RECONSTRUCTION COMBINATOIRE DE RÉSEAUX PHYLOGÉNÉTIQUES

Philippe GAMBETTE

Université Paris Est, Marne-la-Vallée
LIGM, cité Descartes, bât. Copernic, 5, boulevard Descartes
Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2
philippe.gambette@gmail.com

Abstract. – Phylogenetic networks generalize the tree model of Evolution, by adding some edges between branches inside the tree of life, to reflect exchanges of genetic material. Lots of combinatorial approaches have been designed to reconstruct them from data extracted from a set of contradictory gene trees. We will describe reconstruction methods from triplets, quartets and clusters, based on graph decompositions, which rely on an underlying tree structure. Then we will present some limits of these combinatorial methods (complexity explosion, restrictions on the network model, ambiguity in the data). It encourages the development of new approaches mixing combinatorics and statistics.

INTRODUCTION

La prise en compte des transferts de matériel génétiques entre espèces co-existantes par des processus d'hybridation ou de transferts horizontaux a conduit à l'apparition d'un modèle plus complexe que l'arbre : le réseau phylogénétique. Les divers travaux menés sur les arbres phylogénétiques ont alors été étendus, autant que possible, aux réseaux. Aux études sur les distances d'arbres et les ultramétriques (qui supposent que toutes les feuilles se trouvent à une même distance du centre de l'arbre) ont succédé des travaux sur la décomposition des métriques («*split decomposition*» dans BANDELT & DRESS, 1992 ; «*block decomposition*» dans DRESS *et al.*, 2008), ou des réalisations optimales (c'est-à-dire de la recherche d'un graphe minimal qui reflète une matrice de distance, voir ALTHÖFER, 1986 et LESSER, 2007). À cet ensemble de résultats théoriques se sont ajoutées les premières implémentations de méthodes rapides de construction de réseaux phylogénétiques : SplitsTree pour la split decomposition (HUSON, 1998), T-Rex pour les réticulogrammes (une heuristique qui ajoute des arêtes-«raccourcis» dans un arbre de départ pour représenter au mieux la matrice de distance fournie en entrée), comme détaillé dans MAKARENKOV (2001), ou encore deux méthodes dérivées de Neighbor-Joining (SAITOU & NEI, 1987) : NeighborNet (BRYANT & MOULTON, 2002) et plus récemment ConstNJ (MATSEN, 2010). Quelques premières procédures de reconstruction basées sur la parcimonie ou la vraisemblance ont également été proposées. Toutefois, les approches exactes dans ce contexte sont limitées par les temps de calcul, et les méthodes approchées présentent aussi des inconvénients analysés dans VELASCO & SOBER (2010).

MÉTHODES COMBINATOIRES

Les motivations

Dans ce qui suit, nous nous intéresserons donc uniquement aux méthodes combinatoires, c'est-à-dire basées sur l'étude d'ensemble finis d'objets mathématiques. Cette approche peut se justifier par plusieurs arguments.

Tout d'abord, l'hypothèse d'une structure arborée sous-jacente, même en présence de réticulations, incite à exploiter les propriétés mathématiques et algorithmiques de la structure d'arbre. Il existe en effet de nombreux théorèmes à propos des *X*-arbres (c'est-à-dire dont les feuilles sont étiquetées bijectivement par un ensemble *X* de taxons). Par exemple, une caractérisation des ensembles de quadruplets (des *X*-arbres à 4 feuilles) ou de splits (bipartitions des taxons) compatibles avec un *X*-arbre non enraciné est fournie par BANDELT & DRESS (1986). Des résultats similaires existent avec les clades (c'est-à-dire des ensembles d'espèces apparaissant dans un même sous-arbre) ou les triplets (*X*-arbres à trois feuilles) pour les arbres enracinés. Outre ces caractérisations, qui se traduisent parfois en algorithmes, les problèmes de consensus d'arbres ou de définition d'une distance entre deux arbres font aussi intervenir la connaissance combinatoire de cette structure.

