Grenoble – 03/09/2009 – SFC'09

Structure des réseaux phylogénétiques de niveau borné

Philippe Gambette
Vincent Berry, Christophe Paul

Plan

- Classification et réseaux phylogénétiques
- Décomposition des réseaux de niveau *k*
- Construction des générateurs de niveau *k*
- Nombre de générateurs de niveau *k*
- Simulation de réseaux de niveau *k*

Plan

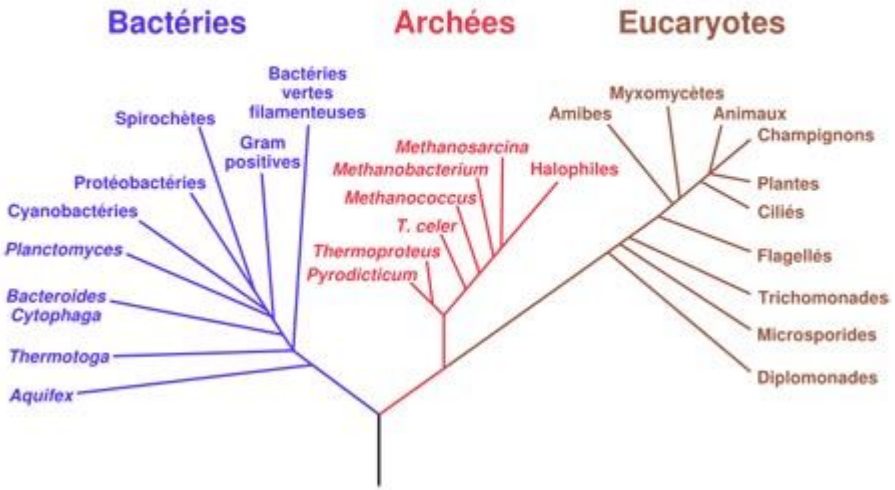
- **Classification et réseaux phylogénétiques**
- Décomposition des réseaux de niveau *k*
- Construction des générateurs de niveau *k*
- Nombre de générateurs de niveau *k*
- Simulation de réseaux de niveau *k*

Les arbres phylogénétiques

Arbre phylogénétique

Un **arbre phylogénétique** est un arbre schématique qui montre les relations de parentés entre des entités supposées avoir un ancêtre commun.

Arbre phylogénétique de la vie

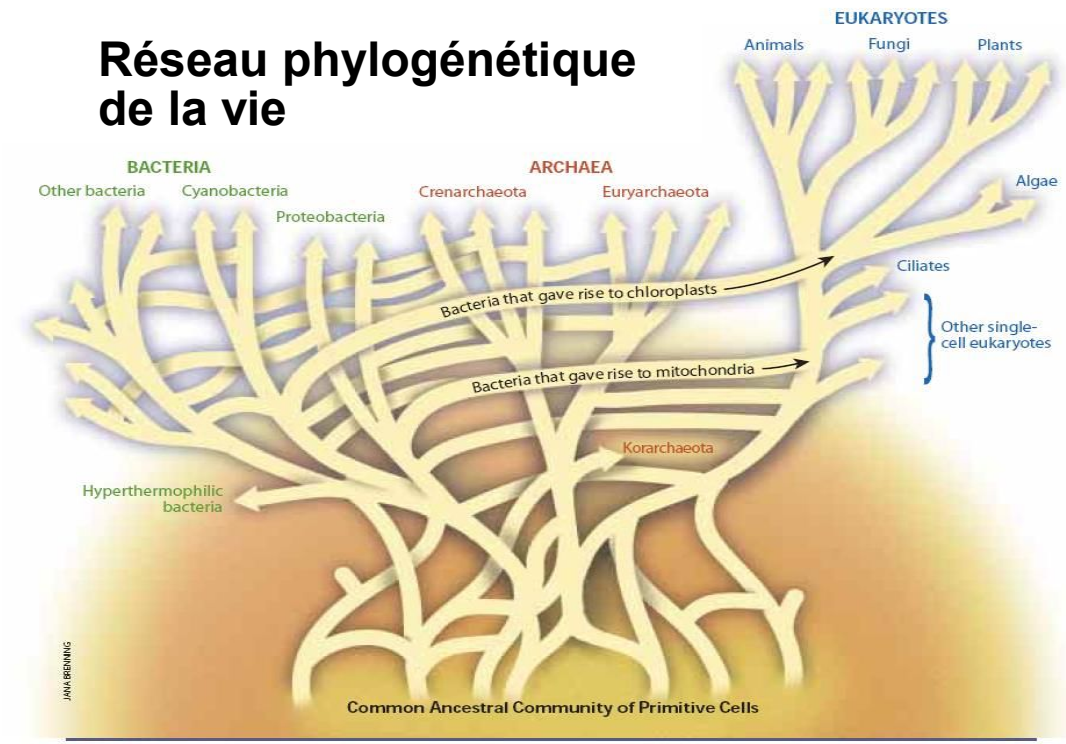


D'après Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

Les réseaux phylogénétiques

Réseau phylogénétique

Un **réseau phylogénétique** désigne un graphe utilisé pour visualiser les relations liées à l'évolution entre des espèces ou des organismes. Il doit être employé quand interviennent des événements d'hybridations, de transferts horizontaux de gènes, ou de recombinaisons génétiques.



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

Les réseaux phylogénétiques

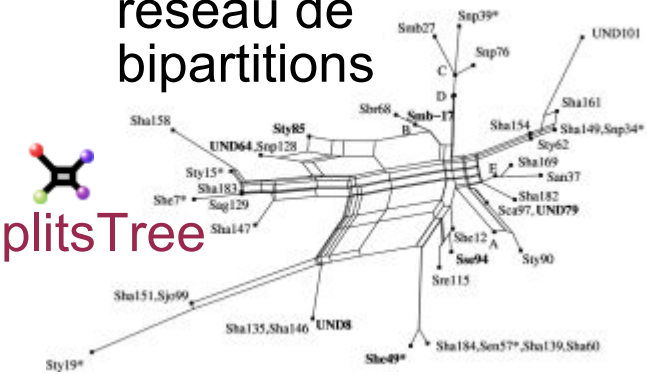
Réseau phylogénétique

Un **réseau phylogénétique** désigne un graphe utilisé pour visualiser les relations liées à l'évolution entre des espèces ou des organismes. Il doit être employé quand interviennent des événements d'hybridations, de transferts horizontaux de gènes, ou de recombinaisons génétiques.

