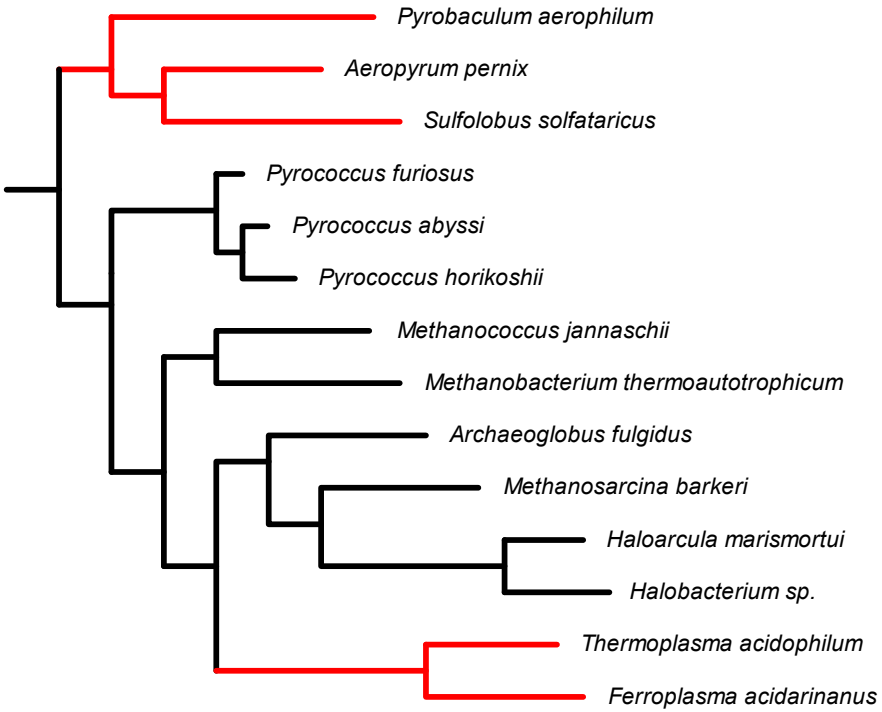
Détection des transferts horizontaux de gènes

Alix Boc
Université du Québec à Montréal

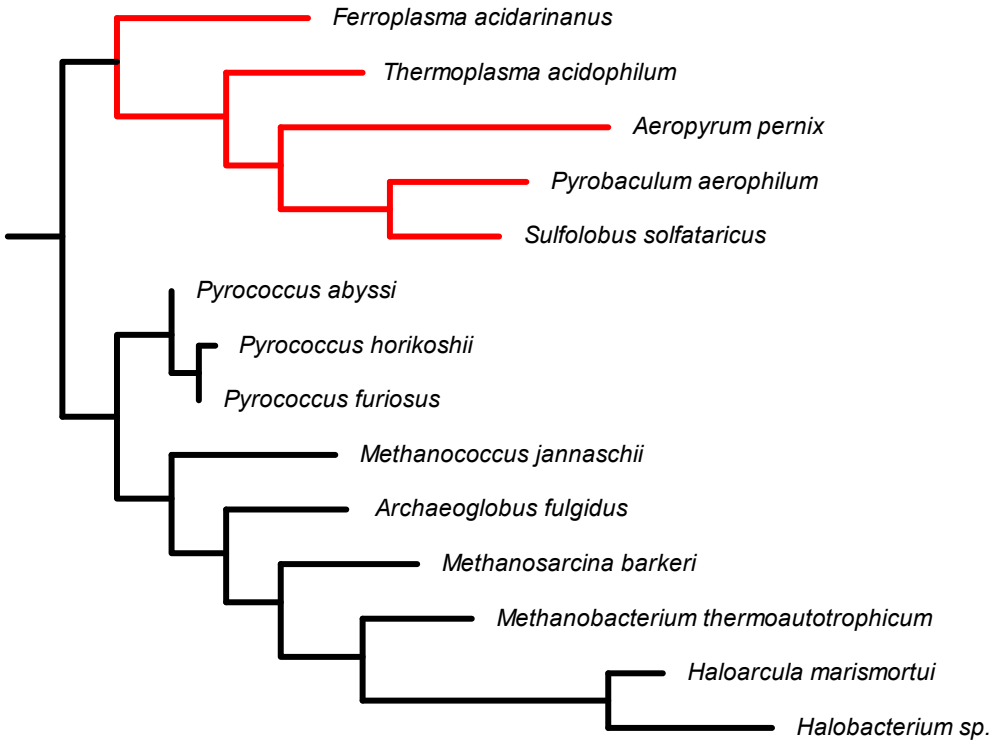
SOMMAIRE

- Phylogénie
- Problématique
- Détection des transferts horizontaux de gènes complets
- Détection des transferts horizontaux de gènes partiels
- Biolinguistique
- Conclusion

TRANSFERTS HORIZONTAUX DU GÈNE *RPL12E*



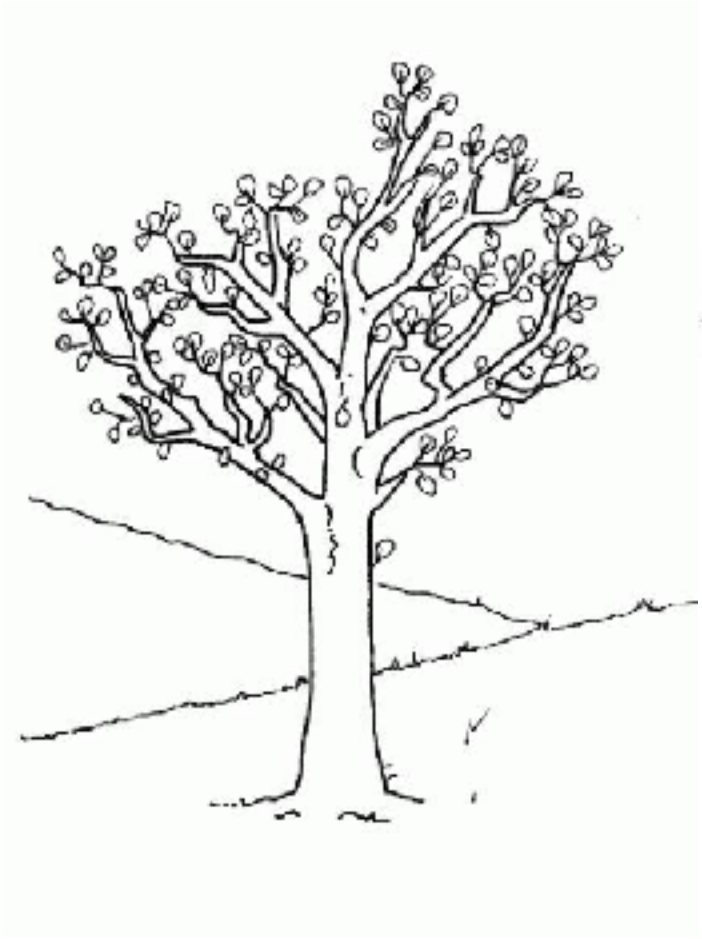
arbre d'espèces

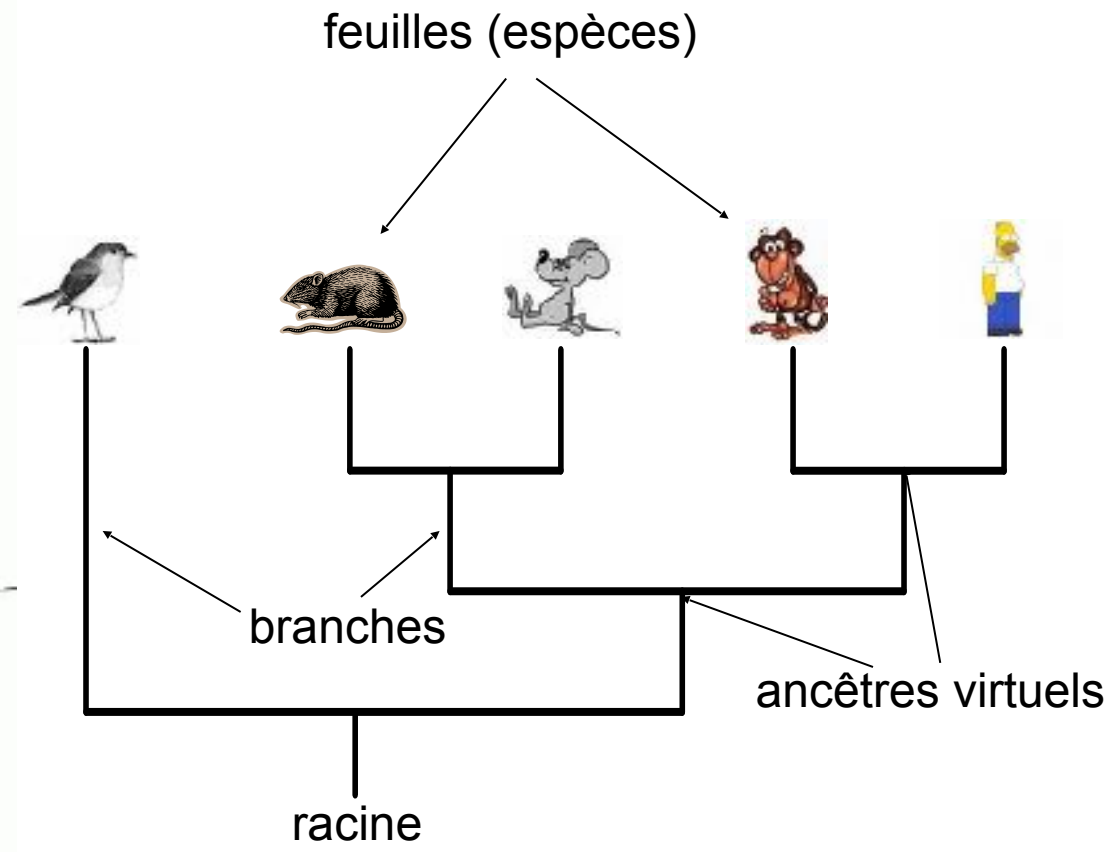


arbre du gène *rpl12e*

Phylogénie

La phylogénie (ou phylogénèse) étudie la parenté entre différents êtres vivants en vue de comprendre leur évolution.



