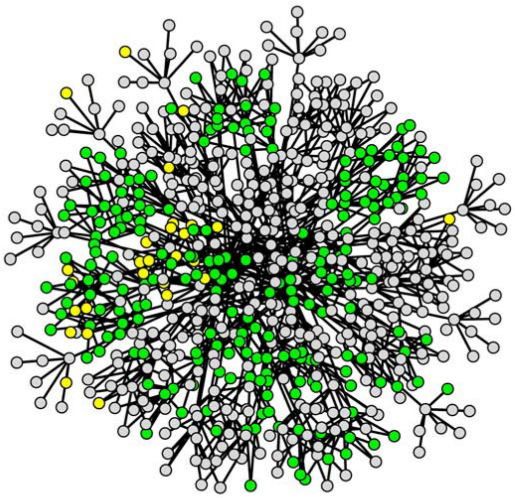
Théorie des graphes

Sujet de synthèse

Scénarios d'évolution de gènes et décomposition modulaire des graphes de permutation



Bérénice Batut

$2^{ème}$ année Master d'Informatique fondamentale

ENS DE LYON

Table des matières

1 Introduction

Afin de comprendre l'évolution du génome de groupes d'espèces, de nombreux outils mathématiques ont été développés récemment. C'est le cas de la reconstruction de scénarios d'évolution basés sur des réarrangements génomiques, que nous étudierons dans ce sujet de synthèse. Pour cette méthode, les génomes sont codés par des permutations signées. L'objectif est alors de retrouver la bonne séquence d'inversions qui transforme une permutation d'un génome en une autre. Ce sont des scénarios dits parfaits d'évolution.

La construction de scénarios parfaits est basée sur la structure combinatoire d'intervalles communs. La définition d'intervalles communs peut être reliée au concept de la décomposition modulaire de graphes de permutation [de Montgolfier, 2003].

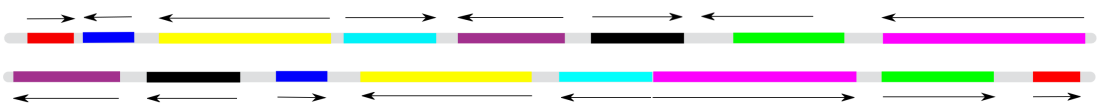
Dans ce sujet de synthèse, nous allons tout d'abord nous intéresser à la construction d'évolution de gènes comme décrite dans Berard et al. [2007]. Nous développerons ensuite le lien entre intervalles communs et décomposition modulaire. En particulier, nous essayerons de montrer que certains résultats de la première section sont des conséquences directes ou des corollaires de résultats théoriques bien connus dans la décomposition modulaire.

2 Scénarios d'évolution de gènes [Berard et al., 2007]

Dans cette première partie, nous nous intéressons à la construction de scénarios hypothétiques d'évolution de génomes.

Les génomes sont codés par des permutations signées de l'ensemble des entiers. Chaque élément représente un segment génomique stable, c'est-à-dire conservé, entre plusieurs génomes d'étude. Et, le signe de l'élément correspond à l'ordre du segment, comme le sens de lecture lors de la transcription, dans le génome étudié par rapport à un génome de référence comme dans la figure suivante :

$(1\,2\,3\,4\,5\,6\,7\,8)$



$(5\,\bar{6}\,\bar{2}\,3\,\bar{4}\,8\,\bar{7}\,1)$

A partir de deux permutations représentant les génomes à étudier, nous souhaitons retrouver le scénario d'évolution. Nous nous intéresserons plus particulièrement aux scénarios d'appariement entre deux génomes unichromosomiques qui ne brisent pas la structure combinatoire présente dans les deux permutations. Cette structure est basée sur les intervalles communs aux permutations signées, c'est-à-dire les ensembles d'éléments qui forment des intervalles dans les deux permutations, comme nous le verrons dans la suite. Ces intervalles sont alors des ensembles de segments génomiques conservés dans les deux

génomes étudiés, jusqu'au réarrangement local. La conservation de tels groupes de segments est un caractère probablement présent dans le génome de l'ancêtre commun.

Dans un premier point, nous définissons un certain nombre de notions utiles comme les permutations, les inversions et le problème du tri par inversion. Le second point est consacré à la notion d'intervalles forts et d'arbre des intervalles forts d'une permutation signée. Dans la troisième section, les scénarios parfaits sont caractérisés précisement en terme de sommets de l'arbre des intervalles forts. Deux algorithmes de calcul de scénarios parsimonieux sont présentés et illustrés par la comparaisons des chromosomes X de l'homme, de la souris et du rat.

2.1 Tri par inversion et intervalles communs

Definition 1. Une **permutation** P de n éléments est un ordre linéaire complet sur l'ensemble des entiers $\{1, \ldots, n\}$. Id_n est la permutation identité $(1, \ldots, n)$.

Une **permutation signée** de n éléments est une permutation sur l'ensemble des entiers $\{1, \ldots, n\}$ dans laquelle chaque élément a un signe, positif ou négatif.

Par soucis de notation et de lisibilité, les entiers négatifs sont représentés par une barre au-dessus d'eux.

Definition 2. Un **intervalle** d'une permutation signée est un ensemble d'éléments consécutifs de la permutation. Un intervalle peut être défini par l'ensemble de ses éléments non signés, appelé **contenu.**

Exemple 1. Soit une permutation signée de six éléments $P = (5\,\bar{6}\,\bar{2}\,3\,\bar{4}\,8\,\bar{7}\,1)$, $\{2, 3, 6\}$ est un intervalle de P.

Dans une permutation P, tout ensemble d'entiers ne correspond ainsi pas à un intervalle.

Definition 3. L'**inversion** d'un intervalle d'une permutation signée est une permutation signée d'ordre et de signes inversés des éléments de l'intervalle.

Si P est une permutation, $\bar{P}$ est la permutation obtenue en inversant complètement la permutation P.

Comme toute inversion est un intervalle de la permutation, elle est souvent traitée et représentée comme telle.

Exemple 2. Soit une permutation signée de six éléments $P = (5\bar{6}\bar{2}3\bar{4}8\bar{7}1)$, son inversion est $\bar{P} = (\bar{1}\,7\,\bar{8}\,4\,\bar{3}\,2\,6\,\bar{5})$.

Definition 4. Soient P et Q deux permutations signées de n éléments. Un **scénario** entre P et Q est une séquence d'inversions distinctes qui transforment P en Q ou P en $\bar{Q}$.

La **longueur** d'un tel scénario est le nombre d'inversions qu'il contient.

Quand Q est la permutation identité, un scénario entre P et Q est appelé simplement un **scénario** pour P.

En génomique comparative, les permutations sont utilisées pour représenter des chromosomes car l'ensemble des scénarios entre P et Q contient des séquences d'inversions qui transforment P en $\bar{Q}$. En effet, inverser un chromosome complet ne modifie pas sa structure.

