Centro de Información de COVID (CIC): Charlas Cientificas Relámpago

Transcripción de una presentación de Niema Moshiri (University of California, San Diego), April 24, 2023

Título: Alineación de secuencias múltiples de genomas virales guiada por referencia escalable masivamente

Niema Moshiri CIC Perfil

NSF Award #: 2028040

Grabación de YouTube con diapositivas

Primavera 2023 CIC Webinar Información

Transcript Editor: Karem Coca and Lylybell Teran

Transcripción

Diapositiva 1

Niema Moshiri:

Impresionante, sí, gracias por la introducción. Con suerte, la gente puede ver mi pantalla. Sí, hola a todos. Como se mencionó, mi nombre es Niema Moshiri. Soy Profesor Asistente en el Departamento de Ciencias de la Computación e Ingeniería en UC San Diego. Mi charla se centrará en algunos métodos que mi laboratorio desarrolló usando los fondos de la NSF para acelerar el análisis genómico viral. Específicamente, la charla de hoy se va a centrar en cómo hemos habilitado el alineamiento de secuencias múltiples guiado por referencias escalables de genomas virales completos. De hecho, hemos hecho muchas otras aceleraciones, así que no tuve tiempo de hablar de hoy. Terminaré con un enlace a mi sitio web si la gente tiene curiosidad acerca de cómo otra cosa se podría acelerar otros aspectos de este tipo de análisis.

Diapositiva 2

Así que empecemos. Solo para dar un poco de contexto - aquí hay un marco para un flujo de trabajo de filogenia viral estándar. Y, sabes, antes de hablar de esto, la filogenia viral es muy importante para poder estudiar cómo el virus está mutando con el tiempo. ¿Cómo es, una especie de ramificación? ¿Cómo están relacionadas las diferentes muestras que recolectamos en todo el mundo? Hay un montón de usos en el mundo de la epidemiología molecular viral que están fuera del alcance de esta charla, pero generalmente, tener una filogenia inferida a partir de genomas virales es muy útil. Normalmente el flujo de trabajo comienza así cuando se empieza con un

montón de secuencias virales sin alinear, que estoy mostrando aquí. El primer paso es por lo general la alineación de secuencias múltiples donde se trata de tipo de colocar estas lagunas en el tipo diferente de posiciones de cada una de las secuencias para conseguir que se alinean mejor. Esto te da una noción de homología de secuencias después de hacer esto. Luego, dada la alineación de secuencias múltiples, podemos realizar inferencia filogenética para tratar de inferir una relación evolutiva no enraizada entre estas secuencias. Luego, típicamente después de eso, hacemos lo que se llama enraizamiento para determinar cuál es el ancestro común más probable de todas las secuencias. Ese tipo de entonces nos dice cuál fue el avance en la historia evolutiva del tiempo de estas secuencias. Luego, tal vez haga algunos análisis adicionales. Tal vez haga un clúster de transmisión. Hay muchos otros análisis que puedes hacer sobre la filogenia y las secuencias. Pero este es el tipo de bloques de construcción de cómo hacer todos estos otros análisis. Así que típicamente estos pasos aquí son los cuellos de botella computacionales clave. La alineación de secuencias múltiples y luego la inferencia filogenética. En la charla de hoy, no voy a hablar de inferencia filogenética voy a estar acercándome a la alineación de secuencias múltiples.

Diapositiva 3

Así que, algún contexto - alineación de secuencias múltiples esto es lo que se llama un problema computacional NP-Complete. Lo que eso significa - hay un término muy técnico de informática - pero básicamente lo que esto significa es que no hay una solución exacta de tiempo polinomial. Básicamente me dio un montón de secuencias y me pidió que se me ocurriera la alineación óptima de secuencias múltiples. No hay manera de hacer esto en tiempo polinomial. Es muy muy lento. Heurística se han desarrollado para proporcionar optim- para proporcionar soluciones aproximadas. Por ejemplo, es posible que haya oído hablar de ClustalOmega, MUSCLE y MAFFT. Estas son algunas herramientas estándar que se utilizan en el espacio. Sin embargo, incluso estas heurísticas - generalmente escalan cuadráticamente con respecto al número de secuencias. Para el contexto, la base de datos GISAID, que es la base de datos donde la mayoría de la gente está almacenando sus genomas completos de SARS-CoV-2, esta base de datos está creciendo extremadamente rápidamente y a partir de hoy tenemos más de 15 millones de secuencias de SARS-CoV-2 disponibles de todo el mundo. La próxima epidemia será secuenciar genomas en tiempo real. Esta va a ser una herramienta que esperamos seguir utilizando en las epidemias virales por venir. Podemos esperar que esto sea aún más significativo de un problema de big data. Actualmente con estas herramientas como ClustalOmega, MAFFT y MUSCLE, estamos viendo tiempos de ejecución de décadas a siglos, que, ya sabes, por razones obvias, si estamos tratando de hacer análisis molecular en tiempo real, décadas o siglos es un poco demasiado lento. Entonces, ¿cómo podemos acelerar esto? Bueno, resulta que el problema es en realidad un poco más fácil de lo que estamos tratando de resolver. La alineación de secuencias múltiples, en general, es como asumir que no hay homología de las secuencias en absoluto. Este es el tiempo que lleva alinear secuencias completamente arbitrarias. Pero el SARS-CoV-2 y con los virus en general tenemos un problema mucho más simple, ¿verdad? Tenemos mucha homología de secuencias. Incluso si el virus está mutando, ya sabes, significativamente en todo