C'est tout naturellement que, depuis les années 1990 et la prise de conscience sur l'importance des événements de réticulation dans le contexte de la reconstruction phylogénétique, ces résultats ont été étendus autant que possible des arbres aux réseaux, ou plutôt à des sous-classes de réseaux qui présentent certaines restrictions (réseaux planaires, réguliers, normaux, de niveau borné, « tree-child », « tree-sibling », pyramides, hiérarchies faibles… voir GAMBETTE, 2008 pour une liste exhaustive). La généralisation étant alors plus guidée par le modèle mathématique choisi que par le modèle biologique d'évolution, on a pu assister au développement de méthodes de reconstruction de réseaux abstraits. Ces réseaux, en particulier les réseaux de bipartitions ou split networks, qui ont bénéficié de la popularité du logiciel SplitsTree (HUSON, 1998), ne peuvent être interprétés directement comme un ensemble d'événements biologiques, mais permettent d'avoir un aperçu visuel de la complexité des réticulations, et de leur localisation approximative, comme on peut le voir en figure 1 (i). Cette démarche a aussi conduit à l'apparition de restrictions n'ayant pas de sens du point de vue biologique, mais utiles visuellement ou algorithmiquement, comme la planarité.

Un autre argument en faveur des méthodes combinatoires provient de l'explosion des données biologiques disponibles. Il devient impossible d'utiliser des algorithmes sophistiqués les considérant toutes en même temps. C'est pourquoi on assiste de plus en plus à des reconstructions de phylomes, c'est-à-dire d'ensembles d'arbres phylogénétiques, chaque arbre correspondant à un ensemble de gènes homologues (HUERTA-CEPAS *et al.*, 2007). Ces données sont parfois accessibles librement par internet, comme la base HOGENOM (http://pbil.univ-lyon1.fr/databases/hogenom.php), qui concerne 513 espèces et contient plus de 70 000 arbres de gènes.

Le modèle d'arbre en filigrane

L'arbre comme modèle de transmission des gènes

L'arbre est à la base de plusieurs méthodes de réseaux phylogénétiques. Tout d'abord, on considère le plus souvent que dans le cas de transferts horizontaux, ce sont des gènes entiers qui sont transmis. Ainsi, l'histoire de l'évolution d'un ensemble de gènes homologues peut se représenter de manière arborée. C'est cette remarque qui est à la base de la définition des clades souples dans un réseau : on considère comme un clade non pas l'ensemble de toutes les feuilles situées sous un nœud du réseau (comme on le fait pour un arbre), mais l'ensemble des feuilles d'un *X*-arbre inclus dans le réseau. Ainsi, dans le deuxième réseau de la figure 4 (i), en considérant un des deux arbres inclus dans le réseau, on constate que $\{a, b\}$ est un clade souple, et en considérant l'autre, on constate que $\{b, c\}$ est un clade souple. On peut alors voir un réseau comme l'ensemble d'arbres de gènes qu'il contient. Ceci conduit aux problématiques de déterminer s'il est possible de retrouver un arbre précis dans un réseau (KANJ *et al.*, 2008), ou de reconstruire le réseau le plus parcimonieux (avec le moins de nœuds de réticulation) contenant un certain nombre d'arbres (BORDEWICH & SEMPLE, 2007). Ces deux problèmes sont difficiles (NP-complets, c'est-à-dire qu'il est peut probable qu'un algorithme puisse les résoudre de façon exacte en temps polynomial), en particulier le second l'est même dans le cas où on ne souhaite réunir que deux arbres, ce qui incite

