Moleculaire Biologie in Code, Fase II, script 3

In deze opdracht ga je een bestaand online bio-informatica tool nabouwen dat kenmerkende delen van sequenties uit GenBank-bestanden verzamelt en hiervan een samenvatting weergeeft.

Verplichte onderdelen:

- Kopieer de inhoud van je taaislijmziekte mRNA GenBank-bestand naar de online GenBank Feature Extractor op http://www.bioinformatics.org/sms2/genbank_feat.html (*). Upload deze informatie met de Submit-knop en kopieer de inhoud van de pagina die dan verschijnt naar een tekst- bestand. (Hint: kopieer alleen de inhoud in monospaced font; de titel "GenBank Feature Extractor results" hoef je niet mee te kopiëren.)
- Bestudeer de uitvoer nauwkeurig en vergelijk deze met de invoer die je had ingegeven. Analyseer hoe de uitvoer is verkregen uit de invoer. Verifieer je conclusies met GenBank-bestanden van mRNA sequenties behorende bij andere aandoeningen die je in dit practicum hebt bekeken. (Hint: de secties DEFINITION, FEATURES, en ORIGIN zijn belangrijk.)
- Schrijf een programma dat een door de gebruiker gespecificeerd willekeurig mRNA GenBank- bestand kan inlezen en dat een tekstbestand met alle features uitvoert op exact dezelfde wijze zoals de GenBank Feature Extractor dat doet. Als de invoer bestaat uit een bestand genaamd *mrna.gb*, dan dient het programma de uitvoer weg te schrijven in een bestand genaamd *mrna_features.txt*. (Hint: maak zelf eerst op papier een stappenplan hoe je dit gaat aanpakken.)
- Ga na hoe de uitvoer verandert als je in de online GenBank Feature Extractor de optie met het uitvoerformaat verandert van "separated" in "uppercased". (Hint: kijk niet alleen naar het hoofdlettergebruik, maar ook naar de bestandsgrootte van het geproduceerde bestand.) Voorzie je eigen programma van een optie waarmee de gebruiker kan kiezen voor het "separated" of "uppercased" formaat, en zorg dat de uitvoer daarmee overeenkomt.
- Ga na dat je programma eveneens de exacte uitvoer reproduceert voor de mRNA GenBank bestanden van de genen behorende bij dystrofine, C-C chemokine receptor type 5, Hemofilie A stollingsfactor, sikkelcelanemie, early-onset borstkanker 1, en neurofibromatose.

Bonus-onderdelen:

- Zet met de online GenBank Feature Extractor ook de inhoud van een GenBank-bestand met DNA om in een tekstbestand met features. Pas je eigen programma zo nodig aan dat het ook deze uitvoer kan reproduceren.
- Zoek een eiwit-sequentie op en download deze in GenPept-formaat. Probeer ook deze om te zetten met de online GenBank Feature Extractor. Ga na dat dit (meestal) niet helemaal goed gaat, maar zorg ervoor dat je eigen programma deze omzetting wél kan maken.

Lever je programma in via eJournal in de vorm van een python-bestand genaamd `script2_3.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.

* Soortgelijke online GenBank Feature Extractors zijn te vinden op:
http://www.geneinfinity.org/sms/sms_gbfeatures.html,
http://www.cellbiol.com/sequence_manipulation_suite/genbank_feat.php,
http://groups.molbiosci.northwestern.edu/matouschek/links/sms2/genbank_feat.html (en elders).