STOCKAGE DE DONNÉES NGS PROVENANT DE GÉNOMES HAUTEMENT SIMILAIRES

ENCADRANTS

- Raluca Uricaru (Maître de conférences, LaBRI) - raluca.uricaru@labri.fr
- Guillaume Blin (Professeur d'Université, LaBRI) - guillaume.blin@labri.fr

MOTS-CLÉS

Bioinformatique, Next Generation Sequencing, Compression

DESCRIPTION

Le séquençage de l'ADN, c'est à dire la détermination de la succession des nucléotides le composant, est aujourd'hui une technique de routine pour les laboratoires de biologie. En effet, les progrès tout à fait spectaculaires dans le domaine du séquençage au cours des dernières années, avec l'apparition des techniques de séquençages dites de nouvelle génération (NGS), ont permis la publication de nombreuses séquences brutes des génomes provenant d'espèces variées. Les mêmes technologies ont rendu possibles des projets de grande envergure, pour lesquels les génomes de centaines à des milliers d'individus d'une même espèce sont séquencés, dans le but de fournir des informations sur la diversité génétique des populations entières (*e.g.*, 1000 Genomes Project, the UK10K Project, the Personal Genome Project ou l'étude systématique des génomes du cancer portée par The International Cancer Genome Consortium). Par conséquent, le stockage et l'analyse des données génomiques provenant d'une telle démarche de reséquençage est devenu un sujet de recherche très actif au cours des dernières années.

Étant donné que deux organismes d'une même espèce sont extrêmement similaires (*e.g.*, les génomes de deux personnes sont identiques à 99.5%), il est évident que, pour être efficace, le stockage des données de ce type devrait exploiter leur fort taux de redondance. En raison de cette redondance, stocker seules les différences par rapport à une séquence de référence s'est révélée être une solution hautement adaptée qui a engendré de nombreux outils, comme SlimGene (ref. 1) ou MZip (ref. 2). Lorsque l'on n'a pas accès à une telle référence, des solutions qui encodent les différences par rapport à des séquences faisant partie du jeu de données à compresser sont en général utilisées, *e.g.*, ReCoil (ref. 3).

Dans le cadre de ce projet on vise une nouvelle approche destinée à représenter et à analyser simultanément les données NGS de reséquençage (séquençage des nombreux individus provenant d'une même espèce), sans utiliser une séquence de référence. En effet, la grande limitation dans le stockage dépendant d'une référence est l'impossibilité d'utiliser les données, an l'absence de cette référence. Une première idée qui peut être envisagée serait de stocker les données dans un graphe de De Bruijn coloré, i.e. chaque individu étant représenté par une couleur. Afin d'évaluer cette nouvelle approche du point de vue de son efficacité (en fonction de la taille des jeu de données stockés), ainsi que de la facilité d'analyse de ces données, on devra procéder à une évaluation comparative avec les méthodes existantes.

De manière connexe à la question principale, plusieurs problématiques pourraient être envisagées, comme :

- l'analyse des polymorphismes (SNPs, indels) entre les individus séquencés
- le cas particulier des données des génomes tumoraux provenant de plusieurs ou d'un même patient malade de cancer

LE CANDIDAT IDÉAL

- possédera des solides compétences et savoirs-faire en algorithmique, ceci incluant de préférence les domaines des structures de données et de la théorie des graphes. Des compétences supplémentaires dans les domaines suivants représenteraient un plus : programmation, optimisation combinatoire, bioinformatique. Une motivation pour travailler en contexte interdisciplinaire est nécessaire.
- aura obtenu un master (de préférence en informatique) ou un diplôme d'ingénieur et présentera une forte motivation pour les activités de recherche.
- aura des compétences solides en C/C++. Des langages de script comme bash, perl, python seront également utilisés.
- sera capable de travailler en équipe.

Un intérêt pour le domaine de la biologie ou le domaine de la recherche biomédicale est nécessaire. Des compétences déjà acquises en génétique, génomique et biologie ne sont pas requises, bien qu'elles seront un plus.

Un candidat titulaire d'un master de bioinformatique, et possédant des compétences en algorithmique et en programmation approfondies, doublées d'une motivation pour la conception, l'implémentation et le test de méthodes algorithmiques avancées, pourrait convenir. Une expérience significative en programmation est requise, ainsi qu'une rigueur dans les développements applicatifs.

Références :

1. Compressing genomic sequence fragments using SlimGene, Kozanitis C. *et al.*, J.Comput.Biol. 2011
2. Efficient storage of high throughput DNA sequencing data using reference-based compression, Hsi-Yang Fritz M. *et al.*, Genome Res. 2011
3. ReCoil - an algorithm for compression of extremely large datasets of dna data, Yanovsky V., AMB 2011