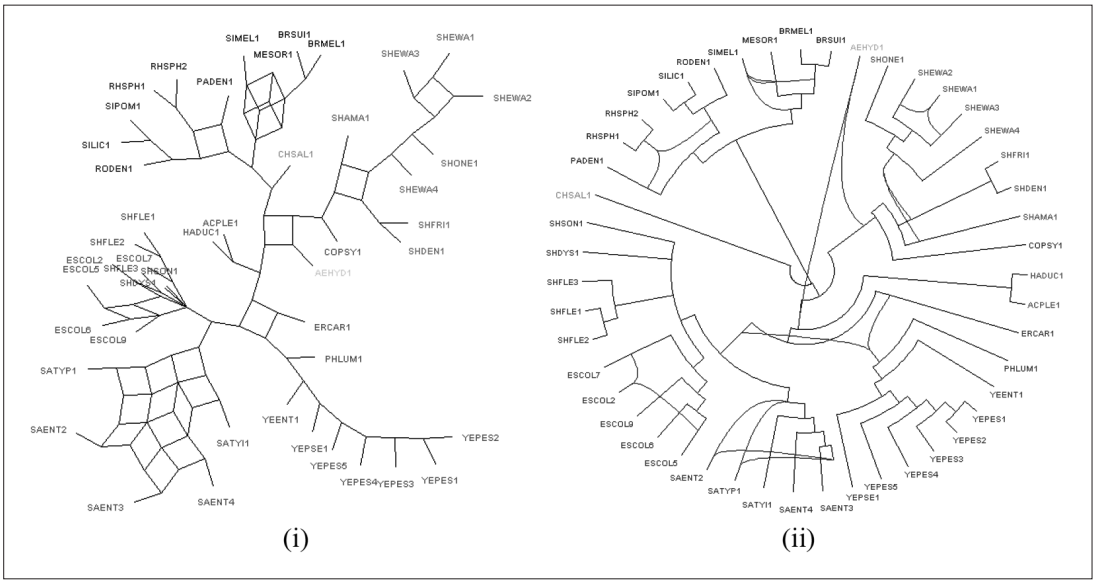

Figure 1. (i) Un split network, exemple de réseau phylogénétique abstrait ; et (ii) un réseau phylogénétique explicite construits respectivement par SplitsTree et Dendroscope, à partir de bipartitions et de clades apparaissant dans 20 % de 16 arbres de gènes de la base HOGENOM concernant 46 espèces d'alpha et gamma-protéobactéries

à se pencher sur des méthodes utilisant des données intermédiaires comme les clades ou les triplets, sur lesquelles nous reviendrons plus ci-dessous.

L'arbre comme modèle de simplicité

Les méthodes combinatoires se fondent aussi sur le principe que les réseaux à reconstruire ont une structure proche de celle d'un arbre, et fonctionnent d'autant plus vite que c'est le cas. La structure de réseau phylogénétique la plus simple est celle qui ne contient qu'un seul nœud de réticulation (c'est-à-dire un nœud ayant deux parents), et une formule de dénombrement est par exemple fournie dans Semple & Steel (2006). Toutefois, c'est la plupart du temps la reconstruction d'un « galled tree » qui est le premier objectif. Ces réseaux phylogénétiques sont tels que tous les cycles sont disjoints et ont été introduits dans Ma *et al.* (1998). Cette propriété très forte permet généralement d'obtenir des algorithmes rapides (Jansson *et al.*, 2006), voire des théorèmes de caractérisation (Song, 2006, Gupta *et al.*, 2006).

Elle a ensuite été généralisée avec la définition d'un paramètre de niveau pour les réseaux enracinés (Jansson & Sung, 2006) : le niveau est le nombre maximum de nœuds de réticulation par blob, un blob étant une partie entièrement réticulée du réseau aphylogénétique (c'est-à-dire qui ne contient pas d'arête dont la suppression déconnecte le réseau, fig. 2). Ainsi, un réseau de niveau 0 est un arbre, un de niveau 1 est un « galled tree » (les blobs sont

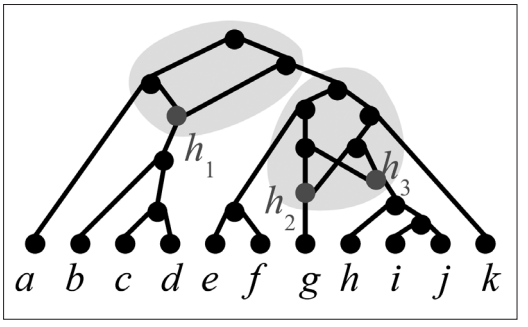

Figure 2. Un réseau phylogénétique de niveau 2 : les blobs sont soulignés par la couleur grise, le premier ne contient qu'un nœud de réticulation, h_1, et le second en contient 2, h_2 et h_3.

