Rapport de stage

Philippe Gambette

20 septembre 2004

Résumé

Les arbres de duplication sont des outils performants pour reconstruire de l'histoire des séquences de gènes répétées en tandem, problème introduit par Fitch en 1977. On a récemment trouvé des algorithmes pour les construire à partir de la séquence finale selon différents critères, ou les générer au hasard de façon uniforme. On s'intéresse ici aux réarrangements topologiques de ces arbres, les suppressions de feuilles, de sous-arbres, ou les mouvements SPR (Subtree Pruning and Regrafting), pour tester la solidité du modèle et trouver les transformations qui permettent de rester dans l'ensemble des arbres de duplication.

1 Introduction

La moitié de l'ADN humain est constitué de séquences répétées. Parmi elles, on trouve des séquences répétées en tandem, c'est à dire des segments adjacents presque identiques, issus d'une suite de *duplications en tandem* d'un segment original. Lorsque le motif répété est assez long (contient un gène), on considère que c'est la recombinaison inégale (figure 1) qui est à l'origine de cette duplication en tandem ([1], [2], [3] et [4]).

Ce processus est un mécanisme d'évolution des plus importants, puisque les gènes ainsi dupliqués, seront par la suite modifiés et pourront ainsi acquérir de nouvelles fonctionnalités. Reconstruire l'histoire des séquences dupliquées est donc fondamental pour comprendre leur évolution et leur fonction. L'outil de modélisation de cette évolution est l'*arbre de duplication*, qui permet de représenter la suite des événements de duplication subis pour obtenir la séquence finale, en étant éventuellement enraciné par le segment ancêtre. Ces arbres se distinguent en *arbres de duplication simple* si l'événement de duplication ne

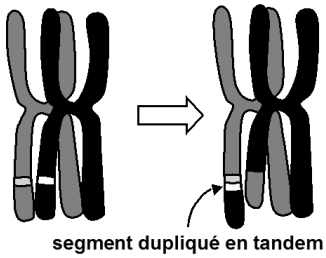

FIG. 1 – Recombinaison inégale pendant la prophase de la méiose

concerne toujours qu'un segment, et *arbres de duplication multiple* s'il existent des événements qui concernent plusieurs segments adjacents.

Si la reconstruction de l'histoire des duplications en tandem a été abordée par Fitch dès 1977 [2], les algorithmes pour l'automatiser sont assez récents. Retrouver l'arbre de duplication simple le plus parcimonieux est NP-complet [6], tout comme pour les phylogénies. On utilise donc des algorithmes heuristiques pour optimiser divers critères : la parcimonie [7] ou la distance [8, 9]. Les arbres obtenus par ces algorithmes peuvent être améliorés par l'application de méthodes de recherche locale. Elles donnent de bons résultats en phylogénie, ce qui nous permet d'espérer des améliorations sensibles aussi pour les arbres de duplication. Il faut donc étudier les mouvements topologiques sur ces arbres. Nous nous focaliserons sur les SPR (*Subtree Pruning Regrafting*), en identifiant les conditions de validité de ces mouvements : il faut vérifier qu'après ces réarrangements locaux on obtient bien un arbre de duplication, et pas seulement un arbre phylogénétique. En effet, si les arbres de duplication sont similaires aux phylogénies, le fait que les feuilles soient ordonnées (en tant que segments adjacents dans la séquence finale) implique que leur proportion par rapport aux phylogénies est faible (3% environ pour des arbres de neuf segments par exemple).

En outre, avant d'étudier les SPR, on pourrait observer le devenir d'un arbre de duplication si l'on enlève une ou plusieurs feuilles. Pour les arbres de duplication simple, le problème est trivial, il en va autrement pour certains arbres de duplication multiple qui ne le sont plus lorsqu'on enlève certaines feuilles. Identifier les feuilles qui font perdre leur caractère de duplication aux arbres permettra d'obtenir une proportion de feuilles qu'on ne peut enlever, ce qui pourra donner une indication sur la robustesse du modèle des arbres de duplication. En effet, ce modèle ne représente pas les délétions de segments éventuelles que la séquence aurait pu subir au cours de son histoire. Cette hypothèse se justifie en général de manière biologique, puisqu'en cas de suppression de segments, on a une perte de diversité, mais il est intéressant de voir comment réagit le modèle si elle n'est pas vérifiée.

Ainsi, après avoir défini formellement les arbres de duplication dans la section 2, nous examinerons l'impact des suppressions de feuilles (section 3) ou de sous-arbres (section 4) de ces arbres, puis nous caractériserons les différents types de SPR dans la section 5.

2 Modèle

2.1 Histoire de duplication

Le modèle de duplication que nous utiliserons est fondé sur le processus de recombinaison inégale que l'on suppose être l'unique mécanisme d'évolution (excepté les mutations ponctuelles) agissant sur les séquences. Le *slipped strand mispairing* [5, 7] donne des contraintes identiques sur les histoires de duplication, mais il s'applique principalement aux motifs courts sur lesquels agissent d'autres forces évolutives.

Soit $O = (1, 2, ..., n)$ un ensemble ordonné de segments représentant le locus tel qu'on peut l'observer aujourd'hui. Initialement, la séquence contenait un unique segment, et ce locus a grandi au cours d'une série de duplications. Une

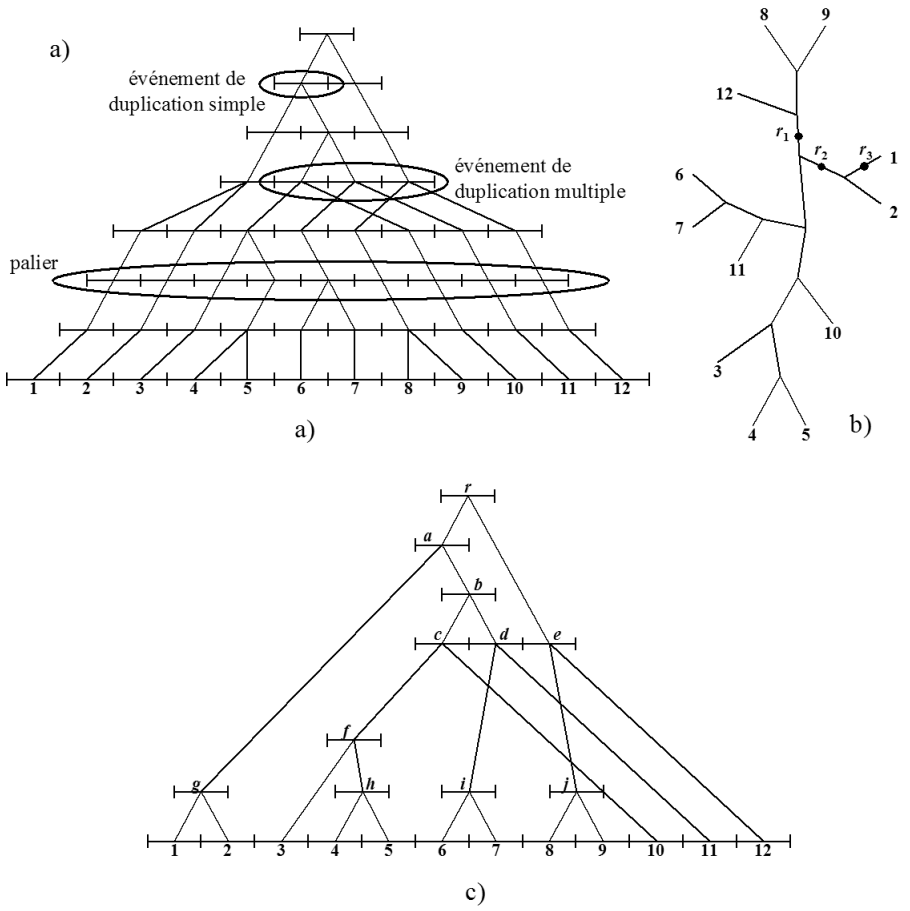

FIG. 2 – (a) Histoire de duplication : chaque segment représente un duplicat, les segments contemporains sont numérotés de 1 à 12. (b) Arbre de duplication (DT) : les points noirs indiquent les trois positions valides de la racine. (c) Arbre de duplication enraciné (RDT) correspondant à l'histoire (a) et dont la position de la racine est r_1 sur le DT (b).