Exemple 3. Inverser successivement les intervalles $\{2,3,4,5,6\}$, $\{1,7,8\}$, $\{1,5,6\}$, $\{1,2,3,4\}$, $\{1\}$, $\{2\}$, $\{4\}$, $\{6\}$, $\{8\}$ est un scénario de longueur 9 pour la permutation $P = (5\ \bar{6}\ \bar{2}\ 3\ \bar{4}\ 8\ \bar{7}\ 1)$.

Definition 5. Deux intervalles distincts, I et J, **commutent** si leur contenu se coupe trivialement, c'est-à-dire $I \subset J$, $J \subset I$ ou $I \bigcap J = \emptyset$. Si les intervalles I et J ne commutent pas, ils se **chevauchent**.

La notion suivante d'intervalle commun, introduite par Uno and Yagiura [2000], a été étudiée pour modéliser le fait qu'un groupe de gènes peut être réarrangé dans un génome tout en restant connecté [Heber and Stoye, 2001].

Definition 6. Soit P une permutation signée de n éléments. Un **intervalle commun** de P est un ensemble d'un ou plusieurs entiers, à la fois intervalle de P et intervalle de la permutation identité Id_n.

Un tel ensemble est aussi un intervalle de $\bar{P}$ et de $I\bar{d}_n$.

Les singletons et l'ensemble $\{1,2,\ldots,n\}$ sont des **intervalles communs triviaux**.

Exemple 4. Les intervalles communs de la permutation $P = (5\ \bar{6}\ \bar{2}\ 3\ \bar{4}\ 8\ \bar{7}\ 1)$ sont $\{2,3\}$, $\{3,4\}$, $\{2,3,4\}$, $\{5,6\}$, $\{2,3,4,5,6\}$, $\{7,8\}$, $\{2,3,4,5,6,7,8\}$, $\{1,2,3,4,5,6,7,8\}$ et les singletons.

Definition 7. Soit P une permutation signée. Un scénario S pour P est un **scénario parfait** si toutes les inversions de S commutent avec tous les intervalles communs de P. Un scénario parfait de longueur minimale est appelé **scénario parfait parsimonieux**.

Exemple 5. Le scénario contenant les inversions $\{2,3,4,5,6\}$, $\{1,7,8\}$, $\{1,5,6\}$, $\{1,2,3,4\}$, $\{1\}$, $\{2\}$, $\{4\}$, $\{6\}$, $\{8\}$ est un scénario parfait pour la permutation $P = (5\ \bar{6}\ \bar{2}\ 3\ \bar{4}\ 8\ \bar{7}\ 1)$.

Pour une permutation signée P donnée, il existe toujours un scénario parfait .

D'un point de vue biologique et évolutif, les scénarios qui ne brisent pas un sous-ensemble précis d'un intervalle commun sont intéressants : ils conservent ainsi les segments génomiques entre les permutations.

2.2 Arbre des intervalles forts

de Montgolfier [2003] fait remarquer la correspondance entre les intervalles communs des permutations et le concept de modules de graphes. Les résultats de cette section et de la section 2.3.3 peuvent ainsi être vus comme des conséquences directes des résultats théoriques de la décomposition modulaire. Cette correspondance sera détaillée et démontrée dans la seconde partie de ce rapport.

On peut remarquer qu'un intervalle commun d'un intervalle I ne dépend pas du signe des éléments de I. Ainsi, tous les résultats structuraux présentés dans cette section sont

valides pour des permutations signées et non signées. Ainsi, par soucis de simplification, les signes sont omis.

Soient I un intervalle commun de la permutation P de n éléments et $x \in \{1, \ldots, n\}$ tel que $x \notin I$. D'après la définition d'un intervalle commun, x est soit supérieur, soit inférieur à tous les éléments de I. Cette relation d'ordre entre x et I est notée $x > I$ ou $x < I$.

De façon similaire, pour deux intervalles communs disjoints I et J, $I < J$ signifie que tous les éléments de I sont inférieurs à ceux de J.

Definition 8. Un intervalle commun I d'une permutation P est un **intervalle fort** de P s'il commute avec tous les intervalles communs de P.

Exemple 6. Les singletons et les intervalles $\{2, 3, 4\}$, $\{5, 6\}$, $\{7, 8\}$, $\{2, 3, 4, 5, 6\}$, $\{2, 3, 4, 5, 6, 7, 8\}$, $\{1, 2, 3, 4, 5, 6, 7, 8\}$ sont les intervalles forts de $P = (5\,\bar{6}\,\bar{2}\,3\,\bar{4}\,8\,\bar{7}\,1)$.

Definition 9. L'ordre d'inclusion de l'ensemble des intervalles forts définit un arbre à n feuilles, appelé **arbre des intervalles forts** de P, noté $T_S(P)$, dont les feuilles sont les singletons et la racine est l'intervalle contenant tous les éléments d'une permutation. Un sommet de $T_S(P)$ est identifié par l'intervalle fort qu'il représente.

Exemple 7. Soit la permutation $P = (5\,6\,2\,3\,4\,8\,7\,1)$, ses intervalles forts, $\{2, 3, 4\}$, $\{5, 6\}$, $\{7, 8\}$, $\{2, 3, 4, 5, 6\}$, $\{2, 3, 4, 5, 6, 7, 8\}$, $\{1, 2, 3, 4, 5, 6, 7, 8\}$ et les singletons, permettent de construire un arbre des intervalles forts :

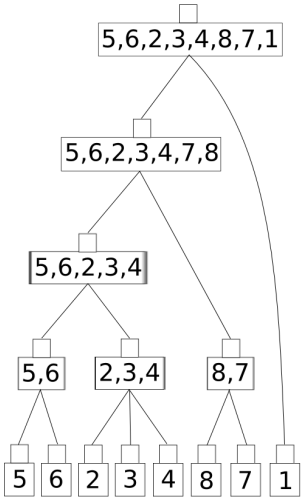


FIGURE 1 – Arbre des intervalles forts $T_S(P)$ de la permutation $P = (5\,6\,2\,3\,4\,8\,7\,1)$.

Definition 10. Soit P une permutation. Une partition $\mathcal{I} = \{I_1, \ldots, I_k\}$ des éléments de P en intervalles communs est une **partition de congruence**.

La **permutation quotient** associée à $\mathcal{I}$, notée $P_{|\mathcal{I}}$, est définie selon :

i précède j dans $P_{|\mathcal{I}}$ si et seulement si I_i précède I_j.

Exemple 8. Pour la permutation $P = (5\,6\,2\,3\,4\,8\,7\,1)$ de la figure 1, la partition $\mathcal{I} = \{\{5,6\},\{2,3,4\},\{7,8\},\{1\}\} = \{I_3, I_2, I_4, I_1\}$ est une partition de congruence de P en intervalles communs, avec $I_1 < I_2 < I_3 < I_4$. La permutation quotient associée à $\mathcal{I}$ est $P_{|\mathcal{I}} = (3,2,4,1)$.

