Une nouvelle méthode pour la détection de transferts horizontaux de gène : la réconciliation topologique d'arbres de gène et d'espèces

Alix Boc, Vladimir Makarenkov, Abdoulaye Baniré Diallo,

Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8
Courriels : boc.alix@courrier.uqam.ca, makarenkov.vladimir@uqam.ca et diallo.abdoulaye_banire@courrier.uqam.ca

Résumé

L'évolution des espèces est un problème complexe qui nécessite la prise en compte de nombreux mécanismes d'évolution tels que le transfert horizontal de gène (THG), la duplication et la perte de gènes, l'hybridation et l'homoplasie. Dans cet article, nous proposons une nouvelle méthode pour la détection des THGs. Cette méthode est basée sur la réconciliation des topologies de l'arbre du gène considéré et de l'arbre d'espèces. La méthode décrite dans cet article considère un modèle d'évolution phylogénétique en réseau. Le critère d'optimisation utilisé est la distance topologique de Robinson et Foulds (Robinson et Foulds, 1981), qui permet de mesurer la similarité entre deux arbres phylogénétiques. La nouvelle méthode génère un scénario de propagation du gène testé pour un ensemble d'espèces considérées. Dans la section application, nous verrons son utilité pour la détection des transferts du gène rubisco dans une phylogénie de 40 espèces comprenant des plantes, des cyanobactéries et des protéobactéries.

Introduction

Le processus d'évolution d'espèces a longtemps été modélisé uniquement à l'aide d'arbres phylogénétiques. Dans de tels arbres, chaque espèce ne peut être reliée qu'avec son ancêtre le plus proche et toutes autres relations inter-espèces, comme par exemple, des transferts horizontaux de gènes (i.e. transferts latéraux de gènes) ne sont pas représentables. Cependant, les transferts horizontaux de gènes jouent un rôle clé dans l'évolution d'espèces, et en particulier, des bactéries. En effet, de nombreux projets de séquençages de génomes des bactéries ont renforcé l'idée que l'analyse phylogénétique d'un groupe d'espèces doit prendre en compte des phénomènes suivants : la convergence évolutive, la duplication, la perte et le transfert latéral de gènes. Ces importants mécanismes ne peuvent alors être représentés qu'à l'aide d'un model en réseau. Plusieurs tentatives d'utiliser des modèles en réseaux pour détecter des transferts latéraux peuvent être trouvées dans la littérature scientifique, voir par exemple Hein (1990) ou Page et Charleston (1998). Un nouveau modèle de transfert latéral permettant d'inscrire des arbres phylogénétiques de gènes dans un arbre phylogénétique d'espèces correspondant a été proposé par Hallet et Lagergren (2001). Dans Boc et Makarenkov (2003) et Makarenkov, Boc et Diallo (2004) nous avons proposé deux algorithmes pour la détection des transferts horizontaux de gène. Ces algorithmes sont basés sur un modèle d'évolution dans lequel plusieurs règles d'évolution ont été incorporées. Le critère des moindres carrés a été choisi comme critère d'optimisation de base dans les deux derniers articles.

Dans ce papier, nous décrivons de nombreuses améliorations apportées au modèle d'évolution de base. De plus, nous examinons comment un critère de similarité topologique entre deux arbres phylogénétiques, la distance de Robinson et Foulds (1981), pourrait être utilisé pour la détection de transferts horizontaux. Nous montrerons comment les différences topologiques entre les arbres de gène et d'espèces peuvent être exploitées pour déterminer un scénario possible de transferts latéraux du gène considéré survenus au cours de l'évolution. À la différence de nos précédents travaux, ici nous considérons le cas quand le gène transféré du donneur *supplante au complet* le gène homologue de l'espèce hôte.

