**מגישים**

מחמד חדר זעתרי – 314929811

ספאא נתשה – 207711359

**◄ חלק א' - איסוף ועיבוד מידע אודות גנום החיידק בצילוס סבטיליס**

1.

יצרנו מילון, עברנו על כל הגנים ובדקנו את הסוג שלו וספרנו אותו לתוך המילון. אחר כך הוצאנו את ה source כי הוא לא רלוונטי.

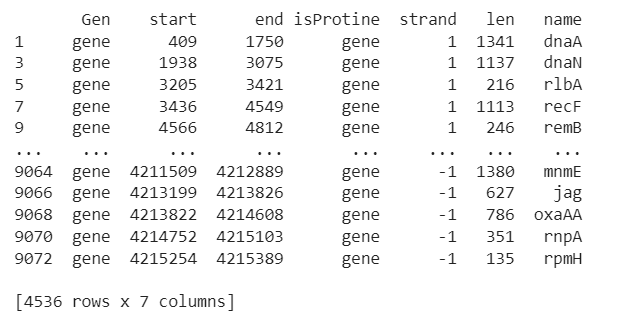
את התוצאה: 

2.

2.1 יצרנו מילון והוספנו לו את הפרטים של כל גן

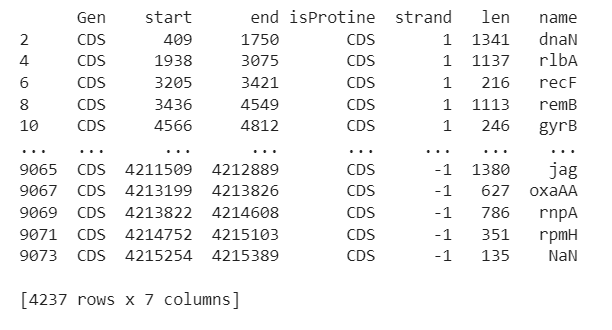
(סוג הגן, התחלת הגן, סוף הגן, האם הוא מקודד לחלבון או לא, סטרנד, אורך הגן, שם הגן) אחר כך המרנו את המילון לדאתה פראם וחסרנו את האינדקס של סוף הגן להתחלה וכך הוצאנו את האורך של כל גן.

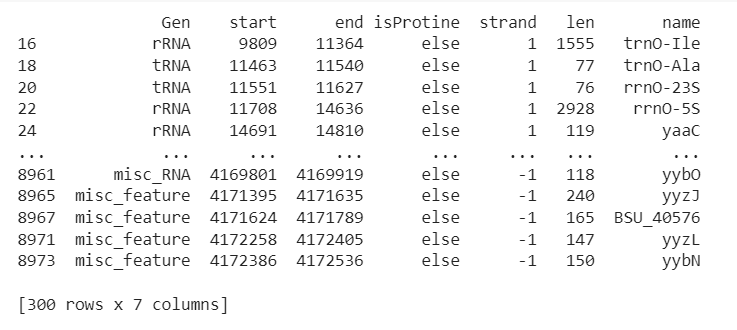
טבלת כל גן ברצף הדנ"א – כולל האורך של כל גן



2.2 תוך כדי הוספת הפרטים למילון בדקנו כל גן אם הוא חלבון או לא ועדכנו את הטבלה בהתאם.

