Algorithmique pour le consensus d'ordres

Étude bibliographique

Master *Sciences et Technologies*,
Mention *Informatique*,
Parcours Théorique

Auteur
Lisa de Mattéo

Superviseurs
Sèverine Bérard
Vincent Ranwez

Lieu de stage
ISE-M UMR5554 - CNRS, Université de Montpellier

version 1 17 mars 2017

Table des matières

1 Introduction

Un problème bien connu de la littérature est celui du *consensus d'ordres* : il s'agit, étant donné m ordres sur un ensemble $\mathcal{D}$, de proposer un *ordre consensus* qui contredise le moins possible chacun des ordres d'entrée. Ce problème possède moult champs d'application : il permet, entre autres, d'aider un groupe d'individus à prendre une décision commune basée sur les préférences des individus, d'améliorer le résultat associé à une requête, de classer un ensemble d'objets en prenant en compte simultanément plusieurs critères de similarité, etc. Ainsi, de nombreuses communautés se sont intéressées à ce problème. L'état de l'art de ce mémoire vise à recenser les différentes communautés, définir les différentes versions du problème, et proposer une classification des méthodes de résolution existantes.

Cette bibliographie est ordonnée de la façon suivante. D'abord, la section 2 situe le domaine de ce stage. Dans la section 3 est décrit l'état de l'art réalisé lors de cette étude bibliographique. Enfin, la problématique ainsi que la direction de recherche envisagée pour la poursuite du stage sont détaillées dans la section 4.

2 Présentation du domaine

L'ensemble des chromosomes présents dans les cellules d'un organisme composent son génome, c'est-à-dire son patrimoine génétique. Un chromosome contient et ordonne un ensemble de marqueurs moléculaires (*e.g.*, gènes). Avoir accès au génome des organismes est un enjeux crucial de la génomique : cela permet d'en comprendre l'organisation, les différentes transformations qu'il a connues, mais aussi de pouvoir le comparer à d'autres. Il est pour cela possible d'utiliser des techniques de séquençage, qui fragmentent le génome en de petits morceaux qu'il faut ensuite ré-assembler. Il a été montré que cette phase d'assemblage est un problème **NP**-difficile : dès lors, les heuristiques utilisées fournissent un accès imparfait au génome. Par conséquent, l'ordre entre les marqueurs fournit par l'assemblage est *partiel*, et peut même contenir des erreurs.

Une manière visuelle de représenter schématiquement ce que l'on sait de l'organisation des marqueurs moléculaires est de regrouper ensemble ceux que l'on sait être sur un même chromosome. Un partitionnement en classes d'équivalences, appelées groupes de liaison dans ce contexte biologique, est ainsi obtenu. Il est ensuite possible de représenter chacun de ces groupes de liaison par un segment, sur lequel la position (relative ou absolue) de chaque marqueur est indiquée par un tiret : il s'agit d'une *carte génomique*. Idéalement, une carte est composée d'autant de segments qu'il y a de chromosomes, et contient autant de tirets que de marqueurs. Cependant, l'information dont on dispose est généralement incomplète : des cartes parcellaires contenant plus de segments que de chromosomes (on ne dispose pas de l'information qui permettrait de les regrouper), et seulement un sous-ensemble des marqueurs, sont obtenus (pour les autres marqueurs, aucune information n'est disponible).

Une carte génomique peut être obtenue *via* différents types de données biologiques et différentes approches méthodologiques, comme par exemple des techniques de cartogra-

phie, de déséquilibre de liaison, ou encore par des logiciels de prédiction. Par conséquent, pour une espèce donnée, de nombreuses cartes peuvent avoir été obtenues par différentes équipes de recherche, en utilisant différentes approches, ainsi qu'avec différents types de marqueurs biologiques. Des travaux récents [75, 26, 27, 66] montrent qu'il est possible, en confrontant ces différents ordonnancements, d'en proposer une synthèse plus riche et plus fiable que ce qui est obtenu par une unique approche. En effet, si les cartes ne couvrent pas les mêmes marqueurs, les combiner peut permettre de couvrir une plus grande partie du génome. De plus, des informations concordantes entre plusieurs cartes se verront accorder plus de crédit. Ce stage s'intéresse à la combinaison de cartes génomiques existantes, contenant des informations partielles et/ou contradictoires, en une unique, appelée *carte consensus*. Cette dernière doit représenter l'ordre le plus plausible entre les marqueurs. Plus de détails sur la problématique du stage sont donnés dans la section 4 de ce mémoire.

3 État de l'art

3.1 Les différentes applications

Le consensus d'ordres est un problème ayant des multitudes d'applications, c'est pourquoi il a motivé de nombreux travaux, et ce dans différentes communautés. Cette section vise à donner au lecteur un aperçu de ces différents champs d'applications.

Théorie du choix social Cette théorie s'intéresse à définir les préférences collectives d'un groupe à partir des préférences individuelles de ses individus. de Borda [19] et Condorcet [15] se sont par exemple intéressés, dans un vote démocratique, à comment élire le *meilleur* candidat par rapport à l'ensemble des préférences des électeurs.

Systèmes de recommendation pour un collectif Un système de recommendation est un outil permettant de suggérer à un utilisateur un ensemble d'items (un produit, une musique, un journal d'actualité, etc.) correspondant à ses goûts et préférences. Cependant, certains types d'items qu'un système peut recommender tendent à être appréciés par des groupes d'individus plutôt que par une seule personne : c'est le cas des restaurants ou des musées par exemple, où il est plus courant de s'y rendre à plusieurs. Afin de pouvoir faire des recommendations à un groupe, un consensus des préférences des individus doit être réalisé.

De nombreux systèmes de recommendation pour un collectif ont été développés, comme Flytrap [17], qui sélectionne de la musique à jouer dans un espace public en fonction des goûts musicaux des personnes présentes, Pocket RestaurantFinder [56] qui aide un groupe de personnes à choisir un restaurant, ou encore Travel Decision Forum [42, 43] qui demande aux utilisateurs leur ressenti sur un ensemble d'attributs pour les aider à choisir une destination de vacances commune.

Priorités aux urgences Une préoccupation majeure dans les services d'ugences (SU) des hôpitaux et des cliniques est l'ordre de passage des patients. De façon à soigner les patients qui en ont le plus besoin en premier, la plupart des SU utilisent un système de *triage*[1] des patients. Il s'agit d'un processus de prise de décision dans lequel les patients sont priorisés en fonction de leur état de santé ainsi que de leur chance de survie.

Gilboy *et al.* [37] ont développé l'index de gravité d'urgence (ESI)[2] : il s'agit d'un support de prise de décision sur lequel peuvent se baser les infirmières pour catégoriser l'état de santé des patients. Le travail de Fields *et al.* [33] a permis de se rendre compte que l'affectation d'un patient à une catégorie de l'ESI dépend de l'expérience, du savoir, ainsi que de l'intuition de l'infirmière. Dès lors, en combinant les affectations données par différentes infirmières, un consensus de priorisation plus fiable [34] peut être obtenu.

Langages naturels La désambiguïsation lexicale consiste à déterminer le sens d'un mot w dans une phrase. Lorsque w possède de nombreuses significations, la tâche de désambiguïsation devient difficile (il s'agit d'un problème ouvert). Plusieurs algorithmes de désambiguïsation existent, chacun d'entre eux retournant une liste ordonnée des sens possibles de w en fonction du contexte dans lequel w apparaît. Le travail de Brody *et al.* [11] s'intéresse à combiner les résultats de différents algorithmes, de façon à améliorer la précision de la désambiguïsation.

Recherche d'informations Il s'agit, étant donné une requête q, de classer un ensemble de documents en fonction de leur *degré de pertinence* par rapport à q : plus un document est pertinent par rapport à la requête, mieux il est classé. De plus, différentes stratégies de récupération d'informations amènent à différents résultats. C'est pourquoi de nombreuses recherches récentes [35, 51, 46, 40, 8, 52] se sont concentrées sur le gain de performance qui pouvait être réalisé en combinant les résultats obtenus par différents systèmes. Cette combinaison est en fait un consensus d'ordres dans lequel il s'agit d'agréger chacun des classements retournés par différents systèmes de récupération d'informations.

Méta-recherches Étant donné une requête q posée par un utilisateur, un méta-moteur de recherche transmet cette même requête à différents moteurs de recherche (Google, DuckDuckGo, Yahoo, etc.). Chacun de ces moteurs retourne un classement des pages web en réponse à q. Le méta-moteur de recherche combine alors ces listes pour produire une liste agrégée : c'est cette dernière qui est présentée à l'utilisateur comme réponse à sa requête. Différents méta-moteurs de recherche ont été développés, comme MetaCrowler [64], SavvySearch [23], Inquirius [50], ou encore ProFusion [36].

Les différents travaux sur les méta-recherches [73, 5, 24, 57, 60, 21] sont motivés par un ensemble de déficiences des moteurs de recherche, comme la rapidité à laquelle une grande quantité d'informations doit être indexée, les pages web payant certains moteurs

[1] Le mot anglais est emprunté.
[2] De l'anglais Emergency Severity Index.

pour apparaître "plus haut" dans les recherches, ou encore les mauvaises indexations [24], volontaires ou non, de pages web.

