Thematoets Kans 2

Bi2a-T Course 2a 2016-2017
Bio-informatica 1: Donderdag 13-04-2017 09.30-13.30 uur.

1 Casus

Je loopt stage bij een onderzoeksafdeling waarbij gistcellen, *Saccharomyces cerevisiae*, gebruikt worden voor het produceren van olie. *S. cerevisiae* is een goed te onderzoeken model organisme, omdat het genoom uit slechts iets meer dan 6000 bekende genen bestaat. Je hebt het bestand **yeast_orf_coding.fasta** tot je beschikking, waarin alle coding sequences van alle ruim 6000 genen van gist beschreven staan. Hieronder vind je een stukje van het bestand.

```
>YAL001C TFC3 SGDID:S000000001, Chr I from 151006-
147594,151166-151097, Genome Release 64-2-1, reverse
complement, intron sequence removed.....
ATGGTACTGACGATTTATCCTGACGAACTCGTACAAATAGTGTCTGATAAAA
TTGCTTCAAATAAGGGAAAAATCACTTTGAATCAGCTGTGGGATATATCTGG
TAAATATTTTGATTTGTCTGATAAAAAAGTTAAACAGTTCGTGCTTTCATGC
GTGATATTGAAAAAGGACATTGAG...

>YAL002W VPS8 SGDID:S000000002, Chr I from 143707-
147531, Genome Release 64-2-1, Verified ORF, "Membrane-
binding component of the ...
ATGGAGCAAAATGGCCTTGACCACGACAGCAGATCTAGCATCGATACGACTA
TTAATGACACTCAAAAGACTTTCCTAGAATTTAGATCGTATACCCAATTAAG
TGAAAAACTGGCATCTAGTTCTTCATATACGGCACCTCCCCTGAACGAAGAT
GGTCCTAAAGGGGTAGCTTCTGCA...
```

2 Taak

Van je stagebegeleider heb je eerder een opdracht gekregen om de status van bepaalde genen te bekijken, maar nu is je stagebegeleider benieuwd naar een aantal eigenschappen van de sequenties. Hij wil graag voor iedere sequentie de volgende informatie weten:

- GC%
- AT%
- Lengte van de sequentie

Verder zou hij graag de volgende informatie gemiddeld over alle sequenties willen weten:

- Ratio van stopcodon gebruik (dus hoe vaak TAA/TGA/TAG voorkomt)
- Gemiddelde lengte sequenties

Omdat het hier om coding sequences gaat, ga je er vanuit dat elke sequentie met 'ATG' begint en met een stopcodon eindigd. Dit wil je natuurlijk wel controleren voor iedere sequentie voordat je verder gaat met de sequentie.

De output kan er als volgt uitzien, maar voel je vrij om je eigen draai eraan te geven.

```
Eigenschappen gist:
*******************************************
Gemiddelde lengte coding seq:    512 bp
Ratio stopcodons (TAA:TAG:TGA): 2458:784:3012
*******************************************
************************
Acc.code:        YAL002W
GC%:             54%
AT%:             46%
Lengte:          487 bp
************************
.....
```

Tip: Begin met het extraheren van de informatie per gen en deze naar de output te printen voordat je aan de totalen begint.

3 Opdracht

Ga uit van het bestand **yeast_orf_coding.fasta**.

Het op te leveren programma dient aan de volgende **functionele eisen** te voldoen:

1. Leest het bestand.
2. Bij ontbreken van het bestand of een leesfout genereert het programma een nette foutmelding.
3. Het script doorzoekt het bestand en berekend de gevraagde eigenschappen per gen en voor het totaal.
4. Het programma geeft output waarbij de gebruiker een overzicht krijgt van de eigenschappen van de genen en het totaal aan genen

Het op te leveren programma dient aan de volgende **technische (niet functionele) eisen** te voldoen:

1. Er is een flowchart.
   - Ontwerp een flowchart voor de functie *get_genes()* en schrijf de bijhorende code.
   - De flowchart mag digitaal of op papier.
   - De in te leveren flowchart beschrijft de functie *get_genes()* nauwgezet.
2. Het script is opgedeelt in functies. De volgende functies zijn in ieder geval aanwezig.
   - De functie *get_genes()* leest het bestand aan en maakt een lijst met hierin alle genen, voor ieder gen is er een accessiecode en een sequentie
   - De functie *get_gene_info()* accepteert een gen en berekent hiervoor de gewenste eigenschappen.
   - De functie *get_total()* accepteert de gehele lijst aan genen en berekend over het totaal de gewenste gegevens.
   - De functie *main()* roept al deze functies aan en verwerkt eventuele output naar het scherm.
3. Alle te verwachten excepties worden afgevangen.
4. De snelheid van het programma doet er niet toe.
5. Het programma is geschreven in Python.

Naam student:

4 Beoordeling

In te leveren:

- Programma met bestandsnaam Bi1-OWE2a-<studentnaam>.py
- Flowchart voorzien van naam. Digitaal in gangbaar bestandsformaat (JPEG/PNG/BMP/PDF/DOC).
- Onderstaand formulier voorzien van inschatting van eigen functioneren en naam.

Beoordelingsformulier Thematoets

		Punten	Beoordeling	
			Student	Docent
I.	Code			
	a. Maakt gebruik van variabelen en gebruikt de juste datatypes	10		
	b. Past de juste controlestructuren toe	10		
	c. Bouwt programma op volgens standaard met een onderverdeling in functies	10		
	d. Schrijft functies met parameterisering	10		
	e. Schrijft fucties die waardes retourneren	10		
	f. Documentatie conform de standaard	10		
II.	Juiste Werking			
	a. Bestand lezen	10		
	b. Exception Handling	20		
	c. Voert de gevraagde bewerking juist uit	30		
III.	Flowchart			
	a. Flowchart is volgens standaard	10		
	b. Flowchart stemt overeen met de code	20		
Eindoordeel		150		