Reconstitution de l'ordre des gènes dans les génomes ancestraux à partir de données sur les familles multigéniques

David Sankoff et Nadia El-Mabrouk

Centre de recherches mathématiques
et Département d'informatique et de recherche opérationnelle
Université de Montréal, CP 6128 Montréal, Québec H3C 3J7

Résumé

Une méthode pour rendre compte de l'ordre des gènes dans N génomes à l'aide d'un arbre d'espèces donné $\mathcal{A}$, sans restriction sur le nombre de copies d'un gène dans un génome. Les données, sous forme d'arbres de gènes réconciliés à $\mathcal{A}$, sont soumises à une analyse qui combine le concept de chaînes de signets avec la recherche de médianes à points de cassure.

1 Introduction

Les distances de remaniements génomiques, les arbres phylogéniques, et l'évolution des familles de gènes sont trois avenues de recherche, chacune avec ses propres buts et méthodologies. Nous aborderons ici la phylogénie des espèces basée sur des données de remaniement de génomes comportant des familles multigéniques. Cette problématique nécessitera l'adaptation et l'intégration d'algorithmes provenant de ces trois domaines.

1.1 Les remaniements

La théorie des remaniements génomiques, dont les percées majeures sont les algorithmes polynômiaux de Hannenhalli and Pevzner (1995a,b); Hannenhalli (1995) pour les distances d'inversion et de translocation entre deux génomes signés, et la démonstration de Caprara (1997) de la NP-difficulté de la distance d'inversion entre génomes non signés, s'applique à la comparaison de deux ordres différents du même ensemble de gènes. Ces distances de remaniement mesurent le nombre d'opérations élémentaires nécessaires pour transformer un ordre de gènes en un autre. Les opérations modélisent les processus d'évolution génomique, tels que l'inversion d'un segment de chromosome, la transposition d'un segment entre deux sites sur le même chromosome, et la translocation (échange) de segments terminaux entre deux chromosomes.

1.2 Les signets

L'hypothèse de travail, que deux génomes différents sont formés du même ensemble de n gènes différents, sous-entend que n relations biunivoques d'homologie entre les gènes des deux génomes ont été établies au préalable, pour former n paires de gènes dorenavant considérés comme identiques. Cette hypothèse *de gènes uniques* paraît réaliste dans le cas de quelques petits génomes, comme par exemple des virus ou des mitochondries, mais n'est pas justifiée dans le cas d'espèces divergentes contenant plusieurs copies du même gène ou plusieurs gènes hautement homologues (paralogues), éparpillés le long des génomes.

Dans une publication récente (Sankoff, 1999), nous avons formulé une version généralisée du problème des remaniements génomiques, qui permet à un génome de contenir plusieurs copies de chaque gène. L'idée centrale, qui se base sur un modèle de déplacements de copies de gènes, consiste à effacer de chaque famille de gènes tout les membres sauf un – son *signet*, de façon à minimiser la distance de remaniement entre les deux génomes réduits, ainsi dérivés.

1.3 La médiane

La solution du *problème de la médiane* est d'une importance clé pour rétablir les états ancestraux dans un arbre phylogénique. Étant donné une distance d, trois génomes A, B et C, ainsi qu'un ensemble $\mathcal{E}$ de génomes, la médiane est un génome $M \in \mathcal{E}$ tel que la somme $d(A, X) + d(B, X) + d(C, X)$ est minimale pour $X = M$. Un algorithme de recherche de la médiane pourrait servir à optimiser les ancêtres dans une phylogénie donnée. Malheureusement, le problème de la médiane s'avère NP-difficile, déjà dans le cas de gènes uniques, pour toute distance de remaniement d connue, même pour la distance d'inversion signée (Caprara, 1999). Par contre, dans le cas de la distance d de points de cassure, où $d(Y, Z)$ est le nombre de paires de gènes qui sont adjacents dans le génome Y mais pas dans Z, le problème de la médiane peut se résoudre de façon relativement simple pour trois génomes A, B et C ayant le même ensemble de gènes. En effet, dans ce cas, le problème se ramène à celui du commis voyageur (Sankoff and Blanchette, 1997). Bien que, théoriquement, ce problème soit également NP-difficile, il existe des algorithmes et des logiciels applicables dans des contextes particuliers (Reinelt, 1991). Ceux-ci trouvent, en quelques minutes, la médiane de trois génomes de taille approximative $n = 100$. Récemment, nous avons développé des heuristiques pour ce problème, dans le cas où les génomes n'ont pas le même ensemble de gènes (Sankoff et al., 2000). Ces algorithmes, sans garantie d'optimalité, requièrent un temps de calcul d'une heure ou plus pour un problème de même taille.