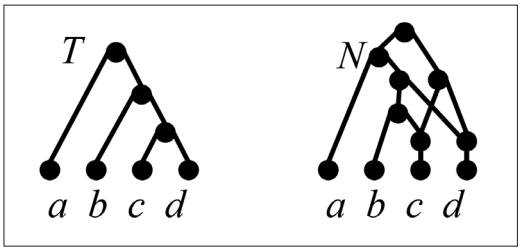

Figure 3. Un arbre *T* et un réseau phylogénétique *N* qui contient tous les triplets et clusters de *T*, mais pas *T* lui-même.

alors de simples cycles disjoints), et des algorithmes existent pour reconstruire un réseau de niveau *k* compatible avec un ensemble de triplets (VAN IERSEL *et al.*, 2008, TO & HABIB, 2009), s'il en existe un. Ces algorithmes s'exécutent en temps exponentiel par rapport à *k*, et peuvent être utilisés en pratique uniquement pour de petites valeurs du niveau, dans le cas où l'ensemble de triplets est dense, c'est-à-dire qu'il contient une (ou plusieurs) topologies sur tout ensemble de trois feuilles. Une partie de ces algorithmes ont été généralisés aux quadruplets et aux réseaux non enracinés (GAMBETTE *et al.*, 2011), pour éviter les erreurs dues à un mauvais enracinement : il est possible de reconstruire un réseau non enraciné de niveau 1 à partir de son ensemble de quadruplets. Cette propriété ne semble toutefois pas utilisable directement en pratique : il est improbable que les données biologiques conduisent à déterminer un ensemble de quadruplets sans erreur ni omission.

Des méthodes existent aussi pour reconstruire des réseaux compatibles avec un ensemble de clades souples. Par exemple, celle implémentée dans Dendroscope (HUSON *et al.*, 2007) permet de trouver un réseau phylogénétique à structure restreinte (« galled network »), comme celui de la figure 1 (ii). La construction de ce réseau se fait en deux étapes : premièrement, identifier l'ensemble maximum de taxons tels que les clades concernant cet ensemble sont alors compatibles avec un arbre ; deuxièmement, reconstruire cet arbre et y attacher les taxons manquants en utilisant le minimum de nœuds de réticulation (HUSON *et al.*, 2009). Bien que chacune de ces deux étapes d'optimisation corresponde à un problème théoriquement difficile (NP-complet), elles fonctionnent très rapidement sur des exemples d'une dizaine d'arbre concernant une centaine de taxons.

LES LIMITES DES APPROCHES COMBINATOIRES

Des propriétés contre-intuitives par rapport aux arbres

Certaines propriétés naturelles dans les arbres ne s'étendent pas aux réseaux et peuvent expliquer le saut de complexité d'un modèle à l'autre. Par exemple, l'unicité du plus petit ancêtre commun de deux taxons n'est pas assurée, *c* et *d* ayant deux plus petits ancêtres communs dans le réseau phylogénétique *N* de la figure 3. On peut également être surpris par le fait que même si un réseau contient l'ensemble de tous les triplets, ou tous les clades, d'un arbre, il ne contient pas nécessairement cet arbre, comme illustré en figure 3. Ainsi, si l'on reconstruit le réseau phylogénétique *N'* le plus parcimonieux compatible avec tous les triplets (ou tous les clades) d'un ensemble *A* d'arbres, on ne peut affirmer qu'il contiendra tous les arbres de *A*. Si ce n'est pas le cas, on peut juste assurer que le réseau phylogénétique *N* qui contient tous les arbres de *A* est plus complexe (moins parcimonieux) que *N'*, puisqu'il doit également contenir tous les triplets (ou clades) des arbres de *A*, et que *N'* fournit donc une borne inférieure de la complexité du réseau que l'on veut reconstruire à partir de *A*. On peut également considérer que le réseau *N'* a un intérêt en tant que tel quand les données de triplets ou de clades considérées en entrée des algorithmes de reconstruction proviennent d'un consensus d'arbres suivi d'un filtrage (pour garder par exemple uniquement ceux qui apparaissent dans au moins 30 % des arbres de *A*).

