Production de graphes d'échafaudage à partir d'une structure d'indexation des $k-$mers pairés

Alban Mancheron – Annie Chateau – 2014

Mots-clés bioinformatique, algorithmique du texte, graphes.

Description La production d'une séquence génomique passe par une première étape, appelée séquençage, au cours de laquelle on récupère plusieurs millions (voire milliards) de séquences de taille réduite (une centaine de caractères) issues du génome. Ces séquences sont appelées des *reads*. Elles constituent les pièces d'un gigantesque puzzle qu'il s'agit par la suite de reconstituer grâce à des algorithmes d'assemblage. Certains de ces algorithmes, par exemple [8], utilisent le découpage des reads en mots de taille fixe k, appelés k-mers. Ces k-mers sont ensuite stockés dans une structure de données particulière qui permet de retrouver facilement leurs chevauchements. Les algorithmes d'assemblage reconstruisent de façon incomplète ces génomes, et on récupère en sortie des séquences plus longues, de longueur variables, appelées *contigs* (voir figure 1).

Figure 1: Les contigs (en-dessous) sont reconstruits en utilisant les chevauchements entre les reads (au-dessus).

Afin d'étudier les propriétes de ces contigs, on réalise souvent une opération qui consiste à aligner les k-mers des reads sur les séquences des contigs. Cela permet notamment de regarder dans quelle mesure une position particulière d'un contig est couverte par ces k-mers. En scriptant les sorties de l'outil d'alignement CRAC, on obtient ce qu'on appelle des profils de couverture des contigs par les k-mers des reads, ce sont des sortes de courbes irrégulières comme dans la figure 2. Les pics de couverture peuvent indiquer une répétition de la zone du génome concerné, qui est indétectable par les logiciels d'assemblage actuels.