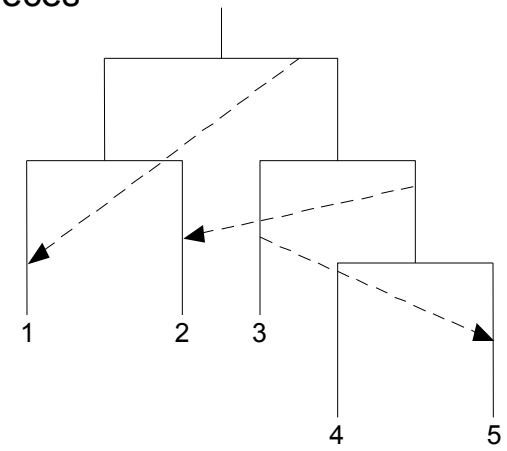


Problématique

LE TRANSFERT HORIZONTAL DE GENES

Définition : **Le transfert horizontal de gènes (THG) est un processus permettant à une espèce d'acquérir du matériel génétique d'une autre espèce**

- C'est un mécanisme important qui joue **un rôle clé** dans l'évolution des espèces
- Il permet aux organismes de **s'adapter** à leur environnement
- C'est un processus dominant chez les procaryotes (e.g., bactéries)
- Il s'explique par le phénomène de **l'évolution réticulée**
- Il ne peut pas être représenté par un **arbre phylogénétique** traditionnel

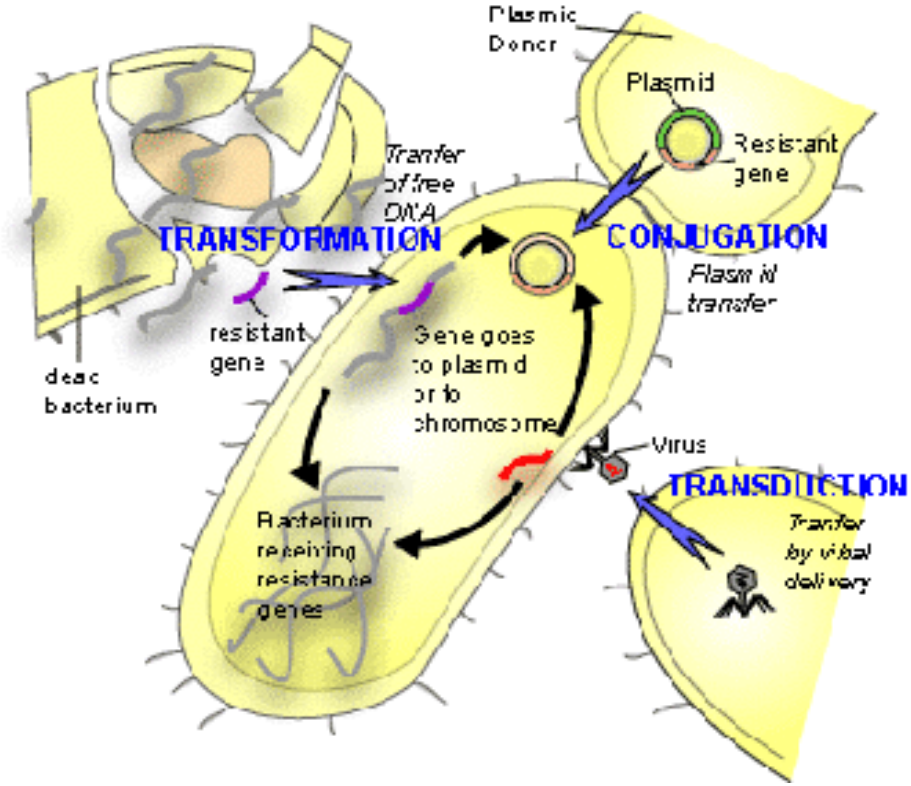


Nous proposons donc une méthode de détection des THG qui se base sur les arbres phylogénétiques d'espèces et de gène

Autres processus importants s'expliquant par l'évolution réticulée :

- La duplication ancestrale et la perte partielle de gènes (Delwiche et Palmer, 1996)
- L'hybridation (Huson, 1998, Bryant et Moulton, 2004)
- L'homoplasie et la convergence de gènes (Legendre et Makarenkov, 2002)

Trois principaux mécanismes de transfert horizontal de gènes



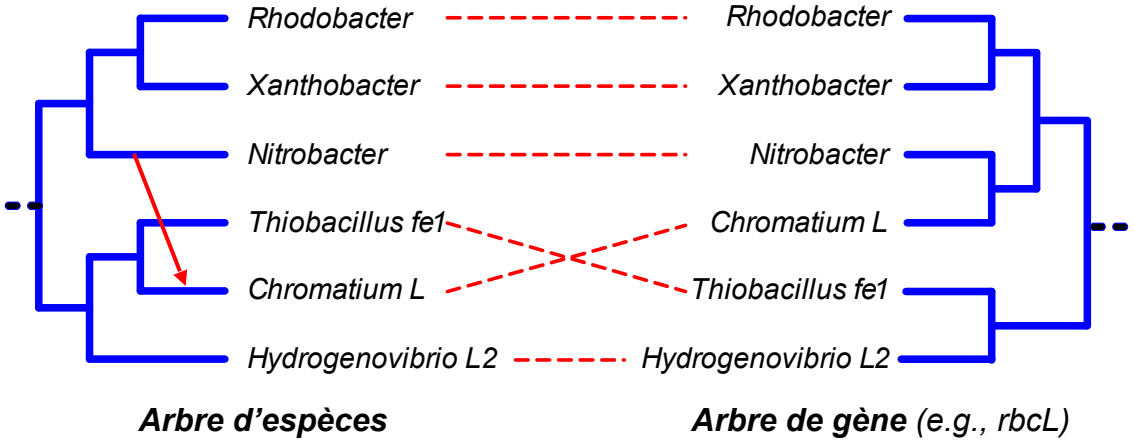
- **Transformation :** intégration d'ADN libre
- **Conjugaison :** échange de matériel entre deux organismes
- **Transduction :** transmission par l'intermédiaire de bactériophages

Le transfert horizontal de gène est considéré comme un des facteurs principaux de l'augmentation de la **résistance** des bactéries aux **antibiotiques**.

Détection des transferts complets★

Nom : ***HGT-Detection***

Méthode : utilisation du principe de **réconciliation** par des **déplacements de sous-arbres** (mouvements *SPR*) afin de transformer l'arbre d'espèce en l'arbre de gène.



Données en entrée :

- arbre phylogénétique d'espèces
- arbre phylogénétique du gène étudié (*pour le même ensemble d'espèces*)

Données en sortie: nombre minimal de déplacements de sous-arbres dans l'arbre d'espèces permettant de le transformer en l'arbre de gène (=> scénario de réconciliation)

Contraintes :

- incorporer les règles d'évolution dans le modèle mathématique
- maintenir la complexité algorithmique polynomiale

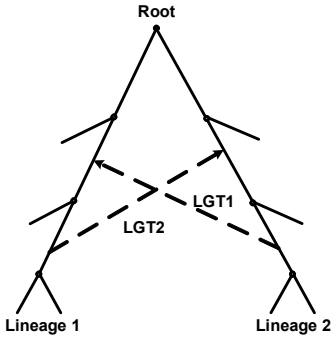
(le problème SPR, subtree pruning and regrafting, a été montré NP-complet par Hein et al., 1996)

RÈGLES D'ÉVOLUTION

Les règles d'évolution assurent qu'un transfert respecte les contraintes biologiques définies

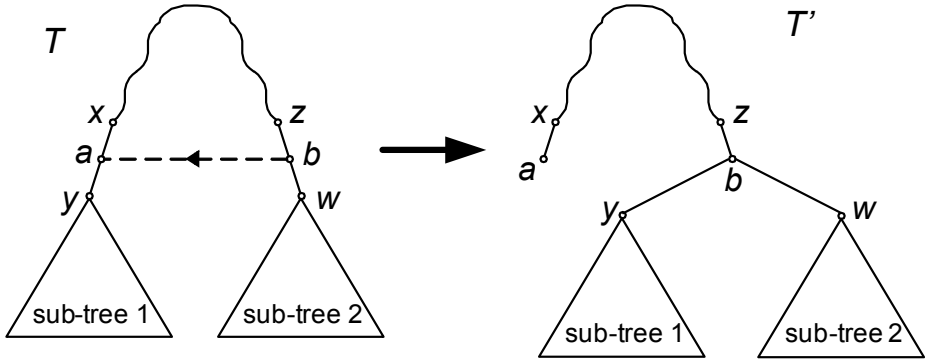
Règles d'évolution : 2 exemples

Les transferts sur la même lignée sont interdits.