Une partition de congruence « hérite » des intervalles communs de P :

Lemme 1. Soit $\mathcal{I} = \{I_1, \ldots, I_k\}$ une partition de congruence d'une permutation P. $J = \{j, \ldots, h\}$ est un intervalle commun de la partition quotient $P_{|\mathcal{I}}$ si et seulement si $K = \bigcup_{j \leq i \leq k} I_i$ est un intervalle commun.

Le théorème de décomposition suivant montre l'importance de la partition de congruence dont les intervalles communs sont les intervalles forts maximaux. Ce théorème peut être vu comme un corollaire d'un théorème de décomposition modulaire, il sera donc démontré comme tel dans la seconde partie du rapport.

Théorème 1. Soient P une permutation de n éléments et $\mathcal{I} = \{I_1, \ldots, I_k\}$ la partition de P en intervalles communs forts autres que P. Ainsi,

1. soit n'importe quel ensemble d'éléments consécutifs de $P_{|\mathcal{I}}$ est un intervalle commun de $P_{|\mathcal{I}}$
2. soit les intervalles communs de $P_{|\mathcal{I}}$ sont triviaux

De plus, dans le cas 1, soit $P_{|\mathcal{I}} = Id_k$ soit $P_{|\mathcal{I}} = I\bar{d}_k$. Ainsi, $P_{|\mathcal{I}}$ est dit **linéaire** s'il satisfait le cas 1 et **premier** sinon.

Le théorème 1 induit une classification des sommets d'un arbre des intervalles forts.

Proposition 1. Un intervalle I d'une permutation signée P est un intervalle commun si et seulement si il est soit un sommet de $T_S(P)$, soit l'union de fils consécutifs d'un sommet linéaire de $T_S(P)$.

Cette proposition sera démontrée dans la seconde partie du rapport.

Definition 11. Soit $P_{|\mathcal{I}}$ la permutation quotient définie par les fils d'un sommet interne I de $T_S(P)$. Le sommet I ou, de façon équivalente, l'intervalle fort I de P est

- soit **linéraire croissant**, si $P_{|\mathcal{I}}$ est la permutation identité
- soit **linéaire décroissant**, si $P_{|\mathcal{I}}$ est l'inverve de la permutation identité
- soit **premier**, sinon

Exemple 9. Dans la figure 1, les sommets sont rectangulaires : ce sont des sommets linéaires. Aucun sommet n'est premier.

2.3 Calcul de scénarios parfaits d'évolution

La base dans le calcul de scénarios parfaits parsimonieux est l'utilisation de l'arbre des intervalles forts comme un guide.

Proposition 2. Un scénario S d'une permutation P est parfait si et seulement si chaque inversion de S est soit un sommet de $T_S(P)$, soit l'union de fils d'un sommet premier de $T_S(P)$.

Le calcul d'un scénario parfait S revient à identifier les feuilles, les sommets linéaires et l'union de fils de sommets premiers de $T_S(P)$, qui sont des inversions de S.

Definition 12. Un arbre des intervalles forts est **non ambigu** si tous les sommets premier ont un parent linéaire et **ambigu** sinon. Si $T_S(P)$ n'a pas de sommets premiers, il est **défini**.

Un arbre défini est ainsi non ambigu.

Definition 13. Un **arbre signé** est un arbre des intervalles forts $T_S(P)$ dans lequel est associé un signe, $+$ ou $-$, aux sommets selon les règles suivantes

- le signe d'une feuille est le signe de l'élément correspondant dans P
- le signe d'un sommet linéaire est $+$ si le sommet est croissant et $-$ sinon
- un sommet premier hérite du signe de son père si ce dernier est linéaire.

Certains sommets peuvent ainsi ne pas avoir de signe, c'est le cas dans des arbres ambigus.

2.3.1 Calcul de scénarios parfaits pour des arbres non ambigus

Pour un arbre $T_S(P)$ non ambigu, il y a une unique façon d'affecter des signes à tous les sommets de $T_S(P)$.

Lemme 2. (Lemme de parité) Soit I un sommet de l'arbre $T_S(P)$ d'une permutation signée P. Si I a un père linéaire et un signe différent du signe de son père, alors il appartient à un scénario parfait pour P.

Le lemme précédent identifie les inversions qui doivent appartenir à tout scénario parfait et ainsi, à tout scénario parfait parsimonieux. Il s'applique à tous les arbres définis, ambigus ou non.

Théorème 2. Soit P un permutation signée. Si $T_S(P)$ est défini, l'ensemble des sommets dont le signe est différent de leur père est un scénario parfait parsimonieux pour P. De plus, aucune autre inversion n'appartient à un scénario parfait de P.

Ainsi, le calcul d'un scénario parfait parsimonieux pour P semble immédiat quand l'arbre des intervalles forts est défini.

Exemple 10. Soient 16 blocs génomiques sur le chromosome X de la souris et du rat, codant une permutation signée :

Mouse = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Rat = $\bar{4}$ $\bar{3}$ $\bar{2}$ 1 $\bar{13}$ $\bar{15}$ 14 $\bar{16}$ 8 9 10 $\bar{11}$ 12 5 6 7

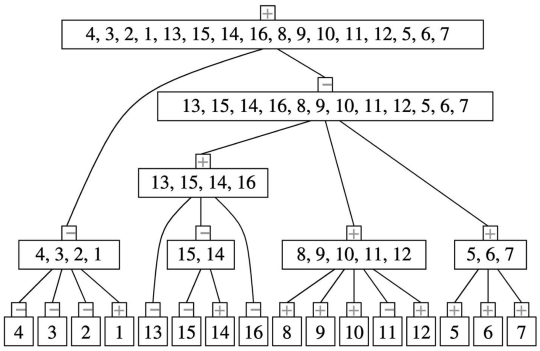


FIGURE 2 – Comparaison des chromosomes X du rat et de la souris [Berard et al., 2007].

L'arbre des intervalles forts est définie, le scénario correspondant est obtenu en comparant les signes des sommets.

L'ensemble de sommets qui ont un signe différent de celui de leur parent forme un scénario parfait parsimonieux qui transforme le chromosome X du rat en celui de la souris via 11 inversions : $(4,3,2,1)$,(1), $(13,15,14,16)$, $(13,15,14,16,8,9,10,11,12,5,6,7)$, (13), $(15,14)$, (14), (16), $(8,9,10,11,12)$, (11), $(5,6,7)$.

Dans le cas le plus général des arbres non ambigus, un sommet premier hérite son signe de son père et toute inversion, qui est une union des fils d'un sommet premier, commute avec tous les intervalles communs et ainsi, peut appartenir à un scénario parfait.

L'algorithme 1 décrit comment obtenir un scénario parfait parsimonieux dans le cas d'arbres non ambigus. L'idée de base est de calculer, pour chaque sommet premier I de l'arbre, un scénario parsimonieux qui trie les fils du sommet I dans l'ordre croissant ou décroissant, selon le signe de I. Alors, il suffit de traiter les sommets linéaire dont les parents sont linéaires selon le même principe que pour des arbres définis.