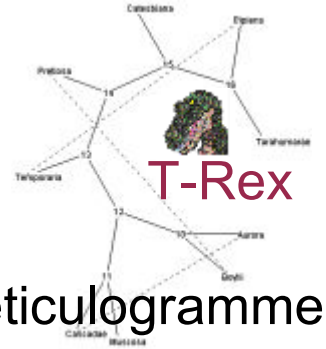
réseau de niveau 2

Level-2

réseau de bipartitions



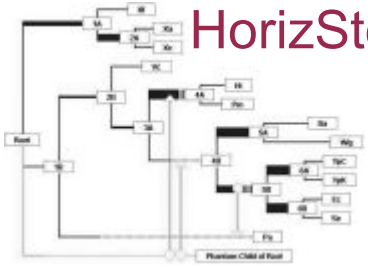
SplitsTree



T-Rex

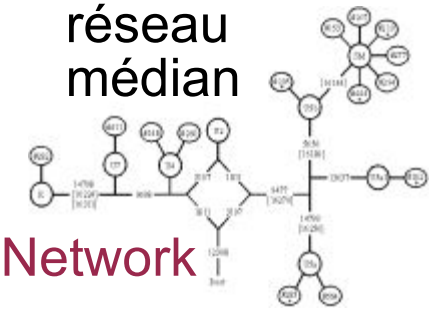
réticulogramme

diagramme de synthèse



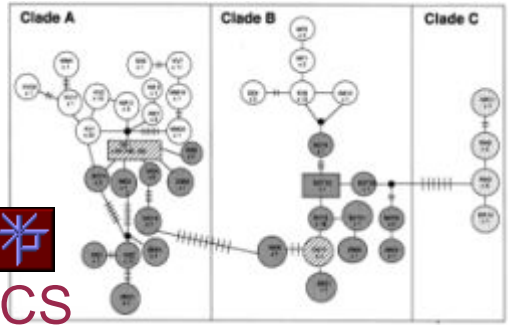
HorizStory

réseau médian



Network

réseau couvrant minimum

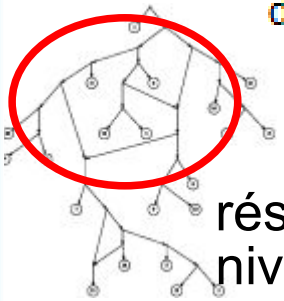


TCS

Les réseaux phylogénétiques

Réseau phylogénétique

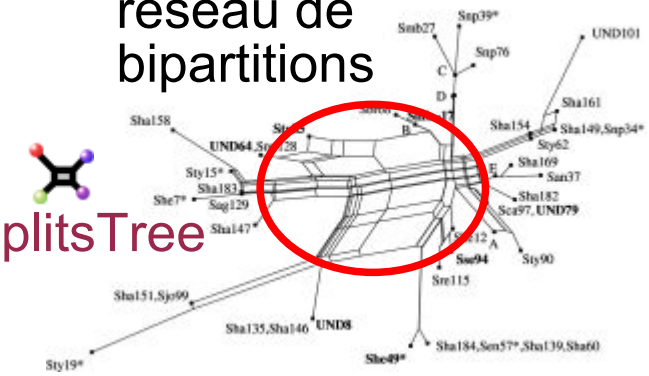
Un **réseau phylogénétique** désigne un graphe utilisé pour visualiser les relations liées à l'évolution entre des espèces ou des organismes. Il doit être employé quand interviennent des événements d'hybridations, de transferts horizontaux de gènes, ou de recombinaisons génétiques.



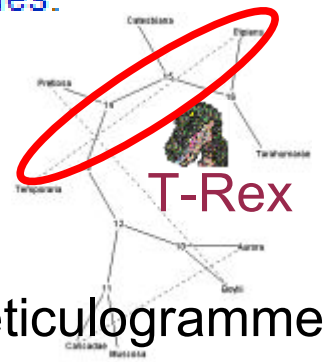
réseau de niveau 2

Level-2

réseau de bipartitions



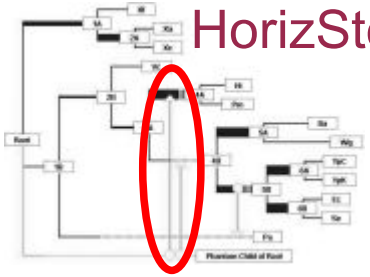
SplitsTree



T-Rex

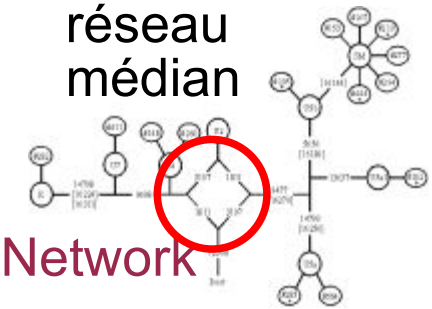
réticulogramme

diagramme de synthèse



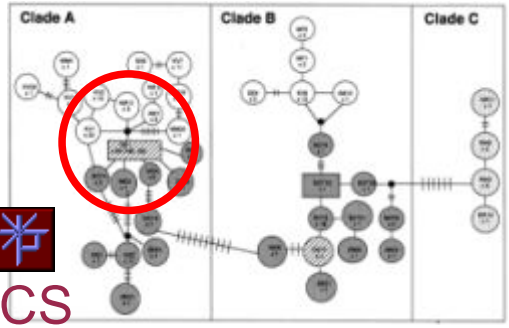
HorizStory

réseau médian



Network

réseau couvrant minimum



TCS

Réseaux phylogénétiques et classification

Arbre : hiérarchie correspondant à :
- un ensemble de **clusters** sans chevauchement
- un ensemble de **triplets** évitant certaines obstructions
- une distance d'arbre

Réseaux phylogénétiques et classification

Réseau : correspondant à :
- un ensemble de clusters avec propriétés plus faibles : **hiérarchies faibles**, **pyramides**...

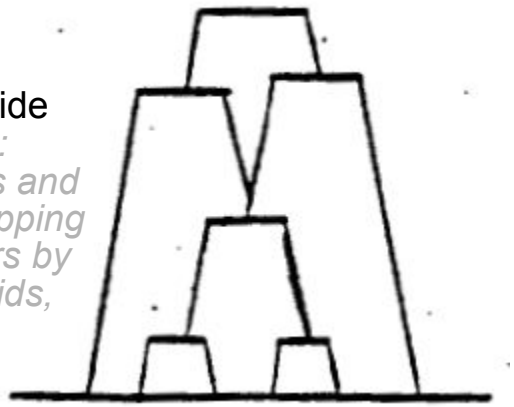
Bertrand & Diday 1985, Bandelt & Dress 1989

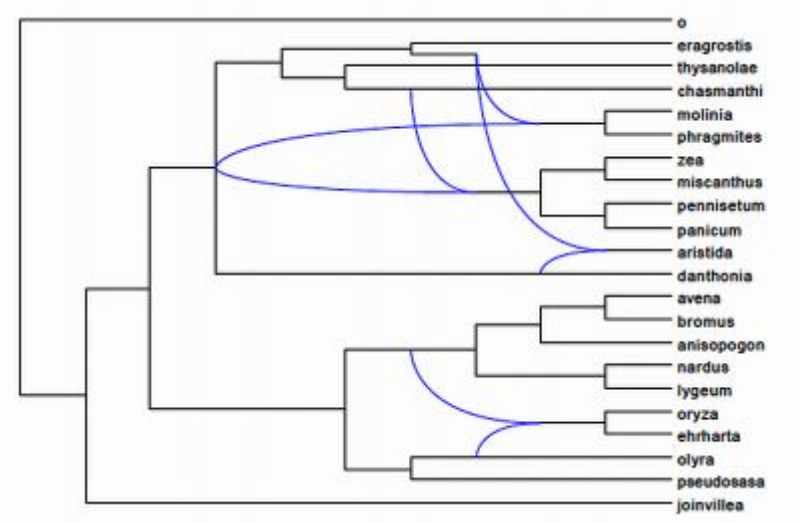
- un ensemble de clusters avec modèle d'évolution différent : **réseaux "softwired"**

Rupp & Huson 2008

- un modèle **interprétable biologiquement** : noeuds de degré au plus 3

Pyramide
Diday : Orders and overlapping clusters by pyramids, 1987



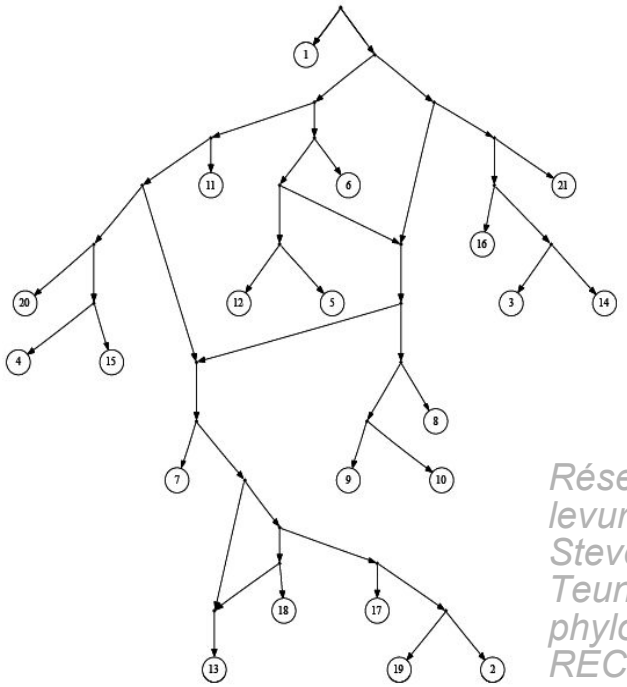


Réseau "softwired"
Huson, Rupp, Berry, Gambette, Paul, Computing galled networks from real data, 2009

Réseaux abstraits ou explicites

Un **réseau phylogénétique explicite** est un réseau phylogénétique dont tous les noeuds correspondent à des événements biologiques précis.

Un **réseau phylogénétique abstrait** reflète des signaux phylogénétiques sans nécessairement représenter explicitement des événements biologiques.