1.4 Réconciliation

L'analyse à signets s'applique à des remaniements génomiques comportant des familles multigéniques, et non pas à des arbres de phylogénie. Inversement, le problème de la médiane est une composante importante de l'analyse phylogénique basée sur les remaniements, mais ne s'applique pas aux familles multigéniques. Un troisième domaine de recherche portant sur les familles multigéniques et la phylogénie, est la *réconciliation* d'un arbre de gènes avec un arbre d'espèces donné (Chen et al., 2000; Eulenstein et al., 1997; Guigó et al., 1996; Ma et al., 1998; Page, 1994; Page and Charleston, 1997). Étant donné un arbre de gènes, construit à partir d'un alignement des séquences de nucléotides ou d'acides aminés du même gène considéré dans plusieurs espèces, et un arbre d'espèces, la réconciliation explique la non-congruence entre ces deux arbres par des évènements de duplication qui se seraient produits pour certaines lignées dans l'arbre mais pas pour d'autres. Cette analyse, contrairement aux deux précédantes, ne tient pas compte de l'ordre des gènes dans les génomes.

1.5 L'intégration

Nous disposons de trois concepts ainsi que des méthodes algorithmiques associées, pour retrouver les génomes ancestraux d'un ensemble de N génomes actuels, à l'aide de remaniements génomiques, et en tenant compte de familles multigéniques :

- les signets, pour analyser des familles de gènes lors de la comparaison de deux génomes,
- la médiane selon la distance de points de cassure, qui permet d'optimiser l'ordre des gènes pour les génomes ancestraux, dans le contexte de gènes uniques, c.-à-d., sans familles multigéniques, et
- la réconciliation, qui évalue la parenté des gènes à l'intérieur d'une famille multigénique.

Dans cette communication, nous proposons d'intégrer ces trois idées, comme stratégie pour reconstituer l'ordre des gènes dans les génomes ancestraux, sans restriction sur les copies de gènes.

2 Les données et le problème

Notre point de départ est un arbre phylogénique (binaire et avec racine) qui récapitule les relations d'évolution entre les N espèces considérées. Notre problème porte sur les données génomiques de deuxième et troisième niveaux, c.-à-d., les séquences génomiques déjà transformées par diverses analyses :

1. La détection de gènes.
2. La détermination des relations d'homologie pour
   (a) définir les familles multigéniques, et
   (b) produire des ordres de gènes comparables entre génomes.
3. L'analyse phylogénique à l'intérieur de chaque famille pour produire des arbres de gènes.

Les analyses 1 et 2 permettent de réécrire les génomes comme des chaînes de symboles. Ces chaînes représentent l'ordre des gènes dans les génomes, où des gènes homologues dans différents génomes (orthologues) ou à l'intérieur d'un même génome (paralogues), sont representés par le même symbole. L'analyse 3 permet de déterminer des relations d'évolution entre tous les gènes, orthologues et paralogues, d'une même famille multigénique.

Dans le cadre de notre étude, le rôle de la réconciliation est d'allouer, à chaque ancêtre de l'arbre des espèces, le bon nombre de copies de chaque géne, et d'indiquer, pour chaque copie d'un gène dans un génome, sa copie parentale dans l'ancêtre immédiat. La Figure 1 illustre ces propos. Bien qu'elles soient étiquetées par le même symbole dans la graphique, chaque copie d'un gène dans un génome ancestral se distingue des autres par les positions de ses descendants dans un ou dans plusieurs des N génomes donnés.

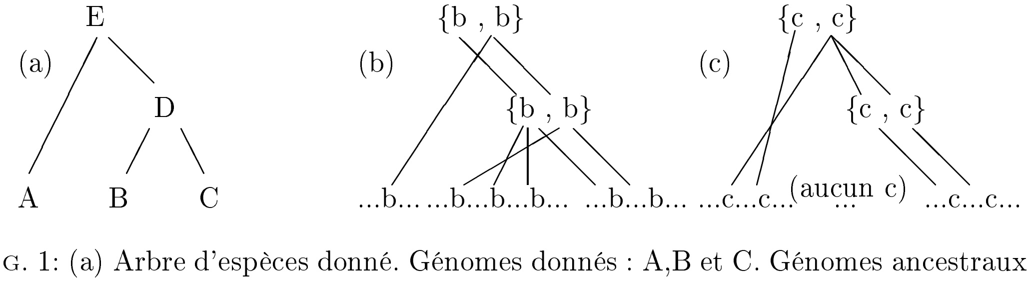

FIG. 1: (a) Arbre d'espèces donné. Génomes donnés : A,B et C. Génomes ancestraux : D et E (racine). (b) et (c) Deux familles de gènes, avec une indication sur la parenté de chaque copie. Information tirée des arbres de gènes réconciliés.