Une explosion de complexité pas toujours maîtrisée

La proximité du réseau à reconstruire avec un arbre est un moyen, nous l'avons vu, d'améliorer la rapidité des algorithmes de reconstruction de réseaux phylogénétiques, en traitant indépendamment chaque

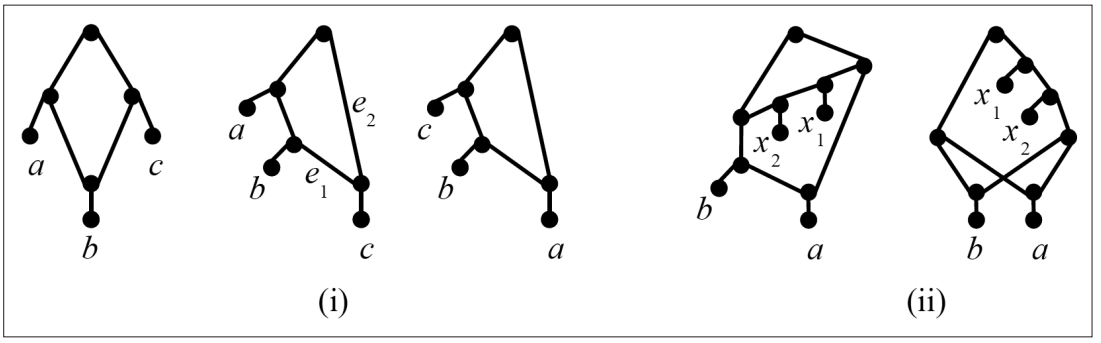

Figure 4. (i) Trois réseaux phylogénétiques de niveau 1 ayant exactement le même ensemble de triplets $\{a|bc, c|ab\}$ et de clades souples $\{\{a\},\{b\},\{c\},\{a,b\},\{b,c\},\{a,b,c\}\}$; et (ii) deux réseaux phylogénétiques de niveau 2 ayant exactement le même ensemble de triplets $\{a|x_1b, b|x_1a, x_1|ab, a|x_2b, b|x_2a, x_2|ab, x_1|x_2a, a|x_1x_2, x_1|x_2b, b|x_1x_2\}$.

partie non arborée par exemple. Ce point de vue permet par exemple de considérer tout réseau de niveau k comme un arbre reliant un nombre fini de générateurs, c'est-à-dire de petits motifs de graphes (Gambette *et al.*, 2009). Nous avons toutefois prouvé que cet ensemble de générateurs était de taille exponentielle par rapport à k. Ainsi, la décomposition en arbre simplifie le problème mais pas suffisamment. De plus, par une étude de simulation, nous avons montré que si les échanges horizontaux de matériel génétique peuvent concerner deux points du réseau phylogénétique indépendamment de leur éloignement dans le réseau, les nœuds de réticulation se trouvent finalement presque tous dans une même partie réticulée de grande taille. Ainsi, l'idée de traitement indépendant de toutes les parties réticulées du réseau ne permet pas un grand gain de temps.

Fiabilité des réseaux reconstruits par les méthodes combinatoires

Une dernière limite des approches combinatoires concerne la fiabilité du réseau reconstruit. Quand on cherche à reconstruire un réseau compatible avec un ensemble de clades souples, qui respecte certaines restrictions topologiques, pour peu que l'évolution ait suivi la forme d'un réseau sans restriction, c'est un réseau différent que l'on reconstruira.

De même, dans le cas où certaines données de triplets sont manquantes, plusieurs réseaux peuvent être compatibles. Toutefois, cette contrainte est prise en compte par certaines méthodes qui fournissent l'ensemble de tous les réseaux solution. Pour raccourcir les temps de calcul, certains logiciels choisissent de renvoyer un des réseaux solution, en choisissant le plus parcimonieux. Par exemple, s'il existe un réseau de niveau 2 compatible avec des données de triplets, le logiciel Simplistic (van Iersel & Kelk 2008) peut fournir celui qui contient le minimum de nœuds de réticulation.