Description du modèle de transfert

Dans un arbre phylogénétique, il existe toujours un chemin unique reliant toute pair de nœuds. Par contre, l'addition d'une branche représentant un transfert horizontal de gène (THG) crée un autre chemin entre certains nœuds, en transformant l'arbre phylogénétique en réseau. La Figure 1 illustre le cas où le chemin de poids minimum entre les taxa *i* et *j* changera après l'ajout d'une nouvelle branche orientée (*a*,*b*), dirigée de *b* vers *a*. D'un point de vue biologique, il est plausible de considérer que le transfert horizontal de gène entre *b* et *a* affecte la distance évolutive entre le taxon *i* et le taxon *j*, dont la position dans l'arbre phylogénétique est fixe, si et seulement si *i* est situé au dessous de la branche de transfert (Figure 1). Par contre, la distance évolutive entre i_1 et *j* ne sera pas affectée par ajout de la branche (*a*,*b*).

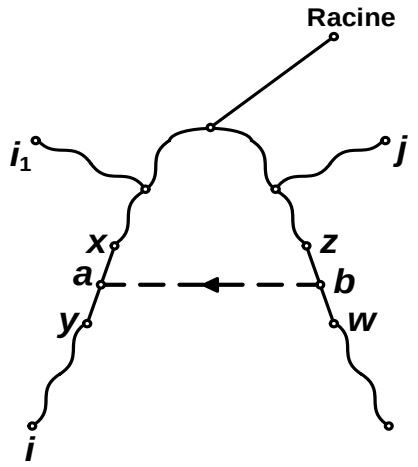

Figure 1. La distance d'évolution entre les taxa *i* et *j* sera affectée par l'ajout d'une nouvelle branche (*a*,*b*) représentant le transfert horizontal de gène entre les branches (*z*,*w*) et (*x*,*y*) dans l'arbre d'espèces; la distance d'évolution entre les taxa i_1 et *j* ne sera pas affectée par l'ajout de la nouvelle branche (*a*,*b*).

De nouvelles règles biologiques ont été ajoutées à notre modèle de base. Ces règles permettent de prendre en considération les interactions entre plusieurs THGs, et de

respecter le sens d'évolution allant de la racine vers les feuilles qui sont associées aux espèces contemporaines.

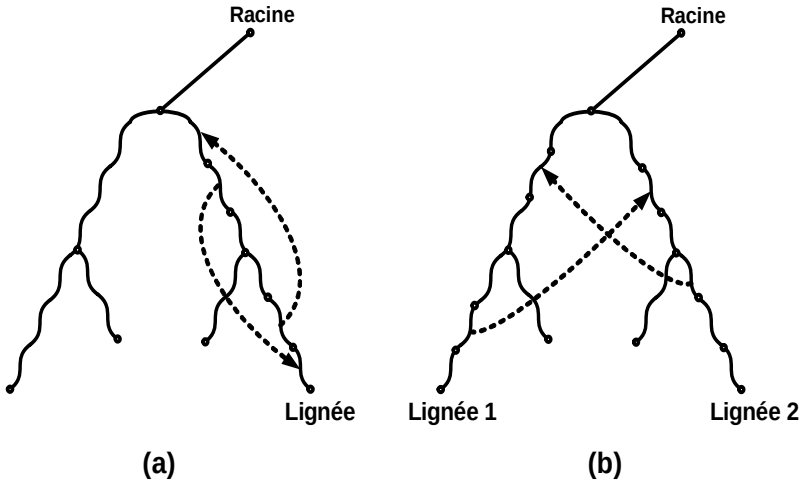

Figure 2. Le transfert horizontal ne peut être considéré quand les branches en question sont situées sur la même lignée (a), i.e. sur le même chemin venant de la racine; ou quand deux transferts, affectant deux lignées considérées, se croisent (b).

