COVID Information Commons (CIC) Research Lightning Talk

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

Title: Massively scalable reference-guided Multiple Sequence Alignment of viral genomes

Niema Moshiri CIC Database Profile

NSF Award #: 2028040

YouTube Recording with Slides

Spring 2023 CIC Webinar Information

Transcript Editor: Shikhar Johri

Transcript

नीमा मोशीरी:

बहुत बढ़िया, हाँ, परिचय के लिए धन्यवाद। उम्मीद है, लोग मेरी स्क्रीन देख सकते हैं। हाँ, तो हे सब लोग। जैसा कि उल्लेख किया गया है, मेरा नाम नीमा मोशीरी है। मैं यूसी सैन डिएगो में कंप्यूटर विज्ञान और इंजीनियरिंग विभाग में एक सहायक शिक्षण प्रोफेसर हूं। मेरी बात कुछ तरीकों पर केंद्रित होने जा रही है जो मेरी प्रयोगशाला ने वायरल जीनोमिक विश्लेषण को तेज करने के लिए एनएसएफ फंडिंग का उपयोग करके विकसित किया है। विशेष रूप से, आज की बात सिर्फ इस बात पर ध्यान केंद्रित करने जा रही है कि हमने पूर्ण वायरल जीनोम के बड़े पैमाने पर स्केलेबल संदर्भ निर्देशित कई अनुक्रम संरेखण को कैसे सक्षम किया है। हमने वास्तव में कई अन्य त्वरण भी किए हैं जिनके बारे में मेरे पास आज बात करने का समय नहीं था। मैं अपनी वेबसाइट के लिंक के साथ समाप्त करूंगा यदि लोग इस बारे में उत्सुक हैं कि आप इस प्रकार के विश्लेषण के अन्य पहलुओं को कैसे गति दे सकते हैं।

स्लाइड *2*

तो चलो शुरू करते है। बस संदर्भ का एक छोटा सा देने के लिए - यहाँ एक मानक वायरल phylogenetics कार्यप्रवाह के लिए एक रूपरेखा है. और, आप जानते हैं, इससे पहले कि मैं इस बारे में बात करूं, वायरल फ़ाइलोजेनेटिक्स यह अध्ययन करने में सक्षम होने के लिए बहुत महत्वपूर्ण है कि वायरस समय के साथ कैसे उत्परिवर्तित हो रहा है। यह कैसे है, एक तरह से, शाखाओं में बंटी? विभिन्न नमूने जो हम दुनिया भर में एकत्र करते हैं, कैसे संबंधित हैं? वायरल आणविक महामारी विज्ञान की दुनिया में एक टन उपयोग है जो इस बात के दायरे से बाहर हैं, लेकिन आम तौर पर, वायरल जीनोम से अनुमानित फाइलोजेनी होना बहुत उपयोगी है। आमतौर पर वर्कफ़्लो इस तरह से शुरू होता है जहां आप असंरेखित वायरल जीनोम अनुक्रमों के एक समूह के साथ शुरू करते हैं, जिसे मैं यहां दिखा रहा हूं। पहला कदम आमतौर पर कई अनुक्रम संरेखण है जहां आप इन अंतरालों को प्रत्येक अनुक्रम के विभिन्न प्रकार के पदों में रखने की कोशिश करते हैं ताकि उन्हें बेहतर लाइन अप मिल सके। ऐसा करने के बाद यह आपको अनुक्रम होमोलॉजी की कुछ

धारणा देता है। फिर, कई अनुक्रम संरेखण को देखते हुए, हम तब इन अनुक्रमों के बीच एक अनियंत्रित विकासवादी संबंध का अनुमान लगाने की कोशिश करने के लिए फ़ाइलोजेनेटिक अनुमान लगा सकते हैं। फिर, आम तौर पर उसके बाद, हम यह निर्धारित करने के लिए रूटिंग कहलाते हैं कि सभी अनुक्रमों का सबसे संभावित सामान्य पूर्वज क्या है। इस तरह का तब हमें बताता है कि इन अनुक्रमों के समय विकासवादी इतिहास में आगे क्या था। फिर, शायद आप कुछ अतिरिक्त डाउनस्ट्रीम विश्लेषण करेंगे। शायद आप ट्रांसमिशन क्लस्टरिंग करते हैं। वहाँ अन्य विश्लेषण है कि आप phylogeny पर और दृश्यों पर कर सकते हैं की एक बहुत कुछ है. लेकिन यह एक तरह का निर्माण खंड है कि आप इन सभी अन्य विश्लेषणों को कैसे करते हैं। तो आम तौर पर यहां पर ये कदम प्रमुख कम्प्यूटेशनल बाधाएं हैं। एकाधिक अनुक्रम संरेखण और फिर फ़ाइलोजेनेटिक अनुमान। आज की बात में, मैं phylogenetic अनुमान मैं सिर्फ कई अनुक्रम संरेखण पर में zooming जा रहा हूँ के बारे में बात नहीं करेंगे.

