Bioinformatyka II, Lista 1C

11 marca 2019

1. W programie BD_pr1 przygotowane jest wczytywanie sekwencji nukleinowych zapisanych w pliku typu fasta. Dla wybranej sekwencji, korzystając z opcji Menu – Obliczenia wyznacz liczbę:
   (a) wszystkich nukleotydów lub aminokwasów w sekwencji – pierwsza opcja Menu Obliczenia;
   (b) wszystkich par nukleotydów lub aminokwasów w sekwencji – druga opcja Menu Obliczenia.

   Wyniki zapisz w przygotowanym polu typu wx.StaticText lub wx.ListBox. Wynik ma być postaci: A, k, gdzie A oznacza jeden lub parę nukleotydów, zaś n liczbe ich występowania. Pamiętaj o zachowaniu porządku w przedstawianych wynikach.
2. Wzorując się na procedurze wczytywania plików typu fasta, napisz procedurę wczytującą pliki typu GenBank, wywołaj ją w opcji Menu – Dane/GenBank. W pliku GenBank w opcji
   (a) FEATURES/CDS/translation: zapisana jest "Sekwencja kodująca" – region nukleotydów odpowiadający sekwencji aminokwasów w białku (translation);
   (b) ORIGIN zapisywane są sekwencje nukleoinowe; Korzystając z tabeli standardowych kodonów porównaj obie sekwencje. Pamiętaj o niekompletności sekwencji.

   Dokładny opis rekordu w bazie GenBank znajdziesz pod adresem:

   `http://www.uwm.edu.pl/bioinfo/dydaktyka/ncbi/samplerecord.html`
3. Pod adresem

   `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC321090/pdf/nar00344-0324.pdf`

   w pracy pt: "The diagonal-traverse homology search algorithm for locating similarities between two sequences" przedstwiono algorytm porównania dwóch sekwencji DNA. Napisz skrypt w pythonie kodujacy powyższy algorytm.
4. Przedstaw graficznie dopasowanie dwóch sekwencji wykorzystując metodę dot-matrix. Dokładniejsze informacje znajdź pod adresem:

   `http://www.srmuniv.ac.in/sites/default/files/files/5(6).pdf`

   Zaproponuj metodę zaznaczania obszarów (linii) wysokiego dopasowania.
5. Algorytm Needelmana-Wunsha w wersji dopasowania lokalnego często nosi nazwę algorytmu Smitha-Watermana:

$$H(i,j) = \max \begin{cases} H(i-1,j-1) + S(a_i, b_j) \\ H(i-1,j) - g \\ H(i,j-1) - g \\ 0 \end{cases}$$

   gdzie $S(\alpha, \beta)$ jest macierzą punktacji, a g jest odpowiedzialna za funkcję kary. Napisz program wyszukujące wszystkie dopasowane podsekwencje składające się z ponad k aminokwasów.
6. Algorytm Needelmana-Wunsha pozwala na porównanie dwóch sekwencji aminokwasowych. Zastosuj go do porównania k, $k > 2$ sekwencji. Zastanów się jak uporządkować porównane sekwencje.

7. Napisz funkcje odległości Penga, Doolittle'a pomiędzy dwoma sekwencjami:

$$d = -100 \ln \left(\frac{S - S_r}{S_{id} - S_r} \right)$$

gdzie:
S - ocena dopasowania;
S_r - średnia ocena dopasowania par losowych sekwencji – jednakowej długości;
S_{id} - średnia ocena dopasowania identycznych par.
S_r i S_{id} oblicz przyjmując, że ich długość losowej sekwencji jest równa długości X_0.
Niech n oznacza liczbę aminokwasów w dłuższej sekwencji. Aby wykonać to zadanie należy:
   - Wylosować przynajmniej 100 par sekwencji (każda długości n). Algorytmem Needelmana-Wunsha obliczyć ich ocenę, a następnie wyznaczyć średnią S_r.
   - Wylosować przynajmniej 100 sekwencji (każda długości n). Dla każdej z wylosowanych sekwencji algorytmem Needelmana-Wunsha obliczyć ocenę dopasowania sekwencji do siebie, a następnie wyznaczyć średnią S_{id}.
   - Obliczyć ocenę dopasowania badanych par sekwencji S i wyniki podstawić do wzoru na odległość Penga.
8. Wczytywanie długich plików
9. Dla różnych macierzy punktacji dokonać porównania tych samych sekwencji aminokwasowych dokonać porównania wyników.