En revanche, un résultat plus étonnant concerne le cas où toutes les données sont présentes et correctes, il est alors possible que plusieurs réseaux tout aussi parcimonieux soient compatibles. En effet, on peut vérifier que les trois réseaux de niveau 1 de la figure 4 (i) correspondent exactement au même ensemble de triplets et de clades souples, il y a donc ambiguïté entre ces trois configurations (et notamment sur le taxon qui provient de la recombinaison) si l'on se base uniquement sur les triplets, ou les clades. En fait, il est possible de caractériser quels réseaux parmi ceux de niveau 1 présentent cette ambiguïté, et lesquels sont au contraire encodables par leur ensemble de triplets ou de clades : un réseau de niveau 1 ne contenant que des cycles à au moins 4 sommets est encodable par ses triplets et ses clades (Gambette & Huber, 2011).

Pour les réseaux de niveau 1, les ambiguïtés possibles quand on se base uniquement sur les triplets sont donc bien caractérisées. On peut toutefois

remarquer que les différents réseaux possibles ont une structure assez proche : comme pour ceux de la figure 4 (i), il est possible de passer de l'un à l'autre par une simple réorientation des arcs. En revanche, l'exemple de la figure 4 (ii) montre deux réseaux de niveau 2 qui ont exactement le même nombre de nœuds de réticulation, de nœuds, d'arêtes, qui ne sont pas équivalents à réorientation des arcs près, et qui pourtant ont exactement le même ensemble de triplets. Ceci incite à une grande prudence sur le réseau reconstruit par des méthodes combinatoires, puisqu'elles peuvent fournir un résultat différent mais tout aussi parcimonieux que le réseau réel, en présence de données exactes et complètes. Il est donc préférable de reconstruire la totalité des réseaux possibles en fonction des données, puis de les considérer comme des candidats à la configuration du réseau réel, et revenir aux données de séquences pour choisir le meilleur de ces candidats, à l'aide de méthodes statistiques par exemple.

PERSPECTIVES

Il reste encore de nombreux problèmes combinatoires concernant la reconstruction de réseaux phylogénétiques. Une grande partie d'entre eux concerne les arbres de gènes « multi-étiquetés », c'est-à-dire qui contiennent non seulement des gènes orthologues mais aussi des paralogues. Outre les implications mathématiques, puisque l'objet étudié change (certaines étiquettes de taxons-espèces pouvant être répétées dans l'arbre de gènes), ceci oblige à prendre en compte les implications biologiques, et intégrer la duplication, ainsi que la perte de gènes, dans le modèle.

Le développement des bases de données de type phylome pose également la question de la navigation parmi les arbres de ces bases. Nous travaillons actuellement sur un système de visualisation pour aider au choix des arbres à utiliser pour la création d'un réseau phylogénétique.

En ce qui concerne la reconstruction de réseaux elle-même, si les approches combinatoires semblent indispensables, un couplage avec des méthodes statistiques semble prometteur : l'algorithme combinatoire propose un candidat ou un ensemble de réseaux candidats, et le calcul statistique permet d'en privilégier un. On peut même imaginer un dialogue, dans une procédure de construction incrémentale du réseau, en ajoutant des feuilles successivement.

MATÉRIEL COMPLÉMENTAIRE

Le matériel complémentaire concernant la figure 1 (données, procédure détaillant les logiciels et scripts utilisés) est disponible à l'adresse suivante http://hogenom.gambette.com.

REMERCIEMENTS

Je tiens à remercier Vincent Berry et Celine Scornavacca pour leurs scripts de traitement de données, Daniel Huson pour d'utiles indications sur Dendroscope, et Delphine Amstutz pour son aide à la récupération des informations taxonomiques sur UniProt.

BIBLIOGRAPHIE

ALTHÖFER I., 1986. On optimal realizations of finite metric spaces by graphs. *Disc. Comp. Geometry*, 3 : 103-122.

BANDELT H.-J. & DRESS A.W., 1992. Split decomposition : a new and useful approach to phylogenetic analysis of distance data. *Mol. Phyl. Evol.*, 1 : 242-252.

BORDEWICH M. & SEMPLE C., 2007. Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Applied Mathematics*, 155 : 914-918.

BRYANT D. & MOULTON V., 2002. Neighbor-Net : an agglomerative method for the construction of planar phylogenetic networks. *Proceedings of WABI'02, L.N.C.S.*, 2452 : 375-391.

