Motorola Science Cup^3
Zadania II Etapu

Bioinformatyka

GARŚĆ HISTORII

Bioinformatyka jest dyscypliną naukową łączącą biologię i informatykę. Jednakże czerpie również z innych dziedzin nauk ścisłych, jak chemia i fizyka. W tym zadaniu spróbujecie zmierzyć się z wyzwaniami stawianymi przez tą gałąź nauki.

Bioinformatyka jako nauka wyłoniła się z biologii, a szczególnie z biochemii i biologii molekularnej w drugiej połowie XX wieku. Katalizatorem było określenie przez podwójnego noblistę Frederica Sangera pierwszorzędowej struktury insuliny - jej sekwencję aminokwasową.
Wraz z ustalaniem sekwencji kolejnych białek stało się jasne, iż generowane dane są zbyt duże, by mogły być wygodnie przechowywane i analizowane bez wyspecjalizowanych narzędzi.

Pionierką informatyzacji biologicznych danych została Margaret Oakley Dayhoff, twórczyni pierwszego zbioru wszystkich znanych sekwencji białkowych. Jej opublikowany w 1965 roku *Atlas of Protein Sequence and Structure*, zawierający 65 sekwencji, wydany był jeszcze w wersji książkowej. Po wielokrotnych aktualizacjach stał się podstawą opublikowanej w 1984 roku pierwszej bazy sekwencji białkowych online (Protein Information Resource).

Prace Dayhoff objęły nie tylko zbieranie sekwencji białkowych, lecz również utworzenie pierwszych algorytmów i programów porównujących sekwencje białkowe, tworząc podwaliny analizy porównawczej i filogenetycznej.

Równolegle powstawały analogiczne zbiory sekwencji nukleotydowych. Pierwszą opublikowaną online bazą sekwencji DNA jest GenBank, uruchomiony 1982 roku, a więc niedługo przed PIR. Uruchomienie GenBank można uznać za symboliczną datę początków współczesnej bioinformatyki.

OMÓWIENIE

Jako rozwiązanie zadania prosimy o przesłanie **gotowej do uruchomienia aplikacji** oraz wszystkich plików źródłowych. Źródła możecie udostępnić jako archiwum w serwisie chmurowym (Dysk Google, OneDrive, etc) lub projekt w systemie kontroli wersji (np. GitHub).

Rozwiązania desktopowe będą testowane w systemie Windows 11 i do pracy w takim systemie powinny być skompilowane. W przypadku aplikacji mobilnych testować będziemy w symulatorze Androida pod kontrolą systemu w wersji 10. Serwery aplikacji webowych będą uruchamiane pod kontrolą systemu Windows 11, a same aplikacje będą testowane w przeglądarce Firefox w wersji 104 lub nowszej.

CO OCENIAMY

1. Zgodność rozwiązania ze specyfikacją (zobacz opis etapów zadania)
2. Zaimplementowaną architekturę aplikacji (**0-10**)
3. Dokumentację projektu (**0-10**). Dokumentacja może zawierać wszystko czym chcielibyście się pochwalić, czego się nauczyliście, na co powinniśmy zwrócić uwagę oceniając wasz program. Dopuszczalna jest forma pamiętnika ;)
4. Zaproponowany interfejs aplikacji. Jego wygląd, czytelność i ogólne wrażenie użytkownika z użytkowania aplikacji (**0-25**)
5. Testy automatyczne (**0-5**)

SPECYFIKACJA

Po uruchomieniu aplikacja powinna pozwolić na ręczne wprowadzenie sekwencji RNA lub wczytanie jej z pliku. Jeśli wprowadzona sekwencja jest sekwencją DNA, należy na potrzeby projektu zamienić Tyminę (T) na Uracyl (U).

Na podstawie wprowadzonej sekwencji program powinien wygenerować sekwencje aminokwasów, a w efekcie znaleźć wszystkie potencjalne białka kodowane w zadanej sekwencji RNA. Forma i sposób prezentacji aminokwasów i białek jest dowolna. Jednak będziemy oceniać zaproponowany przez was sposób wizualizacji, jej czytelność i użyteczność.

