**פרוייקט סוף קורס – קורס ביולוגיה חישובית**

**מגישים : אליעזר רווח - \_\_\_\_\_\_\_\_\_ , ינון עמוסי - \_\_\_\_\_\_ , יאן גנם - \_\_\_\_\_\_\_\_\_**

**חלק א – איסוף ועיבוד מידע אודות גנום החיידק בצילוס סבטיליס**

**סעיף 1 -** הכרת וספירת האלמנטים בגנום  
הסבר: בשאלה זו נדרשנו למנות את מספר סוגי אזורי הגנים ולחשב כמה גנים יש מכל אחד מהאיזורים.  
דרך: קריאת קובץ Genbank משיכת נתוני שדה ה'features' וקריאה לפונקציה 'countDnaArea' העוברת על כל רשומת features שהיא מסוג גן; בודקת את סוג הגן ומונה. בסופו של דבר הפונקציה מחזירה מילון כנדרש.  
תוצאה: ניתו לראות את פלט הפונקציה בקונסול:תמונה שמכילה טקסט, צילום מסך, צג, שחור

התיאור נוצר באופן אוטומטי

**סעיף 2 -** אפיון אורכי גנים  
הסבר: בשאלה זו נדרשנו למנות את אורכי כל הגנים, לחלק את הגנים למתורגמים לחלבון ואלה שלא, לדווח סטטיסטיקות אודות שתי הקבוצות ולצייר היסטוגרמות של האורכי הגנים.  
דרך: קריאת קובץ Genbank משיכת נתוני שדה ה'features' וקריאה לפונקציה 'getAllGeneLength' העוברת על כל רשומת features שהיא מסוג גן; ומחשבת עבור כל גן את האורך שלו לפי ערך הlocations. בסופו של דבר הפונקציה מחזירה list of list ולאחר מכן מומרת לdataframe.  
תוצאה: הדפסה של האורכים – ראה **נספח א(2)**ג': סטטיסטיקה של הגנים המקודדים בהשוואה לאלו שלא:   
תמונה שמכילה טקסט, צג, צילום מסך, מקורה

התיאור נוצר באופן אוטומטי

את ההיסטוגרמות ניתן לראות ב**נספח א(2)**. ניתן לראות כי ההיסטוגרמות משקפות נכונה את הסטטיסטיקה שחישבנו בסעיף הקודם. בעוד ציר הX של ה-coding DNA נמשך עד למעלה מ15000 הnon coding DNA רק עד אזור 3000. כמו כן ניתן לראות כי יש הרבה יותר coding DNA מאשר .non coding DNA. ניתן לראות שקיימת שכיחות נמוכה מעל לאורך 2000 coding dna hist' '. ושכיחות נמוכה בגני הnon coding מעל אורך 300.

**סעיף 3 -** חישוב אחוז GC בגנים:  
הסבר: בשאלה זו נדרשנו לחשב אחוז GC בגנים.  
דרך: לשם חישוב הGC בכל הDNA משכנו את הרצף כולו, מנינו את מספר הGC וחילקנו באורך הרצף כולו. לשם חישוב ה%GC רק בגנים שמתוגרמים לחלבון משכנו רק את הגנים האלה וספרנו את כמות הGC וחילקנו באורך הכולל של כל הגנים האלה.  
תוצאה: ממוצע על פני כל הרצף: 43.51440813017155%, ממוצע על פי הגנים המקודדים: 44.25276114196909%  
הערות: ההיסטוגרמה בסעיף זה ו5 הגנים עם אחוז הGC הגבוה והנמוך בנספחים, בקובץ 'part\_a.csv' מופיעים הגנים המקודדים וממוצע הGC (ללא הכפלה ב100) לכל אחד מהם.  
תשובה לשאלה המילולית: הגנים מתחלקים לכאלה שמקודדים לחלבון וכל השאר. מצאתי מההיסטוגרמות, שכמות הגנים הלא מקודדים קטן לעומת כמות הגנים המקודדים. האורך הממוצע של גן שלא מתורגם לחלבון קטן יותר מהאורך הממוצע של גן המתורגם לחלבון. ניתן לראות את זה מהתוצאה של 2.ג. (סטטיסטיקות לגבי ממוצע האורכים של גנים).  
%GC קשור לרמת העמידות של הגן - ככל ש GC % גבוה יותר כך העמידות של הגן גבוהה יותר. משום שאורכי האזורים המקודדים לחלבון גדולים יחסית כפי שניתן לראות מההיסטוגרמות של סעיף ב' ומהתוצאה של שאלה 2.ג, ומשום שצריך דיוק בתהליך התרגום של אזורים המקודדים לחלבון הייתי מצפה שיהיה באזורים אלו %GC גבוה. העובדה שיש מעט גנים שלא מקודדים לחלבון יחסית לאלה שכן, גרמה לי לצפות לGC% דומה פחות או יותר בכל הגנום כולו(כי ההשפעה שלהם נמוכה).  
ואכן: בסעיף א' התוצאה: the %GC of whole DNA seq is 43.51440813017155