Comme nous l'avons déjà mentionné, afin de calculer la distance entre deux génomes X et Y, nous utilisons l'analyse à signets, qui ne retient qu'un membre – le signet – de chaque famille de gènes dans un génome. L'idée de cette analyse est d'identifier les gènes – l'ensemble des *vrais* signets – qui se sont déplacés le moins, chacun par rapport aux autres, pendant l'évolution divergente de X et Y. En quelque sorte, ces gènes marquent mieux, ou reflètent mieux, les positions originales des gènes dans l'ancêtre commun de X et Y, d'où l'appellation "signet".

Les mêmes concepts seront utilisés pour formaliser le problème de reconstitution de l'ordre des gènes dans les ancêtres. Supposons qu'un génome G contienne $r \geq 1$ gènes homologues (copies), pouvant être considérés comme les descendants d'un gène x dans l'ancêtre immédiat A de G. Une copie parmi les r est désignée comme étant le signet de x dans G. Si plusieurs copies de x sont présentes dans A, alors chacune pourra avoir son signet dans G, en autant qu'elle a au moins un descendant dans G.

Étant donné

- un arbre de phylogénie $\mathcal{A}$ pour N espèces ;
- leurs N génomes représentés par des chaînes de symboles appartenant à un alphabet de taille F ;
- F arbres de gènes, chacun construit à partir de toutes les occurrences du symbole considéré ;
- une mesure non-négative d de la distance entre deux ordres de gènes ne contenant que des gènes uniques,

le problème est de trouver, dans chaque génome ancestral (nœud interne) de $\mathcal{A}$,

- son ensemble de gènes ainsi que
- leurs liens de parenté par rapport aux gènes dans l'ancêtre immédiat,
- l'ordre de ces gènes dans le génome, et
- un gène dans chaque famille, désigné comme le signet,

tel que la somme des longueurs des branches de l'arbre $\mathcal{A}$ soit minimale. La longueur de la branche reliant un génome G à son ancêtre immédiat A est $d(G', A)$, où G' est le génome construit à partir de G en effaçant tout sauf les signets de chaque famille.

Remarquons que les signets dans un génome se définissent à partir du génome de l'ancêtre immédiat. Il n'y a donc pas de signets dans la racine R, sauf si on modifie notre problème pour tenir compte d'un outgroup. En fait, dans la somme des longueurs des branches de $\mathcal{A}$, les distances $d(R, S')$ et $d(R, T')$ entre R et ses deux fils S et T sont remplacées par $d(S', T')$. Par contre, les signets peuvent et doivent être déterminés dans les N génomes considérés.

3 Les outils

Avant d'aborder nos solutions au problème, nous présentons plus en détail les concepts de signet, de médiane et de réconciliation. Nous formulons les idées de signet et de médiane en toute généralité et nous donnons des précisions pour l'analyse à points de cassure.

3.1 L'analyse à signets

Considérons un alphabet de F symboles. Soient G et H deux chaînes de symboles signés ($+$ ou $-$), de longueurs respectives n_G et n_H. Pour chaque symbole a, soit $k_X(a)$ le nombre d'occurrences ($+$ ou $-$) de a dans le génome X. On peut supposer, sans perte de généralité, que $k_G(a) > 0$ et $k_H(a) > 0$, et donc $F \leq n_G$ et $F \leq n_H$. L'ensemble des occurrences du symbole a dans les deux génomes constitue une famille de gènes (famille multigénique).

À partir des deux génomes G et H, on construit deux chaînes réduites respectives g et h, en effaçant tout sauf une occurrence de chaque famille de gènes. Ce sont des chaînes de signets. Notons que h est une permutation des symboles de g.

