

Moleculaire Biologie in Code, Fase II, script 2

In deze opdracht zoek je informatie over taaislijmziekte en analyseer je aan de hand van een ProteinDataBank-bestand de drie-dimensionale structuur van een eiwitmolecuul.

Verplichte onderdelen:

- Download het ProteinDataBank-bestand behorende bij het eiwit betrokken bij het ontstaan van taaislijmziekte in de mens (met ID: 5UAK).
- Schrijf een programma dat het molekuulgewicht van het eiwit schat door de atoomgewichten van de samenstellende atomen op te tellen. (Hint: gebruik de regels met tag "ATOM"; PDB-documentatie hierover vind je op <https://www.wwpdb.org/documentation/file-format-content/format23/sect9.html#ATOM>.) De gebruiker geeft de gewenste bestandsnaam aan met een command-line argument (bv. `script_2_2.py protein.pdb`). Waarom zal je geschatte molekuulgewicht systematisch te laag uitvallen?
- Breid je programma uit zodat het ook de aminozuursequentie inleest. Vertaal hierbij de 3-letter aminozuur codes uit het PDB bestand naar de bijbehorende 1-letter codes, en sla het resultaat op in een string. (Hint: zoek de regels met tag "SEQRES"; zie de documentatie op <https://www.wwpdb.org/documentation/file-format-content/format23/sect3.html#SEQRES>.) Laat je programma een nauwkeurigere schatting van het molekuulgewicht berekenen door de molekuulgewichten van de samenstellende aminozuren op te tellen; let hierbij op de gevolgen van condensatie/dehydratatie.
- Bepaal vervolgens middels het PDB-bestand waar zich alpha-helices en beta-sheets bevinden. Selecteer alleen de aminozuren die zich in een helix en sheet bevinden en plak die aan elkaar tot één helix- en één sheet-sequentie. Bepaal voor elke aminozuur of het een voorkeur heeft voor een helix of een sheet.
- Hint: de relevante regels worden met de labels "HELIX" en "SHEET" aangegeven; lees <https://www.wwpdb.org/documentation/file-format-content/format23/sect5.html> om te bepalen welke delen je moet hebben om het begin en eind van een helix of sheet uit te knippen.)
- Download eveneens de bijbehorende aminozuur-sequentie van het taaislijmziekte-eiwit van NCBI (met ID: NP_000483.3). Lees ook deze aminozuur-sequentie in en print beide sequenties naar scherm, op zo'n manier dat de gebruiker eenvoudig de eventuele verschillen kan zien.
- Bepaal de gemiddelde hydrofobiciteit van het eiwit. De hydrofobiciteit of hydrophobiciteit geeft aan hoeveel een aminozuur "van water houdt", oftewel erin oplost, met consequenties voor het eiwit in z'n omgeving in de cel. Doe ook weer hetzelfde met de helix- en sheet-sequenties afzonderlijk. (Hint: op http://en.wikipedia.org/wiki/Hydrophilicity_plot staat een tabel met de waarden per aminozuur.)
- Bepaal tenslotte van de eiwit-sequentie als geheel de compositie en geef die in de terminal weer middels een histogram (m.b.v: '#', bepaal zelf hoeveel procent elke # moet weergeven). Doe hetzelfde met de helix- en sheet-sequenties afzonderlijk. Bijvoorbeeld:

```
A ###
L ###
Q #
T #####
R #
...
```

Bonus-onderdelen:

- Download ook PDB-bestanden (en overeenkomstige FASTA-bestanden) behorende bij een aantal andere eiwitten en bekijk of je programma deze bestanden correct kan verwerken. Wat gaat er mis?

Lever je programma in via eJournal in de vorm van een python-bestand genaamd `script2_2.py`.

Let hierbij op dat je code:

- ✓ op tijd wordt ingeleverd;
- ✓ zonder foutmeldingen uitvoerbaar is;
- ✓ de verplichte onderdelen van de opdracht correct en volledig verricht;
- ✓ liefst zo efficiënt mogelijk geïmplementeerd is;
- ✓ een netjes leesbare programmeerstijl gebruikt;
- ✓ en – optioneel – de bonus-onderdelen juist uitvoert.