Instytut Informatyki i Matematyki Komputerowej UJ
opracowanie: Ewa Matczyńska, Jacek Śmietański

Zadania bioinformatyki

1. Program kursu i zasady zaliczenia

Repozytorium: https://github.com/dadoskawina/Bioinformatics_lecture_2018

Program kursu przedstawiony został na pierwszym wykładzie. Tematyka laboratoriów będzie się mniej więcej pokrywać z tematyką wykładów.

Szczegółowe zasady zaliczenia dostępne są w osobnym dokumencie na repozytorium.
Tam też będą umieszczane prezentacje z wykładów oraz materiały z laboratoriów.

2. Cele i znaczenie bioinformatyki

Zadanie 1. (2 pkt)
Na podstawie wiadomości z wykładu oraz dyskusji na forum grupy napisz jakie znaczenie **Twoim zdaniem** ma bioinformatyka w codziennym życiu. Czy potrafisz wskazać obszary życia, które rozwój badań bioinformatycznych może zmienić (lub już zmienił) na lepsze albo na gorsze.
Jakie zagadnienia bądź problemy bioinformatyczne są Twoim zdaniem najważniejsze? Jakie uważasz za najbardziej interesujące?

3. Budowa nici DNA

Organizmy zbudowane są z komórek. Każda komórka to skomplikowany system współdziałających ze sobą elementów. Szacuje się, że ciało człowieka zawiera ok $6*10^{23}$ komórek różnych typów. Istnieją organizmy jednokomórkowe, np. bakterie czy drożdże jak i wielokomórkowe, np. człowiek.
Cała informacja potrzebna do konstrukcji i funkcjonowania komórki zakodowana jest w DNA znajdującym się w jądrze komórkowym. DNA *(Deoxyribonucleic acid)* jest nośnikiem informacji genetycznej u niemal wszystkich znanych organizmów żywych, ma strukturę podwójnej helisy.

Helisa jest zbudowana z dwóch nici, z których każda składa się z powtarzalnych elementów tzw. nukleotydów, które są zbudowane z cukru, reszty fosforanowej i zasady azotowej. W DNA mamy 4 rodzaje zasad azotowych, są to adenina (A), guanina (G), cytozyna (C) i tymina (T). Z powodu struktury chemicznej zasady azotowe mogą łączyć się w ściśle określone pary: A łączy się z T, a C łączy się z G, jest to tzw. zasada komplementarności. Wiązania wodorowe, które wytwarzają się między komplementarnymi zasadami stabilizują helisę. Zróćmy uwagę, że obydwie nici tworzące helisę są w pewien sposób zorientowane. Na schematycznym rysunku można zauważyć tzw. końce 5' i 3' – pochodzą one z nazewnictwa

chemicznego i informują jaka grupa jest dołączona na końcu nici. Jeśli jest to grupa fosforanowa to mamy koniec 5', jeśli hydroksylowa (-OH) to mamy koniec 3'. Komórka zawsze „czyta" sekwencję od końca 5' do końca 3' i domyślnie wszystkie sekwencje w bazach danych zapisywane są w tą stronę. Długość nici DNA mierzona jest w ilości zasad azotowych, ang. *base pairs* (bp).

Zadanie 2.

Obejrzyj jak wygląda struktura DNA w trójwymiarze:
http://www.rcsb.org/pdb/explore/jmol.do?structureId=1BNA

- najlepiej ustaw *display options*: Style - *Ball and Stick*, Color - *By Element*, zaznacz checkbox *H-bonds*.
Jakie kolory odpowiadają poszczególnym atomom?

Autorami modelu podwójnej helisy są James Watson i Francis Crick, zaproponowali oni tą strukturę w 1953 roku, otrzymując za nią później nagrodę Nobla.

W komórce organizmów eukariotycznych DNA zorganizowane jest w chromosomy, które są bardzo dobrze widoczne podczas podziału komórki, gdyż wtedy DNA jest w nich bardzo ciasno upakowane. Gdy komórka nie podlega podziałowi, DNA jest w luźnej postaci.
U człowieka mamy 23 pary chromosomów.
Gdyby rozwinąć DNA z 23 par chromosomów i połączyć, uzyskalibyśmy podwójną helisę długości ok.2m.

Genomem nazywamy całą sekwencję DNA zawartą w pojedynczym zestawie chromosomów. U człowieka genom ma długość ok 3.2 mld bp, przy czym sekwencja genomu jest w 99.9% taka sama dla wszystkich ludzi. Natomiast 98% sekwencji jest identyczne jak u szympansa.

Wydawać by się mogło, że im bardziej skomplikowany organizm tym dłuższy powinien być genom, jednak znane są genomy organizmów wiele razy dłuższe niż genom człowieka.

Rysunek 1. Wielkość wybranych genomów (źródło: http://blogs.biomedcentral.com/on-biology/wp-content/uploads/sites/5/2014/03/genomelog030.jpg)

Zadanie 3.
Jakie dostrzegasz korzyści z tego, że DNA ma budowę podwójnej helisy? (w stosunku do nici pojedynczej; weź pod uwagę, że DNA jest nośnikiem informacji)

Zadanie 4.
Odległość między nukleotydami na helisie DNA wynosi 0,34nm = 0,34 * 10^9m. Oblicz ile bajtów można zapisać na 1 cm DNA?