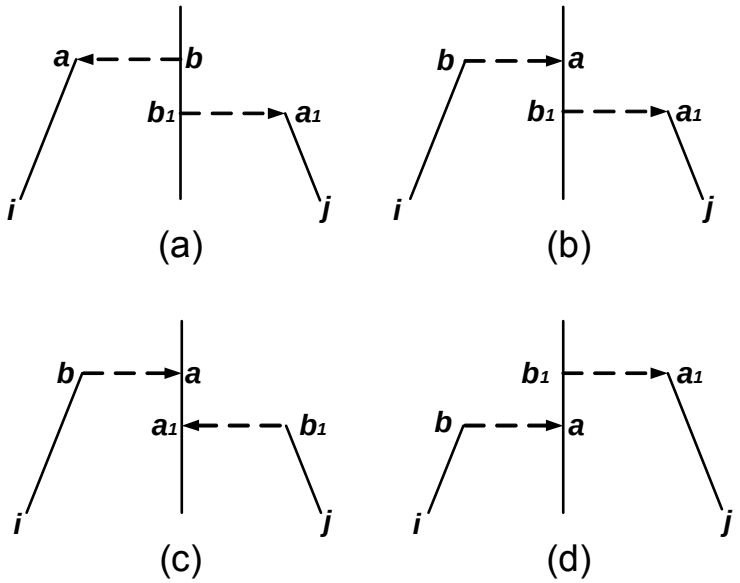

Figure 3. Cas (a) et (b): la distance évolutive entre les feuilles *i* et *j* peut passera par le chemin incluant les deux HGTs (*a*,*b*) et (a_1,b_1). Cas (c) et (d): la distance évolutive entre les feuilles *i* et *j* ne passera pas par le chemin incluant les deux HGTs (*a*,*b*) et (a_1,b_1).

Tout d'abord, tout transfert entre des espèces d'une même lignée (Figure 2a) ne est pas permis. De même, on interdit les cas de transferts doubles entre deux lignées, quand deux THGs se croisent, comme c'est illustré sur la Figure 2b. Tous les deux THGs sur la Figure 2b affecte l'ancêtre de l'espèce se trouvant à la base de l'autre THG. Ces

restrictions permettent d'éviter les scénarios d'évolution n'ayant aucune explication biologique. Par la suite, nous identifions deux cas où la distance évolutive entre les espèces *i* et *j* peut être évaluée à travers de multiples transferts (voir Figures 3a et b), ainsi que deux cas où la distance évolutive entre les espèces *i* et *j* ne peut être affectée par l'ajout des mêmes transferts multiples (voir Figures 3c et d).

Distance topologique de Robinson et Foulds

La distance topologique de Robinson et Foulds (1981) représente le nombre minimum d'opérations élémentaires de contraction et d'expansion de branches nécessaires pour transformer un arbre phylogénétique en un autre. La Figure 4 ci-dessous illustre les deux opérations nécessaires pour transformer l'arbre $\boldsymbol{T}$ en $\boldsymbol{T}_1$. La distance topologique de Robinson et Foulds sera utilisée pour évaluer la similarité topologique entre l'arbre d'espèces et l'arbre de gène à réconcilier. Les THGs minimisant le plus cette distance seront retenus comme solutions possibles (voir la section suivante).

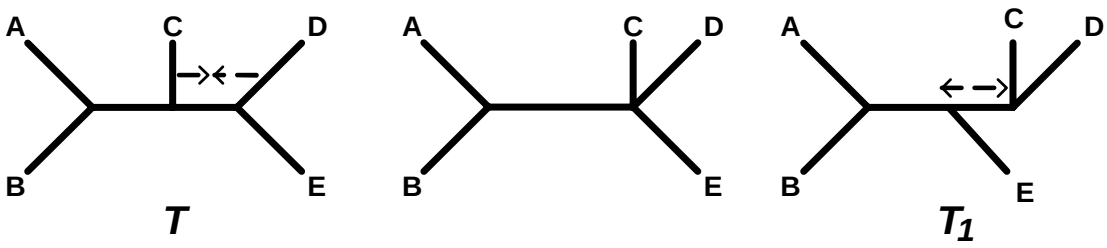

Figure 4. La distance de Robinson et Foulds entre $\boldsymbol{T}$ et $\boldsymbol{T}_1$ est égale à 2. Deux opérations de base, une contraction et une expansion de branches, sont nécessaires pour transformer l'arbre $\boldsymbol{T}$ en $\boldsymbol{T}_1$.

