BKAP-582
[1-3]

RESEARCH ARTICLE Bhartiya Krishi Anusandhan Patrika

सी.एन.वी. पहचान हेतु डीप लर्निंग आधारित कार्यप्रणाली

नीतेश कुमार शर्मा[1], उमा[1], एम.ए. इकबाल[1], सारिका जायसवाल[1], अनिल राय[1], दिनेश कुमार[1] 10.18805/BKAP582

सारांश

पृष्ठभूमि: कॉपी नंबर वैरिएंटस (सी.एन.वी. CNVs) से हमें जैनेटिक वेरिएशन के बारे में पता चलता है। चूंकि कई सी.एन.वी. में डिफ्रेंशियली एक्सप्रेस्ड जीन शामिल होते हैं, जिनके कारण सामान्य फेनोटाइपिक वेरिएशन जाना जा सकता है। वर्तमान प्रयासों को सी.एन.वी.के अधिक व्यापक लक्षण वर्णन की ओर केन्द्रित किया गया है जिससे यह निर्धारित करने में मदद मिल सके कि जीनोमिक विविधता मानव के साथ–साथ पौधों में जैविक कार्य, विकास और सामान्य बीमारियों को कैसे प्रभावित करती है।

विधियाँ: नेक्स्ट–जेनरेशन सीक्वेंसिंग (एन.जी.एस.) में विश्लेषणात्मक वेरिएबिलिटी, कवरेज डेटा में आर्टिफैक्ट्स और सी.एन.वी. पहचान में जैव सूचना विज्ञान उपकरणों के आभाव ने सी.एन.वी.पहचान करने के लिए टार्गेटेड एन.जी.एस डेटा की उपयोगिता को सीमित कर दिया है। साहित्य में डीप लर्निंग आधारित पाइपलाइन का विविरण है, जिसमें टार्गेटेड एन.जी.एस डेटा से सी. एन.वी.की पहचान करने की मशीन लर्निंग कंपोनेंट शामिल है।

परिणाम: यह माना जाता है कि क्लीनिकल "गोल्ड स्टैंडर्ड" (जैसे एफ.आई.एस.एच. FISH) के आंकड़ों के साथ, सी.एन.वी. पहचान अधिक सटीक हो सकता है। इससे मौजूदा एन.जी.एस. विधियों के पूरक के रूप में शोध को नई दिशा मिलेगी।

शब्दकुंजी: कॉपी नंबर वेरिएशन, नेक्स्ट–जेनरेशन सीक्वेंसिंग, डीप लर्निंग, टारगेट सीक्वेंसिंग।

CNV Deep Learning based Methodology for Recognition

Nitesh Kumar Sharma[1], Uma[1], Mir Asif Iquebal[1], Sarika Jaiswal[1], Anil Rai[1], Dinesh Kumar[1]

ABSTRACT

Background: Copy number variants (CNVs) account for a significant amount of genetic variation. Since many CNVs include genes that result in differential levels of gene expression, substantial normal phenotypic variation can be explained. Current efforts are directed toward a more comprehensive characterization of CNVs that will provide the basis for determining how genomic diversity impacts biological function, evolution and common diseases in human as well as plants.

Methods: The analytical variability in next generation sequencing (NGS) and artifacts in coverage data along with lack of robust bioinformatics tools for CNV detection have limited the utility of targeted NGS data to identify CNVs. Literature has the evidence of development of deep learning-based pipeline that incorporates a machine learning component to identify CNVs from targeted NGS data.

Result: It is believed that combining this with clinical "gold standard" (*e.g.* FISH) information, the CNV detection could be more accurate. This would lead to a new research direction, supplementing the existing NGS methods.

Key words: Convolutional neural network, Copy number variation, Deep learning, Next generation sequencing, Target sequencing.

प्रस्तावना

कॉपी नंबर वेरिएंट (सी.एन.वी.) कॉपी नंबर चेंज का प्रतिनिधित्व करता है जिसमें डी.एन.ए. का भाग शामिल होता है जो एक किलोबेस (केबी) या उससे बड़ा होता है। अधिकांश सी.एन.वी. उच्च सीक्वेंस सिमिलेरिटी की लो–कॉपी रिपीट्स के बीच गैर–एलीलिक समरूप रिकोंबिनेशन का परिणाम हैं (Liu *et al.*, 2012)। सी.एन.वी. की पहचान अब अनेक आनुवंशिक विकारों के रोगजनन में महत्वपूर्ण भूमिका निभाने के रूप में हो रही है और यह बौद्धिक असमर्थता का एक सामान्य कारण भी है

[1]Agricultural Bioinformatics Centre, ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110 012, India.

