Stage ÉNSL 2013 – 2014

Complexité et approximation sur le problème de l'échafaudage de génomes

Mots clés : Complexité, graphes, approximation

1 Contexte

La production des séquences génomiques passe par de nombreux traitements informatiques visant à manipuler de grandes quantités de données en un temps raisonnable. Parmi ces problèmes, l'étape consistant à réarranger et orienter des morceaux de génomes entre eux, appelée *échafaudage* (scaffolding dans la littérature), passe par l'analyse d'un graphe non orienté sans boucle, valué sur les arêtes, et disposant d'un couplage parfait. Le problème que l'on étudie sur les graphes d'échafaudage, qui est en réalité une généralisation d'un problème de voyageur de commerce, est NP-complet. Pour des raisons de passage à l'échelle évident, il est important de définir des heuristiques efficaces en terme de temps de calcul, mais aussi donnant une garantie de performance sur le résultat obtenu. Le stage a pour but d'étudier, dans le cadre d'une mesure d'approximation particulière, appelée mesure différentielle, quel ratio d'approximation on peut prouver sur les heuristiques les plus naturelles du problème (notamment l'algorithme glouton) et dans des classes de graphes particulières.

2 Présentation du problème

Définition 2.1 *On appelle* cycle bicoloré *(resp.* chemin bicoloré*) dans G, relativement à un couplage parfait M^* de G, un cycle (resp. un chemin) dont les arêtes sont alternativement dans M^* ou en-dehors.*

La classe de problèmes $(\sigma_p, \sigma_c)-$Scaffold est définie comme suit :

> $(\sigma_p, \sigma_c)-$Scaffold Problem :
> **Instance** : Soit $G = (V, E)$ un graphe d'ordre pair n et M^* un couplage parfait de G. Soit $(\sigma_p, \sigma_c) \in \mathbb{N} \times \mathbb{N} \backslash \{(0, 0)\}$.
> **Question** : Existe-t-il une collection d'exactement σ_p chemins bicolorés et σ_c cycles bicolorés, disjoints, qui couvrent tous les sommets du graphe ?

La classe Min/Max$-(\sigma_p, \sigma_c)-$Scaffold Problems est définie comme suit :

> Min/Max-$(\sigma_p, \sigma_c)-$Scaffold Problem :
> **Instance** : Soit $G = (V, E)$ un graphe d'ordre pair n et M^* un couplage parfait de G. Soit $(\sigma_p, \sigma_c) \in \mathbb{N} \times \mathbb{N} \backslash \{(0, 0)\}$.
> **Question** : Trouver une collection d'exactement σ_p chemins bicolorés et σ_c cycles bicolorés, disjoints, qui couvrent tous les sommets du graphe, et de poids total minimal (resp. maximal).

L'échafaudage de génome correspond à la version maximisante de ce problème.

3 Présentation de la mesure différentielle

Lorsque l'on étudie un algorithme heuristique pour un tel problème NP-complet, on s'intéresse classiquement, pour toutes instances I du problème, au ratio entre la valeur donnée par l'algorithme approché, notée $A(I)$, et la valeur donnée par la solution exacte $opt(I)$. On dit que l'approximation est garantie lorsque ce ratio peut être borné par une même borne pour toutes les instances.

La mesure différentielle pour l'approximation suit une approche légèrement différente. Pour une instance fixée I, elle mesure la position de la valeur approchée $A(I)$ entre la meilleure valeur $opt(I)$ et la pire valeur $\omega(I)$. La valeur $A(I)$ qui se réfère à la solution de l'algorithme A sur une instance I est comprise entre $[opt(I), \omega(I)]$.

Le principe de l'approximation différentielle, est naturellement de se rapprocher au maximum de $opt(I)$ ou de s'éloigner au maximum de $\omega(I)$. Ainsi, on ne compare pas la valeur de la solution à celle de l'optimum, mais le chemin déjà parcouru depuis la pire solution à l'étendue des valeurs possibles, donnée par le diamètre $diam(I) = |opt(I) - \omega(I)|$. La qualité d'une solution approchée ρ^A_{diff} pour un algorithme A, en approximation différentielle, est donc donnée par la valeur du rapport

$$\rho^A_{diff} = \min_I \frac{|\omega(I) - A(I)|}{|\omega(I) - opt(I)|}$$

Cette mesure est intéressante et les résultats existants en approximation différentielle sont parfois à l'opposé des résultats d'approximation classique. Nous pouvons citer par exemple le problème de la coloration de sommets, un des problèmes de base dans l'optimisation combinatoire dans lequel le but est de colorier, avec le minimum de couleurs, les sommets du graphe tel que deux sommets reliés par une arête ne peuvent avoir la même couleur. Dans la théorie de l'approximation avec la mesure classique, ce problème est non-approximable c'est à dire qu'il n'existe pas d'algorithme qui garantisse un ratio à valeur constant (on peut s'éloigner tant qu'on veut de la solution optimale). En revanche, dans le cas de l'utilisation de la mesure différentielle, il existe un algorithme ayant un ratio à facteur constant (en fait il existe un algorithme 1/2-approché et un autre à facteur 2/3). Notons la stabilité de la mesure différentielle, à la différence de la mesure classique, par rapport à la mesure d'optimisation en min ou en max. Par exemple, les rapports différentiels pour le problème du stable de taille maximum et d'un transversal de taille minimum coïncident.

La mesure différentielle semble bien définie pour les problèmes d'optimisation combinatoire pour lesquels le pire des cas est facile à caractériser, ce qui semble le cas pour le problème de l'échafaudage.

4 Travail à faire

Le but de ce projet est d'étudier le problème selon la complexité classique et de l'approximation polynomiale (classique et différentielle) dans les graphes quelconques et les graphes particuliers (biparti, chordaux, ...).

Encadrement : Annie Chateau (`annie.chateau@lirmm.fr`) et Rodolphe Giroudeau (`rgirou@lirmm.fr`) sont maîtres de conférences au LIRMM (Laboratoire d'Informatique, Robotique et Microélectronique de Montpellier), respectivement dans les équipes MAB (Méthodes et Algorithmes pour la Bioinformatique) et MAORE (Méthodes et Algorithmes pour l'Ordonnancement et les Réseaux).