תרגיל תכנות – מבוא ביולוגיה של התא – 2021

תאריך הגשה: 20.1.22

*הדרכה על קריאת רצפי DNA ופונקציות מומלצות מופיעה בסוף התרגיל.

**אנא הוסיפו הסברים ותיעוד (documentation) לפונקציות שאתם כותבים. מהם הinputs והoutputs איזה סוג אובייקט הם, ומה המשמעות שלהם. בשורות חשובות בפונקציה הוסיפו הערה עם הסבר על משמעות השורה (לא צריך בכל שורה ובפרט לא בשורות פשוטות שניתן להבין מקריאה של הקוד). חוסר תיעוד יוכל להוביל להורדת ציון.

1) מצורף לכם קובץ המכיל קטעי Open Reading Frame (ORF) של גנים, כלומר הקטע בגן שמקודד לחומצות אמינו ומתורגם לחלבון, עבור הבקטריה E.Coli.

a. כתבו פונקציה count_codons(orf) אשר מקבלת רצף (string של אותיות המסמנות נוקלאוטידים) ORF (ניתן להניח שכבר השלשה הראשונה היא קודון) ומוציאה array כאשר כל שורה בה הוא וקטור של כמות החזרות של הקודונים המקודדים את כל אחת מחומצות האמינו Lysine, Leucine , Isoleucine:

קודון	חומצת אמינו
AAA, AAG	Lysine
CTA, CTC, CTG, CTT, TTA, TTG	Leucine
ATA, ATC, ATT	Isoleucine

כלומר הפלט יהיה:

[[#AAA, #AAG], [#CTA, #CTC, #CTG, #CTT, #TTA, #TTG], [#ATA, #ATC, #ATT]]

למשל עבור הORF הבא (עם פסיקים לשם קריאות – בקובץ אין פסיקים):

ATG,AAA,ATA,ATT,ATC,CCA,ATA,TAA

נקבל:

[[1, 0], [0, 0, 0, 0, 0, 0], [2, 1, 1]]

b. כתבו פונקציה נוספת calc_codon_bias(file_name) המקבלת קובץ המכיל מספר OFRים של אותו אורגניזם, מחשבת את כמות החזרות של כל קודון בכל

ORF ע"י קריאה לcount_codons ועם המידע הזה- מחשבת את ההסתברות לקבל כל קודון מתוך כל הקודונים האפשריים עבור כל אחת מחומצת האמינו בכל הORFים שבקובץ. הפונקציה תוציא array באשר כל שורה בה הוא וקטור של הסתברות של הקודונים המקודדים את כל אחת מחומצות האמינו Lysine, Leucine , Isoleucine. על הפונקציה גם לייצר bar graph של התפלגות הקודונים עבור כל אחת מחומצות האמינו. לדוגמא, נניח שיש לנו 2 ORFs:

ATG,AAA,ATA,ATT,ATC,CCA,ATA,TAA
ATG,CTA,ATA,TTG,TTA,AAA,GGG,TAG

אז נקבל את ההסתברויות הבאות:

קודון	חומצת אמינו
AAA=2/2=1, AAG=0	Lysine
CTA=1/3=0.33, CTC=0, CTG=0, CTT=0, TTA=1/3=0.33, TTG=1/3=0.33	Leucine
ATA=3/5=0.6, ATC=1/5=0.2, ATT=1/5=0.2	Isoleucine

כלומר הפלט יהיה:

[[1, 0], [0.33, 0, 0, 0, 0.33, 0.33], [0.6, 0.2, 0.2]]

c. האם ההתפלגויות שקיבלתן.ם הן מה שהיה מצופה אם הגנום היה מורכב מנוקלאוטידים באופן רנדומלי (כל נוקלאוטיד הופיע מספר זהה של פעמים בגנום באופן רנדומלי)? הסבירו
d. לכל יצור יש הטייה כלשהי בבחירת הקודונים (Codon Usage Bias), והיא משפיעה על היבטים רבים בתהליכים הקשורים לביטוי גנים. תארו מקרה בו שינוי בנוקלאוטידים שמהווים קודון עבור חומצת אמינו כלשהי (נוקלאוטידים בDNA), יכול להשפיע על תהליך השחבור (splicing) של אותו גן.

2) בקובץ winrar בשם Bacillus Genomes נתונים לכם.ן 11 גנומים של בקטריות מסוג bascilus. בנוסף, נתון לכן.ם בקובץ Bascilluis_Opt_T.csv את הטמפ' האופטימלית לגידול שלהן.

a. כתבו פונקציה GC_cont = calc_GC_cont(genome) המקבלת רצף ומחשבת את רמת הGC Content שיש בו.

$$GC\ Content = \frac{\#\ of\ times\ G\ appeared +\ \#\ of\ times\ C\ appeared}{\#(of\ times\ A\ or\ G\ or\ C\ or\ T\ appeared)}$$

שימו לב שמכיוון הGC content הוא של גנום, אז המכנה במשוואה לחישובו הוא אורך הגנום. למשל עבור הרצף באורך 10: TTCACACCGG הGC Content יהיה 6/10=0.6

b. כתבו פונקציה:

multi_GC_cont , optimal_temp = calc_multi_GC(folder_path)
המקבלת כתובת של תיקייה ומוציאה וקטורים של multi_GC_cont ו-optimal_temp בגודל כמות הרצפים (קבצי fasta) המופיעים בתיקייה.
השתמשו בפונקציה calc_GC_cont בשביל החישובים בפונקציה הנ"ל. כמו כן, הפונקציה הנ"ל תצטרך ליצור גרף (scatter plot) של הGC Content כתלות בטמפ' גידול האופטימלית של אותו אורגניזם. לגרף צריך להיות כותרת, שמות מתחת לצירים (עם יחידות – צלזיוס לטמפ' ואחוזים לGC Content). צרו תיקייה בה יש את הגנומים (לאחר שהוצאו מהwinrar) ואת Bascilluis_Opt_T.csv והשתמשו בפונקציה הנ"ל כשהיא מקבלת כאינפוט את הכתובת של התיקייה.

c. מהגרף שיצרתם, מה הקשר המסתמן בין טמפ' הגידול האופטימלית ל GC Content? מה יכולה להיות הסיבה הביולוגית לכך?

