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: Shikhar Johri

Transcript

हो-जून ली:

स्लाइड *1*

सभी को नमस्कार, मेरा नाम येल स्कूल ऑफ मेडिसिन से हो-जून ली है और मैं मशीन लर्निंग द्वारा *SARS-CoV-2* वायरस-मानव प्रोटीन-प्रोटीन इंटरैक्शन के एक इंटरएक्टोम परिदृश्य के बारे में बात करने जा रहा हूं।

स्लाइड *2*

इसके दो उद्देश्य हैं। पहला Viruses.STRING डेटाबेस का उपयोग करके सबूत या आत्मविश्वास स्तर की भविष्यवाणी के लिए प्रोटीन अनुक्रम आधारित बहु-वर्ग मशीन लर्निंग या गहन शिक्षण क्लासिफायर विकसित करना है। दूसरा है - उन क्लासिफायरों का उपयोग करके हम साइटोस्केलेटल वायरस मानव प्रोटीन-प्रोटीन इंटरैक्शन का एक मसौदा इंटरैक्टिव परिदृश्य बनाना चाहते हैं।

स्लाइड *3*

तो यहाँ हमारे मशीन लर्निंग और डीप लर्निंग वर्कफ़्लो का अवलोकन है। इसलिए हम Viruses.STRING डेटाबेस का उपयोग करते हैं, जिसमें विश्लेषण के समय SARS-CoV-2 शामिल नहीं था। यह पीपीआई वायरस-मानव पीपीआई का नेटवर्क है जिसमें 102 वायरस प्रजातियों के लगभग 1,200 वायरस प्रोटीन और लगभग 8,500 मानव प्रोटीन के बीच 80,000 से अधिक इंटरैक्शन होते हैं। और प्रत्येक इंटरैक्शन में शून्य से एक हजार तक का एक संयुक्त स्कोर होता है जिसे हम पांच साक्ष्य वर्गों के टुकड़ों में परिवर्तित

करते हैं। और यह साक्ष्य वर्गों के लिए संख्या पीपीआई का वितरण है। और हम प्रयोगात्मक पीपीआई पर ध्यान केंद्रित करने जा रहे हैं जो यहां शून्य सूचकांक के आधार पर साक्ष्य वर्ग 3 या 2 से संबंधित हैं। और डेटा के आधार पर, हम पहले नोड सुविधाओं को निकालते हैं, एक और प्रोटीन विशेषताएं जो 20 अमीनो एसिड की आंशिक रचनाएं हैं। और इस बिंदु पर हम दो अलग-अलग मॉडल विकसित कर रहे हैं - एक इस मामले में रैंडम फॉरेस्ट और XGBoost जैसे अधिक विहित [अश्रव्य] मॉडल हैं। और दूसरा गहरी शिक्षा पर आधारित है। हम विशेष रूप से ग्राफ तंत्रिका नेटवर्क जैसे ग्राफसेज या हिनसेज के डेटालाइज्ड संस्करण का उपयोग करते हैं।का उपयोग कर सकते हैं कनेक्टेड मशीन लर्निंग के लिए, हम एज फीचर्स भी निकालते हैं जो वायरस प्रोटीन और मानव प्रोटीन के बीच अमीनो एसिड संरचना प्रोफाइल के बीच 72 दूरी या समानता के उपाय हैं। और सुविधाओं के आधार पर, हमने रैंडम फ़ॉरेस्ट और XGBoost विकसित किए। रैंडम फ़ॉरेस्ट के लिए, हम अस्थायी अनुरोध विनियमन के साथ अनुसंधान द्वारा 36 मॉडल और कार्यकारी स्थान के लिए 432 मॉडल का अनुकूलन करते हैं जिसमें समान समकालीन प्रत्यारोपण होता है। और, संक्षेप में, हम यादृच्छिक वन मामलों के लिए 67% सटीकता और 37% सटीकता और XGBoost मामलों के लिए 74% सटीकता और 67% सटीकता प्राप्त करते हैं। और यह काम, यह हिस्सा, हाल ही में एक प्रीप्रिंट के रूप में प्रकाशित किया गया है। तो आप पेपर को विस्तार से देख सकते हैं [https://www.biorxiv.org/content/10.1101/2021.11.07.467640v2]। और यहां ग्राफसेज के लिए, अभी भी उन्नत पढ़ने और तैयारी में है, लेकिन मैं आपको दिखाने जा रहा हूं, संक्षेप में आपको दिखाता हूं, ग्राफसेज के परिणाम भी। क्योंकि यह 70% से अधिक सटीकता दिखाता है जो काफी आशाजनक भी है।