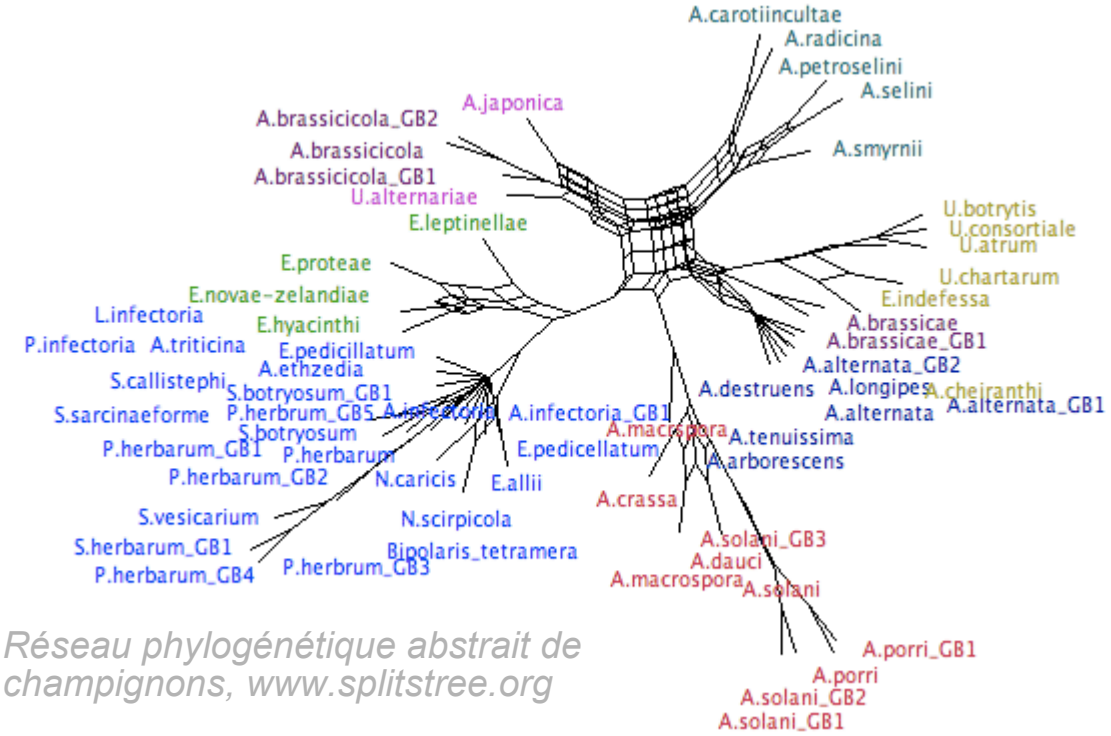


Réseau phylogénétique explicite de levures, Leo van Iersel, Judith Keijsper, Steven Kelk, Leen Stougie, Ferry Hagen, Teun Boekhout : Constructing level-2 phylogenetic networks from triplets. RECOMB'08

Réseaux abstraits ou explicites

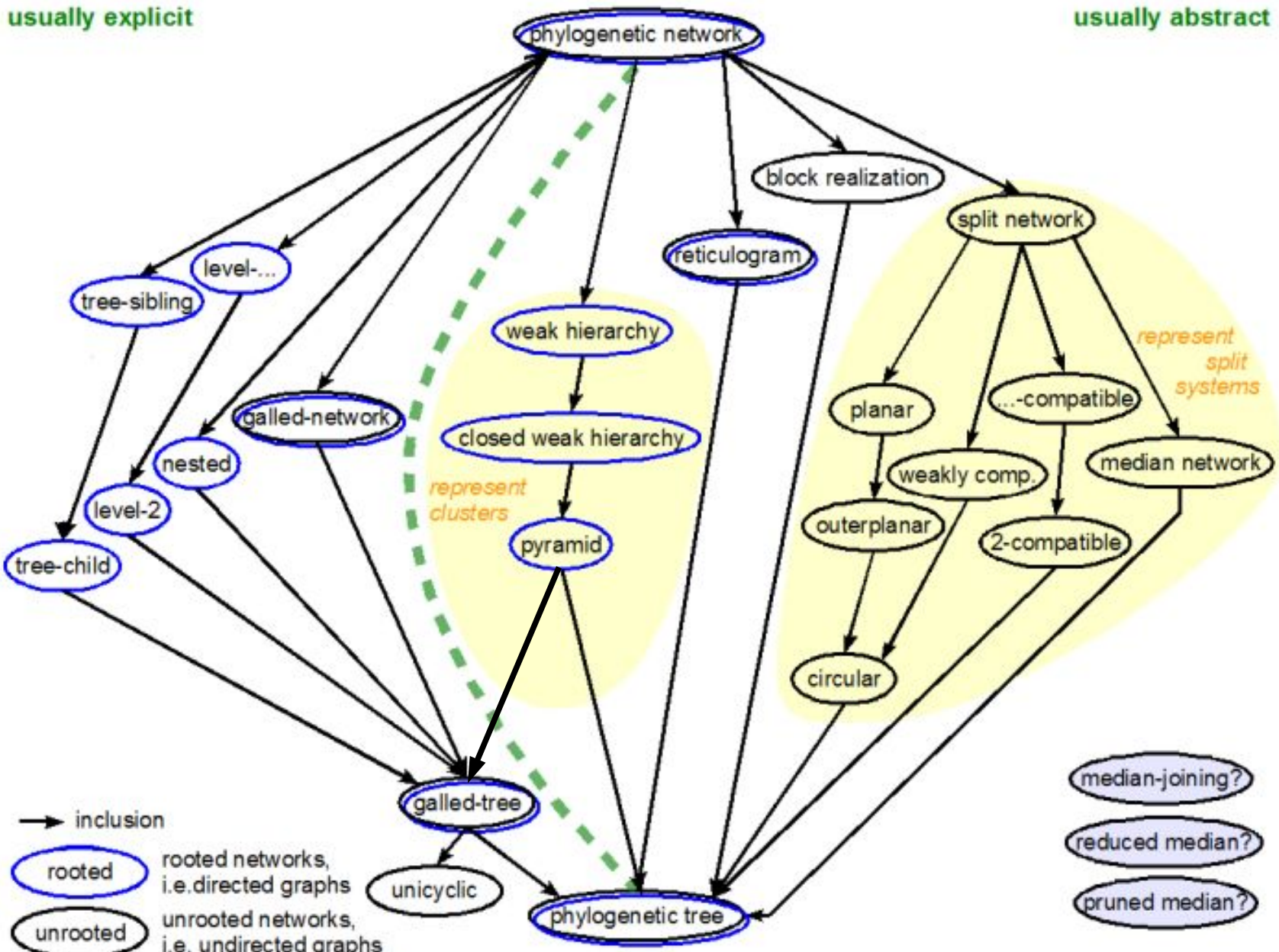
Un **réseau phylogénétique explicite** est un réseau phylogénétique dont tous les noeuds correspondent à des événements biologiques précis.

Un **réseau phylogénétique abstrait** reflète des signaux phylogénétiques sans nécessairement représenter explicitement des événements biologiques.



