COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Ali Rahnavard (George Washington University), October 2022

Title: *A novel platform for data integration and deep learning on COVID-19*

Ali Rahnavard CIC Database Profile

NSF Award #: 2028280

Youtube Recording with Slides

October 2020 CIC Webinar Information

Transcript Editor: Julie Meunier

Transcript

Slide 1

Ali Rahnavard:

Merci beaucoup pour cette opportunité. C'est un travail de collaboration au Computational Biology Institute de l'Université George Washington.

Slide 2

Brièvement, dans mon laboratoire, nous nous concentrons sur les technologies à haut débit, qui nous permettent de mesurer de nombreux paramètres "-omiques" différents. Notre objectif est de voir si nous pouvons obtenir une meilleure image en intégrant ces différentes données "-omiques", y compris la génomique, la métabolomique, la protéomique et les génomes viraux.

Slide 3

Donc en effet, l'objectif ou la proposition NSF était de développer une plateforme computationnelle qui inclut deux ensembles - deux approches analytiques différentes de l'investissement - d'approches analytiques et de logiciels pour étudier les données "-omiques" liées à la COVID-19.

Slide 4

Comme vous avez pu le constater plus tôt dans les différentes présentations, la principale cause de cette pandémie est le virus, et nous nous concentrons sur le génome du virus en tant que ressource pour étudier le comportement du virus et déterminer s'il existe des protéines ou des régions spécifiques dans le génome du virus que nous devons cibler pour le développement d'un vaccin.

Slide 5

Notre manière d'exploiter les données est d'utiliser les données de séquençage, nous disposons du génome du virus pour tous les individus infectés de la pandémie. Et nous nous concentrons aussi sur les variations du génome, dans des régions spécifiques du génome, incluant les protéines et d'autres régions comme les protéines non structurales. Notre méthode pour le réaliser : nous obtenons les séquences génomiques de tous les individus de notre population, nous essayons de calculer la variation expliquée dans chacun de ces échantillons par rapport aux autres échantillons, et nous essayons d'évaluer et de voir ce que signifient ces variations que nous observons.

Slide 6

Je vous montre donc ici une feuille de route sur comment nous pouvons calculer, comment nous pouvons regarder les variations du génome dans des régions spécifiques. Sur la première ligne, ce que vous voyez est la variation du génome, la corrélation avec d'autres régions du génome. Et nous observons quatre régions, cinq ici, qui sont très corrélées à la variation du génome qu'elles expliquent à travers les populations. Et deux de ces régions sont très intéressantes. L'une d'entre elles est la région de la protéine de spicule. Le nombre que vous voyez, 32.5, veut dire que la variation que nous observons dans le génome du virus dans la population est corrélée entre la protéine S et le génome entier.

Et aussi pour le NSP nous observons trois régions. Donc les trois autres régions que vous observez en jaune, ce sont des grandes régions et elles ont des sous-ensembles. C'est pourquoi nous ne nous sommes pas trop concentrés sur ces éléments, car nous nous attendions à ce qu'il y ait une forte corrélation entre une grande région du génome et la façon dont elle transmet l'information que le génome entier contient. Donc nous voyons ici qu'il y a des endroits spécifiques dans le génome du virus, comme cette protéine spiculaire, qui facilite son entrée dans les cellules humaines. C'est très corrélé avec la protéine non-structurale 3, et ces deux protéines sont en corrélation avec la variation de l'ensemble du génome que nous observons dans la population.

Dans le cadre de notre subvention de la NSF, nous avons donc développé une approche appelée omeClust. Ce qu'il fait, c'est que vous lui donnez un ensemble de points ici, les points sont les individus ou les souches du virus, et les trois - lorsque nous l'exécutons en utilisant les informations des trois régions différentes, premièrement, le génome entier du virus. Deuxièmement, la protéine de spicule et troisièmement, la protéine non structurale. Nous constatons que ces trois régions nous donnent les mêmes communautés. Et c'est le cas pour toutes les régions. Cela suggère qu'il est important de cibler ces deux distributions et de les étudier plus avant.

Slide 7

De plus, nous observons cette variation que nous avons calculé à travers la population pour voir comment elle est corrélée avec les données épidémiologiques. Nous ne disposions donc pas de beaucoup d'informations, à l'exception du sexe et de l'âge, qui constituent un bon exemple : le génome du virus n'est pas vraiment en corrélation avec le sexe et l'âge, comme on s'y attendrait. Le résultat pourrait être sans rapport, mais pas le génome de la variation du génome.

Slide 8

Dans le cadre de nos travaux, nous intégrons donc en partie les données “-omiques” que nous mesurons chez les personnes infectées, comme les protéines ou le métabolisme. Nous voulons voir comment ces informations, lorsque nous les rassemblons, nous donnent des informations qui nous permettent d'émettre des hypothèses à partir des données que nous pouvons cibler. Et, par exemple, ce que vous pouvez voir ici, il y a une approche que nous développons appelée ‘btest’. Le principe : on vous donne les informations sur les métabolites des patients et les protéines des patients et le corps , il nous donne ensuite un bloc de relations, comment, quels sont les métabolites qui sont liés aux protéines dans les individus infectés. Nous sommes aussi en train de développer une approche d’apprentissage profond (*Deep learning*), qui prend en entrée le séquençage des données et qui donne en sortie individuellement, chez la personne infectée, comment la gravité va se manifester en fonction du génome du virus que nous recevons de la personne infectée. Ainsi, comme vous le voyez, nous développons différentes approches, méthodes, approches d'apprentissage profond, pour étudier les données “-omiques”. Nous nous sommes d'abord concentrés sur le génome, le génome viral, et maintenant nous passons aux protéines métabolisées et nous essayons d'intégrer toutes ces informations en utilisant l'approche d'apprentissage profond.

Slide 9

Il y a donc ici plusieurs étudiants de mon laboratoire ainsi que des cochercheurs principaux de l'Institut de biologie computationnelle de Washington, qui collaborent à ce travail. C'est un travail d'équipe, vous pouvez donc en savoir plus sur nos approches et sur les résultats de COVID-19. Nous les publions sur une page web du site de mon laboratoire. Nous mettons également à votre disposition les logiciels que nous développons et que vous pouvez utiliser pour étudier vos données. Sur ce, je vous remercie de votre attention.