Corresponding Author: Sarika Jaiswal, Agricultural Bioinformatics Centre, ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110 012, India. Email: aijaiswal@gmail.com

How to cite this article: Sharma, N.K., Uma, Iquebal, M.A., Jaiswal, S., Rai, A. and Kumar, D. (2023). CNV Deep Learning based Methodology for Recognition. Bhartiya Krishi Anusandhan Patrika. doi: 10.18805/BKAP582.

Submitted: 25-08-2022 **Accepted:** 05-01-2023 **Online:** 19-04-2023

(Coe *et al.*, 2012)। आजकल, क्लीनिकल सेटिंग्स में सी.एन.वी. का पता लगाने के लिए "गोल्ड स्टैंडर्ड" का तरीके का साथ–साथ साधारणतः उपयोग किए जाने वाला तुलनात्मक जीनोमिक संकरण भी है (Freeman *et al.*, 2006)। दोनों पारंपरिक तरीकों से केवल पूरी जीन रिपीट्स [(बड़ी रिपीट्स > 40kb)] का पता लगाया जा सकता है। पिछले एक दशक में छोटी रिपीट्स (<40 kb) का पता लगाने के लिए, एन.जी.एस. आधारित तरीकों को लागू किया गया है। ज़्यादातर चार सीक्वेंस–आधारित विधियाँ सी.एन.वी.का पता लगाने के लिए उपयोग की जाती हैं: (1) रीड डेप्थ (आर.डी.) विधि, जो कॉपियों के लाभ और हानि का अनुमान लगाने के लिए नोर्मलाइज्ड रीड डेप्थ में परिवर्तन पर निर्भर करती है (2) रीड पेयर (आर.पी.) विधि, जो असंगत रूप से मैप किए गए रीड पेयर पर आधारित है (3) स्पलिट रीड (एस.आर.) विधि, जो गैप्ड रीड एलाइनमेंट का उपयोग करती है और (4) असेंबली (ए.एस.) विधि, जिससे कॉन्टिग्स/स्कैफोल्ड प्राप्त होते हैं, जिनकी स्ट्रक्चरल वेरिएशन की खोज के लिए संदर्भ जीनोम के साथ तुलना की जाती है (Teo *et al.*, 2012)। उच्च स्तर पर, एन.जी.एस. डेटा से कॉपी नंबर वैरिएंटस का पता लगाना मशीन लर्निंग में क्लासिफिकेशन प्रॉबलम के रूप में माना जा सकता है। सीक्वेंसिंग डेटा के साथ डीप लर्निंग का उपयोग करके सी.एन.वी. जैसी स्ट्रक्चरल वेरिएशन का पता लगाना एक नई शोध की ओर ले जाता है। हाल ही में, गूगल के डीपवैरिएंट (Poplin *et al.*, 2016) को सीक्वेंसिंग डेटा से एस.एन.पी. और इंडेल्स को कॉल करने के लिए विकसित किया गया था। यह सीक्वेंसिंग डेटा प्रोसेसिंग डोमेन में डीप लर्निंग की क्षमता को प्रदर्शित करता है। डीपवैरिएंट तरीका एक स्वाभाविक शोध प्रश्न उठाता है: क्या सी.एन.वी.जैसे अन्य प्रकार के जेनिटिक वेरिएशन को सीक्वेंसिंग डेटा से कॉल करने के लिए डीप लर्निंग लागू किया जा सकता है जो एस.एन.पी. और शॉर्ट इंडल्स से अधिक जटिल हैं? अपने अध्ययन में, Zhang *et al.* (2019) इस प्रश्न का एक सकारात्मक उत्तर प्रदान करते हैं: वे दिखाते हैं कि सीक्वेंसिंग डेटा से कॉपी नंबर वैरिएंटस (सी.एन.वी.) को सटीक रूप से कॉल करने के लिए डीप लर्निंग का उपयोग किया जा सकता है।