Algorithme 1 Calcul d'un scénario parfait parsimonieux pour un arbre non ambigu $T_S(P)$.

S est un scénario vide
Pour tout sommet premier I de $T_S(P)$
P_I est la permutation quotient de I sur ses fils
Si le signe de I est positif **Alors**
calculer d'un scénario parsimonieux T de P_I à Id
Sinon
calculer d'un scénario parsimonieux T deP_I à $\bar{Id}$
Déduire le scénario correspondant T' sur les fils de P_I
Ajouter l'inversion de T' à S
Fin pour
Ajouter à S les sommets linéaires et les feuilles ayant un parent linéaire et un signe différent de celui de leur père

Exemple 11. Soient 16 blocs génomiques sur le chromosome X de la souris et de l'homme.

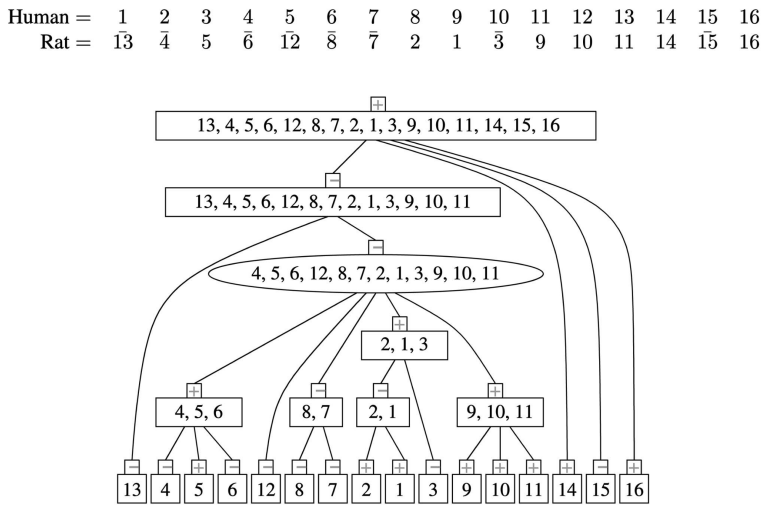


FIGURE 3 – Comparaison des chromosomes X de l'homme et de la souris [Berard et al., 2007].

L'arbre des intervalles forts est non ambigu : il possède un sommet premier $(4, 5, 6, 12, 8, 7, 2, 1, 3, 9, 10, 11)$ dont le père est un sommet linéaire décroissant. La permutation quotient de ce sommet sur ces 5 fils est $P_I = (2\ \bar{5}\ \bar{3}\ 1\ 4)$ et un scénario parsimonieux qui trie P_I en $\bar{I}d$ est donné par : $\{1, 3, 4\}, \{1, 3\}, \{1\}, \{2, 3, 4, 5\}, \{3, 4, 5\}$.

2.3.2 Calcul de scénarios parfait avec des arbres ambigus

Quand l'arbre des intervalles forts est ambigu, le signe des sommets premier n'est pas défini. On peut quand même appliquer un algorithme brute, généralisation de l'algorithme 1, déjà décrit selon un autre formalisme [Figeac and Varre, 2004].

Algorithme 2 Calcul d'un scénario parfait parsimonieux pour un arbre ambigu $T_S(P)$.

Soient $I_1, \ldots, I_k$ les sommets de $T_S(P)$ dont les signes sont indéfinis
Pour tout mot binaire W de longueur k faire
Donner à chaque sommet non signé I_j le signe + si $W[j] = 1$ et le signe − si $W[j] = 0$
Appliquer l'algorithme 1 à l'arbre signé résultant
Fin pour
Retourner un scénario parsimonieux parmi l'ensemble résultat de 2^k scénarios parfaits.

Exemple 12. Soient 16 blocs génomiques sur le chromosome X de la souris et de l'homme.

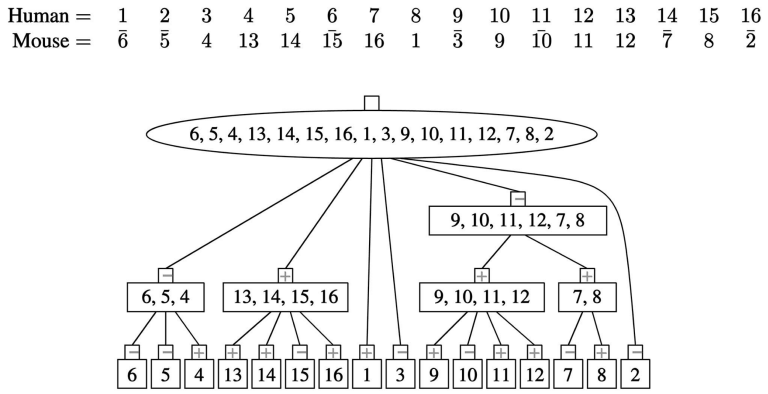


FIGURE 4 – Comparaison des chromosomes X de l'homme et de la souris [Berard et al., 2007].

L'arbre des intervalles forts est ambigu : la racine est un sommet premier. Les fils de la racine peuvent être triés en Id et $\bar{Id}$ avec 6 inversions utilisant un scénario parsimonieux qui trie la permutation quotient $P_I = (\bar{4}\,6\,1\,\bar{3}\,\bar{5}\,2)$. Un scénario parsimonieux doit aussi contenir les inversions (4), (15), (9, 10, 11, 12), (10), (7, 8), (7). La longueur totale d'un scénario est 12.

2.3.3 Calcul de scénarios parfaits pour un sous-ensemble d'intervalles communs

L'intérêt de calculer des scénarios ne cassant pas les intervalles communs repose sur l'idée que les gènes ou d'autres marqueurs génomiques s'assemblent en groupes pour des raisons fonctionnelles. Cependant, il est possible que les groupes de marqueurs génomiques apparaissent par « chance » dans les données ou ne sont supportés par aucune évidence fonctionnelle. Il ne serait alors pas utile d'imposer que des intervalles ne se cassent pas durant le scénario d'évolution. Ceci conduit à la généralisation suivante du problème précédent.

Definition 14. Soient une permutation P de longueur n et $\mathcal{C}$ l'ensemble des intervalles communs. Soit $\mathcal{F} \subseteq \mathcal{C}$ un sous-ensemble. Un scénario parsimonieux S pour P qui ne casse aucun intervalle commun de $\mathcal{F}$ est appelé un **scénario parfait parsimonieux pour** P **selon** $\mathcal{F}$. S **respecte** les intervalles de $\mathcal{F}$.

Etant donné un ensemble d'intervalles communs supposés pertinents d'un point de vue biologiques, un arbre des intervalles peut être construit, avec les même propriété que l'arbre des intervalles forts.