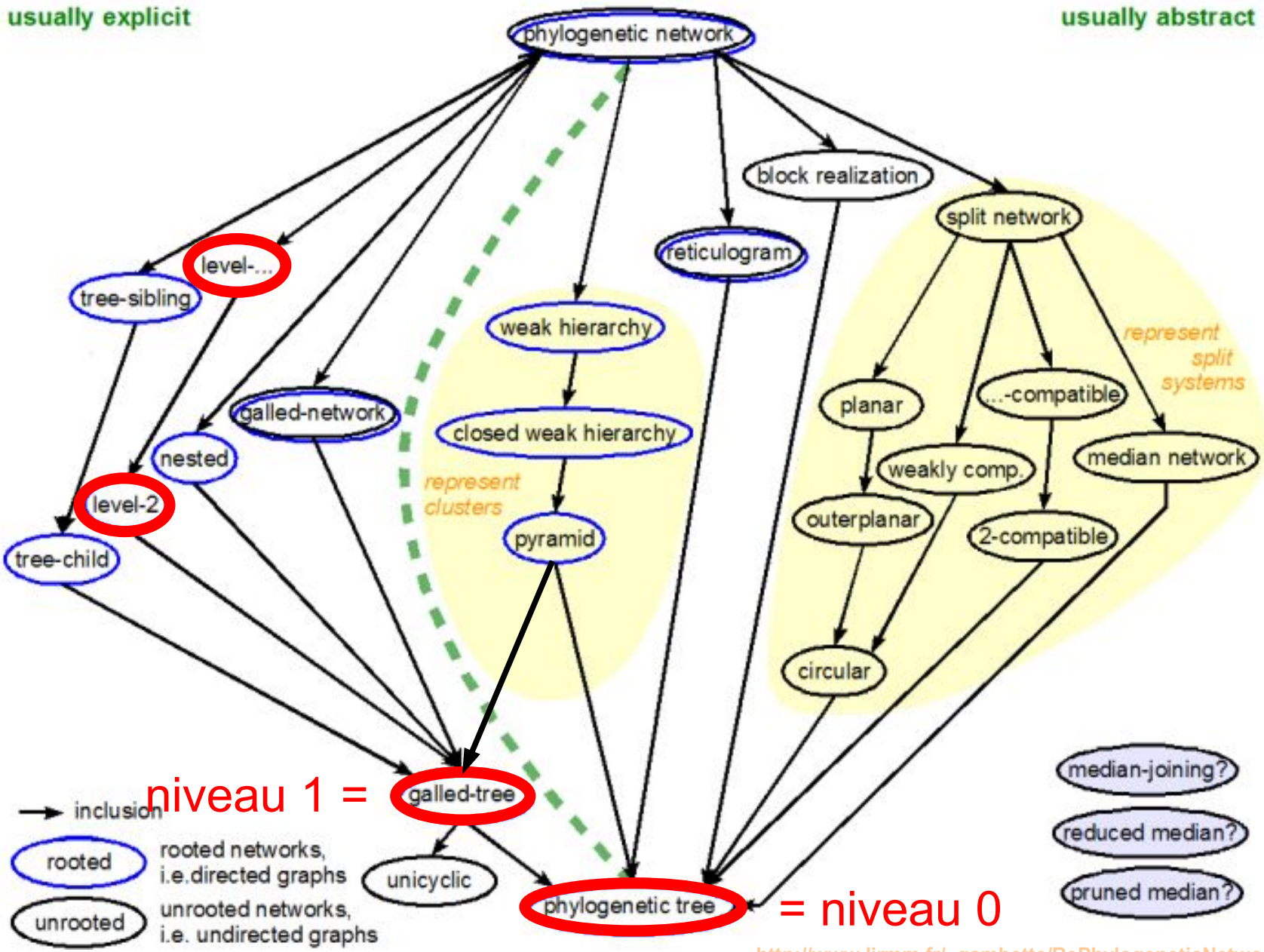
Réseau phylogénétique abstrait de champignons, www.splitstree.org

Hiérarchie de sous-classes de réseaux



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

Réseaux phylogénétiques de niveau *k*



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

Réseaux phylogénétiques de niveau *k*

Motivation : généraliser les “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

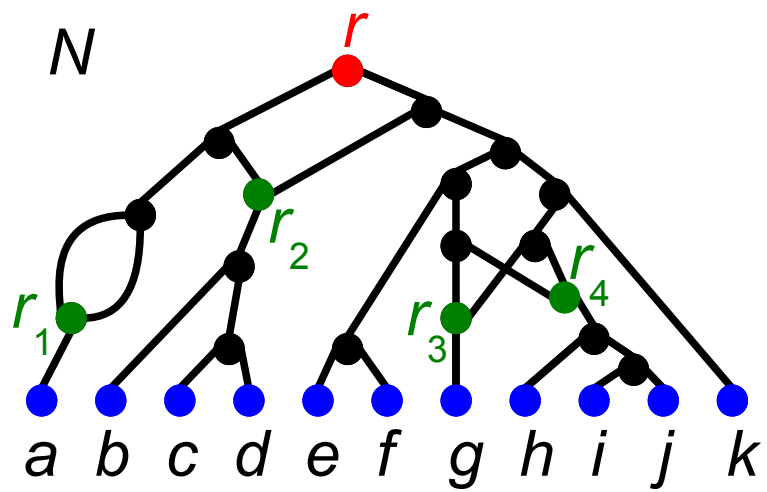
Table 2: Number of simulated networks falling in each class as a function of the recombination rate $\rho = 0, 1, 2, 4, 8, 16, 32$, for sample size $n = 50$.

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.

Réseaux phylogénétiques de niveau *k*

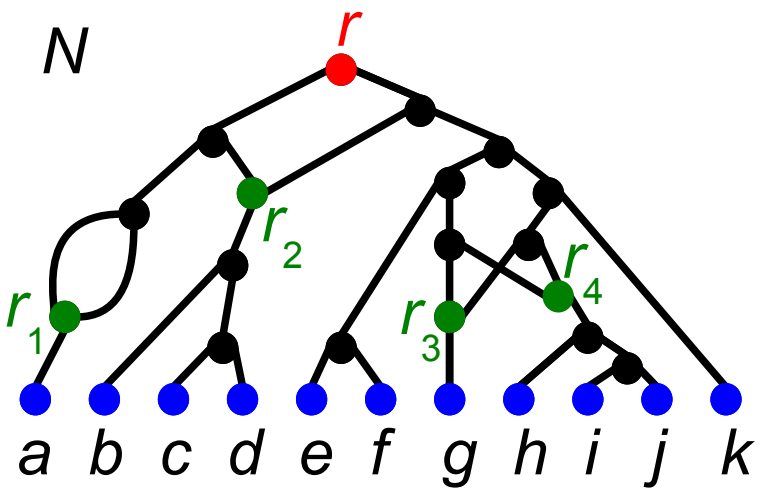
Un **réseau phylogénétique *N* de niveau *k*** sur un ensemble *X* de *n* taxons est un multigraphe orienté dans lequel :
- exactement un sommet a degré entrant 0 et sortant 2 : **racine**,
- tous les autres sommets ont :
  - degré entrant 1 et sortant 2 : **sommets de spéciation**,
  - degré entrant 2 et sortant ≤ 1: **sommets de réticulation**,
  - ou degré entrant 1 et sortant 0 : **feuilles** étiquetées par *X*,
- tout **blob** a au plus *k* sommets de réticulation.



Arcs orientés vers le bas.

Réseaux phylogénétiques de niveau *k*

Un **réseau phylogénétique *N* de niveau *k*** sur un ensemble *X* de *n* taxons est un multigraphe orienté dans lequel :
- exactement un sommet a degré entrant 0 et sortant 2 : **racine**,
- tous les autres sommets ont :
  - degré entrant 1 et sortant 2 : **sommets de spéciation**,
  - degré entrant 2 et sortant ≤ 1: **sommets de réticulation**,
  - ou degré entrant 1 et sortant 0 : **feuilles** étiquetées par *X*,
- tout **blob** a au plus *k* sommets de réticulation.

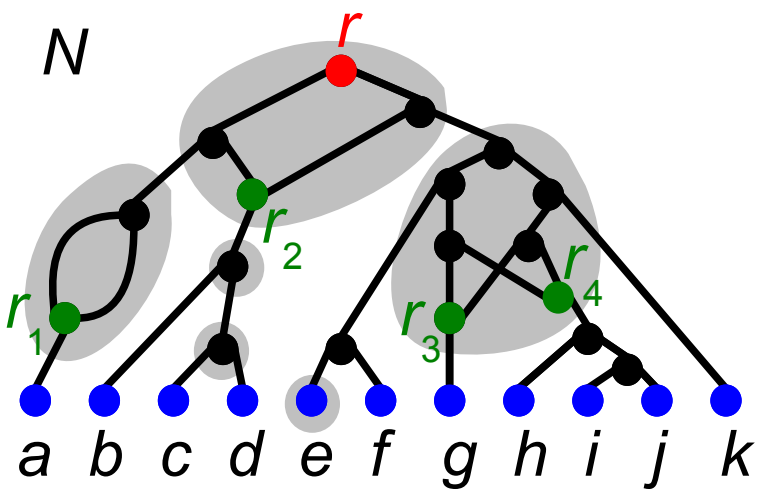


Un **blob** est un sous-graphe induit connexe maximal sans isthme.

Un **isthme** est un arc qui déconnecte le graphe.

Réseaux phylogénétiques de niveau *k*

Un **réseau phylogénétique *N* de niveau *k*** sur un ensemble *X* de *n* taxons est un multigraphe orienté dans lequel :
- exactement un sommet a degré entrant 0 et sortant 2 : **racine**,
- tous les autres sommets ont :
  - degré entrant 1 et sortant 2 : **sommets de spéciation**,
  - degré entrant 2 et sortant ≤ 1: **sommets de réticulation**,
  - ou degré entrant 1 et sortant 0 : **feuilles** étiquetées par *X*,
- tout **blob** a au plus *k* sommets de réticulation.



