Michał Humiński
Kacper Kaszuba
Bioinformatyka 2 rok studiów licencjackich

Sprawozdanie z listy 1 - zadania 1 oraz 2

Zadania 1 oraz 2 są dosyć podobne, oba polegają na wczytaniu danych w formacie FASTA lub Genbank w celu wyciągnięcia z nich określonych danych. W związku z tym napisaliśmy 1 program, który obsługuje wszystkie zawarte w nich polecenia.

W naszym kodzie użyliśmy pakietów re (celem wyszukiwania określonych sekwencji tekstowych), wx (jako interfejs graficzny) oraz wybranych funkcji z pakietu Biopython.

W kodzie na początku zawarliśmy krótką informację czego dotyczy oraz przykładowe numery sekwencji GenBank do pobrania celem sprawdzenia jego działania.
Ponadto podsyłamy powyższe sekwencje celem ich wczytania z określonej lokalizacji. Jedna sekwencja zapisana jest w formacie *.fasta*, zawiera ją plik *„sequence.fasta”*.
3 pozostałe pliki dotyczą sekwencji w formacie *.gb*. Plik *„sequence.gb”* zawiera sekwencję genu niepodzielonego (ciągłego, z jedną sekwencją CDS), plik *„sequence2.gb”* sekwencję genu podzielonego (nieciągłego, z wieloma sekwencjami CDS), natomiast plik *„sequence3.gb”* sekwencję genu podzielonego ze sztucznie wprowadzonymi mutacjami w aminokwasach (zamiana kilku par aminokwasów) celem ukazania różnic w translatowanej sekwencji CDS oraz sekwencji aminokwasowej.

Instrukcja użytkowania programu
Po otworzeniu programu widać puste okno z paskiem menu na górze.

Wczytanie danych
Najpierw musimy wczytać nasze dane w wybranym przez nas formacie FASTA lub GenBank. Dane możemy załadować z pobranego pliku lub też ściągnąć z bazy danych. Odpowiadają za to odpowiednio zatytułowane funkcje: *„OtworzGenbank”* i *„PobierzGenBank”* oraz *„OtworzFasta”* i *„PobierzFasta”* w menu *„Dane”*.

W funkcjach dotyczących załadowania wcześniej pobranych danych otwiera się okno dialogowe, w którym możemy wybrać docelową lokalizację naszego pliku. W przypadku funkcji pobierających dane bezpośrednio z bazy GenBank wyskakują 2 okna dialogowe. W pierwszym z nich musimy podać nasz adres email, który jest niezbędny celem pobrania danych, a następnie numer identyfikacyjny rekordu, który zamierzamy analizować (numer *„Version”* - tylko on jednoznacznie wskazuje na interesujący nas rekord).
Dalsze postępowanie w przypadku analogicznych funkcji dla poszczególnych formatów jest identyczne.

W przypadku formatu FASTA otwieramy nasz plik, który dzielimy na linijki. Następnie badamy każdą z nich. Jeżeli zaczyna się od znaku „>”, to jest ona nagłówkiem. W przeciwnym przypadku zawiera ona sekwencję nukleotydową lub aminokwasową, które to sklejamy ze sobą.
Należy zauważyć, że możemy wczytać plik FASTA zawierający dane dla więcej niż 1 organizmu, co zostało przez nas uwzględnione przy ich obrabianiu. Na koniec uzyskujemy interaktywną listę z nazwami (jako nagłówki) wczytanych rekordów.

Format GenBank jest znacznie bardziej skomplikowany, wymaga więcej kodu celem wyciągnięcia interesujących nas wielkości - sekwencji nukleotydowej (*„origin”*), sekwencji kodujących (*„CDS”*) oraz sekwencji białkowych (*„translation”* w *„CDS”*).
Celem wyszukania danych otwieramy nasz plik, który następnie redukujemy do jednej linijki (usuwamy znaki końca wierszy).

Najpierw wyciągamy sekwencję nukleotydową *„origin”*, co jest najprostszym zadaniem. Za pomocą pakietu re wyszukujemy odpowiednią sekwencję w tekście. Zawiera ona sekwencję nukleotydową, z której usuwamy przerwy, a następnie tworzymy *string* zawierający i łączący wyłącznie litery (nukleotydy). Będzie to zapis znajdujący się na pierwszej pozycji w utworzonej liście.

Następnie wyciągamy sekwencje CDS - to najtrudniejsze zadanie. Najpierw musimy wyszukać odpowiednią sekwencję w tekście z pomocą pakietu re. Podobnie jak poprzednio usuwamy przerwy, jednak tym razem wyciągamy wyłącznie liczby. Otrzymujemy listę zawierającą same liczby numery oznaczające położenie sekwencji CDS. Geny mogą być także podzielone - zawierać introny i eksony, co zostało uwzględnione w naszym kodzie. Następnie iterując po 2 elementy z listy numerów pobieramy odpowiednio początek i koniec poszczególnych sekwencji CDS z obrobionej wcześniej sekwencji nukleotydowej. Wszystkie sekwencje na koniec sklejamy ze sobą. Należy również zauważyć, że sekwencje mogą również znajdować się na drugiej nici, która nie jest podana w bazie danych. To również uwzględniliśmy w naszym kodzie. Podczas „wycinania” sekwencji kodujących z pełnej sekwencji nukleotydowej i ich „sklejania”, za pomocą pakietu re badamy czy w określonych CDS znajduje się informacja (zapisana jako: *„complement”*). Kiedy zostanie ona znaleziona, to czytamy daną sekwencję od końca i tworzymy nową nić na zasadzie komplementarności. Dopiero po takiej obróbce dodawana jest ona do łącznej sekwencji kodującej.
W ten sposób otrzymujemy zapis znajdujący się na drugiej pozycji listy.

