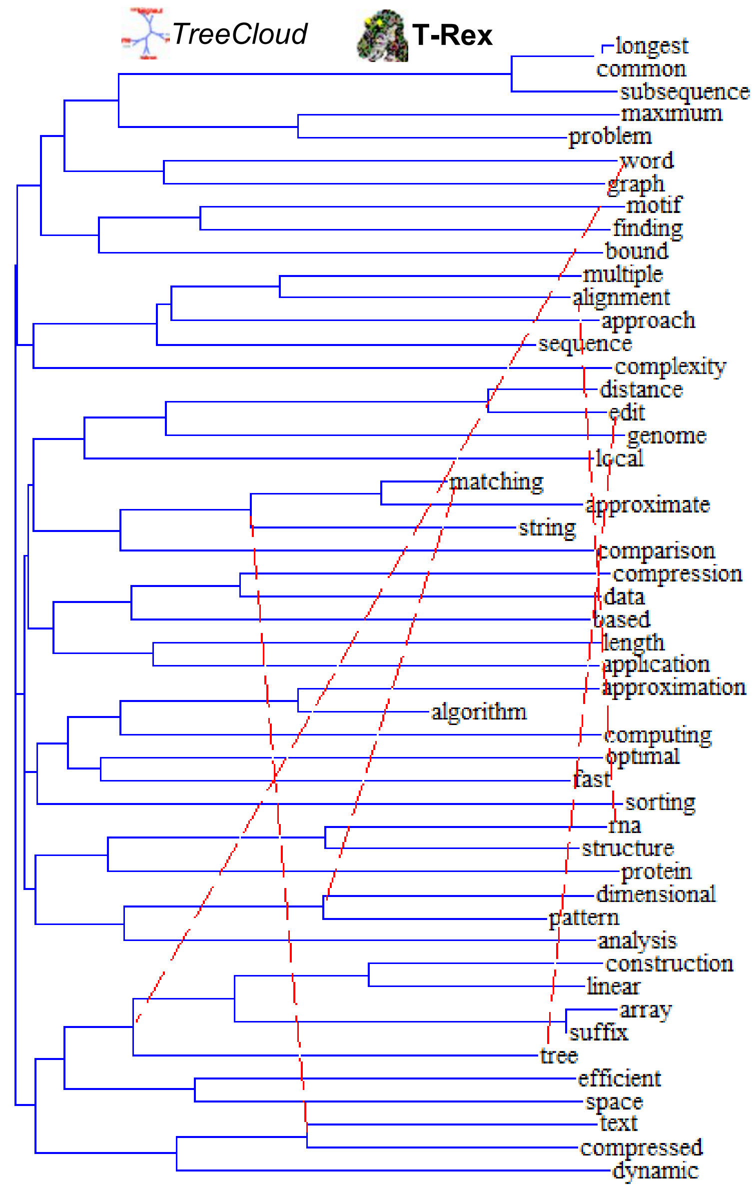
Limits:

- number of generators exponential in the level
- complex structure of generators
- when recombination is not local, level doesn't help

Questions?

Thank you for your attention!





Split network and reticulogram of the 50 most frequent words in CPM titles, hyperlex cooccurrence distance.

TreeCloud available at http://treecloud.org - Slides available at **http://www.lirmm.fr/~gambette/RePresentationsENG.php**