histoire de duplication représente donc l'évolution de la séquence au cours du temps. Comme on peut le voir sur la figure 2(a), une *histoire de duplication* contient des *événements de duplication simple* qui transforment un segment en deux segments adjacents, et des *événements de duplication multiple* qui transforment k segments en $2k$ segments adjacents. Suivant ce modèle, une histoire de duplication est un arbre enraciné possédant n feuilles ordonnées, dans lequel chaque sommet interne de degré 3 appartient à un événement de duplication. L'ensemble des sommets de même hauteur représente une configuration complète et ordonnée du génome à un instant donné de son évolution. Nous appellerons ces ensembles des *paliers*. Deux sommets i et j sont *adjacents* ($i \prec j$) s'il existe un palier pour lequel i et j sont consécutifs.

2.2 Arbres de duplication DT et RDT

Cependant, si le mode d'évolution ne suit pas l'horloge moléculaire (ce qui est vrai dans la plupart des cas), l'ordre chronologique entre les événements de duplication de deux lignées différentes est impossible à retrouver à partir des séquences. Dans ce cas, nous ne pouvons inférer à partir des séquences contemporaines qu'un *arbre de duplication* (*duplication tree*, *DT*, figure 2(b)) ou un *arbre de duplication enraciné* (*rooted duplication tree*, *RDT*, figure 2(c)).

Contrairement aux phylogénies, un DT ne peut être enraciné sur n'importe laquelle de ses arêtes. La racine d'un arbre de duplication est nécessairement située sur le chemin entre les répétitions les plus distantes du locus, 1 et n ; de plus, la racine est toujours située "au dessus" des duplications multiples éventuelles. Ainsi, sur la figure 2(b), il n'existe que trois positions valides de la racine, qui ne peut être ancêtre direct de 12.

Un arbre de duplication est une phylogénie dont les feuilles sont ordonnées et dont la topologie est compatible avec au moins une histoire de duplication. Ceci suggère une définition récursive, qui reconstruit progressivement une histoire de duplication compatible, à partir d'une phylogénie T et d'un ordre sur les feuilles O. Pour toute paire (u, v) de O, nous utiliserons la notation $u \prec v$ pour exprimer que u est *adjacent* à v dans O. On définit une *cerise* (g, u, d) comme une paire de feuilles de T, (g, d), séparée par un unique sommet u dans T, et nous appelons $C(T)$ l'ensemble des cerises de T. La définition récursive suit une procédure inverse à celle de l'évolution : elle cherche un *événement de duplication visible*, *agglomère* cet événement et vérifie si l'*arbre réduit* est un arbre de duplication.

Définition 1
(T, O) est un *arbre de duplication* (RDT) de racine ρ si et seulement si :

(i) (T,O) contient uniquement ρ, ou

(ii) il existe dans $C(T)$ une série de cerises (g_i, u_i, d_i), $(g_{i+1}, u_{i+1}, d_{i+1}), \ldots,$ (g_k, u_k, d_k) avec $k \geq i$ et $g_i \prec g_{i+1} \prec ... \prec g_k \prec d_i \prec d_{i+1} \prec ... \prec d_k$ dans O, telle que (T', O') soit un arbre de duplication de racine ρ, où T' est obtenu à partir de T en enlevant $g_i, g_{i+1}, ..., g_k, d_i, d_{i+1}, ..., d_k$, et O' est obtenu en remplaçant $(g_i, g_{i+1}, ..., g_k, d_i, d_{i+1}, ..., d_k)$ par $(u_i, u_{i+1}, ..., u_k)$ dans O.

Si $k > i$, on dit alors que l'on a *aggloméré* l'*événement de duplication multiple* $(u_i, ..., u_k)$, l'opération inverse est la *dissociation* de l'événement de duplication.

Si $k = i$, on dit que l'on a *aggloméré la cerise* (g_i, u_i, d_i), et le noeud u_i est un *événement de duplication simple*. Nous appelons *1-RDT* un arbre de duplication ne comportant que des événements de duplication simple.

En construisant (T_1, O_1) à partir de (T, O) et pour tout $i > 1$ (T_{i+1}, O_{i+1}) à partir de (T_i, O_i) de même que l'on a construit (T', O') à partir de (T, O), on obtient un ensemble d'*arbres de duplication intermédiaires* pour (T, O) : $\{(T, O), (T_1, O_1), ..., (T_k, O_k)\}$ avec (T_k, O_k), l'arbre racine ρ.

Cette définition récursive, dont une version similaire existe pour les arbres DT (non enracinés) permet de vérifier si une phylogénie quelconque est un arbre de duplication, en agglomérant chaque événement de duplication trouvé. Si par ces agglomérations successives, on arrive à obtenir l'arbre racine, alors la phylogénie proposée était bien un arbre de duplication. Sinon, elle n'en était pas un. L'algorithme, appelé *PDH* (*Possible Duplication History*) est implémentable en

$O(n)$ [10, 11] (n étant le nombre de feuilles de l'arbre), et permet d'obtenir la liste des événements de duplication d'un RDT, ainsi que l'ordre gauche-droite des fils des sommets internes en vue de leur représentation comme dans la figure 2(c).

2.3 Adjacence possible

Dans un RDT, les sommets internes ne sont plus ordonnés entre eux horizontalement comme dans les histoires de duplication. Mais toutes les relations d'adjacence ne sont pas possibles. Par exemple, le RDT de la figure 2(c) n'est compatible avec aucune histoire de duplication pour laquelle les sommets f et j seraient adjacents. Par contre, les sommets g et c sont adjacents dans l'histoire de duplication (figure 2(a)) compatible avec ce RDT. Bien entendu, de nombreuses histoires de duplication sont compatibles avec un RDT donné. Nous dirons que deux sommets i et j ne faisant pas partie d'un même événement de duplication d'un RDT T sont *adjacents possibles* ($i \prec_p j$) si et seulement si il existe une histoire de duplication compatible avec T pour laquelle $i \prec j$. Il est impossible d'énumérer toutes les histoires de duplication associées à un RDT pour obtenir tous les liens d'adjacence possible. Une manière simple d'obtenir les relations d'adjacence possible consiste à vérifier, pour chaque couple de sommets, si l'algorithme PDH peut obtenir, en choisissant bien l'ordre des agglomérations successives, un arbre pour lequel ces deux sommets sont feuilles et adjacents. Sur la figure 2(c), on peut montrer que $g \prec_p 3$ en agglomérant l'événement g : dans l'arbre obtenu, nous avons $g \prec 3$. De même, on peut obtenir la relation $f \prec_p i$ en agglomérant successivement les événements h, f et i. Pour un arbre à n feuilles, on obtient un algorithme simple en $O(n^3)$ (n^2 couples de sommets à examiner, algorithme d'agglomération en O(n)) pour calculer l'ensemble des relations d'adjacence possible. Cette complexité peut être améliorée à l'aide des observations suivantes.