सी.एन.वी. कॉल करने हेतु डीप लर्निंग का कार्यान्वयन

सैंपल स्रोत एवं डेटासेट की तैयारी

हम यहाँ Zhang *et al.* (2019) के अध्ययन से सी.एन.वी. पहचान में डीप लर्निंग की उपयोगिता को समझेंगे। Zhang *et al.* (2019) ने Illumina Nextseq500 प्लेटफॉर्म के साथ ERBB2+MET पूर्ण –एक्सॉन डिजाइन किए गए एन.जी.एस. पैनल सीक्वेंसिंग के माध्यम से 1301 ERBB2 और 1148 MET सैंपल डेटा एकत्र किया। ERBB2 के सी.एन.वी. की आवृत्ति लगभग 16% एवं MET के सी.एन.वी. की आवृत्ति लगभग 2% थी। एन.जी.एस. पैनल पर आधारित आयनकॉपी टूल के उपयोग से सी.एन.वी. सेंपिलों को पॉसिटिव और नेगेटिव रूप में लेबल किया गया। एफ.आई.एस.एच. (FISH) परिणामों के संयोजन में, सी.एन.वी. पॉजिटिव सैंपल के टारगेट क्षेत्र में लगभग हमेशा गेन सिग्नल होता है, इसलिए ऐसे अनुमानित सी.एन.वी. को पॉजिटिव सैंपल माना गया। अंत में, क्रमशः 272 ERBB2 और 63 MET सी.एन.वी.–पॉजिटिव सैंपल और 1029 ERBB2 और 1085 MET सी.एन.वी को नकारात्मक सैंपल के तौर पर एकत्र किए गए।

कवरेज डेप्थ कैलकुलेशन और नोर्मेलाइजेशन

इस विधि में सर्वप्रथम विशिष्ट जीन या टारगेट रीजन के रीड नंबरों की मैट्रिक्स तैयार करना महत्वपूर्ण कार्य है। इसलिए, संदर्भ एक्सॉन को 40 बी.पी. की एक स्ट्राइड के साथ कई 50 बी.पी. विंडो में विभाजित किया जाता है, जिसमें साथ की विंडो की प्रत्येक जोड़ी के बीच 10 बी.पी. ओवरलैप होता है। एक्सॉन सीक्वेंस के पूरे बिन से मेल खाने वाले रीड्स को प्रत्येक बिन पर रीड नंबर के रूप में गिना जाता है। रीड्स नंबर से युक्त एक्सॉन को दर्शाने वाली पंक्तियों को ट्रेनिंग के लिए इनपुट मैट्रिक्स बनाने के लिए एकत्रित किया जाता है। रीड नंबर डेटा या मैट्रिस में मौजूद दवपेम के प्रभाव को कम करने के लिए रीड नंबर मैट्रिसेस को नॉर्मलाइज किया गया।

फ्रेमवर्क डिजाइन

सी.एन.वी. कॉलिंग एल्गोरिथम को लागू करने के लिए एक डीपलर्निंग फ्रेमवर्क TensorFlow का उपयोग किया गया है। मॉडल, LeNet–5 जिसका उपयोग MNIST डेटा (LeCun *et al.*, 1998) को वर्गीकृत करने के लिए किया गया था, उसी मॉडल का उपयोग सी.एन.वी. सेंपिलों को वर्गीकृत करने के लिए किया गया है। इसमें एक convolutional लेयर, एक मैक्स पूलिंग लेयर, एक ड्रॉपआउट लेयर और दो पूरी तरह से कनेक्टेड लेयर्स होते हैं। कर्नेल या फीचर डिटेक्टर का आकार $n \times n$ पर सेट है जहां कनवल्शन में n का मान $\{3,5,7,9,11,13,15,17,19,21\}$ और पूलिंग प्रक्रिया में $n = 2$ हो सकता है। पहली, पूरी तरह से जुड़ी हुई लेयर में 1024 न्यूरॉन्स होते हैं और शेष दूसरी लेयर में 2 न्यूरॉन्स होते हैं, जो आउटपुट लेयर के रूप में यह वर्गीकृत करता है कि सैंपल सी.एन.वी.–पॉजिटिव है या नहीं। ड्रॉपआउट मान को $lr \in \{0.3, 0.$

5, 0.7, 0.9, 1.0} पर सेट किया गया था, और लेयर को दो पूरी तरह से जुड़ी लेयर्स के बीच रखा गया। Adam एल्गोरिथ्म का उपयोग ग्रेडिएंट–डिसेन्ट–ओप्टिमाइजेशन विधि के रूप में किया गया है। लर्निंग रेट 1×10–n के रूप में निर्धारित किया गया है जहाँ n का मान {4, 5, 6, 7} हो सकता है। मॉडल को 30 सैम्पलों के बैच साइज के साथ 10,000 epochs के लिए प्रशिक्षित किया गया।

प्रशिक्षण और क्रॉस वेलीडेशन

मॉडल की क्षमता को जाँचने के लिए, 272 एन.जी.एस.–पॉजिटिव और 1029 एन.जी.एस.–नेगेटिव ERBB2 सैंपल एवं 63 एन.जी.एस.–पॉजिटिव और 1085 एन.जी.एस.–नेगेटिव MET सैंपल मूल्यांकन हेतु लिए गए। ERBB2 CNN मॉडल के प्रशिक्षण के लिए 223 एन.जी.एस.–पॉजिटिव और 817 एन.जी.एस.–नेगेटिव ERBB2 सैंपल और MET सी.एन.एन. मॉडल के लिए 51 एन.जी.एस.–पॉजिटिव और 867 एन.जी.एस.–नेगेटिव MET सैंपल का इस्तेमाल किया गया है। ERBB2 सी.एन.एन. मॉडल और MET सी.एन.एन. मॉडल के प्रशिक्षण के लिए क्रमशः 10 फोल्ड और 5 फोल्ड क्रॉस–वेलीडेशन किया गया।

