Reconstruction d'un génome dupliqué ancestral

Nadia El-Mabrouk
Département d'informatique et de recherche opérationnelle, Université de Montréal, CP6128 succursale Centre-Ville, Montréal, Québec H3C 3J7

Résumé

Cette présentation traite d'un phénomène d'évolution particulier : la duplication complète de génomes. Nous présentons des méthodes de parsimonie pour retrouver un génome dupliqué ancestral donnant lieu à un minimum de réarrangements génomiques. Les opérations de réarrangement considérées sont les inversions et les translocations, et les algorithmes obtenus sont exacts et en temps polynômial.

1 Introduction

La duplication complète de génome est l'un des mécanismes les plus spectaculaires qui survienne au cours de l'évolution. Son importance dépasse celle des autres mécanismes responsables de la production de gènes dupliqués. En effet, non seulement une copie de chaque gène est libre d'évoluer, mais surtout, des ensembles de milliers de gènes peuvent évoluer de concert. La duplication de génome est, de ce fait, une source d'évolution rapide. Bien que le phénomène soit rare, il est d'une importance majeure. Des traces d'un tel phénomène ont été mises en évidence chez de nombreuses espèces d'eucaryotes [11; 12; 13; 15; 20] et surtout chez les plantes [1; 5; 10; 14; 16; 17]. Bien que le fait soit controversé, le génome des vertébrés est connu pour avoir subit deux duplications complètes il y a plus de 200 millions d'années [2; 9; 13]. Wolfe et Shields [19] ont suggéré que le génome de *Saccharomyces cerevisiae* est également le produit d'une duplication complète d'un génome ancestral, survenue il y a environ 100 millions d'années.

À la suite du phénomène de duplication, le génome subit une suite de mutations interchromosomales et intrachromosomales qui modifie grandement l'ordre et l'orientation des gènes dans le génome. Au bout du compte, on ne détecte plus que des segments, plus ou moins longs, de chromosomes, apparaissant exactement deux fois dans le génome complet. Le problème est alors de retrouver le génome ancestral juste après la duplication, ainsi que le nombre et le type de mutations survenues. La méthode de parsimonie consiste à considérer le nombre minimal d'opérations de réarrangement nécessaire pour retrouver un génome dupliqué. La complexité et les méthodes de résolution de ce problème sont étroitement liées au type d'opérations de réarrangement et de données biologiques considérées. Dans [4], nous avons développé une heuristique dans le cas d'ensembles synténiques. Nous avons ensuite, dans [3], développé un algorithme polynômial de réarrangement par translocations, dans le cas où l'on connaît l'ordre et l'orientation des gènes. Cet algorithme était exact lorsque certaines conditions étaient vérifiées.

Dans ce résumé, nous montrons comment obtenir un algorithme polynômial, exact dans tous les cas, pour les translocations, et nous généralisons la méthode au cas inversions+translocations.

2 Problème

On appellera **séquence** une suite de termes (**gènes**) signés ($+$ ou $-$) appartenant à un ensemble $\mathcal{B}$. Un **génome dupliqué réarrangé** G est un ensemble de séquences $C_1, \cdots, C_{2N}$ (**chromosomes**), tel que chaque gène de $\mathcal{B}$ apparaît exactement deux fois dans le génome (i.e. une copie dans deux chromosomes différents, ou deux copies dans le même chromosome).

Le problème consiste à trouver le nombre minimal d'opérations de réarrangement nécessaire pour transformer un génome dupliqué réarrangé G en un **génome parfaitement dupliqué** (on dira simplement génome dupliqué) inconnu H, défini comme suit : H est formé de $2M$ chromosomes $K_1, \cdots, K_{2M}$ tels que, pour tout $i \in \{1, \cdots, 2M\}$, il existe un unique $j \in \{1, \cdots, 2M\} \backslash \{i\}$ tel que $K_i = K_j$.

Dans ce résumé, les génomes considérés sont multichromosomaux, et les opérations de réarrangement considérées sont les inversions et les translocations. Une **inversion** est une opération qui transforme une certaine sous-séquence $X = x_1 x_2 \cdots x_r$ d'un chromosome en son **inverse** $(-X) = -\ x_r -\ x_{r-1} \cdots -\ x_1$. Soient X_1, X_2, Y_1 et Y_2 des séquences non vides. Une **translocation réciproque** entre les deux chromosomes $X = X_1 X_2$ et $Y = Y_1 Y_2$ est une opération de la forme $X_1 X_2, Y_1 Y_2 \longrightarrow X_1 Y_2, Y_1 X_2$ ou de la forme $X_1 X_2, -Y_2 - Y_1 \longrightarrow X_1 - Y_1, -Y_2 X_2$ (voir Figure 2).

3 Méthode de Hannenhalli et Pevzner

Étant donnés deux génomes H_1 et H_2 contenant les mêmes gènes d'un ensemble $\mathcal{B}$, chacun de ces gènes apparaissant une fois et une seule dans chaque génome, le problème classique du réarrangement de génomes consiste à trouver le nombre minimal d'opérations de réarrangement nécessaire pour transformer H_1 en H_2 (ou de façon symétrique H_2 en H_1). Hannenhalli et Pevzner (HP) ont développé des algorithmes polynômiaux pour résoudre le problème du réarrangement par inversions seules [7] (dans le cas de génomes à un seul chromosome), translocations seules [6] ou inversions+translocations [8](dans le cas de génomes multichromosomaux). Dans tous les cas, les génomes H_1 et H_2 doivent avoir le même ensemble d'extrémités. À défaut, des extrémités artificielles sont rajoutées. Dans tous les cas, leurs algorithmes se basent sur un graphe de cycles, bicoloré, non-orienté, $\mathcal{G}_{12}$ défini comme suit. Pour tout gène x dans H_1, si x est de signe positif, le remplacer par la paire $x^t x^h$, et s'il est négatif, par $x^h x^t$. Les nœuds de $\mathcal{G}_{12}$ sont alors les x^t et x^h pour tous les x dans $\mathcal{B}$. Deux nœuds quelconques, autres que x^t et x^h pour le même x, adjacents dans un chromosome de H_1, sont connectés par un **arc noir**, et deux nœuds adjacents dans H_2, sont connectés par un **arc gris**. Dans le cas du réarrangement par inversion pour des génomes à un seul chromosome, le graphe est réduit à une seule ligne, i.e. tous les arcs noirs sont contenus sur une seule ligne. Dans le cas du réarrangement par translocations pour des génomes contenant chacun N chromosomes, le graphe contient N lignes. Dans le cas du réarrangement par inversions+translocations, une concaténation de tous les chromosomes de chacun des deux génomes ramène le problème à un problème de réarrangement par inversion, et le graphe est représenté sur une seule ligne. Étant donné que chaque nœud est incident à exactement un arc noir et un arc gris, il existe une décomposition unique de $\mathcal{G}_{12}$ en c_{12} cycles alternés disjoints. On appellera **taille d'un cycle** le nombre d'arcs noirs qu'il contient. Noter que $c_{21} = c_{12} = c$ est maximal lorsque $H_1 = H_2$, auquel cas chaque cycle est de taille 1 (un arc noir et un arc gris).

