# Bioinformatyka - laboratorium 4

**Czas na oddanie finalnego raportu**: 6 tygodni

**Nazwa pliku**: imie_nazwisko_4_bio.pdf

**Typ ćwiczenia**: jednotygodniowe

**Cel**: Celem ćwiczenia jest zapoznanie studenta z podstawami analizy i przewidywania struktury białek.

## Zadania

> *Homo sapiens* jest nazwą gatunkową człowieka. "Homo" to rodzaj wskazujący na przynależność do rodziny hominidów, czyli ssaków naczelnych zaś "sapiens" oznacza "rozumniejszy" lub "mądry". Jest to nazwa, która została nadana człowiekowi przez naukowca Carolusa Linneausa w XVIII wieku.

1. Zainstaluj PyMOL 2.5 lub skorzystaj z tego narzędzia, jeśli znajduj się już na komputerze.
   - Wejdź do bazy danych: www.rcsb.org a następnie wyszukaj "protein homo sapiens"
     - w "Scientific Name of Source Organism" zaznacz Homo sapiens
     - w "Polymer Entity Type" zaznacz "Protein"
     - w "SCOP Classification" zaznacz "Small Proteins"
   - Pobierz plik PDB z dowolnego ze znalezionych rekordów (podaj link do wybranego rekordu [**task 1**]) i otwórz go w PyMOL. Zrób zrzut ekranu uzyskanego obrazu struktury białka i umieść go w raporcie [**task 2**].
   - Z prawej strony obok nazwy pliku widnieją przyciski A S H L C, kliknij S, następnie "as", następnie "sticks". Zrób zbliżenie tej struktury, a następnie wykonaj zrzut ekranu [**task 3**]. Wpisz w google "sticks chemical visualization" i wyjaśnij, co widać na ekranie, tzn. na czym polega ten sposób wizualizacji białka [**task 4**].
   - Wróć do poprzedniego wyglądu struktury białka, klikając "as", następnie "cartoon". Znajdź i krótko opisz co to za sposób prezentacji struktury białka [**task 5**].
   - Zmień widok struktury na "ribbon" i również wykonaj zrzut ekranu [**task 6**] i skrótowo wyjaśnij co to za sposób prezentacji struktury białka [**task 7**].

Algorytm GOR (predictive algorithm for protein secondary structure) jest jednym z wielu algorytmów używanych do predykcji struktury białka na podstawie sekwencji aminokwasów. Jego działanie polega na analizie sekwencji aminokwasów i identyfikacji wzorców, które są charakterystyczne dla różnych rodzajów struktur białka.

2. Z wykorzystaniem narzędzia GDR wykorzystującego algorytm GOR dokonaj predykcji struktury białka, którego sekwencja znajduje się tutaj: link. Po wklejeniu sekwencji kliknij "start prediction".
   - Spróbuj zinterpretować na podstawie wykresu położenie alpha-helixes, beta-loops oraz tzw."hot spots" [**task 8**]. Podpowiedź: jeśli obie krzywe są pod osią X to najprawdopo-

dobniej mamy do czynienia z tzw. “hot spot” (a connecting point between the regular secondary structures (alpha-helixes and beta-sheets)).
- Na podstawie tych materiałów www.gersteinlab.org lub innych, które znajdziesz, podaj, co oznaczają litery w kolumnie “St” tabeli [**task 9**].

3. Z wykorzystaniem narzędzia TMHMM, nie zmieniając ustawień, dokonaj predykcji struktury białka podanego w zakładce “Guide” jako przykład.

   - Co oznacza w raporcie pole “Number of predicted TMHs” [**task 10**]?
   - Podaj liczbę widoczną obok “Exp number of AAs in TMHs” i zinterpretuj ją (czyli napisz co ona oznacza) [**task 11**].
   - Zinterpretuj wykres “Plot of probabilities” [**task 12**].
   - Przeprowadź tę samą analizę narzędziem DeepTMHMM i porównaj oraz krótko opisz wyniki obu analiz [**task 13**].
   - Zwizualizuj to białko z użyciem MembraneFold i wykonaj zrzut ekranu [**task 14**], jakie narzędzia wykorzystuje MembraneFold celem tworzenia wizualizacji [**task 15**].