Classification automatique Différents problèmes concernent la classification d'objets. Il s'agit, étant donné un ensemble de classes, de chercher à quelle classe l'objet en question est le plus susceptible d'appartenir. Chaque classificateur retourne la probabilité qu'a chaque classe de contenir l'objet considéré. Comme de nombreuses stratégies de classification existent, il est possible de combiner les différentes décisions prises par chacune d'entre elles dans un consensus final. Différentes études [41, 68] tentent de montrer sous quelles conditions le consensus final peut être meilleur que les classements d'entrée.

Consensus de cartes génomiques Étant donné un ensemble de marqueurs moléculaires, une carte génomique est un ordonnancement (partiel) de ces marqueurs. L'objectif est d'agréger, dans une carte consensus, un ensemble de cartes génomiques (qui peuvent avoir des types différents : génétiques, physiques, cytologiques, optiques, synténiques, basées sur les séquences, etc.) fournissant chacune un ordre partiel entre les marqueurs moléculaires. Des conflits (*i.e.*, inconsistances) peuvent apparaître dans les cartes : il faut tenter de les résoudre en supprimant un minimum d'informations contradictoires. Une telle agrégation permet de déduire un ensemble d'informations qui n'était donné par aucune des cartes d'entrée prise individuellement ; en agrégeant plusieurs cartes génomiques, une cartographie plus précise peut ainsi être obtenue. De nombreuses études [75, 76, 77, 26, 66] utilisent les outils de la théorie des graphes pour obtenir des cartes consensus de bonne qualité (voir la troisième catégorie (page 12) des méthodes de résolution de la section 3.3).

3.2 Les différentes variantes du consensus d'ordres

Afin de lister un ensemble de variantes du consensus d'ordres, le problème dans sa version la plus générale est d'abord donné.

Problème d'optimisation 1. Consensus d'ordres (version générale)

Entrée	*Un ensemble de m ordres $\mathcal{R} = \{\sigma_1, \ldots, \sigma_m\}$ sur un domaine $\mathcal{D}$.*
Sortie	*Un ordre σ^* sur $\mathcal{D}$ qui soit le plus "proche"[a] de $\mathcal{R}$.*

---

[a] Les versions plus spécifiques du problème préciseront cette notion de proximité

Ce problème existe sous différentes variantes dans la littérature. Nous avons recensé et synthétisé un ensemble de spécifications pouvant servir à leur définition, et nous le décrivons dans cette section.

Type des ordres d'entrée Tout d'abord, le type des ordres d'entrée constitue une première spécification du problème : il peut s'agir d'un ordre total, d'un ordre à *buckets*[3], d'un ordre à intervalles, ou encore d'un ordre partiel. Les définitions de ces différents types d'ordres, données ci-dessous, sont basées sur celles fournies dans [9] et illustrées dans l'exemple 1.

Une *relation binaire* R sur un ensemble fini non vide $\mathcal{D}$ (appelé *domaine*) est un sous-ensemble de $\mathcal{D} \times \mathcal{D}$; on note $(x, y) \in R$ par $x \prec_R y$.

Une relation binaire σ est un *ordre partiel (strict)*[4] sur $\mathcal{D}$ si, pour tout $x, y, z \in \mathcal{D}$, elle est :

- *antiréflexive*, *i.e.*, $x \not\prec_\sigma x$,
- *antisymétrique*, *i.e.*, $x \prec_\sigma y \Rightarrow y \not\prec_\sigma x$,
- *transitive*, *i.e.*, ($x \prec_\sigma y$ et $y \prec_\sigma z$) $\Rightarrow x \prec_\sigma z$.

Un ordre partiel κ sur $\mathcal{D}$ est un *ordre à intervalles* s'il existe une bijection I de $\mathcal{D}$ vers un ensemble d'intervalles, *i.e.*, $I(x) = [l_x, r_x]$, telle que $x \prec_\kappa y \Leftrightarrow r_x < l_y$.

Un ordre partiel π sur $\mathcal{D}$ est un *ordre à buckets* s'il est *négativement transitif*, c'est-à-dire que pour tout $x, y, z \in \mathcal{D}$, ($x \not\prec_\pi z$ et $z \not\prec_\pi y$) $\Rightarrow x \not\prec_\pi y$. Par conséquent, le domaine $\mathcal{D}$ est partitionné en une séquence de *buckets* $\mathcal{B}_1, \ldots, \mathcal{B}_t$ telle que $x \prec_\pi y$ s'il existe i et j avec $i < j$ et $x \in \mathcal{B}_i$ et $y \in \mathcal{B}_j$. Ainsi, dans un tel ordre, deux éléments sont incomparables si, et seulement si, ils sont dans le même *bucket*.

Un ordre partiel τ sur $\mathcal{D}$ est un *ordre total*[5] s'il est *complet*, *i.e.*, pour tout $x, y \in \mathcal{D}$ avec $x \neq y$, $x \prec_\tau y$ ou $y \prec_\tau x$. Autrement dit, tous les éléments de τ sont comparables ; τ représente une permutation des éléments de $\mathcal{D}$.

Finalement, tous les ordres considérés dans ce mémoire sont strict, *i.e.*, désignent une relation antiréflexive.

Exemple 1. *Sur le domaine* $\mathcal{D} = \{A, B, C, D\}$, *l'ensemble* $\{A \prec_\tau B \prec_\tau C \prec_\tau D\}$ *représente un ordre total,* $\{\{A, B\} \prec_\pi \{C, D\}\}$ *un ordre à buckets (mais pas un ordre total),* $\{A \prec_\kappa C,\ A \prec_\kappa D,\ B \prec_\kappa D\}$ *un ordre à intervalles (mais pas un ordre à buckets puisque les conditions* $A \not\prec_\kappa B$ *et* $B \not\prec_\kappa C$ *sont respectées bien que* $A \prec_\kappa C$*), et* $\{A \prec_\sigma C,\ B \prec_\sigma D\}$ *un ordre partiel (mais pas un ordre à intervalles, ce qui peut être montré par l'absurde de la façon suivante. Supposons qu'il s'agisse d'un ordre à intervalles. Puisque* $A \prec_\sigma C$, $l_A \leq r_A < l_C \leq r_C$. *En outre, comme* B *est incomparable à* A *et* C, $l_B \leq r_A < l_C \leq r_B$. *De plus, comme* $B \prec_\sigma D$, $r_B < l_D$. *Finalement, comme* $r_A < l_D$, *alors* $A \prec_\sigma D$ *: il s'agit d'une contradiction).*

[3] Le mot anglais est emprunté.

[4] Dans la plupart des communautés, un ordre partiel est une relation réflexive alors qu'un ordre partiel strict est une relation antiréflexive ; de la même façon que dans [9, 10], un abus de langage est fait dans ce mémoire, où un ordre partiel désigne une relation antiréflexive.

[5] Dans la plupart des communautés, un ordre est *total* si tous ses éléments sont comparables, et il est *partiel* sinon (autrement dit, un ordre ne peut pas être total et partiel à la fois). Dans ce mémoire et de la même façon que dans [9, 10], un ordre total désigne une généralisation d'un ordre partiel.

Finalement, il est facile de s'apercevoir qu'un ordre total est un ordre à *buckets* (avec un seul élément par *bucket*), qu'un ordre à *buckets* est un ordre à intervalles (les intervalles des éléments d'un même *bucket* se chevauchent alors que les intervalles des éléments contenus dans deux *buckets* consécutifs se suivent), et qu'un ordre à intervalles est un ordre partiel (tout ordre est partiel).

Avec ou sans enrichissement des données Il arrive parfois qu'en plus des ordres d'entrée, des informations supplémentaires, appelées *labels*, soient disponibles sur les éléments à classer (comme leur importance) et/ou sur les ordres d'entrée (comme leur qualité). Dans ce cas, les données sont dites *enrichies*.

Avec ou sans apprentissage Le consensus peut être réalisé avec ou sans *apprentissage* préalable. Cette technique, détaillée dans [39], consiste à construire des *modèles de consensus*, en créant plusieurs entrées du problème (sur un même domaine $\mathcal{D}$), puis en les faisant analyser par un oracle supposé connaître la vérité absolue. Ces instances analysées sont appelées *données préparées*. Des consensus d'ordres peuvent ensuite être réalisés sur $\mathcal{D}$, en se basant sur les informations contenues dans les "vrais" classements fournis par l'oracle.

Méthode de consensus La dernière spécification prise en compte dans ce mémoire concerne la manière de réaliser le consensus. D'après le théorème de l'impossibilité d'Arrow [4], il n'existe pas de méthode d'agrégation indiscutable. Pour être plus précis, dès lors qu'il y a trois options de choix (*i.e.*, $|\mathcal{D}| \geq 3$) et deux votants (*i.e.*, deux ordres d'entrée), aucun consensus ne peut satisfaire à la fois :

- le principe d'unanimité (ou de Pareto faible) [53] : si tous les individus préfèrent x à y alors le consensus préfère x à y,
- la non-dictatorialité [53] : le consensus ne coïncide pas en permanence avec les préférences d'un des votants indépendamment des préférences des autres votants,
- l'indépendance des alternatives [53] : dans le consensus, l'ordre relatif entre deux alternatives de choix ne doit dépendre que de leur position relative dans les préférences des différents votants (et non de la position des autres options de choix).