Une opération de réarrangement (inversion ou translocation) ρ est déterminée par deux "cassures" dans le génome, correspondant à deux arcs noirs $e = (a, b)$ and $f = (c, d)$. On dit que ρ **est déterminée par les deux arcs noirs** e et f. D'après ce qui précède, le problème est d'accroître le plus rapidement possible le nombre de cycles du graphe. Soi ρ une opération de réarrangement, et $\Delta(c)$ la différence entre le nombre de cycles avant et après avoir appliqué l'opération ρ. HP montrent que les seules valeurs possibles de $\Delta(c)$ sont 0, 1 ou -1. ρ est une opération dite **propre** si $\Delta(c) = 1$, **impropre** si $\Delta(c) = 0$ et **mauvaise** si $\Delta(c) = -1$. De façon générale, une opération déterminée par deux arcs noirs contenus dans deux cycles différents est mauvaise, et une opération agissant sur deux arcs noirs appartenant à un même cycle est soit propre soit impropre, dépendant du type de cycle et d'arcs considérés. Deux cycles, dont tous les arcs noirs sont contenus sur une seule ligne, sont dits **connectés**, s'ils contiennent deux arcs gris (un dans chaque cycle) qui se chevauchent. Une **composante** de $\mathcal{G}_{12}$ est un sous-ensemble de cycles, non réduit à un seul cycle de taille 1, construit récursivement à partir d'un cycle, en rajoutant, à chaque étape, les cycles connectés à l'un des cycles du sous-ensemble obtenu à l'étape précédente. Une composante est dite **bonne** si elle peut être transformée en un ensemble de cycles de taille 1 par une suite d'opérations propres, et elle est dite **mauvaise** si non. Le résultat de HP dans les trois cas peut être résumé par la formule suivante :

$$D(G, H) = b(\mathcal{G}_{12}) - c(\mathcal{G}_{12}) + m(\mathcal{G}_{12}) + f(\mathcal{G}_{12}) \quad \mathbf{(1)}$$

où $D(G, H)$ est le nombre minimal d'opérations de réarrangement considérées (inversions, translocations ou inversions+translocations), $b(\mathcal{G}_{12})$ est le nombre d'arcs noirs, $c(\mathcal{G}_{12})$ le nombre de cycles, $m(\mathcal{G}_{12})$ le nombre de mauvaises composantes du graphe $\mathcal{G}_{12}$, et $f(\mathcal{G}_{12})$ est un terme correcteur de valeurs 0, 1 ou 2.

Les mauvaises composantes sont les **sous-permutations minimales** dans le cas du réarrangement par translocations, les **hurdles** dans le cas des inversions, et les **knots** dans le cas des inversions+translocations. Pour plus de précisions sur les différents termes de l'équation (1) et sur les hurdles et les knots, voir [7; 8], ou une récapitulation claire dans le cas du réarrangement par inversions dans [18].

4 Préliminaires

Afin d'utiliser le graphe HP pour notre problème de duplication de génome, on introduit une distinction arbitraire entre les deux copies de chaque gène x, en étiquettant l'une x_1 et l'autre x_2. On rajoute également deux extrémités fictives $+O_{i1}$ et $+O_{i2}$ à chaque chromosome C_i. Cela permet de ne pas tenir compte de la contrainte de HP stipulant que les deux génomes doivent avoir le même ensemble d'extrémités. Dans le cas du réarrangement par translocations, cela permet de traiter toutes les translocations, y compris celles qui réduisent (par **fusion**, i.e. X_1Y_2 nulle) ou augmentent (par **fission**, i.e. X_1X_2 nulle) le nombre de chromosomes du genome, comme des translocations réciproques. Cela permet également de considérer un génome formé d'un nombre impair de chromosomes, et de rajouter un chromosome fictif, formé exclusivement de deux extrémités, pour obtenir $2N$ chromosomes. De plus, chaque x_j est remplacé par x_j^t et x_j^h. Considérons les ensembles : $O = \{O_{i1}, O_{i2}\}_{i=1,\cdots,2N}$, $V = \{x_j^s\}^{s\in\{h,t\}}_{\substack{x\in\mathcal{B}\\ j=1,2}}$, $\mathbf{V} = O \cup V$. On note $\overline{1} = 2$, $\overline{2} = 1$, $\widetilde{t} = h$, $\widetilde{h} = t$. Pour tout $u = x_j^s \in V$, son **symétrique**, noté $\overline{u}$, est $x_{\overline{j}}^s$, et son **inverse**, noté $\widetilde{u}$, est $x_j^{\tilde{s}}$. On appelle **graphe partiel associé à** G, le graphe $\mathcal{G}(V, A)$ où A est l'ensemble des arcs noirs (non-orientés) reliant les nœuds adjacents dans G, autres que les paires de nœuds inverses (voir exemple, Figure 3). Le but est de compléter ce graphe par des arcs gris correspondant à un génome dupliqué H, de telle sorte que chaque nœud soit incident à exactement un arc noir et un arc gris. On appelle **graphe complété** de $\mathcal{G}(V, A)$ le graphe $\mathcal{G}_D(V, A, D)$ obtenu en rajoutant au graphe partiel un ensemble D d'arcs gris. Afin de donner lieu à un génome dupliqué, les arcs gris doivent satisfaire certaines contraintes, explicitées dans le lemme 1.