el mundo, cada secuencia viral que obtenemos va a ser casi idéntica al gen de referencia. No va a ser exactamente idéntico, pero va a ser casi idéntico. Así que en realidad nos enfrentamos a un problema computacional mucho más simple que es la alineación de secuencias múltiples de secuencias muy similares. Entonces, ¿cómo podemos usar esa función para acelerar este análisis?

Diapositiva 4

Podemos hacer lo que se llama un enfoque de alineación a referencia. Así que en lugar de tratar de alinear todo con el otro a la vez, lo que podríamos hacer es alineaciones individuales en cuanto a pares contra una unidad de referencia. Así que en esta figura, la barra verde más gruesa en la parte superior representa la referencia a nuestro genoma de ámbito 2, y cada uno de estos otros genomas de color representa una secuencia que recojo del mundo real. Quiero alinear cada uno de estos con el genoma de referencia. Lo que podría hacer es uno por uno puedo alinear independientemente cada una de estas secuencias del genoma contra el genoma de referencia, que podría hacer cada uno de estos bastante rápido y puedo hacer una paralelización masiva porque cada una de estas alineaciones de pares a la referencia se puede hacer de forma completamente independiente. Puedo paralelizar sin embargo muchos núcleos mi computadora tiene, puedo lanzar que muchos en este problema. Luego, una vez que haya competido con todas esas líneas de emparejamiento a la referencia, podría usar el genoma de la llave - puedo usar sus anclajes, sus posiciones como anclajes para crear las columnas de mi línea de secuencia múltiple. Por ejemplo, tal vez empezaré con la primera posición del genoma de referencia y veré, bueno, en la secuencia roja esta es la letra que se alinea con esa posición. En la secuencia naranja, esta es la letra. En la secuencia rosa, esta es la letra. En la secuencia azul, esta es la letra. Y puedo combinar todas esas letras en una columna de mi alineación de secuencia múltiple. Y podría hacer lo mismo para la segunda posición de mi genoma de referencia, lo mismo para la tercera posición, cuarta posición, todo el camino. Y tipo de posición por posición puedo construir mi alineación de secuencia múltiple. Esta idea, esto es realmente bueno porque es masivamente paralelizable y se escala linealmente con el número de secuencias en lugar de cuadrática. Así que también tiene una escalabilidad mucho mejor. ¿Tenemos que implementar este enfoque desde cero? En realidad no.

Diapositiva 5

Resulta que si uno da un paso atrás y piensa en este problema, esto es realmente equivalente, en cierto sentido, al problema del mapeo de lectura larga. Demos un paso atrás y volvamos a pensar cuál es el problema que estamos abordando. Nuestra entrada es un genoma de referencia y un montón de secuencias largas que son muy similares al genoma de referencia. Nuestra salida es una alineación de cada una de esas secuencias contra el genoma de referencia. Este es exactamente el mismo problema computacional que la asignación de lecturas largas. En lugar de tener que reinventar la rueda, podríamos simplemente construir a partir de todas estas técnicas realmente avanzadas que la gente ha construido para resolver el problema del mapeo de lectura larga y simplemente aplicarlo a este contexto.

Diapositiva 6

Con ese objetivo, desarrollé una herramienta llamada ViralMSA y lo que hace es simplemente envolver los mapeadores de lectura largos existentes para realizar esta alineación de secuencias múltiples guiada por referencia. Trata cada uno de esos genomas que he recopilado como lecturas largas y trata el genoma de referencia como un genoma de referencia. Solo llama a ese mapeador de lectura - Me envuelvo contra algunos mapeadores de lectura diferentes solo para demostrar flexibilidad - pero principalmente sugiero que la gente use Minimap2 para la velocidad y la precisión. Luego, dados esos resultados de mapeo leídos, puedo entonces - o dados los resultados de mapeo - puedo simplemente compilarlos en una sola línea de secuencia múltiple.