Definition 15. Soit $\mathcal{F}$ un ensemble d'intervalles communs d'une permutation signée P. La **clôture** $\mathcal{F}^*$ de $\mathcal{F}$ est le plus petit ensemble d'intervalles communs de P qui contient $\mathcal{F}$, tous les intervalles communs triviaux de P et tel que, pour tout $I_1 \in \mathcal{F}^*$ and $I_2 \in \mathcal{F}^*$, si I_1 et I_2 se chevauchent, alors $I_1 \cap I_2$, $I_1 \cup I_2$, $I_1 \setminus I_2$ et $I_2 \setminus I_1$ appartiennent à $\mathcal{F}^*$.

Lemme 3. Soit P une permutation signée et $\mathcal{F}$ un ensemble d'intervalles communs de P. Un scénario pour P est un scénario parfait selon $\mathcal{F}$ si et seulement si c'est un scénario parfait selon $\mathcal{F}^*$.

Definition 16. Un intervalle de $\mathcal{F}^*$ est **fort selon** $\mathcal{F}^*$ s'il ne chevauche aucun autre intervalle de $\mathcal{F}^*$.

Un arbre d'inclusion des intervalles forts de $\mathcal{F}^*$ peut ainsi être défini immédiatement, comme pour $T_S(P)$, noté $T_S^{\mathcal{F}^*}(P)$. Quand $\mathcal{F}^* = \mathcal{C}$, l'arbre correspond à l'arbre des intervalles forts.

Exemple 13. Soient 16 blocs génomiques sur le chromosome X de la souris et du rat. Les intervalles $(13, 14, 15)$, $(14, 15, 16)$, $(13, 14, 15, 16)$, $(10, 11)$, $(9, 10, 11)$, $(10, 11, 12)$, $(8, 9, 10, 11)$ et $(9, 10, 11, 12)$ sont supprimés de l'ensemble des intervalles communs.

Mouse =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Rat =	$\bar{4}$	$\bar{3}$	$\bar{2}$	1	$\bar{13}$	$\bar{15}$	14	$\bar{16}$	8	9	10	$\bar{11}$	12	5	6	7

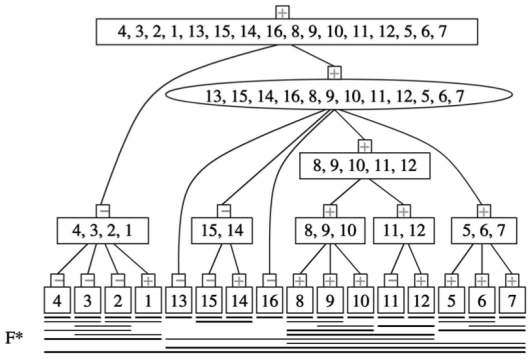


FIGURE 5 – Arbre des intervalles forts obtenu en comparant les chromosomes X du rat et de la souris avec l'ensemble des intervalles communs privé des intervalles précédents[Berard et al., 2007].

L'arbre $T_S^{\mathcal{F}^*}(P)$ a une structure similaire à l'arbre des intervalles forts de P pour les sommets linéiares et premiers, si le sous-ensemble $\mathcal{F}^*$ d'intervalles communs d'une permutation signée est strict. Le théorème suivant est une généralisation du théorème 1 :

Théorème 3. Soient P une permutation de n éléments et $\mathcal{I} = \{I_1, \ldots, I_k\}$ la partition de congruence de P en intervalles forts maximaux de $\mathcal{F}^*$ autre que P lui-même. Une des trois propositions suivantes est exacte :

- toute union d'éléments consécutifs $I = \{i, \ldots, j\}$ de $P_{|\mathcal{I}}$ est un intervalle commun de $P_{|\mathcal{I}}$ et $K = \cup_{i \leq h \leq j} I_h$ appartient à $\mathcal{F}^*$ (de plus, $P_{|\mathcal{I}} = Id_k$ ou $P_{|\mathcal{I}} = I\bar{d}_k$)
- aucune union d'intervalle de $\mathcal{I}$ n'appartient à $\mathcal{F}^*$.

Ce théorème sera démontré dans la partie suivante.

3 Décomposition modulaire des graphes de permutation

Dans la section 1.2, nous avons mentionné le lien entre les intervalles communs d'une permutation et la décomposition modullaire de graphes. Nous présenterons dans cette partie cette correspondance, en rappelant tout d'abord les principe de la décomposition modulaire. Nous définirons ensuite le concept de graphes de permutation. Enfin, nous étudierons le lien entre la décomposition modulaire de graphes de permutation et la notion d'intervalle commun. En particulier, nous montrerons que les résultats des sections 2.2 et 2.3.3 sont des conséquences des résultats théoriques de la section 3.1 lorsqu'ils sont appliqués aux graphes de permutation.

3.1 Décomposition modulaire

Definition 17. Un **module** d'un graphe $G = (V, E)$, direct, fini et sans boucle, est un sous-ensemble S de sommets tel que tout sommet $x \notin S$ est adjacent soit à tous les sommets de S, soit à aucun.

Les sommets singletons et l'ensemble de tous les sommets du graphe sont des **modules triviaux**. Un graphe dont les modules sont précisement les modules triviaux est appelé **graphe premier**.

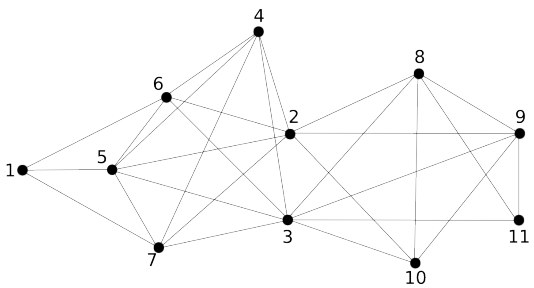
Un **graphe complet** possède des arêtes entre chaque paire de sommets. Un **graphe stable** n'a aucune arête. Un **graphe dégénéré**, au sens de la décomposition modulaire, est soit un graphe complet, soit un graphe stable : tout sous-ensemble de sommets d'un graphe dégénéré est un module.

Les composantes connexes d'un graphe G et les composantes connexes du complémentaires de G sont des modules de G.

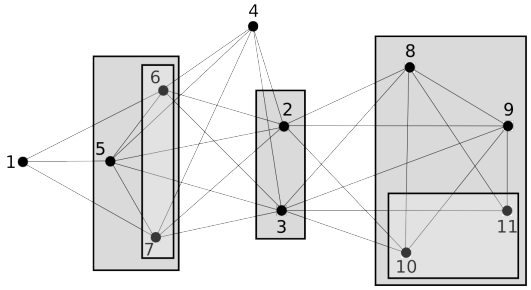
Definition 18. Un module M est **fort** s'il ne chevauche aucun autre module, c'est-à-dire tout module $M' \neq M$ satisfait $M \bigcap M' = \emptyset$, $M \subset M'$ ou $M' \subset M$.

Ainsi, les modules triviaux sont des modules forts.