**NCBI BLAST** (Basic Local Alignment Search Tool) to narzędzie stosowane do wyszukiwania podobieństw sekwencji nukleotydów lub aminokwasów w bazach danych. Jest to jedno z najbardziej popularnych narzędzi do porównywania sekwencji i jest szeroko stosowane w biologii molekularnej i genetyce. Algorytm korzysta z metody kontekstowej, która pozwala na znalezienie fragmentów sekwencji, które są podobne do siebie, a następnie przypisuje im punkty za podobieństwo. Na końcu algorytm tworzy raport, w którym wyświetla wyniki wyszukiwania, w postaci listy sekwencji, które są najbardziej podobne do sekwencji wejściowej.

**Accession number** jest unikalnym numerem identyfikacyjnym przypisanym do każdego rekordu w bazie danych NCBI (National Center for Biotechnology Information). Ten numer jest używany do identyfikacji i odnalezienia konkretnego rekordu w bazie danych, takiego jak sekwencja nukleotydów, sekwencja białka lub publikacja naukowa. Przykład “accession number” to np. NM_001114443.1

4. Badaną sekwencję białkową tj.:

```
MSEAAHVLITGAAGQIGYILSHWIASGELYGDRQVYLHLLDIPPAMNRLTALTMELEDCAF
PHLAGFVATTDPKAAFKDIDCAFLVASMPLKPGQVRADLISSNSVIFKNTGEYLSKWAKPSV
KVLVIGNPDNTNCEIAMLHAKNLKPENFSSLSMLDQNRAYYEVASKLGVDVKDVHDIIVWG
NHGESMVADLTQATFTKEGKTQKVVDVLDHDYVFDTFFKKIGHRAWDILEHRGFTSAASPT
KAAIQHMKAWLFGTAPGEVLSMGIPVPEGNPYGIKPGVVFSFPCNVDKEGKIHVVEGFKVN
DWLREKLDFTEKDLFHEKEIALNHLAQGG
```

umieść jako input w narzędziu: NCBI Blast Proteins tzn. wklej w pole z nagłówkiem “Enter accession number(s), gi(s), or FASTA sequence(s)” , w database, wybierz bazę „pdb” i kliknij “Blast”. Skopiuj accession number pierwszego rekordu, który został znaleziony (Zakładka “Descriptions” –> accession number pierwszego rekordu z kolumny “Accession”). W bazie danych RCSB wklej pierwsze 4 znaki ze znalezionego uprzednio accession number i znajdź rekord dla sekwencji, która okazała się najbardziej homologiczna (w uproszczeniu – najbardziej podobna, umieszczona jest jako pierwszy rekord w wynikach) do tej, którą wkleiłaś/eś w NCBI Blast (po Accession Number) i pobierz plik pdb ze strukturą tego białka. Do raportu podaj jej PDB DOI z bazy RCSB [**task 16**].

- Załóż konto tutaj: robetta.bakerlab.org i w zakładce submit we właściwym polu umieść

swoją badaną sekwencję.
- Zaznacz pole CM, a jako strukturę homologiczną wybierz pobrany uprzednio plik `pdb`.
- Po skompletowaniu analizy otrzymasz e-mail. Wejdź w odnośnik, który prowadzi z e-maila do prognozowanej struktury. Zrób zrzut ekranu uzyskanej struktury białka [**task 17**].

Uzyskana struktura białka zapisana jest w formacie PDB (Protein Data Bank). Znajduje się w przysłanym do Ciebie e-mailu, ale możesz też zobaczyć ją, klikając „View" w raporcie na stronie internetowej. Każdy wiersz w pliku PDB opisuje jedną część białka (np. atom, grupę atomów itp.). W tym fragmencie pliku, każdy wiersz "ATOM" opisuje pojedynczy atom. Kolumny w wierszu "ATOM" opisują informacje takie jak numer identyfikacyjny atomu, typ atomu, nazwę reszty aminokwasowej, pozycję przestrzenną (współrzędne x, y, z) oraz inne parametry.

