Structure des réseaux phylogénétiques de niveau borné

Philippe Gambette, Vincent Berry, Christophe Paul

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier.
C.N.R.S., Université Montpellier 2.
161 rue Ada, 34392 Montpellier Cedex 5 France

RÉSUMÉ. *Les* réseaux phylogénétiques *généralisent les arbres phylogénétiques en représentant des échanges de matériel génétique entre espèces par des branches qui se rejoignent pour former des parties réticulées. Le* niveau *est un paramètre introduit sur les réseaux phylogénétiques enracinés pour décrire la complexité de leur structure par rapport à un arbre [JAN 04]. Des algorithmes polynomiaux ont récemment été proposés pour reconstruire un réseau de niveau borné compatible avec un ensemble de triplets fournis en entrée [IER 08, TO 09]. Nous étudions la structure d'un réseau de niveau borné pour montrer qu'il peut être décomposé en un arbre de générateurs choisis parmi un ensemble fini. Nous nous intéressons alors à la pertinence du paramètre de niveau dans le cadre d'un modèle d'évolution avec recombinaisons : le modèle coalescent.*

MOTS-CLÉS : *Combinatoire, Décomposition, Graphe, Réseau phylogénétique.*

1. Introduction et définitions

Un *arbre phylogénétique* est un arbre binaire enraciné avec des arcs (orientés, donc) et des feuilles étiquetées bijectivement par un ensemble X de *taxons*, qui représentent le plus souvent des espèces ou des gènes. Un *réseau phylogénétique explicite* est une généralisation d'arbre phylogénétique qui permet de prendre en compte les échanges de matériel génétique entre espèces, qui sont très fréquents entre les bactéries [DOO 99] mais aussi présents chez les végétaux ou même les animaux [HUB 55]. Ces échanges peuvent correspondre à divers événements biologiques : hybridation, recombinaison, transferts horizontaux...

On peut définir formellement un réseau phylogénétique explicite comme un multigraphe orienté acyclique, contenant : exactement un sommet a degré entrant 0 et degré sortant 2 (la *racine*) ; des sommets de degré entrant 1 et de degré sortant 2 (*sommets de spéciation*) ; des sommets de degré entrant 2 et de degré sortant au plus 1 (*sommets hybrides*) ; des sommets étiquetés bijectivement par un ensemble X de taxons, de degré entrant 1 et de degré sortant 0 (*feuilles*). Dans la Figure 2(a) est représenté un réseau phylogénétique explicite N de racine ρ et d'ensemble de taxons $X = \{a, b, c, d, e, f, g, h, i\}$. Les sommets h_i sont des sommets hybrides et ceux non étiquetés sont des sommets de spéciation.

Notons que parler de multigraphe, c'est à dire autoriser la présence de plusieurs arcs entre deux sommets, est un détail technique qui permet la présence de cycles à deux sommets dans le réseau phylogénétique, comme celui contenant h_1 en figure 2(a).

Un graphe orienté est dit *biconnexe* s'il ne contient aucun sommet d'articulation (dont la suppression déconnecte le graphe). Une *composante biconnexe* (ou *blob*) d'un réseau phylogénétique N est un sous-graphe biconnexe maximal de N. Pour tout arc (u, v) de N, on appelle u un père de v, et v un fils de u.

Un réseau phylogénétique explicite est dit de *niveau* k [JAN 04] si toute composante biconnexe contient au plus k sommets hybrides. Un réseau de niveau k qui n'est pas de niveau k-1 est dit strictement de niveau k. Par exemple, dans la Figure 2(a), la composante biconnexe de N qui contient le plus de sommets hybrides est située dans la zone grise (elle contient h_3 et h_4), donc N est strictement de niveau 2.

