Ordre et désordre dans l'algorithmique du génome

Laurent Bulteau [1]

Accessit du prix de thèse Gilles Kahn 2013

Laurent Bulteau a soutenu sa thèse en juillet 2013 à l'université de Nantes, sous la direction de Guillaume Fertin et Irena Rusu. Il effectue actuellement un stage post-doctoral au sein de la "Technische Universität" de Berlin (Allemagne).

La génomique comparative est le domaine de la bioinformatique visant à explorer les informations présentes dans les génomes de différentes espèces afin de mieux comprendre les espèces étudiées. Les objectifs peuvent être de retracer l'évolution de certaines espèces ou gènes, de mieux comprendre les rôles et les interactions entre les gènes, et même de corriger ou compléter les informations génomiques. Le but de cette thèse est d'explorer certains problèmes de génomique comparative en utilisant des outils algorithmiques apportant des garanties de précision tout en contrôlant le temps d'exécution. Le premier de ces outils est la théorie de la complexité, qui permet de rapidement identifier les difficultés inhérentes aux différents problèmes. Des solutions algorithmiques sont ensuite recherchées. Les problèmes les plus simples admettent des algorithmes polynomiaux : rapides, et fournissant un résultat exact. Pour les cas plus difficiles – qui sont généralement les plus pertinents – on recherche des algorithmes d'approximation (rapides

1. http://www.user.tu-berlin.de/l.bulteau/index.html

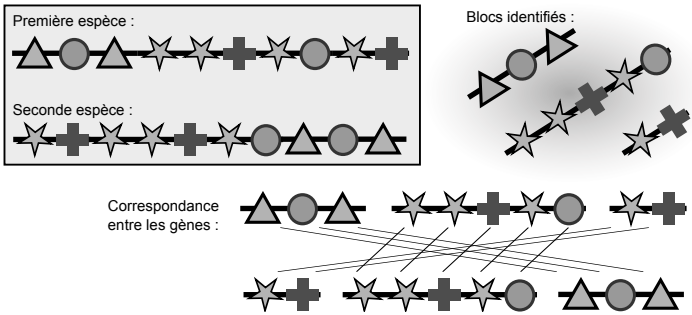

FIGURE 1. Problème MCSP : à partir des génomes de deux espèces (encadré), identifier un petit nombre de *blocs* de façon à pouvoir représenter chacun des génomes comme une concaténation de ces blocs. Cette représentation induit une correspondance entre les gènes des deux espèces.

mais inexacts), ou paramétrés : appelés *FPT*, ces derniers sont des algorithmes exacts exploitant la structure profonde des données pour éviter une explosion combinatoire.

Les questions les plus ambitieuses cherchent à recréer, étape par étape, l'histoire évolutive la plus probable expliquant les génomes contemporains. Le problème de tri par transpositions cherche à transformer un génome en un autre où chaque étape consiste à déplacer un fragment de chromosome dans la séquence. De tels problèmes nécessitent des données particulièrement purifiées, où l'on connaît une correspondance exacte entre les gènes des deux espèces. Dans la première partie de cette thèse, j'ai étudié de tels problèmes du point de vue de la théorie de la complexité, et apporté des résultats de NP-difficulté longtemps attendus.

La seconde partie de la thèse cherche à construire la correspondance entre les gènes de deux espèces, lorsque les séquences ADN ne permettent pas de différencier certaines familles de gènes. Le problème de plus petite partition commune (MCSP) cherche à découper deux séquences de gènes en un même ensemble de blocs (Figure 1) afin d'obtenir une correspondance pertinente entre les gènes de deux espèces. Malgré la complexité algorithmique du problème, il a été possible de créer deux algorithmes paramétrés exploitant notamment le fait que, en pratique, le nombre de blocs reste faible.

Enfin, la troisième partie de la thèse s'intéresse à plusieurs problèmes visant à corriger des données génomiques incorrectes ou incomplètes, en se servant d'autres génomes comme référence de comparaison. L'objectif est de nettoyer autant que possible les données afin de pouvoir, dans un second temps, appliquer des algorithmes plus poussés. À nouveau, les techniques d'algorithmique théorique permettent de maîtriser le temps de calcul et les approximations faites pour résoudre ces problèmes.

En marge de cette thèse, et à présent en post-doctorat à la *Technische Universität* de Berlin, j'étudie des problèmes sortant du cadre de la bioinformatique (allant du clustering de matrices au partitionnement de graphes) en appliquant diverses méthodes algorithmiques, et tout particulièrement la complexité paramétrée.