स्लाइड *3*

तो, कुछ संदर्भ - एकाधिक अनुक्रम संरेखण इसे एनपी-पूर्ण कम्प्यूटेशनल समस्या कहा जाता है। इसका क्या अर्थ है - एक बहुत ही तकनीकी कंप्यूटर विज्ञान शब्द है - लेकिन मूल रूप से इसका मतलब यह है कि कोई बहुपद समय सटीक समाधान नहीं है। असल में यह मुझे दृश्यों का एक गुच्छा दिया और मुझे इष्टतम एकाधिक अनुक्रम संरेखण के साथ आने के लिए कहा. बहुपद समय में ऐसा करने का कोई तरीका नहीं है। यह बहुत बहुत धीमा है। अनुमानित समाधान प्रदान करने के लिए हेरिस्टिक्स विकसित किए गए हैं। उदाहरण के लिए, आपने ClustalOmega, MUSCLE और MAFFT के बारे में सुना होगा। ये कुछ प्रकार के मानक उपकरण हैं जो अंतरिक्ष में उपयोग किए जाते हैं। हालांकि, यहां तक कि ये हेरिस्टिक्स - वे आम तौर पर अनुक्रमों की संख्या के संबंध में चतुर्भुज रूप से स्केल करते हैं। संदर्भ के लिए, GISAID डेटाबेस, जो वह डेटाबेस है जहां अधिकांश लोग अपने पूर्ण SARS-CoV-2 जीनोम को संग्रहीत कर रहे हैं, यह डेटाबेस बहुत तेजी से बढ़ रहा है और आज तक हमारे पास दुनिया भर से 15 मिलियन से अधिक SARS-CoV-2 अनुक्रम उपलब्ध हैं। अगली महामारी वास्तविक समय में जीनोम को अनुक्रमित करने के लिए अधिक और तरह की होने जा रही है। यह एक ऐसा उपकरण बनने जा रहा है जो उम्मीद है कि हम आने वाली वायरल महामारियों में उपयोग करना जारी रखेंगे। हम उम्मीद कर सकते हैं कि यह एक बड़ी डेटा समस्या का और भी महत्वपूर्ण होगा। वर्तमान में ClustalOmega, MAFFT, और MUSCLE जैसे इन उपकरणों के साथ, हम दशकों से सदियों तक के रनटाइम को देख रहे हैं, जो, आप जानते हैं, स्पष्ट कारणों से, अगर हम वास्तविक समय आणविक विश्लेषण करने की कोशिश कर रहे हैं, तो दशकों या सदियों बस थोड़ा बहुत धीमा है। तो हम इसे कैसे गति दे सकते हैं? खैर, यह पता चला है कि समस्या वास्तव में थोड़ा आसान है जो हम हल करने की कोशिश कर रहे हैं। एकाधिक अनुक्रम संरेखण, सामान्य रूप से, अनुक्रमों की कोई समरूपता नहीं मानने की तरह है। यह वह समय है जो पूरी तरह से मनमाने दृश्यों को संरेखित करने में लगता है। लेकिन SARS-CoV-2 और सामान्य रूप से वायरस के साथ हमें बहुत सरल समस्या है, है ना? हमारे पास बहुत सारे अनुक्रम होमोलॉजी हैं। यहां तक कि अगर वायरस उत्परिवर्तित हो रहा है, तो आप जानते हैं, दुनिया भर में, हर एक वायरल अनुक्रम जो हम प्राप्त करते हैं, वह संदर्भ जीन के लगभग समान होने वाला है। यह बिल्कुल समान नहीं होने जा रहा है, लेकिन यह लगभग समान होने जा रहा है। इसलिए हम वास्तव में एक बहुत ही सरल कम्प्यूटेशनल समस्या का सामना कर रहे हैं जो अत्यधिक समान अनुक्रमों के कई अनुक्रम संरेखण है। तो हम इस विश्लेषण को गति देने के लिए उस सुविधा का उपयोग कैसे कर सकते हैं?

