Journée de la science du 5 mai 2010 organisée par l'Institut des maladies infectieuses et immunitaires (IMII) des Instituts de recherche en santé du Canada (IRSC) (largeur: 48 pouces, hauteur: 36 pouces)

Assemblage parallèle de génomes avec des séquences très courtes en paires

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

Introduction

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 numériques, 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 obtenues avec un séquenceur de nouvelle génération comme le Illumina Genome Analyzer. Les nouvelles technologies de séquençage permettent d'obtenir des millions de paires de séquences digitales (voir la Figure 1).

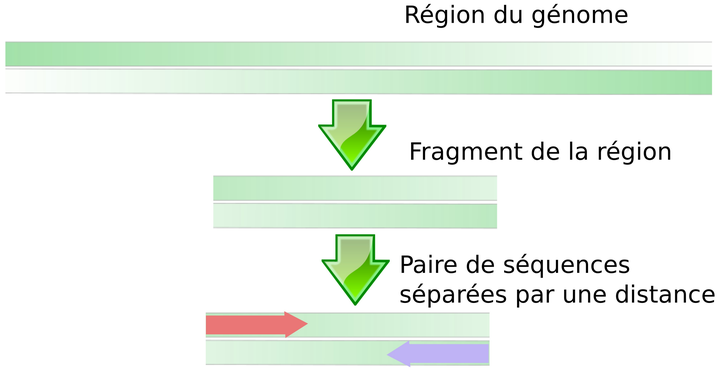

Figure 1: Les séquences en paires sont issues de fragments.

Avec Ray, chaque séquence est transformée en une structure discrète telle qu'illustrée à la Figure 2. Lorsque deux sous-séquences sous liées, la première sous-séquence sans sa première lettre et la deuxième sous-séquence sans sa dernière lettre sont identiques.

En transformant toutes les séquences en structures discrètes, et en les regroupant, il est possible de reconstruire la séquence du génome. Ainsi, chaque sous-séquence sera observée un nombre particulier de fois dans les séquences obtenues par séquençage à très haut débit.

Puisque les données obtenues par séquençage à très haut débit contiennent des erreurs, nous allons analyser la distribution du nombre d'observations des sous-séquences ainsi générées.

Ray utilise le passage de messages (Open-MPI), est codé en C++, et est testé avec GNU/Linux. Les ordinateurs sont numérotés de 0 à n-1, et chacun contient 4096 arbres "Splay" pour accueillir des sous-séquences. La communication peut se faire par mémoire partagée, par TCP/IP, par Infiniband, ou simplement par copie de mémoire ("self").

Ray isole les sous-séquences avec beaucoup d'observations, et construit les graines -- lesquelles permettent de démarrer l'assemblage. Ces graines sont étendues en utilisant les distances entre les séquences en paires, et le génome est aisément reconstruit de cette façon.

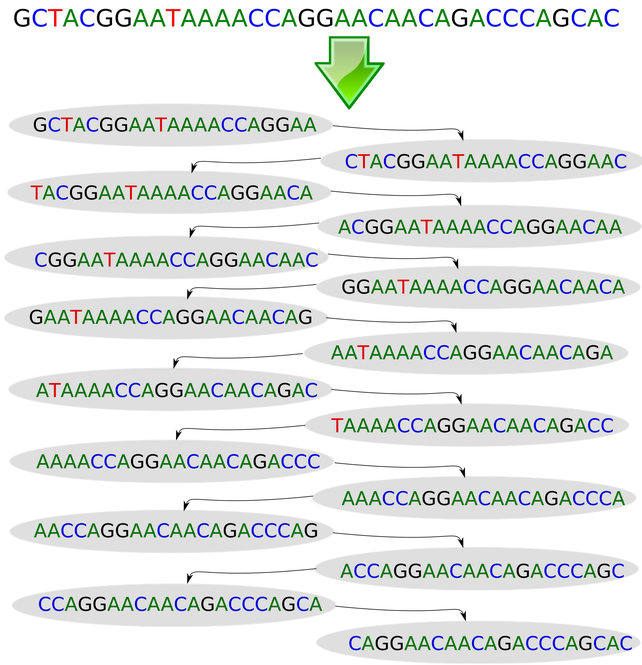

Figure 2: Transformation d'une séquence numérique en structure discrète.

Résultats

Pour la bactérie modèle Escherichia coli K-12 MG1655 (génome: NC_000913), nous avons utilisé 21.9 millions de paires de séquences de 36 nucléotides générées avec le Illumina Genome Analyzer (séquences: SRA001125) pour assembler son génome. La distribution du nombre d'observations des sous-séquences est affichée dans la Figure 3: le minimum local est à 33 observations, et le maximum local est à 109 observations, ce qui veut dire, qu'en moyenne, une position quelconque du génome est couverte par 109 séquences.