Les transferts croisés suivants sont interdits.

Contrainte de sous-arbres ★



Description : Le transfert entre les branches (*z*,*w*) et (*x*,*y*) de l'arbre d'espèces *T* sera permis si et seulement si le sous-arbre regroupant les deux sous-arbres affectés, et enraciné par la branche (*z*,*b*) dans *T'*, est présent dans l'arbre de gène.

Avantages :

- Prend en compte automatiquement toutes les règles d'évolution
- Améliore le taux de détection des THG
- Permet une détection chronologique des transferts (du plus récent au plus ancien).

CRITÈRES D'OPTIMISATION

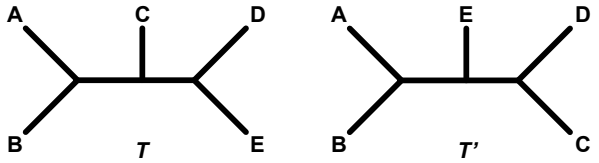
Permet de déterminer <u>les meilleurs transferts</u> à chaque étape du processus de réconciliation

Moindres carrés (Gauss, 1811)

$$Q = \sum_i \sum_j (d(i,j) - \delta(i,j))^2 \rightarrow \min$$

d(i,j) - distance entre *i* et *j* dans l'arbre d'espèces
$\delta(i,j)$ - distance entre *i* et *j* dans l'arbre de gène

Robinson et Foulds (1981)



La distance topologique de Robinson et Foulds entre deux arbres phylogénétiques est égale au nombre minimal d'opérations élémentaires de fusion et de séparation de noeuds, nécessaires pour transformer un arbre en un autre.

<u>Exemple</u> : la distance de Robinson et Foulds entre les arbres *T* et *T'* est égale à 2.

Dissimilarité de bipartitions ★

<u>Description</u> : Cette dissimilarité reflète mieux la différence topologique entre deux tables de bipartitions décrivant deux arbres donnés.

<u>Avantages</u> :

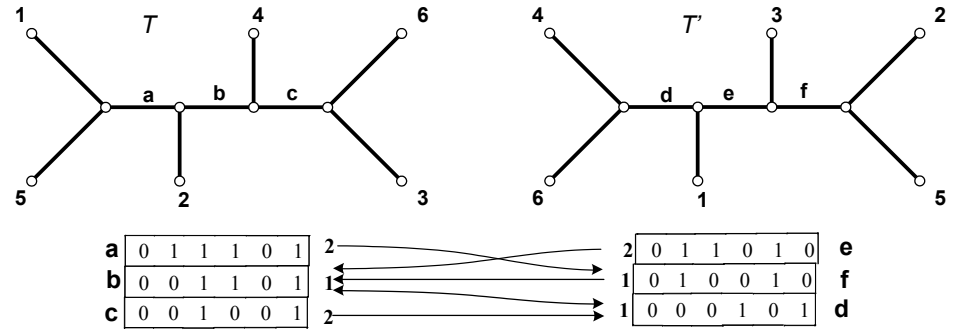
- Donne une mesure topologique plus précise (que les mesures similaires)
- Améliore le taux de détection des THG

Elle est définie comme suit :

$$bd = \left(\sum_{a \in BT} \underset{b \in BT'}{Min} (Min(d(a,b); d(a,\bar{b}))) + \sum_{b \in BT'} \underset{a \in BT}{Min} (Min(d(b,a); d(b,\bar{a}))) \right) / 2$$

où d(a,b) est la distance de Hamming entre les vecteurs de bipartitions a et b.

<u>Exemple</u> *: bd(T,T')*= ((2 + 1 + 2) + (2 + 1 + 1))/2 = 4,5

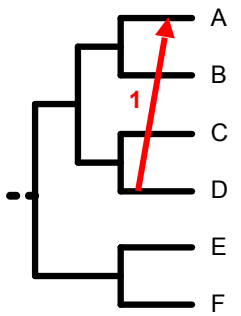


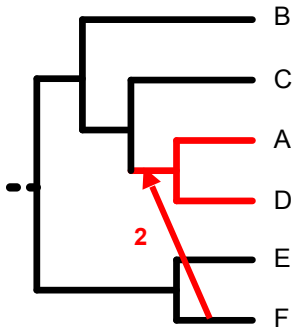
ALGORITHME : UN EXEMPLE

L'exemple ci-dessous montre comment l'arbre d'espèces *T* est transformé en l'arbre de gène *T'*

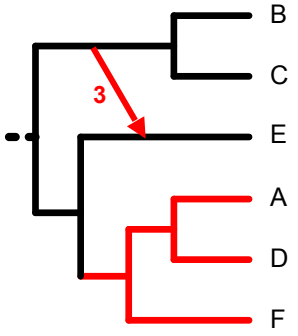
RÉCONCILIATION

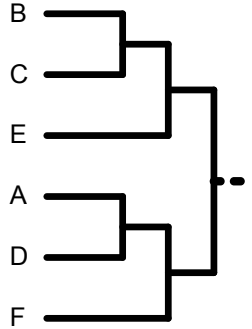
Arbre d'espèces *T*





Arbre de gène *T'*

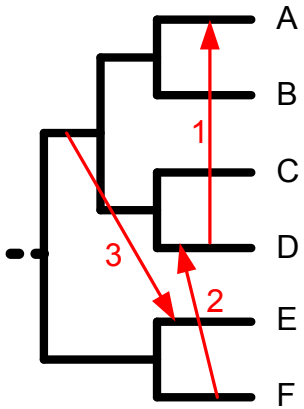




Données en sortie :

Scénario de taille minimale trouvé :

1 - transfert de (D) vers (A)
2 - transfert de (F) vers (D,A)
3 - transfert de (B,C) vers (E)



ALGORITHME « *HGT-DETECTION* »

Inférer les arbres d'espèces et de gène T et T' sur le même ensemble d'espèces;

Enraciner T et T' selon des évidences biologiques, par outgroup ou par midpoint;

Si (il existe des sous-arbres identiques avec au moins deux feuilles dans T et T') **alors**
Réduire la taille du problème en réduisant ces sous-arbres à une seule espèce dans T et T';

Sélectionner le critère d'optimisation : OC = LS (moindres carrés), RF (distance de Robinson et Foulds), QD (distance des quartets) ou BD (dissimilarité de bipartitions);

Calculer la valeur initiale de OC entre T et T';
$T_0 = T$;
$k = 1$;

Tant que ($OC \neq 0$)
{
Trouver l'ensemble de tous les THG éligibles à l'étape k (noté E_THG_k);

Tant que (les THG satisfaisant les conditions des Théorèmes 2 et 1 existent)
{
Si (il existe des THG appartenant à E_THG_k et satisfaisant les conditions du Théorème 2) **alors**
Effectuer les opérations SPR correspondant à ces THG;

Si (il existe des THG appartenant à E_THG_k et satisfaisant les conditions du Théorème 1) **alors**
Effectuer les opérations SPR correspondant à ces THG;
}
Effectuer toutes les opérations SPR correspondant aux THG satisfaisant la contrainte de sous-arbres;