הגנים המקודדים לחלבון

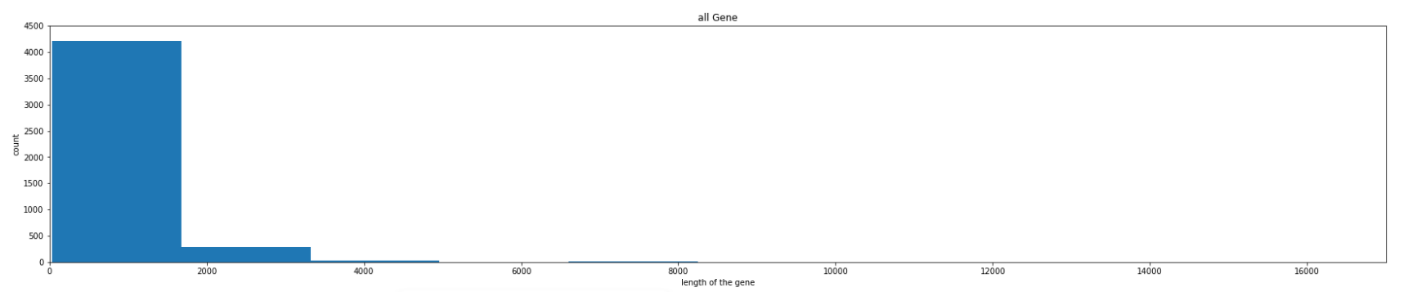


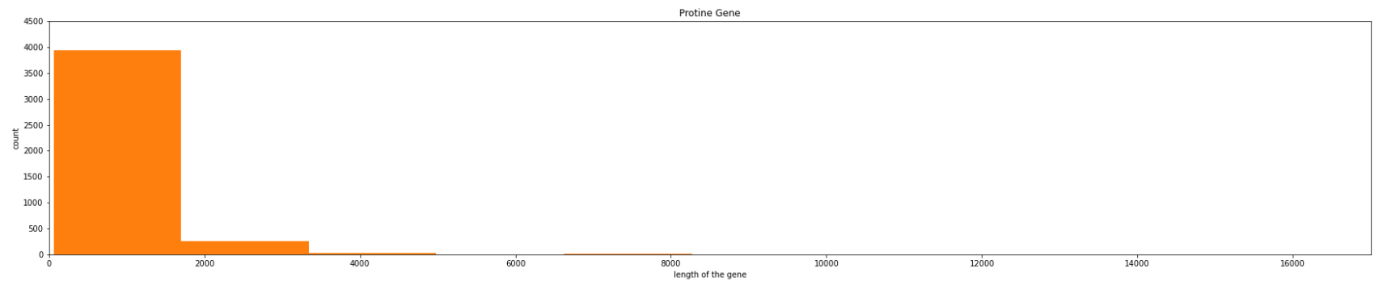
הגנים הלא מקודדים לחלבון

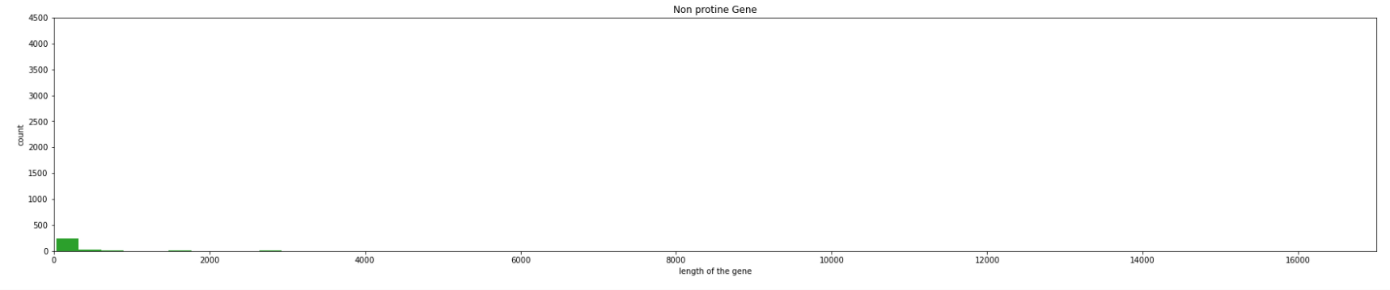
|  |  |
| --- | --- |
| שאר הקבוצות | קבוצת הגנים המקודדים לחלבון |
| Text  Description automatically generated | Text  Description automatically generated |

2.3

2.4 ציירנו את הגרפים, מספר הגנים יצא 4536, מספר הגנים המקודדים לחלבון יצא 4237 - קטן ממספר הגנים כולו, זה תואם לפי הציפיות שלנו כי למדנו שתמיד מספר הגנים המקודדים לחלבון קטן או שווה למספר הגנים כולו ונותרו 4536 – 4237 = 299 שזה מספר הגנים הלא מקודדים לחלבון.