Soit (T, O) un RDT. L'événement de duplication E est un *événement antérieur commun* (*eac*) aux sommets i et j s'il contient deux sommets u et v ancêtres respectifs de i et de j (u pouvant être égal à v). Nous dirons que l'événement E est *le plus proche événement antérieur commun* (*peac*) de i et j si aucun sommet de E n'est l'ancêtre d'un sommet appartenant à un *eac* de i et j. Le peac est unique pour chaque paire de sommets. Sur la figure 2(c), les sommets 2 et 10 possèdent a comme peac, en revanche, les sommets 4 et 11 possèdent comme peac l'événement (c,d,e).

Pour obtenir toutes les relations d'adjacence possible, nous allons calculer le peac de tous les couples de sommets i et j adjacents, que ce soient des feuilles ou des éléments consécutifs d'un même événement. A l'aide de celui-ci, nous otenons deux *chaînes d'ancêtres* de i et j : C_i (resp. C_j), composées de tous les ancêtres de i (resp j) descendant du peac de i et j. La définition du peac implique qu'aucun sommet de C_i ne peut être dans le même événement qu'un élément de C_j. L'algorithme PDH peut donc agglomérer, de manière indépendante, les éléments des deux chaînes. Nous pouvons donc obtenir un RDT possédant un élément de C_i, s_i, et un élément de C_j, s_j, pour feuilles. Comme i et j étaient adjacents dans le RDT étudié, s_i et s_j le seront aussi dans le RDT obtenu après agglomération par l'algorithme PDH, donc si $s_i \prec_p s_j$, et cette relation est vérifiée pour chaque paire d'éléments de C_i, C_j.

Remarquons que chaque sommet appartient à au plus deux chaînes d'an-

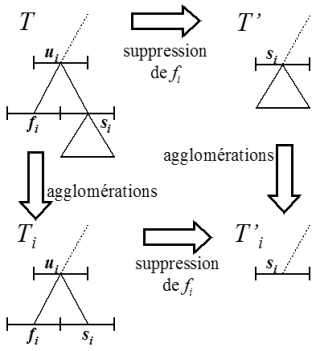

FIG. 3 – Suppression d'une feuille dans un 1-RDT

cêtres, au sens défini ci-dessus, l'une correspondant au sommet gauche dans la paire (i, j) considérée, et l'autre au sommet droit. La détermination de toutes les chaînes est donc linéaire sur n, tandis que la détermination de toutes les relations d'adjacence possible est linéaire sur le nombre des relations, c'est à dire en $O(n^2)$.

3 Suppression de feuilles

3.1 Suppression de feuilles dans un 1-RDT

Théorème 1
Si l'on supprime une ou plusieurs feuilles d'un arbre 1-RDT, on obtient un arbre 1-RDT.

Démonstration
Soit (T, O), l'arbre initial, avec $O = (f_1, ..., f_n)$, et (T',O'), l'arbre résultant de la suppression d'une feuille f_i (figure 3). Supposons par exemple que f_i est fils gauche de u_i. Par définition, il existe une manière d'agglomérer (T, O) telle que s_i soit une feuille. On appelle (T_i, O_i) l'arbre intermédiaire obtenu. Les mêmes agglomérations s'appliquent à (T', O') pour obtenir (T'_i, O'_i). Si maintenant nous agglomérons la cerise (f_i, s_i) de (T_i, O_i), nous obtenons justement l'arbre (T'_i, O'_i) qui est donc bien un 1-RDT. On en déduit que T' est bien un 1-RDT. ■

3.2 Suppression de feuilles dans un 1-DT

Théorème 2
Si l'on supprime une ou plusieurs feuilles d'un arbre 1-DT, on obtient un arbre 1-DT.

Démonstration
Soit un arbre 1-DT. On peut l'enraciner pour obtenir un 1-RDT. Alors si l'on supprime la feuille voulue dans ce 1-RDT, l'arbre obtenu reste un 1-RDT, et son équivalent non enraciné est l'arbre 1-DT original privé de sa feuille. ■

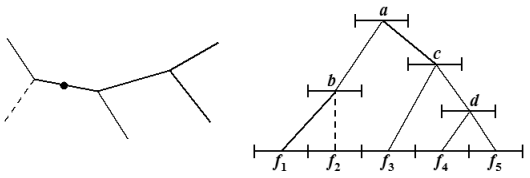

FIG. 4 – Suppression d'une racine dans un 1-DT et 1-RDT associé

3.3 Suppression d'une feuille dans un RDT quelconque

Introduisons d'abord une définition qui permet de distinguer plusieurs types parmi les feuilles ou noeuds issus de duplication multiple. Soit un événement de duplication multiple $(u_1, ..., u_k)$, avec pour tout $1 < i < k$, g_i et d_i les fils respectivement gauche et droit de u_i. g_1 et d_k sont des feuilles ou noeuds *extrêmes* issus de la duplication multiple, g_k et d_1 sont des feuilles ou noeuds *centraux* et $g_2...g_{k-1}$ et $d_2...d_{k-1}$ sont des feuilles ou noeuds *internes* issus de l'événement de duplication.

Théorème 3
Soit (T, O), un arbre de duplication enraciné. Soit f, la feuille à supprimer dans cet arbre. Soit p, son père, et p' le père éventuel de p, u étant alors le second fils de p'. Alors (T', O'), l'arbre résultant de la suppression de la feuille f dans (T, O) est un arbre de duplication si et seulement si l'une des trois *conditions de délétion* suivantes est vérifiée :

(i) f n'est pas une feuille issue de duplication multiple.

(ii) f est une feuille centrale issue de duplication multiple.

(iii) f est une feuille extrême issue de duplication multiple, u est adjacent possible d'une des deux feuilles extrêmes issues de cette duplication multiple, et p est supprimable.

Démonstration
Afin de démontrer ce résultat, nous devons déterminer à laquelle des 6 catégories suivantes la feuille f, supposée feuille gauche (on traite le cas feuille droite par symétrie) appartient :

1- f est n'est pas issue de duplication multiple.

2- f est feuille centrale issue d'une duplication multiple.

3- f est feuille interne issue d'une duplication multiple.

4- f est une feuille extrême issue de duplication multiple, et u n'est pas adjacent possible d'une feuille extrême issue de cette duplication.

5- f est une feuille extrême issue de duplication multiple, et u est adjacent possible d'une feuille extrême issue de cette duplication, et p est supprimable.

6- f est une feuille extrême issue de duplication multiple, et u est adjacent possible d'une feuille extrême issue de cette duplication, et p n'est pas supprimable.

Nous allons maintenant vérifier qu'après suppression d'une feuille dans les cas 1, 2, et 5 (correspondant à i, ii, et iii), l'arbre obtenu est un RDT correct, et pas dans les cas 3, 4 et 6. Dans tous les cas, on considérera que l'on a déjà réalisé

les agglomérations nécessaires pour que l'événement de duplication $(u_1, ..., u_k)$ (constitué des cerises (g_1, u_1, d_1), ... , (g_k, u_k, d_k)) duquel f est issue soit prêt à être aggloméré. Il s'agit donc de savoir si après suppression de la feuille, l'agglomération de ce qui reste de cet événement de duplication est toujours possible (si tel est le cas, on peut alors "redérouler" les dissociations pour reconstruire l'arbre privé de f_i qui est alors bien un RDT).

1 - f n'est pas issue directement de duplication multiple, donc en la supprimant, de même que dans la démonstration du 3.1, on se contente de supprimer un événement de duplication : l'arbre reste un RDT.

2 - f est une des deux feuilles centrales de la duplication multiple, elle appartient à la cerise (f, p, s) (figure 5). L'événement de duplication duquel elle était issue contient désormais un noeud de moins (le noeud p), et peut bien être aggloméré, et s est désormais à la place de p, donc l'événement de duplication dont s est issu pourra aussi être aggloméré : l'arbre reste donc un RDT.

