Genomowe bazy danych

Mateusz Oleszek, nr 144608

Genomowe bazy danych odgrywają kluczową rolę w dziedzinie genetyki, zapewniając badaczom ogromne repozytoria informacji genetycznych do analizy i eksploracji. Bazy te są kompleksowymi zbiorami danych genomicznych, w tym sekwencji DNA, adnotacji genów, wariantów genetycznych i innych związnych informacji. Służą one jako cenne zasoby do badania struktury, funkcji i ewolucji genomów różnych gatunków. Genomowe bazy danych przechowują dane z różnych źródeł, takich jak projekty sekwencjonowania genomów, badania naukowe i wkład publiki. Umożliwiają one naukowcom dostęp i analizę danych genetycznych na dużą skalę, ułatwiając odkrywanie genów związanych z chorobami, identyfikację funkcji genów i badanie relacji genetycznych między różnymi organizmami.

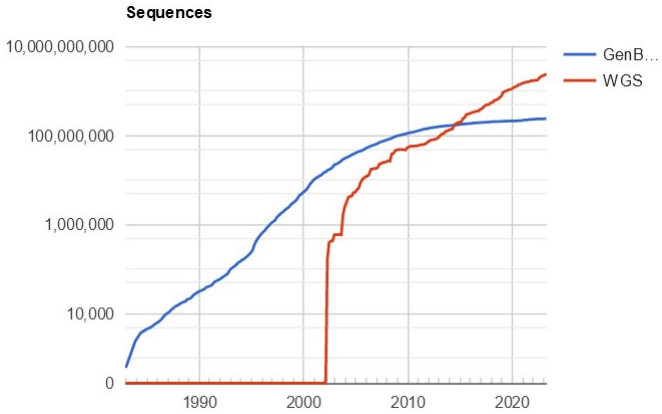
Baza danych sekwencji Los Alamos, założona w 1979 roku i później znana jako GenBank, służyła jako wczesne repozytorium sekwencji biologicznych. W 1982 roku GenBank został przeniesiony do National Center for Biotechnology Information (NCBI, część Narodowego Instytutu Zdrowia USA), gdzie obecnie się znajduje. Pod koniec 1983 roku GenBank przechowywał ponad 2000 sekwencji z prawie 1 milionem par zasad. Jednocześnie NCBI, EMBL i DDBJ utworzyły Międzynarodową Współpracę Baz Danych Sekwencji Nukleotydów (INSDC) w celu zarządzania rosnącą ilością danych sekwencji nukleotydów i aminokwasów. Bazy danych INSDC odnotowały wykładniczy wzrost, mieszcząc obecnie ponad 95 miliardów par zasad.

Aby poradzić sobie z ogromną ilością surowych danych sekwencyjnych, stworzono wyspecjalizowane bazy danych, w tym przeglądarki genomów, bazy danych organizmów modelowych i bazy danych specyficzne dla procesów. Liczba genomicznych baz danych i narzędzi znacznie wzrosła. Szacuje się, że istnieje około 3000 różnych zasobów genomicznych, narzędzi i baz danych dostępnych online.

Repozytoria danych sekwencji

International Nucleotide Sequence Database Collaboration (INSDC) obejmuje GenBank, EMBL i DDBJ, które przyjmują zgłoszenia sekwencji od naukowców z całego świata. Każde zgłoszenie otrzymuje unikalny numer identyfikacyjny i jest przechowywane w bibliotece. Jednak wykładniczy wzrost ilości danych stanowi wyzwanie w utrzymaniu dokładności i dostępności. Powszechnymi problemami są duplikacja oraz niekompletne lub nieprawidłowe informacje. Wysiłki takie jak RefSeq mają na celu selekcjonowanie i poprawianie danych sekwencji, ale ogromna ilość danych utrudnia dodawanie adnotacji i organizowanie każdego wpisu.