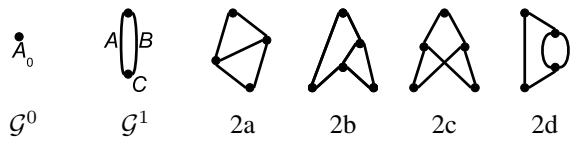


FIGURE 1. *Le générateur $\mathcal{G}^0$ de niveau 0 (a), le générateur $\mathcal{G}^1$ de niveau 1 (b), et les générateurs de niveau 2, appelés 2a, 2b, 2c et 2d [IER 08]. Tous les arcs sont dirigés vers le bas (l'orientation n'est pas indiquée pour un souci de lisibilité.)*

Ce paramètre exprime à quel point le réseau est proche d'un arbre : un réseau de niveau 0 est un arbre phylogénétique, un réseau de niveau 1 est communément appelé *galled tree*. De nombreux problèmes NP-complets peuvent être résolus en temps polynomial sur ces classes de réseaux phylogénétiques [GAM], ce qui motive l'étude des niveaux supérieurs.

En section 2, on étudie la structure de ces réseaux en montrant qu'ils peuvent avoir une grande complexité intrinsèque. Nous considérons ensuite, en section 3, un ensemble de réseaux simulés selon le modèle coalescent avec recombinaison pour montrer que dans ce contexte, les réseaux ont un niveau élevé, ce qui réduit l'application pratique des algorithmes de reconstruction existants.

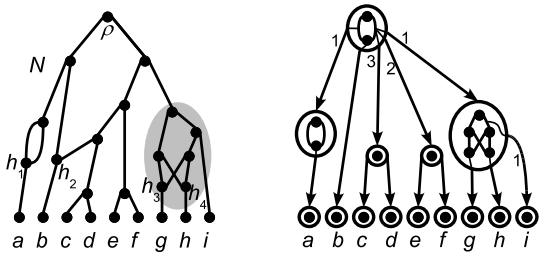
2. Décomposition des réseaux de niveau k

Définition 1 ([IER 08]) *Un* générateur de niveau k *est un réseau phylogénétique biconnexe strictement de niveau k (voir figure 1).*

Ces générateurs ont été introduits à l'origine dans le contexte d'une sous-classe de réseaux phylogénétiques dits *simples*, contenant une seule composante biconnexe. Nous montrons dans le théorème suivant qu'ils permettent de décomposer tout réseau de niveau k en un arbre de générateurs.

Théorème 1 (décomposition des réseaux de niveau k) *Tout réseau N de niveau k peut être décomposé de façon unique en un arbre de générateurs de niveau au plus k.*

L'arbre de décomposition en générateurs d'un réseau de niveau k est illustré en figure 2(b). Il correspond globalement à l'arbre de décomposition en composantes biconnexes du graphe, avec un intérêt supplémentaire dans notre cas : pouvoir étiqueter les noeuds de l'arbre de décomposition par un générateur extrait d'un ensemble fini (à niveau fixé).



(a) (b)

FIGURE 2. *Un réseau phylogénétique de niveau 2 (a) et son arbre de décomposition en générateurs (b) : la numérotation sur les arcs de l'arbre de décomposition indique dans quel ordre les générateurs sont attachés aux arcs du générateur du noeud père.*

Ce théorème de décomposition en générateurs demande donc une étude un peu plus précise de la structure des générateurs. Nous fournissons quelques propriétés sur leur taille, leur nombre, ainsi qu'un algorithme pour construire tous les générateurs de niveau k à partir des générateurs de niveau $k-1$.

Propriété 1 *Pour $k \geq 1$, un générateur de niveau k a au plus $3k-1$ sommets et $4k-2$ arcs.*

Propriété 2 *Le nombre g_k de générateurs de niveau k est compris entre 2^{k-1} et $k!^2 50^k$.*

Ces bornes très peu fines servent essentiellement à montrer que le nombre de générateurs est exponentiel en fonction du niveau. Ceci indique que la vision d'un réseau phylogénétique comme un arbre de blobs cache derrière l'apparente simplicité de l'arbre une grande complexité de structure à l'intérieur des blobs. Toutefois, elles permettent aussi de noter qu'il semble possible de construire automatiquement l'ensemble des générateurs de niveau 4, alors que jusqu'alors seuls ceux de niveau au plus 3 avaient été construits [KEL].