Description de la méthode

Dans cette section, nous décrivons une stratégie heuristique pour réconcilier l'arbre de gène et celui d'espèces en se basant sur le modèle d'évolution définie ci-dessus. Cette méthode est divisée en deux parties principales. La première partie décrit la procédure de reconstruction des deux arbres phylogénétiques représentant l'évolution du gène considéré et de l'arbre d'espèces. L'arbre d'espèces est généralement inféré à partir d'un gène ribosomal, e.g. 16S ARNr ou 23S ARNr, dont l'évolution n'a pas été affectée par des transferts horizontaux. La seconde partie de la méthode décrit le processus de réconciliation de deux arbres.

Partie 1. Soit *T* une phylogénie d'espèces dont les feuilles sont étiquetées selon l'ensemble *X* de *n* taxa. *T* peut être inféré à partir de données de séquences ou de distances à l'aide d'une méthode de reconstruction appropriée. Sans perte de généralité, nous supposons que *T* est un arbre binaire, dont les nœuds internes sont tous de degré 3 et qui possède $2n-3$ branches. Cet arbre doit être explicitement enraciné car la position de la racine est importante dans notre modèle.

Soit T_1 un arbre de gène dont les feuilles sont étiquetées selon le même ensemble X de n taxa utilisé pour étiqueter l'arbre d'espèces. Comme la phylogénie d'espèces T, la phylogénie de gène T_1 peut être inférée à partir de données de séquences ou de distances caractérisant ce gène particulier. Si les topologies de T et T_1 sont identiques, aucun transfert horizontal du gène donné ne devrait pas être indiqué dans l'arbre d'espèces. Par contre, si les deux phylogénies sont topologiquement différentes, i.e. la distance de Robinson et Foulds entre les arbres T et T_1 est supérieure à zéro (voir Figure 5, par exemple), cela pourrait être dû aux transferts horizontaux de gène.

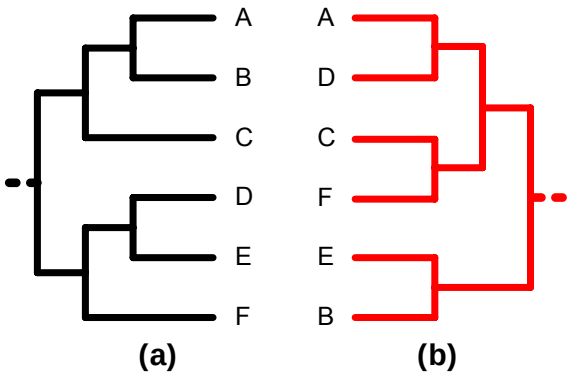

Figure 5. La phylogénie d'espèces (a) est différente de celle de gène (b). Il existe un scénario d'opérations de transferts permettant de réconcilier les topologies des deux arbres.

Partie 2. Le but de cette étape est de déterminer un scénario de transferts horizontaux nécessaires pour transformer T en T_1. Tous les THGs possibles entre les paires de branches de l'arbre T qui sont en accord avec notre modèle d'évolution sont évalués. À la première itération, le transfert diminuant le plus la distance de Robinson et Foulds entre les deux arbres est considéré comme le plus probable. Il est par la suite ajouté à l'arbre T. Comme dans ce modèle nous considérons le transfert du gène au complet, la branche reliant l'espèce affectée par ce transfert et son ancêtre directe est supprimée de T.