4. Budowa i znaczenie białek

Instrukcje zawarte w genomie potrzebne do funkcjonowania komórki jak i całego organizmu to w istocie przepisy na budowę odpowiednich białek. Białka są kluczowe dla działania naszego organizmu:
- katalizują reakcje chemiczne: enzymy
- odpowiadają za układ odpornościowy
- regulują pracę organizmu: hormony, np. insulina
- są receptorami światła, neuroprzekaźników
- pełnią funkcje transportowe, np. hemoglobina
- budują mięśnie, ścięgna, włosy

Białka są zbudowane z aminokwasów, których mamy 20 rodzajów oznaczanych przez litery alfabetu: A R N D C E Q G H I L K M F P S T W Y V. Proces budowy białka polega na sekwencyjnym dołączaniu kolejnych aminokwasów, które tworzą łańcuch. Na skutek oddziaływań fizycznych, łańcuch aminokwasów zwija się i formuje strukturę przestrzenną białka. Sekwencja aminokwasów determinuje strukturę 3D białka, a co za tym idzie, jego funkcję.

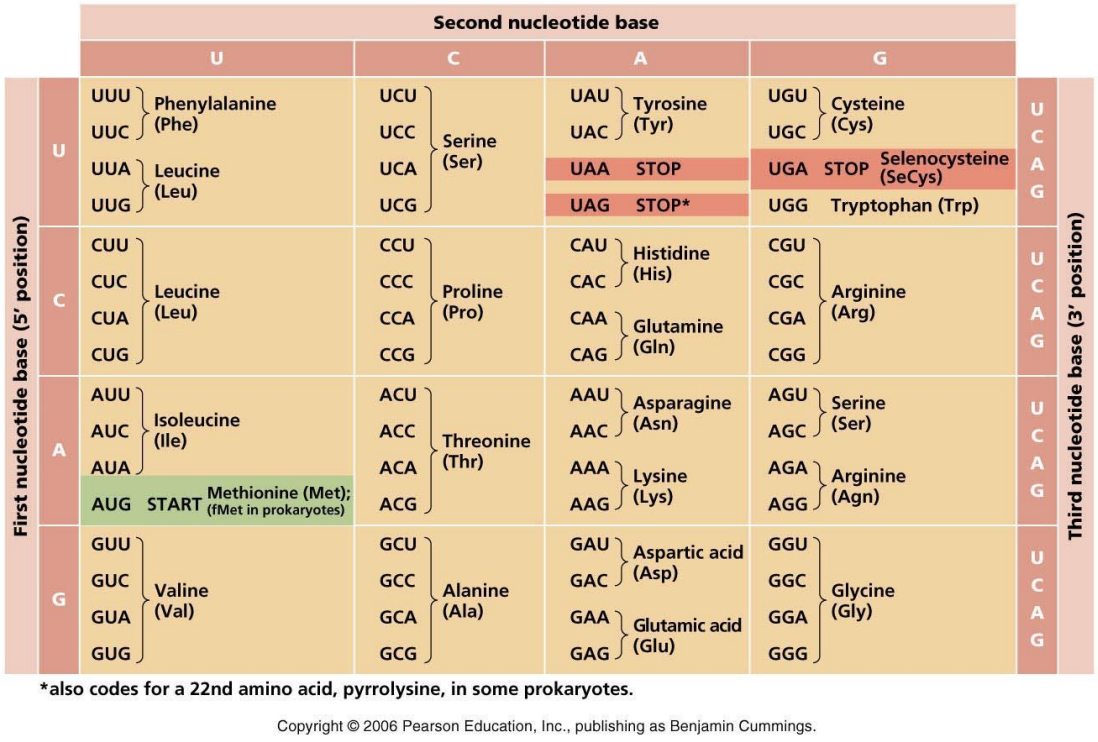
Zadanie 5.
Obejrzyj strukturę białka. Struktury przechowywane są w bazie PDB (Protein Data Bank): http://www.rcsb.org
Możesz wybrać sobie np. hemoglobinę: nieutlenowana, id: 2HHB; utlenowana, id: 1HHO.
Dla zainteresowanych lektura o znaczeniu i roli hemoglobiny:
http://pdb101.rcsb.org/motm/41
Jak oceniasz funkcjonalność bazy i łatwość nawigacji?

5. Kod genetyczny i centralny dogmat

Instrukcja budowy każdego białka czyli sekwencja aminokwasów białka jest zakodowana w genomie. Fragment DNA, który zawiera przepis na taką sekwencję nazywamy genem. Geny stanowią ok. 2% genomu człowieka, a ich liczbę szacuje się na ok. 20 tysięcy. Komórka „wie" gdzie w genomie znajduje się gen danego białka, przepisuje tą instrukcję na chwilową, jednoniciową kopię nazywaną mRNA. Cząsteczka mRNA jest bardzo podobna do DNA, z tą różnicą, że zamiast nukleotydu z tyminą (T), mamy nukleotyd z uracylem (U).