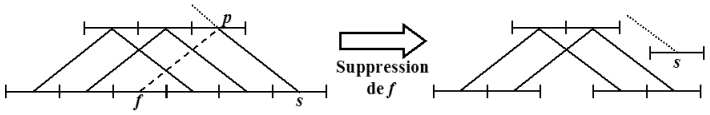

FIG. 5 – Cas 2 : Suppression d'une feuille centrale

3 - f n'est ni une feuille centrale, ni extrême, de la duplication multiple, donc elle appartient à une cerise (g_i, u_i, d_i) qui est entourée par deux autres : $(g_{i-1}, u_{i-1}, d_{i-1})$ et $(g_{i+1}, u_{i+1}, d_{i+1})$. Ainsi, après suppression de f_i, il est impossible d'agglomérer l'événement de duplication multiple contenant les noeuds u_{i-1} et u_{i+1}, puisque d_i s'intercale entre d_{i-1} et d_{i+1} : l'arbre obtenu n'est donc pas un RDT.

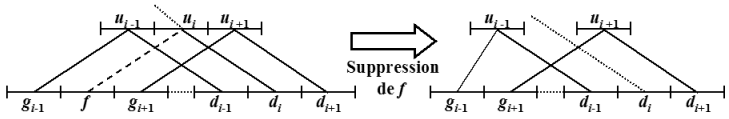

FIG. 6 – Cas 3 : Suppression d'une feuille interne

4 - Il faut agglomérer la duplication multiple privée de p. Or d_1 ne peut être issu de cette duplication multiple. En effet, s'il l'était, u le serait aussi, donc serait adjacent possible d'une des feuilles extrêmes de la duplication, ce qui n'est pas le cas. d_1, en s'insérant ainsi parmi les feuilles de la duplication multiple, empêche donc son agglomération : l'arbre obtenu n'est pas un RDT.

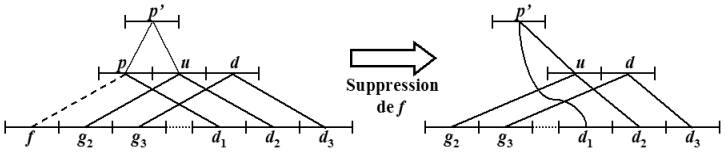

FIG. 7 – Cas 4 : u n'est pas adjacent possible des feuilles extrêmes.

5 (droit) - Supposons que p est fils droit de p'. u et f étant adjacents possibles, poursuivons les agglomérations jusqu'à ce que $u \prec f$. Après suppression

de f, c'est alors u et g_2 qui sont adjacentes, et on souhaite "déplacer" la cerise (u, p', d_1) de l'événement de duplication auquel elle appartient actuellement vers l'événement de duplication multiple duquel f était issue. L'événement de duplication duquel p est issu reste valide après la suppression de p si et seulement si celui-ci est supprimable. Or p est supprimable (en effet, $u \prec f$, donc u adjacent possible de p. Or u et p ont même père, donc ils sont nécessairement issus d'une duplication simple, et p est bien supprimable). On peut alors "déplacer" la cerise (u, p', d_1) comme on le souhaitait (figure 8) : l'arbre obtenu est un RDT.

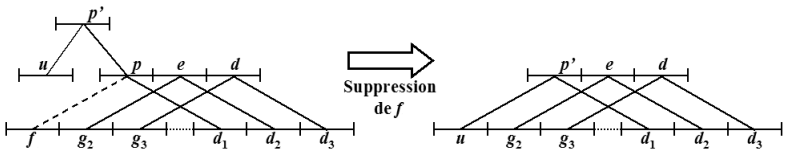

FIG. 8 – Cas 5 droit : Feuille extrême supprimable avec père droit

5 et 6 (gauche) - Supposons que p est fils gauche de p'. u et d_k étant adjacents possibles, poursuivons les agglomérations jusqu'à ce que u et d_k soient des feuilles adjacentes. Après suppression de f, on souhait "déplacer" la cerise (d_1, p', u) à l'extrême droite l'événement de duplication multiple duquel f était issue. L'événement de duplication duquel p est issu reste valide après la suppression de p si et seulement si celui-ci est supprimable. Si c'est le cas, on peut alors "déplacer" la cerise (d_1, p', u) comme on le souhaitait (figure 9) : l'arbre obtenu est un RDT. Si ce n'est pas le cas, la cerise (d_1, p', u) est nécessaire à

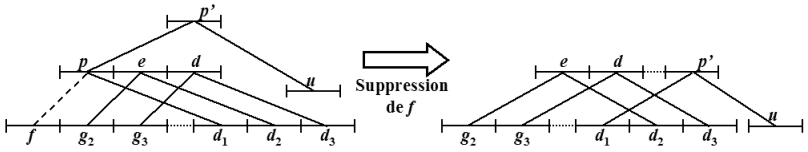

FIG. 9 – Cas 5 gauche : Feuille extrême supprimable avec père gauche

l'agglomération de l'événement de duplication multiple duquel p est issu, donc ne peut rejoindre l'événement de duplication duquel f était issue et empêche son agglomération (figure 10) : l'arbre obtenu n'est pas un RDT. ■

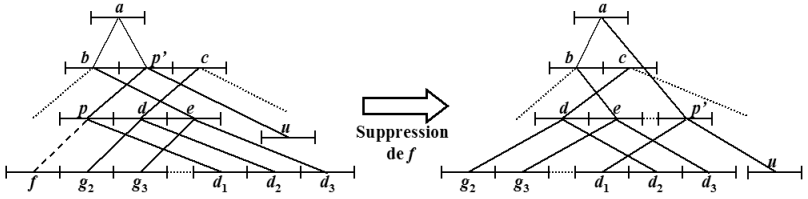

FIG. 10 – Cas 6 : Père non supprimable

On peut remarquer que bien qu'elle soit valide, la suppression des feuilles dans le cas iii) induit de grands changements topologiques sur l'arbre de duplication, puisque l'événement de duplication duquel f est issue n'est pas le seul modifié (figure 11).

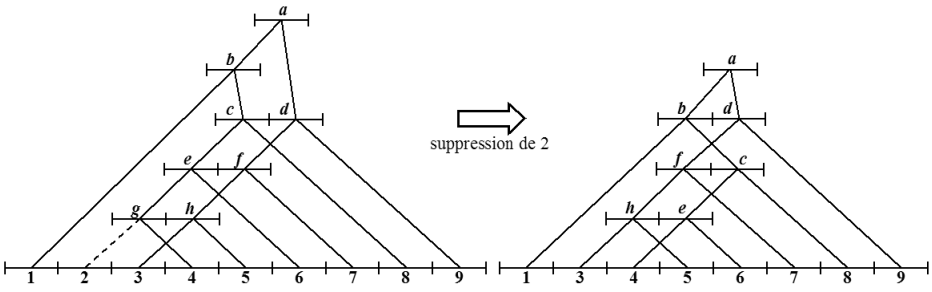

FIG. 11 – La suppression d'une feuille d'un RDT peut modifier plusieurs événements de duplication

3.4 Aspects quantitatifs

Nous cherchons à quantifier le pourcentage de suppressions de feuilles qui permettent de garder des arbres de duplication. Nous avons généré 1000 arbres de n feuilles pour $n = 4, 5, 6, 7, 8, 9, 10, 12, 15, 18, 20, 25, 30, 40$, et 100 arbres de n feuilles avec $n = 50, 60, 70, 80, 90, 100, 200$. On constate que le pourcentage d'arbres qui ne sont plus des arbres de duplication quand on enlève une feuille augmente progressivement pour atteindre environ 26% pour les arbres de plus de 100 feuilles. La proportion d'arbres résistants à toute délétion de feuille passe de 50% pour des arbres de 6 feuilles à 20% pour des arbres de 10 feuilles et 5% pour des arbres de 15 feuilles.