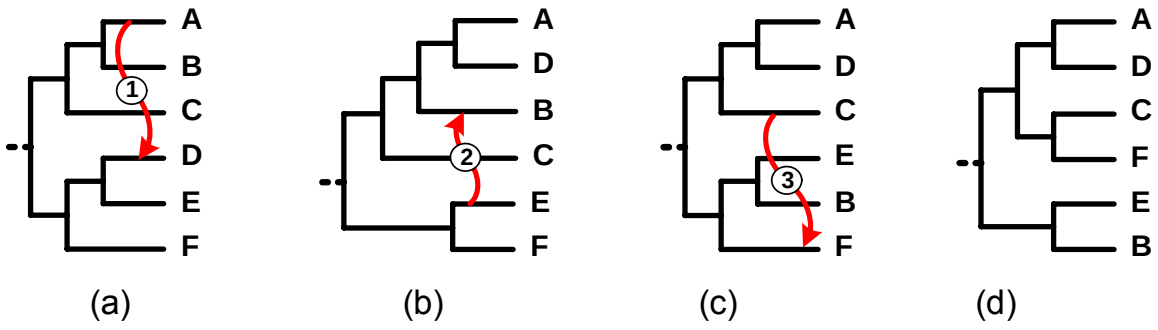

Figure 6. L'arbre d'espèces (a) est transformé en l'arbre de gène (d) en trois étapes.

Ensuite, à la deuxième itération, cet arbre T modifié est considéré pour déterminer le prochain THG à ajouter. La procédure algorithmique s'arrête lorsqu'un nombre prédéfini de THGs est ajouté dans T ou lorsque la distance de Robinson et Foulds entre les arbres T modifié et T_1 atteint zéro. La liste des THGs produite par cette procédure représente un

scénario possible de propagation du gène étudié. La Figure 6 ci-dessus illustre comment notre algorithme transforme l'arbre d'espèces (a) et l'arbre de gène (d). Chaque transfert trouvé tient compte des transferts ajoutés précédemment. Remarquons que le calcul de la distance topologique de Robinson et Foulds (pour le calcul optimal de cette distance, voir Makarenkov et Leclerc, 2000) entre l'arbre de gène T_1 et l'arbre d'espèces T modifié devient possible, car l'ajout d'une branche de transfert dans l'arbre d'espèces est toujours suivi par la suppression d'une branche, ce qui veut dire que T reste connexe est sans cycles et est donc toujours un arbre. La méthode décrite dans cette section nécessite $O(kn^4)$ opérations pour ajouter k transferts horizontaux dans l'arbre phylogénétique de n espèces.

Détection des transferts horizontaux du gène *rbcL*

La méthode introduite dans les sections précédentes a été appliquée au problème de la détection des transferts horizontaux du gène rubisco dans une phylogénie comprenant des algues marines, des cyanobactéries et des protéobactéries.

