Moleculaire Biologie in Code, Fase II, script 1

In deze practicum-opdracht zoek je informatie over borstkanker en maak je een script dat – net als een ribosoom – reeksen nucleotiden omzet in reeksen aminozuren.

Verplichte onderdelen:

- Het BASC-complex is een samenstelling van eiwitten die tezamen fouten in DNA kan repareren. Als één van de eiwitten een afwijking bevat kan dit tot early-onset borstkanker 1 leiden. Zoek hierover informatie en download de DNA-, mRNA-, en aminozuur-referentiesequentie(s) van het (gezonde) gen en eiwit die deze vorm van borstkanker veroorzaken in FASTA-formaat.
- Schrijf een programma dat de nucleotidesequentie uit een FASTA-bestand weergeeft op het scherm. De sequentie dient te worden opgesplitst in tripletten, elk bestaande uit 3 nucleotiden. Laat het programma alle tripletten achtereenvolgens op een eigen regel op het scherm tonen. De header regel dient niet getoond te worden.
- Laat het programma elk triplet vervolgens omzetten in het bijbehorende aminozuur. Bijvoorbeeld: *ACATTTGCTTCTGACA...* wordt *TFASD...*. Pas je programma aan zodat niet de tripletten maar de aminozuren worden weergegeven. Als er een aantal nucleotiden aan het eind van de sequentie overblijven dan dient je programma dit aan de gebruiker te melden. (Hint: bewaar de codon-tabel op https://en.wikipedia.org/wiki/DNA_codon_table in een geschikt type variabele.)
- Laat het programma de gevormde aminozuur-sequentie opslaan in een nieuw FASTA-bestand op schijf in plaats van deze weer te geven op het scherm. Geef het uitvoerbestand dezelfde header als die van het ingelezen DNA-bestand. Zorg dat de aminozuur-sequentie wordt weggeschreven in regels van 70 tekens. De gebruiker dient de gewenste bestandsnamen aan te geven voor zowel in- als uitvoerbestanden door middel van command-line argumenten (bv. `script_2_1.py nucleotides.fasta aminoacids.fasta`).
- Open de gedownloade aminozuur sequentie en vergelijk deze met de sequentie die door jouw programma gegenereerd is. Zijn er verschillen? Hangt dit af van of het om DNA of mRNA gaat, en in welke vorm je de nucleotidesequentie hebt bewaard (“Complete Record”/“Coding Sequences”/“Gene Features”)?
- Geef de gebruiker de mogelijkheid om met een extra (optioneel) argument het leesraam (“reading frame”) aan te geven in de vorm van startpositie 1, 2, of 3 van het eerste codon. Vanaf startpositie 2 wordt *ACATTTGCTTCTGACA...* bijvoorbeeld *HLLLT...*.
- Generaliseer je programma zodat het met multi-FASTA bestanden kan omgaan. Een invoerbestand met meerdere nucleotide-sequenties dient dan te worden omgezet in een uitvoerbestand met meerdere overeenkomstige aminozuur-sequenties. (Hint: informatie over het multi-FASTA formaat vind je op https://en.wikipedia.org/wiki/FASTA_format.)
- Organiseer je programma zodat het gebruik maakt van functies. Gebruik bijvoorbeeld afzonderlijke functies die:
  - een bestand inlezen;
  - een string omzetten;
  - een resultaat wegschrijven.

Lever je programma in via eJournal in de vorm van een python-bestand genaamd `script_2_1.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.