בגרף נוכל לראות באופן ברור שאורכי הגנים המקודדים לחלבון מאוד קרוב לאורכי הגנים כולו וזה מכיוון שרוב הגנים יצאו מקודדים לחלבון וחלק קטן מאוד יצא שהם לא מקודדים לחלבון, לכן יש הבדל משמעותי בגרף (הירוק – של הגנים הלא מקודדים לחלבון) לשאר הגרפים.



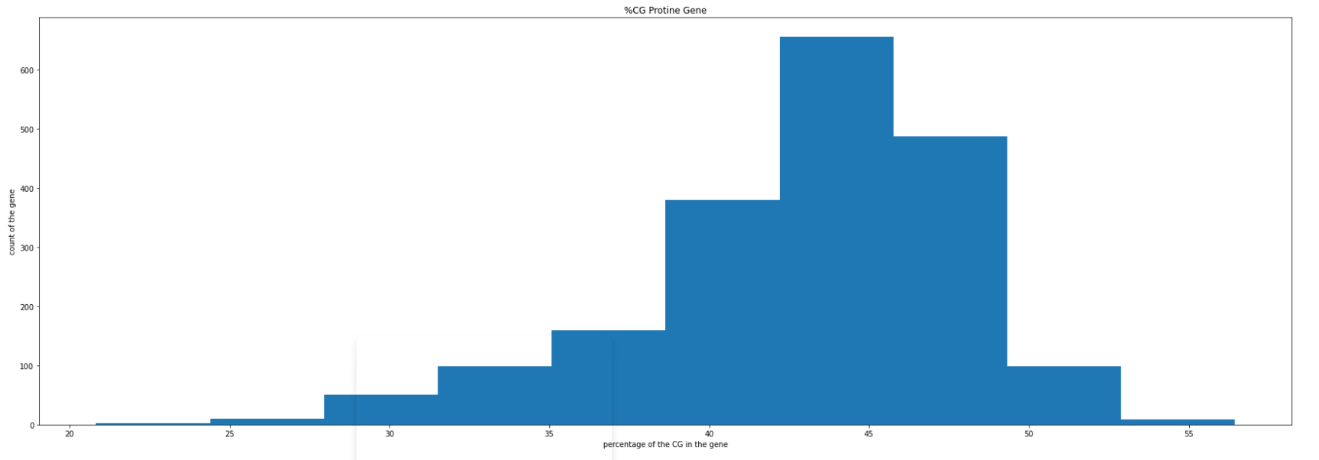


3.

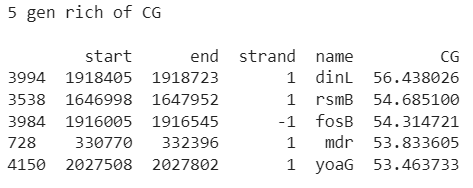
3.1 התוצאה: 43.51440813017155

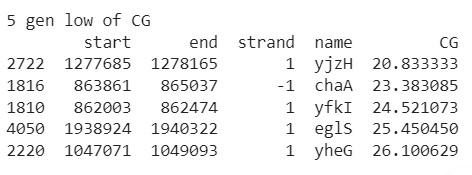
3.2 התוצאה: 42.99335683927281

3.3 התוצאה תואמת לציפיות שלנו, כי בחלק הקודם מצאנו שמספר ואורכי הגנים המקודדים לחלבון קרוב למספר ואורכי הגנים (הגנום כולו), לכן ציפינו שאחוז ה GC בגנים המקודדים לחלבון יהיה קרוב מאוד לאחוז ה GC בגנים(הגנום כולו) וזה מה שיצא לנו.

3.4

3.5 מיינו את המילון שיצרנו אותו בהתחלה לפי אחוז ה GC והוצאנו את הגנים העשירים ביותר ב GC והגנים עם הרכב נמוך ביותר.





4.

מצאנו שתי סתירות, בגן secA ובגן srdFi הסתירה שהרצף לא מתחלק ב 3 ובגן nrdEB יש לנו שגיאה בקודון העצירה.