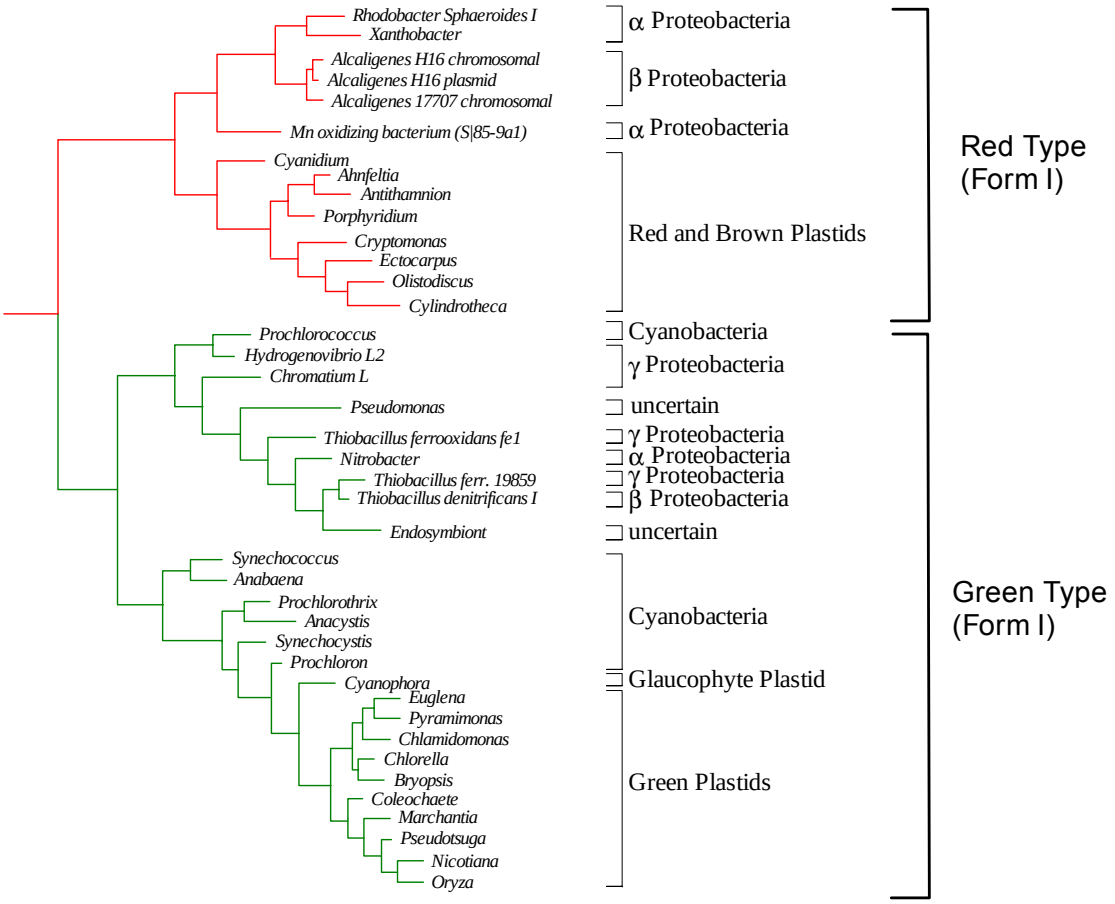

Figure 7. Phylogénie du gène *rbcL* (40 espèces) inférée par Delwiche et Palmer (1996). 8 espèces de phylogénie inférée par Delwiche et Palmer (1996, fig. 2), dont 6 de la forme II utilisées pour enraciner l'arbre et 2 étant présentes deux fois chacune dans l'arbre de gène, ont été supprimées de la phylogénie d'origine de 48 espèces.

Les données considérées proviennent de l'article Delwiche et Palmer (1996). En effectuant l'analyse phylogénétique du gène *rbcL* (rubisco) Forme I et II pour 48 espèces (voir Delwiche et Palmer, 1996 fig. 4), les dernières auteurs ont formulé des hypothèses sur les transferts horizontaux qui auraient pu influencer l'évolution de ce gène. Plus précisément, des transferts horizontaux du gène *rbcL* suivants ont été indiqués : entre les cyanobactéries et les γ-protéobactéries, les γ-protéobactéries et les α-protéobactéries, les γ-protéobactéries et les β-protéobactéries et les α-protéobactéries et les algues rouges et brunes. De plus, deux hypothèses de transfert horizontaux supplémentaires ont aussi été formulées par Delwiche et Palmer : la première – le transfert entre les γ-protéobactéries et *Prochlorococcus* et le deuxième – le transfert entre les α-protéobactéries et le groupe d'*Alcaligènes*.