परिणाम

Zhang इत्यादि ने अपने इस अध्ययन में सिंगल सैंपल से सी.एन.वी. को कॉल करने के लिए एक डीप लर्निंग पाइपलाइन पेश किया एवं इसकी क्षमता को अन्य विधियों की क्षमता के साथ तुलना की। Zhang *et al.* (2019) द्वारा एक नवीन डीप लर्निंग आधारित विधि विकसित की गई है। दो इंडिपेंडेंट मॉडलों को क्रमशः दो सैंपल डेटासेट, ERBB2 और MET पर प्रशिक्षित और परीक्षण किया गया है। नतीजतन, सी.एन.वी. को कॉल करने में सी.एन.एन. अल्गोरिथम अन्य मौजूद मॉडलों के मुकाबले ज्यादा दक्ष पाया गया। प्रशिक्षित सी.एन.एन. मॉडल एक मैट्रिक्स से लोकल फीचर्स को समझ सकता है और उन्हें विभिन्न सैंपलों से अलग–अलग मैट्रिक्स में जेनिरलाइज कर सकता है। मॉडलों की क्षमता का मूल्यांकन करने के लिए एक्यूरेसी, रिकॉल, प्रेसिजन और स्पेसिफिटी जैसे मापों का प्रयोग किया गया। इस कार्यप्रणाली में अन्य विधियों से अधिक एक्यूरेसी पाई गई।

निष्कर्ष

यह सर्वविदित है कि एक डिलिशन या डुप्लिकेशन कई लगातार विंडोज में रीड नंबर में कमी या वृद्धि के रूप में स्पष्ट होती है। सी.एन.वी. डिटेक्शन की यह कार्यप्रणाली सी.एन.वी. की गोल्ड स्टैंडर्ड विधि (FISH) के साथ मिलकर और भी दक्ष होकर अधिक सटीक रूप से कॉल करने में सक्षम हो सकती है। एन.जी.एस. विधियों और डीप लर्निंग विधि के बीच समरूपता दर के आधार पर, जो बहुत अधिक है, यह निष्कर्ष निकाला गया है कि रुचि की जीन को कवर करने वाले रीड में कॉपी नंबर वैरिएशंस का पता लगाने के लिए लगभग पर्याप्त जानकारी होती है। इसलिए, डीप लर्निंग पाइपलाइन कॉपी नंबर वैरिएशंस का पता लगाने के लिए मौजूदा एन.जी.एस. विधियों के लिए एक बेहतर पूरक हो सकती है। अंततः यह निष्कर्ष निकाला गया है कि वर्णित विधि से सी.एन.वी., विशेष रूप से छोटे रिपीट्स का पता लगाने में अन्य विधियों से बेहतर प्रदर्शन करती है और डीप लर्निंग पाइपलाइन मौजूद अन्य सी.एन.वी. पहचान के टूल्स जैसे एफ.आर.ई.ई.सी और सी.एन.वी.नेटर से बेहतर है।

संदर्भ

Coe, B.P., Girirajan, S. and Eichler, E.E. (2012). The genetic variability and commonality of neurodevelopmental disease. In American Journal of Medical Genetics Part C: Seminars in Medical Genetics. 160(2): 118-129.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M. and Carter, N.P. (2006). Copy number variation: New insights in genome diversity. Genome Research. 16(8): 949-961.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86(11): 2278-2324.

Liu, P., Carvalho, C.M., Hastings, P.J. and Lupski, J.R. (2012). Mechanisms for recurrent and complex human genomic rearrangements. Current Opinion in Genetics and Development. 22(3): 211-220.

Poplin, R., Newburger, D., Dijamco, J., Nguyen, N., Loy, D., Gross, S.S. and DePristo, M.A. (2016). Creating a universal SNP and small indel variant caller with deep neural networks. BioRxiv. 092890.

Teo, S.M., Pawitan, Y., Ku, C.S., Chia, K.S. and Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 28(21): 2711-2718.

Zhang, Y.X., Jin, L.C., Wang, B., Hu, D., Wang, L.Q., Li, P. and Yang, J. (2019). DL-CNV: A deep learning method for identifying copy number variations based on next generation target sequencing. Mathematical Biosciences and Engineering. 17(1): 202-215.