Lemme 1 *D ne contient aucun arc de la forme $(u, \overline{u})$, pour tout $u \in V$. De plus, supposons que $(u, v) \in D$ et $v \in V$. Si $u \in V$ alors $(\overline{u}, \overline{v}) \in D$. Si non ($u \in O$), $\overline{v}$ est aussi relié, par un arc gris, à un élément de O.*

Les méthodes de résolution des deux problèmes de réarrangement considérés dans cet article (translocations et inversions+translocations), bien que présentant des aspects différents, se basent sur un même graphe de cycles, et utilisent la même méthodologie générale. La première étape de la méthode consiste à compléter le graphe partiel $\mathcal{G}(\mathbf{V}, A)$ de telle sorte à maximiser le nombre $c(\mathcal{G}_D)$ de cycles ; un tel graphe sera appelé **graphe complété maximal**. Pour cela, une décomposition de $\mathcal{G}(\mathbf{V}, A)$ en un ensemble de sous-graphes est considérée. Elle permet de calculer une borne supérieur pour $c(\mathcal{G}_D)$. Cette décomposition a été détaillée dans [3], et sera rappelée brièvement à la section 5. La section 6 présente un algorithme de construction d'un graphe complété maximal. La deuxième étape consiste à minimiser le nombre de mauvaises composantes. La section 7 présente la méthodologie générale utilisée aussi bien pour les translocations que pour les inversions+translocations. Les différences entre les deux problèmes de réarrangement sont abordées aux sections 8 et 9.

Pour une raison d'espace, les preuves seront omises dans ce résumé, mais seront détaillées dans la version longue de l'article.

5 Majoration du nombre de cycles

Définition 1 *Soit $e = (u, v) \in A$. Soit A_e le sous-ensemble de A défini récursivement par : $(u, v) \in A_e$; Si $(x, y) \in A_e$ et $x \notin O$ alors l'arc de A adjacent à $\overline{x}$ est aussi dans A_e (même chose pour y). Soit $\mathbf{V}_e$ le sous-ensemble de $\mathbf{V}$ formé des nœuds incidents aux arcs de A_e. Alors $\mathcal{G}_e(\mathbf{V}_e, A_e)$ est le* **sous-graphe naturel** *(de* **taille** *$|A_e|$) de $\mathcal{G}(\mathbf{V}, A)$ engendré par e. Noter que si $f \in A_e$, alors $A_f = A_e$.*

Soit GCE l'ensemble des graphes naturels de taille paire, et GCO l'ensemble de ceux de taille impaire de $\mathcal{G}(\mathbf{V}, A)$. Les graphes de GCO sont fusionnés de la façon suivante : fusionner deux par deux, de façon arbitraire, les graphes (de taille impaire) ne contenant pas de nœuds dans O (ensemble $\mathcal{CO}_+$) ; fusionner deux par deux, de façon arbitraire, les graphes de taille impaire restants (ensemble $\mathcal{CO}_-$). Notons $\mathcal{CE} = GCE \cup \mathcal{CO}_+$. Les graphes de $\mathcal{CE} \cup \mathcal{CO}_-$ sont appelés **graphes supernaturels**.

Dans la suite, on considère une décomposition de $\mathcal{G}(\mathbf{V}, A)$ en un ensemble $\mathcal{SS}$ de graphes supernaturels. Quelque soit cette décomposition, le nombre de graphes de $\mathcal{CE}$ est toujours le même. On note $\boldsymbol{\alpha(G)} = |\boldsymbol{\mathcal{CE}}|$ (voir exemple, Figure 4). Le théorème suivant est prouvé dans [3].

Théorème 1 *Soit $\mathcal{G}_D(\mathbf{V}, A, D)$ un graphe complété du graphe partiel $\mathcal{G}(\mathbf{V}, A)$. Soit $c(\mathcal{G}_D)$ le nombre de cycles de $\mathcal{G}_D(\mathbf{V}, A, D)$. Alors, $c(\mathcal{G}_D) \leq \frac{|A|}{2} + \alpha(G)$.*

6 Construction d'un graphe complété maximal

Une caractéristique importante de notre algorithme est que tout arc gris construit relie toujours deux nœuds d'un même graphe supernaturel. À chaque étape, on note F l'ensemble des fragments du génome dupliqué (à construire) H résultant des étapes précédentes. Au départ, F est formé des **fragments unitaires**, i.e. $x^t x^h$, pour tout $x \in \mathcal{B}$, plus les $2N$ éléments de O. Au cours de la construction, à chaque fois qu'un arc (gris, non-orienté) (x, y) est construit, le fragment contenant x et celui contenant y sont reliés pour n'en former qu'un seul. Un **fragment intermédiaire** est unitaire, ou contient une extrémité dans O. Un **fragment terminal** a ses deux extrémités dans O. Les seuls fragments considérés au cours de la construction sont intermédiaires ou terminaux. L'algorithme procède en choisissant deux éléments de O, et en les prolongeant de façon symétrique jusqu'à obtenir deux fragments terminaux. Dès que de tels fragments sont obtenus, l'algorithme recommence avec deux autres éléments de O. À la fin, on obtient un ensemble de $2N$ fragments terminaux, représentant les $2N$ chromosomes dupliqués du génome recherché.

Soient x, y deux nœuds non encore reliés par des arc gris, tels que x est l'extrémité dans V du fragment intermédiaire courant. Afin de pouvoir construire l'arc gris (x, y), les trois conditions suivantes doivent être satisfaites : 1. Si $x \in V$ alors $y \neq \widetilde{\overline{x}}$; 2. $y \neq \overline{x}$; 3. Si $y \in O$ et F contient au moins un fragment unitaire dans V, alors F doit contenir au moins 4 fragments unitaires dans O (incluant y).