N a niveau 2.

Un **blob** est un sous-graphe induit connexe maximal sans isthme.

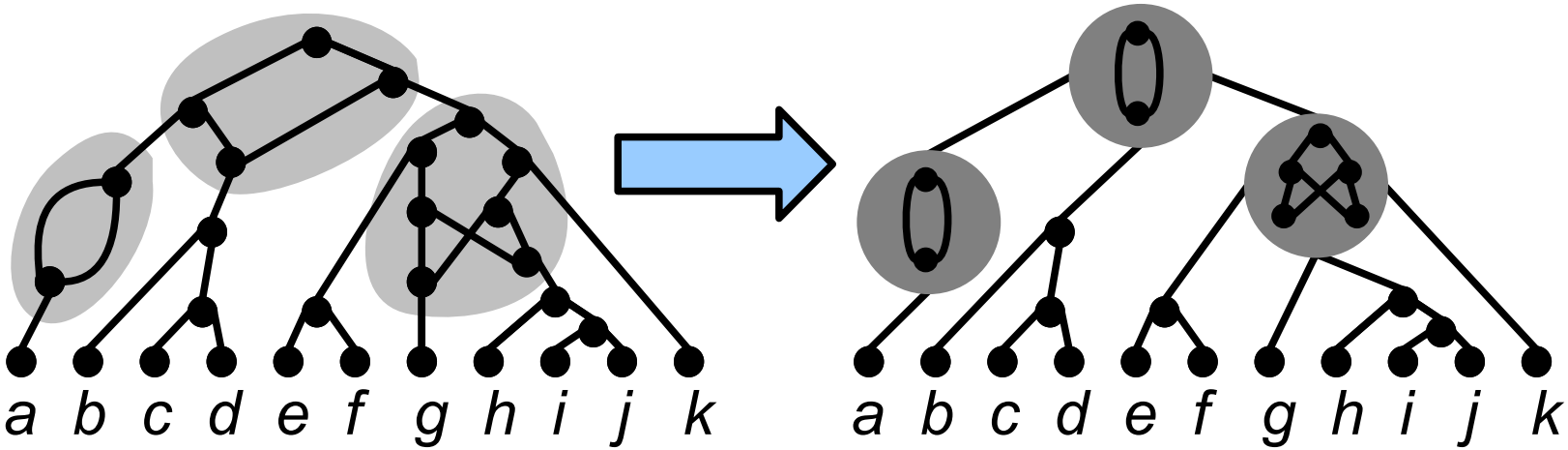
Un **isthme** est un arc qui déconnecte le graphe.

Plan

- Classification et réseaux phylogénétiques
- **Décomposition des réseaux de niveau *k***
- Construction des générateurs de niveau *k*
- Nombre de générateurs de niveau *k*
- Simulation de réseaux de niveau *k*

Décomposition des réseaux de niveau *k*

On formalise la décomposition en blobs :



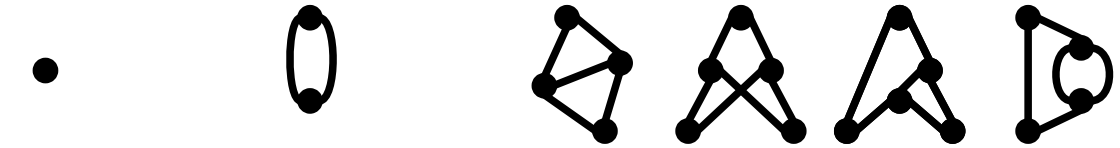
N, un réseau de niveau *k*.

Décomposition arborée de *N* en générateurs.

Générateurs introduits par van Iersel & al (Recomb 2008) pour la classe restreinte des réseaux *simples* de niveau *k*.

Générateurs de niveau *k*

Un **générateur de niveau *k*** est un réseau de niveau *k* sans isthme.



Les **côtés** d'un générateur sont :
- ses arcs,
- ses sommets de réticulation de degré sortant 0.

Décomposition des réseaux de niveau *k*

N est un réseau de niveau *k*

ssi

il existe une suite $(l_j)_{j\in[1,r]}$ de *r* emplacements
(arcs ou sommets hybrides de degré sortant 0)
et une suite $(G_j)_{j\in[0,r]}$ de générateurs de niveau au plus *k*, telles que :

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

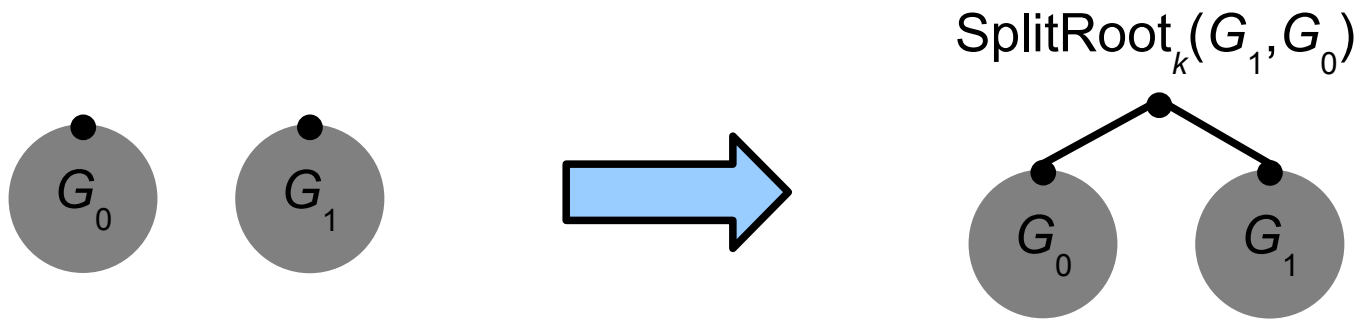
Décomposition des réseaux de niveau *k*

N est un réseau de niveau *k*

ssi

il existe une suite $(I_j)_{j\in[1,r]}$ de *r* emplacements
(arcs ou sommets hybrides de degré sortant 0)
et une suite $(G_j)_{j\in[0,r]}$ de générateurs de niveau au plus *k*, telles que :

- $N = \mathrm{Attach}_k(I_r,G_r,\mathrm{Attach}_k(\ldots \mathrm{Attach}_k(I_2,G_2,\mathrm{Attach}_k(I_1,G_1,G_0))\ldots))$,
- or $N = \mathrm{Attach}_k(I_r,G_r,\mathrm{Attach}_k(\ldots \mathrm{Attach}_k(I_2,G_2,\mathrm{SplitRoot}_k(G_1,G_0))\ldots))$.



Décomposition des réseaux de niveau *k*

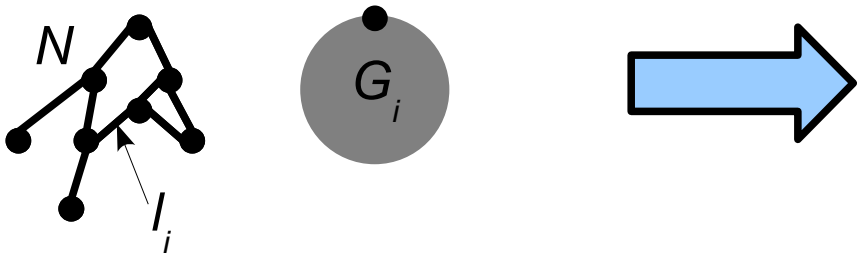
N est un réseau de niveau *k*

ssi

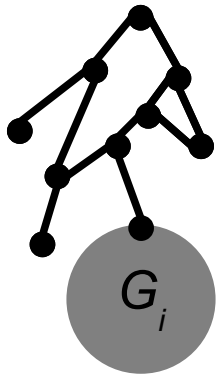
il existe une suite $(I_j)_{j\in[1,r]}$ de *r* emplacements
(arcs ou sommets hybrides de degré sortant 0)
et une suite $(G_j)_{j\in[0,r]}$ de générateurs de niveau au plus *k*, telles que :

- $N = \mathrm{Attach}_k(I_r, G_r, \mathrm{Attach}_k(\ldots \mathrm{Attach}_k(I_2, G_2, \mathrm{Attach}_k(I_1, G_1, G_0))\ldots))$,
- or $N = \mathrm{Attach}_k(I_r, G_r, \mathrm{Attach}_k(\ldots \mathrm{Attach}_k(I_2, G_2, \mathrm{SplitRoot}_k(G_1, G_0))\ldots))$.