**סעיף 4 -** בדיקות עקביות בקובץ הדאטה  
בשאלה זו בדקתי עקביות החלבון המוצג ביחס לתרגום עצמאי של קטעי הDNA וניסיתי להתחקות אחר הסיבה לכך. עברתי על כל הגנים שמתרגמים לחלבון ותרגמתי עצמאית מתוך הDNA וערכתי השוואה בין החלבון המוצג בקובץ GB לבין התרגום שלי. מבין אלה שאין התאמה בדקתי לפי קריטריונים מהי הבעיה וכתבתי את התוצאות לקובץ כנדרש. הקריטריונים הם:  
Inconsistent in protein length- חוסר עקביות באורך של החלבון.  
Wrong nucleotide- האורך שווה אך חוסר התאמה של חומצת אמינו אחת או יותר.  
Missing start/end codon- בדיקה האם חסר בגן קודון התחלה או קודון עצירה (עוד לפני התרגום).  
nonNucleotide in gene- בדיקה שאין אותיות חוץ מארבעת הנוקלאוטידים(ACGT).  
התוצאות בקובץ gene\_exceptions.csv.

**חלק ב – אנליזת חלבונים בעזרת אתר ה UniProt**

**סעיף א**הסבר: בסעיף זה התבקשנו להצליב בין הנתונים שיש לנו בקובץ ה GeneBank לבין הקובץ שהורדנו מ UniProt. את ההצלבה ביצעתי ע"י המזהה 'GOA'. מזהה זה הינו משותף ל2 המאגרים ומזהה חלבונים באופן ייחודי.  
דרך: על מנת לבצע את ההצלבה באופן האידיאלי מבחינת זמני ריצה, הכנתי 2 מילונים. מיליון עבור הערכים מ UniProt, ועבור הערכים מ GeneBank. בחרתי במילון על מנת לבצע שליפה ב O(1) על מנת לבצע את פעולת ההצלבה המלאה ב O(n). המפתח במילונים אלו היה המזהה 'GOA' ו Name והערכים היו רצף הגנים. כלומר סהכ 4 מילונים.  
תוצאות:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| מקור | ערכים קיימים | ערכים עם מזהה GOA | ערכים עם מזהה Name | הצלבה ע"פ GOA | הצלבה ע"פ Name |
| UniProt | 8541 | 8541 | 6938 | 3381 | 3715 |
| GeneBank | 4237 | 3381 | 4171 |

הערות: נראה כי הרישומים הקיימים ב GeneBank עבור החלבונים - חסרים ולא מושלמים. כאשר קיימים חלבונים שלא כוללים את המזהה 'GOA'. להערכתי המזהים החסרים נובעים מכך שמאגר ה GeneBank עיקרו הינו גנים ולא חלבונים, לעומת UniProt שזה עיקר ייעודו. בנוסף אף על פי שיש מיפוי רחב יותר של GOA ב UniProt ההצלבה המוצלחת יותר בוצעה ע"י שם החלבון (כנדרש בתרגיל).  
צילומי מסך: ראה נספח **חלק ב(א)**