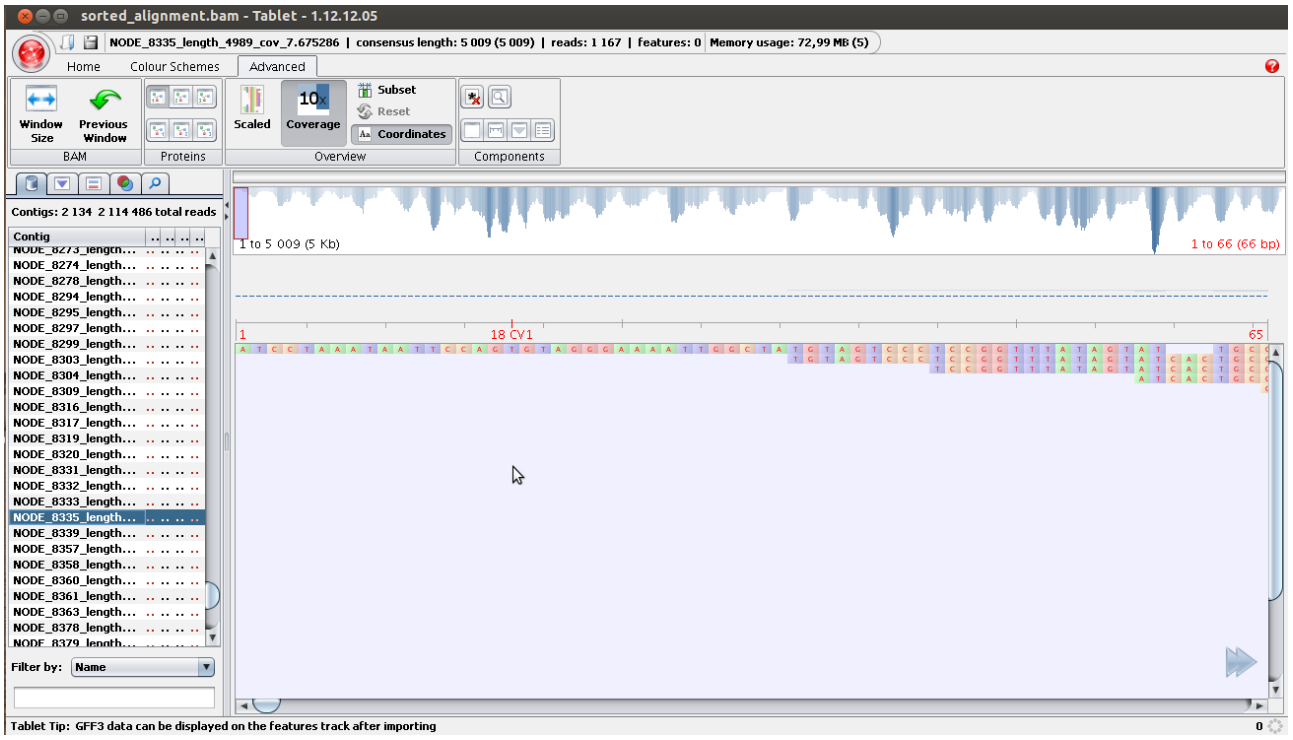

Figure 2: Profil de couverture d'un contig de la Peste Noire.

Le sujet a pour cadre le traitement de données à grande échelle en vue de la production *de novo* de génomes ou de transcriptomes de qualité, en tenant compte des zones répétées. À cette fin, on s'intéresse à une structure de données d'indexation de k-mers, les Gk-arrays [7], pour stocker l'information portée par les lectures appariées issues du séquençage et on cherche à l'adapter et l'améliorer pour l'assemblage [8] et l'échafaudage [5, 3, 1, 4, 2] de génomes avec répétitions. Dans un premier temps, on étudiera comment utiliser l'outil d'alignement CRAC [6], utilisant les Gk-arrays, afin de déterminer des profils de couverture permettant de détecter les répétitions dans les contigs d'un assemblage déjà produit. Puis on considèrera la possibilité de définir, directement à partir de la structure d'indexation, le graphe d'échafaudage. L'accent sera mis sur l'efficacité algorithmique à toutes les étapes, la parallélisabilité des traitements ainsi que des structures de données, et le contrôle de la qualité des échafaudages produits.

Le sujet proposé peut s'orienter vers deux directions complémentaires (ou les deux si le temps le permet) :

- Détection des profils de couverture des contigs via les données d'alignement fournies par CRAC et implémentation d'une méthode efficace d'échafaudage à partir de ces données de multiplicités.

- Amélioration de la structure de données de Gk-arrays, et étude de la construction du graphe d'échafaudage directement à partir de cette structure de données.

Les tests pourront être menés sur des jeux de données génomiques ou transcriptomiques, dont les profils sont mieux marqués. L'objectif final est de mettre au point un outil intégrant les aspects indexation, assemblage et scaffolding, qui soit efficace en terme de temps de calcul, bon en qualité de génome ou transcriptome produit, et parallélisable le plus longtemps possible dans la chaîne des traitements.

Contexte du stage Les encadrants sont membres de l'équipe MAB du Lirmm et de l'IBC, dans l'Axe 1, dans le cadre d'un projet Jeunes Chercheurs IBC. Le stage se déroulera au sein de l'équipe MAB du Lirmm.

References

[1] A. Dayarian, T. P. Michael, and A. M. Sengupta. SOPRA: scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics*, 11:345, 2010.

[2] N. Donmez and M. Brudno. SCARPA: scaffolding reads with practical algorithms. *Bioinformatics*, 29(4):428–34, 2013.

[3] S. Gao, W. Sung, and N. Nagarajan. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J Comput Biol*, 18(11):1681–91, 2011.

[4] A. A. Gritsenko, J. F. Nijkamp, M. J.T. Reinders, and D. de Ridder. GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. *Bioinformatics*, 2012.

[5] D. H. Huson, K. Reinert, and E. W. Myers. The greedy path-merging algorithm for contig scaffolding. *J. ACM*, 49(5):603–615, September 2002.

[6] Nicolas Philippe, Mikaël Salson, Thérèse Commes, and Eric Rivals. CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biology*, 14(3):R30, March 2013.

[7] Nicolas Philippe, Mikaël Salson, Thierry Lecroq, Martine Léonard, Thérèse Commes, and Eric Rivals. Querying large read collections in main memory: a versatile data structure. *BMC Bioinformatics*, 12(1):242, June 2011.

[8] Kamil Salikhov, Gustavo Sacomoto, and Gregory Kucherov. Using cascading bloom filters to improve the memory usage for de brujin graphs. In *WABI*, page to appear, 2013.

Prérequis Intérêt pour la bioinformatique, Programmation

Contacts Alban Mancheron (alban.mancheron@lirmm.fr), Annie Chateau (annie.chateau@lirmm.fr).