Cela permet finalement à une grande variété de méthodes d'exister : une partie d'entre elles est décrite dans la section suivante.

3.3 Catégorisation des méthodes de résolution existantes

Du fait des nombreuses communautés travaillant sur ce sujet, les différentes méthodes de résolution existantes se recoupent parfois sans que cela soit clairement mentionné dans les publications les présentant. Notre état de l'art nous a permis de mettre en évidence quatre grandes catégories de méthodes que nous présentons dans cette section, et qui nous permettent de structurer la littérature existante.

La première classe regroupe les méthodes de résolution basées sur l'assignation d'un score aux éléments du domaine. Les résolutions basées sur la recherche d'une médiane sont décrites dans la deuxième catégorie. Aussi, les méthodes basées sur les outils de la théorie des graphes sont exposées dans la troisième classe. Par ailleurs, des méthodes probabilistes ont été introduites pour résoudre le problème du consensus d'ordres : elles sont expliquées dans la quatrième catégorie. Enfin, d'autres types de méthodes de résolution sont donnés, bien qu'ils ne s'adaptent pas directement à la direction de recherche envisagée pour ce stage.

Méthodes basées sur un score assigné aux items Dans les méthodes décrites dans ce paragraphe, un ordre d'entrée peut contenir soit juste une information d'ordre, soit en plus de l'ordre, l'écart entre ses éléments. En outre, chaque ordre ne classe pas forcément tous les éléments de $\mathcal{D}$.

Un consensus d'ordres respecte le *critère de Condorcet* [15] s'il classe en premier le(s) *gagnant(s) de Condorcet* : il s'agit de l'élément qui est le plus petit (*i.e.*, meilleur élément) ou qui est incomparable à tous les autres éléments du domaine selon les ordres d'entrée. La procédure de Condorcet construit le consensus final en assignant le(s) gagnant(s) de Condorcet à la meilleure place, en le(s) retirant du domaine et des ordres d'entrée, puis en assignant le(s) deuxième(s) gagnant(s) de Condorcet à la deuxième meilleure place, etc. Dans l'exemple 2, les éléments C, D, B, A, E sont de tels gagnants (à différentes itérations de la procédure). Comme un tel gagnant n'existe pas toujours (c'est le cas lors de la 3$^{\text{e}}$ itération de l'exemple 2), de nombreuses variantes de la méthode de Condorcet ont été proposées. Par exemple, une des variantes définit le gagnant de Condorcet comme l'élément le mieux classé par rapport à (ou incomparable à) tous les autres éléments par une majorité absolue d'ordres d'entrée. Cette variante est utilisée lors de la 3$^{\text{e}}$ itération de l'exemple 2. De plus, Truchon *et al.* [67] ont introduit le critère de Condorcet étendu, qui stipule que s'il existe une partition (A, B) du domaine $\mathcal{D}$ telle que pour tout $x \in A$ et tout $y \in B$, une majorité absolue préfère x à y, alors le consensus final doit positionner tous les éléments de A avant ceux de B. Aussi, Jean [44] a introduit la *Black rule*, qui utilise la méthode Condorcet si un gagnant existe, sinon utilise la *méthode de Borda* décrite ci-dessous. La procédure de Condorcet et ses variantes ont été utilisées en théorie du choix social [15] et pour les méta-recherches [24, 57].

Exemple 2. *Soient le domaine* $\mathcal{D} = \{A, B, C, D, E, F, G, H\}$. *La table 1 illustre une procédure de Condorcet réalisant un consensus des ordres à buckets* $\pi_1, \ldots, \pi_4$ *suivants :*

$\pi_1 :\ \{C \prec_{\pi_1} A \prec_{\pi_1} \{G, F\}\},$

$\pi_2 :\ \{D \prec_{\pi_2} A \prec_{\pi_2} G \prec_{\pi_2} F\},$

$\pi_3 :\ \{B \prec_{\pi_3} A \prec_{\pi_3} F \prec_{\pi_3} G\},$

$\pi_4 :\ \{B \prec_{\pi_4} \{E, A\} \prec_{\pi_4} F \prec_{\pi_4} H\}.$

La méthode de Borda [19] ou *méthode de Borda-Kendall*, initialement conçue pour agréger des ordres totaux sur le domaine $\mathcal{D}$, consiste à assigner un score à chaque élément

Étape	π_1	π_2	π_3	π_4	Consensus de Condorcet
1	$C \prec A \prec \substack{F \\ G}$	$D \prec A \prec G \prec F$	$B \prec A \prec F \prec G$	$B \prec \substack{A \\ E} \prec F \prec H$	$\substack{B \\ C \\ D}$
2	$A \prec \substack{F \\ G}$	$A \prec G \prec F$	$A \prec F \prec G$	$\substack{A \\ E} \prec F \prec H$	$\substack{B \\ C \\ D} \prec \substack{A \\ E}$
3	$\substack{F \\ G}$	$G \prec F$	$F \prec G$	$F \prec H$	$\substack{B \\ C \\ D} \prec \substack{A \\ E} \prec F$ [Variante majoritaire]
4	G	G	G	H	$\substack{B \\ C \\ D} \prec \substack{A \\ E} \prec F \prec \substack{G \\ H}$

TABLE 1 – Illustration d'une procédure de Condorcet. Les gagnants de Condorcet sont d'abord B, C et D : ils sont assignés à la meilleure place du consensus (étape 1). Une fois ces éléments retirés du domaine, deux autres gagnants sont trouvés : A et E (étape 2). À cette étape de la procédure, plus aucun gagnant (selon la définition initiale) n'est trouvé. En effet, il n'est pas possible de départager les éléments F et G. Dès lors, en utilisant la variante qui définit le gagnant de Condorcet comme l'alternative la mieux classée par rapport à (ou incomparable à) toutes les autres alternatives par une majorité absolue d'ordres d'entrée, F est gagnant (étape 3). Finalement, il est impossible de départager G et H, c'est pourquoi ils arrivent en dernière position du consensus (étape 4).

du domaine. Le score ω assigné à un élément e est la somme du nombre d'éléments f moins bien classés que lui dans chaque ordre d'entrée σ_i, *i.e.*, $\omega(e) = \sum_{i=1}^{m} |\{f, e \prec_{\sigma_i} f\}|$. Le consensus final est obtenu en classant les éléments du domaine par ordre décroissant de leur score, avec les plus grands en premier. L'exemple 3 illustre cette procédure. La méthode de Borda fait partie de celles les plus largement utilisées, et ce dans différents champs d'application, comme les méta-recherches [5, 60, 21], la classification automatique [41, 68, 61], les systèmes de recommendation pour un collectif [6], l'établissement des priorités des patients aux urgences [34], la récupération d'informations [54, 59], ainsi que le traitement des langages naturels [54]. Cette procédure peut être implémentée en temps linéaire, mais ne permet pas de garantir l'indépendance des alternatives (voir l'exemple 3). Des méthodes permettant de normaliser les scores $\omega(.)$ peuvent être appliquées quand les ordres d'entrée ne sont pas tous de même taille.

Exemple 3. *Soient le domaine $\mathcal{D}$ et les ordres d'entrée $\pi_1, \ldots, \pi_4$ de l'exemple 2. Les scores finaux assignés aux éléments du domaine selon la méthode de Borda sont $\omega(A) = 8$ (car A possède deux éléments plus petits que lui dans chacun des quatre π_i), $\omega(B) = 7$ (car B possède trois éléments plus petits que lui dans π_3, quatre dans π_4, et aucun dans π_1 et π_2), $\omega(C) = \omega(D) = 3$, $\omega(E) = \omega(F) = 2$, $\omega(G) = 1$, et $\omega(H) = 0$. Le consensus construit avec la méthode de Borda est donc $A \prec B \prec \{C, D\} \prec \{E, F\} \prec G \prec H$. Il est intéressant de remarquer que l'indépendance des alternatives n'est pas satisfaite. En effet, si dans π_2, F est déplacé entre D et A, et que dans π_3, F est déplacé entre B et A, le nouveau consensus obtenu selon la méthode de Borda classe B avant A, alors que leur position relative n'a été changé dans aucun des ordres d'entrée modifiés.*

Une généralisation de la méthode de Borda est la méthode d'*agrégation linéaire* : elle est utilisée quand des scores donnant un écart entre les éléments sont disponibles (par exemple, si l'on considère le classement d'élèves en fonction de leurs notes dans différentes matières). Différentes stratégies d'agrégation linéaire, *CombSUM*, *CombMIN*, *CombMAX*, *CombANZ*, et *CombMNZ*, ont été proposées par Shaw et Fox [65]. CombSUM classe les éléments du domaine selon la somme des scores qu'ils ont obtenus dans chacun des ordres d'entrée (équivalent à la méthode de Borda quand les scores $\omega(.)$ sont utilisés, voir le paragraphe précédent) ; CombMIN (resp. CombMAX) les classe selon le score minimum (resp. maximum) qu'ils ont reçu à travers tous les ordres d'entrée ; avec CombANZ (resp. CombMNZ), les éléments sont classés par ordre croissant de leur score CombSUM divisé (resp. multiplié) par le nombre d'ordres d'entrée contenant l'élément. Cette approche ne garantit pas le critère de Condorcet (voir l'exemple 4). Par ailleurs, de la même façon que pour la méthode de Borda, cette approche a plus de sens quand les ordres d'entrée sont totaux ; elle peut être étendue à des ordres plus spécifiques (*e.g.*, à *buckets*, à intervalles, etc.) en normalisant les scores par rapport à la taille des ordres d'entrée. Cette méthode d'agrégation linéaire a été utilisée pour la classification automatique [41], pour les méta-recherches [5, 60], pour les systèmes de recommendation pour un collectif [6], ainsi que pour la récupération d'informations [65, 31, 59].