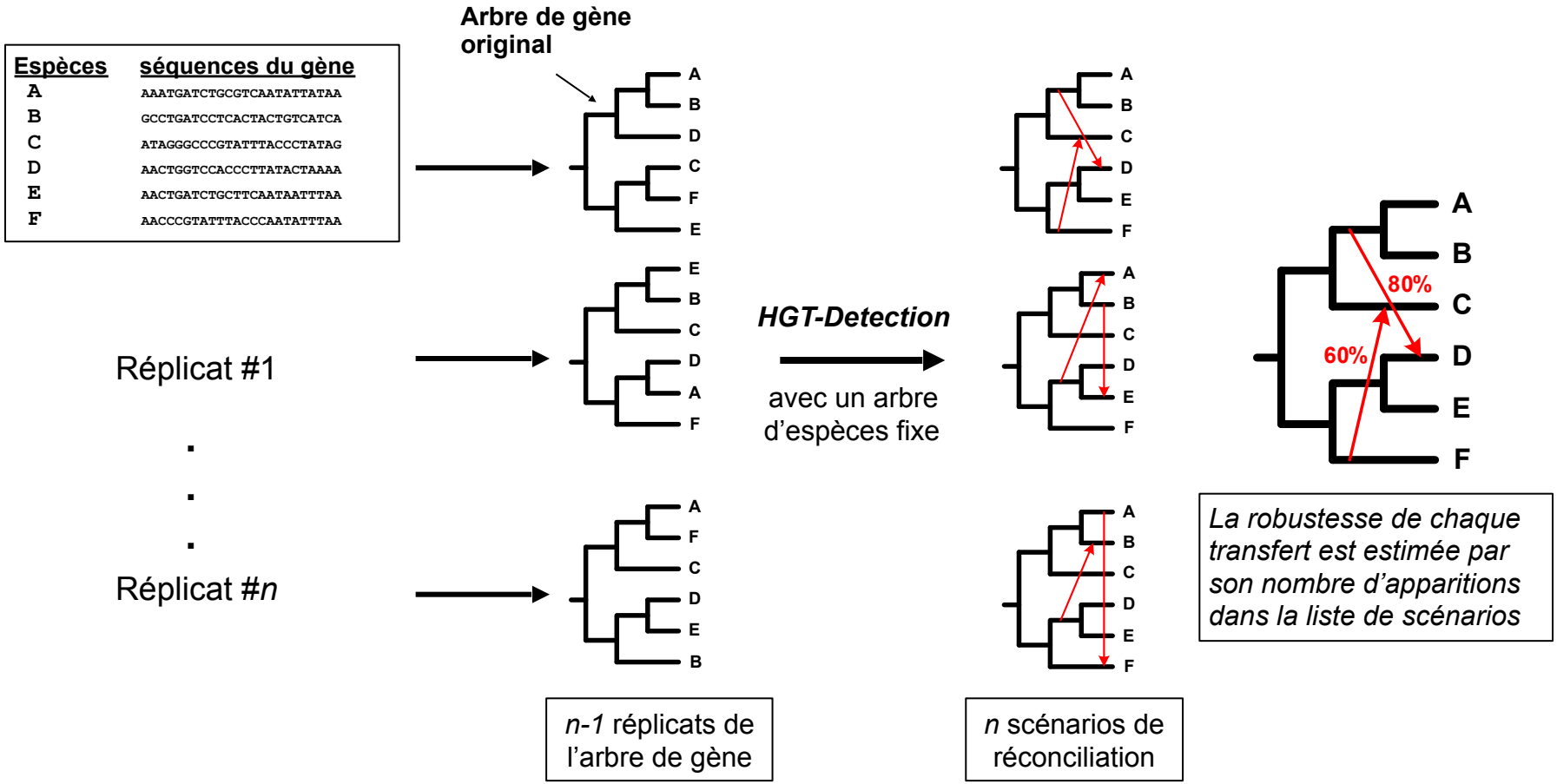
$k = k + 1$;
Décrémenter la taille du problème en réduisant en une arête tous les sous-arbres identiques dans T_k et T';
Calculer la valeur de OC entre T_k et T';
}

Éliminer tous les transferts inutiles;

Complexité algorithmique

$$O(\tau \times n^4)$$

Validation par bootstrap : Permet de déterminer la fiabilité de chaque transfert du scénario original en le comparant avec les scénarios obtenus à partir des réplicats de l'arbre de gène



Trois stratégies possibles de validation par bootstrap :

1. Les séquences utilisées pour construire les arbres d'**espèces** et de **gène** sont répliquées

2. **Seules** les données de séquences utilisées pour construire **l'arbre de gène** sont répliquées

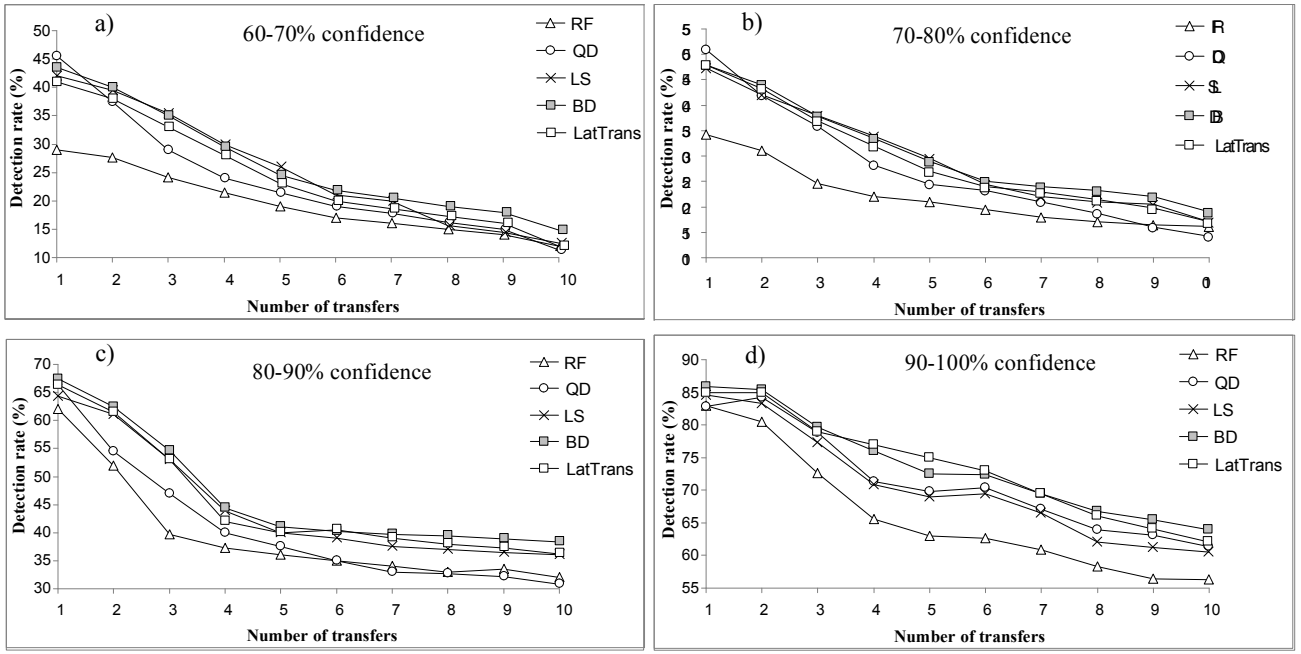
3. Le bootstrap des transferts peut être calculé entre deux topologies d'arbres seulement

$$HGT\_BS(t) = \frac{\sum_{1 \le i \le N_T} \sum_{1 \le j \le N_{T'}} \left(\sum_{1 \le k \le N_{ij}} \frac{\sigma_{k,ij}(t)}{N_{ij}} \times 100\% \right)}{(N_T \times N_{T'})}$$

$$\sigma_{k,ij}(t) = \begin{cases} 1, \text{ si } t \text{ est dans le scénario de coût minimal } k \text{ pour les arbres d'espèces } T_i \text{ et de gène } T'_j, \\ 0, \text{ sinon.} \end{cases}$$

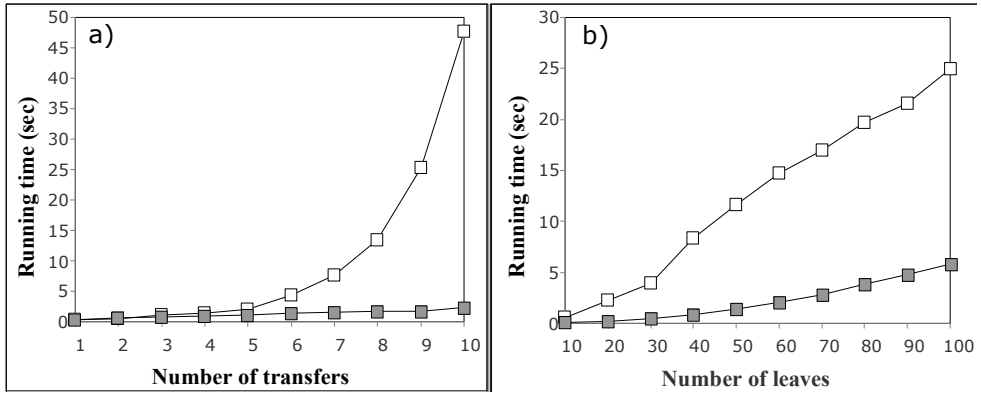
où N_T et $N_{T'}$ sont, respectivement, le nombre d'arbres d'espèces et de gène générés à partir des réplicats et N_{ij} est le nombre de scénarios de coût minimal obtenus quand l'algorithme est appliqué à l'arbre d'espèces T_i et l'arbre de gène T_j'.

Taux de détection en fonction du nombre de transferts.