स्लाइड *4*

हम वह कर सकते हैं जिसे संरेखित-से-संदर्भ दृष्टिकोण कहा जाता है। इसलिए एक बार में एक-दूसरे के साथ सब कुछ संरेखित करने की कोशिश करने के बजाय, हम जो कर सकते हैं वह एक संदर्भ इकाई के खिलाफ व्यक्तिगत जोड़ी-वार संरेखण है। तो इस आंकड़े में, शीर्ष पर मोटी हरी पट्टी हमारे दायरे 2

जीनोम के संदर्भ का प्रतिनिधित्व करती है, और इनमें से प्रत्येक अन्य रंगीन जीनोम एक अनुक्रम का प्रतिनिधित्व करता है जिसे मैं वास्तविक दुनिया से एकत्र करता हूं। मैं इनमें से प्रत्येक को संदर्भ जीनोम में संरेखित करना चाहता हूं। मैं जो कर सकता था वह एक-एक करके एक-एक करके मैं स्वतंत्र रूप से संदर्भ जीनोम के खिलाफ इन जीनोम अनुक्रमों में से प्रत्येक को संरेखित कर सकता हूं, जो मैं इनमें से प्रत्येक को काफी जल्दी कर सकता हूं और मैं बड़े पैमाने पर समानांतर कर सकता हूं क्योंकि संदर्भ के लिए इनमें से प्रत्येक जोड़ीदार संरेखण पूरी तरह से स्वतंत्र रूप से किया जा सकता है। मैं अपने कंप्यूटर के कई कोर को समानांतर कर सकता हूं, मैं इस समस्या पर कई फेंक सकता हूं। फिर, एक बार जब मैंने संदर्भ के लिए उन सभी जोड़ीदार लाइनों का मुकाबला किया है, तो मैं रैप रिंच जीनोम का उपयोग कर सकता हूं - मैं अपने कई अनुक्रम लाइन के कॉलम बनाने के लिए एंकर के रूप में इसके एंकर, इसकी स्थिति का उपयोग कर सकता हूं। उदाहरण के लिए, शायद मैं संदर्भ जीनोम की पहली स्थिति से शुरू करूँगा और मैं देखूंगा, ठीक है, ठीक है, लाल अनुक्रम में यह वह अक्षर है जो उस स्थिति से जुड़ा हुआ है। नारंगी अनुक्रम में, यह पत्र है। गुलाबी अनुक्रम में, यह पत्र है। नीले अनुक्रम में, यह पत्र है। और मैं उन सभी अक्षरों को अपने एकाधिक अनुक्रम संरेखण के एक कॉलम में मर्ज कर सकता हूं। और मैं अपने संदर्भ जीनोम की दूसरी स्थिति के लिए एक ही काम कर सकता था, तीसरी स्थिति, चौथी स्थिति, सभी तरह से एक ही बात। और स्थिति से स्थिति की तरह मैं अपने एकाधिक अनुक्रम संरेखण का निर्माण कर सकते हैं. यह विचार, यह वास्तव में अच्छा है क्योंकि यह बड़े पैमाने पर समानांतर है और यह चतुर्भुज के बजाय अनुक्रमों की संख्या के साथ रैखिक रूप से मापता है। तो इसमें बहुत बेहतर मापनीयता भी है। क्या हमें इस दृष्टिकोण को खरोंच से लागू करना है? हम वास्तव में नहीं करते हैं।

स्लाइड *5*

यह पता चला है कि यदि आप एक तरह से पीछे हटते हैं और इस समस्या के बारे में सोचते हैं, तो यह वास्तव में एक अर्थ में, लंबे समय तक पढ़ी जाने वाली मैपिंग समस्या के बराबर है। आइए बस एक तरह से पीछे हटें और पुनर्विचार करें कि हम किस समस्या से निपट रहे हैं। हमारा इनपुट एक संदर्भ जीनोम और लंबे अनुक्रमों का एक गुच्छा है जो संदर्भ जीनोम के समान हैं। हमारा आउटपुट संदर्भ जीनोम के खिलाफ उन अनुक्रमों में से प्रत्येक का एक संरेखण है। यह बिल्कुल वही कम्प्यूटेशनल समस्या है जो लंबे समय तक पढ़ती है। पहिया को फिर से शुरू करने के बजाय, हम इन सभी वास्तव में उन्नत तकनीकों का निर्माण कर सकते हैं जो लोगों ने लंबे समय से पढ़ी गई मैपिंग समस्या को हल करने के लिए बनाया है और इसे इस संदर्भ में लागू किया है।