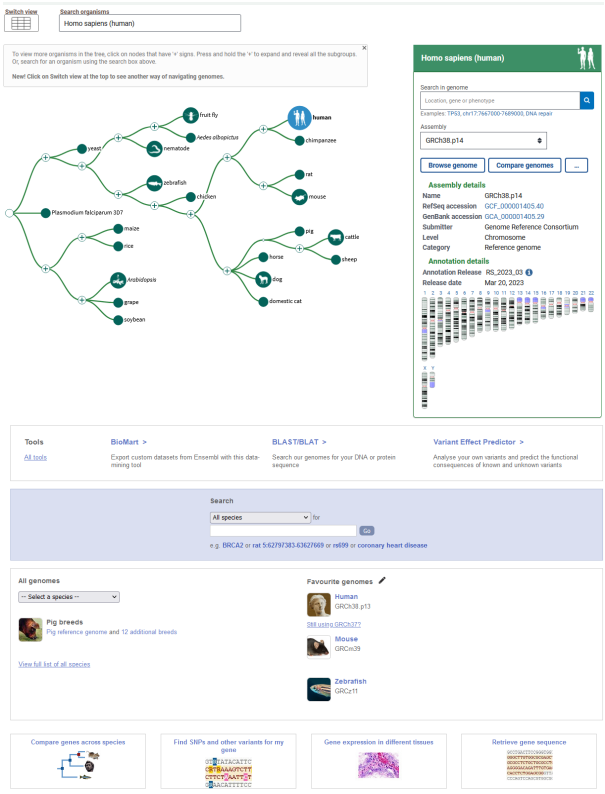
Kolejnym wyzwaniem jest brak kontekstu i adnotacji w danych repozytorium, co czyni je mniej przydatnymi do badań. Ich twórcy zmienili przeznaczenie danych znajdujących się w GenBanku, tworząc wyspecjalizowane bazy danych, dostosowane do konkretnej estetyki, historii projektu i potrzeb społeczności. Bazy te zapewniają lepszą organizację i dostęp do danych genomowych.

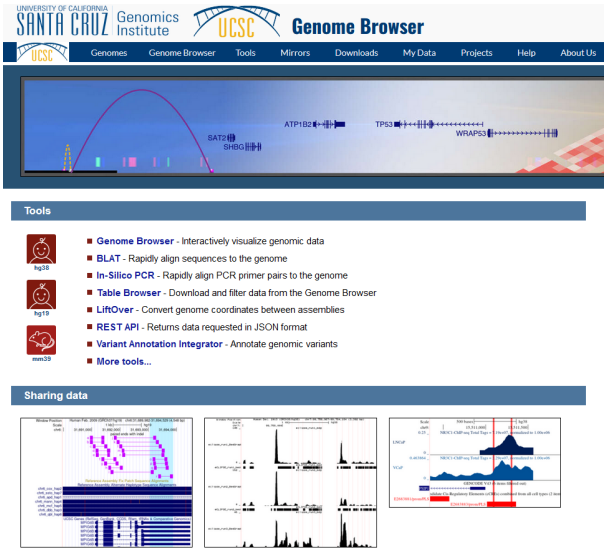


Wzrost ilości sekwencji w bazie GenBank na przestrzeni lat

Ogólne przeglądarki genomu

Aby zaspokoić potrzeby badawcze naukowców, pojawiły się ogólne przeglądarki genomu. Przykłady udanych przeglądarek obejmują UCSC Genome Browser, Ensembl firmy EBI i MapViewer firmy NCBI. Przeglądarki te przepakowują dane genomowe i adnotacje genów z baz danych, takich jak GenBank, oferując kompleksowy kontekst genomowy dla określonych cech genomu, takich jak geny lub loci chorobowe. Ponadto przeglądarki te ułatwiają porównania międzygatunkowe poprzez wyświetlanie informacji we wspólnych formatach, umożliwiając łatwiejszą wizualizację i ekstrakcję danych. Na przykład użytkownicy mogą wyszukiwać określony region genomu, taki jak gen choroby, i uzyskać dostęp do wizualnych wyświetlaczy odpowiedniej sekwencji i adnotacji. Wyświetlacze te są powiązane z dodatkowymi danymi i bazami danych w celu dalszego badania, a także zapewniają linki do oryginalnych źródeł danych.





