COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Praveen Rao (University of Missouri-Columbia), October 16, 2020

Title: Démocratiser l'analyse de la séquence génomique pour le COVID-19 à l'aide de CloudLab

Praveen Rao CIC Database Profile

NSF Award #: 2034247

YouTube Recording with Slides

October 2020 CIC Webinar Information

Transcript Editor: Cora Cole

Transcript

Slide 1

Bonjour à tous, il est encore 11 h 50, et je vais vous expliquer comment je vais m'y prendre pour démocratiser l'analyse de la séquence du génome pour COVID-19, en utilisant CloudLab, un banc d'essai expérimental financé par la NSF pour l'informatique en nuage. Je travaille à l'université du Missouri-Columbia.

Slide 2

La motivation de notre travail est assez simple : il est de plus en plus intéressant de comprendre comment le génome d'un individu influe sur les symptômes de la maladie COVID-19, sur la gravité de la maladie et sur l'issue finale - survie ou non à la maladie. En analysant le génome des patients atteints de COVID-19, nous pouvons donc améliorer notre compréhension de la maladie, ce qui peut nous permettre d'élaborer de nouvelles stratégies de traitement et d'accélérer la découverte de médicaments. Un certain nombre de publications sont à venir, dont l'une, parue dans le New England Journal of Medicine, porte sur une étude d'association à l'échelle du génome - il s'agissait d'utiliser environ 1 900 patients et d'étudier leurs variants génétiques dans les génomes. Une autre initiative a été lancée, le COVID Human Genetic Effort, un consortium international dont l'objectif est d'identifier l'impact du génome d'un individu sur sa réponse au COVID-19. Nos génomes peuvent donc détenir les réponses à la lutte contre le COVID-19, et il s'agit d'un domaine important sur lequel il faut se concentrer.

Slide 3

Les objectifs de notre projet sont essentiellement doubles : le premier est de permettre aux chercheurs d'effectuer des analyses de variants à grande échelle sur les séquences du génome humain, et l'objectif est de leur fournir ces ressources gratuitement. L'analyse des variantes détecte essentiellement les variations dans le génome d'un individu - par exemple les polymorphismes d'un seul nucléotide ou les petites insertions et suppressions, ainsi que, nous pouvons même penser à des variantes structurelles telles que les variations du nombre de copies. L'autre partie de la recherche sera axée sur le développement d'un assemblage de novo efficace des génomes humains afin de pouvoir effectuer une analyse plus approfondie des variantes des génomes des individus, l'un appartenant à un groupe qui n'a pas été touché par la maladie, et l'autre appartenant au groupe qui a été touché par la maladie - et dans ce contexte particulier, nous nous intéressons à COVID-19.

Slide 4

Pour atteindre nos objectifs, nous allons donc développer une infrastructure logicielle à l'aide de CloudLab. CloudLab existe depuis plusieurs années, il a été conçu à l'origine pour la recherche sur les systèmes informatiques et n'a pas vraiment été prévu pour les charges de travail intensives en données, mais dans le cadre de cet effort particulier, nous allons montrer comment nous pouvons tirer parti de CloudLab et contourner certaines de ses limites pour construire une infrastructure capable de prendre en charge l'analyse génomique à grande échelle à l'aide de technologies de calcul en grappes, ainsi que d'outils open source, nous allons donc examiner les meilleures pratiques qui existent pour les pipelines génomiques. L'une d'entre elles est GATK, nous allons également examiner le projet BD Genomics, nous allons utiliser Apache Spark pour atteindre le parallélisme, et nous allons également utiliser certains des outils open source qui sont largement utilisés dans la communauté de la génomique. La deuxième partie consistera à mettre au point un algorithme efficace qui nous aidera à effectuer ce que nous appelons une " analyse exhaustive des variantes " à l'aide d'un assemblage *de novo*.

Slide 5

Il s'agit essentiellement de deux groupes de patients et, à l'aide de la modélisation des graphes bipartites, nous allons examiner la comparaison par paire entre ces individus et procéder à une analyse plus approfondie des variantes dans leurs génomes, ce qui nous aidera à mieux comprendre la maladie. Sur le côté droit, vous voyez essentiellement l'ensemble de l'écosystème que nous sommes en train de mettre en place - en tirant parti de tout ce qui est disponible en termes de logiciels libres et en construisant nos propres composants (tels que le moteur d'analyse exhaustive des variants). L'objectif est qu'en fin de compte, les chercheurs n'aient pas à s'inquiéter de devoir payer des coûts élevés pour des ressources informatiques en nuage, que ce soit par l'intermédiaire de vendeurs commerciaux ou d'autres moyens. CloudLab est donc une plateforme universitaire gratuite et nous aimerions l'exploiter pour donner aux chercheurs la possibilité d'effectuer des analyses génomiques à grande échelle dans le but de trouver un remède au COVID-19. Nous aimerions également comprendre l'impact des charges de travail génomiques sur les systèmes informatiques et de réseau. Comment pouvons-nous construire de futurs systèmes mieux adaptés au traitement des charges de travail génomiques à grande échelle ?

Slide 6

Voici un site de projet - nous l'hébergeons activement sur Github - et nous sommes en mesure de permettre aux utilisateurs de s'inscrire sur CloudLab et d'effectuer des analyses de variants sur une charge unique, ainsi que sur une grappe. Nous avons accès à deux ressources publiques : le Thousand Genomes Project, qui n'est évidemment pas lié à COVID-19, mais qui nous fournit beaucoup de données pour tester notre logiciel, et le portail de données COVID-19, où certains projets répertorient les séquences disponibles pour que nous puissions travailler avec elles. Nous présentons également une évaluation de la performance de nos efforts initiaux - à quelle vitesse pouvons-nous effectuer l'analyse de variants sur ces séquences en mode groupé - ainsi que l'analyse *de novo*.

Slide 7

Veuillez donc suivre ce lien [https://github.com/MU-Data-Science/EVA] si vous souhaitez effectuer gratuitement une analyse du génome à grande échelle. Voici une interface utilisateur simple que nous sommes en train de mettre au point pour que vous puissiez fournir l'accès ou les adresses URL de vos fichiers, puis vous n'avez qu'à dire " exécuter " et donner votre adresse électronique.

Slide 8

Voici notre équipe, qui comprend un ensemble diversifié de chercheurs allant de la pathologie à la génomique, en passant par la bioinformatique et l'épidémiologie. Mon doctorant Arun Zachariah participe activement à la construction du logiciel avec moi, alors n'hésitez pas à nous contacter si vous avez des questions ou si vous êtes intéressé par l'utilisation de notre plateforme.