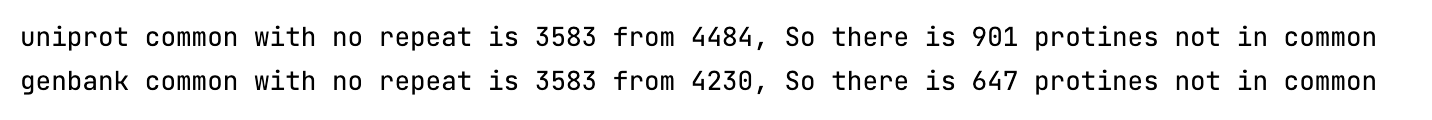
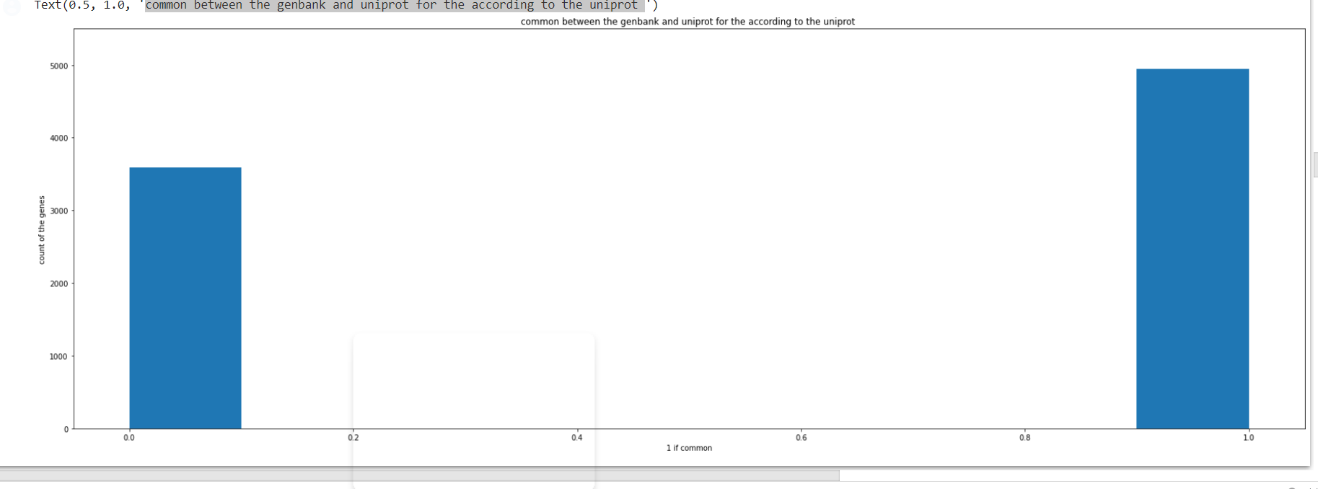
את התוצאה:

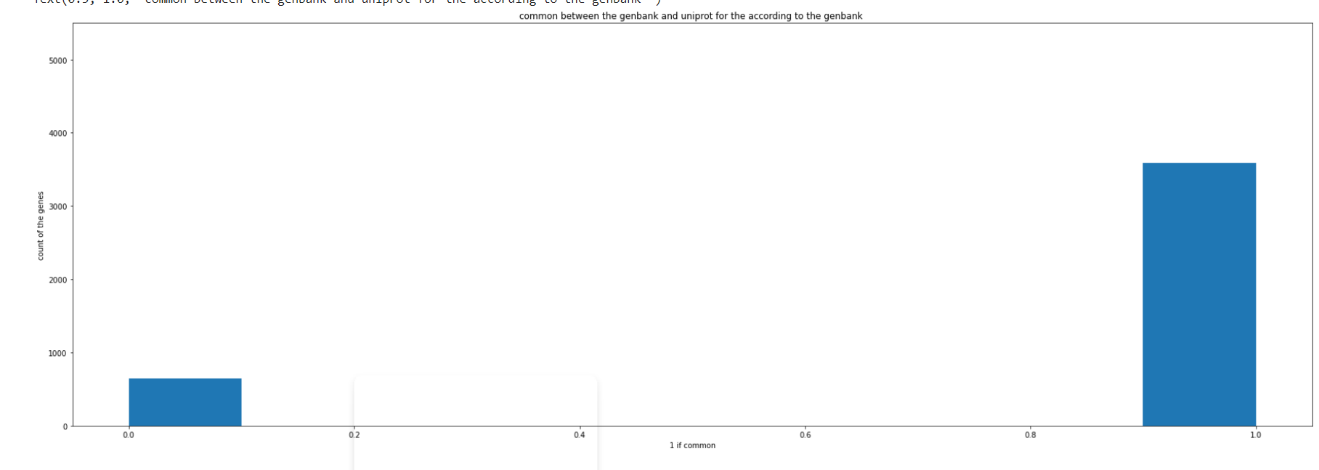
ואפשר גם לראות אותה בקובץ בשם gene\_exceptions.csv