Exemple 14. Soit le graphe $G = (V, E)$ suivant :

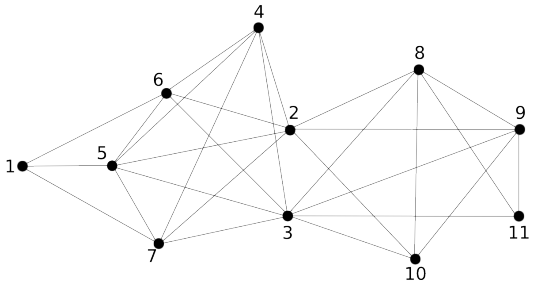


Les modules forts non triviaux sont $\{2, 3\}$, $\{6, 7\}$, $\{5, 6, 7\}$, $\{10, 11\}$ et $\{8, 9, 10, 11\}$.

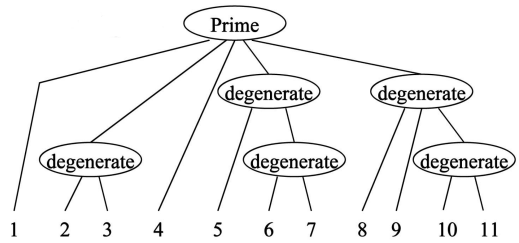


Definition 19. La famille de modules forts définit un arbre d'inclusion, appelé **arbre de décomposition modulaire**, noté $MD(G)$.

Exemple 15. Soit le graphe $G = (V, E)$ suivant :



Il fournit l'arbre de décomposition modulaire :



Lemme 4. [Mohring and Radermacher, 1984] Tout module d'un graphe G est soit un module fort soit l'union de modules forts qui sont tous fils d'un sommet dégénéré de $MD(G)$.

Ainsi, d'après le lemme précédent, la famille des modules forts est une base de l'ensemble des modules d'un graphe.

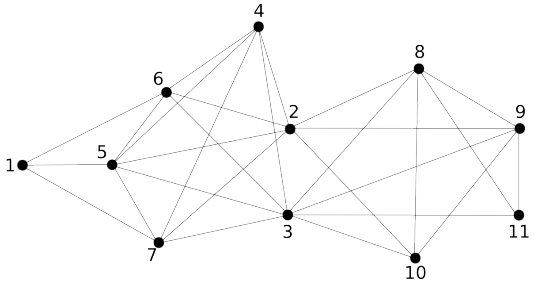
Definition 20. Soit $\mathcal{P} = \{M_1, \ldots, M_k\}$ une partition de l'ensemble des sommets du graphe G. Si, pour tout $1 \leq i \leq k$, M_i est un module de G alors $\mathcal{P}$ est une **partition de congruence**.

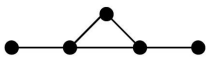
La notion de partition de congruence joue un rôle important dans la théorie de la décomposition modulaire.

Proposition 3. Si M et M' sont des modules d'une partition de congruence, alors ils sont soit adjacents (tout sommet de M est voisin de tous les sommets de M') soit non-adjacents dans G.

Definition 21. Le **graphe quotient** $G_{|\mathcal{P}}$ d'une partition de congruence $\mathcal{P}$ est le sous-graphe induit G[S] avec $S = \{x_1, \ldots, x_k\} \subseteq V$ et $\forall i \epsilon\, [1; k]$, $x_i \epsilon M_i$, c'est-à-dire $\left|V(G_{|\mathcal{P}}) \bigcap M_i\right| = 1$ pour tout $1 \leq i \leq k$.

Exemple 16. Soit le graphe $G = (V, E)$ suivant :



Le graphe représentatif de la racine est : 

Il correspond au graphe quotient $G_{|\mathcal{P}}$ avec $\mathcal{P} = \{\{1\}, \{2, 3, 4\}, \{5\}, \{6, 7\}, \{8, 9, 10, 11\}\}$.

Théorème 4. [Chein et al., 1981] Soient G un graphe et $\mathcal{P} = \{M_1, \ldots, M_k\}$ la partition de congruence contenant les modules forts non triviaux maximaux. Une des deux propositions suivantes est exactes :

1. G est non connexe ($G_{|\mathcal{P}}$ *est un graphe stable)*
2. Le complémentaire de G est non connexe ($G_{|\mathcal{P}}$ *est une clique)*
3. G et son complémentaire sont connexes ($G_{|\mathcal{P}}$ *est un graphe premier)*

Un graphe représentatif est associé à tout module fort M : le sous-graphe quotient $G[M]$ induit par la partition de congruence $\mathcal{M}$ de M en modules forts maximaux inclus dans M.

Les modules forts dont le graphe représentatif est une clique ou un stable sont étiquetés dégénérés, tandis que les autres sont premier.

3.2 Graphes de permutation

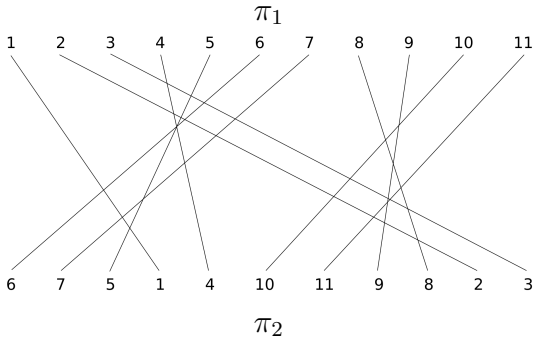
Pour établir le lien entre les modules d'un graphe et intervalles communs, nous allons d'abord définir comment une permutation définit un graphe.

Definition 22. Soit π une permutation de l'ensemble $\{1, \ldots, n\}$. Le **graphe de permutation** $G_\pi = (V, E_\pi)$ a un ensemble de sommet $V = \{1, \ldots, n\}$ et un ensemble d'arêtes $E_\pi = \{(i, j) \mid i < j \, et \, \pi(i) > \pi(j)\}$. Le graphe représenté est alors le graphe d'inversion de

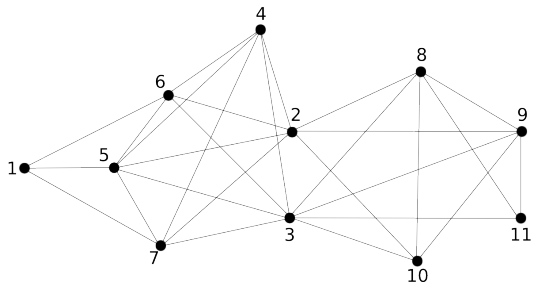
la permutation π, c'est-à-dire qu'il y a une arête entre i et j si et seulement si i et j sont inversés par π.

Le modèle d'intersection pour un graphe de permutation est appelé **diagramme de permutation** ou **réaliseur**. Il est représenté par les deux ordres π_1 et π_2.