स्लाइड *6*

उस उद्देश्य के लिए, मैंने वायरलएमएसए नामक एक उपकरण विकसित किया है और यह क्या करता है यह इस संदर्भ-निर्देशित एकाधिक अनुक्रम संरेखण को करने के लिए मौजूदा लंबे समय तक पढ़े गए मैपर्स के चारों ओर लपेटता है। यह उन जीनोम में से प्रत्येक का इलाज करता है जिन्हें मैंने लंबे समय तक पढ़ा है और यह संदर्भ जीनोम को संदर्भ जीनोम के रूप में मानता है। यह सिर्फ उस रीड मैपर को कॉल करता है - मैं लचीलेपन को प्रदर्शित करने के लिए कुछ अलग-अलग रीड मैपर्स के खिलाफ लपेटता हूं - लेकिन मैं मुख्य रूप से लोगों को गति और सटीकता दोनों के लिए मिनिमैप 2 का उपयोग करने का सुझाव देता हूं। फिर, उन पढ़ने वाले मैपिंग परिणामों को देखते हुए, मैं तब कर सकता हूं - या मैपिंग परिणाम दिए गए हैं - फिर मैं उन्हें एक एकल एकाधिक अनुक्रम लाइन में संकलित कर सकता हूं।

तो वायरलएमएसए को चलाने के लिए आपको बस इतना करना है कि आप वायरलएमएसए को एक संदर्भ जीनोम और संरेखित करने के लिए अनुक्रमों का एक गुच्छा दें। यह स्वचालित रूप से संदर्भ जीनोम को अनुक्रमित करने का काम संभालेगा, हो सकता है कि यदि आप इसे एक सहायक संख्या देते हैं तो यह संदर्भ जीनोम को डाउनलोड और अनुक्रमित करने का काम करेगा। यह पूर्व प्रसंस्करण और सभी बहाव

सामान के सभी संभाल लेंगे और यह सिर्फ उत्पादन - यह पढ़ा मैपर कॉल करेंगे, यह कई SQL संरेखण में परिणाम विलय कर देंगे, और यह सिर्फ एक एकल मानक फ़ाइल है कि अपने एकाधिक अनुक्रम संरेखण है आउटपुट.

स्लाइड *7*

यह मौजूदा उपकरणों के खिलाफ कैसे करता है? हमने एक बेंचमार्क प्रयोग किया जहां हमने वायरलएमएसए के रनटाइम की तुलना में मिनिमैप 2 के चारों ओर रैपिंग की तुलना की, जो कि संदर्भ दृष्टिकोण के लिए एक मौजूदा संरेखण है, लेकिन यह संदर्भ के लिए गठबंधन किए गए खरोंच से अपने स्वयं के उपकरणों की तरह है। हमने MAFFT के खिलाफ भी तुलना की, जिसे आमतौर पर सबसे अधिक इस्तेमाल किए जाने वाले कई अनुक्रम खंड उपकरणों में से एक माना जाता है। इस प्लॉट में, क्षैतिज अक्ष पर मेरे पास अनुक्रमों की संख्या है। ऊर्ध्वाधर अक्ष पर मेरे पास सेकंड में कुल निष्पादन समय है। यह पूर्ण SARS-CoV-2 जीनोम अनुक्रमों पर किया गया था, इसलिए जीनोम की लंबाई लगभग 29,000 थी। जैसा कि हम देख सकते हैं, नीली रेखा, जो वायरलएमएसए है, मौजूदा उपकरणों की तुलना में तेजी से परिमाण के आदेश हैं। VIRULIGN की तुलना में, जो रैखिक रूप से भी स्केलिंग कर रहा है, हम प्राप्त कर रहे हैं - और वैसे, यह प्लॉट एक लॉग स्केल प्लॉट है - इसलिए VIRULIGN की तुलना में, हम लगभग एक हजार गुना तेज-ईश की तरह हैं। और MAFFT के साथ, हम उतने तेज़ नहीं हैं, लेकिन आप देख सकते हैं कि क्योंकि MAFFT चतुष्कोणीय रूप से बढ़ता है, MAFFT के संबंध में हमारी गति वास्तव में समय बढ़ने के साथ बढ़ रही है। यहां तक कि सिर्फ एक हजार दृश्यों पर भी। हम लगभग एक हजार गुना तेजी से हिट करते हैं और यह अंतर बढ़ जाता है।