**◄ הערה: התוצאות הועברו לקובץ בשם part\_a.csv**

◄ חלק ב' - אנליזת חלבונים בעזרת אתר ה UniProt-

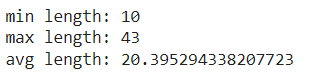
א.

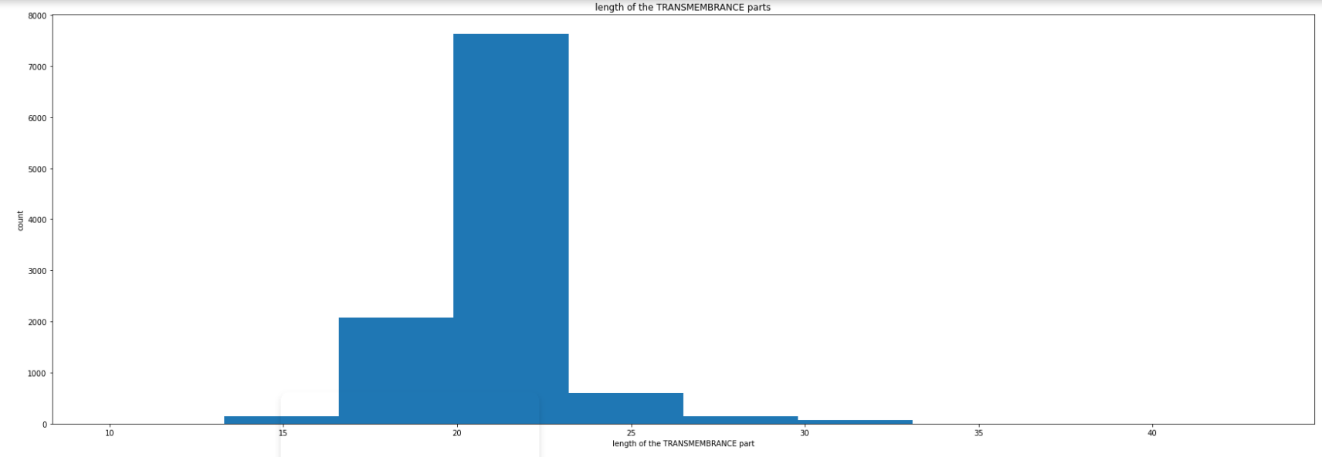
אחרי שקראנו את הקבצים מ Uniprot ומ Genbank, מצאנו את הגנים המשותפים לפי השם של הגן, מספר הגנים המשותפים הוא 3583, ויש בשני הקבצים גנים לא משותפים, שם העמודה שהשתמשנו בה להצלבה Gene name primary

