BLAST krok po kroku: przewodnik po przeszukiwaniu sekwencji

Jak uzywac BLAST do identyfikacji sekwencji DNA i bialek - od zapytania do interpretacji wynikow

Agnieszka Wi?niewska · 19.10.2026

W 2026 roku baza danych GenBank przechowuje ponad 3,7 biliona par zasad sekwencji nukleotydowych. Ktos musi przez to wszystko przeszukiwac. I tu pojawia sie BLAST - narzedzie, ktore od 1990 roku jest absolutnym standardem w laboratoriach bioinformatycznych, a liczba cytowan oryginalnej publikacji Altschula i wspolpracownikow przekroczyla 100 000. To prawdopodobnie najczesciej cytowany artykul naukowy w historii biologii molekularnej.

Kiedy po raz pierwszy zetknelam sie z BLAST podczas pracy nad projektem bioremediacji w dolinie Odry, poczulam cos w rodzaju ulgi - zamiast recznie porownywac setki sekwencji bakteryjnych wyizolowanych z probek osadow, moglam w kilka minut dowiedziec sie, z jakim organizmem mam do czynienia. Dzis pokaze, jak to zrobic krok po kroku.

Czym wlasciwie jest BLAST i jak dziala algorytm

BLAST to skrot od Basic Local Alignment Search Tool. Nazwa mowi duzo: narzedzie szuka lokalnych dopasowan, czyli nie wymaga, zeby cala sekwencja zapytania pasowala do sekwencji w bazie danych. Wystarczy, ze fragment - nawet kilkadziesiat nukleotydow - jest wystarczajaco podobny. To kluczowa roznica w stosunku do wczesniejszych metod globalnego dopasowania, takich jak algorytm Needlemana-Wunscha.

Algorytm dziala w trzech etapach. Najpierw dzieli sekwencje zapytania na krotkie slowa (words) o dlugosci k - domyslnie 11 nukleotydow dla DNA i 3 aminokwasy dla bialek. Nastepnie wyszukuje w bazie danych dokladne lub niemal dokladne trafienia tych slow. Wreszcie rozszerza znalezione trafienia w obie strony, dopoki wynik dopasowania rosnie powyzej ustalonego progu. Ten mechanizm pozwala ominac porownywanie kazdej sekwencji z kazda - co przy miliardach sekwencji w GenBank byloby obliczeniowo niemozliwe w rozsadnym czasie.

Heurystyczny charakter BLAST oznacza, ze narzedzie moze pominac bardzo slabe, ale biologicznie istotne dopasowania. W praktyce jednak dla typowych analiz identyfikacyjnych jest to marginalny problem.

Wazne jest rozroznienie miedzy BLAST a dokladnym przeszukiwaniem sekwencji. BLAST nie gwarantuje znalezienia absolutnie najlepszego dopasowania - gwarantuje znalezienie dopasowan powyzej okreslonego progu statystycznego w czasie wielokrotnie krotszym niz metody dokladne.

Piec programow BLAST - ktory wybrac

Na stronie NCBI BLAST (blast.ncbi.nlm.nih.gov) dostepnych jest kilka wariantow narzedzia. Wybor wlasciwego programu to pierwszy krok, ktory decyduje o sensownosci wynikow.

Dla wiekszosci poczatkujacych uzytkownikow: jesli masz sekwencje DNA z sekwencjonera i chcesz wiedziec, co to jest - zacznij od blastn. Jesli interesujesz sie funkcja bialka - uzyj blastp. Blastx jest szczegolnie przydatny, gdy sekwencja jest nowa i nie wiesz, czy ma homologi bialkowe.

Krok po kroku: pierwsze wyszukiwanie w BLAST

Krok 1: Przygotowanie sekwencji zapytania

Sekwencja musi byc w formacie FASTA lub jako surowy ciag nukleotydow/aminokwasow. Format FASTA wyglada tak: pierwsza linia zaczyna sie od znaku > i zawiera nazwe sekwencji, kolejne linie to sama sekwencja. Przyklad dla fragmentu genu 16S rRNA bakterii izolowanej z probki glebowej:

>Probka_gleba_Odra_2026_izolat3
AGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGTAACAGG

Sekwencja nie powinna zawierac spacji wewnatrz ciagu. Znaki niejednoznaczne (N dla nukleotydow, X dla aminokwasow) sa akceptowane, ale ich nadmiar obniza jakosc wynikow. Jesli sekwencja pochodzi z sekwencjonowania Sangera, usun regiony o niskiej jakosci - zazwyczaj pierwsze 20-50 i ostatnie 20-50 nukleotydow chromatogramu.