**סעיף ב**הסבר: בסעיף זה התבקשנו לזהות את החלבונים המכילים חלקים טרנסממברנלי,ולאפיין אותם.  
דרך: בעזרת מאגר UniProt הוספתי עמודה בשם Transmembrane, ע"פ עמודה זו ניתן לזהות את החלבונים המכילים את האיזורים האלו.  
תוצאות: עבור חלבונים אלו -> האורך הממוצע - 238.58, האורך המקסימלי- 493, האורך המינימלי- 134. צילום מסך של ההיסטוגרמה ניתן למצוא ב**נספח חלק ב(ב).**הערך הממוצע של חומצות האמינו ההידרופוביות בחלבונים אלו הינו : 0.496, זוהי כמות נכבדת מאוד של חומצות אמינו הידרופוביות, וזה אכן תואם את הציפיות שלנו מחלבונים המכילים מקטעים טרנסממברנלי. פירוט על כל חלבון ניתן לראות ב**נספח חלק ב(ב)**

**סעיף ג**הסבר: בסעיף זה נדרשנו לנתח את רצפי הגנים המקודדים (CDS), עם השוואה בין טרנסממברנלי לרגילים.  
דרך: סיוע בספריית Matplolib על מנת להכין את ההיסטוגרמות.  
תוצאות: את התוצאות ניתן לראות בנספח **חלק ב(ג)**

**נספחים**

**נספח חלק א(2)**

תמונה שמכילה טקסט, צג, צילום מסך, מקורה

התיאור נוצר באופן אוטומטי

ההיסטוגרמות משאלה 2 בחלק הראשון:

גרף של אורכי הגנים המקודדים:

Chart, histogram

Description automatically generated

גרף של אורכי הגנים הלא מקודדים:

Chart, histogram

Description automatically generated

אורכי כל הגנים:

Chart, histogram

Description automatically generated

הערה: כל מלבן מחזיק בתוכו טווח של 100 אורכים בלבד.(0-100,100-200....)

שאלה 3:

היסטוגרמת %GC

Chart, histogram

Description automatically generated

5 הגנים העשירים\דלילים בGC:

5 הראשונים הכי דלילים בGC

והאחרונים הכי עשירים.

תמונה שמכילה טקסט, צילום מסך, צג, אלקטרוניקה

התיאור נוצר באופן אוטומטי

**נספח חלק ב(א)**

להלן פלט של התכנית, הערה חשובה את gn len הינו מספר החלבונים המכילים את המזהה 'GOA'

Text

Description automatically generated

**נספח חלק ב(ב)**היסטוגרמה עבור האורכים של החלבונים המכילים קטעים מסוג טרנסממברנלי

Chart, histogram

Description automatically generated

התפלגות חומצות האמינו ההידרופוביות על-פי החלבונים המכילים מקטעים טרנסממברנלי  
{'yrbG': 0.44954128440366975, 'ydbS': 0.5094339622641509, 'ydfS': 0.4723404255319149, 'ydbT': 0.44421906693711966, 'yshB': 0.5706214689265536, 'yxlG': 0.5077519379844961, 'ydfR': 0.4888888888888889, 'yetF': 0.42857142857142855, 'yshE': 0.4626865671641791, 'ykjA': 0.4732510288065844, 'yoaK': 0.5911111111111111, 'ecfT': 0.5622641509433962}

**נספח חלק ב(ג)**מניסוח השאלה לא היה ברור האם הנתונים הרצויים הינם עבור המופעים של GC או עבור האורך של החלבונים. ולכן עשיתי את שניהם:

עבור האורכים: Graphical user interface, chart

Description automatically generated

עבור ה GC:  
Graphical user interface, chart, histogram

Description automatically generated  
נתונים נוספים:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | קבוצה | מינימום | מקסימום | ממוצע | חציון |
| אורך חלבונים | A | 63 | 16467 | 884.881 | 765 |
| B | 405 | 1482 | 718.75 | 696 |
| ספירת GC | A | 21 | 7597 | 44.230 | 765 |
| B | 179 | 654 | 42.828 | 698 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Text  Description automatically generated | Text  Description automatically generated | Text  Description automatically generated | Text  Description automatically generated |  |