Exemple 17. Soit la permutation $\pi = (6, 7, 5, 1, 4, 10, 11, 9, 8, 2, 3)$, le diagramme de permutation associé est :



Le graphe de permutation associé à la permutation π est :



3.3 Décomposition modulaire et graphes de permutation

Nous allons maintenant montrer que les résultats des sections 2.2 et 2.3.3 sont des conséquences de la décomposition modulaire appliquée aux graphes de permutation. En particulier, les théorèmes 1 et 3 sont des corollaires du théorème général 4. Il en est de même pour la proposition 1 vis-à-vis du lemme 4. N'ayant pas trouvé les « démonstrations » de ces résultats, nous essayons d'en donner ici une idée la plus juste possible.

En avant de montrer ces affirmations, il faut faire le lien entre les modules d'un graphe de permutation et les intervalles communs d'une permutation.

Proposition 4. [Crespelle, 2007] Les intervalles communs d'un réaliseur d'un graphe de permutation G sont des modules de G.

La réciproque est fausse mais :

Théorème 5. [de Montgolfier, 2003] Les modules forts d'un graphe de permutation $G = (V, E)$ sont exactement les intervalles forts d'un quelconque de ses réaliseurs.

De façon équivalente :

Lemme 5. [de Montgolfier, 2003, Xuan et al., 2005] Un intervalle I est un intervalle fort de π si et seulement si I est un module fort de G_π.

A partir des ces affirmations, nous pouvons montrer que les différents résultats des sections 2.2 et 2.3 sont démontrables par l'applications des résultats de décomposition de modulaire aux graphes de permutation.

Tout d'abord, le théorème 1 :

Soient P une permutation de n éléments et $\mathcal{I} = \{I_1, \ldots, I_k\}$ la partition de P en intervalles communs forts autres que P. Ainsi,

- *soit n'importe quel ensemble d'éléments consécutifs de $P_{|\mathcal{I}}$ est un intervalle commun de $P_{|\mathcal{I}}$*
- *soit les intervalles communs de $P_{|\mathcal{I}}$ sont triviaux.*

*De plus, dans le cas 1, soit $P_{|\mathcal{I}} = Id_k$ soit $P_{|\mathcal{I}} = I\bar{d}_k$. Ainsi, $P_{|\mathcal{I}}$ est dit **linéaire** s'il satisfait le cas 1 et **premier** sinon.*

Il peut être vu comme le corollaire du théorème 4 :

Soient G un graphe et $\mathcal{P} = \{M_1, \ldots, M_k\}$ la partition de congruence contenant les modules forts non triviaux maximaux. Une des deux propositions suivantes est exactes :

- *G est non connexe ($G_{|\mathcal{P}}$ est un graphe stable)*
- *Le complémentaire de G est non connexe ($G_{|\mathcal{P}}$ est une clique)*
- *G et son complémentaire sont connexes ($G_{|\mathcal{P}}$ est un graphe premier)*

Démonstration. Soient π une permutation de n éléments et $\mathcal{I} = \{I_1, \ldots, I_k\}$ la partition de π en intervalles forts maximaux non triviaux.

Soit G_π le graphe de permutation associé à π, $\mathcal{I}$ correspond, d'après le lemme 5, à la partition de G_π en modules forts maximaux non triviaux, c'est-à-dire une partition de congruence.

$\pi_{|\mathcal{I}}$ est la permutation quotient de $\mathcal{I}$: i précède j dans $\pi_{|\mathcal{I}}$ si et seulement si I_i précède I_j. $G_{\pi|\mathcal{I}}$ est le graphe quotient associé à $\mathcal{I}$, c'est-à-dire, le sous-graphe induit G_π[S] avec $S = \{x_1, \ldots, x_k\} \subseteq V$ et $\forall i \epsilon\, [1; k]$, $x_i \epsilon I_i$, I_i étant un module fort. $G_{\pi|\mathcal{I}}$ est alors le graphe de permutation associé à $\pi_{|\mathcal{I}}$.

D'après le théorème 4, une des propositions suivantes est vraie :

- G_π n'est pas connexe ($G_{\pi|\mathcal{I}}$ est stable)
- $\bar{G}_\pi$ n'est pas connexe ($G_{\pi|\mathcal{I}}$ est une clique)
- G_π et $\bar{G}_\pi$ sont connexes ($G_{\pi|\mathcal{I}}$ est premier)

Nous étudions alors les trois cas.

Si G_π n'est pas connexe, $G_{\pi|\mathcal{I}}$ est stable et ne possède donc aucune arête. La permutation associée est donc l'identité et tout ensemble d'éléments consécutifs est un intervalle commun de $\pi_{|\mathcal{I}}$.

Lorsque $\bar{G}_\pi$ n'est pas connexe, le graphe quotient est un graphe complet qui possède des arêtes entre chaque paire de sommets. Tout ensemble de sommets est alors un module et tout ensemble d'éléments consécutifs de $\pi_{|\mathcal{I}}$ est un intervalle commun. $\pi_{|\mathcal{I}}$ est alors $I\bar{d}_k$.

Dans le dernier cas, si G_π et $\bar{G}_\pi$ sont connexes, le graphe quotient $G_{\pi|\mathcal{I}}$, premier, ne possède que des modules triviaux. D'après la proposition 4, les intervalles communs de $\pi_{|\mathcal{I}}$ sont des modules $G_{\pi|\mathcal{I}}$. Comme les modules de $G_{\pi|\mathcal{I}}$ sont tous triviaux, tous les intervalles communs de $\pi_{|\mathcal{I}}$ donnant ces modules ne peuvent être alors que triviaux.

En conclusion,

- soit n'importe quel ensemble d'éléments consécutifs de $\pi_{|\mathcal{I}}$ est un intervalle commun de $\pi_{|\mathcal{I}}$
- soit les intervalles communs de $\pi_{|\mathcal{I}}$ sont triviaux.

□

De même, nous pouvons montrer, comme cas particulier du théorème 4, le théorème 3 :

Soient P une permutation de n éléments et $\mathcal{I} = \{I_1, \ldots, I_k\}$ la partition de congruence de P en intervalles forts maximaux de $\mathcal{F}^*$ autre que P lui-même. Une des trois propositions suivantes est exacte :

- *toute union d'éléments consécutifs $I = \{i, \ldots, j\}$ de $P_{|\mathcal{I}}$ est un intervalle commun de $P_{|\mathcal{I}}$ et $K = \cup_{i \leq h \leq j} I_h$ appartient à $\mathcal{F}^*$ (de plus, $P_{|\mathcal{I}} = Id_k$ ou $P_{|\mathcal{I}} = I\bar{d}_k$)*
- *aucune union d'intervalle de $\mathcal{I}$ n'appartient à $\mathcal{F}^*$.*

Démonstration. Soient π une permutation de n éléments, $\mathcal{F}$ un ensemble d'intervalles communs de π et $\mathcal{F}^*$ la clôture associée. Soit $\mathcal{I} = \{I_1, \ldots, I_k\}$ la partition de congruence de $\mathcal{F}^*$ en intervalles forts maximaux autre que P.