ETAPY ZADANIA

Odczyt kodu genetycznego

Za ten etap zadania uzyskacie do **35** punktów.

Kod genetyczny odczytywany jest przez rybosomy. Jest to “urządzenie” komórkowe, które potrafi odczytywać kod RNA i przetworzyć go na serię aminokwasów, które na końcu procesu replikacji staną się białkami. W kodzie RNA występują cztery zasady azotowe: adenina (oznaczona literą A), cytozyna (C), guanina (G) oraz uracyl (U). Zasady te w strukturze RNA nazywamy **nukleotydami**. Trzy kolejne nukleotydy nazywamy **kodonem**, który to koduje określony aminokwas.

Przykładowy kod RNA wygląda następująco:

AAAUGAACGAAAAUCUGUUCGCUUCAUUCAUUGCCCCCACAAUCCUAGGCCUACCC

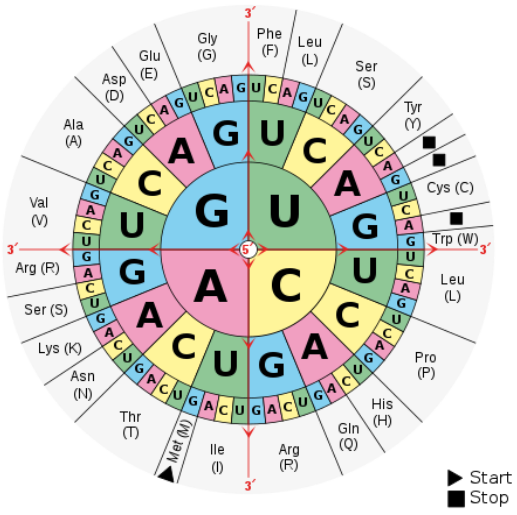
Ponieważ rybosom może rozpocząć replikację RNA od dowolnego miejsca kodu, odczyt należy wykonać trzykrotnie z odpowiednim przesunięciem. Poniżej trzy zestawy zakodowanych nukleotydów dla trzech różnych przesunięć:

+1: AAA UGA ACG AAA AUC UGU UCG CUU CAU UCA UUG CCC CCA CAA UCC UAG GCC UAC

+2: AAU GAA CGA AAA UCU GUU CGC UUC AUU CAU UGC CCC CAC AAU CCU AGG CCU ACC

+3: AUG AAC GAA AAU CUG UUC GCU UCA UUC AUU GCC CCC ACA AUC CUA GGC CUA CCC

Powyższe zestawy kodonów należy zamienić na serię aminokwasów zgodnie z poniższym schematem:



Aminokwasu szukamy od środka okręgu. Na przykład kodon UCA koduje Serynę (S), a GAC koduje Kwas asparginowy (D).

Niektóre aminokwasy mają specjalne znaczenie. Metionina (M) opisana kodonem AUG jest znakiem początku białka, a kodony UAA, UAG i UGA są znacznikiem końca przepisu na białko.

W naszym przykładzie trzeci nukleotyd zaczyna się od kodu AUG, ale kod ten może wystąpić w dowolnym miejscu łańcucha RNA. Waszym zadaniem jest znalezienie wszystkich potencjalnych białek.