- Jaki jest pierwszy aminokwas z tego białka (zobacz na pierwsze wiersze "ATOM" w pliku pdb, a następnie trzyliterowy symbol powtarzający się w pierwszych 19 wierszach i za pomocą wyszukiwarki google dowiedz się co to za aminokwas) [**task 18**]?
- W każdym wierszu rozpoczynającym się od „ATOM" znajdują się trzy cyfry, jak np. w pierwszym jest „12.556 20.247 27.259" Poszukaj dokumentacji plików PDB i odpowiedz na pytanie, co to jest [**task 19**]?

5. Przeprowadź analizę powyższego białka z użyciem narzędzia https://modbase.compbio.ucsf.edu/modweb/ i podając "Modeller license key" jako "MODELIRANJE", a w "Fold assignment methods" wybierz "Fast". Wyjaśnij co oznacza wybór opcji „Fast" [**task 20**].

    - Po uzyskaniu wyników wklej dane znajdujące się pod "Run-Name:" a nad "This dataset has been successfully loaded into ModBase." [**task 21**].
    - Wejdź w link "This dataset has been successfully loaded into ModBase." Kliknij zielony pasek, żeby poznać szczegóły modelu.
    - Na podstawie dokumentacji oceń czy zastosowany model należy uważać za wiarygodny (m.in. spójrz na wartości parametrów: E-Value, GA341, MPQS, zDOPE) [**task 22**].
    - Pobierz plik `pdb` i obejrzyj strukturę swojego białka w PyMOL, wykonaj zrzut ekranu tej struktury [**task 23**].

### Zadanie dla chętnych

[**task 24**] Napisz w Pythonie skrypt, który w sekwencji jakiegokolwiek białka, które wprowadzi na wejściu użytkownik (input) zaznaczy prolinę (symbol jednoliterowy: P), np. poprzez wstawienie symbolu * przed każdą proliną. Prolina nie występuje w strukturach alpha-helix. Zbudujesz w ten sposób proste narzędzie do bardzo wstępnej predykcji struktury białka:

Przykładowe dane wejściowe:

```
MSEAAHVLITGAAGQIGYILSHWIASGELYGDRQVYLHLLDIPPAMNRLTALTMELEDCAFPHLAGFVA
TTDPKAAFKDIDCAFLVASMPLKPGQVRADLISSNSVIFKNTGEYLSKWAKPSVKVLVIGNPDNTNCEI
AMLHAKNLKPENFSSLSMLDQNRAYYEVASKLGVDVKDVHDIIVWGNHGESMVADLTQATFTKEGKTQK
VVDVLDHDYVFDTFFKKIGHRAWDILEHRGFTSAASPTKAAIQHMKAWLFGTAPGEVLSMGIPVPEGNP
YGIKPGVVFSFPCNVDKEGKIHVVEGFKVNDWLREKLDFTEKDLFHEKEIALNHLAQGG
```

Przykładowe dane wyjściowe:

```
MSEAAHVLITGAAGQIGYILSHWIASGELYGDRQVYLHLLDI*P*PAMNRLTALTMELEDCAF*PHLAG
FVATTD*PKAAFKDIDCAFLVASM*PLK*PGQVRADLISSNSVIFKNTGEYLSKWAK*PSVKVLVIGN*
```

PDNTNCEIAMLHAKNLK*PENFSSLSMLDQNRAYYEVASKLGVDVKDVHDIIVWGNHGESMVADLTQAT
FTKEGKTQKVVDVLDHDYVFDTFFKKIGHRAWDILEHRGFTSAAS*PTKAAIQHMKAWLFGTA*PGEVL
SMGI*PV*PEGN*PYGIK*PGVVFSF*PCNVDKEGKIHVVEGFKVNDWLREKLDFTEKDLFHEKEIALN
HLAQGG