Soit $G_{\pi\mathcal{F}^*}$ le graphe de permutation associé à π réduit à la clôture $\mathcal{F}^*$. $\mathcal{I}$ correspond, d'après le lemme 5, à la partition de $G_{\pi\mathcal{F}^*}$ en modules forts maximaux de $\mathcal{F}^*$.

$\pi\mathcal{F}^*_{|\mathcal{I}}$ est la permutation quotient de $\mathcal{I}$. $G_{\pi\mathcal{F}^*|\mathcal{I}}$ est le graphe quotient, c'est-à-dire, le sous-graphe induit $G_{\pi\mathcal{F}^*}[\mathrm{S}]$ avec $S = \{x_1, \ldots, x_k\} \subseteq V$ et $\forall i \epsilon\, [1; k]$, $x_i \epsilon I_i$. $G_{\pi\mathcal{F}^*|\mathcal{I}}$ est le graphe de permutation associé à $\pi\mathcal{F}^*_{|\mathcal{I}}$.

D'après le théorème 4, une des propositions suivantes est vraie :

- $G_{\pi\mathcal{F}^*}$ n'est pas connexe ($G_{\pi\mathcal{F}^*|\mathcal{I}}$ est stable)
- $\bar{G}_{\pi\mathcal{F}^*}$ n'est pas connexe ($G_{\pi\mathcal{F}^*|\mathcal{I}}$ est une clique)
- $G_{\pi\mathcal{F}^*}$ et $\bar{G}_{\pi\mathcal{F}^*}$ sont connexes ($G_{\pi\mathcal{F}^*|\mathcal{I}}$ est premier)

Comme ensuite le raisonnement est identique à la démonstration précédente, nous ne le répétons pas.

En conclusion,

- soit toute union d'éléments consécutifs $I = \{i, \ldots, j\}$ de $\pi\mathcal{F}^*_{|\mathcal{I}}$ est un intervalle commun. D'après la définition de la clôture, toute union d'éléments de $\mathcal{F}^*$ appartient à $\mathcal{F}^*$. Ainsi, $K = \cup_{i \leq h \leq j} I_h$ appartient à $\mathcal{F}^*$.
- soit aucune union d'éléments consécutifs $I = \{i, \ldots, j\}$ de$\pi\mathcal{F}^*_{|\mathcal{I}}$ n'est un intervalle commun. Aucune union d'intervalles de $\mathcal{I}$ n'appartient donc à $\mathcal{F}^*$.

□

A partir de l'application à des graphes de permutation du lemme 4 :

Tout module d'un graphe G est soit un module fort soit l'union de modules forts qui sont tous fils d'un sommet dégénéré de $MD(G)$.

nous pouvons montrer la proposition 1 :

Un intervalle I d'une permutation signée P est un intervalle commun si et seulement si il est soit un sommet de $T_S(P)$, soit l'union de fils consécutifs d'un sommet linéaire de $T_S(P)$.

Démonstration. Soient π une permutation de n éléments et G_π le graphe de permutation associé. Soit $T_S(\pi)$ l'arbre des intervalles forts, où les feuilles sont les singletons de π, la racine l'intervalle contenant tous les éléments de π et les sommets les intervalles forts. Soit $MD(G_\pi)$ l'arbre de décomposition modulaire de G_π, arbre d'inclusion de la famille des modules forts.

D'après le théorème 5, les modules forts de G_π sont exactements les intervalles forts de π. Ainsi, $T_S(\pi)$ et $MD(G_\pi)$ sont identiques.

D'après le lemme 4, tout module de G_π est soit un module fort, soit l'union de modules forts qui sont fils d'un noeud dégénéré de $MD(G_\pi)$.

Si I est un module fort de G_π, il est alors intervalle fort de π (d'après le lemme 5) et donc un sommet de l'arbre des intervalles forts $T_S(\pi)$.

Si I est l'union de modules forts, tous fils d'un noeud dégénéré de $MD(G_\pi)$, I est l'union d'intervalles forts, fils consécutifs d'un sommet linéaire de $T_S(\pi)$. Comme $T_S(\pi)$ et $MD(G_\pi)$ sont identiques, une union de modules forts tous fils d'un même noeud de $MD(G_\pi)$ correspond à une union d'intervalles forts qui sont tous fils d'un noeud $T_S(\pi)$, de même nature que celui de $MD(G_\pi)$. De plus, un noeud dégénéré de $MD(G_\pi)$ est module fort stable ou complet, ce qui correspond à un intervalle fort J où n'importe quel ensemble d'élément de J est un intervalle commun de J (d'après le théorème 1), c'est-à-dire un sommet linéaire.

Ainsi, un intervalle commun est soit un sommet de $T_S(\pi)$, soit l'union d'intervalles forts fils consécutifs d'un sommet linéaire de $T_S(\pi)$. □

4 Conclusion

Dans ce sujet, nous avons décrit la structure combinatoire pour le calcul de scénarios parfaits mais aussi le lien entre cette structure et la notion de décomposition modulaire. Nous avons aussi exhibé quelques algorithmes permettant un calcul rapide de scénarios parfaits. Nous ne nous sommes cependant pas attardés dessus : ils étaients présentés comme illustration de l'utilité de la décomposition en intervalles forts pour le calcul de scénarios d'évolution de génomes.

Ce sujet de synthèse m'a permis de me plonger dans un sujet actuel de la recherche en évolution génétique et de voir quels sont les outils utilisés, en particulier en combinatoire et théorie des graphes.

Références

S. Berard, A. Bergeron, C. Chauve, and C. Paul. Perfect sorting by reversals is not always difficult. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(1) : 4–16, 2007.

M. Chein, M. Habib, and M.-C. Maurer. Partitive hypergraphs. *Discrete Math.*, 37 :1981, 1981.

C. Crespelle. *Representation dynamique de graphes.* PhD thesis, Universite Montpellier II, France, 2007.

F. de Montgolfier. *Decomposition Modulaire des Graphes. Theorie, Extensions et Algorithmes.* PhD thesis, Universite Montpellier II, France, 2003.

M. Figeac and J.-S. Varre. Sorting by reversal with common intervals. *Proc. Int'l Workshop Algorithms in Bioinformatics (WABI '04)*, pages 26–37, 2004.

S. Heber and J. Stoye. Finding all common intervals of k permutations. *Proc. 12th Int'l Symp. Combinatorial Pattern Matching (CPM '01)*, pages 207–218, 2001.

R.H. Mohring and F.J. Radermacher. Substitution decomposition for discrete structures and connections with combinatorial optimization. *Annals of Discrete Math*, 19 :257–356, 1984.

T. Uno and M. Yagiura. Fast algorithms to enumerate all common intervals of two permutations. *Algorithmica*, 26(2) :290–309, 2000.

B.M.B. Xuan, M. Habib, and C. Paul. Revisiting uno and yagiuraÕs algorithm. *Proc. IntÕl Symp. Algorithms and Computation (ISAAC Õ05)*, pages 146–155, 2005.