Exemple 4. *Soient le domaine* $\mathcal{D} = \{A, B, C, D, E, F, G, H\}$*, ainsi que les ordres d'entrée* $\pi_1, \ldots, \pi_4$ *(déjà vus dans les exemples 2 et 3) et les scores d'entrée associés* $\omega_{\pi_1}, \ldots, \omega_{\pi_4}$ *suivants :*

π_1 : $\{C \prec_{\pi_1} A \prec_{\pi_1} \{G, F\}\}$, *avec* $\omega_{\pi_1}(C) = 3$, $\omega_{\pi_1}(A) = 2$, *et* $\omega_{\pi_1}(G) = \omega_{\pi_1}(F) = 0$,

π_2 : $\{D \prec_{\pi_2} A \prec_{\pi_2} G \prec_{\pi_2} F\}$, *avec* $\omega_{\pi_2}(D) = 3$, $\omega_{\pi_2}(A) = 2$, $\omega_{\pi_2}(G) = 1$, *et* $\omega_{\pi_2}(F) = 0$,

π_3 : $\{B \prec_{\pi_3} A \prec_{\pi_3} F \prec_{\pi_3} G\}$, *avec* $\omega_{\pi_3}(B) = 3$, $\omega_{\pi_3}(A) = 2$, $\omega_{\pi_3}(F) = 1$, *et* $\omega_{\pi_3}(G) = 0$,

π_4 : $\{B \prec_{\pi_4} \{E, A\} \prec_{\pi_4} F \prec_{\pi_4} H\}$, *avec* $\omega_{\pi_4}(B) = 4$, $\omega_{\pi_4}(E) = \omega_{\pi_4}(A) = 3$, $\omega_{\pi_4}(F) = 1$, *et* $\omega_{\pi_4}(H) = 0$.

Le consensus construit selon CombMAX en utilisant les scores d'entrée $\omega_{\pi_i}(.)$ *est* $B \prec \{C, D\} \prec \{A, E\} \prec \{F, G\} \prec H$*. En effet,* $\max_{i=1}^{4} \omega_i(B) = 4$, $\max_{i=1}^{4} \omega_i(C) = \max_{i=1}^{4} \omega_i(D) = 3$, $\max_{i=1}^{4} \omega_i(A) = \max_{i=1}^{4} \omega_i(E) = 2$, $\max_{i=1}^{4} \omega_i(F) = \max_{i=1}^{4} \omega_i(G) = 1$, *et* $\max_{i=1}^{4} \omega_i(H) = 0$. *Dans cette procédure, le critère de Condorcet n'est pas respecté (voir l'exemple 2 pour comparaison).*

Les méthodes décrites jusqu'ici traitent tous les ordres d'entrée comme égaux, alors qu'il peut être intéressant d'accorder à chacun plus ou moins de crédit. Yager [72] a introduit les opérateurs pondérés, très utilisés par les domaines dans lesquels une décision, selon différents critères, doit être prise. Par exemple, une personne souhaitant acheter un véhicule va s'intéresser au prix, mais aussi à la puissance du moteur, ainsi qu'à la consommation du véhicule. Pour prendre sa décision, la personne souhaite se baser sur

ces trois critères ; cependant, il peut vouloir accorder plus d'importance à certains qu'à d'autres (le prix est plus important que la puissance du moteur, par exemple). La prise de décision multi-critères peut être vue comme un problème de consensus d'ordres, dans lequel chaque critère joue le rôle d'un ordre d'entrée. Pour cette méthode, chaque élément e se voit attribuer un score final $\Omega(e)$ en combinant les scores $\omega(e, \sigma_i)$ de cet élément dans chaque ordre d'entrée σ_i. La résolution proposée par Yager est basée sur la famille de fonctions d'utilités paramétrée par p :

$$\Omega(e) = \left(\frac{\sum_{i=1}^{m} \omega(e, \sigma_i)^p}{m} \right)^{\frac{1}{p}}$$

Cette famille procure une grande variété de fonctions d'agrégation, par exemple :

- $p = 1$: moyenne arithmétique,
- $p = -1$: moyenne harmonique,
- $p \to 0$: moyenne géométrique,
- $p \to +\infty$: *max* (généralise l'opérateur "OU"),
- $p \to -\infty$: *min* (généralise l'opérateur "ET"),

La généralisation de l'opérateur "ET" préfère l'élément qui n'a pas de gros point faible selon les différents critères, alors que celle de l'opérateur "OU" préfère celui qui a un gros point fort sur un critère. Finalement, les domaines d'application ayant recours à cette méthode de résolution sont ceux qui doivent prendre une décision selon plusieurs critères, comme l'établissement des priorités aux urgences [34], la récupération d'informations [28], ou les méta-recherches [25].

Méthodes par recherche d'une médiane Il est possible d'utiliser la notion de *distance* pour obtenir un consensus d'ordres. Une distance entre deux ordres d'un ensemble $\mathcal{O}$ est une fonction $d : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$ qui satisfait, pour tout $\sigma_x, \sigma_y, \sigma_z \in \mathcal{O}$, les conditions de :

- non-négativité (*i.e.*, $d(\sigma_x, \sigma_y) \geq 0$),
- identité des indiscernables (*i.e.*, $d(\sigma_x, \sigma_y) = 0 \Leftrightarrow \sigma_x = \sigma_y$),
- symétrie (*i.e.*, $d(\sigma_x, \sigma_y) = d(\sigma_y, \sigma_x)$),
- inégalité triangulaire (*i.e.*, $d(\sigma_x, \sigma_y) \leq d(\sigma_x, \sigma_z) + d(\sigma_z, \sigma_y)$).

Étant donnés une distance d et un ensemble de m ordres σ_i sur $\mathcal{D}$, un *ordre médian*, aussi appelé *médiane*, est un ordre σ^* sur $\mathcal{D}$ qui minimise la somme des distances d entre σ^* et les σ_i. Autrement dit, une médiane est un ordre σ^* qui minimise $\sum_{i=1}^{m} d(\sigma^*, \sigma_i)$.

Une des distances très largement étudiée pour le consensus d'ordres est celle de *Kendall-tau*, notée $\mathcal{K}$. Elle est définie sur les ordres totaux, c'est-à-dire des ordres représentant des permutations du domaine. Cette distance mesure le nombre de paires

d'éléments $\{e, e'\}$ pour lesquelles deux ordres totaux τ_1 et τ_2 sur un domaine $\mathcal{D}$ sont en désaccord, *i.e.*, $\mathcal{K}(\tau_1, \tau_2) = |\{\{e, e'\}, e \prec_{\tau_1} e' \text{ et } e' \prec_{\tau_2} e \text{ ou } e' \prec_{\tau_1} e \text{ et } e \prec_{\tau_2} e'\}|$.

Exemple 5. *Soit le domaine $\mathcal{D} = \{A, B, C, D\}$ et deux ordres totaux τ_1 et τ_2 sur $\mathcal{D}$ représentés par $\{B \prec_{\tau_1} D \prec_{\tau_1} A \prec_{\tau_1} C\}$ et $\{A \prec_{\tau_2} D \prec_{\tau_2} C \prec_{\tau_2} B\}$. La distance de Kendall-tau entre τ_1 et τ_2 est $\mathcal{K}(\tau_1, \tau_2) = 4$, puisque τ_1 et τ_2 sont en désaccord pour l'ordre concernant les paires $\{A, B\}$, $\{A, D\}$, $\{B, C\}$ et $\{B, D\}$.*

La recherche d'une médiane sous la distance de Kendall-tau est appelé le problème de *l'agrégation optimale de Kemeny*, et permet de définir une notion de "consensus optimal" [24]. Cependant, obtenir une telle agrégation est **NP**-complet [7, 24], même avec seulement 4 ordres d'entrée (*i.e.*, $m = 4$).

Des heuristiques et algorithmes d'approximation ont été donnés par Dwork *et al.* [24] concernant le problème de l'agrégation optimale de Kemeny. De plus, différentes recherches [7, 18, 16, 47, 69, 2, 62, 38] ont étudié le problème de l'agrégation optimale de Kemeny comme un cas particulier de *minimum (weighted) feedback arc set*[6]. Les méthodes de résolution pour ce dernier ont donc pu être utilisées pour résoudre ou approcher l'agrégation optimale de Kemeny. C'est de cette façon qu'un algorithme exact de *branch and bound* [18], des heuristiques [18], des algorithmes d'approximation [2, 69, 16] ainsi qu'un PTAS[7] [47] ont pu être introduits pour la résolution de l'agrégation optimale de Kemeny. Comme la distance de Kendall-tau est définie sur des ordres totaux (*i.e.*, des permutations), des extensions de l'agrégation optimale de Kemeny ont été proposées [14, 58] pour des ordres à *buckets*, puis pour des ordres à intervalles et partiels [9, 10]. Différents domaines d'application se sont intéressés à ce problème : la théorie du choix social [7, 18, 16, 47, 69, 2, 62, 9, 10, 38, 45, 58], l'établissement des priorités aux urgences [34], les méta-recherches [24, 21], la classification automatique [61], ou encore le consensus de cartes génomiques [14, 74].

