Centro de Información de COVID (CIC): Charlas científicas de relámpago

Transcripción de una presentación de Praveen Rao (Universidad de Missouri-Columbia), 16 de octubre de 2020

Título: Democratizando el análisis de secuencias genómicas para COVID-19 usando CloudLab

Perfil de Praveen Rao en la base de datos del CIC

Subvención de La Fundación Nacional de Ciencias (NSF, por sus siglas en inglés) #: 2034247

Grabación de YouTube con diapositivas

Información del seminario web del CIC de octubre 2020

Editora de la Transcripción: Cora Cole

Editora de la Translación: Isabella Graham Martínez

Transcripción

Praveen Rao:
Diapositiva 1
Buenas tardes a muchos de ustedes, todavía son las 11:50 [a.m.] aquí, y voy a hablar de cómo voy a trabajar en la democratización del análisis de secuencias genómicas para COVID-19, usando CloudLab, que es un banco de pruebas experimental financiado por la NSF para la computación en nube. Estoy en la Universidad de Missouri-Columbia.

Diapositiva 2
Y así la motivación para nuestro trabajo es bastante sencilla: hay un creciente interés en entender cómo el genoma de un individuo realmente impacta los síntomas que una persona ve debido a COVID-19, así como la gravedad de la enfermedad, así como el resultado final - si sobreviven a la enfermedad o no. Por lo tanto, al hacer un análisis genómico del genoma de los pacientes con COVID-19, podemos mejorar nuestra comprensión de la enfermedad, y esto puede permitirnos tener nuevas estrategias de tratamiento y un descubrimiento de medicamentos más rápido. Hay un número de publicaciones que están viniendo para arriba, y uno de ellos está en el New England Journal of Medicine que hizo un estudio de la asociación genoma-ancha - y éste era sobre el uso de cerca de 1.900 pacientes y el estudio de sus variantes genéticas en los genomas. Otro esfuerzo que ha existido es el Esfuerzo Genético Humano COVID, y es un consorcio internacional y el objetivo es básicamente identificar cómo el genoma de un individuo impacta su respuesta a COVID-19. Así que nuestros genomas pueden contener las respuestas para luchar contra COVID-19, y este es un área importante en la que centrarse.

Diapositiva 3
Así que los objetivos de nuestro proyecto son básicamente, ya sabes, dos: el primero es permitir a los investigadores realizar análisis variantes a escala en secuencias del genoma humano, y el objetivo es darles los recursos sin cargo. Así que el análisis de variantes esencialmente detecta variaciones en el genoma del individuo - por ejemplo, polimorfismos de nucleótidos individuales o pequeñas inserciones y elimina, así como, incluso podemos pensar en variantes estructurales como las variaciones del número de copia. Ahora, la otra parte de la investigación se va a centrar en el desarrollo de un eficiente *ensamblaje de novo* de los genomas humanos para que podamos hacer un análisis más profundo de las variantes de los genomas de los individuos, ya sabes, uno que pertenece a un grupo que no fueron afectados por la enfermedad, y el otro perteneciente al grupo que se vieron afectados por la enfermedad - y en este contexto particular, estamos viendo COVID-19.

Diapositiva 4
Así que para lograr nuestros objetivos, lo que vamos a hacer es desarrollar una infraestructura de software utilizando CloudLab. Y CloudLab ha existido durante varios años, fue diseñado originalmente para la investigación de sistemas informáticos y no fue realmente planeado para las cargas de trabajo de datos intensivos, pero en este esfuerzo en particular vamos a mostrar cómo podemos aprovechar CloudLab y tener soluciones alternativas en torno a algunas de las limitaciones que tiene para construir una infraestructura que pueda soportar el análisis genómico a gran escala utilizando tecnologías de computación de clúster, así como herramientas de código abierto, así que vamos a estar mirando las mejores prácticas que hay por ahí para las tuberías genómicas. Uno de ellos es el GATK, también vamos a ver el proyecto BD Genomics, vamos a utilizar Apache Spark para lograr el paralelismo, y también vamos a utilizar algunas de las herramientas de código abierto que se utilizan ampliamente en la comunidad de genómica. Y la segunda parte sería desarrollar un algoritmo eficiente que nos va a ayudar a realizar lo que llamamos un análisis de variantes exhaustivo usando *ensamblaje de novo*.

Diapositiva 5
Así que esencialmente estamos hablando de dos grupos de pacientes aquí, y usando el modelado gráfico bipartito, estaremos mirando la comparación por parejas entre estos individuos y tendremos un análisis más profundo de las variantes en sus genomas que nos ayudarán a entender mejor la enfermedad. Y en el lado derecho lo que se ve es esencialmente todo el ecosistema que estamos poniendo juntos - aprovechando lo que está disponible en términos de software de código abierto, y la construcción de nuestros propios componentes (como el motor de análisis de variantes exhaustiva). Y el objetivo es, al final del día, los investigadores no deben preocuparse por tener que pagar altos costos de recursos de computación en nube, ya sea a través de proveedores comerciales u otros, ya sabes, lugares. Así que CloudLab es una plataforma académica gratuita y nos gustaría aprovechar eso para empoderar básicamente a los investigadores con la capacidad de hacer análisis genómicos a gran escala en un esfuerzo por encontrar una cura para COVID-19. También nos gustaría entender: ¿cómo afectan las cargas de trabajo genómicas a la computadora y la red sistemas, ya sabes? ¿Cómo podemos construir sistemas futuros que estén mejor orientados hacia el procesamiento de cargas de trabajo genómicas a escala?

Diapositiva 6
Ahora aquí hay un sitio de proyecto - lo tenemos alojado activamente en Github - y podemos permitir a los usuarios registrarse en CloudLab y hacer análisis de variantes en una sola carga, así como en un clúster. También pueden hacer *ensamblaje de novo* sobre secuencias, tenemos acceso a dos recursos disponibles públicamente: uno es el Proyecto Mil Genomas, que por supuesto no está relacionado con COVID-19 pero nos da muchos datos para probar nuestro software, y luego tenemos acceso a la

COVID-19 portal de datos donde algunos de los algunos de los proyectos enumeran secuencias que están disponibles para nosotros, ya sabes trabajar con, también informamos alguna evaluación de rendimiento en nuestros esfuerzos iniciales - qué tan rápido podemos hacer realmente el análisis de variantes en estas secuencias en un modo de clúster - y así como *análisis de novo*.

Diapositiva 7
Así que por favor, siga este enlace [https://github.com/MU-Data-Science/EVA] si está interesado en hacer análisis de genoma a gran escala sin costo alguno, y aquí hay una interfaz de usuario sencilla que estamos construyendo para que pueda proporcionar acceso, o proporcionar las URL de sus archivos, y luego puede decir ejecutar y dar su ID de correo electrónico y luego le enviaremos de vuelta el archivo de análisis de variantes una vez que se complete el proceso.

Diapositiva 8
Aquí está nuestro equipo, que incluye un conjunto diverso de investigadores que van desde la patología a la genómica a la bioinformática a la epidemiología, y mi Ph.D. estudiante Arun Zachariah está activamente involucrado en la construcción del software junto conmigo, así que no dude en ponerse en contacto con nosotros si tiene alguna pregunta o interés en usar nuestra plataforma, y muchas gracias por su atención.