Strony wymienionych genomowych baz danych

Znormalizowane narzędzia bazy danych genomu: GMOD

GMOD to standardowa baza danych i zestaw narzędzi do wizualizacji typu open-source, który promuje ustandaryzowane zapytania, przeglądanie i korzystanie z baz danych genomów różnych gatunków. Kilka baz danych genomów specyficznych dla gatunków i taksonów przyjęło narzędzia GMOD w celu ulepszenia ich adnotacji, wizualizacji i opcji zapytań. Celem jest ułatwienie badań i studiów porównawczych poprzez zapewnienie wspólnych ram dla rozwoju baz danych i eksploracji danych.

Specyficzne bazy danych:

Bazy danych genomów specyficznych dla gatunku lub taksonu zostały opracowane w celu zapewnienia głębszych informacji o różnych genomach. Te bazy danych są publicznie dostępne i często kuratorowane. Oferują one dokładne adnotacje i zawierają typy danych specyficzne dla danego gatunku. Naukowcy mogą uzyskać dostęp do tych baz danych w celu bardziej dogłębnej analizy określonych genomów.

Bazy danych tematycznych koncentrują się na określonych kategoriach danych biologicznych, takich jak domeny białkowe, struktury białkowe, dane dotyczące ekspresji i badania asocjacyjne obejmujące cały genom. Te bazy danych służą jako cenne zasoby dla naukowców pracujących w określonych obszarach badań. Jednak duża liczba tych baz danych może prowadzić do redundancji i braku integracji.

Przykład bazy danych poświęconej specyficznemu tematowi - szczurom

Wyzwania

Zapewnienie dokładnych danych oraz efektywnego zarządzania i opieki nad nimi to ciągłe wyzwania dla genetycznych baz danych. Proponowane rozwiązania obejmują edukację biokuratorów baz danych, znormalizowane włączanie danych sekwencyjnych i odniesień do publikacji oraz kuratelę społeczności. Podczas gdy niektóre bazy danych przyjęły kuratelę społeczności, inne, takie jak GenBank, mają obawy i utrzymują autorytatywne repozytoria. Wysiłki mające na celu zachęcenie naukowców do wnoszenia wkładu i poprawy wiarygodności kurowania trwają i będą trawły cały czas.

Jeśli chodzi o bazy danych zawierające ludzkie dane genomowe, pojawia się również duży problem prywatności i bezpieczeństwa, stanowiący wyzwanie zarówno techniczne, jak i społeczne. Takie jak wystarczająca anonimizacja, poufność i zarządzanie dostępem. Wymaga to projektowania systemów od podstaw z uwzględnieniem prywatności. Obawy te spowodowały w przeszłości konieczność usunięcia dużych zbiorów danych z publicznych baz danych.

Kolejną przeszkodą jest przytłaczająca liczba dostępnych zasobów genomicznych, co utrudnia badaczom ich zlokalizowanie i efektywne wykorzystanie.

Wnioski:

Zasoby danych genomowych pchnęły dziedzinę bioinformatyki do przodu. Zapewniają one badaczom nieocenione narzędzia do eksploracji i analizy informacji genetycznych. Jednak nadal istnieją wyzwania, takie jak utrzymanie dokładności, dostępności i organizacji danych.
Pomimo tych wyzwań, korzyści i możliwości zapewniane przez dane genomiczne były znaczące, umożliwiając nowe odkrycia i relacje interdyscyplinarne.

Źródła

Praca została oparta na artykule "Genomic Data Resources: Challenges and Promises" by Warren C. Lathe III (OpenHelix), Jennifer M. Williams (OpenHelix), Mary E. Mangan (OpenHelix), and Donna Karolchik (University of California, Santa Cruz Genome Bioinformatics Group), published in Nature Education in 2008 (Lathe, Williams, Mangan, & Karolchik, 2008, p. 2).

Wan, Z., Hazel, J.W., Clayton, E.W. et al. Sociotechnical safeguards for genomic data privacy. Nat Rev Genet 23, 429–445 (2022)