Étant donné que l'on procède en prolongeant un fragment intermédiaire, un nœud $x \in V$ n'est considéré qu'après avoir construit un arc gris adjacent à $\widetilde{x}$. La condition 1 est donc toujours satisfaite. Pour s'assurer de toujours pouvoir considérer un couple de nœuds vérifiant les conditions précédentes, il faut faire attention, en construisant un arc gris (x, y) où y est le nœud unitaire que l'on insère dans le fragment intermédiaire, à ne pas créer une **situation impossible**, vérifiant l'une des trois propriétés suivantes. Une situation ne vérifiant aucune de ces propriétés sera dite **possible**.

Propriété I. Il ne reste plus que des fragments unitaires dans V.

Propriété II. Le graphe supernaturel contenant $\widetilde{y}$ contient deux fragments unitaires de O, et deux fragments unitaires de V. D'autre part, F contient d'autres fragments de V mais aucun autre fragment de O.

Propriété III. $z = \widetilde{y}$ est dans un chemin dont la deuxième extrémité est $\overline{z}$. D'autre part, le graphe supernaturel contenant ce chemin contient un autre chemin C_2 d'extrémités t et $\overline{t}$, et (z, t) crée une situation qui vérifie la propriété I ou II.

Un arc (x, y) est dit **impossible** s'il crée une situation impossible, et il est dit **possible** si non. Si (x, y) est possible, alors $(x, \overline{y})$, $(\overline{x}, y)$ et $(\overline{x}, \overline{y})$ le sont également. Notons $\mathcal{G}$ le graphe semi-complété obtenu à une étape donnée de l'algorithme. Soit x un nœud de $\mathbf{V}$ non encore relié par un arc gris. Alors x est l'extrémité d'un chemin alterné (un arc noir suivit d'un arc gris) C dans $\mathcal{G}$, contenu complètement dans le graphe supernaturel de x. On note x^c la deuxième extrémité de ce chemin. La Figure 1 décrit un algorithme pour la construction d'un graphe complété

$\mathcal{G}_D(\mathbf{V}, A, D)$. Dans cette description, il sera sous-entendu qu'à chaque fois qu'un arc (x, y) est crée, l'arc $(\overline{x}, \overline{y})$ est également crée. (Voir exemple Figure 5.)

Algorithm dédouble :

Tant que $\mathcal{G}$ contient des fragments unitaires faire
  Choisir un nœud x correspondant à un fragment unitaire de O ;
  Faire
    Si (x, x^c) est possible
      Construire l'arc (x, x^c)
      Poser $x = \widetilde{x^c}$;
    Si non
      Si x et $\overline{x}$ n'appartiennent pas à un même chemin de $\mathcal{G}$
        Construire l'arc $(\overline{x}, \overline{x}^c)$
        Poser $x = \widetilde{\overline{x}^c}$;
      Si non, le graphe supernaturel $\mathcal{G}'$ contenant x est dans $\mathcal{CO}_-$. Soient $\mathcal{G}'_1$ et $\mathcal{G}'_2$
      les deux graphes naturels qui constituent $\mathcal{G}'$, où $x \in \mathcal{G}'_1$.
        Si il n'existe que deux nœuds y et $\overline{y}$ de $\mathcal{G}'_2$ non encore reliés par des arcs gris
          Construire l'arc (x, y) ;
          Poser $x = \widetilde{y}$;
        Si non, il existe des nœuds $y, \overline{y}, z, \overline{z}$ de $\mathcal{G}'_2$ non encore reliés par des arcs gris
          Si (x, y) est possible
            Construire l'arc (x, y) ;
            Poser $x = \widetilde{y}$;
          Sinon
            Construire l'arc (x, z) ;
            Poser $x = \widetilde{z}$;
  Tant que $x \notin O$

FIG. 1: Algorithme de construction d'un graphe complété maximal.

Théorème 2 *L'algorithm dédouble donne lieu à un graphe complété maximal $\mathcal{G}_D(\mathbf{V}, A, D)$, contenant $c(\mathcal{G}) = \alpha(G) + \frac{|A|}{2}$ cycles.*

7 Mauvaises composantes

On veut maintenant minimiser le nombre de mauvaises composantes d'un graphe complété. Bien que les notions de mauvaises composantes soient différentes dans le cas du réarrangement par translocations seules ou inversions+translocations, elles sont toutes les deux basées sur la notion de sous-permutations. Étant donnés deux génomes H_1 et H_2 ayant le même ensemble de gènes, chaque gène apparaissant une fois et une seule dans chaque génome, une **sous-permutation** SP de H_1, H_2 (ou, de façon équivalente, du graphe des cycles $\mathcal{G}_{12}$ associé à H_1, H_2) est une sous-séquence $S = u_1 u_2, \cdots u_{p-1} u_p$ d'un chromosome X de H_1 telle qu'il existe une permutation P et une sous-séquence $T = P(S) = u_1 v_2 \cdots v_{p-1} u_p$ d'un chromosome Y de H_2, où $v_2 \neq u_2$ et $v_{p-1} \neq u_{p-1}$. Une **sous-permutation minimale** est une sous-permutation ne contenant aucune autre sous-permutation, et une **sous-permutation maximale** est une sous-permutation inclue dans aucune autre sous-permutation.

Une composante C du graphe de cycles $\mathcal{G}_{12}$ correspondant à H_1, H_2, peut être identifiée par un intervalle $I = [u_l, u_r]$ où u_l et u_r sont les deux nœuds extrémités de C, tel qu'aucun arc gris ne relie un nœud à l'intérieur de I à un nœud à l'extérieur de I. Un tel intervalle est appelé **intervalle de la composante** C. Une **composante minimale** est une composante dont l'intervalle ne contient aucune autre composante. Il existe une bijection entre l'ensemble des sous-permutations de $\mathcal{G}_{12}$ et les composantes de $\mathcal{G}_{12}$. Plus précisément, soit S une sous-permutation de $\mathcal{G}_{12}$, $\Pi = \{\pi_1, \cdots \pi_p\}$ l'ensemble des composantes de $\mathcal{G}_{12}$ contenant des nœuds de S, et pour tout i, $\mathbf{V_i}$ l'ensemble des nœuds de π_i : **(1)** S_i est une sous-permutation contenue dans S (possiblement S lui même) si et seulement si S_i correspond à un intervalle d'une composante π_i de Π : **la composante de la sous-permutation** S_i ; **(2)** S_i est une sous-permutation minimale contenue dans S si et seulement si S_i correspond à un intervalle d'une composante minimale de Π. (Voir exemple Figure 6.)