Pozostaje pytanie, jak zakodowana jest sekwencja 20 aminokwasów za pomocą 4 nukleotydów. Tłumaczenie sekwencji mRNA na sekwencję aminokwasów odbywa się za

pomocą kodu genetycznego, który jest uniwersalny dla wszystkich organizmów. Każde 3 nukleotydy kodują aminokwas, bądź oznaczają koniec tłumaczenia, są to tzw. kodony stop. Ponieważ mamy $4^3 = 64$ możliwości na zakodowanie 20 aminokwasów, niektóre aminokwasy są kodowane przez więcej niż jedną trójkę nukleotydów, jest to tzw. degeneracja kodu.



Rysunek 2. Kod genetyczny (źródło: http://adegene.com/content/standard-genetic-code)

Podsumowując, otrzymujemy tzw. centralny dogmat biologii molekularnej, który określa przepływ informacji w komórce. Sekwencja genu w DNA jest przepisywana na mRNA, na którego podstawie tworzone jest białko. Jeśli w danej chwili na podstawie genu produkowane jest białko to mówimy, że ulega on ekspresji. Każda komórka organizmu posiada ten sam genom, jednak różnią się one od siebie dlatego, że w komórkach różnych typów ekspresji ulegają różne białka.

Bioinformatyka zajmuje się stosowaniem narzędzi matematycznych i informatycznych do rozwiązywania problemów z nauk biologicznych. Obejmuje takie zagadnienia jak badanie struktury, funkcji, ewolucji genów, białek, całych genomów. Podstawowe problemy to analiza sekwencji, analiza struktur 3D białek, analiza ekspresji, interakcje molekularne, projektowanie leków, modelowanie systemów biologicznych, analiza obrazów z wysokoprzepustowych technik eksperymentalnych, obrazów medycznych i wiele innych.

Zadanie 6.

Jak sądzisz, dlaczego kod genetyczny jest zdegenerowany?

Zadanie 7. (2pkt)

Napisz funkcję **`translate(sequence)`** zwracającą sekwencję aminokwasów, przetłumaczoną zgodnie z kodem genetycznym dla zadanej na wejściu sekwencji nukleotydowej.

Swoją funkcję wywołaj dla sekwencji z załączonego pliku (hemoglobin.txt). Wykonaj tłumaczenie, rozpoczynając również od drugiego i od trzeciego nukleotydu, zobacz jak wpływa to na zmianę wyprodukowanego białka.

Wskazówki:

1. Tabela kodu genetycznego:

```
gencode = { 'ATA':'I', 'ATC':'I', 'ATT':'I', 'ATG':'M',
            'ACA':'T', 'ACC':'T', 'ACG':'T', 'ACT':'T',
            'AAC':'N', 'AAT':'N', 'AAA':'K', 'AAG':'K',
            'AGC':'S', 'AGT':'S', 'AGA':'R', 'AGG':'R',
            `CTA':'L', 'CTC':'L', 'CTG':'L', 'CTT':'L',
            'CCA':'P', 'CCC':'P', 'CCG':'P', 'CCT':'P',
            'CAC':'H', 'CAT':'H', 'CAA':'Q', 'CAG':'Q',
            'CGA':'R', 'CGC':'R', 'CGG':'R', 'CGT':'R',
            'GTA':'V', 'GTC':'V', 'GTG':'V', 'GTT':'V',
            'GCA':'A', 'GCC':'A', 'GCG':'A', 'GCT':'A',
            'GAC':'D', 'GAT':'D', 'GAA':'E', 'GAG':'E',
            'GGA':'G', 'GGC':'G', 'GGG':'G', 'GGT':'G',
            'TCA':'S', 'TCC':'S', 'TCG':'S', 'TCT':'S',
            'TTC':'F', 'TTT':'F', 'TTA':'L', 'TTG':'L',
            'TAC':'Y', 'TAT':'Y', 'TAA':'*', 'TAG':'*',
            'TGC':'C', 'TGT':'C', 'TGA':'*', 'TGG':'W' }
```

Gwiazdka (*) oznacza kodon STOP. Po jego przeczytaniu, translacja powinna się zakończyć.

2. Funkcja **`line.rstrip()`** usunie znaki końca linii z pliku.

3. Funkcja **`range(start, stop, step)`** ułatwi iterowanie po sekwencji

Rozwiązania zadań 1 i 7 prześlij mailem do wtorku, **9.10.2018** włącznie, na adres:
jacek.smietanski@ii.uj.edu.pl
Temat wiadomości proszę opatrzyć przedrostkiem **[Bio] Lab 01**. Rozwiązaniem mają być **dwa pliki** – dokument PDF dla zadania 1 oraz skrypt zgodny z Pythonem w wersji 3.x**,** zawierający implementację zadania 7. Proszę o nazwanie pliku wg schematu**: Imie.Nazwisko.nr_zadania.rozszerzenie**.