Krok 2: Wybor bazy danych

Na stronie NCBI BLAST dostepnych jest kilka baz. Dla sekwencji nukleotydowych najwazniejsze to:

- nr/nt - baza nieredundantna, najwieksza, zawiera sekwencje ze wszystkich organizmow; dobra do identyfikacji ogolnej
- RefSeq - sekwencje referencyjne, zweryfikowane i adnotowane; lepsza do analiz porownawczych
- 16S ribosomal RNA sequences - dedykowana do identyfikacji bakterii i archeonow na podstawie genu 16S; szybsza i dokladniejsza dla mikrobiologow
- Whole-genome shotgun contigs (wgs) - dane z projektow sekwencjonowania genomow, przydatna gdy szukasz sekwencji z projektow metagenomicznych

Dla bialek glowna baza to nr (non-redundant protein sequences) lub UniProtKB/Swiss-Prot - ta ostatnia zawiera tylko manualnie zweryfikowane wpisy, co zmniejsza ryzyko propagacji blednych adnotacji.

Krok 3: Ustawienia parametrow

Domyslne parametry dzialaja dobrze dla wiekszosci analiz, ale kilka ustawien warto rozumiec:

- Word size - dlugosc poczatkowych slow; mniejsza wartosc = wieksza czulosc, dluzszy czas obliczen. Dla sekwencji bardzo podobnych (>95% identycznosci) mozna zwiekszyc do 28.
- Expect threshold (E-value) - domyslnie 0,05; im nizszy, tym bardziej rygorystyczny prog statystyczny. Dla wstepnej eksploracji mozna zwiekszyc do 10, dla pewnych identyfikacji zmniejszyc do 1e-10.
- Max target sequences - domyslnie 100 wynikow; przy analizach filogenetycznych

warto zwiekszyc do 500.
- Low complexity filter - domyslnie wlaczony; maskuje powtorzenia i regiony o niskiej zlozonosci, ktore daja falszywe trafienia.

Krok 4: Uruchomienie wyszukiwania

Po wklejeniu sekwencji i ustawieniu parametrow klikasz BLAST. Czas oczekiwania zalezy od dlugosci sekwencji, wybranej bazy i obciazenia serwerow NCBI. Dla pojedynczej sekwencji 16S (~1500 bp) w bazie 16S rRNA wyniki pojawiaja sie zazwyczaj w 15-60 sekund. Dla blastx w bazie nr przy dlugiej sekwencji genomowej moze to potrwac kilka minut.

NCBI generuje unikalny identyfikator wyszukiwania (RID), ktory pozwala wrocic do wynikow pozniej - wazny przez 36 godzin.

Jak interpretowac wyniki BLAST

Sekcja graficzna i tabela wynikow

Strona wynikow sklada sie z kilku czesci. Na gorze widac graficzna reprezentacje dopasowan - kolorowe paski pokazuja, gdzie sekwencje z bazy pokrywaja sie z sekwencja zapytania. Czerwony kolor oznacza najlepsze dopasowania (najwyzszy score), niebieski - slabsze.

Pod grafika znajduje sie tabela z lista trafien. Kazdy wiersz to jedno dopasowanie z nastepujacymi kolumnami:

- Description - nazwa organizmu i genu/bialka z bazy danych
- Max Score - najwyzszy wynik bitowy (bit score) dla danego trafienia; wyzszy = lepsze dopasowanie
- Total Score - suma wynikow dla wszystkich fragmentow dopasowania (HSP)
- Query Cover - procent sekwencji zapytania objety dopasowaniem; wartosc ponizej 50% powinna wzbudzic ostroznosc
- E-value - wartosc oczekiwana; kluczowy parametr statystyczny
- Per. Ident - procent identycznych pozycji w dopasowaniu

E-value - najwazniejszy parametr

E-value (wartosc oczekiwana) mowi, ile dopasowan o danym wyniku mozna by oczekiwac przypadkowo przy przeszukiwaniu bazy danych tej wielkosci. E-value = 0,001 oznacza, ze statystycznie raz na tysiac wyszukiwan takie dopasowanie pojawi sie przez przypadek. Im nizsze E-value, tym bardziej istotne statystycznie trafienie.

Praktyczna zasada: E-value ponizej 1e-5 to dopasowanie warte uwagi. Ponizej 1e-20 - silny sygnal homologii. Powyzej 0,01 - interpretuj z duza ostroznoscia, moze byc przypadkowe.

Uwaga: E-value zalezy od rozmiaru bazy danych. To samo dopasowanie bedzie mialo wyzsze E-value w wiekszej bazie niz w mniejszej. Dlatego nie mozna porownywac E-values z roznych wyszukiwan bez uwzglednienia kontekstu.