Dans le cas du réarrangement par translocations, toutes les sous-permutations minimales constituent des mauvaises composantes du graphe. Autrement dit, si $s(\mathcal{G}_{12})$ est le nombre sous-permutations minimales de $\mathcal{G}_{12}$, alors dans la formule (1) de HP, $m(\mathcal{G}_{12}) = s(\mathcal{G}_{12})$. Dans le cas du réarrangement par inversions+tranlocations, certaines sous-permutations peuvent être résolues par inversions. Les **knots** [8] représentent des "mauvaises" sous-permutations (intrachromosomales) minimales et maximales.

Revenons à notre problème de duplication de génome. On veut déterminer, à partir du génome G, le nombre minimal de mauvaises sous-permutations d'un graphe complété. Pour cela, nous avons besoin d'une notion supplémentaire.

Définition 2 *Soit $S = x_1x_2 \cdots x_{n-1}x_n$ une sous-séquence d'un chromosome de G. On dit que S est une* **sous-permutation locale** *de G si S est une* **vraie sous-permutation locale** *ou une* **sous-permutation locale potentielle***, où :* **(1)** *S est une vraie sous-permutation locale de G si $\{x_1, \cdots, x_n\} \cap O = \emptyset$, et il existe une autre sous-séquence d'un chromosome de G de la forme $\overline{S} = \overline{x_1}P(\overline{x_2}, \cdots, \overline{x_{p-1}})\overline{x_p}$, où P est une permutation autre que l'identité ;* **(2)** *S est une sous-permutation locale potentielle si l'une des deux conditions suivantes est vérifiée :* **i.** *$\{x_1, x_n\} \subset O$, et il existe un chromosome de G contenant une sous-séquence $\overline{S} = O_1P(\overline{x_{i+1}}, \cdots, \overline{x_{j-1}})O_2$, où P est une permutation différente de l'identité et $\{O_1, O_2\} \in O$;* **ii.** *$x_1 \in O$, et il existe un chromosome Y contenant une sous-séquence $\overline{S} = O'P(\overline{x_2}, \cdots, \overline{x_{n-1}})x_n$, où P est une permutation différente de l'identité et $O' \in O$. Condition symétrique pour x_n.*

$\overline{S}$ est appelée **la séquence complémentaire** *de S. On dit qu'une sous-permutation locale (vraie ou potentielle) S est* **minimale** *si elle ne contient pas une autre sous-permutation locale.*

On montre maintenant comment corriger l'algorithme *dédouble* de telle sorte qu'appliqué à un génome ne contenant pas de sous-permutation locale, il produise un graphe complété maximal ne contenant aucune sous-permutation. Soit $\mathcal{G}_D(\mathbf{V}, A, D)$ un graphe complété maximal produit par *dédouble* et $S = x_1 \cdots x_n$ une sous-permutation de $\mathcal{G}_D(\mathbf{V}, A, D)$. La procédure suivante est appliquée à la sous-permutation S.

Procedure SP($x_1 \cdots x_n$)

Supprimer tous les arcs de D adjacents aux nœuds $\{x_1, \cdots, x_n, \overline{x_1}, \cdots, \overline{x_n}\}$;

Construire les arcs (x_k, x_{k+1}) et $(\overline{x_k}, \overline{x_{k+1}})$ pour tout k, $1 \leq k < n$.

Lemme 2 *Supposons que l'algorithme dédouble produise un graphe complété* $\mathcal{G}_D(\mathbf{V}, A, D)$ *contenant une sous-permutation* $S = x_1 x_2 \cdots x_{n-1} x_n$. *Si* S *ne correspond pas à une sous-permutation locale de* G, *alors* SP($x_1 \cdots x_n$) *donne lieu à un graphe complété* $\mathcal{G}_{D'}(\mathbf{V}, A, D')$ *contenant le même nombre de cycles que* $\mathcal{G}_D(\mathbf{V}, A, D)$, *et une sous-permutation en moins.*

Soit $\mathcal{G}_D(\mathbf{V}, A, D)$ le graphe complété maximal produit par *dédouble*. Soit $\mathcal{S}$ l'ensemble des sous-permutations de $\mathcal{G}_D(\mathbf{V}, A, D)$ qui ne correspondent pas à des sous-permutations locales de G, et **destruction-SP** la procédure qui applique *SP* à chaque $S \in \mathcal{S}$.

Théorème 3 *Si* G *ne contient aucune sous-permutation locale, alors le graphe complété obtenu par l'algorithme dédouble suivit de la procédure destruction-SP, est maximal et ne contient aucune sous-permutation.*

Dans la suite de ce résumé, il sera implicite que la procédure *destruction-SP* est rajoutée à la fin de l'algorithme *dédouble*. On s'intéresse maintenant au cas où G contient des sous-permutations locales.

Lemme 3 *Supposons que* G *contienne une vraie sous-permutation locale* $S = x_1 \cdots x_n$. *Alors tout graphe complété maximal* $\mathcal{G}_D(\mathbf{V}, A, D)$ *contient nécessairement une sous-permutation formée des nœuds* $\{x_1, \cdots, x_n\}$.

Les sous-permutations potentielles posent plusieurs problèmes supplémentaires. Il s'agit de minimiser le nombre de ces sous-permutations potentielles qui se transforment en sous-permutations du graphe complété final. Les sections suivantes détaillent les modifications à apporter à la méthode générale pour les deux cas de réarrangement considérés.