स्लाइड *8*

अब, आप सोच रहे होंगे, ठीक है, ठीक है, उपवास अच्छा है लेकिन क्या बात है अगर यह मुझे अच्छे संरेखण नहीं देता है? हमने सटीकता की तुलना भी की। हमने जो किया वह हमने एमएएफएफटी द्वारा गणना किए गए कई अनुक्रम संरेखण को लिया, एचआईवी, इबोला से हाथ से क्यूरेटेड संरेखण के एक समूह पर वायरलएमएसए द्वारा गणना किए गए कई अनुक्रम संरेखण को लिया, और मैं तीसरे वायरस पर खाली कर रहा हूं, लेकिन मूल रूप से हमने वायरस लिया कि हमने लॉस अलामो से संरेखण को क्यूरेट किया था- ओह वास्तव में, नहीं - यह साजिश सिर्फ एचआईवी -1 से है। लॉस एलामोस नेशनल लैब से हमने उनके क्यूरेटेड कई अनुक्रम संरेखण लिए और हम इसे जमीनी सच्चाई के रूप में उपयोग करते हैं। फिर हमने देखा कि वायरलएमएसए में मैप किए गए जमीनी सच्चाई के खिलाफ तुलना कैसे करते हैं, कई अनुक्रम संरेखण को क्यूरेट किया जाता है। यदि हम अपने संरेखण में प्राप्त अनुक्रमों की जोड़ीवार दूरी की गणना करते हैं, और फिर हम सटीकता के लिए एक मेंटल परीक्षण करते हैं - हम अपनी जोड़ीदार दूरियों के बीच संबंध पाते हैं, और जोड़ीवार की गणना सीधे सही एकाधिक अनुक्रम संरेखण से की जाती है, तो हम देखते हैं कि सहसंबंध नगण्य रूप से भिन्न है। यहां, आप जानते हैं, हमें जोड़ीदार दूरी गणना के लिए 0.997 की तुलना में वायरल एमएसए के लिए 0.994 जैसा सहसंबंध गुणांक मिलता है। दरअसल, जब हमने फाइलोजेनियों की गणना की, तो वायरलएमएसए कई अनुक्रम तत्वों का उपयोग करके अनुमान लगाया गया फ़ाइलोजेनीज़ वास्तव में एमएएफएफटी से अनुमानित लोगों की तुलना में थोड़ा अधिक टोपोलॉजिकल सटीकता है। तो, नगण्य रूप से, लेकिन फिर भी हम जो दिखा रहे हैं वह यह है कि ये सभी इरादों और उद्देश्यों के लिए सटीकता के मामले में अनिवार्य रूप से समकक्ष हैं।

स्लाइड *9*

निष्कर्ष - वायरलएमएसए एक उपकरण है जो अल्ट्रा बड़े वायरल डेटासेट के तेजी से कई अनुक्रम संरेखण को सक्षम बनाता है। यह खुला स्रोत है, आप इसे GitHub पर पा सकते हैं, और आप जानते हैं, कृपया इसे अपने वायरल विश्लेषणों में उपयोग करने पर विचार करें।

स्लाइड *10*

और आभार - मैं हेंग ली को धन्यवाद देना चाहता हूं, वह मिनिमैप 2 के डेवलपर हैं और यह वास्तव में मिनिमैप 2 को विकसित करने में उनकी विशेषज्ञता है जो वायरलएमएसए की गति और प्रदर्शन को सक्षम बनाता है। मैं इस परियोजना का समर्थन करने वाले अनुदान के लिए एनएसएफ को धन्यवाद देना चाहता हूं। और Google क्लाउड प्लेटफ़ॉर्म अनुसंधान क्रेडिट का उपयोग करके अनुसंधान का भी समर्थन किया गया था।

स्लाइड *11*

इसलिए मैं किसी भी प्रश्न के लिए समय बचाऊंगा या मुझे इसे यहां समाप्त करने में खुशी होगी।