L'utilisation de la dissimilarité de bipartitions (BD) permet d'obtenir un meilleur taux de détection des THG

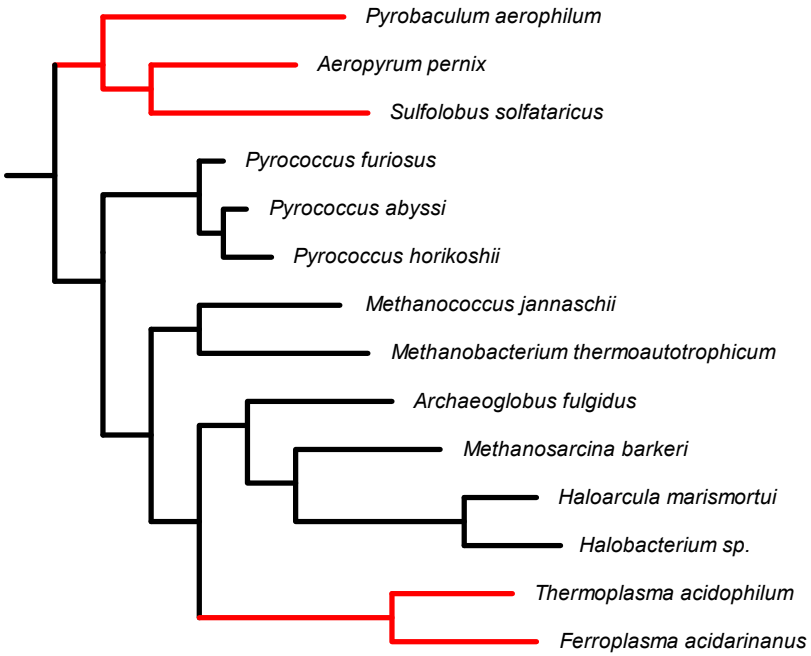
Comparaison de la stratégie basée sur BD avec l'algorithme ***LatTrans*** *(Hallett et Lagergren, 2001)*



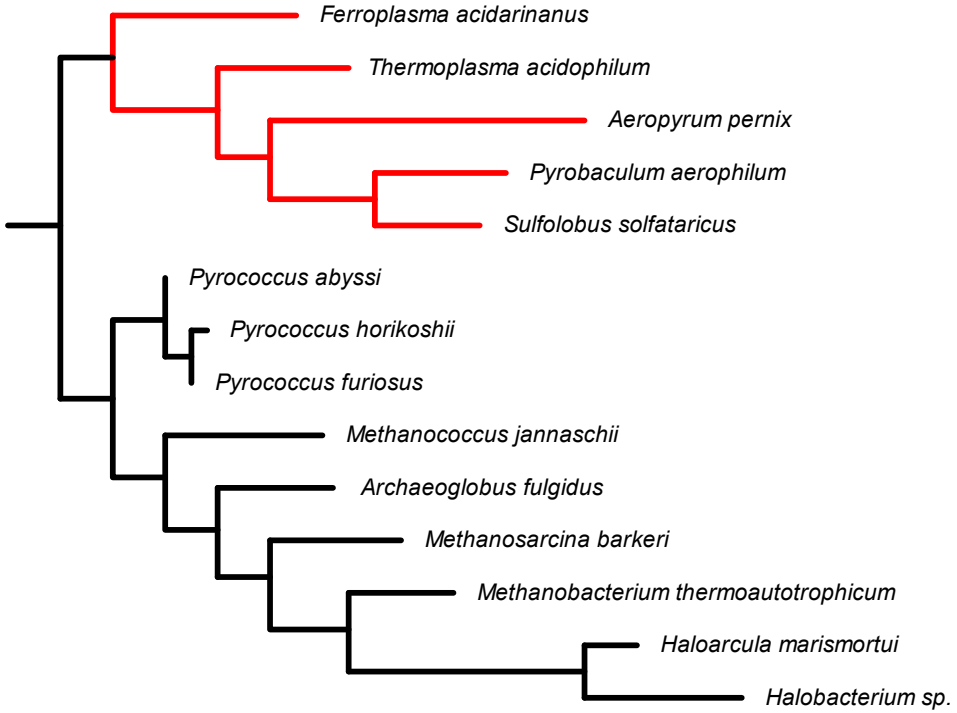
Exemple : évolution du gène *rpl12e* (Matte-Tailliez *et al.*, 2002)

TRANSFERTS HORIZONTAUX DU GÈNE *RPL12E*

<u>Hypothèse</u> : des transferts du gène *rpl12e* sont survenus entre les groupes des Crenarchaeota et Thermoplasmatales (Matte-Tailliez *et al.*, 2004)



arbre d'espèces

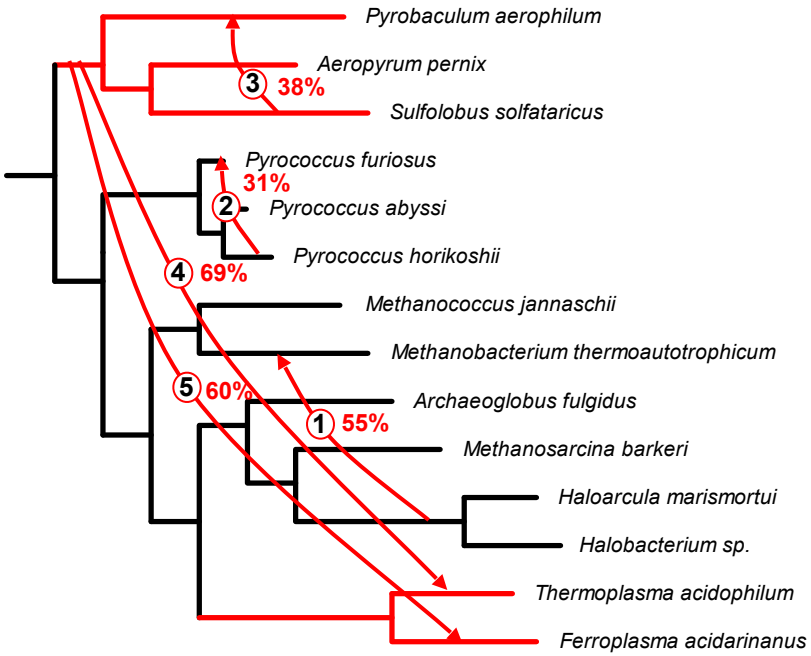


arbre du gène *rpl12e*

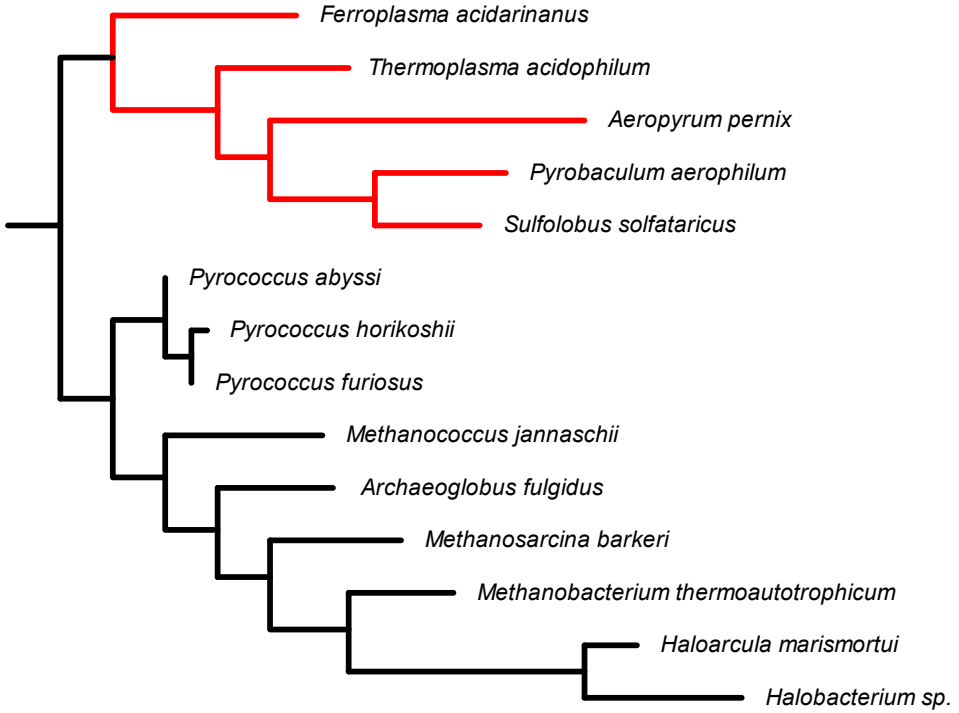
SCÉNARIO DE RÉCONCILIATION RETROUVÉ POUR LE GÈNE *RPL12E*

Résultat de l'application de notre méthode (*HGT-Detection)* :

Les transferts détectés correspondent aux hypothèses des auteurs



arbre d'espèces
(incluant les 5 THG détectés)