8 Réarrangement par translocations

Il faut différencier entre deux sortes de sous-permutations potentielles. Une sous-permutation potentielle $S = O_1 x_1 \cdots x_n O_2$ (respec. $S = O_1 x_1 \cdots x_n$) est **paire** (PES) si les deux sous-graphes naturels (respec. le sous-graphe naturel) contenant des éléments de O et des nœuds de $\{x_1, \cdots x_n, \overline{x_1}, \cdots, \overline{x_n}\}$ sont des sous-graphes de GCE, autrement dit, des sous-graphes de taille paire. S est une **sous-permutation potentielle impaire** (POS) si non. On note $\mathcal{PES}$ l'ensemble des graphes (ou paires de graphes) correspondant aux PES.

Afin d'éviter que des POS ne se transforment en sous-permutations du graphe complété final, un certain arbitraire dans le choix des sous-graphes impaires à fusionner (pour former les graphes supernaturels), doit être supprimé. Une procédure de fusion particulière pour les graphes correspondant aux POS doit être déterminée. D'autre part, afin d'éviter de transformer des PES en sous-permutations, une procédure de fusion pour les graphes correspondant aux PES doit également être introduite. Cette fusion concerne des graphes de taille paire. La conséquence est que le graphe complété obtenu par l'algorithme *dédouble* appliqué à ce nouvel ensemble de graphes supernaturels, n'est plus maximal, ce qui introduit une difficulté supplémentaire.

Les détails liés à ces procédures de fusion et à la résolution précise des sous-permutations potentielles sont omits dans ce résumé, ainsi que le calcul exact de tous les paramètres du théorème suivant.

Théorème 4 *Le nombre minimal de translocations nécessaire pour transformer* G *en un génome dupliqué est* $T(G) = \frac{|A|}{2} - \alpha(G) + sp(G) + \psi(G) + s(G) + sr(G) + f(G)$ *où* $s(G)$ *est le* *nombre de vraies sous-permutations minimales de G ; $sp(G)$ est le nombre de graphes obtenus en fusionnant les graphes de $\mathcal{PES}$; $\psi(G)$, $sr(G)$ et $f(G)$ sont des paramètres qui ne dépendent que de G, et qui prennent les valeurs 0,1, ou 2.*

De plus, l'algorithme dédouble appliqué aux graphes supernaturels appropriés, génère un génome dupliqué H tel que $D(G,H) = T(G)$.

9 Réarrangement par inversions et translocations

HP réduisent le problème du réarrangement par inversions+translocations à un problème de réarrangement par inversions (réduction d'un génome multichromosomal à un génome unichromosomal). Dans ce cas, toutes les sous-permutations du graphe complété final ne représentent pas nécessairement des mauvaises composantes. Un **arc gris orienté** relie le nœud gauche (ou droit) d'un arc noir au nœud gauche (respec. droit) d'un autre arc noir, et un **arc gris non orienté** relie deux cotés différents de deux arcs noirs. Un **cycle non-orienté** est un cycle ne contenant aucun arc gris orienté. Une **composante non-orientée** est une composante ne contenant que des cycles non-orientés. On dit qu'une composante U **sépare** deux composantes U' et U'' si tout arc fictif qui relierait un nœud de U' à un nœud de U'' couperait un arc gris de U. Un **knot** est une composante non-orientée (intrachromosomale) qui ne sépare pas deux composantes non-orientées. HP font également la différence entre deux types de knots : les **real-knots** et les **semi-knots**.

Afin de déterminer les composantes qui forment des knots, on a besoin d'une définition supplémentaire. Soit $\mathcal{U} = \{u_1, \cdots, u_p\}$ un sous-ensemble de $\mathcal{B}$, et $\overline{\mathcal{U}} = \{\overline{u_1}, \cdots, \overline{u_p}\}$. On dit que $\mathcal{U}$ est **non-orienté** si, soit les gènes u_i et $\overline{u_i}$ ont le même signe pour tout i dans G, soit u_i et $\overline{u_i}$ ont des signes différents pour tout i dans G.

Soit $\pi_i \in \Pi$ une composante et $\mathbf{V_i}$ son ensemble de nœuds. On dit que $\mathbf{V_i}$ est orienté si et seulement si l'ensemble des gènes décrit par $\mathbf{V_i}$ est orienté.

Lemme 4 *π_i est orienté si et seulement si $\mathbf{V_i}$ est orienté.*

On dit qu'une **vraie sous-permutation locale est non-orientée** si l'ensemble des gènes de sa composante est non-orienté. Les real-knots produits par *dédouble* correspondent aux vraies sous-permutations locales non-orientées minimales, et à au plus une autre sous-permutation, que nous ne détaillerons pas ici. Toutes ces sous-permutations seront appelées **mauvaises vraies sous-permutations**. Quant aux semi-knots, ils sont associés aux sous-permutations potentielles de G. Il s'agit de minimiser le nombre de ces sous-permutations potentielles qui se transforment en mauvaises sous-permutations du graphe complété final. Pour cela, une procédure de fusion particulière des graphes naturels correspondant aux POS et aux PES est également introduite, mais n'est pas détaillée dans ce résumé.

Théorème 5 *Le nombre minimal d'inversions+translocations nécessaire pour transformer G en un génome dupliqué est $IT(G) = \frac{1}{2}|A| - \alpha(G) + brs(G) + \left\lceil \frac{s(G)-gr(G)+fr(G)}{2} \right\rceil$ où $brs(G)$ est le nombre de mauvaises sous-permutations de G ; $s(G)$ est le nombre de sous-permutations locales qui se transforment en semi-knots ; $fr(G)$ et $gr(G)$ sont des paramètres qui prennent les valeurs 0 ou 1. Tous ces paramètres ne dépendent que de G.*

De plus, l'algorithme dédouble appliqué aux graphes supernaturels appropriés (procédure de fusion particulière) produit un génome H tel que $D(G,H) = IT(G)$.

10 Analyse du génome de la levure

L'analyse des gènes dupliqués du génome de *Saccharomyces cerevisiae* a permit à Wolfe et Shields [19] de suggérer que ce génome est le produit d'une duplication complète survenue il y a environ 10^8 années. Ils ont identifié 55 régions dupliquées, qui représentent environ 50% du génome complet.