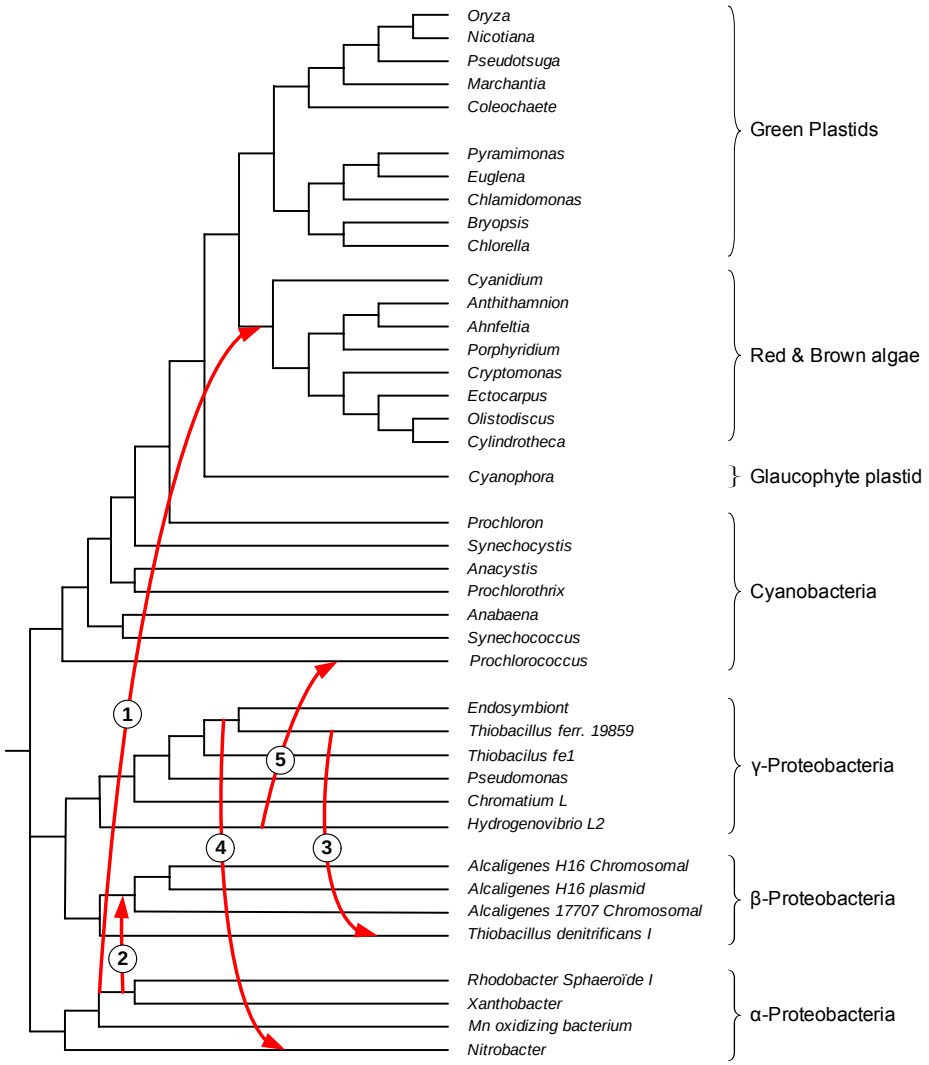

Figure 8. L'arbre phylogénétique de 40 espèces (arbre d'espèces). Les flèches représentent les 5 transferts horizontaux trouvés et les numéros associés leur ordre d'apparition.

Avant d'appliquer notre méthode, nous avons réduit le nombre d'espèces de 48 à 40. Le but étant de garder un représentant du gène *rbcL* par espèce. Nous avons donc supprimés tous les représentants de la Forme II du *rbcL* (6 espèces) utilisés par Delwiche et Palmer pour enraciner l'arbre de gène ainsi que les espèces *Chromatium A* et *Hydrogénovibrio* L_1. La phylogénie du gène *rbcL* pour les 40 espèces retenues est illustrée sur la Figure 7. L'arbre d'espèces (Figure 8), inféré à partir des séquences ribosomales 16S ARNr, 23S ARNr et d'autres évidences biologiques, a été construit en utilisant la méthode NJ (Saitou et Nei, 1987).

