COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Niema Moshiri (University of California, San Diego), April 24, 2023

Title: Alignement de séquences multiples guidé par référence et massivement évolutif des génomes viraux

Niema Moshiri CIC Database Profile

NSF Award #: 2028040

YouTube Recording with Slides

Spring 2023 CIC Webinar Information

Transcript Editor: Lauren Close

Transcript

Slide 1

Génial, oui, merci pour cette présentation. J'espère que les gens peuvent voir mon écran. Oui, bonjour à tous. Comme je l'ai dit, je m'appelle Niema Moshiri. Je suis professeur adjoint au département d'informatique et d'ingénierie de l'université de San Diego. Mon exposé portera sur certaines méthodes que mon laboratoire a développées grâce au financement de la NSF pour accélérer l'analyse génomique virale. Plus précisément, l'exposé d'aujourd'hui se concentrera sur la manière dont nous avons permis un alignement multiple de séquences guidées par référence massivement extensible de génomes viraux complets. Nous avons également réalisé de nombreuses autres accélérations dont je n'ai pas eu le temps de parler aujourd'hui. Je terminerai par un lien vers mon site Web si les gens sont curieux de savoir comment accélérer d'autres aspects de ce type d'analyse.

Slide 2

Commençons. Pour situer un peu le contexte, voici le cadre d'un flux de travail standard en phylogénétique virale. Vous savez, avant même que je ne parle de ce sujet, la phylogénétique virale est très importante pour étudier la façon dont le virus mute au fil du temps. Comment se ramifie-t-il ? Comment les différents échantillons que nous collectons dans le monde entier sont-ils liés ? L'épidémiologie moléculaire virale a une multitude d'utilisations qui sortent du cadre de cet exposé, mais il est généralement très utile de disposer d'une phylogénie déduite des génomes viraux. En règle générale, le flux de travail commence comme suit : on part d'un

tas de séquences de génomes viraux non alignées, que je montre ici. La première étape consiste généralement en un alignement de séquences multiples où l'on essaie de placer ces lacunes dans les différentes positions de chacune des séquences afin de les aligner au mieux. Cela permet d'obtenir une certaine notion de l'homologie des séquences. Ensuite, compte tenu de l'alignement de séquences multiples, nous pouvons effectuer une inférence phylogénétique pour tenter de déduire une relation évolutive non enracinée entre ces séquences. Ensuite, nous procédons généralement à ce que l'on appelle l'enracinement pour déterminer quel est l'ancêtre commun le plus probable de toutes les séquences. Cela nous indique en quelque sorte l'histoire de l'évolution de ces séquences dans le temps. Ensuite, vous ferez peut-être d'autres analyses en aval. On peut par exemple procéder à des regroupements de transmissions. Il y a beaucoup d'autres analyses que l'on peut faire sur la phylogénie et sur les séquences. Mais il s'agit en quelque sorte des éléments de base qui permettent d'effectuer toutes ces autres analyses. En règle générale, ces étapes sont les principaux goulets d'étranglement en matière de calcul. L'alignement des séquences multiples et l'inférence phylogénétique. Dans l'exposé d'aujourd'hui, je ne parlerai pas de l'inférence phylogénétique, je me contenterai de faire un zoom sur l'alignement de séquences multiples.

Slide 3

Un peu de contexte - l'alignement de séquences multiples est ce qu'on appelle un problème informatique NP-Complet. Cela signifie qu'il n'existe pas de solution exacte en temps polynomial - il existe un terme informatique très technique - mais en gros, cela signifie qu'il n'existe pas de solution exacte en temps polynomial. En fait, on m'a donné un tas de séquences et on m'a demandé de trouver l'alignement optimal de séquences multiples. Il n'y a aucun moyen de faire cela en temps polynomial. C'est très très lent. Des heuristiques ont été développées pour fournir des solutions optimales et approximatives. Par exemple, vous avez peut-être entendu parler de ClustalOmega, MUSCLE et MAFFT. Il s'agit d'outils standard utilisés dans ce domaine. Toutefois, même ces outils heuristiques s'adaptent généralement de manière quadratique au nombre de séquences. À titre d'exemple, la base de données GISAID, qui est la base de données dans laquelle la plupart des gens stockent leurs génomes complets de SARS-CoV-2, croît extrêmement rapidement et, à ce jour, nous disposons de plus de 15 millions de séquences de SARS-CoV-2 provenant du monde entier. La prochaine épidémie sera celle du séquençage des génomes en temps réel. Il s'agira d'un outil qui, je l'espère, continuera à être utilisé dans les épidémies virales à venir. On peut s'attendre à ce que cela devienne un problème de big data encore plus important. Actuellement, avec ces outils comme ClustalOmega, MAFFT et MUSCLE, les temps d'exécution sont de l'ordre de plusieurs décennies, voire de plusieurs siècles, ce qui, pour des raisons évidentes, si nous essayons de faire une analyse moléculaire en temps réel, est un peu trop lent. Alors, comment accélérer les choses ? Il s'avère que le problème est en fait un peu plus simple que ce que nous essayons de résoudre. L'alignement de séquences multiples, en général, suppose qu'il n'y a aucune homologie entre les séquences. C'est le temps qu'il faut pour aligner des séquences complètement arbitraires. Mais avec le SRAS-CoV-2 et les virus en général, le problème est beaucoup plus simple, n'est-ce pas ? Nous disposons d'une grande

