

Laboratorium 6

Czas na oddanie finalnego raportu: 6 tygodni

Nazwa pliku: imie_nazwisko_6_bio.pdf

Typ ćwiczenia: jednotygodniowe

Cel: Celem ćwiczenia jest zapoznanie studentów z podstawowymi sposobami analizy sekwencji DNA. Ćwiczenie zawiera elementy analizy kryminalistycznej.

Zadania

Globalne przyrównanie sekwencji i lokalne przyrównanie (ang. global and local alignment) to dwa rodzaje algorytmów stosowanych w analizie sekwencji nukleotydów lub aminokwasów. Oba typy algorytmów służą do porównywania sekwencji z jednego lub więcej organizmów w celu zidentyfikowania podobieństw, lub różnic między nimi. Globalne przyrównanie sekwencji jest wykorzystywane do porównywania dwóch lub więcej sekwencji, w których każde porównanie sekwencji jest przeprowadzane na całej ich długości. Algorytm porównuje każdy nukleotyd lub aminokwas w jednej sekwencji z odpowiadającym nukleotydem, lub aminokwasem w drugiej sekwencji. Globalne przyrównanie sekwencji jest przydatne, gdy analizowane sekwencje są podobne w całej swojej długości. Z kolei lokalne przyrównanie sekwencji wykorzystywane jest do porównywania krótszych segmentów sekwencji, które mogą różnić się od siebie w różnych miejscach. Algorytm porównuje sekwencje jedna po drugiej i identyfikuje regiony, w których sekwencje są identyczne lub podobne. Lokalne przyrównanie sekwencji jest przydatne w przypadku sekwencji, które różnią się w kilku miejscach i ma na celu znalezienie regionów podobieństw, takich jak sekwencje konsensusowe lub motywy sekwencyjne. (patrz algorytmy Needlemana–Wunscha i Smitha-Watermana)

1 Znajdź w bazie danych NCBI Genbank sekwencję DNA dowolnego genu człowieka (podpowiedź: gatunek *Homo sapiens*) znajdującego się w mitochondrialnym DNA (podpowiedź: mtDNA), rekord musi zawierać sekwencję dłuższą niż 1000 nukleotydów. Podpowiedź: Jeśli wpiszesz w wyszukiwarce w bazie NCBI GenBank frazę “homo sapiens gene mtDNA” to na samej górze zobaczysz wizualizację mtDNA, kliknij w symbol jednego z genów (na przykład: ND1). Podaj Accession Number znalezionej sekwencji [**task 1**].

- Znajdź w tej samej bazie odpowiednik tego genu u szympansa (gatunek: *Pan troglodytes*) - nie musi być to cały gen. Kieruj się nazwą genu. Podaj Accession Number znalezionej sekwencji [**task 2**].
- Dokonaj porównania za pomocą narzędzia dotmatcher, ale w taki sposób, żeby na wykresie “dot matrix” uwidocznili przede wszystkim globalne przyrównanie. W tym celu zmieniaj wartość znajdującą się w polu “Threshold” oraz “Window size over which to test threshold”.
- Jako wynik realizacji tego zadania wklej wygenerowany wykres [**task 3**]. Skomentuj co najmniej jednym zdaniem uzyskany wynik [**task 4**]. Wskaż widoczne ograniczenia tego typu analizy (jak widać - dość prymitywny algorytm) [**task 5**]. Na podstawie tej analizy spróbuj skrótkowo zdefiniować “threshold” oraz “Window size over which to test threshold” w tym algorytmie [**task 6**].

Predyspozycja genetyczna to warunek, w którym obecność określonych sekwencji lub po prostu fragmentu DNA zwiększa ryzyko wystąpienia choroby, lub określonej cechy fenotypowej. W

badaniach genetycznych, markery genetyczne, które są związane z predyspozycją do danej choroby lub cechy fenotypowej, są używane do określenia ryzyka wystąpienia danej choroby lub cechy.

2 Na SNPedia.com znajdź podstronę opisującą marker genetyczny związany z częściową opornością na wirusa HIV (odnośnik z głównej strony), podaj jego oznaczenie [task 7], a następnie przejdź do bazy danych Ensembl (ramka po prawej stronie na podstronie markera genetycznego, o którym mowa powyżej) i tam znajdź sekwencję flankującą dla tego fragmentu. Pobierz ją w formacie FASTA poprzez ścieżkę Download sequence → Preview (nie zmieniaj ustawień), co pozwoli pozbyć się zmienności widocznej bezpośrednio pod odnośnikiem “Flanking sequence”. Wklej tę sekwencję jako odpowiedź na zadanie [task 8].