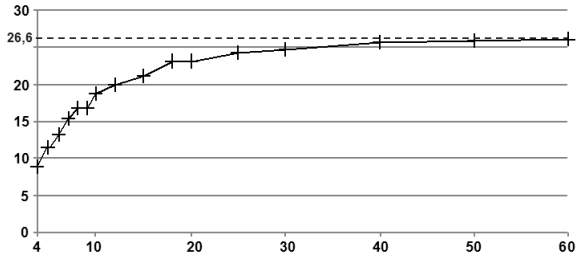

FIG. 12 – Pourcentage d'arbres qui perdent le caractère de duplication après suppression d'une feuille, en fonction du nombre de feuilles

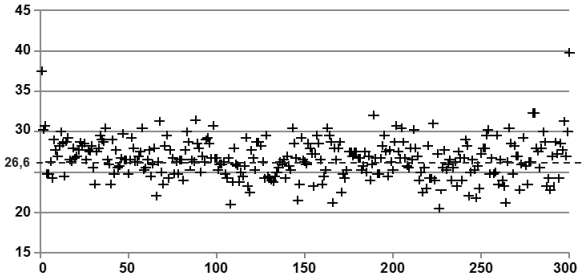

FIG. 13 – Pourcentage d'arbres de 300 feuilles qui perdent le caractère de duplication après suppression d'une feuille, en fonction de la feuille enlevée

Ces informations ont été trouvées à l'aide d'une implémentation (C++) de

l'algorithme de vérification décrit dans [10] complété pour réaliser les tests, ainsi qu'un programme Caml qui prend en entrée un fichier texte contenant des arbres codés sous forme de ses feuilles bien parenthésées, en déduit les arbres représentés, enlève une feuille et retraduit l'arbre en format parenthésé. Les résultats obtenus montrent donc que le modèle n'est pas extrêmement résistant aux délétions.

Nous avons distingué 6 cas dans la démonstration des conditions de délétion d'une feuille. Leur proportion est représentée dans la figure 14 : 500 arbres de 8, 10, 12, 18 et 24 feuilles ont été générés, ainsi que trois lots de 100 arbres de 400 feuilles, afin d'avoir une idée des variations de résultats entre des tests portant sur des arbres identiques. Les trois zones sombres représentent les cas où la feuille n'est pas supprimable, et les trois zones claires, le contraire. Le cas 6 qui entraîne de lourdes modifications des événements de duplication apparaît en général dans moins de 1% des cas, et il est nécessaire de vérifier les conditions de délétion sur le père dans environ 8% des cas.

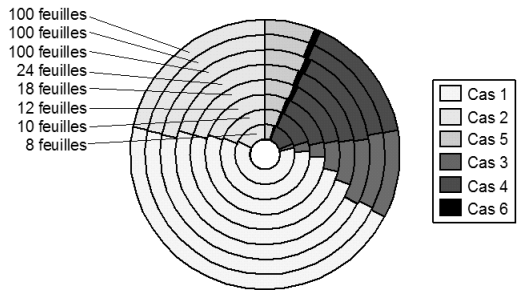

FIG. 14 – Proportion des différents cas présentés dans la démonstration des conditions de délétion d'une feuille

Les conditions de suppression de feuilles dans des arbres DT n'ont en revanche pas été trouvées. Si les conditions de suppression pour les RDT restent valides pour les DT, certains DT qui, enracinés, perdent leur caractère d'arbre de duplication après suppression d'une feuille, peuvent le retrouver en étant enracinés autrement. Ce n'est jamais le cas dans le cas 3 vu ci-dessus (suppression d'une feuille interne), mais ça arrive dans les cas 4 et 6.

4 Suppression d'un sous-arbre

Théorème 4
Soit (T, O), un arbre de duplication enraciné. Soit r, la racine du sous-arbre à supprimer dans cet arbre. L'arbre (T', O') résultant de la suppression du sous-arbre de racine r dans (T, O) est un arbre de duplication si et seulement si l'une des trois conditions de délétion (c.f. 3.3) est vérifiée.

Démonstration
Il existe une manière d'agglomérer (T, O) en (T_i, O_i) tel que r y soit une feuille (figure 15). Pour toute agglomération d'événement de duplication $(u_1, ..., u_l)$ effectuée :
Soit $J = \{j_1, ..., j_m\}$ tel que $j \in J \iff u_j$ n'appartient pas au sous-arbre

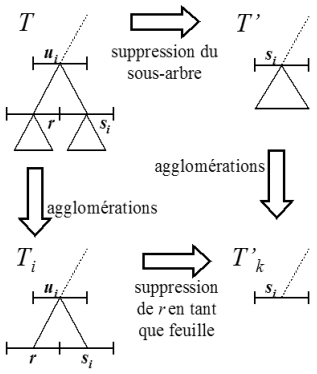

FIG. 15 – Suppression d'un sous-arbre dans un 1-RDT

de racine r. Si $J \neq \emptyset$ alors on peut effectuer dans (T', O') (ou l'arbre intermédiaire qui en dérive suite aux agglomérations déjà réalisées) l'agglomération de $(u_{j_1}, ..., u_{j_m})$.
On obtient finalement l'arbre intermédiaire (T'_k, O'_k). Or (T'_k, O'_k) est en fait l'arbre (T_i, O_i) duquel on a supprimé la feuille r. (T'_k, O'_k) est donc un RDT si et seulement si r satisfait une des conditions de délétion énoncées en 3.3. Finalement, (T', O') est donc bien un RDT si et seulement si r satisfait une des conditions de délétion. ■

5 Réarrangements topologiques

Il existe de nombreuses manières d'effectuer des réarrangements topologiques sur les phylogénies, comme les *NNI* (*Nearest Neighbour Interchange*) ou les *SPR* (*Subtree Pruning Regrafting*).

Le NNI consiste à permuter deux sous-arbres autour d'une arête interne. Ce mouvement permet d'explorer l'espace des phylogénies, c'est à dire qu'il existe une suite de mouvements NNI permettant de transformer toute phylogénie P_1 en une phylogénie P_2 quelconque. Toutefois, lorsque l'on applique un NNI à un RDT, l'arbre obtenu n'est pas toujours un RDT valide.

Le mouvement SPR (*Subtree Pruning Regrafting*) quant à lui, consiste à détacher un sous-arbre de racine r d'un arbre de duplication (T, O) et à le rattacher par sa racine sur une arête (x, y) de l'arbre résultant, x étant le père de y, avec, en notant p le père de f, $x \neq p$ et $y \neq p$ (on n'étudie pas les SPR-identité). Ce SPR est noté : $\text{SPR}(r, (x, y), (T, O))$. L'ensemble des NNI est inclus dans l'ensemble des SPR. Ainsi, tous les SPR sur les RDT ne permettent pas d'obtenir un RDT : un SPR est dit *valide* si l'arbre résultant est un arbre de duplication. On peut démontrer que contrairement aux NNI, les SPR valides permettent d'explorer l'espace des DT [12]. En effet, il est possible de trouver une suite de SPR pour passer d'un RDT à un autre, en passant par une forme intermédiaire : un 1-RDT appelé *peigne*.

Nous allons caractériser les SPR valides, en remarquant tout d'abord que le problème de validité des SPR sur des sous-arbres se réduit au problème de validité des SPR sur les feuilles. Puis nous identifierons les contraintes permettant de déplacer une feuille dans un RDT.

Soit (T, O) un arbre de duplication. On introduit *l'arbre d'agglomération minimale* pour r et (T, O) : c'est l'arbre dans lequel r est feuille obtenu après un minimum d'agglomérations.

Théorème 5
Soit (T, O) un arbre de duplication, r, x et y des sommets de T, et (T_i, O_i), l'arbre d'agglomération minimale pour r et (T, O). SPR$(r, (x, y), (T, O))$ est valide si et seulement si SPR$(r, (x, y), (T_i, O_i))$ est valide.