I_i arc de N



$\mathrm{Attach}_k(I_i, G_i, N)$



Décomposition des réseaux de niveau *k*

N est un réseau de niveau *k*

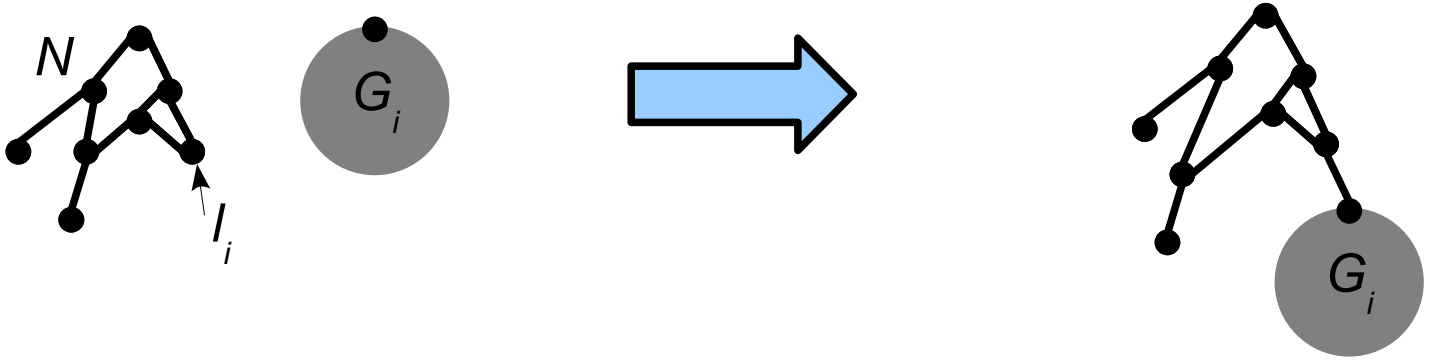
ssi

il existe une suite $(I_j)_{j\in[1,r]}$ de *r* emplacements
(arcs ou sommets hybrides de degré sortant 0)
et une suite $(G_j)_{j\in[0,r]}$ de générateurs de niveau au plus *k*, telles que :

- $N = \mathrm{Attach}_k(I_r,G_r,\mathrm{Attach}_k(\ldots \mathrm{Attach}_k(I_2,G_2,\mathrm{Attach}_k(I_1,G_1,G_0))\ldots))$,
- or $N = \mathrm{Attach}_k(I_r,G_r,\mathrm{Attach}_k(\ldots \mathrm{Attach}_k(I_2,G_2,\mathrm{SplitRoot}_k(G_1,G_0))\ldots))$.

I_i sommet de réticulation de *N*

$\mathrm{Attach}_k(I_i,G_i,N)$



Décomposition des réseaux de niveau *k*

N est un réseau de niveau *k*

ssi

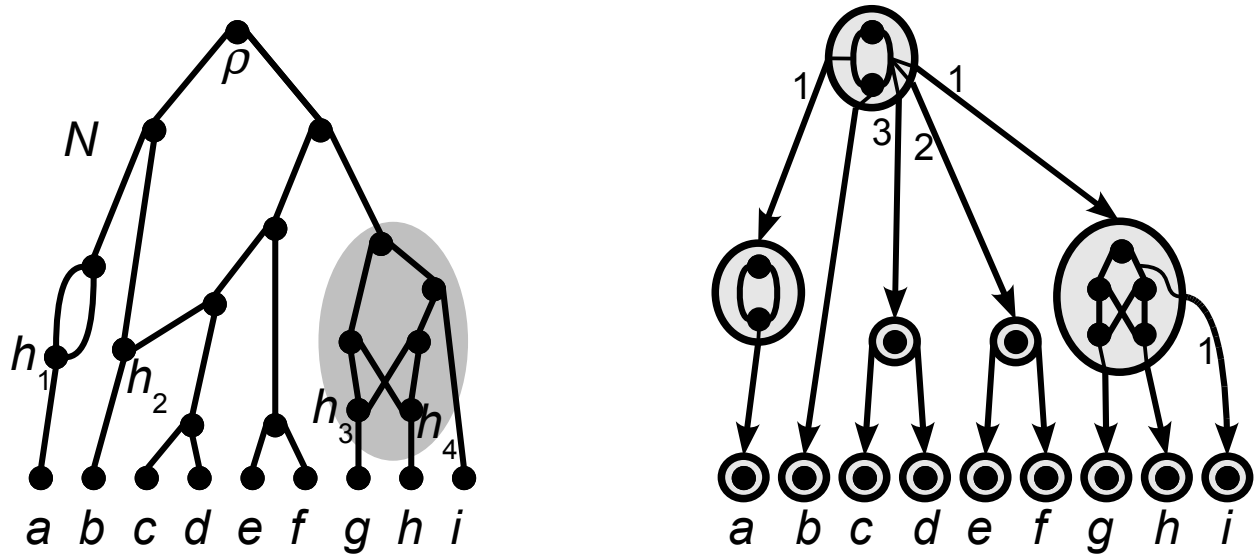
il existe une suite $(I_j)_{j\in[1,r]}$ de *r* emplacements
(arcs ou sommets hybrides de degré sortant 0)
et une suite $(G_j)_{j\in[0,r]}$ de générateurs de niveau au plus *k*, telles que :

- $N = \text{Attach}_k(I_r, G_r, \text{Attach}_k(\ldots \text{Attach}_k(I_2, G_2, \text{Attach}_k(I_1, G_1, G_0))\ldots))$,
- or $N = \text{Attach}_k(I_r, G_r, \text{Attach}_k(\ldots \text{Attach}_k(I_2, G_2, \text{SplitRoot}_k(G_1, G_0))\ldots))$.

Cette décomposition **n'est pas unique !**

Décomposition des réseaux de niveau *k*

Arbre de décomposition **unique** étiqueté par graphes :



Applications possibles :
- génération exhaustive des réseaux de niveau *k*
- énumération des réseaux de niveau *k*

Plan

- Classification et réseaux phylogénétiques
- Décomposition des réseaux de niveau *k*
- **Construction des générateurs de niveau *k***
- Nombre de générateurs de niveau *k*
- Simulation de réseaux de niveau *k*

Construction des générateurs

Analyse de cas par Van Iersel & al pour trouver les 4 générateurs de niveau 2.
Généralisation par Steven Kelk en un algorithme exponentiel pour trouver les 65 générateurs de niveau 3.

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, 480107741329872181766 5, 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

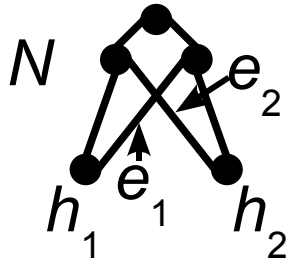
Construction des générateurs

Analyse de cas par Van Iersel & al pour trouver les 4 générateurs de niveau 2.

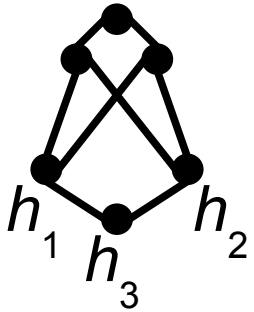
Règles de construction des générateurs de niveau *k*+1 à partir de ceux de niveau *k* ?