Jako ostatnie wyciągamy sekwencje aminokwasowe. Wykonujemy to podobnie za pomocą pakietu re, wyszukując odpowiednie sekwencje tekstowe. Znajdują się one w sekcji *„CDS”* jako *„translation”*. Tutaj zadanie jest prostsze - usuwamy przerwy i końcowo sklejamy ze sobą wszystkie zalezione sekwencje.

Na sam koniec otrzymujemy interaktywną listą z 3 zapisami zawierającymi określone informacje - odpowiednio: sekwencję nukleotydową, sekwencję kodująąa oraz sekwencję aminokwasową.

Podstawowe obliczenia

Na wczytanych danych możemy wykonać różne operacje w podmenu *„Obliczenia”*. Warto zaznaczyć, że ich odpowiednia obróbka i przygotowanie pozwala użyć tych samych funkcji do analizy zarówno sekwencji pochodzących z formatu *.fasta*, jak i sekwencji pochodzących z formatu *.gb*.

Za pomocą funkcji *„LancuchAminokwasow”* możemy zobaczyć łańcuch aminokwasów lub też łańcuch nukleotydów (z dowolnego zapisu) w postaci listy w nowym oknie.

Funkcja *„LiczbaAminokwasow”* ukazuje liczbę aminokwasów lub też nukleotydów występującą w zaznaczonej sekwencji. Wynik otwiera się jako tekst w nowym oknie.

Następnie możemy zliczyć liczbę par aminokwasów (lub też nukleotydów) za pomocą funkcji *„LiczbaParAminokwasow*”. Wyniki są ukazane w podobnej postaci jak w poprzedniej funkcji.

Ostatnią prostą funkcją jest *„LiczbaTrojekNukleotydow”*, która zlicza trójki nukleotydów (lub aminokwasów) analogicznie, jak poprzednia funkcja dotycząca par. Należy tutaj zauważyć, że niniejsza funkcja ukazuje wszystkie możliwe trójki nukleotydów - wszystkie możliwe kodony, co nie odpowiada rzeczywistej sytuacji kodonów, które kodują określone aminokwasy tylko i wyłącznie w jednej fazie odczytu.

Porównania

Zgodnie z treścią 2 zadania, dotyczą one tylko sekwencji pobranych z GenBanku. Odpowiednie opcje znajdują się w podmenu *„Porównanie sekwencji GenBank”*.

Najpierw opiszemy najprostszą funkcję porównującą sekwencje nukleotydowe pełną oraz kodującą (*„origin”* oraz *„CDS”*). Zapisy w liście mają określoną kolejność, zatem aby funkcja *„PorownanieSekwencjiOriginiCDS”* zadziałała prawidłowo, musimy zaznaczyć na liście sekwencję nukleotydową, o czym przypomina wyskakujące okienko.
Funkcja dotyczy porównania procentowej zawartości nukleotydów. Funkcja zlicza ilości podstawowych nukleotydów (*a* - adenina, *t* - tymina, *c* - cytozyna, *g* - guanina) w każdej sekwencji i porównuje je ze sobą. Otrzymujemy zatem informację o różnej zawartości zasad azotowych w sekwencjach kodujących oraz pełnej, która zawiera sekwencje niekodujące. Znajdują się one w osobnym oknie jako różnica zawartości nukleotydów w sekwencji *„origin”* względem sekwencji *„CDS”*.

Druga funkcja *„PorownaniePrzetlumaczonychSekwencjiOriginiCDS”* dotyczy podobnego porównania sekwencji jak w powyższym przypadku, jednak tym razem analizujemy sekwencje białkowe powstałe na podstawie sekwencji nukleotydowych. Podobnie musimy oznaczyć sekwencję nukleotydową w powstałej wcześniej liście, co kontroluje wyskakujące okno.
Powstałym problemem jest wybór pierwszego nukleotydu kodującego, który przekłada się na określoną fazę odczytu. Nasz kod jest przygotowany na taką ewentualność. Następnie obie nasze sekwencje poddajemy translacji zgodnie ze standardowym kodem genetycznym (w tym sekwencję *„origin”* w odpowiedniej fazie odczytu). Końcowo otrzymujemy osobne okno z różnicą zawartości poszczególnych aminokwasów w sekwencji *„origin”* względem sekwencji *„CDS”*. Należy zauważyć, że analizujemy wszystkie możliwe jednoliterowe zapisy aminokwasów w formacie *.gb,* także takie skróty, które oznaczają 2 możliwe lub dowolny aminokwas.

Ostatnie porównanie dotyczy analizy sekwencji nukleotydowej kodującej (CDS) oraz sekwencji aminokwasowej. Funkcja *„PorownanieSekwencjiCDSibialkowej”* wymaga wybrania sekwencji kodującej w liście. Nukleotydy tłumaczone są na aminokwasy i usuwane są znaki *„*”* oznaczające aminokwasy terminalne na końcu każdego CDS. Translacja zachodzi na podstawie tabeli standardowych kodonów.

Przy identycznych sekwencjach aminokwasowych otrzymujemy odpowiednią informację. Przy różnych sekwencjach aminokwasowych otrzymujemy informację o procentowym stopniu podobieństwa obu sekwencji oraz różnic w poszczególnych aminokwasach (zamiana wraz z pozycją). Mogą one wynikać z niekompletności sekwencji lub też różnic pomiędzy poszczególnymi kodami genetycznymi różnych organizmów, np. u bakterii.

Zakończenie programu

Program możemy wyłączyć z menu klikając opcję *„Wyjście”* a następnie *„Koniec”*.