L'algorithme de réarrangement par translocations appliqué aux données de Wolfe et Shields donne lieu à un graphe complété $\mathcal{G}(V, A, D)$ contenant $c = 81$ cycles. Étant donné que le génome initial ne contient aucune sous-permutation locale, l'algorithme ne donne lieu à aucune mauvaise composante. Étant donné que le génome de la levure est formé de 16 chromosomes, le nombre minimal de translocations nécessaire pour transformer ce génome en un génome parfaitement dupliqué est donc $T = 2|\mathcal{B}| + |\mathcal{O}| - 2N - c = 110 + 32 - 16 - 81 = 45$.

D'autre part, l'absence de sous-permutations locales, vraies ou potentielles, implique également que l'algorithme de réarrangement par inversions+translocations n'effectue aucune inversion. Autrement dit, 45 translocations sont également requises dans ce cas.

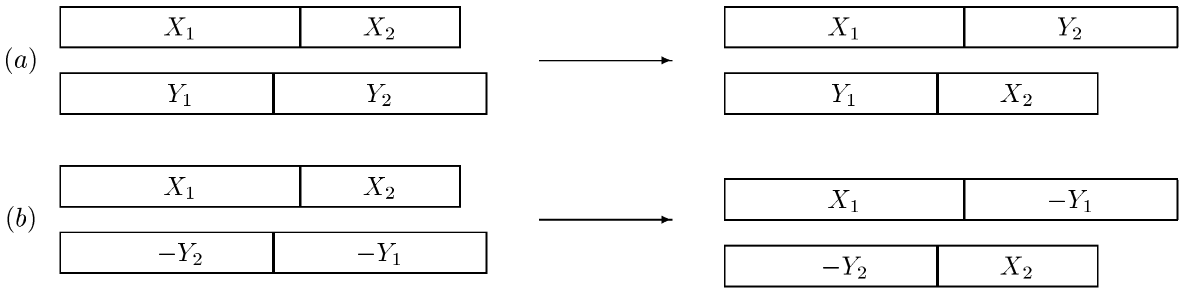

FIG. 2: Translocation réciproque entre deux chromosomes X_1X_2 et Y_1Y_2 ; (a) : tranlocation préfixe-préfixe ; (b) : translocation préfixe-suffixe.

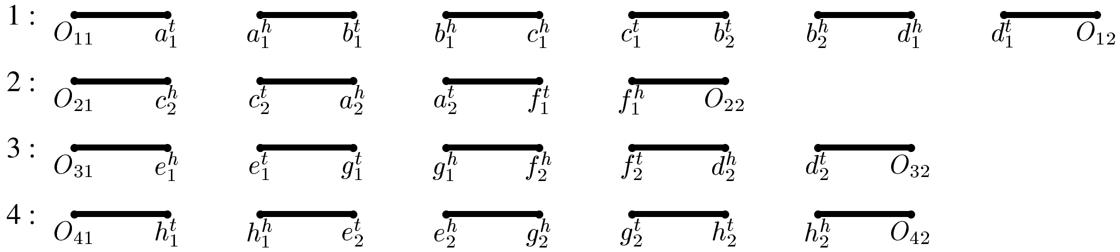

FIG. 3: Graphe partiel $\mathcal{G}(\mathbf{V}, A)$ correspondant au génome dupliqué réarrangé G formé des 4 chromosomes suivants : 1 : $+a\ +b\ -c\ +b\ -d$; 2 : $-c\ -a\ +f$; 3 : $-e\ +g\ -f\ -d$; 4 : $+h\ +e\ -g\ +h$. Pour différencier les occurences de chaque gène x, la première occurence est notée x_1 et la deuxième x_2.

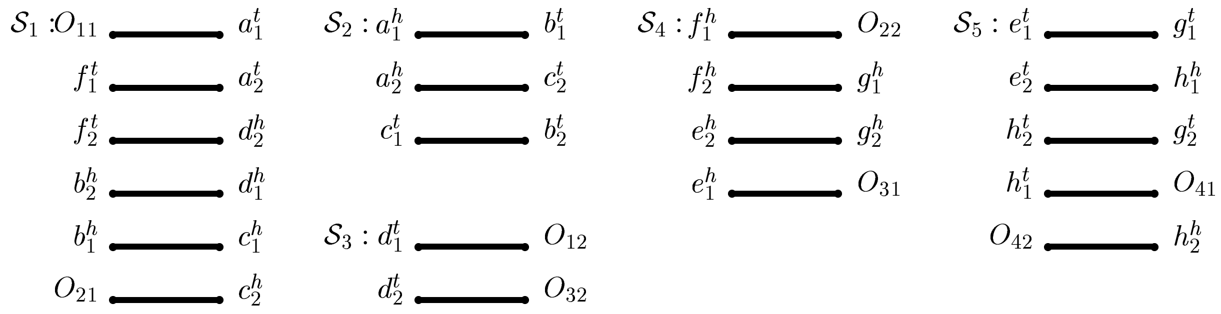

FIG. 4: Sous-graphes naturels du graphe partiel $\mathcal{G}(\mathbf{V}, A)$ de la Figure 3. Pour former des graphes supernaturels, $\mathcal{S}_2$ et $\mathcal{S}_5$ sont fusionnés. Ici, $\alpha(G) = 3$.