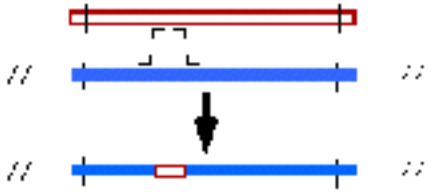


arbre du gène *rpl12e*

Détection des transferts partiels ★

❑ Contexte d'évolution

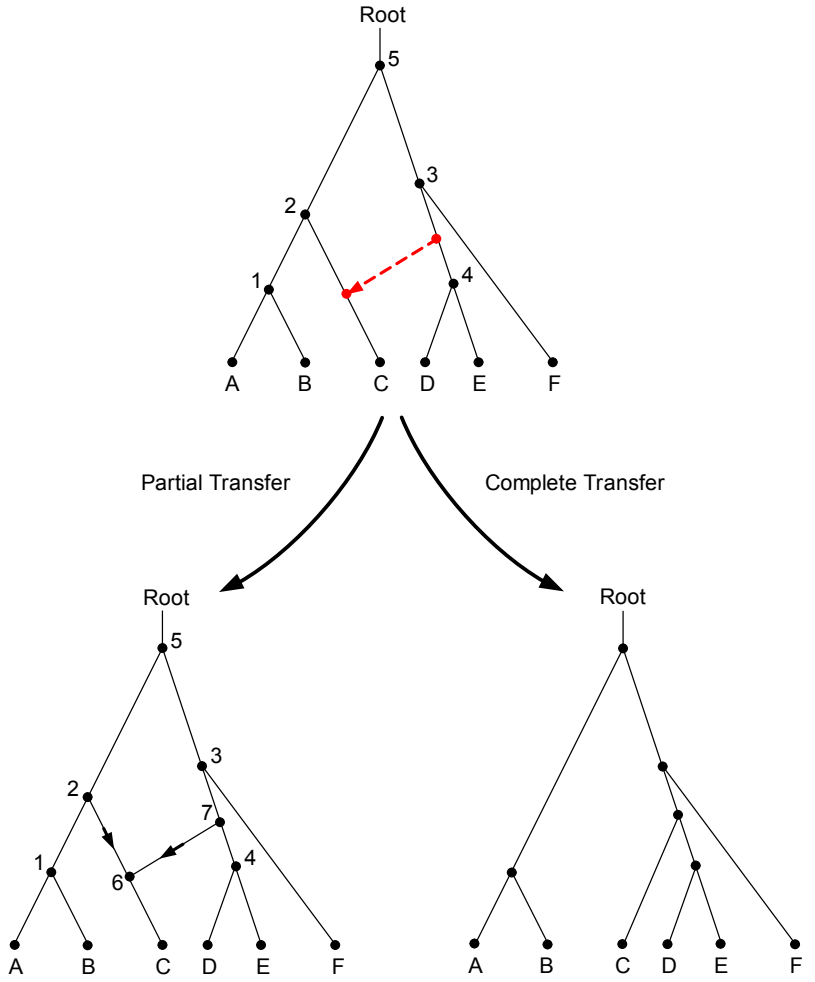
Les bactéries et les archées peuvent évoluer dans différentes conditions en s'échangeant des parties de gènes seulement, ce qui mène à la création des gènes mosaïques.



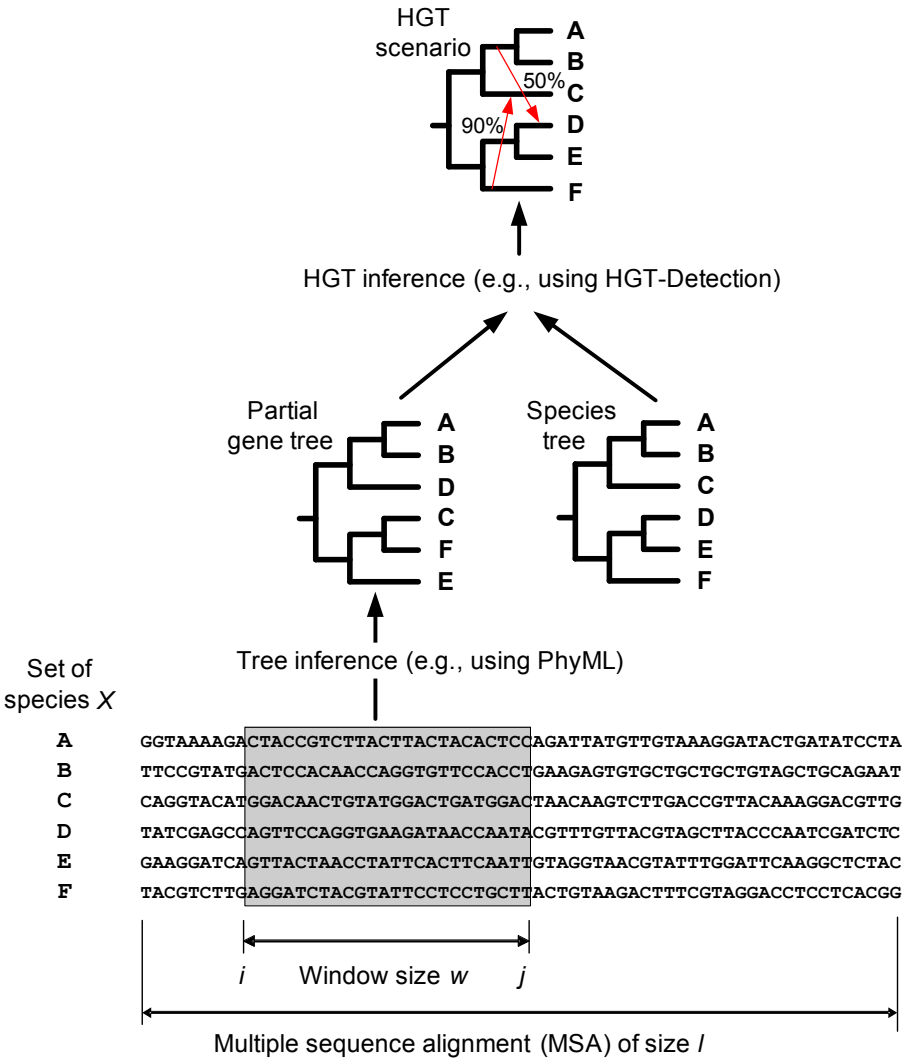
Un gène mosaïque est composé de sous-séquences provenant des espèces ou des souches différentes.

Généralisation : appliqué à l'échelle d'un génome, on peut estimer le taux de transferts horizontaux (complets et partiels) entre les espèces.

❑ Transfert partiel versus transfert complet



Méthode utilisant une fenêtre coulissante



Algorithme

Étape 1. Inférons un arbre phylogénétique ***T*** pour un ensemble d'espèces ***X***.

Étape 2. Pour *i* variant de 1 à |*l*-*w*|, inférons l'arbre de gène ***T'*** en utilisant les sous-séquences situées dans la fenêtre coulissante entre *i* et |*i*+*w*|. Appliquons *HGT-Detection* sur ***T*** et ***T'***.

Étape 3. Établissons une liste de transferts prédits. Les transferts entrelacés ($S_{i,j+w}$ .. $S_{j,j+w}$) seront considérés comme un unique transfert affectant les sites de *i* à |*j*+*w*|.

Complexité algorithmique

$$O(r \times (\frac{(l-w)}{s} \times (C(Phylo\_Inf) + \tau \times n^4)))$$

Une étude de l'évolution des langues Indo-Européennes

Alix Boc

Description :

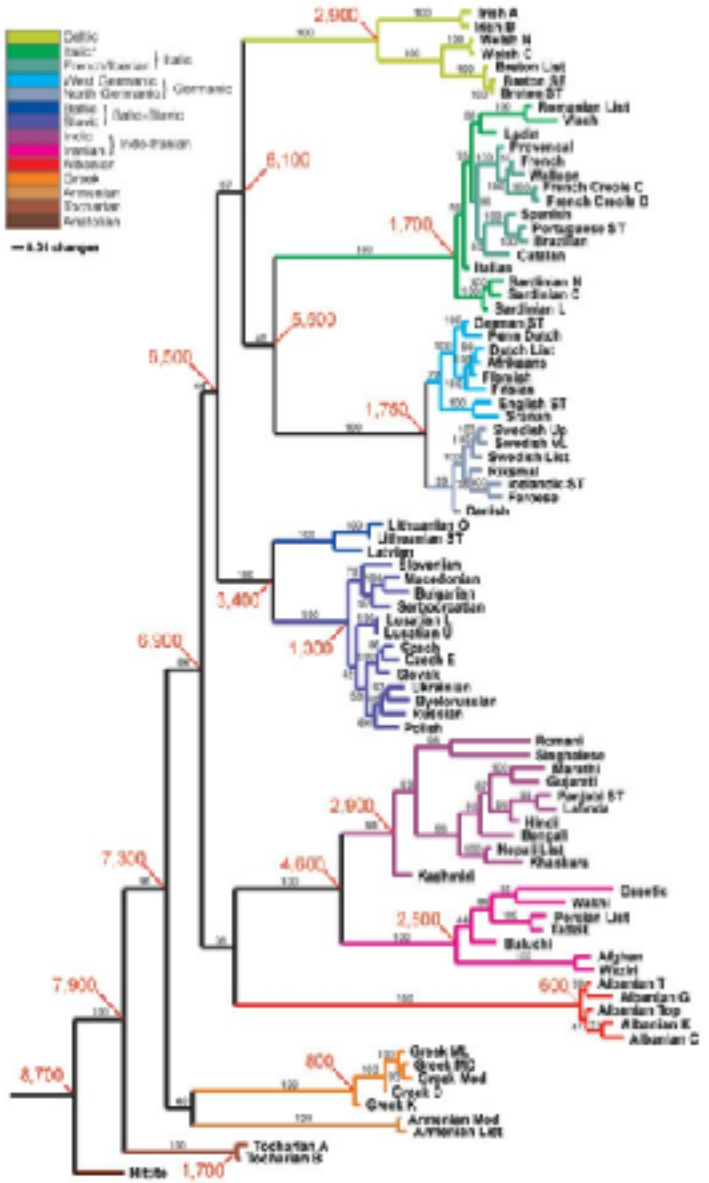
Détecter les plus importantes tendances d'échanges de mots survenus au cours de l'évolution des langues Indo-Européennes (IE) en se basant sur :