homologie de séquences. Même si le virus subit des mutations importantes dans le monde entier, chaque séquence virale obtenue sera presque identique au gène de référence. Elle ne sera pas exactement identique, mais elle sera presque identique. Nous sommes donc confrontés à un problème informatique beaucoup plus simple, à savoir l'alignement multiple de séquences très similaires. Comment pouvons-nous utiliser cette caractéristique pour accélérer l'analyse ?

Slide 4

Nous pouvons adopter ce que l'on appelle une approche d'alignement sur la référence. Au lieu d'essayer d'aligner toutes les séquences les unes sur les autres en une seule fois, nous pourrions procéder à des alignements individuels par paire par rapport à une unité de référence. Dans cette figure, la barre verte plus épaisse en haut représente la référence à notre génome scope 2, et chacun de ces autres génomes colorés représente une séquence que j'ai collectée dans le monde réel. Je veux aligner chacun d'entre eux sur le génome de référence. Ce que je pourrais faire, c'est aligner indépendamment, une par une, chacune de ces séquences génomiques sur le génome de référence, ce que je pourrais faire assez rapidement et je pourrais procéder à une parallélisation massive parce que chacun de ces alignements par paire sur la référence peut être effectué de manière totalement indépendante. Je peux paralléliser autant de cœurs que mon ordinateur en possède, je peux en lancer autant sur ce problème. Ensuite, une fois que j'ai aligné toutes ces lignes par paire sur la référence, je peux utiliser le génome de la clé à molette - je peux en quelque sorte utiliser ses ancres, ses positions comme ancres pour créer les colonnes de ma ligne de séquences multiples. Par exemple, je commencerai peut-être par la première position du génome de référence et je verrai que, dans la séquence rouge, c'est la lettre qui s'est alignée sur cette position. Dans la séquence orange, voici la lettre. Dans la séquence rose, c'est la lettre. Dans la séquence bleue, c'est la lettre. Je peux fusionner toutes ces lettres dans une colonne de mon alignement de séquences multiples. Je pourrais faire la même chose pour la deuxième position de mon génome de référence, la même chose pour la troisième position, la quatrième position, tout le long. Ainsi, position par position par position, je peux construire mon alignement de séquences multiples. Cette idée est vraiment bonne parce qu'elle est massivement parallélisable et qu'elle s'adapte linéairement au nombre de séquences plutôt que quadratiquement. L'évolutivité est donc bien meilleure. Devons-nous mettre en œuvre cette approche à partir de zéro ? En fait, non.

Slide 5

Il s'avère que si l'on prend un peu de recul et que l'on réfléchit à ce problème, il est en fait équivalent, dans un sens, au problème du mappage de la lecture longue. Prenons un peu de recul et repensons au problème auquel nous nous attaquons. Notre entrée est un génome de référence et un ensemble de longues séquences très similaires au génome de référence. Notre résultat est un alignement de chacune de ces séquences par rapport au génome de référence. Il s'agit exactement du même problème de calcul que celui de la cartographie des longues lectures. Au lieu de réinventer la roue, nous pourrions nous inspirer de toutes ces techniques

avancées que les gens ont mises au point pour résoudre le problème de la cartographie des longues séquences et les appliquer à ce contexte.

Slide 6

À cette fin, j'ai mis au point un outil appelé ViralMSA, qui se contente d'envelopper les cartographes à longues lectures existants pour effectuer un alignement de séquences multiples guidé par la référence. Il traite en quelque sorte chacun des génomes que j'ai recueillis comme de longues lectures et il traite le génome de référence comme un génome de référence. Il suffit d'appeler ce cartographe de lecture - j'ai effectué un alignement avec quelques cartographes de lecture différents pour démontrer la flexibilité - mais je suggère principalement aux gens d'utiliser Minimap2 pour sa rapidité et sa précision. Ensuite, compte tenu des résultats de la cartographie des lectures, je peux - ou compte tenu des résultats de la cartographie - les compiler en une seule ligne de séquences multiples.