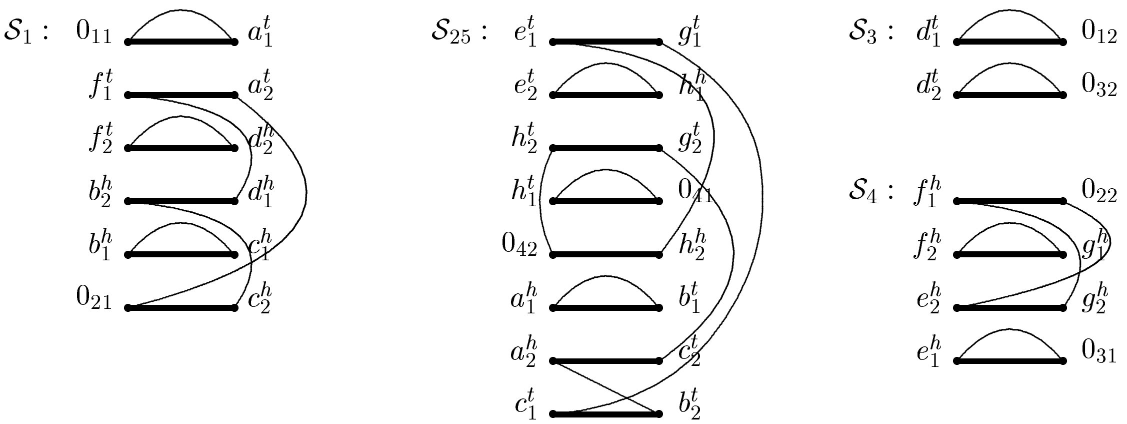

FIG. 5: Graphe complété $\mathcal{G}_D(\mathbf{V}, A, D)$ obtenu par l'algorithme *dédouble* appliqué aux graphes supernaturels de la Figure 4. Le génome dupliqué correspondant H est formé des 4 chromosomes suivants : $1 : 0_{11} + a_1 \ + b_1 \ - c_1 \ + g_1 \ - f_2 \ - d_2 \ 0_{32}$; $2 : 0_{21} + a_2 \ + b_2 \ - c_2 \ + g_2 \ - f_1 \ - d_1 \ 0_{12}$; $3 : 0_{22} \ - e_2 - h_1 \ 0_{41}$; $4 : 0_{31} \ - e_1 - h_2 \ 0_{42}$.

Références

[1] Ahn, S., Tanksley, S.D. : Comparative linkage maps of rice and maize genomes. Proc. Natl. Acad. Sci. USA **90** (1993) 7980-7984.

[2] Atkin, N. B., Ohno, S. : DNA values of four primitive chordates. Chromosoma **23** (1967) 10–13

[3] El-Mabrouk, N., Bryant, B., Sankoff, D. :Reconstructing the Pre-Doubling Genome. Proceedings of the third annual international conference on computational molecular biology (RECOMB'99). (1999) ACM press. 154-163,

[4] El-Mabrouk, N., Nadeau, J.H., Sankoff, D. : Genome halving. CPM. Ninth Annual Symposium (M. Farach-Colton, ed.). LNCS **1448** (1998), 235–250.

[5] Gaut, B.S., Doebley,J.F. : DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci., U.S.A. **94** (1997) 6809– 6814.

[6] Hannenhalli, S. : Polynomial-time algorithm for computing translocation distance between genomes. In CPM. Sixth Annual Symposium (Z. Galil and E. Ukkonen, ed.) LNCS **937** (1995), 162–176.

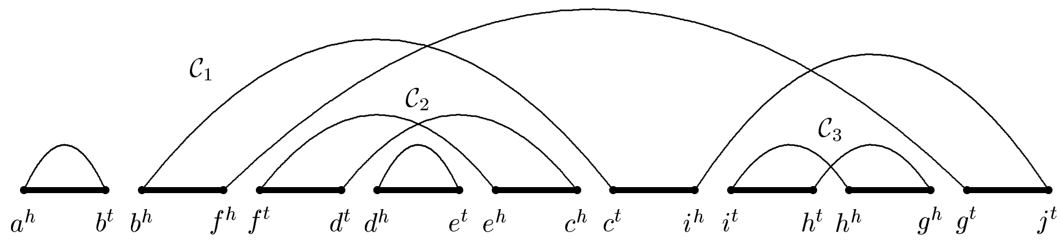

FIG. 6: Partie du graphe HP correspondant aux deux fragments F_1 et F_2 suivants des génomes H_1 et H_2 : $F_1 = +a+b-f+d+e-c-i+h-g+j$; $F_2 = +a+b+c+d+e+f+g+h+i+j$. $\mathcal{C}_1$ est la composante de la sous-permutation $S_1 = +b-f+d+e-c-i+h-g+j$; $\mathcal{C}_2$ celle de $S_2 = -f+d+e-c$; $\mathcal{C}_3$ celle de $S_3 = -i+h-g$. $\mathcal{C}_2$ et $\mathcal{C}_3$ correspondent à deux sous-permutations minimales.

[7] Hannenhalli, S., Pevzner, P.A. : Transforming cabbage into turnip. In : Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing (1995) 178–189.

[8] Hannenhalli, S., Pevzner, P.A. : Transforming men into mice. In : Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science (1995) 581–592.

[9] Hinegardner, R. : Evolution of cellular DNS content in teleost fishes. American Naturalist **102** (1968) 517–523

[10] Moore, G., Devos, K. M., Wang, Z., Gale, M. D. : 1995. Grasses, line up and form a circle. Current Biology **5** (1995) 737–739.

[11] Nadeau, J. H. : Genome duplication and comparative mapping. In Advanced Techniques in Chromosome Research (ed. Adolph, K.T.) (1991) (Marcel Dekker, New York) 269–296

[12] Nadeau, J.H., Sankoff, D. : Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics **147** (1997) 1259–1266

[13] Ohno, S., Wolfe, U., Atkin, N. B. : Evolution from fish to mammals by gene duplication. Hereditas **59** (1968) 169–187

[14] Paterson, A.H., Lan, T.-H., *et al.* : Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. Nature Genetics **14** (1996) 380–382

[15] Postlethwait, J.H, Yan, *et al.* : Vertebrate genome evolution and the zebrafish gene map. Nature Genetics 18 (1998) 345–349.

[16] Scheffler, J. A., Sharpe, *et al.* : Desaturase multigene families of Brassica napus arose through genome duplication. Theoretical and Applied Genetics **94** (1997) 583–591

[17] Shoemaker, R.C.,Polzin, *et al.* : Genome duplication in soybean (Glycine subgenus soja). Genetics **144** (1996) 329–228

[18] Setubal, J., Meidanis, J. : Introduction to Computational Molecular Biology. PWS Publishing Company (1997).

[19] Wolfe, K.H., Shields, D.C. : Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387** (1997) 708–713

[20] Xu, R-H., Kim, *et al.* : Differential regulation of neurogenesis by the two Xenopus GATA-1 genes. Molecular and Cellular Biology **17** (1997) 436–443