Considérons deux chaînes de signets $g = g_1 \ldots g_F$ et $h = h_1 \ldots h_F$. On dit que g_i *précède* g_{i+1} dans g. Si le gène a précède b dans g, mais a ne précède pas b et $-b$ ne précède pas $-a$ dans h, alors a détermine un *point de cassure* dans g. De plus, il y a des points de cassure supplémentaires si $g_1 \neq h_1$ ou $g_F \neq h_F$. La distance à points de cassure entre g et h est le nombre de points de cassure dans g (qui est également le nombre de points de cassure dans h). La *distance à signets* entre G et H est le minimum, sur tout choix de chaînes de signets g et h, de la distance à points de cassure entre g et h.

Exemple 1 Soit $G =$ *-b -a b a -c d c*, $H =$ *a -a c a -c b d*. En se basant sur les chaînes de signets *-b -a -c d* et *c a b d*, la distance à signets est égale à 2.

Notre algorithme "branch-and-bound" (Sankoff, 1999) pour calculer la distance à signets, résoud une famille de gènes à la fois, et insère le nouveau signet dans les deux chaînes partielles déjà obtenues. À chaque étape, le signet choisi est la paire de gènes qui accroît le moins possible la distance entre les deux chaînes partielles. Aussitôt que la distance entre les chaînes partielles dépasse celle des meilleures chaînes de signets complètes déjà trouvées lors des étapes antérieures, ces chaînes partielles ainsi que toute autre paire de chaînes qui les inclut, sont exclues, c.-à-d. ne sont pas considérées par l'algorithme.

Quelles familles doit-on examiner d'abord ? Notre stratégie est de commencer avec les familles de taille 2, par exemple $k_G(a) = 1$ et $k_H(a) = 1$, ensuite celles de taille 3, 4 , $\cdots$. Cette approche permet de s'assurer que les chaînes partielles s'allongent le plus vite possible, afin que la borne inférieure mentionnée ci-dessus soit rapidement dépassée. L'arbre de recherche est élagué, ce qui permet d'éviter l'explosion combinatoire dûe aux familles pour lesquelles les valeurs k_G et k_H sont élevées.

3.2 La médiane de trois génomes

Étant donné

- une distance d définie sur les ordres de gènes,
- trois génomes A, B, C où chaque gène apparaît au plus une fois dans un génome, et un sous-ensemble E inclu dans l'union des gènes de A, B et C.

on cherche un génome M qui contient les gènes de E et qui minimise $d(A,X) + d(B,X) + d(C,X)$ (Sankoff et al., 1996). On appèle M la médiane.

3.2.1 Médianes à points de cassure

Soit d la distance correspondant au nombre de points de cassure. Si A, B, C et M contiennent les même gènes, il existe une réduction simple du problème de la médiane à points de cassure au problème du commis voyageur (Sankoff and Blanchette, 1997). Pour plus de simplicité, on présente la version du problème pour des génomes non-signés ; le cas signé est légèrement plus compliqué. Soit $v(x,y)$ la valeur d'une paire de gènes x, y définie par

$$v(x,y) = |\{X \in \{A,B,C\} : \{x,y\} \text{ adjacents dans } X\}|.$$

Soit également :

$$\Psi(S) = d(A, S) + d(B, S) + d(C, S),$$

Le problème de déterminer le génome S qui minimise $\Psi(S)$ équivaut à déterminer un parcours de longueur minimale par rapport à la matrice de distance δ, où $\delta_{x,y} = 3 - v(x, y)$.

Le problème du commis voyageur est NP-difficile ; le problème de la médiane à points de cassure l'est aussi (Pe'er and Shamir, 1998). Il existe, néanmoins, plusieurs approches relativement efficaces à ce problème (Reinelt, 1991) ; celles-ci peuvent aussi bien s'appliquer à notre problème, en autant que tous les génomes sont formés du même ensemble de gènes.

Pour comparer deux génomes dont les ensembles de gènes diffèrent, nous nous servons du concept de *point de cassure induit* (cf Sankoff and Blanchette (1997)). Tout d'abord, on efface tout gène présent dans un seul des deux génomes. Ensuite on détecte les points de cassure dans les génomes réduits (de compositions identiques). Le calcul des points de cassure induits est une façon plus subtile d'évaluer la correspondance entre deux ordres de gènes que les points de cassure ordinaires. En effet, si un point de cassure existe entre a_i et a_j dans le génome réduit et que $j \neq i + 1$, ce point de cassure ne peut pas être détecté dans le génome originel. (Voir Figure 2.)

génome A	1 4 5 6 7 3 8
génome A, réduit	1| 4 5 7 |3 8 |
génome B, réduit	1 | 3 8| 4 5 7 |
génome B	2 1 3 10 8 9 4 5 7

FIG. 2: Points de cassure induits (lignes verticales) dans deux génomes dont les ensembles de gènes diffèrent.