स्लाइड *4*

और यहां मैं आपको इस यादृच्छिक बीज के साथ [अश्रव्य] के 20% के लिए सर्वश्रेष्ठ मॉडल के लिए एक प्रदर्शन उदाहरण दिखाने जा रहा हूं। हम देखते हैं, इस मामले में, जब वन 60% सटीकता दिखाता है, तो XBG 67.7% सटीकता थी। और यदि आप कंप्यूटर मेट्रिक्स को देखते हैं, तो फिर से, मैं इस [ईसी 3?] पर ध्यान केंद्रित करने जा रहा हूं जिसका अर्थ है ज्यादातर विस्तारित पीपीआई। और अगर हम अलग-अलग वर्गों को देखें, तो f1-स्कोर पर ध्यान केंद्रित करते हुए, अतिरिक्त बूस्टर चार अलग-अलग वर्गों में उच्च f1 स्कोर दिखाता है।

स्लाइड *5*

पर आधारित, इस पर आधारित करने के लिए [अश्रव्य] बूस्ट मॉडल. यहां दो वैकल्पिक तरीकों का उपयोग करके महत्वपूर्ण विशेषताओं की पहचान की गई थी। एक गिनी इंडेक्स द्वारा और दूसरा SHAP विश्लेषण द्वारा, जो SHAP पर आधारित है, SHAP मानों के माध्यम से आया था। और दिलचस्प रूप से पर्याप्त, हम देखते हैं कि सिस्टीन और हिस्टिडाइन सबसे अधिक हैं - दो सबसे महत्वपूर्ण विशेषताएं। जहां यह माइनस [C_minus और H_minus] का मतलब है कि वायरस और मानव के बीच सिस्टीन का अंश। और अनुपात का अर्थ है वायरस और मनुष्यों के बीच अंशों सिस्टीन और हिस्टिडाइन प्रतिक्रियाओं के बीच का अनुपात।

स्लाइड *6*

हमारे द्वारा किया गया एक नियंत्रण प्रयोग प्रयोगात्मक पीपीआई की भविष्यवाणी की तुलना करना है और - वायरस में टेक्स्ट माइनिंग पीपीआई की भविष्यवाणी के साथ [अश्रव्य]। क्योंकि डेटा का आकार, अंतर यहां बहुत बड़ा है, छह वर्ग अंतर, लेकिन हमने यहां जो देखा वह यह है कि XGBoost, वास्तव में, उच्च सटीकता दिखाता है। पाठ खनन मामले के लिए 90% सटीकता की तुलना में 94% सटीकता के साथ।

इसलिए डेटा आकार अंतर के बावजूद, XGBoost यह एक अच्छा पूर्वानुमान प्रदर्शन दिखाता है। और यह ec3 और परीक्षण बाध्यकारी के लिए यादृच्छिक बल गतिविधियों के बीच समझौता है जैसा कि हम उम्मीद करते हैं कि ज्यादातर ec1 या ec2 दिखाता है।

स्लाइड *7*

इसलिए उन उत्साहजनक परिणामों के आधार पर, हमने उन क्लासिफायरों को SARS-CoV-2 में अपने दूसरे उद्देश्य के लिए दो तरीकों से लागू किया। तो सबसे पहले, हम इसे इंटैक्ट डेटाबेस पर लागू करते हैं, जो प्रयोगात्मक पीपीआई का एक संग्रह है। और यहाँ मैं आपको XGBoost द्वारा [अश्रव्य] अनुमानित साक्ष्य के साथ नेटवर्क दिखा रहा हूँ। तो नीले, ec4, लाल के लिए ec3। तो इसे नेटवर्क को प्राथमिकता देने के रूप में देखा जा सकता है। इसलिए यद्यपि ये लिंक प्रयोगात्मक डेटा से लगभग 2,000 लिंक समान रूप से सार्थक होंगे, हम इस मामले में इस [अश्रव्य] वर्ग की भविष्यवाणी के आधार पर उन लिंक को भी प्राथमिकता दे सकते हैं। दूसरे, हम इसे प्रोटीन-वाइड इंटरैक्शन [अश्रव्य] 27 SARS-CoV-2 प्रोटीन और लगभग 20,000 से अधिक मानव प्रोटीन के बीच आधे मिलियन से अधिक के पुराने जोड़े पर भी लागू करते हैं। और यहां मैं आपको कम से कम 2 के साक्ष्य वर्ग के साथ 22,000 पीपीआई का सबसेट दिखा रहा हूं। मैं या तो वास्तव में XGBoost या रैंडम फ़ॉरेस्ट का उपयोग करता हूं। और यह एक और सबसेट है - उच्चतम साक्ष्य वर्ग के साथ 140 पीपीआई, 5, XGBoost द्वारा। और इस इंटरैक्शन नेटवर्क के आधार पर हमने देखा कि कई मानव प्रोटीन संवहनी, चिकनी मांसपेशियों के संकुचन और लक्ष्य और एच 2 ए घटकों को समृद्ध कर रहे हैं।

