

Algorytmy dla edytowalnych grafów pangénomowych

Współczesna biologia coraz częściej wymaga zrozumienia pełnej różnorodności genetycznej w obrębie gatunku. Tradycyjnie naukowcy wykorzystywali pojedynczy *genom referencyjny* — reprezentatywną sekwencję DNA — do badań genetycznych. Jednak takie podejście może pomijać istotne warianty występujące u różnych osobników. Na przykład niektóre fragmenty DNA mogą być całkowicie nieobecne w genomie referencyjnym, co prowadzi do zniekształconych lub niepełnych wniosków.

Aby lepiej uchwycić tę różnorodność, badacze opracowali koncepcję *pangenomu*, który obejmuje wszystkie sekwencje genetyczne występujące u wielu osobników tego samego gatunku. Zamiast jednej liniowej sekwencji, pangenom reprezentuje wszystkie warianty, unikalne geny i różnice strukturalne, dostarczając bogatszego i dokładniejszego źródła informacji do badań genetycznych.

Skutecznym sposobem reprezentowania pangenomów są grafy wariantów. W tym modelu każdy wierzchołek zawiera fragment sekwencji DNA, a ścieżki w grafie odpowiadają indywidualnym genomom. Podejście to pozwala na uliniwienie wspólnych sekwencji, redukując redundancję, a przy tym naturalnie oddaje złożone wariacje, takie jak insercje, delecje i translokacje. Dodatkowo zapewnia wspólny system współrzędnych, który umożliwia precyzyjne porównania genomów i wspiera zastosowania takie jak mapowanie odczytów DNA czy wykrywanie wariantów.

Określenie, które fragmenty DNA powinny być współdzielone między genomami, nie jest trywialnym zadaniem. Istniejące metody zazwyczaj polegają na uliniwieniu między wszystkimi parami genomów oraz heurystycznym wykrywaniu homologii, co wprowadza niejednoznaczność do struktury powstającego grafu. Ponadto większość narzędzi buduje cały graf jednorazowo na podstawie ustalonego zestawu genomów, co utrudnia włączanie nowych danych lub poprawianie błędów. Dodanie lub usunięcie genomów wymaga przebudowy całego grafu od podstaw, co jest kosztowne obliczeniowo i nieefektywne w przypadku dużych lub zmieniających się zbiorów danych.

W poprzednich pracach przedstawiliśmy metodę konstrukcji unikalnej i matematycznie zdefiniowanej struktury grafu wariantów opartej na składzie k -merów genomów wejściowych, eliminującą niejednoznaczności typowe dla metod opartych na uliniwieniach.

W tym projekcie zamierzamy rozszerzyć tę metodę, aby wspierała dynamiczną aktualizację, poprzez opracowanie algorytmów dla edytowalnych grafów wariantów — grafów, które mogą być modyfikowane bez utraty gwarancji matematycznych własności. Celem jest umożliwienie wstawiania i usuwania sekwencji genomowych, jak również łączenia osobno skonstruowanych grafów, przy zachowaniu jednoznacznej i spójnej struktury grafu, niezależnie od kolejności wykonywanych operacji.

Nasze podejście będzie oparte na ścisłych podstawach teoretycznych, które mają na celu zapewnienie, że struktura grafu pozostanie unikalna i spójna, spełniając założenia konstrukcji, pomimo kolejnych modyfikacji. Dzięki wsparciu dla inkrementalnych aktualizacji możliwe stanie się równoległe konstruowanie grafów oraz stopniowe dodawanie nowych danych bez kosztownej przebudowy.

Ta innowacja uczyni grafy pangénomowe bardziej praktycznymi i skalowanymi w rzeczywistych zastosowaniach, zwłaszcza w dużych inicjatywach takich jak Human Pangenome Reference Consortium, które dąży do uchwycenia różnorodności genetycznej setek ludzkich genomów. Skorzysta na niej również genomika bakterii, gdzie precyzyjne odwzorowanie różnorodności genomowej może pomóc w badaniu mechanizmów odporności szczepów na antybiotyki oraz w śledzeniu rozprzestrzeniania się cech oporności. W hodowli zwierząt zaś umożliwi wykorzystanie zaawansowanych metod selekcji genetycznej, które pozwolą na efektywniejsze i bardziej precyzyjne kształtowanie pożądanych cech użytkowych. Ponadto, każda dziedzina wymagająca dokładnego i kompleksowego porównywania genomów zyska na elastyczności i wydajności oferowanych narzędzi.