Démonstration
Montrons tout d'abord que si x appartient à l'un des événements agglomérés lors du passage de (T, O) à (T_i, O_i), le mouvement SPR est invalide. Supposons que le mouvement est valide. Essayons alors d'agglomérer l'événement de duplication contenant $x : E_x$. Comme (T_i, O_i) est l'arbre d'agglomération minimale pour r et (T, O), il est nécessaire d'agglomérer l'événement E_x avant d'agglomérer E_r. Or pour agglomérer E_x après le SPR$(r, (x, y), (T, O))$, il est nécessaire d'agglomérer r qui est fils d'un noeud issu de E_x : absurde ! Donc le mouvement est invalide. Ainsi, les SPR portant sur des arêtes (x, y) de (T, O) mais pas de (T_i, O_i) sont invalides. On peut donc s'intéresser uniquement aux SPR portant sur les arêtes (x, y) de (T_i, O_i).

Soit (T', O') l'arbre après le SPR. Il existe une manière d'agglomérer (T, O) en (T_i, O_i) tel que r soit une feuille (figure 16). Les mêmes agglomérations s'appliquent à (T', O') pour obtenir l'arbre intermédiaire (T'_i, O'_i). Or si l'on effectue le SPR$(r, (x, y), (T_i, O_i))$, on obtient (T'_i, O'_i) qui est donc un RDT ssi le SPR était valide. Finalement, (T', O') est donc bien un RDT si et seulement si le SPR sur la feuille r est valide. ■

Théorème 6
Soit (T,O) un RDT, f une feuille, et x et y deux sommets de T. Le SPR(f, (x, y), (T, O)) n'est pas valide si :

1 - f est une feuille interne issue de duplication simple.

Démonstration
Considérons l'événement de duplication E_f duquel f est issue, dont les feuilles sont $f_1, \ldots, f_{i-1}, f, f_{i+1}, \ldots, f_k$. Pour que l'événement puisse être aggloméré, il est nécessaire que la cerise contenant f reste inchangée, puisqu'elle doit se trouver entre les cerises $(f_{i-1}, \ldots, \ldots)$ et $(f_{i+1}, \ldots, \ldots)$. Aucun SPR portant sur f n'est donc valide. ■

Théorème 7
Soit (T,O) un RDT. Pour toute duplication multiple, on note ci-dessous $g_1, \ldots, g_k, d_1, \ldots, d_k$ les feuilles qui en sont issues. Le SPR(f, (x, y), (T, O)) est valide si :

1 - f est supprimable, f et y sont adjacents possibles.

2 - f est supprimable, et v est telle que $f \prec v$ et $y \prec_p v$ (ou $v \prec f$ et $v \prec_p y$).

3 - f est feuille centrale gauche issue de duplication multiple (resp. droite) et $y \prec_p g_1$ (resp. $d_k \prec_p y$).

4 - f est supprimable selon les conditions de délétion i) et ii) (voir section 3.3), v est une feuille adjacente à f, E est le peac de f et v, a_f est l'ancêtre de

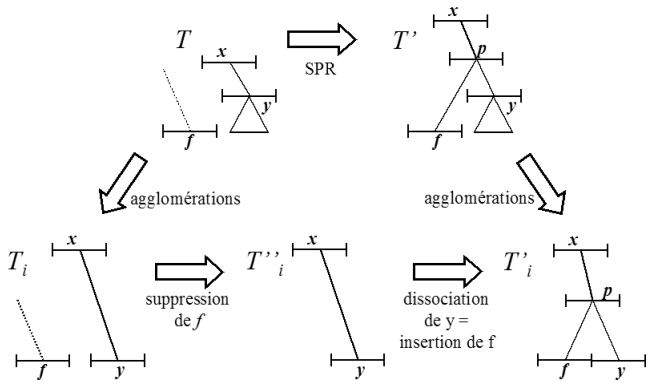

FIG. 16 – SPR1

f fils d'un élément de E, a_v est défini de même pour v, $a_f \neq f$, a_f ou a_v est extrême gauche (resp. droit) issu de E, et y central gauche (resp. droit) issu de E.

5 - f est supprimable selon les conditions de délétion i) et ii) (voir section 3.3), v est une feuille adjacente à f, E est le peac de f et v, a_f est l'ancêtre de f fils d'un élément de E, a_v est défini de même pour v, $a_f \neq f$, et a_f n'est pas le père de f, a_f et a_v sont centraux issus de E, ou appartiennent à une événement de duplication simple, et que $y \prec_p g_1$ ou $d_k \prec_p y$.

Démonstration

1 - Soit (T, O) l'arbre avant SPR, et (T', O'), l'arbre obtenu après. On considère que $f \prec_p y$ (second cas traitable par symétrie), et on effectue des agglomérations dans l'arbre (T, O) jusqu'à ce que $f \prec y$: on obtient l'arbre intermédiaire (T_i, O_i). On effectue les mêmes agglomérations sur l'arbre (T', O') pour obtenir l'arbre intermédiaire (T'_i, O'_i). Or, en supprimant la feuille f de (T_i, O_i), on garde un arbre de duplication (T''_i, O''_i). Agrafer la feuille f sur l'arête (x, y) revient alors à dissocier y pour créer la cerise (f, p, y), p constituant un événement de duplication simple, donc on obtient un arbre de duplication. Or cet arbre est l'arbre (T'_i, O'_i) donc finalement, (T', O') est bien un arbre de duplication.

2 - Soit (T, O) l'arbre avant SPR, et (T', O'), l'arbre obtenu après. On considère que $f \prec v$ et $y \prec_p v$ (second cas traitable par symétrie). Montrons que l'on peut décomposer ce SPR en deux SPR de type 1 : l'arrachage de f puis son insertion sur l'arête (v, p') puis son arrachage et insertion sur l'arête (x, y). Après le premier SPR, on obtient l'arbre (T'', O''). On a alors $f \prec v$, f et v sont issues d'une duplication simple, donc f est supprimable.
De plus, on a désormais $y \prec_p f$. En effet, considérons l'arbre (T, O) duquel on a supprimé f : $y \prec_p v$ donc on agglomère les événements de duplication jusqu'à ce que $y \prec v$ (arbre (T_i, O_i)). Si l'on agglomère les mêmes événements dans l'arbre (T'', O''), on obtient un arbre (T''_i, O''_i). Or l'arbre obtenu après insertion de f sur l'arête (v, p') (ce qui revient à dissocier la feuille v pour créer la cerise (f, p, v)) est exactement l'arbre (T''_i, O''_i). Ainsi, par une suite d'agglomérations et de dissociations, on peut passer de (T', O') à (T''_i, O''_i), où $y \prec f$ donc $y \prec_p f$ dans (T', O').

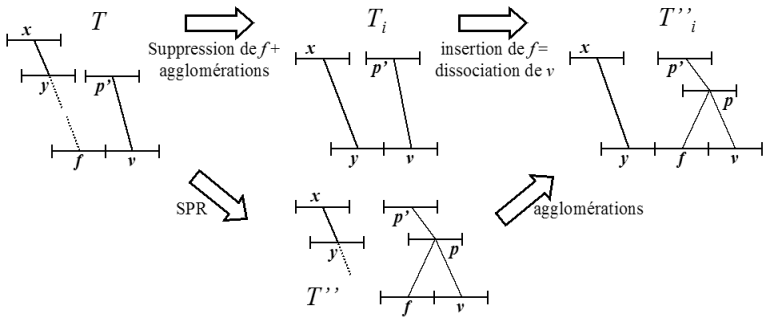

FIG. 17 – SPR2

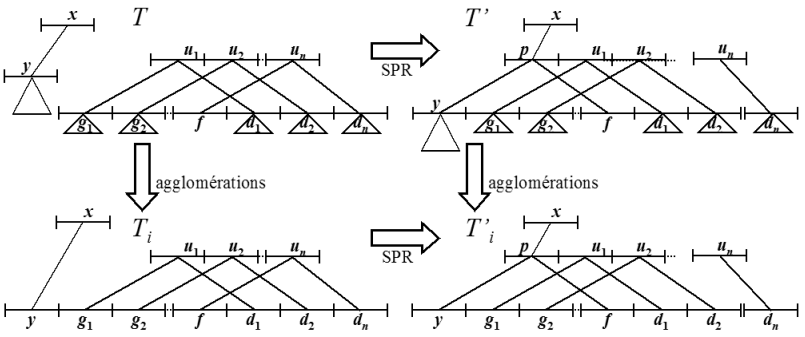

FIG. 18 – SPR3

On peut donc bien effectuer un nouveau SPR de type 1, afin d'obtenir l'arbre (T', O') qui est donc un arbre de duplication.

3 - Soit (T, O) l'arbre avant SPR, et (T', O'), l'arbre obtenu après. On considère que f feuille centrale gauche issue de duplication multiple, et, avec g_1 la feuille extrême gauche issue de cette duplication, $y \prec_p g_1$ (second cas traitable par symétrie). On effectue des agglomérations dans l'arbre (T, O) jusqu'à ce que $y \prec g_1$, et que l'événement de duplication multiple duquel y et g_1 sont issues soit prêt à être aggloméré : on obtient (T_i, O_i). On effectue les mêmes agglomérations sur l'arbre (T', O') pour obtenir l'arbre intermédiaire (T'_i, O'_i). Or, en supprimant la feuille f, on garde un arbre de duplication, qui le reste lorsque l'on insère la feuille f sur l'arête (x, y) créant ainsi la cerise (y, p, f) qui vient rejoindre l'événement de duplication duquel f était initialement issue (dans (T, O)). Ainsi, l'arbre obtenu est bien de duplication. Or c'est l'arbre (T'_i, O'_i) donc (T', O') est bien un arbre de duplication.

4 - Soit (T, O) l'arbre avant SPR, et (T', O'), l'arbre obtenu après. On considère que a_f est noeud extrême gauche issu de duplication multiple (second cas traitable par symétrie). On supprime f dans l'arbre. Si a_f est père de f, il est supprimé lors de la suppression de f, donc, en appelant f' le fils de a_f différent de f, on notera $a_f = f'$ dans la suite de la démonstration. $f \neq a_f$, donc l'événement de duplication duquel a_f est issu reste inchangé. On agglomère alors l'arbre (T, O) jusqu'à ce que $a_f \prec a_v$ (resp. $a_v \prec a_f$), et que l'événement de duplication multiple duquel a_f et a_v sont issues soit prêt à

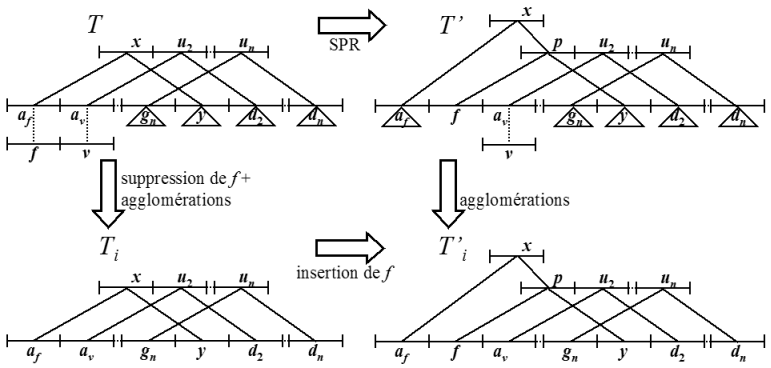

FIG. 19 – SPR4a

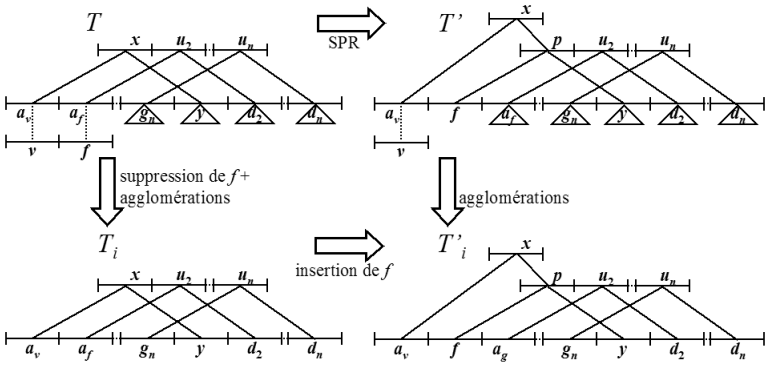

FIG. 20 – SPR4b

être aggloméré : on obtient l'arbre (T_i, O_i). On peut effectuer ces agglomérations sur (T', O') pour obtenir (T'_i, O'_i). En effet, les événements agglomérés ne peuvent contenir à la fois un ancêtre de v et un ancêtre de f (voir la fin de la section 2.3), et ne peuvent concerner des ascendants de y, donc ils existent et sont agglomérables aussi bien dans (T', O') que dans (T, O). On insère alors f sur l'arête (x, y) : x quitte l'événement de duplication duquel a_f et a_v étaient issus, et la cerise (f, p, y) vient s'insérer à l'extrême gauche de cette duplication, comme montré dans la figure 19 (respectivement dans la figure 20). Ainsi, l'arbre obtenu est bien de duplication. Or c'est l'arbre (T'_i, O'_i) donc (T', O') est bien un arbre de duplication.

5 - Soit (T, O) l'arbre avant SPR, et (T', O'), l'arbre obtenu après. On considère que a_f est noeud central gauche issu de duplication multiple, et a_v noeud central droit (second cas traitable par symétrie). On appellera par la suite *ancêtres de* f les ancêtres de f dans cet arbre (T, O) descendants du peac de f et v, de même pour v. On supprime f dans l'arbre. Si a_f est père de f, il est supprimé lors de la suppression de f, donc, en appelant f' le fils de a_f différent de f, on notera $a_f = f'$ dans la suite de la démonstration. $f \neq a_f$, donc l'événement de duplication duquel a_f est issu reste inchangé. On effectue alors une première phase d'agglomérations dans l'arbre (T, O) jusqu'à ce que $a_f \prec a_v$, et que l'événement de duplication multiple duquel a_f