Lors d'une analyse phylogénique, le calcul du nombre de points de cassure induits peut mener à des biais si les génomes diffèrent beaucoup entre eux (par rapport au nombre total de gènes qu'ils contiennent). Pour éliminer ce problème, nous normalisons le nombre de points de cassure par le nombre de gènes en commun dans les deux génomes (Sankoff et al., 2000).

3.2.2 Une heuristique pour le problème de la médiane

La réduction du problème de la médiane à points de cassure (normalisé ou non) au problème du commis voyageur n'est plus valable dans le cas de génomes contenant des gènes différents. Les nombreuses approches au problème du commis voyageur ne sont plus applicables. Dans ce cas on se sert d'un algorithme heuristique, coûteux en termes de temps de calcul, mais qui s'avère assez précis lorsqu'on le vérifie avec des techniques de borne inférieure Sankoff et al. (2000).

L'idée est de construire la médiane un gène à la fois. On commence avec des gènes de E qui sont dans les trois génomes A, B et C. On insère chacun de ces gènes là où il augmente le moins la somme des distances de la médiane partielle aux trois autres génomes. Ensuite on répète ce processus avec les gènes qui ne sont présents que dans deux des trois génomes. Bien entendu, on ne se préoccupe ici que des distances aux deux génomes concernés. Finalement, on ajoute les gènes de E qui sont dans seulement un des trois génomes.

À chaque étape de cet algorithme, il existe généralement plusieurs gènes qui pourraient être insérer. On améliore la solution en répétant l'algorithme plusieurs fois, et en changeant les choix à chaque fois.

3.3 Réconciliation d'un arbre de gènes avec un arbre d'espèces

L'un des progrès récents dans l'étude phylogénique d'un ensemble de N espèces est de considérer toutes les occurrences (orthologues et paralogues) d'un gène dans les N espèces, d'en produire un *arbre de gène*, et de le réconcilier avec un arbre d'espèces donné (cf références à la Section 1.4). L'hypothèse de travail est que, aussi bien l'arbre des espèces que les arbre de gène, sont des arbres corrects. La discordance potentielle entre les arbres s'explique par la duplication de gènes.

La première étape des algorithmes de réconciliation est de trouver ces duplications dans les arbres de gènes, et de les situer par rapport aux évènements de spéciation dans l'arbre des espèces. Dans le cadre de notre étude, seule cette étape des algorithmes de réconciliation nous intéresse.

Les duplications sont identifiées à l'aide d'une application injective M de l'ensemble des nœuds et des feuilles d'un arbre de gène $\mathcal{G}$, à l'ensemble des nœuds et des feuilles de l'arbre des espèces $\mathcal{A}$. De telles applications sont proposées par les auteurs cités ci-dessus. L'observation clé est la suivante :

Observation : Un nœud n de $\mathcal{G}$ correspond à une duplication si et seulement si $M(n) = M(g(n))$ ou $M(n) = M(d(n))$, où $g(n)$ (respec. $d(n)$) est le fils gauche (respec. droit) de n.

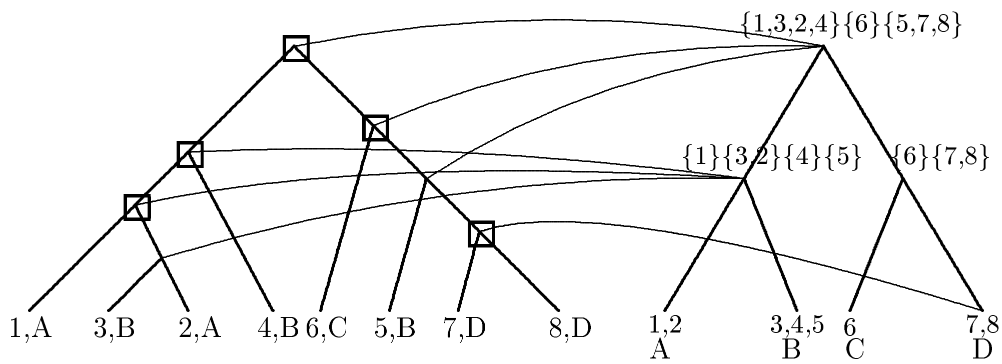

FIG. 3: Application entre l'arbre des gène $\mathcal{G}$ (à gauche), et l'arbre des espèces $\mathcal{A}$ (à droite). Les chiffres correspondent aux différentes copies du gène considéré, et les lettres majuscules correspondent aux espèces. Les nœuds de duplication dans $\mathcal{G}$ sont indiqués par un carré. Les regroupements induits pour les nœuds internes de $\mathcal{A}$ sont indiqués par des ensembles entre accolades.