DRESS A.W., HUBER K.T., KOOLEN J. & MOULTON V., 2008. Compatible Decompositions and Block Realizations of Finite Metrics. *Eur. Jour. of Comb.*, 29 (7) : 1617-1633.

GAMBETTE P., 2008. Who is Who in phylogenetic networks : articles, authors and programs. <http://www.atgc-montpellier.fr/phylnet>.

GAMBETTE P., BERRY V. & PAUL C., 2009. The structure of level-k phylogenetic networks. *Proceedings of CPM'09*, *L.N.C.S.*, 5577 : 289-300.

GAMBETTE P., BERRY V. & PAUL C., 2011. Quartets and unrooted phylogenetic networks, soumis au *Journal of Bioinformatics and Computational Biology*.

GAMBETTE P. & HUBER K., 2011. On encodings of phylogenetic networks of bounded level. *Journal of Mathematical Biology*, à paraître.

GUPTA A., MANUCH J., ZHAO X. & STACHO L., 2006. Characterization of the existence of galled-tree networks. *J. Bioinfo. Comp. Bio.*, 4 : 1309-1328.

HUERTA-CEPAS J., DOPAZO H., DOPAZO J. & GALBADÓN T., 2007. The human phylome. *Genome Biology*, 8 : r109.

HUSON D., 1998. SplitsTree : analyzing and visualizing evolutionary data. *Bioinformatics*, 14 (1) : 68-73.

HUSON D., RICHTER D.C., RAUSCH C., DEZULIAN T., FRANZ M. & RUPP R., 2007. Dendroscope : an interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, 8 : 460.

HUSON D., RUPP R., BERRY V., GAMBETTE P. & PAUL C., 2009. Computing galled networks from real data. *Proceedings of ISMB / ECCB'09*, *Bioinformatics*, 25 : i85-i93.

VAN IERSEL L., KEIJSPER J., KELK S., STOUGIE L., HAGEN F. & BOEKHOUT T., 2008. Constructing level-2 phylogenetic networks from triplets. *Proceedings of RECOMB'08, L.N.C.S.*, 4955 : 450-462.

VAN IERSEL L., KELK S., 2008. Constructing the simplest possible phylogenetic network from triplets. *Proceedings of ISAAC'08, L.N.C.S.*, 5369 : 472-483.

JANSSON J., NGUYEN N.B. & SUNG W.-K., 2006. Algorithms for combining rooted triplets into a galled phylogenetic network. *SIAM Journal on Computing*, 35 : 1098-1121.

JANSSON J. & SUNG W.-K., 2006. Inferring a level-1 phylogenetic network from a dense set of rooted triplets. *Theoretical Computer Science,* 363 : 60-68.

KANJ I.A., NAKHLEH L., THAN C. & XIA G., 2008. Seeing the trees and their branches in the network is hard. *Theoretical Computer Science*, 401 : 153-164.

LESSER A., 2007. *Optimal and hereditarily optimal realizations of metric spaces*. PhD Thesis, Uppsala University, Uppsala.

MA B., WANG L. & LI M., 1998. Fixed topology alignment with recombination. *Proceedings of CPM'98*, *L.N.C.S.*, 1448 : 174-188.

MAKARENKOV V., 2001. T-Rex : reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17 : 665-668.

MATSEN F.A., 2010. ConstNJ : an algorithm to reconstruct sets of phylogenetic trees satisfying pairwise topological constraints. *J. Comput. Biol.*, 17 : 799-818.

SAITOU N. & NEI M., 1987. The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4 : 406-425.

SONG Y.S., 2006. A concise necessary and sufficient condition for the existence of a galled tree. *T.C.B.B.*, 3 : 186-191.

SEMPLE C. & STEEL M., 2006. Unicyclic networks : compatibility and enumeration. *T.C.B.B.*, 3 : 84-91.

TO T.-H. & HABIB M., 2009. Level-k phylogenetic networks are constructable from a dense triplet set in polynomial time. *Proceedings of CPM'09*, *L.N.C.S.*, 5577 : 275-288.

VELASCO J. & SOBER E., 2010. Testing for treeness : lateral gene transfer, phylogenetic inference, and model selection. *Biology and Philosophy*, 25 : 675–687.