De plus, Jiao *et al.* [45] introduisent une méthode permettant de prédire, étant donnée n'importe quelle procédure d'agrégation, à quel point le consensus de sortie fournit par cette même procédure est proche d'une agrégation optimale de Kemeny (sans la calculer). Pour cela, ils s'appuient sur des propriétés géométriques des agrégations optimales de Kemeny dans un espace euclidien.

Par ailleurs, une autre distance très étudiée est celle de *Spearman-footrules*, notée $\mathcal{F}$. Elle est définie entre deux ordres totaux τ_1 et τ_2 sur $\mathcal{D}$ et est égale à la somme, sur tous les éléments j de $\mathcal{D}$, de la différence absolue entre les positions de j dans les deux ordres. Autrement dit, $\mathcal{F}(\tau_1, \tau_2) = \sum_{j=1}^{|\mathcal{D}|} |\tau_1(j) - \tau_2(j)|$, avec $\tau_i(j)$ la position de l'élément j dans l'ordre τ_i. Par ailleurs, l'inégalité

$$\mathcal{K}(\tau_1, \tau_2) \leq \mathcal{F}(\tau_1, \tau_2) \leq 2\,\mathcal{K}(\tau_1, \tau_2) \tag{1}$$

[6] Il s'agit, étant donné un graphe orienté, de trouver un ensemble d'arcs (de poids) minimum qui, une fois retirés, laisse le graphe acyclique. Ce problème est **NP**-complet.

[7] Le mot anglais *Polynomial Time Approximation Scheme* est emprunté. Un PTAS est une famille d'algorithmes paramétrés qui, étant donné un problème Π **NP**-difficile, permet d'approcher -autant que souhaité- une solution optimale de Π, en temps polynomial en la taille de l'entrée. Cependant, plus l'approximation souhaitée est fine, plus le temps d'obtention de la solution est augmenté.

a été montrée par Diaconis et Graham [22]. Dwork *et al.* [24] détaillent un algorithme polynomial, basé sur la recherche d'un couplage parfait de coût minimum dans un graphe biparti pondéré complet, qui retourne une médiane d'un ensemble d'ordres d'entrée totaux sous la distance de Spearman-footrules (plus de détails sont donnés dans [24, 6]). Il est possible de montrer, en utilisant l'équation 1, qu'une médiane pour un ensemble d'ordres d'entrée totaux sous la distance de Spearman-footrules est une 2-approximation efficace du problème de l'agrégation optimale de Kemeny [24]. La recherche de médiane sous la distance de Spearman-footrules est étudiée dans différents domaines d'application, comme la théorie du choix social [22, 62], l'établissement des priorisations des patients aux urgences [33, 34], les méta-recherches [24], les systèmes de recommendation pour un collectif [6], ou encore la fusion de cartes génomiques [74].

Exemple 6. *Soit le domaine $\mathcal{D}$ et les deux ordres totaux τ_1 et τ_2 de l'exemple 5. La distance de Spearman-footrules entre τ_1 et τ_2 est $\mathcal{F}(\tau_1, \tau_2) = |3-1| + |1-4| + |4-3| + |2-2| = 6$ (en regardant les éléments dans l'ordre A, B, C puis D).*

Les distances de Kendall-tau et de Spearman-footrules ont été généralisées, de différentes façons, pour des ordres à *buckets* [29], et pour des ordres partiels [24, 1, 9, 10]. Cependant, la recherche d'une médiane sous la distance de Spearman-footrules devient un problème **NP**-difficile [24, 9] pour des ordres d'entrée partiels. Dans les différentes extensions susmentionnées, l'inégalité de Diaconis et Graham reste vérifiée, ce qui a permis à différentes approximations d'être introduites [1, 9] pour le problème de la recherche d'une médiane.

Par une approche théorique, le livre de Fertin *et al.* [32] formalise et classifie les différents problèmes liés aux calculs de distance entre deux ordres dans le contexte des réarrangements génomiques. Les distances étudiées sont celles correspondant à des événements biologiques : transpositions, translocations, inversions, ou encore des combinaisons de ces événements. La formalisation et classification des problèmes de recherche d'une médiane (sous ces mêmes distances) sont aussi données. Dans ce livre, les événements biologiques pris en compte permettent de retracer l'histoire évolutive des espèces : les ordres d'entrée sont supposés sans erreur (*i.e.*, correspondent à des vrais génomes), et chercher une médiane est vu comme une manière (parcimonieuse) d'inférer l'organisation (carte) du génome de l'espèce ancestrale du groupe étudié. De plus, la distance utilisée permet d'obtenir une prédiction des événements génomiques qui se sont produits au cours de l'évolution de ce groupe d'espèces.

Méthodes de la théorie des graphes Une instance du problème du consensus d'ordres partiels peut être représentée par un graphe orienté $G = (V, A)$ dont les sommets sont les éléments du domaine $\mathcal{D}$; un arc $\overrightarrow{xy}$ existe si, et seulement si, il existe un ordre d'entrée σ_i tel que l'élément x est un prédécesseur strict de y (*i.e.*, $x \prec_{\sigma_i} y$ et il n'existe pas d'élément z dans tel que $x \prec_{\sigma_i} z \prec_{\sigma_i} y$). Un tel graphe est appelé *graphe des préférences* : un exemple est donné dans la figure 1b ci-dessous. Lorsque G est acyclique, les ordres d'entrée ne se contredisent pas : un simple tri topologique de G permet de trouver un consensus optimal. Cependant, lorsqu'il existe un cycle orienté dans G,

il existe au moins deux ordres en désaccord sur au moins deux de leurs éléments. Dès lors, résoudre le problème du consensus d'ordres peut être ramené à trouver un ensemble minimum d'arcs X tel que $G = (V, A - X)$ est acyclique : il s'agit du problème du *minimum feedback arc set* introduit précédemment (cf. note 6 de bas de page 11). Dans le graphe des préférences de la figure 1b, l'ensemble minimum X d'arcs laissant le graphe acyclique est soit $\{\overrightarrow{GF}\}$ soit $\{\overrightarrow{FG}\}$.

De plus, la même représentation, mais avec des pondérations sur les arcs, a été introduite [18] : le poids d'un arc $\overrightarrow{xy}$ correspond au nombre d'ordres d'entrée σ_i tels que $x \prec_{\sigma_i} y$ moins le nombre d'ordres d'entrée σ_j tels que $y \prec_{\sigma_j} x$ (les arcs à pondération strictement négative sont retirés). Une telle pondération est donnée dans la figure 1c.

Montague et Aslam [57] ont introduit le *graphe de Condorcet* : il s'agit du graphe orienté $G = (V, A)$ dont les sommets sont les éléments du domaine $\mathcal{D}$; pour toute paire d'éléments (x, y) de $\mathcal{D}$, il existe un arc $\overrightarrow{xy}$ si, et seulement si, il y a au moins autant d'ordres d'entrée qui considèrent x meilleur que y qu'il y en a qui considèrent y meilleur que x, ou si x et y sont incomparables à travers tous les ordres d'entrée. La figure 1d illustre un graphe de Condorcet. Montague et Aslam décrivent un algorithme permettant de trouver un consensus en $\mathcal{O}(|\mathcal{D}|mlog|\mathcal{D}|)$, basé sur la recherche d'un chemin Hamiltonien (*i.e.*, un chemin qui passe par tous les sommets).

Par ailleurs, Desarkar *et al.* [21] se sont intéressés à assigner des poids aux ordres d'entrée, en donnant plus de poids aux "meilleurs" ordres (moins un ordre est en désaccord avec les autres ordres d'entrée concernant l'ordre entre deux éléments, plus il est bon), sans que les données soient enrichies, avant de réaliser un consensus. Différents travaux se sont intéressés à ce type de représentation, en théorie du choix social [18, 38], en consensus de cartes génomiques [75, 26, 66], ou encore en méta-recherches [57, 21].

Le stage facultatif [20] que j'ai effectué lors de mon Master 1 concernait une variante particulière du consensus d'ordres. Dans cette version du problème, une instance est représentée par un graphe non orienté possédant deux types d'arêtes. Le premier type d'arêtes modélise des adjacences fiables à longue distance (*i.e.*, d'autres éléments peuvent être insérés entre deux éléments adjacents à longue distance), et est contenu dans le premier ordre d'entrée. Le second type d'arêtes représente des adjacences strictes (*i.e.*, aucune insertion n'est possible entre deux éléments adjacents strictement) issues de *prédictions*, c'est-à-dire pouvant être remises en question : ce type d'information est contenu dans les $m-1$ autres ordres d'entrée. L'objectif est alors de trouver un ensemble minimum d'arêtes d'adjacences strictes qui, une fois retirées, permettent d'obtenir un graphe sans contradiction sur les informations d'ordres restantes. Nous avons montré la **NP**-complétude de ce problème et proposé un algorithme heuristique.

