COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Alexander Niema Moshiri (University of California San Diego), November 13, 2020

Title: RAPID: Inférence phylogénétique en temps réel et analyse groupée de transmission du COVID-19

Alexander N Moshiri CIC Profile

NSF Award #: 2028040

YouTube Recording with Slides

November 2020 CIC Webinar Information

Transcript Editor: Macy Moujabber

---

Transcript:

*Slide 1*

Cool. Je vous remercie. J'espère que les gens voient mes diapositives. Je m'appelle Niema Moshiri. Je suis à l'université de San Diego. Je fais partie du département d'informatique et d'ingénierie et je vais vous parler d'un développement récent que nous avons réalisé et qui s'appelle la MSA virale.

*Slide 2*

Fondamentalement, un flux de travail phylogénétique standard - si vous avez déjà vu l'histoire de l'évolution des séquences de virus, vous commencez avec ces séquences initiales qui ne s'alignent pas nécessairement très bien. La première étape consiste à effectuer un alignement de séquences multiples pour les aligner. Cet alignement nous donne des informations sur l'homologie des séquences, sur les relations entre les séquences.

Cet alignement permet d'estimer une phylogénie. Ensuite, on enracine la phylogénie pour déterminer quel est l'ancêtre commun. Il y a ensuite tout un tas d'autres analyses en aval que vous pourriez être intéressé à faire, vous savez, des choses comme quels sont les groupes d'épidémies qui se produisent, vous savez, quelles démographies pensons-nous être plus à risque d'être infecté par une certaine maladie ? Il y a beaucoup de choses que l'on peut faire. En général, je me concentre sur ces deux étapes : l'alignement de séquences multiples et l'inférence phylogénétique, qui sont vraiment les goulots d'étranglement informatiques. Dans cet exposé, je me concentrerai sur cette seule étape : l'alignement de séquences multiples.

*Slide 3*

Le problème de l'alignement de séquences multiples est NP-Complet, ce qui signifie qu'il n'existe pas de solution en temps polynomial et que de nombreuses heuristiques ont été développées pour obtenir des

solutions approximatives. Elles sont raisonnablement précises. Certains des outils que les gens connaissent peut-être sont MUSCLE, ClustalOmega et MAFFT. Ce sont quelques-uns des outils courants qui les mettent en œuvre. Cependant, même ces heuristiques s'échelonnent généralement de manière quadratique en fonction du nombre de séquences. Ainsi, dans le cas du SRAS-CoV-2, nous avons assisté à une croissance exponentielle du séquençage, ce qui est une excellente chose pour nous. Nous obtenons de plus en plus de données de séquences, mais l'inconvénient est que nous devons analyser ces données de séquences et que nos outils ne sont pas suffisamment évolutifs. C'est pourquoi, à l'heure actuelle, ce document est en fait obsolète. Il date d'il y a une ou deux semaines. Aujourd'hui, nous en sommes presque à 200 000 séquences. Les outils ne sont donc pas adaptés pour permettre une analyse en temps réel.

*Slide 4*

Et si, au lieu de cela, nous savions à l'avance que nos séquences seront très similaires et que nous disposons déjà d'un génome de référence représentatif et fiable auquel nous pouvons les comparer ? Ici, je montre le génome de référence sous la forme d'une ligne verte en haut et chacune des autres séquences colorées en dessous sont les séquences que je veux aligner les unes sur les autres. Au lieu de les aligner entre elles, je peux aligner la première séquence sur le génome de référence. Aligner la deuxième séquence sur le génome de référence. Continuer à aligner chacune des séquences indépendamment, directement par rapport au génome de référence. Puis, en utilisant les positions du génome de référence comme points d'ancrage, je peux fusionner les alignements individuels par paire en une ligne de séquences multiples. Prenez donc la première colonne de mon affectation de séquences multiples, voyez - dans la première séquence, c'est la position qui correspond, puis cette position correspond à la deuxième et à la troisième, et regroupez-les ensemble. Je peux faire cela pour chaque position du génome de référence et j'ai maintenant une ligne de séquences multiples. Ce qui est bien, c'est que cela se fait très bien en parallèle. Chaque séquence peut être alignée sur le génome de référence indépendamment et simultanément, et l'échelle est linéaire avec le nombre de séquences plutôt que quadratique.

*Slide 5*

Mais revenons un peu en arrière et réfléchissons à cette approche. Mon entrée est un génome de référence et un ensemble de séquences très similaires à ce génome de référence, et mon résultat est un alignement de chaque séquence par rapport au génome de référence. Si les gens sont un tant soit peu familiers avec le séquençage de longues lectures, il s'agit exactement du même problème de calcul que la mise en correspondance de longues lectures avec un génome de référence. Ma question était donc la suivante : puis-je exploiter les outils de mappage de lectures existants et bien mis en œuvre pour permettre ce type d'alignement de séquences multiples évolutif guidé par une référence ?

*Slide 6*

J'ai donc mis au point un outil appelé ViralMSA qui contourne les outils de mappage de lectures existants afin d'effectuer un alignement de séquences multiples guidé par la référence. J'en utilise plusieurs, mais il y a un outil en particulier, Minimap2, qui est en quelque sorte l'étalon-or pour ce que je fais. Je ne recommande donc l'utilisation de mon outil qu'avec ce cartographe de lecture spécifique, mais j'en utilise plusieurs pour montrer que je peux faire évoluer cet outil naturellement au fur et à mesure que les technologies de cartographie de lecture évoluent elles aussi. En gros, il suffit de fournir le génome de

référence viral de la MSA et les séquences à aligner. L'outil se chargera d'indexer le génome de référence et d'effectuer tout prétraitement nécessaire, puis il appellera le cartographe de lectures et fusionnera les résultats en un alignement de séquences multiples.

*Slide 7*

Vous pouvez voir ici une comparaison de cette approche avec les outils de meilleures pratiques existants. Mon outil est la ligne bleue en bas et les deux autres lignes sont d'autres outils. Vous voyez qu'en général, cette approche est plus rapide de plusieurs ordres de grandeur que ce que les gens font actuellement et qu'elle s'adapte très bien. Et les alignements de séquences que nous obtenons sont très précis.

*Slide 8*

En conclusion, l'outil que j'ai mis au point permet d'aligner rapidement des séquences multiples de génomes viraux. Il s'agit d'un logiciel libre. Vous pouvez l'obtenir en ligne et j'espère que vous envisagerez de l'utiliser si vous effectuez des analyses virales.

*Slide 9*

Quelques remerciements : Heng Li a développé Minimap2, qui est en quelque sorte l'essence de la vitesse, et ce travail a été soutenu par la NSF et Google. Et oui, je vais laisser les questions pour le chat ou plus tard dans la session.