- L'arbre d'évolution des langues IE

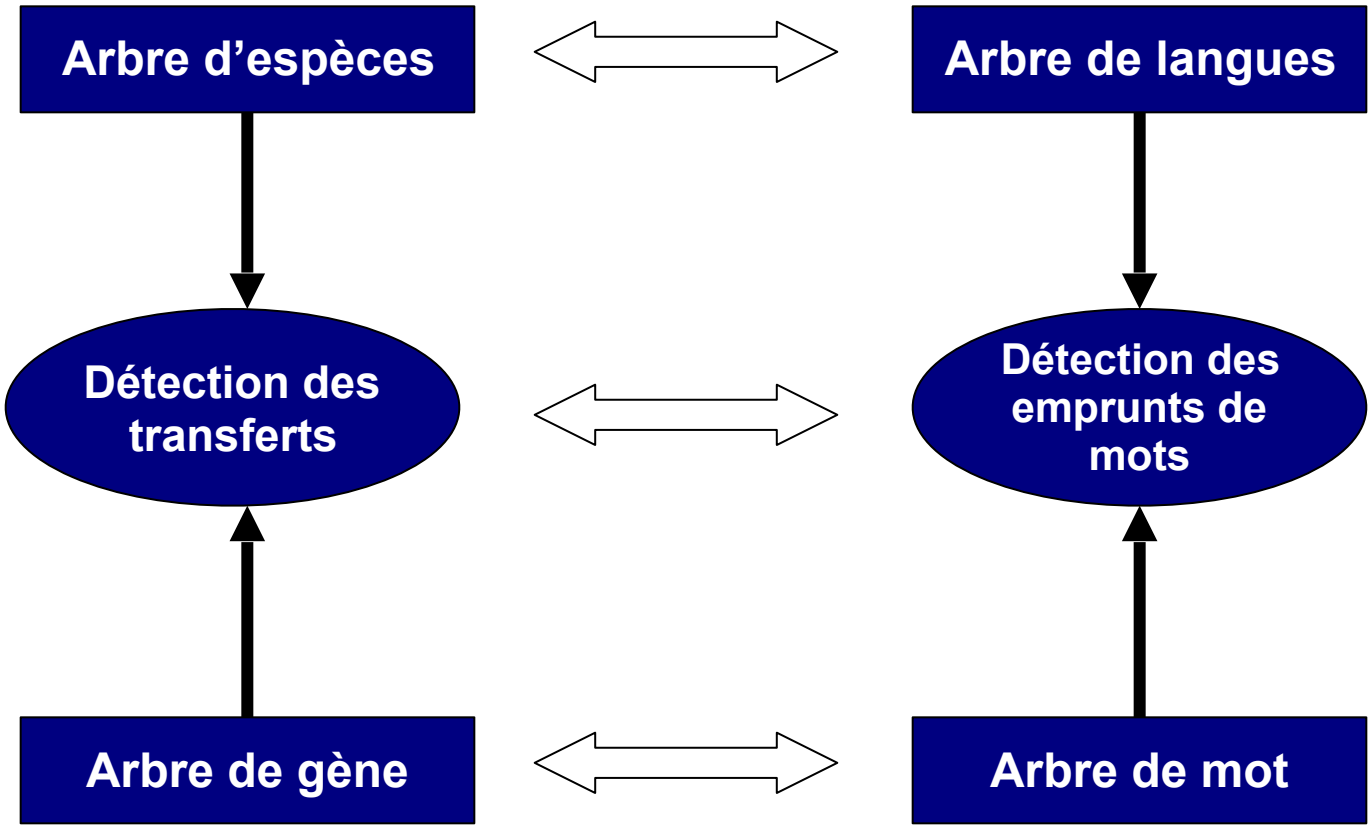
  (Gray et Atkinson, Nature, 2003).

- La base de données organisée par Dyen *et al.* (1997)

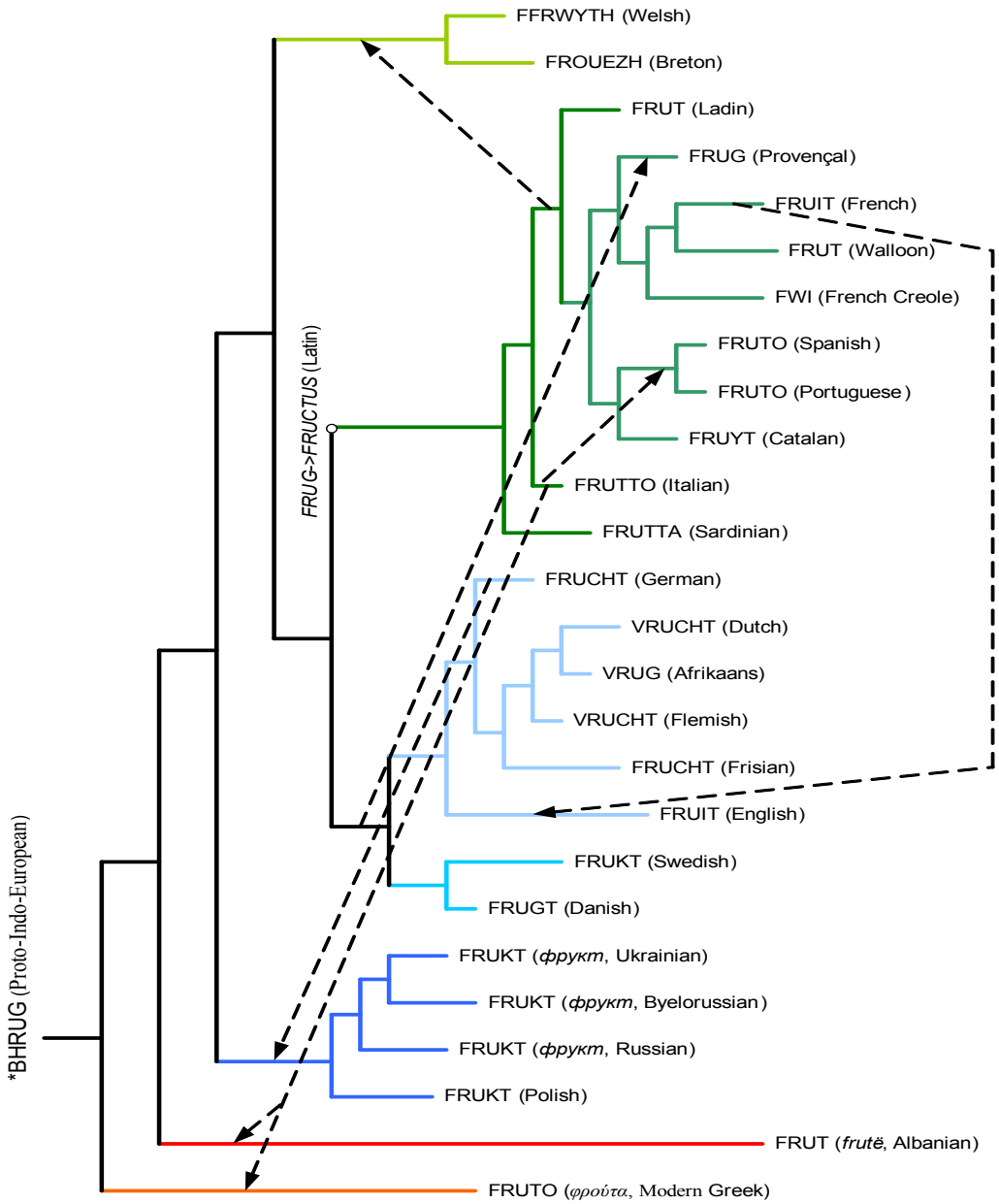
(200 mots de la liste Swadesh traduits dans 87 langues et structurés en 1484 cognats).