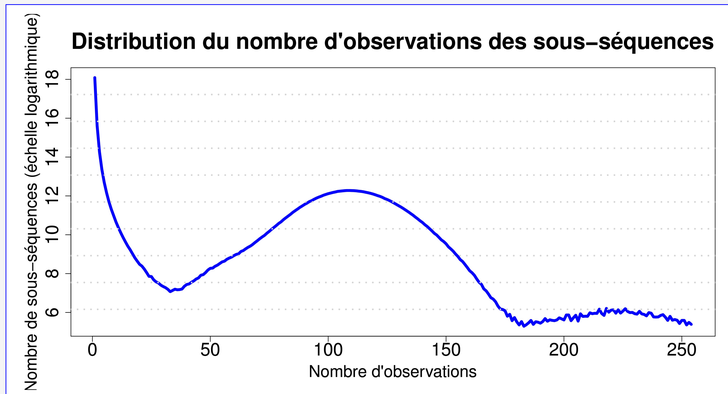

Figure 3: Distribution du nombre d'observations des sous-séquençages. Le logarithme naturel est utilisé.

Les paires de séquences, pour ce jeu de données, sont groupées en deux librairies -- la distribution de la longueur des fragments est illustrée à la Figure 4. La librairie 1 a une moyenne de 215 (écart type: 11) et la librairie 2 a une moyenne de 486 (écart type: 26).

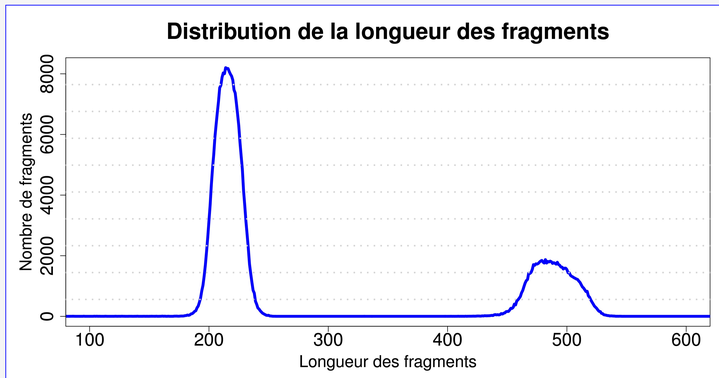

Figure 4: Distribution de la longueur des fragments des deux librairies.

Utiliser Ray est facile: pour assembler ces données, la commande suivante est exécutée:

```
mpirun -np 25 Ray -p 1_1.fastq 1_2.fastq -p 2_1.fastq 2_2.fastq -o Ecoli.fasta
```

Ray calcule la taille moyenne des fragments de chaque librairie, et génère un fichier appellé Ecoli.fasta -- lequel contient la reconstruction du génome. En utilisant Ray, nous avons obtenu un génome réparti en 126 séquences couvrant le génome à 98.17% avec aucune séquence erronée. Nous avons comparé Ray avec les meilleurs assembleurs disponibles pour les séquences courtes en paires (voir Tableau 1).

Tableau 1: Comparaisons des assembleurs pour séquences courtes.

Assembleur	Séquences	Nucléotides	Longueur moyenne	N50	Longueur maximale	Couverture du génome	Séquences erronées	Changements de nucléotide	Insertions et délétions	Temps d'exécution
EULER-SR	1761	4099614	2328	3429	19094	0.8708	13	1173	7355	143m27.647s
Velvet	83	4542631	54730	125611	311586	0.9657	28	456	895	31m51.199s
Ray	**126**	**4590771**	**36434**	**72499**	**174569**	**0.9817**	**0**	**2**	**4**	**47m18.992s**
ABySS	154	4661190	30267	56703	174288	0.9840	0	233	9	44m53.714s

Conclusion

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

Remerciements

Sébastien Boisvert est un étudiant boursier des Instituts de Recherche en Santé du Canada (IRSC) (doctorat: 200910GSD-226209-172830 et maitrise: 200902CGM-204212-172830). Jacques Corbeil est titulaire de la Chaire de recherche du Canada en génomique médicale. Nous remercions la Fondation canadienne pour l'innovation pour le financement de l'infrastructure (ls30.genome.ulaval.ca) et le CLUMEQ et Calcul Canada pour l'accès à colosse.clumeq.ca. Nous remercions également la communauté des logiciels libres GNU, Linux, et Open-MPI.

Références

Bentley et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature. http://dx.doi.org/doi:10.1038/nature07517

Fan et al. (2006) Highly parallel genomic assays. Nature Reviews Genetics. http://dx.doi.org/doi:10.1038/nrg1901

Flicek and Birney (2009) Sense from sequence reads: methods for alignment and assembly. Nature Methods. http://dx.doi.org/doi:10.1038/nmeth.1376

Pevzner et al. (2001) An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences. http://dx.doi.org/doi:10.1073/pnas.171285098

IRSC CIHR
Instituts de recherche en santé du Canada
Canadian Institutes of Health Research

Canada Research Chairs
Chaires de recherche du Canada

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

Canada
CHUQ
CENTRE HOSPITALIER UNIVERSITAIRE DE QUÉBEC
CRCHUQ

UNIVERSITÉ LAVAL

CLUMEQ

compute · calcul
CANADA