Théorème 2 *Un algorithme polynomial permet de construire l'ensemble de tous les générateurs de niveau $k+1$ à partir de l'ensemble S_k^* de tous les générateurs strictement de niveau k fourni en entrée.*

A la base de cet algorithme, et des deux propositions précédentes, se trouvent deux règles permettant d'attacher un nouveau sommet hybride à l'intérieur d'un générateur de niveau k. L'algorithme consiste donc à construire progressivement l'ensemble S_{k+1}^* en considérant à tour de rôle chaque générateur de l'ensemble S_k^*, et en appliquant une règle d'insertion d'un nouveau sommet hybride, et en vérifiant si le générateur de niveau $k+1$ ainsi créé est isomorphe à un des générateurs déjà ajoutés à S_{k+1}^*. Il faut noter que bien que la complexité du test d'isomorphisme de graphes soit encore indéterminée, nous pouvons le faire théoriquement en temps polynomial car nous travaillons sur des graphes orientés de degré maximum 3 [LUK 82, MIL 77]. En fait, l'algorithme de Luks est peu utilisable en pratique, et nous utilisons un algorithme exact exponentiel dans notre implémentation disponible à l'adresse `http://www.lirmm.fr/~gambette/ProgGenerators.php`.

Cette implémentation a permis de déterminer qu'il existait 1993 générateurs de niveau 4. On a ainsi pu vérifier que la séquence du nombre de générateurs de niveau k, 1,4,65,1993, n'était pas présente dans l'Encyclopédie en ligne des séquences d'entiers [SLO 08], alors que deux séquences de cette base de données contenaient 1,4,65.

3. Niveau de réseaux simulés

Arenas, Valiente et Posada ont étudié les propriétés de réseaux phylogénétiques simulés selon le modèle coalescent avec recombinaison [ARE 08], en mesurant la proportion parmi ces réseaux de ceux appartenant à certaines sous-classes, en particulier les arbres, et les réseaux "galled tree", c'est à dire les réseaux de niveau 0 et 1. Nous avons prolongé leur étude en calculant le niveau de tous les réseaux phylogénétiques générés par leur simulation qui a utilisé le programme Recodon [ARE 07]. L'implémentation en Java d'un algorithme basique de décomposition en composantes biconnexes pour calculer le niveau est également disponible à l'adresse `http://www.lirmm.fr/~gambette/ProgGenerators.php`.

Pour de petites valeurs du niveau, les résultats obtenus sont réunis dans la Table 1. Nous observons que les réseaux phylogénétiques avec un petit niveau, comme les autres restrictions étudiées dans la référence [ARE 08], ne couvrent qu'une portion réduite des réseaux phylogénétiques correspondant au modèle coalescent avec de forts taux de recombinaison. En fait, les réseaux simulés selon ce modèle n'ont pas vraiment la structure arborée exprimée dans le Théorème 1, mais sont le plus souvent constitués d'une grosse composante biconnexe qui contient tous les sommets hybrides. Ce phénomène apparaît même pour de faibles taux de recombinaison.

Ainsi, dans ce contexte, de nouvelles structures et techniques algorithmiques doivent être étudiées. Mentionnons toutefois que ce modèle ne convient pas pour décrire tous les cas d'évolution réticulée, et que d'autres peuvent être plus appropriés, comme celui qui insère des transferts horizontaux selon une loi de Poisson [GAL 07], ou ceux utilisés pour la simulation de réseaux phylogénétiques dans NetGen [MOR 06].

r	arbre	niveau 1	niveau 2	niveau 3	niveau 4	niveau 5
0	1000	1000	1000	1000	1000	1000
1	139	440	667	818	906	948
2	27	137	281	440	582	691
4	1	21	53	85	136	201
8	0	1	1	6	7	12
16	0	0	0	0	0	0