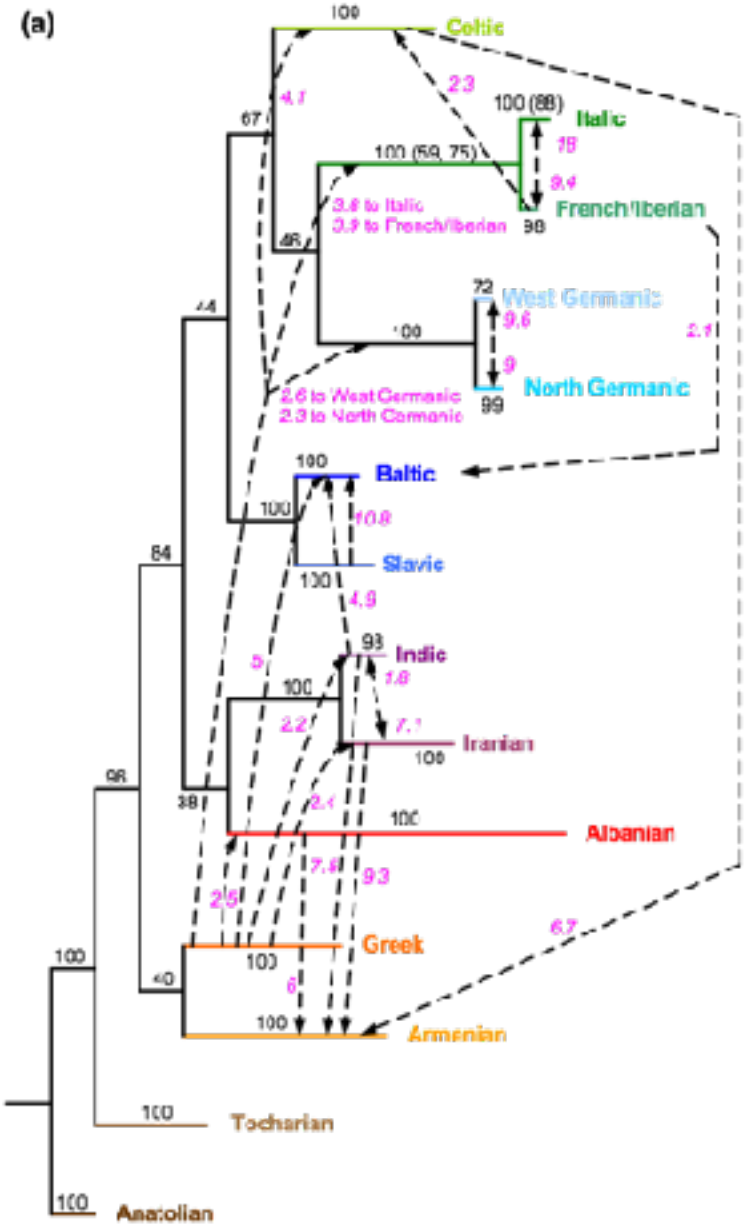
PARALLÈLE ENTRE L'ÉVOLUTION DES ESPÈCES ET CELLE DES LANGUES

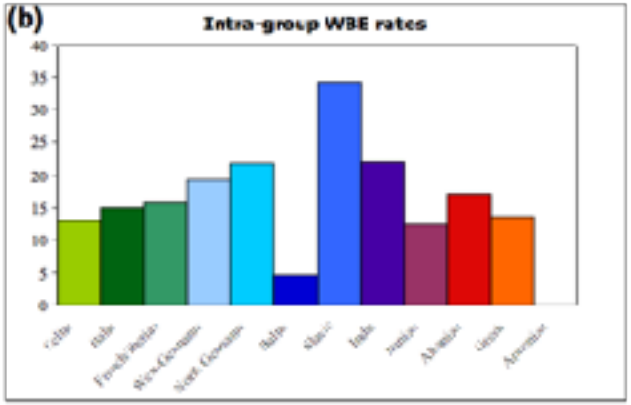


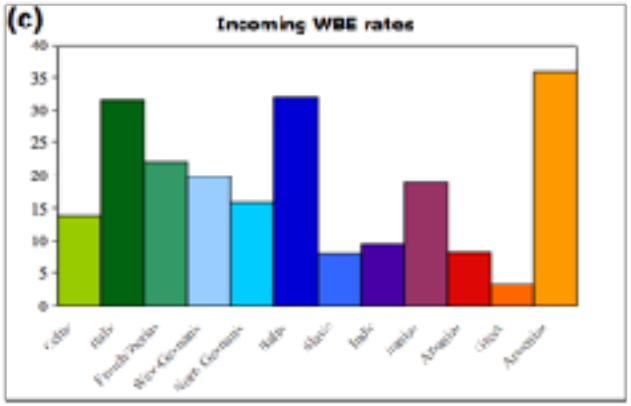
RÉSULTATS: ÉVOLUTION DU MOT "FRUIT"



RÉSULTATS CUMULATIFS







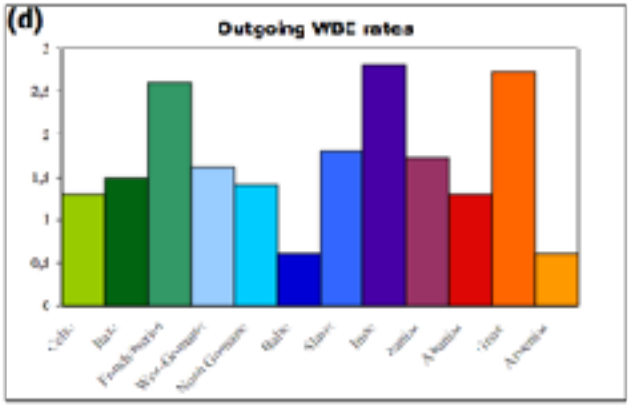


Figure: Taux de transferts entre les groupes de langues obtenus pour les mots des catégories lexicale et fonctionnelle (i.e., le total des 200 mots).

Conclusion : **35,4** % des mot des langues Indo-Européennes ont été affectées par des emprunts de mots.

Conclusion

CONTRIBUTIONS DE LA THÈSE

❑ Dans le cadre de cette recherche nous avons apporté plusieurs contributions importantes dont :

- Un algorithme efficace de détection des transferts horizontaux complets (*HGT-Detection*) (Makarenkov *et al.*, 2006 ; Boc *et al.*, 2010a)

- Une nouvelle mesure de comparaison d'arbres : la dissimilarité de bipartitions (Boc *et al.*, 2010a)

- Validation des transferts horizontaux par bootstrap (Makarenkov *et al.*, 2007 ; Boc *et al.*, 2010a)

- La contrainte de sous-arbres (Makarenkov *et al.*, 2006 ; Boc *et al.*, 2010a)

- Deux algorithmes de détection des transferts partiels, permettant d'identifier des gènes mosaïques (Makarenkov *et al.*, 2008 ; Boc *et al.*, 2011)

❑ Les algorithmes relatifs à la détection des transferts horizontaux de gènes sont librement accessibles à l'adresse URL suivante : *www.trex.uqam.ca* (le serveur web que nous avons développé)

RÉFÉRENCES

- **Boc, A.** et Makarenkov, V. (2003) New Efficient Algorithm for Detection of Horizontal Gene Transfer Events. In Benson G. et Page R. (Eds.). *WABI 2003, Algorithms in Bioinformatics*, Springer-Verlag, pp. 190-201.
- **Boc, A.**, Philippe, H. et Makarenkov, V. (2010a) Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Systematic Biology*, 59, 195-211.
- Delwiche, C.F. et Palmer, J. D. (1996) Rampant Horizontal Transfer and Duplication of Rubisco Genes in Eubacteria and Plastids. *Mol. Biol. Evol.*, 13, 873-882.
- Hallett, M., et Lagergren, J. (2001) Efficient algorithms for lateral gene transfer problems. In El-Mabrouk, N., Lengauer, T. et Sankoff, D. (Eds.), *Proceedings of the fifth annual international conference on research in computational biology*, ACM Press, New-York, pp. 149-156.
- Levenshtein, V. I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* , 10, 707–710.
- Makarenkov,V. (2001),T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17, 664-668.
- Makarenkov, V., **Boc, A.**, Delwiche, C.F. et Philippe, H. (2006) New efficient algorithm for modeling partial and complete gene transfer scenarios. In Batagelj, V., Bock, H.-H., Ferligoj, A. et Ziberna, A. (Eds.). *IFCS 2006, Series: Studies in Classification, Data Analysis, and Knowledge Organization*, Springer Verlag, pp. 341-349.
- Matte-Tailliez, O., Brochier,C., Forterre, P. et Philippe,H. (2002) Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.*, 19, 631-639.
- Robinson, D.R. et Foulds, L.R. (1981) Comparison of phylogenetic trees. *Mathematical Biosciences*, 53, 131-147.
- Than, C. Ruths, D. et Nakhleh, L. (2008) PhyloNet: A Software Package for Analyzing and Reconstructing Reticulate Evolutionary Relationships. *BMC Bioinformatics*, 9, 322.
- Woese, C.R., Olsen, G., Ibba, M. et Söll,D. (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.,* 64, 202-236.