Québec, Canada – le 14 avril 2010

À l'Institut des maladies infectieuses et immunitaires (IMII):

Vous trouverez ci-dessous mon résumé pour la Journée de la science du 5 mai 2010.

ASSEMBLAGE PARALLÈLE DE GÉNOMES AVEC DES SÉQUENCES TRÈS COURTES EN PAIRES

MOTS CLÉS: GÉNOMIQUE, BIOINFORMATIQUE, ALGORITHME

Boisvert, Sébastien[1]; Laviolette, François[3,4] ; Marchand, Mario[3]; Corbeil, Jacques[1,2];

[1] Centre de recherche en infectiologie, Centre hospitalier universitaire de Québec (CHUQ), Pavillon CHUL, 2705 boul. Laurier, Québec (Québec) G1V 4G2, Canada
[2] Faculté de médecine, Pavillon Ferdinand-Vandry, 1050, ave de la Médecine, bureau 4633, Université Laval, Québec (Québec) G1V 0A6, Canada
[3] Département d'informatique et de génie logiciel, Pavillon Adrien-Pouliot, 1065, av. de la Médecine, Université Laval, Québec (Québec) G1V 0A6, Canada,
[4] Department of Computer Science, UCL (University College London), Malet Place, London WC1E 6BT, UK

OBJECTIF : Les organismes vivants encodent leur fonctionnement dans un génome. Un génome bactérien a une longueur de quelques centaines de milliers de nucléotides à quelques millions. Par exemple, *Mycoplasma agalactiae* PG2, un pathogène chez les petits ruminants, a un génome de 877438 nucléotides, alors que *Pseudomonas aeruginosa* PAO1, un agent pathogène en santé humaine, a un génome de 6264404 nucléotides. Le séquençage permet de guider la découverte de cibles thérapeutiques: les protéines encodées dans les génomes. Plusieurs nouvelles technologies de séquençage (454, Illumina, SOLiD) permettent d'obtenir des millions de paires de séquences digitales, lesquelles correspondent aux extrémités de fragments d'ADN provenant d'un génome. L'analyse bioinformatique de ces données est cependant un défi nécessitant les algorithmes adéquats et les structures de données appropriées.

MÉTHODES : Nous avons développé Ray (http://denovoassembler.sf.net/), un logiciel qui calcule en parallèle la séquence d'un génome à partir de millions de paires de séquences digitales obtenues par un séquenceur de nouvelle génération comme le Illumina Genome Analyzer. Nous avons testé Ray sur plusieurs ensembles de données, et avons utilisé le superordinateur CLUMEQ de l'Université Laval.

RÉSULTATS : Pour la bactérie modèle *Escherichia coli* K-12 MG1655, nous avons obtenu un génome réparti en 148 séquences couvrant le génome à 98.1% avec aucune séquence erronée à partir de 21.9 millions de paires de séquences (fragments de 215 +/- 20 nucléotides) de 36 nucléotides (technologie Illumina). Le temps d'exécution est de moins de 30 minutes à l'aide de 28 processeurs AMD Opteron.

CONCLUSION : L'obtention de séquences digitales avec la technologie de séquençage Illumina et l'analyse subséquente avec l'assembleur parallèle Ray permet de décoder un génome bactérien très rapidement et très efficacement. Ray est présentement un des assembleurs les plus performants en distribution libre.

(300 mots)