Comme illustré à la Figure 3, l'application permet de retrouver les évènements de duplication dans l'arbre des gènes, et de relater chacun de ces évènements à un évènement dans l'arbre des espèces. Plus précisément, le fait qu'un nœud de duplication n de $\mathcal{G}$ soit associé à un nœud e dans $\mathcal{A}$ signifie que la duplication n a eu lieu avant l'évènement de spéciation e.

Jusqu'à présent, les méthodes de réconciliation n'ont pas servi à analyser les génomes entiers. En effet, chaque gène est traité de façon indépendante, sans se soucier de sa position dans l'ordre des gènes. Le seul but est de détecter des duplications. Notre démarche est d'utiliser la réconcilation pour retrouver beaucoup plus d'information : le nombre de copies d'un gène à chaque nœud interne de l'arbre des espèces, ainsi que la parenté de chacune de ces copies (préalable à l'analyse de l'ordre de gènes). En fait, notre utilisation de cette méthodologie est innovatrice, dans le cas où plusieurs copies d'un même gène sont présentes dans un même génome.

On représente l'ancêtre commun d'un ensemble de feuilles $a_1, a_2, \cdots, a_i$ par le *regroupement* $\{a_1, a_2, \cdots a_i\}$. Donc, un nœud interne présentant p duplications contiendra p regroupements. Pour déterminer le contenu des nœuds internes, on parcourt l'arbre en préordre. Pour chaque nœud interne e de $\mathcal{A}$:

1. on considère la restriction des regroupements du père de e aux feuilles du sous-arbre de racine e. Le regroupement initial considéré pour la racine contient toutes les feuilles de l'arbre ;
2. on subdivise les regroupements ainsi obtenus en fonction des duplications reliées au nœud e.

Un exemple de cette procédure est illustré à la Figure 3.

4 Solution

Notre méthode pour retrouver les ancêtres d'un arbre d'espèces, basée sur la solution du problème de la médiane, consiste à considérer chaque ancêtre (nœud interne) comme la médiane des trois génomes avoisinants. L'un de ces voisins sera son ancêtre immédiat, les autres ses deux descendants. Une initialisation préalable, aléatoire ou autre, de l'ordre des gènes des ancêtres, est effectuée. Ces ordres sont améliorés un par un, en se servant, à chaque étape, des versions les plus récentes des voisins. Après quelques itérations de tous les ancêtres de l'arbre, on peut s'attendre à ce que l'ensemble des ancêtres convergent vers un minimum local.

4.1 Séparation des recherches de la médiane et des signets

Le calcul des distances de points de cassure n'est pas possible pour des génomes contenant plusieurs copies d'un même gène, simplement parce que la question d'adjacence devient ambiguë. De ce fait, les algorithmes de calcul des médianes ne peuvent pas s'appliquer à de tels génomes. Considérons deux espèces G et H, où H est le père de G dans un arbre d'espèces. Supposons que l'on connaissent les regroupements des gènes dans G et H. Il est nécessaire d'associer chaque regroupement $\mathcal{R}$ de H à un regroupement unique dans G. Pour cela, il suffit de considérer que tous les regroupements de G inclus dans $\mathcal{R}$ forment une famille de gènes. Le problème est alors d'associer, de façon appropriée, chaque regroupement de H à un regroupement de G appartenant à la famille correspondante. Ceci revient à appliquer l'analyse des signets à deux génomes G et H, où H contient un seul représentant de chaque famille de gènes.