Procent identycznosci a homologia

Czesty blad: utozsamianie procentu identycznosci z homologia. Dwie sekwencje sa homologiczne, jesli maja wspolnego przodka ewolucyjnego - to kwestia jakosciowa, nie ilosciowa. Procent identycznosci to miara podobienstwa sekwencji, nie dowod homologii.

Dla identyfikacji bakterii na poziomie gatunku na podstawie genu 16S rRNA przyjmuje sie prog 98,65% identycznosci (wynika z badan porownawczych z 2015 roku, nadal stosowany w 2026). Ponizej tego progu mozna mowic co najwyzej o identyfikacji na poziomie rodzaju. Dla bialek progi sa inne i zaleza od konkretnej rodziny bialkowej.

Typowe bledy i jak ich unikac

Kilka problemow pojawia sie regularnie, zwlaszcza u osob zaczynajacych prace z BLAST.

Zanieczyszczenie sekwencji. Jesli pierwsze trafienia to ludzkie sekwencje, a badales bakterie glebowa, prawdopodobnie sekwencja jest zanieczyszczona materialem ludzkim. Sprawdz procedure izolacji DNA i powtorz sekwencjonowanie.

Chimery w sekwencjach 16S. Sekwencje chimeryczne - powstale z polaczenia fragmentow roznych organizmow podczas PCR - daja niespojne wyniki BLAST: rozne fragmenty sekwencji zapytania dopasowuja sie do roznych organizmow. Przed analiza BLAST warto uzyc narzedzi takich jak UCHIME lub ChimeraSlayer.

Bledne adnotacje w bazie. GenBank zawiera sekwencje z blednymi lub niepelnymi adnotacjami. Jesli kilka najlepszych trafien wskazuje na rozne gatunki, a E-values sa podobne, nie polegaj na nazwie pierwszego trafienia. Sprawdz kilka wynikow i porownaj.

Ignorowanie Query Cover. Trafienie z 99% identycznosci, ale pokryciem zapytania na poziomie 20%, jest malo wiarygodne. Tylko maly fragment sekwencji pasuje do trafienia.

BLAST lokalnie i alternatywy

Serwer NCBI BLAST jest wygodny, ale ma ograniczenia: nie mozna przeszukiwac wlasnych, niepublicznych baz danych, a przy duzej liczbie zapytan czas oczekiwania rosnie. W takich przypadkach warto zainstalowac BLAST+ lokalnie.

BLAST+ to pakiet narzedzi wiersza polecen dostepny bezplatnie ze strony NCBI. Instalacja na systemie Linux zajmuje doslownie kilka minut. Mozna wtedy tworzyc wlasne bazy danych z sekwencji FASTA poleceniem makeblastdb i przeszukiwac je lokalnie - bez limitu zapytan i bez wysylania danych na zewnetrzne serwery, co jest istotne przy pracy z niepublikowanymi danymi genomowymi.

Alternatywy dla BLAST to miedzy innymi DIAMOND - narzedzie do przeszukiwania bialek, nawet 20 000 razy szybsze od blastp przy podobnej czulosci, szczegolnie przydatne w analizach metagenomicznych; oraz MMseqs2, ktory laczy wysoka szybkosc z elastycznoscia w dopasowaniach sekwencji o niskiej identycznosci. Dla analiz filogenetycznych na duza skale te narzedzia coraz czesciej zastepuja klasyczny BLAST.

Praktyczne zastosowania w badaniach srodowiskowych

W projektach bioremediacji BLAST jest narzedziem pierwszego kontaktu z nieznanym materialem biologicznym. Gdy izolujesz bakterie ze skazonego gruntu i sekwencjonujesz jej gen 16S, BLAST pozwala w ciagu minut ustalic, z jakim rodzajem masz do czynienia i czy pokrewne gatunki maja udokumentowane zdolnosci do degradacji konkretnych zwiazkow

chemicznych.

W monitoringu srodowiskowym opartym na sekwencjonowaniu metabarkodingu - na przyklad przy analizie roznorodnosci biologicznej rzek na podstawie eDNA - BLAST jest jednym z etapow klasyfikacji taksonomicznej odczytow. Kazdy odczyt jest porownywany z baza referencyjna, co pozwala przypisac go do taksonu. Przy tysiacach odczytow na probke lokalne uruchomienie BLAST lub DIAMOND jest praktyczna koniecznoscia.

Warto jednak pamietac o ograniczeniu: wyniki BLAST sa tak dobre, jak baza danych, z ktora porownujesz. Dla organizmow slabo zbadanych taksonomicznie - wielu grup grzybow, pierwotniakow czy bakterii z ekstremalnych srodowisk - bazy referencyjne sa niekompletne. Trafienie z 90% identycznosci do sekwencji oznaczonej jako "uncultured bacterium