Méthodes probabilistes Différentes méthodes basées sur les *chaînes de Markov* ont été utilisées pour le consensus d'ordres. Il s'agit d'un automate dans lequel les éléments du domaine sont considérés comme les états ; pour chaque paire d'éléments (i_1, i_2), la probabilité de transition correspondante $(i_1 \to i_2)$ dépend du nombre d'ordres d'entrée contenant les deux éléments, ainsi que du nombre d'ordres d'entrée qui ont ordonné i_1 avant i_2. Quatre méthodes ont été proposées par [24] : MC_1, MC_2, MC_3 et MC_4. Elles

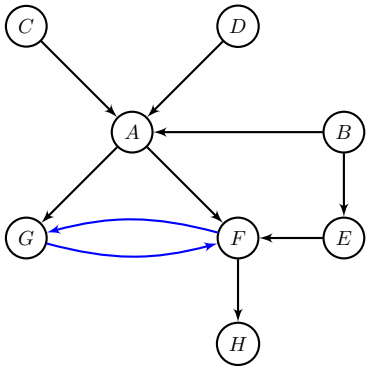
$$\{C \prec_{\pi_1} A \prec_{\pi_1} \{G, F\}\}$$

$$\{D \prec_{\pi_2} A \prec_{\pi_2} G \prec_{\pi_2} F\}$$

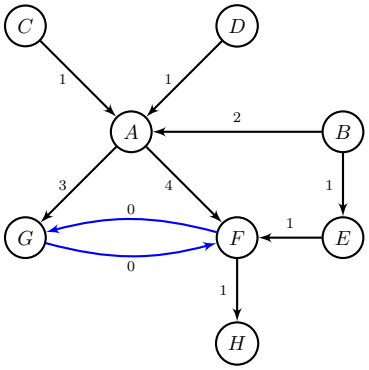
$$\{B \prec_{\pi_3} A \prec_{\pi_3} F \prec_{\pi_3} G\}$$

$$\{B \prec_{\pi_4} \{E, A\} \prec_{\pi_4} F \prec_{\pi_4} H\}$$

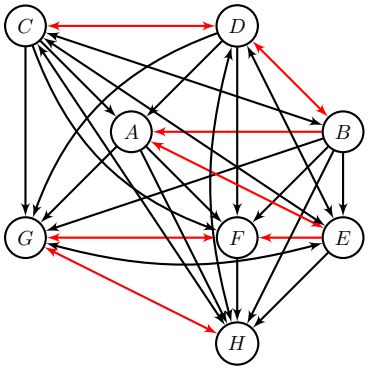
(a) Ordres d'entrée (déjà vus dans les exemples 2 à 4)



(b) Graphe des préférences



(c) Graphe des préférences pondéré



(d) Graphe de Condorcet

FIGURE 1 – Graphe des préférences (b), graphe des préférences pondéré (c) et graphe de Condorcet (d) des ordres d'entrée $\pi_1, \ldots, \pi_4$ (a). Dans (c), le poids associé à un arc $\overrightarrow{xy}$ correspond au nombre d'ordres d'entrée π_i tels que $x \prec_{\pi_i} y$, moins le nombre d'ordres d'entrée π_j tels que $y \prec_{\pi_j} x$. Par exemple, l'arc $\overrightarrow{FG}$ a une pondération de 0 car F est préféré à G dans π_3, et G est préféré à F dans π_2 (π_1 et π_4 ne se prononcent pas). Par ailleurs, un unique cycle (dessiné en bleu), $\{G, F\}$, est présent dans ce graphe. Dans (d), un chemin Hamiltonien (par exemple, le chemin dessiné en rouge) correspond à un consensus qui contredit le moins possible les ordres d'entrée. De plus, pour éviter de surcharger le graphe, deux arcs $\overrightarrow{CD}$ et $\overrightarrow{DC}$ sont condensés en un trait avec deux flèches, *i.e.*, $\overleftrightarrow{CD}$.

différent dans la façon dont elles calculent les probabilités de transition. Ces méthodes ont été utilisées pour les méta-recherches [60, 21], ainsi que pour la récupération d'informations [31, 59]. Par ailleurs, Fang *et al.* [30] soulignent le fait que les méthodes MC_1 à MC_4 ne prennent pas en compte les positions des éléments dans les ordres d'entrée pour construire le consensus final. C'est pour pallier ce défaut que ces auteurs ont introduit la méthode basée sur les *chaînes de Markov pondérées*.

D'autres modèles probabilistes ont été testés, comme le modèle de Mallows [13, 48, 49]. Les algorithmes basés sur ce modèle sont en $\mathcal{O}(|\mathcal{D}|!)$. Les modèles probabilistes ont été utilisés pour les méta-recherches [5, 21].

Par ailleurs, d'autres méthodes de consensus probabilistes peuvent être utilisées lorsqu'une phase d'apprentissage permet de préparer les données. Deux phases composent ce type de méthode : la phase d'apprentissage, et la phase de consensus. La première consiste à construire un modèle de consensus en comparant le consensus de sortie de plusieurs instances du problème (selon une certaine méthode de résolution) avec le consensus fourni par un oracle. La deuxième phase consiste alors à utiliser ce modèle précédemment construit pour améliorer la prédiction des consensus futurs : des scores sont assignés aux éléments du domaine selon le modèle, ce qui permet d'ordonner les éléments selon ces scores, et donc d'obtenir un consensus. Récemment, de nombreux domaines et champs d'application (comme les systèmes de recommendation pour un collectif [70], la récupération d'informations [55, 12, 54, 71], les méta-recherches [63, 12], ou les langages naturels [54]) se sont intéressés au gain de performance entre un consensus d'ordres avec ou sans apprentissage.

Autres types de résolution Ce qui nous intéresse dans ce stage est non seulement d'obtenir une carte consensus, mais également de pouvoir inférer et interpréter les opérations (qu'il s'agisse d'un événement biologique ou d'une erreur lors de la création d'une carte d'entrée) qui ont conduit à des contradictions au sein des cartes d'entrée. Autrement dit, en plus d'un consensus de sortie, une explication détaillant pourquoi il s'agit du meilleur consensus (en terme d'opérations survenues), est attendue. Les méthodes de résolution données dans ce paragraphe ne permettent pas de donner de telles explications, c'est pourquoi nous ne les détaillons pas.

Les méthodes de programmation linéaires ont été utilisées pour agréger un ensemble d'ordres d'entrée. Elles ont par exemple été utilisées pour les méta-recherches [3], pour la priorisation des patients aux urgences [34], ou en théorie du choix social [62]. Par ailleurs, lorsque les ordres d'entrée sont totaux, différents articles de la littérature mentionnent que retourner le "meilleur" des ordres d'entrée (*i.e.*, celui qui est le moins en désaccord avec les autres ordres d'entrée concernant l'ordre entre deux éléments) en tant que consensus de sortie est une 2-approximation de l'agrégation optimale de Kemeny [2]. Cette approximation a surtout été utilisée en théorie du choix social [62, 2, 1]. D'autre part, trois différentes approches basées sur des algorithmes de tri (le tri rapide, le tri fusion, et le tri par insertion) sont utilisés par Schalekamp et van Zuylen [62] pour le consensus d'ordres partiels. La comparaison des éléments dans ces algorithmes de tri est basée sur l'avis de la majorité (non stricte) des ordres d'entrée. Finalement, étant donné une instance du consensus d'ordres, Guénoche [38] s'est intéressé, non plus à chercher un unique consensus, mais à partitionner les ordres d'entrée en plusieurs ensembles "homogènes", puis à chercher un consensus pour chacun d'entre eux. À l'évidence, ce problème de consensus multiples est bien différent du problème de consensus (simple) d'ordres.

La table 2 ci-dessous présente une vue synthétique des références bibliographiques

citées dans cette sous-section, en les regroupant par domaine d'application et méthode de résolution.

Domaine d'application / **Méthode de résolution**		Théorie du choix social (votes)	Systèmes de recommendation pour un collectif	Priorités aux urgences	Langages naturels	Récupération d'informations	Méta-recherches	Classification automatique	Fusion de (ou distance entre) cartes génomiques
Scores sur éléments	Condorcet	[15, 67]					[24, 57]		
	Borda	[19]	[6]	[34]	[54]	[54, 59]	[5, 60, 21]	[41, 68, 61]	
	CombX		[6]			[65, 31, 59]	[5, 60]	[41]	
	Opérateurs pondérés			[34]		[28]	[25]		
Recherche de médiane	Médiane sous $\mathcal{K}$	[22, 7, 18, 16, 47, 69, 2, 62, 9, 10, 58]		[34]			[24, 21]	[61]	[74]
	Médiane sous $\mathcal{F}$	[22, 62]	[6]	[33, 34]			[24]		[32, 74]
	Médiane sous distance d'événements								[32]
Graphes	Graphes	[7, 18, 16, 47, 69, 2, 62, 38]					[57, 21]		[75, 26, 66, 20]
Probabilistes	Méthodes probabilistes					[31, 59]	[5, 60, 21]		
	Avec apprentissage		[70]		[54]	[55, 54, 71]	[63]		
Autres	Prog. linéaire	[62]		[34]			[3]		
	Méthode aléatoire	[62, 2, 1]							
	Algo. de tri	[62]							
	Consensus multiples	[38]							

TABLE 2 – Synthèse des références bibliographiques par domaines d'application et méthodes de résolution. Le domaine d'application de la fusion de (ou distance entre) cartes génomiques regroupe aussi bien de l'étude de distance d'événements pour l'inférence de cartes ancestrales, que le problème de la fusion de cartes génomiques contemporaines. Les méthodes de résolution ne sont pas spécifiques à un domaine d'application : elles sont très souvent utilisées par différentes communautés. Cependant, seulement la communauté des réarrangements génomiques [32] s'est intéressée à la recherche d'une médiane sous une distance correspondant à un ensemble d'événements complexes. Cette référence [32] n'est pas un article de recherche mais un livre regroupant différents travaux sur les réarrangements génomiques, en formalisant et donnant les classes de complexité des problèmes étudiés.