Considérons un sous-arbre de feuilles A, B, et de nœuds internes X et C, tel que représenté à la Figure 4. Avant de pouvoir appliquer l'analyse des médianes à ce sous-arbre, les génomes A, B, C et X doivent être ramenés à une forme qui ne contient qu'un membre de chaque famille. Comme mentionné plus haut, ceci est réalisé par une utilisation judicieuse de l'analyse à signets. Pour commencer, un ordre arbitraire est fixé pour les gènes des nœuds internes de l'arbre, en l'occurrence X et C. L'analyse à signets est appliquée d'abord entre X et A, et entre X et B. Ces analyses choisissent les signets dans A et B qui donnent lieu à un minimum de remaniements (en termes de points de cassure). L'analyse est aussi appliquée à C et X afin de choisir les signets dans X, dans le cas où un gène dans C a plusieurs descendants dans X. Plus précisément, pour les analyses de X avec A ou B, on considère la restriction des regroupements de X aux gènes de A et B, respectivement.

Le signet affecté à un de ces regroupements est le gène de ce regroupement qui donne lieu à un minimum de remaniements. Des ambiguïtés potentielles persistent entre C et X, et l'analyse à signets doit également être effectuée entre ces deux génomes.

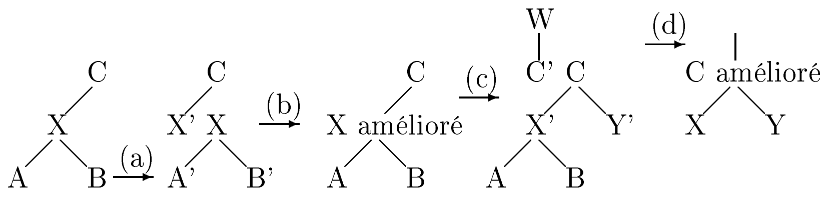

FIG. 4: Analyses alternées des signets et des médianes. (a) et (c) : recherches des signets. (b) et (d) : calcul de la médiane.

À la fin de cette analyse, aucune ambiguïté ne persiste pour le calcul de la médiane des trois génomes A, B, C. Plus précisément, des relations biunivoques sont établies entre chaque regroupement de X et chaque regroupement (ou gène) de A, B et C.

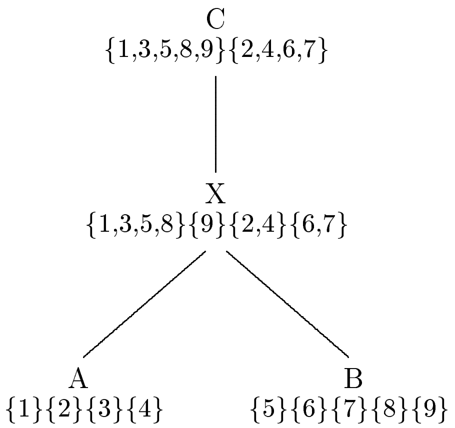

FIG. 5: Sous-arbre de feuilles A, B, et de nœuds internes X et C. Chaque regroupement, entre accolades, représente le père commun des gènes en question.

Exemple 2 Considérons les génomes A, B, X et C de la Figure 5.

- **Analyse de A et X :** Les gènes $1, 3$ sont dans un même regroupement dans X, ainsi que $2, 4$. L'analyse à signet choisit donc, dans A, un signet pour $1, 3$ et un signet pour $1, 2$. Supposons que ces signets soient 1 et 2. Alors, 3 et 4 sont effacés dans A, X et C.
- **Analyse de B et X :** Un signet est choisit pour $5, 8$, et un autre pour $6, 7$. Supposons que ces signets soient 5 et 6, alors 8 et 7 sont effacés de B, X et C.
- **Analyse de C et X :** Après les deux analyses précédentes, les regroupements de C sont $\{1, 5, 9\}, \{2, 6\}$, et les regroupements de X sont $\{1, 5\}, \{9\}, \{2\}\{6\}$. Un signet est alors choisit, dans X, pour $\{1, 5\}, \{9\}$ et pour $\{2\}, \{6\}$.

4.2 Balayage de l'arbre

À la Figure 4, on entrevoit que le choix des signets dans le génome X peut changer lorsque l'on passe de la reconstruction de l'ordre dans X à la reconstruction de l'ordre dans C. Mais un tel changement n'affecte pas les distances $d(A', X)$ et $d(B', X)$. En n'acceptant que les changements qui améliorent $d(X', C)$ on s'assure que la qualité de la solution augmente.

