Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK)
Stiftung des öffentlichen Rechts

KI entschlüsselt neuen Genregulationscode in Pflanzen und macht genaue Vorhersagen für neu sequenzierte Genome

Pressemitteilung

Wissenschaftlicher Kontakt
Dr. Jedrzej Jakub Szymanski
Tel.: +49 39482 5753
szymanski@ipk-gatersleben.de

Medienkontakt
Christian Schafmeister
Tel.: +49 39482 5461
schafmeister@ipk-gatersleben.de

Gatersleben, 25.04.2024 **Die Aufklärung der Beziehung zwischen Sequenzen von regulatorischen Elementen und ihren Zielgenen ist der Schlüssel für das Verständnis der Genregulation und ihrer Variation zwischen Pflanzenarten und Ökotypen. Ein Forschungsteam unter Führung des IPK Leibniz-Instituts und mit Beteiligung des Forschungszentrums Jülich hat jetzt „Deep-Learning"-Modelle entwickelt, die Gensequenzdaten mit der mRNA-Kopienzahl für mehrere Pflanzenarten verknüpfen und die regulatorische Wirkung von Gensequenzvariationen vorhersagen. Die Ergebnisse wurden in der Zeitschrift „Nature Communications" veröffentlicht.**

Mit der Sequenzierungstechnologie ist es heute möglich, jährlich Tausende neue Pflanzengenome zu entschlüsseln. Forscherinnen und Forscher verbinden diese genomischen Informationen mit Daten zu verschiedenen Pflanzenmerkmalen. Ziel ist es, Korrelationen zwischen genetischen Varianten und Pflanzenmerkmalen wie der Anzahl der Samen, der Resistenz gegen Pilzbefall, der Fruchtfarbe oder dem Geschmack zu ermitteln. Bisher fehlt allerdings das umfassende Verständnis dafür, wie genetische Variation die Genaktivität auf molekularer Ebene beeinflusst. Dies schränkt die Möglichkeiten ein, „intelligente Nutzpflanzen" mit verbesserter Qualität und geringeren negativen Auswirkungen auf die Umwelt zu züchten.

Forschern des IPK Leibniz-Instituts und des Forschungszentrums Jülich (FZ) ist jetzt ein Durchbruch bei der Bewältigung dieser Herausforderung gelungen. Unter der Leitung von Dr. Jedrzej Jakub Szymanski trainierte das internationale Forscherteam interpretierbare „Deep-Learning"-Modelle, eine Untergruppe von KI-Algorithmen, auf einem riesigen Datensatz mit genomischen Informationen aus verschiedenen Pflanzenarten. „Diese Modelle waren nicht nur in der Lage, die Genaktivität anhand von Sequenzen genau vorherzusagen, sondern auch festzustellen, welche Sequenzteile diese Vorhersagen ermöglichen", erklärt der Leiter der IPK-Arbeitsgruppe „Netzwerkanalyse und Modellierung". Die von den Forschern angewandte KI-Technologie ist vergleichbar mit derjenigen, die im Bereich des Computersehens eingesetzt wird, wo es darum geht, Gesichtszüge in Bildern zu erkennen und auf Emotionen zu schließen.

Im Gegensatz zu früheren Ansätzen, die auf statistischer Anreicherung beruhen, entwickelten die Forscher hier ein mathematisches Modell, das anhand von genomischen Sequenzmerkmalen die mRNA-Kopienzahl voraussagen kann. Das Modell berücksichtigt die Struktur des Genmodells und die Sequenzhomologie, also die Genevolution.

„Wir waren wirklich erstaunt über die Effektivität. Innerhalb weniger Tage Training haben wir viele bekannte regulatorische Sequenzen wiederentdeckt und festgestellt, dass etwa 50 Prozent der identifizierten Sequenzmerkmale völlig neu waren. Die Modelle ließen sich sogar hervorragend auf Pflanzenarten anwenden, für die sie nicht trainiert wurden. Das

macht sie für die Analyse neu sequenzierter Genome so wertvoll“, sagt Dr. Jedrzej Jakub Szymanski.

„Wir haben speziell ihre Anwendung für verschiedene Tomatensorten mit sogenannten ‚Long-Read-Sequenzdaten‘ getestet. Dabei konnten wir spezifische regulatorische Sequenzvariationen identifizieren, die die beobachteten Unterschiede in der Genaktivität und folglich auch der Form, Farbe und Robustheit der Pflanzen erklären. Und dies ist eine bemerkenswerte Verbesserung gegenüber den klassischen statistischen Assoziationen von Einzelnukleotid-Polymorphismen, bei denen es nur um einzelne DNA-Basen geht.“

Das Team hat seine Modelle öffentlich zugänglich gemacht und eine Webschnittstelle für die Nutzung bereitgestellt. „Um zu optimistische Ergebnisse zu vermeiden, die darauf zurückzuführen sind, dass die KI Abkürzungen findet, mussten wir tief in die Biologie der Genregulation eintauchen, um mögliche Verzerrungen zu beseitigen und Datenverluste und Überanpassungen zu reduzieren“, erläutert Fritz Forbang Peleke, leitender Forscher für maschinelles Lernen und Erstautor der Studie, die in der Zeitschrift „Nature Communications“ veröffentlicht wurde.

Dr. Simon Zumkeller, Mitautor und Evolutionsbiologe am FZ Jülich, sagt: „Die von uns vorgestellten Analyseansätze bieten Möglichkeiten, die Genregulation in Pflanzen besser und sogar auf evolutionärer Ebene zu untersuchen. Auch für die praktische Anwendung gibt es mit der von uns beschriebenen Methode eine neue Basis. Mit ihr nähern wir uns der routinemäßigen Identifizierung regulatorischer Genelemente in bekannten und neu sequenzierten Genomen, in verschiedenen Geweben und unter verschiedenen Umweltbedingungen.“

Originalpublikation:

Peleke *et al.* (2024): Deep learning the cis-regulatory code for gene expression in selected model plants. Nature Communications. DOI: 10.1038/s41467-024-47744-0

Abbildung: