COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Ho-Joon Lee (Yale University), February 2022

Title: A landscape of virus-host protein-protein interactions in SARS-CoV-2 infection in humans by machine learning

Funded by NSF Office of Advanced Cyberinfrastructure, Directorate for Computer & Information Science & Engineering (OAC/CISE) through the Northeast Big Data Innovation Hub Seed Fund Program.

Youtube Recording with Slides

February 2022 CIC Webinar Information

Transcript Editor: Julie Meunier

Transcript

Ho-Joon Lee:

Slide 1

Bonjour à tous, je m'appelle Ho-Joon Lee de la Yale School of Medicine et je vais parler du paysage interactif des interactions protéine-protéine du virus SARS-CoV-2 avec les protéines humaines par l'apprentissage automatique.

Slide 2

Il y a deux objectifs. Le premier est de développer des classificateurs d'apprentissage automatique ou d'apprentissage profond multi-classes basés sur la séquence protéique pour la prédiction du niveau de preuve ou de confiance en utilisant la base de données Viruses.STRING. Le deuxième est d'utiliser ces classificateurs pour créer un paysage interactif préliminaire des interactions protéine-protéine du cytosquelette virus-humain.

Slide 3

Voici un aperçu de notre flux de travail d'apprentissage automatique et d'apprentissage profond. Nous utilisons la base de données Viruses.STRING, qui n'incluait pas le SARS-CoV-2 au moment de l'analyse. Il s'agit du réseau des interactions protéine-protéine virus-humain qui contient plus de 80 000 interactions entre environ 1 200 protéines virales de 102 espèces de virus et environ 8 500 protéines humaines. Chaque interaction a un score combiné allant de zéro à mille, que nous convertissons en cinq classes de

preuve. Nous nous concentrons sur les interactions protéine-protéine expérimentales appartenant aux classes de preuve 3 ou 2. À partir de ces données, nous extrayons d'abord des caractéristiques des nœuds, d'autres caractéristiques de protéines qui sont des compositions fractionnelles des 20 acides aminés. À ce stade, nous développons deux modèles différents : l'un est basé sur des modèles plus canoniques comme Random Forests et XGBoost, dans ce cas. Et l'autre est basé sur l'apprentissage profond. Nous utilisons spécifiquement des réseaux neuronaux graphiques comme GraphSAGE ou la version datalisée de HinSAGE. Pour l'apprentissage automatique connecté, nous extrayons également des caractéristiques d'arêtes qui sont 72 mesures de distance ou de similarité entre les profils de composition en acides aminés entre les protéines virales et humaines. Et sur la base de ces caractéristiques, nous développons les Random Forests et XGBoost. Pour les Random Forests, nous optimisons 36 modèles par recherche avec une régulation temporelle et 432 modèles pour l'espace exécutif qui a la même transplantation contemporaine. En bref, nous obtenons jusqu'à 67 % de précision et 37 % de précision pour les cas de Random Forests et 74 % de précision et 67 % de précision pour les cas de XGBoost. Cette partie a été récemment publiée en préimpression. Vous pouvez vous référer à l'article pour plus de détails [https://www.biorxiv.org/content/10.1101/2021.11.07.467640v2]. Pour GraphSAGE, les résultats sont encore en lecture et préparation avancées, mais je vais vous montrer brièvement les résultats de GraphSAGE également, car cela montre une précision de plus de 70 %, ce qui est également très prometteur.

Slide 4

Ici, je vais vous montrer un exemple de performance pour les meilleurs modèles pour 20 % [inaudible] avec cette graine aléatoire. Nous constatons, dans ce cas, que lorsque Random Forest montre une précision de 60 %, XGBoost affiche une précision de 67,7 %. Et si nous regardons les mesures informatiques, je vais me concentrer à nouveau sur [EC3?] qui implique principalement des PPI expérimentales. Et si nous regardons les classes individuelles, en nous concentrant sur le score f1, le extra booster montre des scores f1 plus élevés pour les quatre classes individuelles.

Slide 5

Sur la base de ce modèle de booster supplémentaire, les caractéristiques importantes ont été identifiées à l'aide de deux méthodes alternatives ici. L'une par l'indice de Gini et l'autre par l'analyse SHAP, qui se base sur les valeurs SHAP. Et de manière intéressante, nous constatons que la cystéine et l'histidine sont les deux caractéristiques les plus importantes. Où ce moins [C_minus et H_minus] signifie que la fraction de cystéine entre le virus et l'homme. Et le ratio signifie le ratio entre les fractions de réactions de cystéine et d'histidine entre le virus et les humains.