Así que todo lo que tienes que hacer para ejecutar ViralMSA es darle a ViralMSA un genoma de referencia y un montón de secuencias para alinear. Se encargará automáticamente de indexar el genoma de referencia, tal vez si le das un número accesorio se encargará de descargar e indexar el genoma de referencia. Se encargará de todo el pre-procesamiento y todas las cosas de downstream y solo saldrá - llamará al mapeador de lectura, fusionará los resultados en la alineación SQL múltiple, y solo generará un único archivo estándar que es tu alineación de secuencias múltiples.

Diapositiva 7

¿Cómo funciona frente a las herramientas existentes? Hicimos un experimento de referencia en el que comparamos el tiempo de ejecución de ViralMSA alrededor de Minimap 2 en comparación con Virulign, que es un enfoque de alineación existente con referencia, pero que solo implementa su propia desde cero alineado con referencia. También comparamos con MAFFT, que se considera típicamente una de las herramientas de segmento múltiple más utilizadas. En este gráfico, en el eje horizontal tengo el número de secuencias. En el eje vertical tengo el tiempo total de ejecución en segundos. Esto se hizo en secuencias completas del genoma del SARS-CoV-2, por lo que la longitud del genoma es de aproximadamente 29.000. Como podemos ver, la línea azul, que es ViralMSA, es órdenes de magnitud más rápido que las herramientas existentes. En comparación con VIRULIGN, que también está escalando linealmente, estamos obteniendo - y por cierto, esta parcela es una parcela de escala de registro - así que en comparación con VIRULIGN, somos aproximadamente mil veces más rápidos. Y con MAFFT, no somos tan rápidos, pero se puede ver que debido a que MAFFT escala cuadrática, nuestra velocidad con respecto a MAFFT está aumentando a medida que avanza el tiempo. Incluso en solo mil secuencias. golpeamos aproximadamente mil veces más rápido y esa brecha aumenta.

Diapositiva 8

Ahora, pueden estar preguntándose, bien, bueno, rápido es bueno pero ¿cuál es el punto si no me da buenas alineaciones? También comparamos la precisión. Lo que hicimos fue tomar la alineación de secuencias múltiples calculada por MAFFT, tomar la alineación de secuencias múltiples calculada por ViralMSA en un montón de alineaciones curadas a mano del VIH, Ébola, y estoy en blanco en el tercer virus, pero básicamente tomamos virus que habíamos curado alineaciones de Los álamo- oh en realidad, no - esta trama es solo de HIV-1. Desde el Laboratorio Nacional de Los álamos tomamos sus alineaciones de secuencia múltiple y las usamos como verdad de tierra. Luego vimos cómo se mapea en ViralMSA comparar contra la tierra verdad comisariada alineación de secuencias múltiples. Si calculamos las distancias de emparejamiento de las secuencias que obtenemos en nuestra alineación, y luego hacemos una prueba de manto para la precisión - encontramos la correlación entre nuestras distancias de emparejamiento, y el emparejamiento se calcula directamente desde la verdadera alineación de secuencias múltiples, vemos que la correlación es insignificantemente diferente. Aquí, ya sabes, obtenemos un coeficiente de correlación como 0.994 para el MSA viral en comparación con 0.997 para el cálculo de distancia por pares. En realidad, cuando calculamos las filogenias, las filogenias inferidas utilizando los elementos de secuencia múltiple ViralMSA son en realidad ligeramente más alta precisión topológica que los estimados a partir de MAFFT. Así que, insignificantemente, pero aún así lo que estamos mostrando es que estos son esencialmente equivalentes en términos de precisión para todos los efectos.

Diapositiva 9

La conclusión - ViralMSA es una herramienta que permite la rápida alineación de secuencias múltiples de ultra grandes conjuntos de datos virales. Es de código abierto, puedes encontrarlo en GitHub, y ya sabes, por favor considera usarlo en tus análisis virales.

Diapositiva 10

Y agradecimientos - Quiero agradecer a Heng Li, él es el desarrollador de Minimap2 y es realmente su experiencia en el desarrollo de Minimap2 que permite la velocidad y el rendimiento de ViralMSA. Quiero agradecer a la NSF por la subvención que apoya este proyecto. Y la investigación también fue apoyada mediante créditos de investigación de la plataforma Google Cloud.

Diapositiva 11

Así que voy a ahorrar tiempo para cualquier pregunta o estoy feliz de terminar aquí.