Tout ce que vous avez à faire pour exécuter ViralMSA, c'est de lui donner un génome de référence et un ensemble de séquences à aligner. Il s'occupera automatiquement de l'indexation du génome de référence, et si vous lui donnez un numéro d'accessoire, il s'occupera du téléchargement et de l'indexation du génome de référence. Il s'occupera de tous les prétraitements et de toutes les opérations en aval, et il sortira simplement - il appellera le cartographe de lectures, il fusionnera les résultats dans l'alignement SQL multiple, et il sortira simplement un fichier standard unique qui est votre alignement de séquences multiples.

Slide 7

Quelle est la performance du logiciel par rapport aux outils existants ? Nous avons réalisé une expérience de référence au cours de laquelle nous avons comparé la durée d'exécution de ViralMSA autour de Minimap 2 par rapport à Virulign, qui est une approche d'alignement sur la référence existante, mais qui met en œuvre son propre alignement sur la référence à partir de zéro. Nous l'avons également comparé à MAFFT, qui est généralement considéré comme l'un des outils de segmentation de séquences multiples les plus couramment utilisés. Sur ce graphique, l'axe horizontal indique le nombre de séquences. Sur l'axe vertical, j'indique le temps d'exécution total en secondes. Cette analyse a été réalisée sur des séquences complètes du génome SARS-CoV-2, soit une longueur de génome d'environ 29 000. Comme on peut le voir, la ligne bleue, qui représente ViralMSA, est beaucoup plus rapide que les outils existants. Par rapport à VIRULIGN, dont l'échelle est également linéaire, nous obtenons - et en passant, ce graphique est un graphique à échelle logarithmique - donc par rapport à VIRULIGN, nous sommes en gros mille fois plus rapides. Avec MAFFT, nous ne sommes pas aussi rapides, mais vous pouvez voir que, parce que MAFFT s'échelonne quadratiquement, notre vitesse par rapport à MAFFT augmente au fur et à mesure que le temps s'écoule. Même avec un millier de séquences, nous sommes environ mille fois plus rapides et l'écart se creuse.

Slide 8

Vous vous demandez peut-être : " D'accord, c'est bien d'être rapide, mais à quoi bon si cela ne me donne pas de bons alignements ? Nous avons également comparé la précision. Nous avons pris l'alignement de séquences multiples calculé par MAFFT, l'alignement de séquences multiples calculé par ViralMSA sur un ensemble d'alignements de VIH, d'Ebola, et je ne sais plus quel est le troisième virus, mais nous avons pris des virus pour lesquels nous avions des alignements curatés du Los Alamo- oh en fait, non - ce graphique ne concerne que le VIH-1. Nous avons pris les alignements de séquences multiples du laboratoire national de Los Alamos et nous les avons utilisés comme vérité de base. Ensuite, nous avons vu comment la cartographie de ViralMSA se compare à l'alignement de séquences multiples de référence. Si nous calculons les distances par paire des séquences que nous obtenons dans notre alignement, et que nous effectuons ensuite un test de manteau pour la précision - nous trouvons la corrélation entre nos distances par paire et les distances par paire calculées directement à partir du véritable alignement de séquences multiples, nous constatons que la corrélation est négligeable. Ici, nous obtenons un coefficient de corrélation de 0,994 pour l'ASM viral contre 0,997 pour le calcul de la distance par paire. En fait, lorsque nous avons calculé les phylogénies, les phylogénies déduites à l'aide des éléments de séquences multiples ViralMSA sont en réalité légèrement plus précises sur le plan topologique que celles estimées à l'aide de MAFFT. C'est donc négligeable, mais ce que nous montrons, c'est qu'elles sont essentiellement équivalentes en termes de précision, à toutes fins utiles.

Slide 9

Conclusion - ViralMSA est un outil qui permet d'aligner rapidement des séquences multiples de très grands ensembles de données virales. Il s'agit d'un logiciel libre, vous pouvez le trouver sur GitHub, et vous savez, n'hésitez pas à l'utiliser dans vos analyses virales.

Slide 10

Remerciements - Je tiens à remercier Heng Li, le développeur de Minimap2, et c'est vraiment son expertise dans le développement de Minimap2 qui permet à ViralMSA d'être aussi rapide et performant. Je tiens à remercier la NSF pour la subvention qui soutient ce projet. La recherche a également été soutenue par les crédits de recherche de la plateforme Google Cloud.

Slide 11

Je vais donc garder du temps pour d'éventuelles questions ou je suis heureux d'en finir ici.