- Za pomocą narzędzia NCBI Blast znajdź sekwencje homologiczne dla tego fragmentu tzn. wklej swoją sekwencję (w ustawieniach zmień tylko w “Algorithm parameters” -> “Max target sequences” na 250) i kliknij “Blast”. Filtrując parametr “Percent Identity” znajdź w obrębie tej bazy jednego z “najmniej podobnych krewnych” *Homo sapiens* (jeden z najniższych %), który posiada gen homologiczny wobec tego, którego szukasz (jedna z najniższych wartości w kolumnie “Per. Ident”). Pamiętaj żeby znaleźć rekord dla DNA, w nazwie powinien mieć “complete cds”, a nie mRNA (podpowiedź: wybierz rekord dla gatunku *Chlorocebus pygerythrus*). Podaj Accession Number znalezionej sekwencji [task 9]. Wejdź następnie do tego rekordu i skopiuj sekwencję DNA. Jako odpowiedź na zadanie wklej sekwencję w formacie FASTA [task 10].
- W narzędziu Emboss Needle wybierz “DNA” i wykonaj przyrównanie obu sekwencji (sekwencji z task 8 oraz drugiej z task 10). Wklej wynik działania programu jako odpowiedź na to zadanie (nie zmieniaj ustawień tego programu) [task 11].
- Usuń te fragmenty (z przodu i z tyłu) sekwencji z gatunku *Chlorocebus pygerythrus*, które nie mają odpowiednika w Twojej sekwencji i powtórz analizę. Wklej wynik działania programu jako odpowiedź na to zadanie [task 12]. Dlaczego teraz podobieństwo jest wyższe? O czym świadczy fakt, że w pierwszym przyrównaniu podobieństwo było niższe? [task 13]

3 Powtórz powyższą analizę z użyciem Emboss Needle, dla dwóch losowo wygenerowanych sekwencji DNA (narzędzia można znaleźć on-line lub napisz własne w Pythonie) i zaznacz na takim przyrównaniu po jednym przykładzie: match, mismatch i gap [task 14].

4 Za pomocą FinchTV lub innej aplikacji otwórz zawartość pliku B93_27F.ab1 pochodzącego z fragmentu gleby pozyskanego z miejsca przestępstwa, a uzyskanego z wykorzystaniem sekwenatora wykorzystującego technologię Sangera, a następnie skopiuj sekwencję nukleotydową i umieść ją w narzędziu BLAST:

- Co oznaczają duże litery “N”, “M” i inne na początku sekwencji (inne niż klasyczne oznaczenia nukleotydów, czyli A, T, C, G?) [task 15]? Usuń fragment, który zawiera takie nukleotydy do sekwencji “TGCAGTCGAG”, aby nie wpływały na algorytm BLAST.
- Bez zmiany ustawień kliknij “BLAST” (aczkolwiek upewnij się, że w “Algorithm parameters” -> “Max target sequences” jest 100, czyli ustawienie default).
- Na podstawie przyrównania z zasobami bazy danych z zakładki “Sequences producing significant alignments” spróbuj odpowiedzieć na pytanie — do jakiego gatunku może należeć ta sekwencja, którą analizujesz (podpowiedź: na przykład “*Bacillus* sp.” to nazwa rodzaju, “*Bacillus subtilis*” to nazwa gatunku) [task 16]? Poszukaj i ustal, kiedy po raz pierwszy

- wyizolowano ten gatunek i z powierzchni jakiego innego organizmu go pozyskano [task 17]?
- Z jakiego genu pochodzi analizowany fragment DNA [task 18]?
 - Zjrzyj do rekordów o największym podobieństwie do sekwencji badanej (wysoki Max Score, Total Score, Per. Ident. i in.) i spróbuj odpowiedzieć na pytanie: z jakiego najbardziej prawdopodobnego środowiska pochodzi sekwencja badana, którą analizujesz (np. człowiek, roślina, zwierzę, a może coś innego?) [task 19]
 - W zakładce “Alignments” ustaw wyświetlania jako “Flat query-anchored with dots for identities”, długość linii na 120 i skopiuj przyrównanie pierwszej linii Twojej sekwencji do pozostałych [task 20]. Co oznacza “.”, “-”, a co oznacza, jeśli w miejscu kropki znajduje się nukleotyd [task 21].
 - Załóżmy, że nie jesteś pewna/pewny identyfikacji do gatunku. Posłużmy się dedykowaną bazą dla rRNA (czyli sekwencji kodujących rybosomalne RNA, którego fragment posiadasz). W oknie wyszukiwania BLAST wybierz bazę danych “rRNA/ITS databases”, a z listy poniżej “16S ribosomal RNA...”. Kliknij BLAST.
 - W wynikach pojawiło się więcej gatunków bakterii z tego rodzaju i rzeczywiście ich podobieństwo do sekwencji badanej jest duże. Załóżmy, że wahasz się pomiędzy gatunkiem *L. fusiformis*, a *L. mangiferihumi* oraz *L. sphaericus*. Załóżmy, że interesują Cię tylko te sekwencje, które wykazują podobieństwo na poziomie ponad 98%. Ustaw w Filter Results → Percent Identity, aby przefiltrowało tylko sekwencje o 98-100% podobieństwa. Następnie w zakładce “alignments” zastosuj takie ustawienia: “Flat query-anchored with dots for identities”, długość linii na 120 i skopiuj całe przyrównanie [task 22]. Zaznacz te nukleotydy, które rozróżniają twój gatunek badany od pozostałych (podpowiedź: jest ich 9) [task 23]. Spróbuj znaleźć taki nukleotyd, który rozróżnia sekwencję badaną od tych analizowanych (to jest marker twojej próbki, w tej pozycji ten nukleotyd występuje tylko w Twojej próbce!) [task 24].