TABLE 1. *Nombre de réseaux simulés selon le modèle coalescent avec recombinaison sur 10 feuilles, ayant niveau 0, 1, 2, 3, 4, 5, en fonction du taux de recombinaison $r = 0, 1, 2, 4, 8, 16$.*

4. Conclusion

Devant l'engouement récent pour les réseaux de niveau k, qui permettent d'obtenir des algorithmes efficaces pour la reconstruction phylogénétique de réseaux à partir de triplets, nous avons présenté des résultats qui permettent de mieux comprendre ces objets : simples par la structure arborée qui apparaît, complexes à l'intérieur des parties réticulées, puisqu'un choix exponentiel de structures y est possible, en fonction du niveau.

La validation des méthodes de reconstruction de réseaux de niveau k sur des données biologiques est en cours, et il est intéressant de voir si les nuances théoriques que nous apportons à propos de leur utilisation seront confirmées en pratique. Ces résultats relancent aussi l'intérêt pour d'autres paramètres sur les réseaux qui permettraient d'obtenir des algorithmes rapides, et quelques pistes semblent déjà prometteuses dans cette optique.

5. Bibliographie

[ARE 07] ARENAS M., POSADA D., Recodon : coalescent simulation of coding DNA sequences with recombination, migration and demography, *BMC Bioinformatics*, vol. 8, 2007, page 458.

[ARE 08] ARENAS M., VALIENTE G., POSADA D., Characterization of Reticulate Networks based on the Coalescent, *Molecular Biology and Evolution*, vol. 25, 2008, p. 2517-2520.

[DOO 99] DOOLITTLE W., Phylogenetic Classification and the Universal Tree, *Science*, vol. 284, 1999, p. 2124-2128.

[GAL 07] GALTIER N., A Model of Horizontal Gene Transfer and the Bacterial Phylogeny Problem, *Systematic Biology*, vol. 56, 2007, p. 633-642.

[GAM] GAMBETTE P., Who is Who in Phylogenetic Networks : Articles, Authors and Programs, `http://www.lirmm.fr/~gambette/PhylogeneticNetworks`.

[HUB 55] HUBBS C., Hybridization Between Fish Species in Nature, *Systematic Zoology*, vol. 4, 1955, p. 1-20.

[IER 08] VAN IERSEL L., KEIJSPER J., KELK S., STOUGIE L., HAGEN F., BOEKHOUT T., Constructing Level-2 Phylogenetic Networks from Triplets, *RECOMB'08*, vol. 4955 de *LNCS*, Springer Verlag, 2008, p. 450-462.

[JAN 04] JANSSON J., SUNG W.-K., Inferring a Level-1 Phylogenetic Network from a Dense Set of Rooted Triplets, *COCOON'04*, vol. 3106 de *LNCS*, Springer Verlag, 2004, p. 462-471.

[KEL] KELK S., `http://homepages.cwi.nl/~kelk/lev3gen/`.

[LUK 82] LUKS E., Isomorphism of Graphs of Bounded Valence Can Be Tested in Polynomial Time, *Journal of Computer and System Sciences*, vol. 25, n^o 1, 1982, p. 42-65.

[MIL 77] MILLER G., Graph Isomorphism, General Remarks, *STOC'77*, 1977, p. 143-150.

[MOR 06] MORIN M., MORET B., NETGEN : Generating Phylogenetic Networks with Diploid Hybrids, *Bioinformatics*, vol. 22, n^o 15, 2006, p. 1921-1923.

[SLO 08] SLOANE N., The On-Line Encyclopedia of Integer Sequences, 2008, Published electronically at `http://www.research.att.com/~njas/sequences/`.

[TO 09] TO T.-H., HABIB M., Level-k Phylogenetic Network can be Constructed from a Dense Triplet Set in Polynomial Time, *CPM'09*, 2009, à paraître.