Slide 6

Une expérience de contrôle que nous avons réalisée est de comparer la prédiction des PPI expérimentales et - avec la prédiction des PPI par text mining dans le virus [inaudible]. En raison de la différence de taille des données, la différence est assez importante ici, un facteur de six, mais ce que nous observons ici, c'est que XGBoost, en fait, montre une précision plus élevée. Avec une précision de 94 % par rapport à une précision de 90 % pour le text mining. Malgré la différence de taille des données,

XGBoost montre une bonne performance de prédiction. Et c'est l'accord entre les activités de la forêt aléatoire pour EC3 et la liaison de test comme nous nous y attendons montre principalement EC1 ou EC2.

Slide 7

Sur la base de ces résultats encourageants, nous avons appliqué ces classificateurs au SARS-CoV-2 pour notre deuxième objectif de deux manières. Premièrement, nous l'appliquons à la base de données IntAct, qui est une collection de PPI expérimentales. Et ici, je vous montre le réseau par XGBoost avec une preuve prédite [inaudible]. Ainsi, EC3 pour le bleu, EC4 pour le rouge. Cela peut être considéré comme une priorisation d'un réseau. Donc, bien que ces liens soient d'environ 2 000 liens provenant de données expérimentales, ils sont tous également significatifs, nous pouvons également prioriser ces liens en fonction de cette classe prédite [inaudible] XGBoost dans ce cas. Deuxièmement, nous l'appliquons également à l'interaction protéique globale [inaudible] les paires d'au moins un demi-million entre 27 protéines SARS-CoV-2 et environ plus de 20 000 protéines humaines. Et ici, je vous montre le sous-ensemble de 22 000 PPI avec une classe de preuve d'au moins 2. J'utilise soit XGBoost, soit Random Forest. Et c'est un autre sous-ensemble - 140 PPI avec la classe de preuve la plus élevée, 5, par XGBoost. Et sur la base de ce réseau d'interactions, nous observons que de nombreuses protéines humaines enrichissent la contraction musculaire lisse vasculaire et les composants H2A.

Slide 8

Il y a quelques autres applications de ce travail qui ont été découvertes au cours du dernier mois, en fait. Donc, Giuseppe Novelli, qui est un généticien de renom à Rome, en Italie, il m'a contacté par e-mail et par surprise, le mois dernier. Il avait lu ma préimpression me parlant de sa publication sur les thérapeutiques de qualité pour les ligases HECT et son idée d'utiliser les résultats de ce travail important dans le cadre de cette recherche en cours. Et nous avons immédiatement réalisé que nous pouvions nous aider mutuellement sur la base de mes résultats sur les résultats du réseau interactif. Et nous avons constaté que les protéines de domaine HECT ont tendance à interagir avec les protéines SARS-CoV-2 avec une classe de preuve supérieure à 2 avec une signification statistique. En d'autres termes, les protéines de domaine HECT sont favorisées par le SARS-CoV-2. Sur la base de cette observation, nous nous demandons s'il existe d'autres familles de protéines favorisées par le SARS-CoV-2. De plus, nous pouvons également étendre cela à d'autres espèces de virus comme le métapneumovirus humain, sur lequel le Dr Novelli travaille également.

Slide 9

Enfin, je vais vous montrer brièvement les réseaux neuronaux graphiques en utilisant les architectures GraphSAGE et HinSAGE. À gauche, les précisions de 15 modèles différents utilisant trois poids Java différents dans les colonnes et cinq méthodes d'incorporation de bord différentes. Comme vous le voyez, sans taux d'abandon, en fait, nous observons des valeurs de précision de plus de 70 %, ce qui est très prometteur. Cela est basé sur Viruses.STRING et si nous l'appliquons à la base de données IntAct SARS-CoV-2, vous voyez que la prédiction est enrichie avec la classe de preuve 2 ou 3, en fait, qui sont principalement des PPI expérimentales. Et ce consensus est le nombre d'accords entre ces 15 modèles

différents. Nous voyons plus de consensus de 8-9 pour contre deux par rapport à 6-7 contre 1, mais je pense que cela est également très important car - nous verrons.

Slide 10

Très bien, avec cela, je tiens à remercier mes collaborateurs pour leurs discussions, leurs retours et leur soutien très utiles. Et le Yale Center for Research Computing pour les ressources informatiques. Et la communauté COVID HASTE de la Yale School of Engineering & Applied Science. Et enfin, le Northeast Big Data Hub Seed Fund pour le soutien à ce travail. Merci beaucoup.