4 Direction de recherche envisagée

Le problème de la fusion de cartes tente d'améliorer la carte du génome d'une espèce en en combinant plusieurs. Comme des erreurs ont pu être introduites lors de la construction des cartes d'entrée, il faut les prendre en compte lors du consensus : on s'intéressera aux erreurs dues à des inversions de segments par les techniques d'assemblage (*i.e.*, inversion de marqueurs contigus), ainsi qu'aux marqueurs mal placés (*i.e.*, délétion et insertion d'un marqueur). Ainsi, la combinaison de cartes génomiques doit prendre en compte ces différentes opérations dans sa "fonction objectif". De plus, comme les cartes d'entrée sont rarement construites à partir d'un même individu, les événements biologiques (*i.e.*, transpositions, translocations, inversions, ou encore des combinaisons de ces événements) pouvant se produire au sein d'une même population pourraient aussi être considérés.

L'objectif du stage est de formaliser et aborder la complexité du problème de la fusion de cartes génomiques. Pour cela, différentes variantes pourront être considérées, suivant :

- le type des ordres d'entrée : partiels, à intervalles, à *buckets*, totaux sur $\mathcal{D}$, ou totaux sur un sous-ensemble de $\mathcal{D}$.
- le type de l'ordre de sortie (*i.e.*, du consensus) : il s'agira le plus souvent d'un ordre total sur $\mathcal{D}$, cependant, un consensus de sortie total seulement sur un sous-ensemble de $\mathcal{D}$ pourra être envisagé.
- avec ou sans enrichissement des données : on pourra s'intéresser à des données non enrichies, mais aussi à des données enrichies. Il est possible de ne pas accorder la même confiance à chacune des cartes d'entrée, ce qui pourra être modélisé par des labels de fiabilité sur les ordres d'entrée. En outre, la position d'un marqueur dans une carte d'entrée peut être plus ou moins fiable, et une information de distance entre deux marqueurs consécutifs d'une carte d'entrée peut être connue : cela pourra être modélisé par des labels sur les éléments de $\mathcal{D}$.
- la réalisation ou non d'une phase d'apprentissage : s'il est possible "d'entraîner" nos données sur des méthodes (il faut pour cela avoir accès à des fragments de génomes *supposés* totalement reconstruits), nous pourrons nous intéresser au gain de performance entre un consensus avec ou sans apprentissage.

Pour aborder ces différentes variantes, la méthode de résolution que nous privilégierons est celle basée sur la théorie des graphes, puisqu'elle permet :

1. d'avoir une vue globale du problème lors de sa résolution (principe de *total evidence*[8]),
2. de connaître toutes les opérations (erreurs dans la création des cartes génomiques d'entrée et/ou événements biologiques décrivant l'histoire évolutive) ayant eu lieu entre le consensus et les cartes d'entrée.

[8]Le mot anglais est emprunté.

Bien que différentes recherches se soient déjà intéressées à la résolution de la fusion de cartes en utilisant les outils des graphes [75, 26, 66], à notre connaissance, ces méthodes de résolution n'ont pas été abordées selon une approche théorique.

Finalement, il pourrait être avantageux d'utiliser le consensus multiple introduit par Guénoche [38], c'est-à-dire de partitionner les ordres d'entrée en des ensembles "homogènes", avant de fusionner indépendamment chacun d'entre eux.

Bibliographie

[1] Nir Ailon. Aggregation of partial rankings, p-ratings and top-m lists. *Algorithmica*, 57(2):284–300, 2010.

[2] Nir Ailon, Moses Charikar et Alantha Newman. Aggregating Inconsistent Information : Ranking and Clustering. 2005.

[3] Gholam R. Amin et Ali Emrouznejad. Optimizing Search Engines Results Using Linear Programming. *Expert Syst. Appl.*, 38(9):11534–11537, 2011.

[4] K.J. Arrow. *Social Choice and Individual Values*. Cowles Commission Monographs. Wiley, 1951.

[5] Javed A. Aslam et Mark Montague. Models for Metasearch. *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 276–284, 2001.

[6] Linas Baltrunas, Tadas Makcinskas et Francesco Ricci. Group recommendations with rank aggregation and collaborative filtering. *In Proceedings of the fourth ACM conference on Recommender systems*, pages 119–126. ACM, 2010.

[7] J. Bartholdi, C. A. Tovey et M. A. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6(2):157–165, 1989.

[8] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, Ophir Frieder, David Grossman et Nazli Goharian. Disproving the Fusion Hypothesis : An Analysis of Data Fusion via Effective Information Retrieval Strategies. *In Proceedings of the 2003 ACM Symposium on Applied Computing ACM-SAC*, pages 1–5, 2003.

[9] Franz J Brandenburg, Andreas Gleissner et Andreas Hofmeier. The nearest neighbor Spearman footrule distance for bucket, interval, and partial orders. *In Frontiers in Algorithmics and Algorithmic Aspects in Information and Management*, pages 352–363. Springer, 2011.

[10] Franz-Josef Brandenburg, Andreas Gleissner et Andreas Hofmeier. Comparing and Aggregating Partial Orders with Kendall tau Distances. *Discrete Math., Alg. and Appl.*, 5(2), 2013.

[11] Samuel Brody, Roberto Navigli et Mirella Lapata. Ensemble Methods for Unsupervised WSD. *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, pages 97–104, 2006.

[12] Shouchun Chen, Fei Wang, Yangqiu Song et Changshui Zhang. Semi-supervised Ranking Aggregation. *Inf. Process. Manage.*, 47(3):415–425, 2011.

[13] Stéphan Clémençon et Jérémie Jakubowicz. *Kantorovich Distances between Rankings with Applications to Rank Aggregation*, pages 248–263. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[14] Sarah Cohen-Boulakia, Alain Denise et Sylvie Hamel. Using medians to generate consensus rankings for biological data. *In International Conference on Scientific and Statistical Database Management*, pages 73–90. Springer, 2011.

[15] Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785.

[16] Don Coppersmith, Lisa Fleischer et Atri Rudra. Ordering by weighted number of wins gives a good ranking for weighted tournaments. *In Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 776–782. Society for Industrial and Applied Mathematics, 2006.

[17] Andrew Crossen, Jay Budzik et Kristian J. Hammond. Flytrap : Intelligent group music recommendation. *In Proceedings of the 7th International Conference on Intelligent User Interfaces*, IUI '02, pages 184–185, New York, NY, USA, 2002. ACM.

[18] Andrew Davenport et Jayant Kalagnanam. A computational study of the Kemeny rule for preference aggregation. *In AAAI*, volume 4, pages 697–702, 2004.

[19] Jean de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences, Paris*, 1781.

[20] Lisa de Mattéo. Étude d'un problème de graphe pour la compatibilité de données génomiques, 2016. Stage de Master 1 Informatique Théorique.

[21] Maunendra Sankar Desarkar, Sudeshna Sarkar et Pabitra Mitra. Preference relations based unsupervised rank aggregation for metasearch. *Expert Systems with Applications*, 49:86 – 98, 2016.

[22] Persi Diaconis et Ronald L Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 262–268, 1977.

[23] Daniel Dreilinger et Adele E Howe. Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems (TOIS)*, 15(3):195–222, 1997.

[24] C. Dwork, R. Kumar, M. Naor et D. Sivakumar. Rank aggregation methods for the Web. *In Proceedings of the 10th International World Wide Web Conference, WWW 10, Hong Kong, China, May 1-5, 2001*, pages 613–622, 2001.

[25] Ali Emrouznejad. MP-OWA : The most preferred OWA operator. *Knowledge-Based Systems*, 21(8):847–851, 2008.

[26] Jeffrey B Endelman. New algorithm improves fine structure of the barley consensus SNP map. *BMC Genomics*, 12(1):407, 2011.

[27] Jeffrey B Endelman et Christophe Plomion. LPmerge : an R package for merging genetic maps by linear programming. *Bioinformatics*, 30(11):1623–1624, 2014.

[28] Ronald Fagin. Combining Fuzzy Information from Multiple Systems. *Journal of Computer and System Sciences*, 58(1):83 – 99, 1999.

[29] Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D Sivakumar et Erik Vee. Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, 20(3):628–648, 2006.

[30] Qiong Fang, Jianlin Feng et Wilfred Ng. Identifying differentially expressed genes via weighted rank aggregation. *In Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 1038–1043. IEEE, 2011.

[31] Mohamed Farah et Daniel Vanderpooten. An Outranking Approach for Rank Aggregation in Information Retrieval. *In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 591–598, New York, NY, USA, 2007. ACM.