W naszym pierwszym przykładzie będzie to seria aminokwasów określonych kodami:

KWTKICSLHSLPPQS[stop]

zaś w trzecim otrzymamy zestaw

M(start)NENLFASFIAPTILGLP

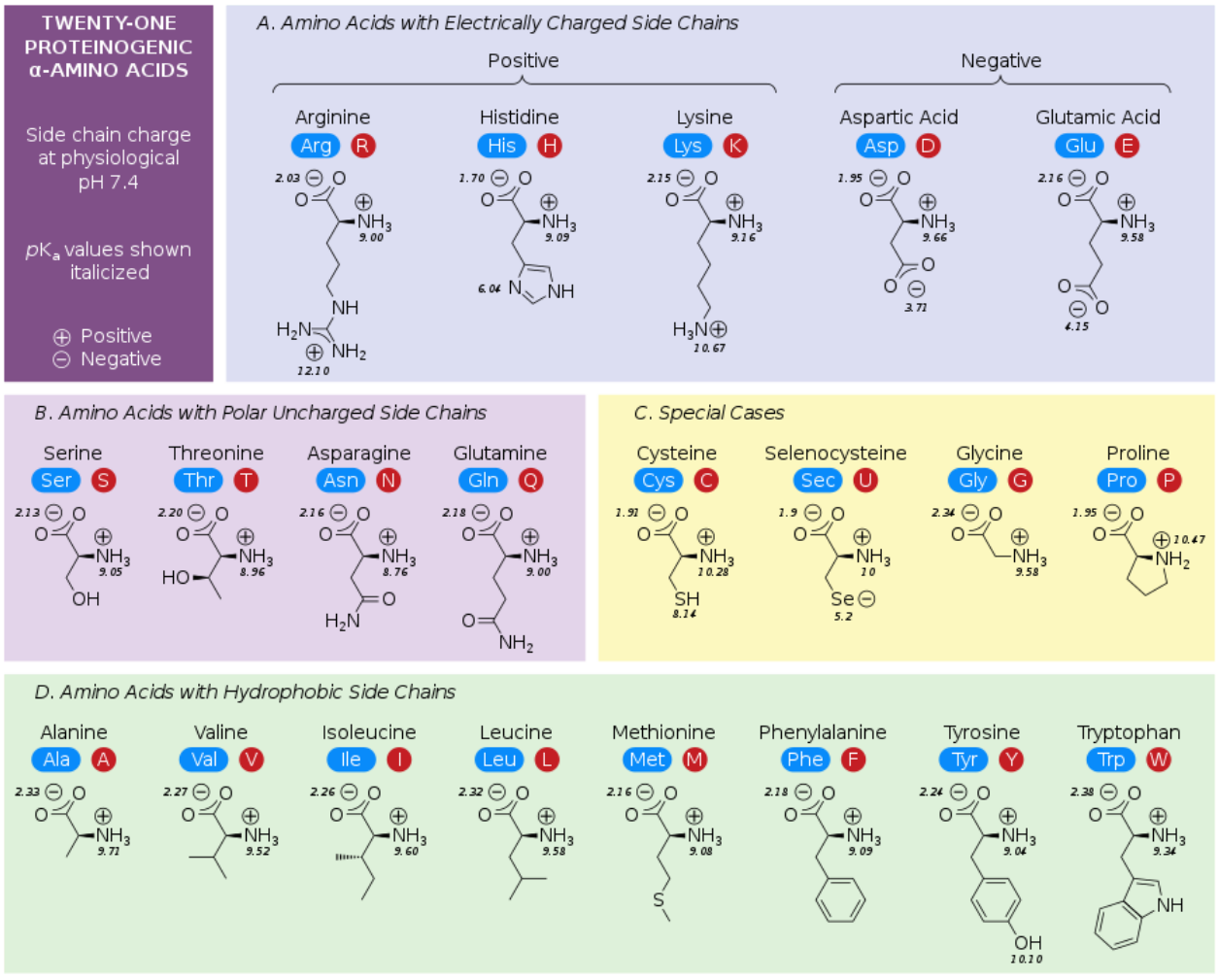
Wasz program jako potencjalne białka pokazuje tylko takie zestawy aminokwasów które zaczynają się kodonem (start) i kończą jednym z kodonów (stop). Niezależnie od długości białek, których łańcuchy mogą być zarówno bardzo długie jak i bardzo krótkie.

Przedstawcie znalezionych kandydatów na ekranie w sposób, który sami wymyślicie.