स्लाइड *8*

इस काम के कुछ और अनुप्रयोग हैं जो वास्तव में पिछले महीने में पाए गए हैं। तो Giuseppe Novelli, जो रोम में प्रसिद्ध आनुवंशिकीविद् हैं, इटली में, वह पिछले महीने ईमेल द्वारा और आश्चर्य से मेरे पास पहुंचे। उन्होंने मेरे प्रीप्रिंट को पढ़ा था जो मुझे एचईसीटी ई 3 लिगेस के लिए उनके गुणवत्ता चिकित्सीय प्रकाशन और इस चल रहे शोध के माध्यम से इस महत्वपूर्ण कार्य के परिणामों का उपयोग करने के उनके विचार के बारे में बता रहा था। और हमने तुरंत महसूस किया कि हम इंटरैक्टिव नेटवर्क परिणामों पर अपने परिणामों के आधार पर एक-दूसरे की मदद कर सकते हैं। और हमने पाया कि HECT-डोमेन प्रोटीन सांख्यिकीय महत्व के साथ 2 से अधिक के साक्ष्य वर्ग के साथ SARS-CoV-2 प्रोटीन के साथ बातचीत करते हैं। दूसरे शब्दों में, HECT-डोमेन प्रोटीन SARS-CoV-2 के पक्षधर हैं। उस अवलोकन के आधार पर, आप पूछ रहे हैं कि क्या SARS-CoV-2 के पक्ष में अन्य प्रोटीन परिवार हैं। इसके अलावा हम इसे मानव मेटान्यूमोवायरस जैसी अन्य वायरस प्रजातियों तक भी बढ़ा सकते हैं, जिस पर डॉ. नोवेली भी काम कर रहे हैं।

स्लाइड *9*

तो अंत में, मैं आपको ग्राफसेज और हिंसेज आर्किटेक्चर का उपयोग करके ग्राफ तंत्रिका नेटवर्क के बारे में संक्षेप में दिखाने जा रहा हूं। बाईं ओर कॉलम में तीन अलग-अलग जावा वेट और पांच अलग-अलग एज एम्बेडिंग विधियों का उपयोग करके 15 अलग-अलग मॉडलों द्वारा सटीकता है। जैसा कि आप देखते हैं, ड्रॉपआउट दरों के बिना, वास्तव में, हम 70% से अधिक सटीकता मान और सटीकता देखते हैं जो बहुत आशाजनक हैं। यह Viruses.STRING पर आधारित है और यदि हम इसे SARS-CoV-2 IntAct डेटाबेस पर लागू करते हैं, तो आप देखते हैं कि भविष्यवाणी साक्ष्य वर्ग 2, या 3 से समृद्ध है, वास्तव में, जो ज्यादातर [अश्रव्य] पीपीआई हैं। और यह आम सहमति इन 15 विभिन्न मॉडलों के बीच समझौतों की

संख्या है। हम 1 के खिलाफ 6-7 की तुलना में दो के खिलाफ 8-9 की अधिक सहमति देखते हैं, लेकिन मुझे लगता है कि यह भी बहुत महत्वपूर्ण है क्योंकि - हम देखेंगे।

स्लाइड *10*

ठीक है, इसके साथ मैं अपने सहयोगियों को बहुत उपयोगी चर्चाओं और प्रतिक्रिया और समर्थन के लिए धन्यवाद देना चाहता हूं। और कम्प्यूटेशनल संसाधनों के लिए येल सेंटर फॉर रिसर्च कम्प्यूटिंग। और येल के स्कूल ऑफ इंजीनियरिंग एंड एप्लाइड साइंस से COVID HASTE समुदाय। और अंत में इस काम का समर्थन करने के लिए पूर्वोत्तर बिग डेटा हब बीज कोष। बहुत-बहुत धन्यवाद।