Construction des générateurs

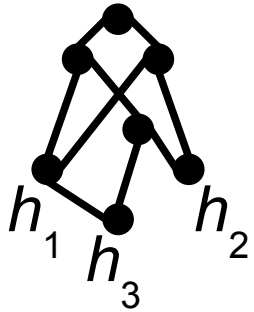
Construction des générateurs de niveau k+1 à partir de ceux de niveau k :



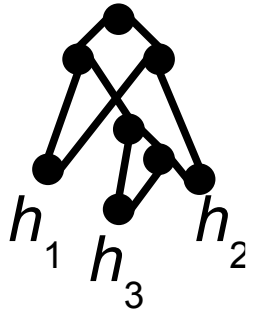
Règle R_1 :



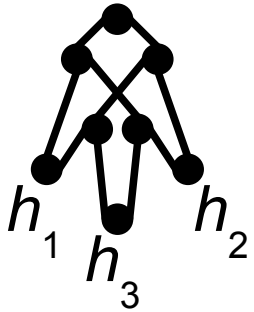
$R_1(N,h_1,h_2)$



$R_1(N,h_1,e_2)$



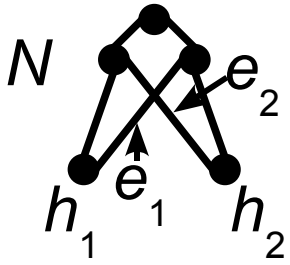
$R_1(N,e_2,e_2)$



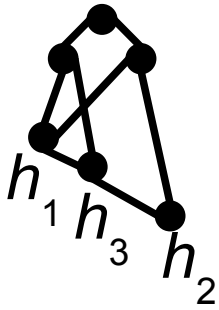
$R_1(N,e_1,e_2)$

Construction des générateurs

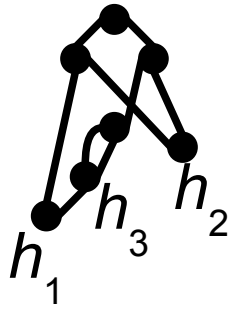
Construction des générateurs de niveau k+1 à partir de ceux de niveau k :



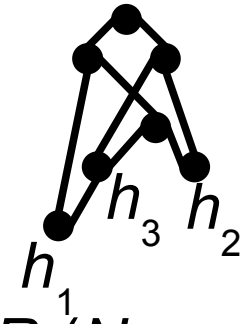
Règle R_2 :



$R_2(N,h_1,e_2)$



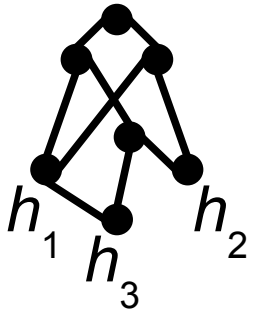
$R_2(N,e_1,e_1)$



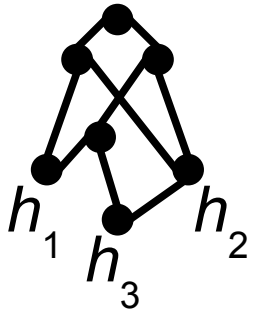
$R_2(N,e_2,e_1)$

Construction des générateurs

Problème !
Certains des générateurs de niveau k+1 obtenus depuis ceux de niveau k sont isomorphes !



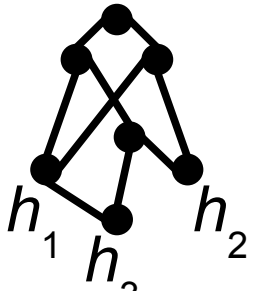
$R_1(2b,h_1,e_2)$



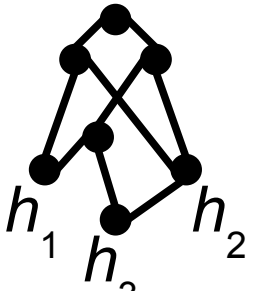
$R_1(2b,h_2,e_1)$

Construction des générateurs

Problème !
Certains des générateurs de niveau k+1 obtenus depuis ceux de niveau k sont isomorphes !



$R_1(N,h_1,e_2)$



$R_1(N,h_2,e_1)$

→ comptage difficile !

Plan

- Classification et réseaux phylogénétiques
- Décomposition des réseaux de niveau *k*
- Construction des générateurs de niveau *k*
- **Nombre de générateurs de niveau *k***
- Simulation de réseaux de niveau *k*

Borne supérieure

R_1 et R_2 peuvent être appliquées sur au plus toutes les paires de côtés.
Un générateur de niveau k a au plus $5k$ côtés :

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

Borne supérieure :

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

Corollaire théorique :
Il existe un algorithme polynomial pour construire l'ensemble des générateurs de niveau k+1 depuis l'ensemble des générateurs de niveau k.

Corollaire pratique :

$$g_4 < 28350$$

→ on peut énumérer tous les générateurs de niveau 4.

Nombre de générateurs de niveau *k*

On peut énumérer tous les générateurs de niveau 4.

Isomorphisme de graphes de degré maximal borné : polynomial

(Luks, FOCS 1980)

Algorithme pratique ?
Simple algorithme exponentiel de backtrack suffisant pour le niveau 4 :

parcourir les deux graphes en parallèle depuis leur racine et identifier leurs sommets : $O(n2^{n-h})$

→ $g_4 = 1993$
→ $g_5 > 71000$

Nombre de générateurs de niveau *k*

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.

Borne inférieure

Borne inférieure :

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

Il y a un **nombre exponentiel** de générateurs !

Idée :
Coder tout nombre entre 0 et $2^{k-1}-1$ par un générateur de niveau k.

Borne inférieure

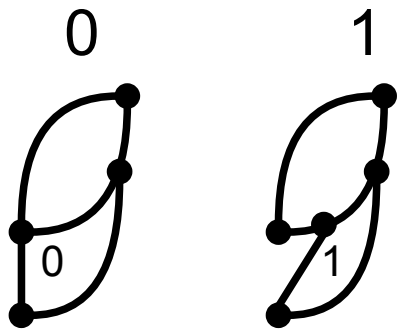
Borne inférieure :

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

Il y a un **nombre exponentiel** de générateurs !

Idée :
Coder tout nombre entre 0 et $2^{k-1}-1$ par un générateur de niveau k.

$k = 1$



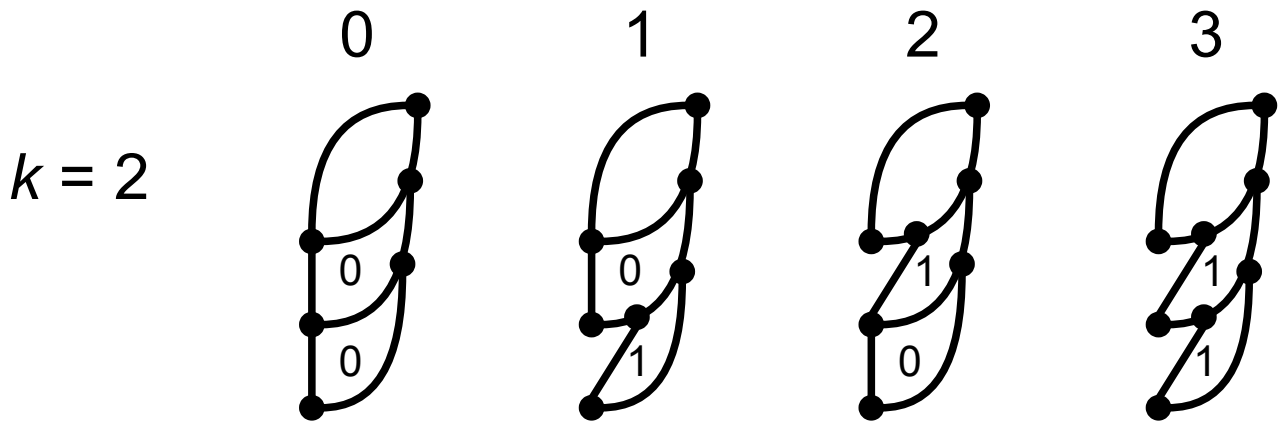
Borne inférieure

Borne inférieure :

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

Il y a un **nombre exponentiel** de générateurs !

Idée :
Coder tout nombre entre 0 et $2^{k-1}-1$ par un générateur de niveau k.