Wizualizacja kandydatów na białka

Ten etap wart jest do **25** punktów

Białka są długimi łańcuchami aminokwasów. Każdy aminokwas można przedstawić w formie wzoru strukturalnego, który pokazuje faktyczny, “przestrzenny” wygląd cząsteczki. Poniżej lista wszystkich interesujących nas aminokwasów oraz ich wzorów strukturalnych:



Pokażcie na ekranie wzór strukturalny znalezionego białka. Tak, będzie to bardzo długi obrazek.

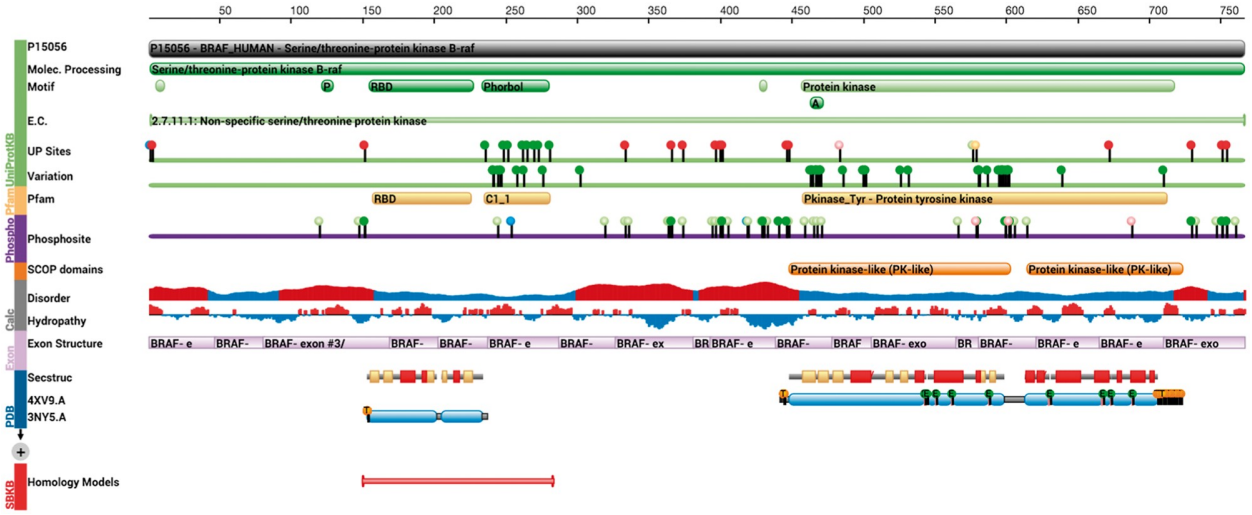
Ponadto policzcie masę znalezionej cząsteczki.

Jeszcze więcej wykresów i diagramów

Za zrealizowanie poniższych wymagań możecie zyskać do **65** punktów

Zbudujcie diagram pokazujący fizyczne właściwości naszego białka. Na przykład indeks hydrofobowy, indeks pH, polarność, punkt izoelektryczny oraz inne wskaźniki.

Punktem wyjścia do tego zadania mogą być informacje zawarte na stronie https://en.wikipedia.org/wiki/Amino_acid, a sam diagram może wyglądać na przykład tak:



Jak zacząć?

Zadanie powstało z inspiracji wykładami doktora Łukasza Lamży:
https://youtu.be/mBvD0qrQtU4
https://youtu.be/SHNRg0_Ly5M
Najlepiej zacząć od ich obejrzenia ;)

Oraz kilka artykułów z Wikipedii:
https://en.wikipedia.org/wiki/Genetic_code
https://en.wikipedia.org/wiki/Protein_biosynthesis
https://en.wikipedia.org/wiki/Amino_acid

Prawdziwe sekwencje RNA i DNA przeróżnych organizmów znajdziecie na stronie GenBank:
https://www.ncbi.nlm.nih.gov/genbank/

A na koniec w ramach inspiracji, komputerowa wizualizacja genomu człowieka.