L'origine du signal phylogénique est dans les N génomes donnés (feuilles de l'arbre). Si on applique nos analyses, premièrement aux ancêtres qui ont deux descendants parmi les N, et ensuite aux ancêtres dont les descendants ont déjà été améliorés, alors le signal est transmis vers la racine. La vitesse de cette transmission sera d'une génération par balayage de l'arbre entier. Donc, pour un arbre équilibré, le temps requis pour propager le signal entre deux génomes, situés des deux côtés opposés de la racine, est de l'ordre de $\log N$. Ceci est une indication du nombre d'itérations du balayage nécessaires avant la convergence.

5 Discussion

On peut envisager l'utilisation de distances autres que les points de cassure dans la phylogénie basée sur les remaniements génomiques. Mais dans toute approche qui requière l'optimisation des ancêtres, il faudra régler la question du contenu (quels gènes, combien de copies) et le problème de familles multigéniques. C'est pourquoi la réconciliation (ou une méthode semblable), ainsi que les signets (ou une analyse analogue), semblent constituer des parties essentielles de l'étude de la phylogénie pour ce genre de données.

Où se trouve de telles données ? Certes, nombreux génomes sont déjà séquencés. Cependant, les analyses primaires – la détection de tous les gènes et la construction des familles multigéniques – sont loin dêtre complètes, sauf pour quelques petits génomes tels que ceux des organelles. Pour le moment, notre méthode demeure donc un outil, en attente de données suffisantes pour être appliqué.

Remerciements

Ce travail a été subventionné par le Conseil de Recherches en Sciences Naturelles et en Génie du Canada. Les auteurs sont membres du Programme en biologie de l'évolution de l'Institut Canadienne de Recherches Avancées.

Références

Caprara, A. (1997). Sorting by reversals is difficult. In *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97)*, pages 75–83, New York. ACM.

Caprara, A. (1999). Formulations and hardness of multiple sorting by reversals. In Istrail, S., Pevzner, P., and Waterman, M., editors, *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99)*, pages 84–93, New York. ACM.

Chen, K., Durand, D., and Farach-Colton, M. (2000). Notung : Dating gene duplications using gene family trees. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, New York. ACM.

Eulenstein, O., Mirkin, B., and Vingron, M. (1997). Comparison of annotating duplications, tree mapping, and copying as methods to compare gene trees with species trees. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 37 :71–93.

Guigó, R., Muchnik, I., and Smith, T. (1996). Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 6 :189–213.

Hannenhalli, S. (1995). Polynomial-time algorithm for computing translocation distance between genomes. In Galil, Z. and Ukkonen, E., editors, *Sixth Annual Symposium on Combinatorial Pattern Matching*, volume 937 of *Lecture Notes in Computer Science*, pages 162–176, Berlin. Springer.

Hannenhalli, S. and Pevzner, P. (1995a). Transforming cabbage into turnip. (polynomial algorithm for sorting signed permutations by reversals). In *Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing*, pages 178–189.

Hannenhalli, S. and Pevzner, P. (1995b). Transforming men into mice. In *Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science*, pages 581–592.

Ma, B., Li, M., and Zhang, L. (1998). On reconstructing species trees from gene trees in term of duplications and losses. In Istrail, S., Pevzner, P., and Waterman, M., editors, *Proceedings of the Second Annual International Conference on Computational Biology (RECOMB 98)*, pages 182–191, New York. ACM.

Page, R. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43 :58–77.

Page, R. and Charleston, M. (1997). Reconciled trees and incongruent gene and species trees. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 37 :57–70.

Pe'er, I. and Shamir, R. (1998). The median problems for breakpoints are np-complete. Electronic Colloquium on Computational Complexity Technical Report 98-071, http ://www.eccc.uni-trier.de/eccc.

Reinelt, G. (1991). *The traveling salesman - computational solutions for TSP applications*. Springer, New York.

Sankoff, D. (1999). Genome rearrangements with gene families. *Bioinformatics*, 15. in press.

Sankoff, D. and Blanchette, M. (1997). The median problem for breakpoints in comparative genomics. In Jiang, T. and Lee, D., editors, *Computing and Combinatorics, Proceeedings of COCOON '97*, number 1276 in Lecture Notes in Computer Science, pages 251–263, Berlin. Springer.

Sankoff, D., Bryant, D., Deneault, M., Lang, B., and Burger, G. (2000). Early eukaryote evolution based on mitochondrial gene order breakpoints. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, New York. ACM.

Sankoff, D., Sundaram, G., and Kececioglu, J. (1996). Steiner points in the space of genome rearrangements. *International Journal of the Foundations of Computer Science*, 7 :1–9.