3) כתבו פונקציה - find_orf(sequence) המקבל רצף ומוצאת ORFים פוטנציאלים עבור רצף כלשהו, והשתמשו בה עבור רצף הDNA (המלאכותי) בקובץ seq_3b.txt. על מנת שתת-רצף יחשב כORF פוטנציאלי, הוא צריך להיות באורך שמתחלק ב3, להתחיל עם קודון התחלה (ATG) ולהסתיים בקודון עצירה (TAA\TAG\TGA). שימו לב כי יש רק קודון עצירה אחד בכל ORF פוטנציאלי, והוא תמיד בסופו - לא יכול להיות למשל:

ATGCGA**TAG**GGG**TGA**

לעומת זאת, ATGCGA**TAG** זה ORF פוטנציאלי תקין (בעיקרון גם ATGTAA אבל אין כאלה...). בנוסף, יכול להיות ATGים גם באמצע הORF (ולא רק אחד בהתחלה).

הפונקציה צריכה להחזיר מטריצה Xnum_ORFs2 שכל עמודה מכילה את קואורדינטות ההתחלה והסיום של כל ORF פוטנציאלי. אפשר להניח שהקלט תקין.

לאחר כתיבת הפונקציות, כתבו script ראשי אשר טוען את הקבצים, קורא לפונקציות ועונה על השאלות.

פונקציות שימושיות לתרגיל:

- על מנת לייצר את הגרפים בצורה נוחה, מומלץ להתקין (pip install) את הספרייה matplotlib ולייבא (import) ממנה את matplotlib.pyplot as plt. בספרייה זו תוכלו להשתמש בפונקציות plt.subplot, axvline, plot, set_title, set_xlim, set_ylim, legend.
- כדאי להשתמש בספריה numpy (import numpy as np) כדי לייצר מערכים ולהשתמש בהם בצורה נוחה (np.array, np.zeros)
- על מנת לקרוא רצפי DNA (אשר מופיעים בפורמט fasta) התקינו את הספריה biopython:

בגוגל colab ניתן להתקין ספריות ע"י:

```
!pip install biopython
```

ואז יבאו את SeqIO:

```
from Bio import SeqIO
```

על מנת לקרוא את הקבצים השתמשו ב:

```
record = SeqIO.read('file_path', "fasta")
```

הפכו את record לרשימה של אותיות:

```
record = str(record.seq)
```

record יהיה רשימה של אותיות

- על מנת לקרוא את רצפי הORFs הנתונים בשאלה 1 בקובץ csv אפשר להשתמש בpandas (שכבר מותקן בגוגל קולאב- אז לא צריך להתקין שם):

יבאו את pandas:

```
Import pandas as pd
```

על מנת לקרוא את הקבצים השתמשו ב:

```
Orfs = pd.read_csv('file_path')
```

על מנת לגשת לORF בשורה ה-iRow נכתוב:

```
Curr_orf = Orfs.iloc[iRow,0]
```

- על מנת לגשת לטמפ' של יצור מסוים בקובץ עם הטמפ' (בשאלה 2)– ניתן גם להשתמש בpandas. נניח שקראנו את הקובץ למשתנה temps אז הגישה לטמפ' בשורה ה-iRow תהיה ע"י temps.iloc[iRow,1] (מכיוון שהן בעמודה השנייה ולא הראשונה). שם היצור אליו שייכת הטמפ' יהיה בtemps.iloc[iRow,0]
- כדי לקרוא קבצים מתיקייה יש לייבא את הפונקציה listdir מהספריה os:

from os import listdir

וכדי ליצור את רשימת הקבצים נקרא ל(files = listdir(mypath.

הגשה:

אנא הגישו את התרגיל (קוד והסברים) בעזרת קובץ (.ipybn) jupyter notebook זהו פורמט אינטראקטיבי בו אפשר לכתוב קוד, לייצר גרפים ולכתוב טקסט על "דף" אחד. ניתן לייצר קובץ זה בקלות דרך google colab (ניתן לכתוב שם את הקוד וההסברים מההתחלה) ע"י לחיצה על:

File -> Download .ipynb.

לקובץ קראו בשם biocell_Name1_ID1_Name2_ID2 לדוגמא:

biocell_OdedScharf_123456789_ShaiCohen_234567890

הערה כללית:

המטרה של תרגיל זה היא להתחיל להקנות לכם.ן הרגלי תכנות נכונים. שניים מההרגלים הבסיסיים הם תיעוד נכון של הקוד וחלוקת הקוד לפונקציות קטנות שכל אחת עושה פעולה קטנה במקום script אחד ארוך שעושה המון דברים. שני ההרגלים האלה מאפשרים קוד מסודר וקל לקריאה ולהבנה. תכונות אלו מאפשרות לנו לחזור ולהשתמש בקוד זמן רב אחרי שכתבנו אותו וגם לאפשר לאחרים להבין ולהשתמש בו. לכן חשוב מאוד להקפיד עליהן כבר מההתחלה