Plan

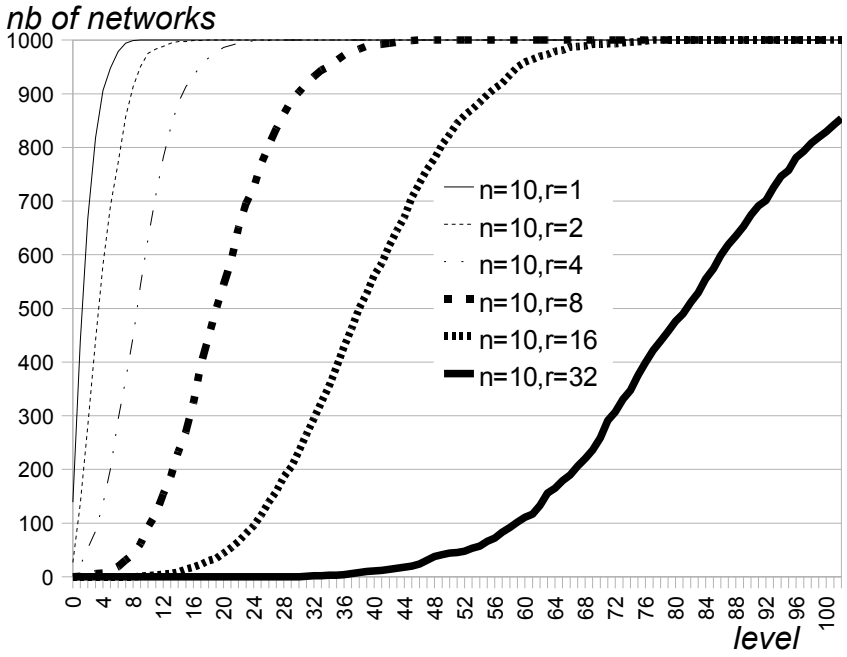
- Classification et réseaux phylogénétiques
- Décomposition des réseaux de niveau *k*
- Construction des générateurs de niveau *k*
- Nombre de générateurs de niveau *k*
- **Simulation de réseaux de niveau *k***

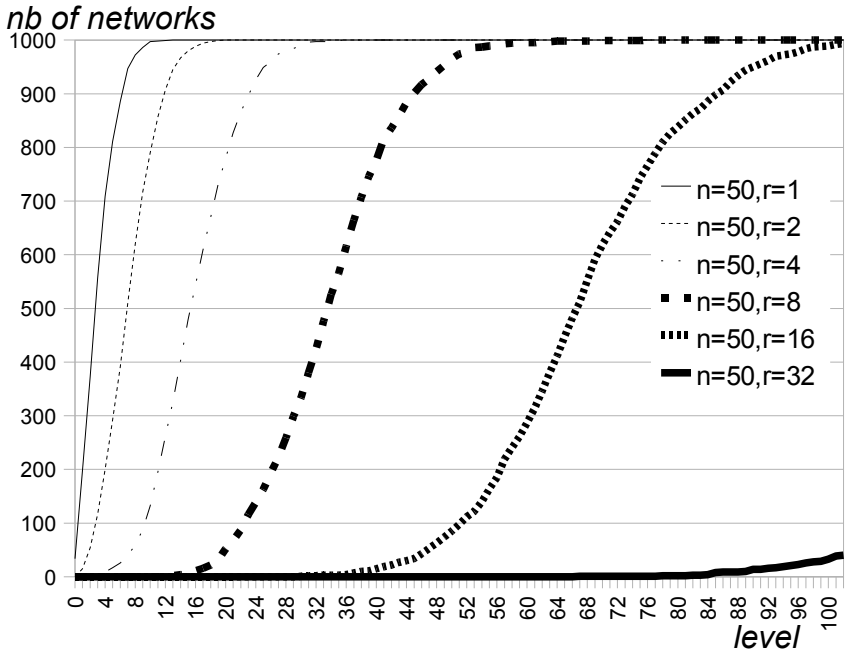
Simulations de réseaux de niveau *k*

Simulation de 1000 réseaux phylogénétiques selon le modèle coalescent avec recombinaison.

Arenas, Valiente, Posada 2008
Program Recodon

Combien sont de niveau 1, 2, 3 ?

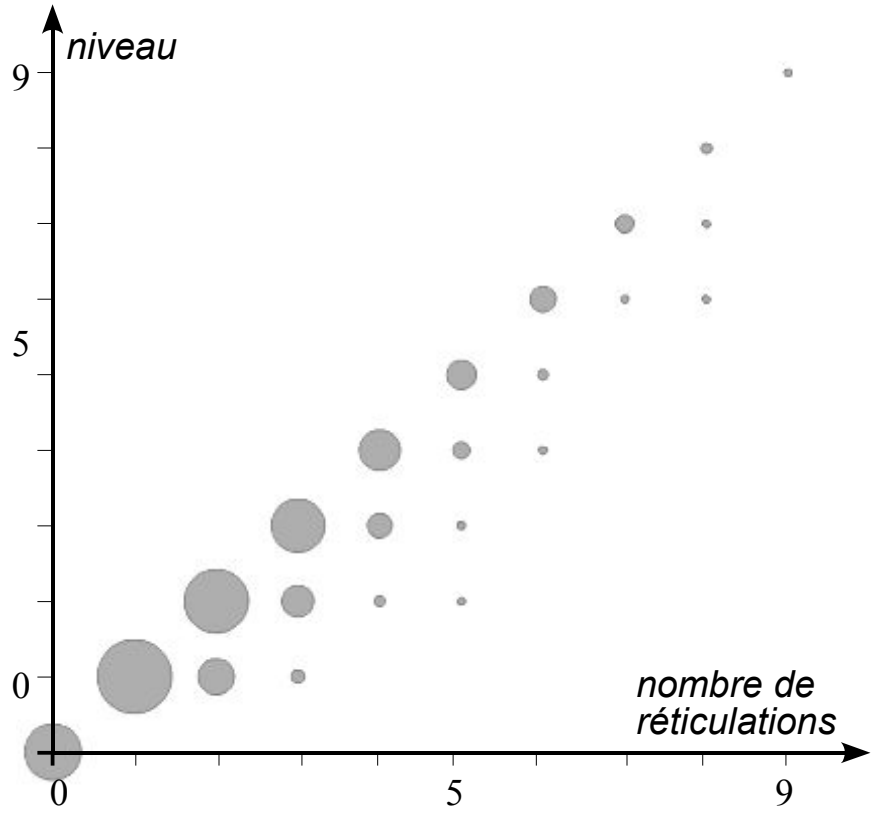




Simulations de réseaux de niveau *k*

Simulation de 1000 réseaux phylogénétiques selon le modèle coalescent avec recombinaison.

Lien entre le niveau et le nombre de réticulations :



Bilan sur les réseaux de niveau *k*

Avantages :

- structure naturelle pour tous les réseaux phylogénétiques explicites
- structure globalement arborée utilisée algorithmiquement : reconstruction à partir de triplets, clusters, quadruplets

(Jansson Nguyen Song 2006, Kelk & al 2008, To & Habib 2009, van Iersel Huson & al 2010, Gambette Berry & Paul 2010)

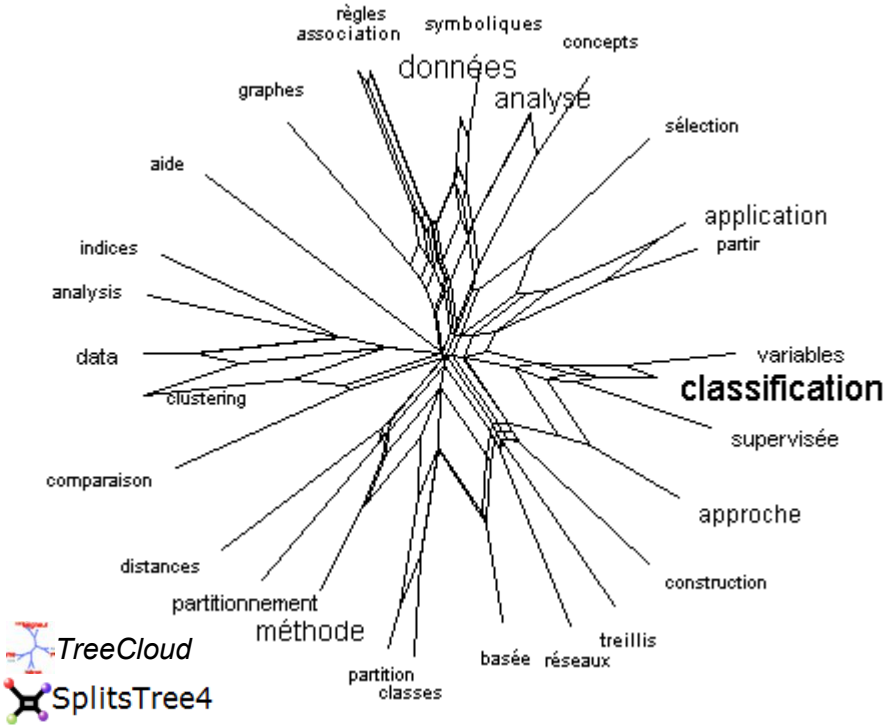
- motifs de graphes finis pour représenter les blobs : les générateurs

Limites :

- nombre exponentiel de générateurs
- structure complexe des générateurs
- quand les réticulations ne sont pas locales, le niveau n'aide pas

Questions ?

Merci de votre attention !



Réseau NeighborNet et réticulogramme des mots les plus fréquents dans les titres de contributions acceptées à SFC'07&09, distance de cooccurrence hyperlex.