

Thèse financée par l'université Lille 1 et la Région Nord-Pas-de-Calais

Indexation de réarrangements lymphoïdes VDJ pour le suivi de la leucémie

Encadrants: Jean-Stéphane Varré, Mathieu Giraud, Mikaël Salson

Contrat doctoral 2014 – 2017

Candidatures (CV, lettre de motivation) à adresser
à mikael.salson@lifl.fr avant le ~~30 avril~~ 5 mai 2014.

1 Contexte

La **leucémie aigue lymphoblastique** (ALL, Acute Lymphoblastic Leukemia) est un cancer liquide touchant principalement les enfants. Le taux de survie des patients affectés par une ALL a nettement augmenté dans les dernières décades, grâce à des diagnostics plus précis et aussi une meilleure stratégie thérapeutique. Le diagnostic et le suivi de ces leucémies demande d'analyser les lymphoblastes et les lymphocytes, en dénombrant ces cellules selon leur réarrangement V(D)J, réarrangement de l'ADN spécifique à ces cellules. Il existe des milliards de combinaisons différentes obtenues à partir d'un répertoire de quelques milliers de gènes.

L'équipe Bonsai (LIFL et Inria Lille) travaille sur l'analyse des répertoires lymphocytaires pour le suivi de la leucémie depuis deux ans avec le département d'hématologie du CHR de Lille et la plateforme de séquençage de Lille 2 hébergée à l'IRCL (Institut pour la Recherche sur le Cancer de Lille). Notre nouvelle approche s'appuie sur les progrès colossaux réalisés dans le domaine du séquençage de l'ADN via les séquenceurs à haut débit, capable de séquencer des millions de séquences en quelques heures. L'équipe a développé de nouveaux algorithmes pour analyser rapidement les **réarrangements V(D)J** : une heuristique à base de « k-mots » identifie les séquences correspondant à des lymphoblastes clonaux et permet d'estimer leur abondance relative. Les réarrangements V(D)J sont un réarrangement de l'ADN qui se produit sur quelques dizaines de nucléotides. Une différence de quelques nucléotides peut aussi bien correspondre à deux lymphocytes complètement différents ou à des erreurs de séquençage n'ayant qu'une importance marginale.

En dehors du champ des réarrangements V(D)J, de nombreux outils pour l'analyse de séquençage à haut débit reposent sur des **structures d'indexation compressées**, telles que la transformée de Burrows-Wheeler (BWT). Ces structures permettent de retrouver très rapidement une information spécifique provenant de grandes quantités de données ainsi indexées. Nous estimons que ces structures seraient profitables dans le cadre des réarrangements V(D)J.

2 Sujet de thèse

Le but de cette thèse est de proposer une structure d'indexation pour des séquences immunologiques, et de l'appliquer à la population clonale complète de patients atteints de leucémie. Concrètement, le travail de la thèse sera de :

- **réaliser une bibliographie** sur les structures d'indexation avancées (BWT, graphes de Bruijn), et algorithmes pour l'analyse de réarrangements V(D)J
- **concevoir et tester une structure d'indexation de « profils V(D)J », et l'appliquer à des problèmes d'analyse de reads.** La structure sera adaptée au stockage de millions de courtes séquences (telle que celles produites par les séquenceurs à haut débit) et à la recherche de données issues de plusieurs points de suivi, dans l'objectif d'apporter des connaissances biologiques. L'index améliorera la *détection de clones*, permettra d'identifier les *évolutions clonales* et aidera à la *reconstruction des séquences germinales* du patient. Il n'existe pas à l'heure actuelle de structure adaptée à la spécificité du problème des réarrangements V(D)J, en prenant en compte les différentes portions d'intérêt de ces séquences.
- **intégrer ces méthodes au logiciel Vidjil** (<http://bioinfo.lifl.fr/vidjil>), logiciel d'analyse de réarrangements V(D)J développé depuis 2012 avec le CHR de Lille. Les résultats seront analysés et diffusés en collaboration avec nos partenaires français (hôpitaux impliqués dans le suivi de la leucémie) et européens (consortium EuroClonality-NGS).

Le doctorant devra avoir de solides compétences en informatique théorique, plus particulièrement en **algorithmique du texte**. Le candidat devra également avoir un goût certain pour la biologie, la thèse demandant d'assimiler certaines notions en immunologie et en hématologie.

Références

- [1] Logan, A. C. et al, High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *PNAS* 108(52), 21194–21199 (2011)
- [2] Gawad, C. et al, Massive evolution of the immunoglobulin heavy chain locus in children with B precursor acute lymphoblastic leukemia. *Blood* (2012)
- [3] Navarro, G., Mäkinen, V., Compressed full-text indexes. *ACM Computing Surveys*, 39(1), 2 (2007)