

Une modélisation par Contrainte de graphe pour résoudre l'échafaudage de génome.

Éric Bourreau¹, Annie Chateau^{1,2}, Clément Dallard³, Rodolphe Giroudeau¹

¹ LIRMM - CNRS UMR 5506 - Montpellier, France

² IBC - Montpellier, France

³ University of Portsmouth, UK

{eric.bourreau,annie.chateau,rodolphe.giroudeau}@lirmm.fr
clement.dallard@port.ac.uk

Depuis le séquençage complet du génome humain, les dernières années ont été marquées par la production de plus en plus rapide de données de séquençage (projet HTS *High Throughput Sequencing* ou NGS *Next Generation Sequencing*). Les techniques de production de séquences complètes des organismes nouvellement séquencés se décomposent en plusieurs parties, dont l'échafaudage de génome [1]. Après le séquençage et l'assemblage, nous nous retrouvons avec un ensemble de contigs (assemblage de *paired-end reads*), ainsi qu'un ensemble d'informations, pas forcément cohérent, sur les relations entre ces contigs. Trouver un ensemble cohérent de telles relations parmi celles qui sont disponibles permet la reconstruction pas à pas des chromosomes cibles.

Ce problème, démontré NP-difficile [2], possède de nombreuses heuristiques [3] mais peu d'approches complètes existent malheureusement [4]. Nous proposons lors de cette présentation de résoudre celui-ci par la Programmation Par Contraintes avec le solveur Choco [5], et plus spécifiquement par l'utilisation judicieuse de variables de graphes [6]. Des expérimentations sont réalisées à la fois sur des données réelles (voir figure 1) ou artificielles [7] afin de valider la performance de l'outil proposé et des différentes stratégies de résolution possibles.

Bibliographie :

- [1] Daniel H. Huson, Knut Reinert, and Eugene W. Myers. The greedy path-merging algorithm for contig scaffolding. *Journal of the ACM (JACM)*, 49(5):603–615, 2002.
- [2] Annie Chateau and Rodolphe Giroudeau. A complexity and approximation framework for the maximization scaffolding problem. *Theor. Comput. Sci.*, 595:92–106, 2015.
- [3] Song Gao, Wing-Kin Sung, and Niranjan Nagarajan. Opera: reconstructing optima genomic scaffolds with high-throughput paired-end sequences. *Journal of Computational Biology*, 18(11):1681–1691, 2011.
- [4] Nicolas Briot, Annie Chateau, Rémi Coletta, Simon De Givry, Philippe Leleux, and Thomas Schiex. An Integer Linear Programming Approach for Genome Scaffolding. In *Workshop Constraints in Bioinformatics*, 2014.
- [5] Jean-Guillaume Fages, Narendra Jussien, and Xavier Lorca and Charles Prud'homme. Choco3: an open source java constraint programming library, 2013.
- [6] Grégoire Dooms, Yves Deville, and Pierre Dupont. CP (graph): Introducing a graph computation domain in constraint programming. In *Principles and Practice of Constraint Programming-CP 2005*, pages 211–225. Springer, 2005.
- [7] Annie Chateau Adel Ferdjoukh, Éric Bourreau and Clémentine Nebut. A ModelDriven Approach to Generate Relevant and Realistic Datasets. In *Software Engineering and Knowledge Engineering (SEKE)*, Jul 2016.

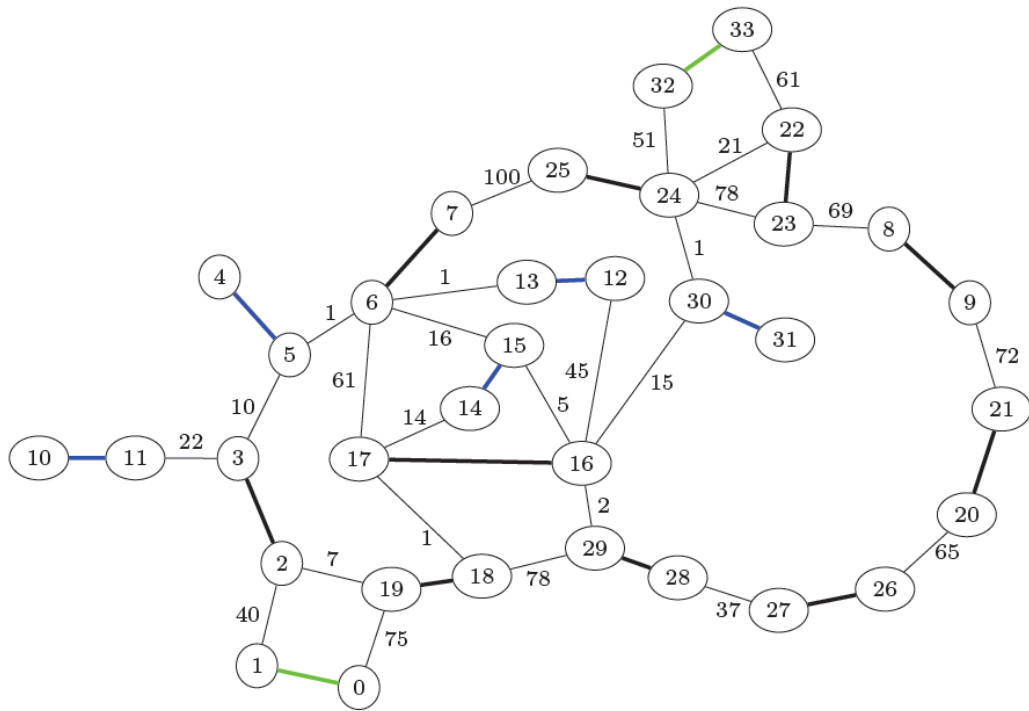


Figure 1 : Graphe d'échafaudage avec 17 contigs (en gras) et 26 liens potentiels pondérés les connectant, correspondant au jeu de données issu d'EBOLA.