[32] Guillaume Fertin, Anthony Labarre, Irena Rusu, Éric Tannier et Stéphane Vialette. *Combinatorics of Genome Rearrangements*. The MIT Press, 1st édition, 2009.

[33] Erica Fields, David Claudio, Gül E Okudan, Carol A Smith et Andris Freivalds. Triage decision making : Discrepancies in assigning the emergency severity index. *In IIE Annual Conference. Proceedings*, page 699. Institute of Industrial Engineers-Publisher, 2009.

[34] Erica B. Fields, GüL E. Okudan et Omar M. Ashour. Rank aggregation methods comparison : A case for triage prioritization. *Expert Syst. Appl.*, 40(4):1305–1311, mars 2013. ISSN 0957-4174.

[35] Edward A. Fox et Joseph A. Shaw. Combination of Multiple Searches. *NIST SPECIAL PUBLICATION SP*, 243, 1994.

[36] Susan Gauch, Guijun Wang et Mario Gomez. ProFusion* : Intelligent fusion from multiple, distributed search engines. *J. UCS*, 2(9):637–649, 1996.

[37] Nicki Gilboy, P Tanabe, DA Travers, DR Eitel et R Wuerz. The Emergency Severity Index implementation handbook : a five-level triage system. *Des Plaines, Illinois : Emergency Nurses Association*, 2003.

[38] Alain Guénoche. Single or Multiple Consensus for Linear Orders. *In Clusters, Orders, and Trees : Methods and Applications*, pages 189–199. Springer, 2014.

[39] LI Hang. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94(10):1854–1862, 2011.

[40] Taher H. Haveliwala, Aristides Gionis, Dan Klein et Piotr Indyk. Evaluation Strategies for Similarity Search on the Web. *In Proceedings of the Eleventh International World Wide Web Conference*, pages 432–442, 2002.

[41] T.K. Ho, J.J. Hull et S.N. Srihari. On Multiple Classifier Systems for Pattern Recognition. *In Pattern Recognition, 1992. Vol. II. Conference B : Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on Pattern Recognition*, pages 84–87. IEEE Comput. Soc. Press, 1992.

[42] Anthony Jameson. More Than the Sum of Its Members : Challenges for Group Recommender Systems. *In Proceedings of the working conference on Advanced visual interfaces - AVI '04*, pages 48–54, New York, New York, USA, 2004. ACM Press.

[43] Anthony Jameson, Stephan Baldes et Thomas Kleinbauer. Two Methods for Enhancing Mutual Awareness in a Group Recommender System. *In Proceedings of the working conference on Advanced visual interfaces - AVI '04*, page 447, New York, New York, USA, 2004. ACM Press.

[44] Meynaud Jean. Black (Duncan) - The theory of committees and elections. *Revue économique*, 12(4):668–668, 1961. ISSN 0035-2764.

[45] Yunlong Jiao, Anna Korba et Eric Sibony. Controlling the distance to a Kemeny consensus without computing it. *In Proceedings of The 33rd International Conference on Machine Learning*, pages 2971–2980, 2016.

[46] Paul B Kantor et Kwong Bor Ng. An investigation of the conditions for effective data fusion in information retrieval : A pilot study. *In Proceedings of the ASIS Annual Meeting*, volume 35, pages 166–78. ERIC, 1998.

[47] Claire Kenyon-Mathieu et Warren Schudy. How to rank with few errors. *In Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 95–103. ACM, 2007.

[48] Alexandre Klementiev, Dan Roth et Kevin Small. Unsupervised Rank Aggregation with Distance-based Models. *In Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 472–479, New York, NY, USA, 2008. ACM.

[49] Alexandre Klementiev, Dan Roth, Kevin Small et Ivan Titov. Unsupervised Rank Aggregation with Domain-specific Expertise. *In Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1101–1106, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[50] Steve Lawrence et C Lee Giles. Inquirus, the NECI meta search engine. *Computer networks and ISDN systems*, 30(1):95–105, 1998.

[51] Joon Ho Lee. Analyses of multiple evidence combination. *ACM SIGIR Forum*, 31:267–276, 1997.

[52] Pei Lee, Laks V.S. Lakshmanan et Jeffrey Xu Yu. On Top-k Structural Similarity Search. *In Data Engineering (ICDE), 2012 IEEE 28th International Conference on Data Engineering*, pages 774–785, 2012.

[53] Marc Lelarge. Notes de cours : Algorithmique des réseaux sociaux, 2013. ENS.

[54] Hang Li. Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies*, 7(3):1–121, 2014.

[55] Tie-Yan Liu, Jun Xu, Tao Qin, Wenying Xiong et Hang Li. Letor : Benchmark dataset for research on learning to rank for information retrieval. *In Proceedings of SIGIR 2007 workshop on learning to rank for information retrieval*, pages 3–10, 2007.

[56] Joseph F McCarthy. Pocket RestaurantFinder : A Situated Recommender System for Groups. *In Workshop on Mobile Ad-Hoc Communication at the 2002 ACM Conference on Human Factors in Computer Systems*, 2002.

[57] Mark Montague et Javed A. Aslam. Condorcet Fusion for Improved Retrieval. *In Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages 538–548, New York, NY, USA, 2002. ACM.

[58] Gonzalo Nápoles, Rafael Falcon, Zoumpoulia Dikopoulou, Elpiniki Papageorgiou, Rafael Bello et Koen Vanhoof. Weighted aggregation of partial rankings using ant colony optimization. *Neurocomputing*, 2017.

[59] Ahmet Murat Ozdemiray et Ismail Sengor Altingovde. Explicit search result diversification using score and rank aggregation methods. *Journal of the Association for Information Science and Technology*, 66(6):1212–1228, 2015.

[60] Elena M. Renda et Umberto Straccia. Web metasearch : rank vs. score based rank aggregation methods. *In SAC '03 : Proceedings of the 2003 ACM symposium on Applied computing*, pages 841–846, New York, NY, USA, 2003. ACM Press.

[61] Chandrima Sarkar, Sarah Cooley et Jaideep Srivastava. Robust feature selection technique using rank aggregation. *Applied Artificial Intelligence*, 28(3):243–257, 2014.

[62] Frans Schalekamp et Anke van Zuylen. Rank Aggregation : Together We'Re Strong. *In Proceedings of the Meeting on Algorithm Engineering & Expermiments*, pages 38–51, Philadelphia, PA, USA, 2009. Society for Industrial and Applied Mathematics.

[63] William W Cohen Robert E Schapire et Yoram Singer. Learning to order things. *Advances in Neural Information Processing Systems*, 10(451):24, 1998.

[64] Erik Selberg et Oren Etzioni. The MetaCrawler architecture for resource aggregation on the Web. *IEEE expert*, 12(1):11–14, 1997.

[65] Joseph A. Shaw et Edward A. Fox. Combination of Multiple Searches. *In Text REtrieval Conference*, pages 243–252, 1993.

[66] Haibao Tang, Xingtan Zhang, Chenyong Miao, Jisen Zhang, Ray Ming, James C Schnable, Patrick S Schnable, Eric Lyons et Jianguo Lu. ALLMAPS : robust scaffold ordering based on multiple maps. *Genome Biology*, 16(1):3, 2015.

[67] Michel Truchon *et al.*. Figure skating and the theory of social choice. *Cahier*, 9814, 1998.

[68] Merijn Van Erp et Lambert Schomaker. Variants of the Borda Count Method for Combining Ranked Classifier Hypotheses. *In Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition*, pages 443–452, 2000.

[69] Anke Van Zuylen et David P. Williamson. Deterministic Algorithms for Rank Aggregation and Other Ranking and Clustering Problems. *In Proceedings of the 5th International Conference on Approximation and Online Algorithms*, WAOA'07, pages 260–273, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 3-540-77917-5, 978-3-540-77917-9.

[70] Jason Weston, Hector Yee et Ron J Weiss. Learning to rank recommendations with the k-order statistic loss. *In Proceedings of the 7th ACM conference on Recommender systems*, pages 245–248. ACM, 2013.

[71] Shengli Wu, Jieyu Li, Xiaoqin Zeng et Yaxin Bi. Adaptive data fusion methods in information retrieval. *Journal of the Association for Information Science and Technology*, 65(10):2048–2061, 2014.

[72] Ronald R. Yager. On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decisionmaking. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, 1988.

[73] Ronald R. Yager et Vladik Kreinovich. On how to merge sorted lists coming from different web search tools. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 3(2):83–88, 1999.

[74] Bo Yang. *Bioinformatics analysis and consensus ranking for biological high throughput data.* Thèse de doctorat, Université Paris Sud-Paris XI, 2014.

[75] Immanuel V. Yap, David Schneider, Jon Kleinberg, David Matthews, Samuel Cartinhour et Susan R. McCouch. A Graph-Theoretic Approach to Comparing and Integrating Genetic, Physical and Sequence-Based Maps. *Genetics*, 165(4): 2235–2247, 2003.

[76] Chunfang Zheng, Aleksander Lenert et David Sankoff. Reversal distance for partially ordered genomes. *Bioinformatics*, 21(1):502–508, 2005.

[77] Chunfang Zheng et David Sankoff. Genome rearrangements with partially ordered chromosomes. *Journal of Combinatorial Optimization*, 11(2):133–144, 2006.