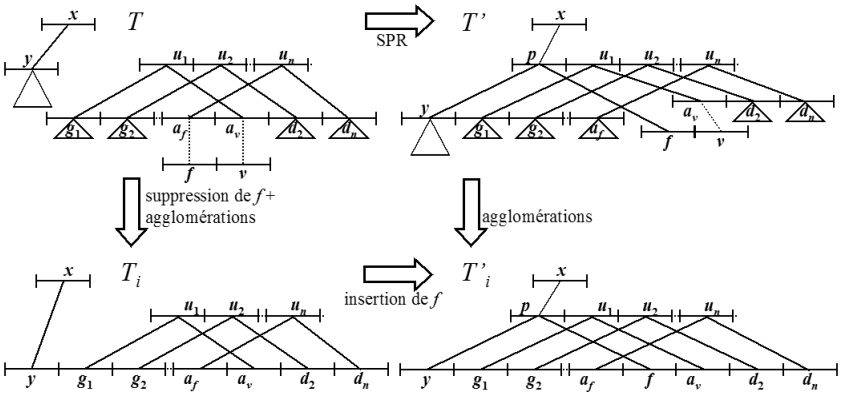

FIG. 21 – SPR5

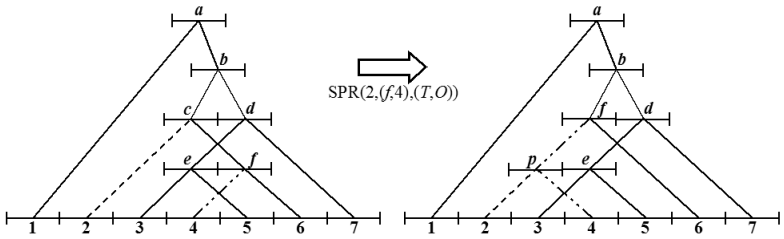

FIG. 22 – SPR valide malgré la feuille non supprimable

et a_v sont issues soit prêt à être aggloméré. On peut effectuer cette première phase d'agglomérations sur (T', O'). En effet, les événements agglomérés ne peuvent contenir à la fois un ancêtre de v et un ancêtre de f (voir la fin de la section 2.3), et ne peuvent concerner des ascendants de y, donc ils existent et sont agglomérables aussi bien dans (T', O') que dans (T, O). Supposons que $y \prec_p g_1$ ($d_k \prec_p y$ traitable par symétrie). On poursuit les agglomérations jusqu'à ce que $y \prec g_1$, pour obtenir (T_i, O_i) (figure 21). On peut effectuer cette seconde phase d'agglomérations sur l'arbre issu de la première phase d'agglomérations sur (T', O'), afin d'obtenir l'arbre (T'_i, O'_i). En insérant la feuille f sur l'arête (x, y), on crée la cerise (y, p, f) qui vient rejoindre à l'extrême gauche l'événement de duplication duquel a_f et a_v sont issus. Ainsi, l'arbre obtenu est bien de duplication. Or c'est l'arbre (T'_i, O'_i) donc (T', O') est bien un arbre de duplication.

■

Des contrexemples aux propriétés suivantes ont été trouvés :

- SPR(f, (x, y), (T, O)) est valide implique f est supprimable (figure 22).
- SPR(f, (x, y), (T, O)) est valide si : f est supprimable selon les conditions de délétion iii) (voir section 3.3), v est une feuille adjacente à f, E est le peac de f et v, a_f est l'ancêtre de f fils d'un élément de E, a_v est défini de même pour v, $a_f \neq f$, et a_f n'est pas le père de f, a_f et a_v sont centraux issus de E, ou appartiennent à une événement de duplication simple, et que $y \prec_p g_1$ ou $d_k \prec_p y$ (figure 23).

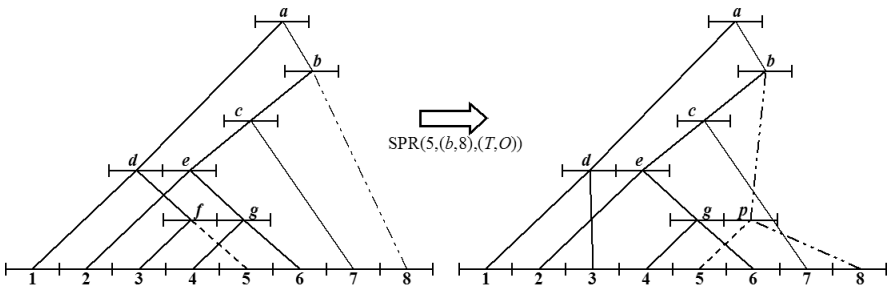

FIG. 23 – SPR invalide

Ainsi, les résultats trouvés pour les SPR ne sont pas très intéressants, puisque l'on n'arrive même pas à mettre en relation ces conditions avec celles de suppression de feuille. Par rapport au nombre total de SPR, pour des arbres de 10 feuilles (test portant sur 100 arbres), il y a 22% de SPR valides environ, soit 16% dans le cas 1, 0.7% dans le 2, 0.3% dans le 3, 0.7% dans le 4, 0.5% dans le 5. Ainsi, 4.6% du total des SPR (soit 20% des SPR valides), n'a pas été détectés par nos cinq cas. Ce sont peut-être des SPR mettant en jeu des feuilles non supprimables.

Ces conditions de validité ses SPR présentent donc peu d'intérêt, puisqu'elles ne couvrent pas tout l'espace des RDT, ni même l'espace des RDT valides, ou invalides. Mais elles seraient peut-être inutiles quand bien même on en aurait la liste exhaustive. Les conditions à vérifier sont en effet trop complexes, et le gain de temps pour les programmes sera limité.

6 Conclusion

Les résultats trouvés sont donc plus complexes qu'espéré. C'est le fait que des cerises puissent quitter des événements de duplication pour s'accrocher à d'autres dans le cas de suppression de feuilles ou de SPR qui complique le problème. Les questions de délétions de feuilles ou de SPR dans les arbres non enracinés sont tout aussi obscures.

Toutefois, la robustesse aux délétions du modèle de l'arbre de duplication est établie. En outre, nous n'avons ici qu'une évaluation pessimiste de cette robustesse (de 75% environ, rappelons-le), puisque les arbres de duplication réellement construits à ce jour semblent contenir moins de duplications multiples. Il faudrait donc corriger les résultats quantitatifs de supprimabilité obtenus en utilisant les données biologiques. Mais peu d'arbres de duplication sont disponibles et vérifiés par les biologistes. Un seul, à 9 feuilles et concernant les lymphocytes, semble certain à ce jour.

Pour le programme d'amélioration des arbres de duplication en utilisant les réarrangements, c'est donc la méthode actuelle qui devra être gardée, en ajoutant éventuellement les conditions simples de validité du SPR pour gagner du temps dans la vérification du caractère de duplication de l'arbre obtenu, dans 20% des cas.

Références

[1] S. Ohno : *Evolution by gene duplication*, Springer Verlag, New York, 1970.

[2] W.M. Fitch : *Phylogenies constrained by crossover process as illustrated by human hemoglobins in a thirteen-cycle, eleven amino-acid repeat in human apolipoprotein A-I*, Genetics 86 (1977), 623-644.

[3] G. P. Smith : *Evolution of repeated DNA sequences by unequal crossover*, Science 191 (1976), 528-535.

[4] A. J. Jeffreys and S. Harris : *Processes of gene duplication*, Nature 296 (1981), 9-10.

[5] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson : *Molecular biology of the cell, 3rd edition*, Garland Publishing Inc. (1995), New York, USA.

[6] D. Jaitly, P. Kearney, G. Lin, and B. Ma : *Methods for reconstructing the history of tandem repeats and their application to the human genome*, in Journal ou Computer and Sustem Sciences (2002), 65, 494-507.

[7] G. Benson and L. Dong : *Reconstructing the duplication history of a tandem repeat*, in Proceedings of Intelligent Systems in Molecular Biology (ISMB 1999), 44-53. AAAI.

[8] O. Gascuel, D. Bertrand, and O. Elemento : *Reconstructing the duplication history of tandem repeated sequences*, in O. Gascuel, Mathematics of Evolution and Phylogeny. Oxford University Press (2004). In press.

[9] M. Tang, M.S. Watermann, S. Yooseph : *Zinc finger gene clusters and tandem gene duplication*, Journal of Computational Biology (2002), 429-446.

[10] O. Gascuel, M. Hendy, A. Jean-Marie, and S. McLachlan : *The combinatorics of tandem duplication trees*, Systematic Biology 52 (2003), 110-118.

[11] L. Zhang, B. Ma, and L. Wang : *Efficient methods for inferring tandem duplication history*, in Proceedinfs of the 2nd International Workshop on Algorithms in Bioinformatics (2002), 97-111.

[12] D. Bertrand and O. Gascuel : *Topological Rearrangements and Local Search Methods for Tandem Duplication Trees* (2004). In press.