Łukasz Birunt 158791,
Dawid Hirsz 154719,
Jakub Janowski 155078

19.12.2018 r.

Sprawozdanie z projektu 1.

Temat 1: Zestawienia optymalne 2 sekwencji. Typ A.

Zadanie polegało na znalezieniu optymalnego zestawienia dwóch sekwencji.
Do programu wczytywane są 2 pliki z sekwencjami.

Użytkownik może podać sekwencje DNA lub RNA.

W przypadku podania DNA wymagane jest również podanie macierzy odległości (Tab. 1.) lub macierzy podobieństw (Tab. 2.).

Macierz odległości zawiera koszt zamiany poszczególnych liter oraz operacji indel. Na jej podstawie tworzona jest tabela przejść dla danych sekwencji. Ich odległość edycyjna znajduje się na jej ostatnim miejscu. Aby odtworzyć najtańsze dopasowanie program znajduje najkrótszą ścieżkę do miejsca 0,0 tabeli**.** Podobna tabela jest tworzona dla wyliczenia optymalnego dopasowania lokalnego. Różni się ona tylko tym, że w miejsca w których miałaby pojawić się wartość ujemna wstawiane jest 0, po to aby nie wpływała ona negatywnie na dalszą część tabeli (w których może zaczynać się optymalne dopasowanie lokalne). Wartość optymalnego lokalnego dopasowania to największa wartość w tabeli. Aby odtworzyć dopasowanie program znajduje najkrótszą ścieżkę, aż do napotkania wartości 0.

Macierz podobieństw zawiera wartości funkcji podobieństwa pomiędzy poszczególnymi literami oraz operacji indel, dla której z reguły przyjmuje się wartości ujemne (kara za wstawienie). Na jej podstawie tworzona jest tabela podobieństw wszystkich przedrostków. Wartości optymalnych dopasowań globalnego i lokalnego są liczone analogicznie jak w poprzednim punkcie z tą różnicą. że rozwiązaniem jest najdłuższa ścieżka. Minimalizacja długości edycyjnej odpowiada maksymalizacji podobieństwa. Równie dobrze można by użyć algorytmu z długości edycyjnej wcześniej mnożąc wartości tabeli przez -1.

W przypadku występowania większej ilości optymalnych dopasowań zwracane jest jedno z nich.

	A	**C**	**G**	**T**	**_**
A	0	2	2	2	1
C	2	0	2	2	1
G	2	2	0	2	1
T	2	2	2	0	1
_	1	1	1	1	0

Tab. 1. Przykładowa macierz odległości dla sekwencji DNA.

	A	C	G	T	_
A	2	-2	-2	-2	-1
C	-2	2	-2	-2	-1
G	-2	-2	2	-2	-1
T	-2	-2	-2	2	-1
_	-2	-2	-2	-2	2

Tab. 2. Przykładowa macierz podobieństw dla sekwencji DNA.

W przypadku podania RNA sekwencje kodonów tłumaczone są na aminokwasy. Uwzględnione są jedynie sekwencje rozpoczynające się kodonem START i kończące kodonem STOP.
Program uwzględnia możliwość istnienia wielu podsekwencji START-STOP w jednej sekwencji, wówczas porównywane są wszystkie możliwe pary, a zwrócony wynik dotyczy najlepszego znalezionego dopasowania wśród wszystkich par.
Dopasowania liczone są tak samo, z tą różnicą, że korzystają z macierzy 21x21, która dotyczy odległości/podobieństw aminokwasów. Następnie dopasowanie aminokwasów zamieniane jest na dopasowanie w alfabecie RNA. Relacja ta jest jednak typu jeden do wielu w związku z czym nie jest jednoznaczna. Jednak można wykorzystać fakt, że wcześniej zamieniliśmy już trójki RNA na aminokwasy, przez co nie ma problemu z jednoznaczną translacją aminokwasów na trójki RNA.

Obsługa programu.

Aplikację uruchamiamy z przekazując w parametrach:
-a *ścieżka do pliku z sekwencją A*
-b *ścieżka do pliku z sekwencją B*
-d *ścieżka do pliku z macierzą dystansu*
-s *ścieżka do pliku z macierzą podobieństw*
-r przekazując ten parametr program będzie oczekiwać sekwencji RNA