Assemblage de génomes à l'aide des réseaux de fonctions de coûts

Thématique : Bioinformatique, Problème de Satisfaction de Contraintes, Optimisation combinatoire
Équipe d'accueil : Statistique et Algorithmique pour la Biologie
Laboratoire d'accueil : Biométrie et Intelligence Artificielle, Institut National de la Recherche Agronomique
Lieu : Auzeville-Tolosane (près de Toulouse), France
Encadrant : Simon de Givry (degivry@toulouse.inra.fr Tel : 05 61 28 50 74)
Gratification : environ 400 euros / mois

Contexte

Le problème étudié qui survient lors de l'assemblage d'un génome est de trouver un ordre total d'un ensemble de *contigs*, i.e. des grands fragments de séquence génomique, ainsi que leur orientation, de telle sorte à minimiser la violation de contraintes de distance minimum entre extrémités de contigs (Gao et al, 2011) (Donmez and Brudno, 2013). Ce problème peut s'exprimer comme un problème d'optimisation quadratique sous contraintes avec des variables discrètes représentant la position et l'orientation de chaque contig. Une contrainte globale de permutation sur les variables de position impose de trouver un ordre total.

La minimisation d'une somme de fonctions de coûts sur des variables discrètes est dans le cas général un problème NP-difficile abordé par les communautés Intelligence Artificielle et Recherche Opérationnelle. Le problème d'assemblage considéré ici a été prouvé comme étant lui-aussi un problème NP-difficile. Ce problème peut se modéliser dans différents formalismes existants. Parmi les plus connus, on peut citer la logique propositionnelle traitant de formules insatisfiables (Max-SAT), les problèmes de satisfaction de contraintes avec prise en compte de préférences (Max-CSP), et aussi la programmation linéaire en nombres entiers (PLNE).

L'équipe d'accueil mène des travaux sur l'optimisation dans les réseaux de fonctions de coûts, *Weighted Constraint Satisfaction Problem* (WCSP) (Cooper et al, 2010), une variante de Max-CSP, et met en oeuvre leur intégration dans une plate-forme logicielle *open-source* C++ toulbar2 (`https://mulcyber.toulouse.inra.fr/projects/toulbar2/`) ayant remporté plusieurs compétitions (*UAI 2008, 2010, and 2011 Challenges* `http://www.cs.huji.ac.il/project/UAI10/` et `http://www.cs.huji.ac.il/project/PASCAL/board.php` *ficolofo*). toulbar2 a récemment été intégré dans une autre plate-forme en python, numberjack (`https://github.com/eomahony/Numberjack/tree/fzn/`), qui dispose d'interfaces vers divers outils d'optimisation issus des formalismes (Max-)CSP, WCSP, Max-SAT et PLNE.

Parallèlement à ces travaux, le laboratoire d'accueil hébèrge la plate-forme bioinformatique de la Génopole Toulouse Midi-Pyrénées (GENOTOUL `http://bioinfo.genotoul.fr/`) qui offre un support matériel (cluster 2000 CPUs) et informatique pour accompagner des projets en bioinformatique, principalement autour du traitement des données issues du séquencage génomique haut-débit.

Sujet

L'objectif du stage est d'étudier différentes modélisations en WCSP du problème d'assemblage en s'inspirant des travaux utilisant la programmation dynamique (Gao et al, 2011) et la PLNE (Donmez and Brudno, 2013). A partir des résultats de cette étape de modélisation, plusieurs pistes algorithmiques en vue d'améliorer les performances seront envisagées, en lien avec les travaux de recherche de l'équipe.
Des expérimentations seront menées sur des données simulées et des données réelles en partenariat avec la plateforme bioinformatique de Toulouse.
Un lien avec le problème du voyageur de commerce est également envisagé.

Bibliographie

Cooper, M.C., de Givry, S., Sanchez, M., Schiex, T., Zytnicki, M., and Werner, T.,
Soft Arc-consistency revisited, Artificial Intelligence, 2010

Song Gao, Wing-Kin Sung, and Niranjan Nagarajan,
Opera: Reconstructing Optimal Genomic Scaffolds with High-Throughput Paired-End Sequences,
Journal of Computational Biology, 18(11): 1681-1691, 2011

Nilgun Donmez and Michael Brudno,
SCARPA: scaffolding reads with practical algorithms, Bioinformatics, 29(4):428-434, 2013