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

Transcripción de una presentación de Alexander Niema Moshiri (Universidad de California San Diego), 13 de noviembre de 2020

Título: RAPID: Inferencia filogenética en tiempo real y análisis de clúster de transmisión de COVID-19

Alexander N Moshiri CIC Profile

Subvención de La Fundación Nacional de Ciencias (NSF, por sus siglas en inglés) #: 2028040

Grabación de YouTube con diapositivas

Información del seminario web del CIC de noviembre 2020

Editora de la Transcripción: Macy Moujabber

Editora de la Translación: Isabella Graham Martínez

Transcripción:

Alexander Moshiri:

Diapositiva 1

Cool. Gracias. Espero que la gente vea mis diapositivas. So yeah I'm Niema Moshiri. I'm at UC San Diego. Estoy en el Departamento de Informática e Ingeniería y voy a estar hablando de un desarrollo reciente que hemos hecho llamado MSA viral.

Diapositiva 2

Básicamente un flujo de trabajo filogenético estándar, si alguna vez has visto una historia evolutiva de secuencias de virus, empiezas con estas secuencias iniciales al principio que no necesariamente se alinean muy bien. Y el primer paso es alinear varias secuencias para que se alineen. Alinearlos así nos dice cosas sobre, ya sabes, cuál es la homología de secuencia, cuáles son las relaciones entre las secuencias.

Usando esta alineación, se estima una filogenia. Y luego arraigas la filogenia para determinar cuál es el ancestro común. Y luego hay un montón de otros análisis aguas abajo que podrían estar interesados en hacer, ya sabes, cosas como cuáles son los grupos de brotes que están sucediendo ¿sabes qué demografía creemos que están más en riesgo de infectarse por alguna enfermedad? Sí, muchas cosas que puedes hacer. Así que en general mi enfoque está en estos dos pasos: la alineación de secuencias

múltiples y la inferencia filogenética que son realmente los cuellos de botella computacionales. Y en esta charla, me centraré solo en este paso: la alineación de secuencias múltiples.

Diapositiva 3

Así que el problema de alineación de secuencias múltiples es NP-Complete, lo que, TLDR, significa que no tenemos una solución de tiempo polinomial, y como resultado, muchas heurísticas se han desarrollado para aproximar soluciones. Son bastante precisos razonablemente. Algunas de las herramientas con las que la gente puede estar familiarizada son MUSCLE, ClustalOmega y MAFFT. Estas son algunas de las herramientas comunes que las implementan. Sin embargo, incluso estas heurísticas generalmente escalan cuadráticamente con respecto al número de secuencias. Así que en el caso del SARS-CoV-2, hemos tenido un crecimiento exponencial de secuenciación sucediendo donde, quiero decir, esto es una gran cosa para nosotros. Estamos obteniendo más y más datos de secuencias, pero la desventaja es que tenemos que analizar esos datos de secuencias y nuestras herramientas no se escalan correctamente. Así que ahora mismo, esto es obsoleto. Esto era de hace una semana o dos. Ahora estamos casi en 200.000 secuencias. Así que las herramientas simplemente no escalan para permitir el análisis en tiempo real.

Diapositiva 4

Entonces, ¿qué pasaría si, en cambio, supiéramos de antemano que nuestras secuencias van a ser súper similares y ya tenemos un genoma de referencia representativo de alta confianza contra el cual podemos compararlas? Así que aquí, estoy mostrando el genoma de referencia como una línea verde en la parte superior y cada una de esas otras secuencias de colores por debajo de ella son la secuencia que quiero alinear entre sí. En lugar de alinearlos entre ellos, lo que puedo hacer es alinear la primera secuencia contra el genoma de referencia. Alinee la segunda secuencia con el genoma de referencia. Siga alineando cada una de las secuencias de forma independiente directamente con el genoma de referencia. Y luego usando las posiciones del genoma de referencia como anclas, puedo fusionar las alineaciones de pares individuales en una línea de secuencia múltiple. Así que tome la primera columna de mi asignación de secuencia múltiple ver- en la primera secuencia esta es la posición que coincidía, luego esta posición coincidió en la segunda, y la tercera, y luego simplemente colapsarlos juntos. Puedo hacer esto para cada posición del genoma de referencia y ahora tengo una línea de secuencia múltiple. Y lo bueno es que este paralelismo es estupendo. Cada secuencia puede alinearse con la referencia de forma independiente y simultánea, y se escala linealmente con el número de secuencias en lugar de cuadráticamente.

Diapositiva 5

Pero permítanme dar un paso atrás muy rápido y pensar en este enfoque. Mi entrada es un genoma de referencia y un montón de secuencias que son muy similares a ese genoma de referencia, y mi salida es una alineación de cada secuencia contra el genoma de referencia. Así que si la gente está familiarizada con la secuencia de lectura larga, este es exactamente el mismo problema computacional que mapear lecturas largas a un genoma de referencia. Así que mi pregunta era: ¿puedo aprovechar las herramientas

de asignación de lectura bien implementadas existentes para habilitar este tipo de alineación de secuencias múltiples guiadas por referencia escalable?

Diapositiva 6

Así que desarrollé una herramienta llamada ViralMSA que envuelve los mapeadores de lectura existentes para realizar la alineación de secuencias múltiples guiadas por referencia. Me envuelvo alrededor de algunos de ellos, pero hay una herramienta específicamente, Minimap2, que es una especie de estándar de oro en este momento para lo que estoy haciendo. Así que solo recomiendo usar mi herramienta con ese mapeador de lectura específico, pero envuelvo algunos de ellos para demostrar que puedo evolucionar esta herramienta naturalmente a medida que las tecnologías de mapeo de lectura evolucionan también. Así que básicamente, simplemente proporcionas el genoma de referencia viral MSA y las secuencias para alinearse y luego se encargará de indexar el genoma de referencia y hacer cualquier pre-procesamiento que necesite hacer, y llamará al mapeador de lectura y luego fusionará los resultados en una alineación de secuencias múltiples.

Diapositiva 7

Y aquí puede ver una comparación de este enfoque con las herramientas de mejores prácticas existentes. Así que mi herramienta es la línea azul en la parte inferior y luego las otras dos líneas son otras herramientas y se ve que, en general, es múltiples órdenes de magnitud más rápido que lo que la gente generalmente está haciendo en este momento, y se escala muy bien. Y las alineaciones de secuencia que obtenemos son muy precisas.

Diapositiva 8

En conclusión, esta herramienta que he desarrollado permite una rápida alineación de secuencias múltiples de genomas virales. Es de código abierto. Se puede obtener en línea, y con suerte si usted hace cualquier análisis viral considerar su uso.

Diapositiva 9

Así que algunos reconocimientos: Heng Li desarrollado Minimap2, que es una especie de la esencia de la velocidad y este trabajo fue apoyado por NSF y Google. Y sí, dejaré preguntas para la charla o más tarde en la sesión.