En observant l'arbre d'espèces et celui du gène *rbcL*, nous pouvons noter d'importantes différences topologiques entre eux. Par exemple, sur la Figure 7 on retrouve un cluster des γ-protéobactéries dans lequel on trouve insérées une α-protéobactérie, une β-protéobactérie et une cyanobactérie. Ces contradictions peuvent être expliquées soit par des transferts latéraux de gène qui auraient pu se produire entre les espèces indiquées, soit par une ancienne duplication du gène *rbcL* suivie de sa perte chez certaines espèces. Ces deux hypothèses ne sont pas mutuellement exclusives (pour plus de détails voir Delwiche et Palmer, 1996). La nouvelle méthode a été appliquée aux deux phylogénies (Figures 7 et 8) et a produit un scénario possible de transferts latéraux du gène *rbcL*. La solution obtenue est décrite par les cinq transferts latéraux de gènes illustrés sur la Figure 8. Le transfert entre les α-protéobactéries et les algues rouges et brunes est le plus significatif, suivi de celui entre les α-protéobactéries et le cluster des *Alcaligènes* et ainsi de suite. Notre méthode a permis de détecter tous les transferts latéraux indiqués par Delwiche et Palmer (1996) sauf celui des cyanobactéria vers les γ-protéobactéries.

Conclusion

Une nouvelle méthode pour la détection des transferts horizontaux de gène, basée sur un principe de réconciliation d'arbres phylogénétiques et sur un modèle d'évolution, a été introduite dans ce papier. Cette méthode exploite efficacement les différences topologiques entre un arbre d'espèces et un arbre de gène construits pour un même ensemble d'organismes. Elle produit un scénario de transferts horizontaux du gène étudié dans la phylogénie d'espèces. De plus, les nouvelles règles biologiques et mathématiques ajoutées au modèle de base ont permis d'améliorer la qualité de nos résultats, tant au niveau de la détection des THGs qu'au niveau du temps d'exécution. L'utilisation de la distance topologique de Robinson et Foulds, en tant que critère d'optimisation, permet la détection des transferts horizontaux pertinents. Dans le futur, il serait important de développer une technique de validation des résultats produits par cette méthode. Une telle technique serait idéalement capable de mesurer un taux de fiabilité à accorder aux résultats obtenus. D'un autre côté, le développement des modèles de THGs basés sur le maximum de vraisemblance et le maximum de parcimonie serait également une piste d'investigation intéressante. La nouvelle méthode a été codée en langage C. L'exécutable Windows de même que le code source en C compilable sous UNIX sont mis à la disposition des chercheurs à l'adresse URL suivante : <http://www.info.uqam.ca/~boca05/software>. Ce programme sera bientôt ajouté au logiciel T-Rex (Makarenkov, 2001) disponible pour Windows et Macintosh et bénéficiant d'une interface utilisateur graphique très conviviale.

Remerciement

Nous tenons à remercier Hervé Philippe pour son aide dans l'analyse phylogénétique des données du gène *rbcL*.

Références

Boc, A. et Makarenkov, V.: New Efficient Algorithm for Detection of Horizontal Gene Transfer Events. *Algorithms in Bioinformatics*, G. Benson and R. Page (Eds.), 3rd Annual WABI'03, Springer-Verlag, pp. (2003) 190-201.

Delwiche, C.F., et J. D. Palmer.: Rampant Horizontal Transfer and Duplication of Rubisco Genes in Eubacteria and Plastids, *Mol. Biol. Evol.* (1996) 13, 873-882.

Hallet, M., et Lagergren, J.: Efficient algorithms for lateral gene transfer problems. RECOMB 2001, Montréal, ACM, (2001) 149-156

Hein, J.: A heuristic method to reconstructing the evolution of sequences subject to recombination using parsimony. *Math. Biosci.* (1990) 185-200.

Makarenkov, V. et Leclerc, B. Comparison of additive trees using circular orders, *Journal of Computational Biology*, (2000) 7, 731-744.

Makarenkov,V.: T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics* (2001) 17, 664-668.

Makarenkov, V. Boc, A. et Diallo, A. B.: Representing lateral gene transfer in species classification. Unique scenario. Accepté pour publication à IFCS 2004, Chicago.

Page, R. D. M. et Charleston, M. A.: From gene to organismal phylogeny: Reconciled trees. *Bioinformatics* (1998) 14, 819-820.

Robinson D.R et Foulds L.R.: Comparison of phylogenetic trees, *Mathematical Biosciences* (1981) 53, 131-147.

Saitou, N. et Nei, M.: The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* (1987) 4, 406-425