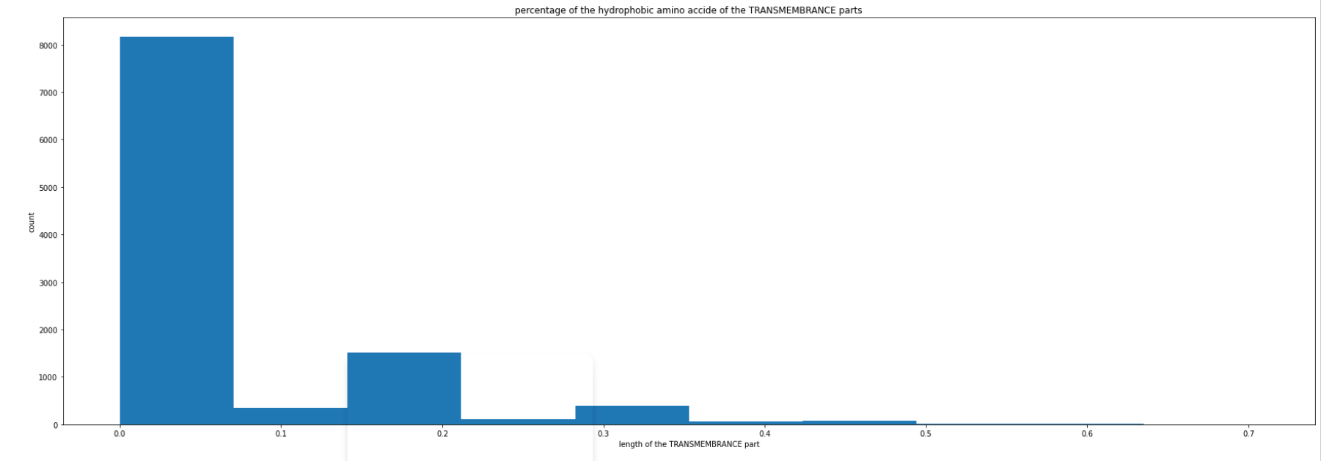


ב.

1. אורך הממוצע, המינימלי והמקסימלי:

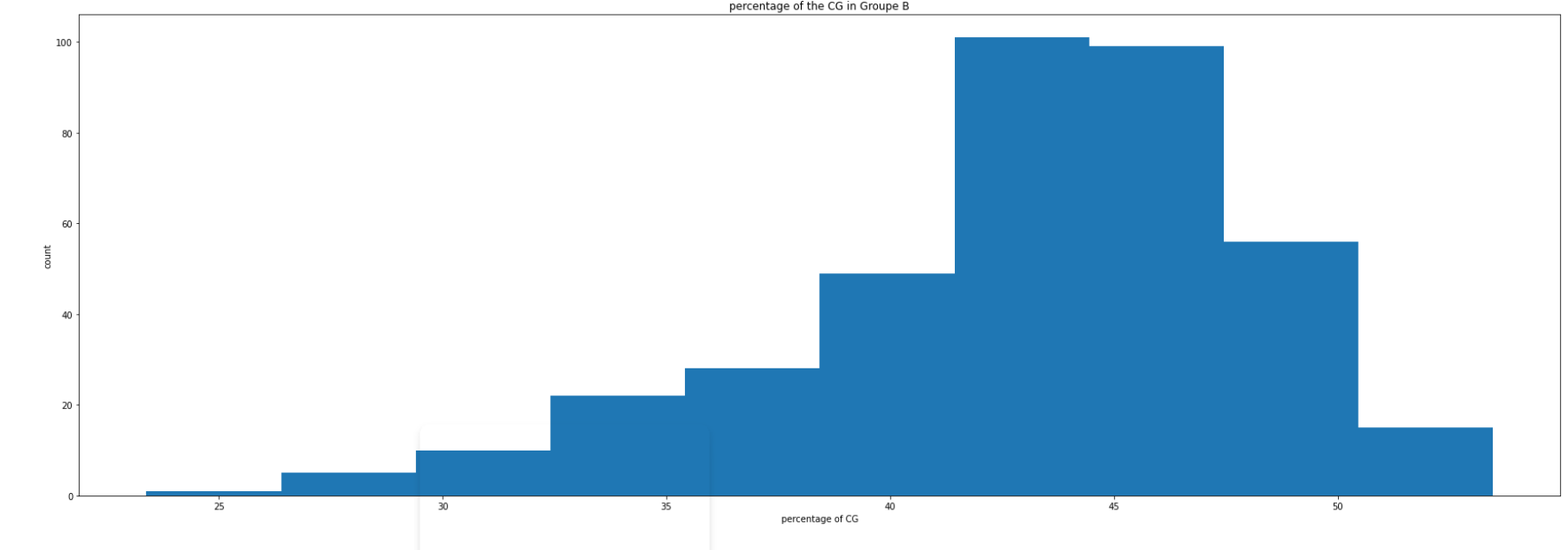


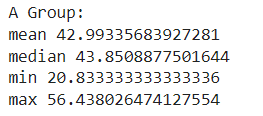
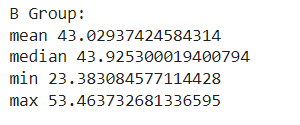


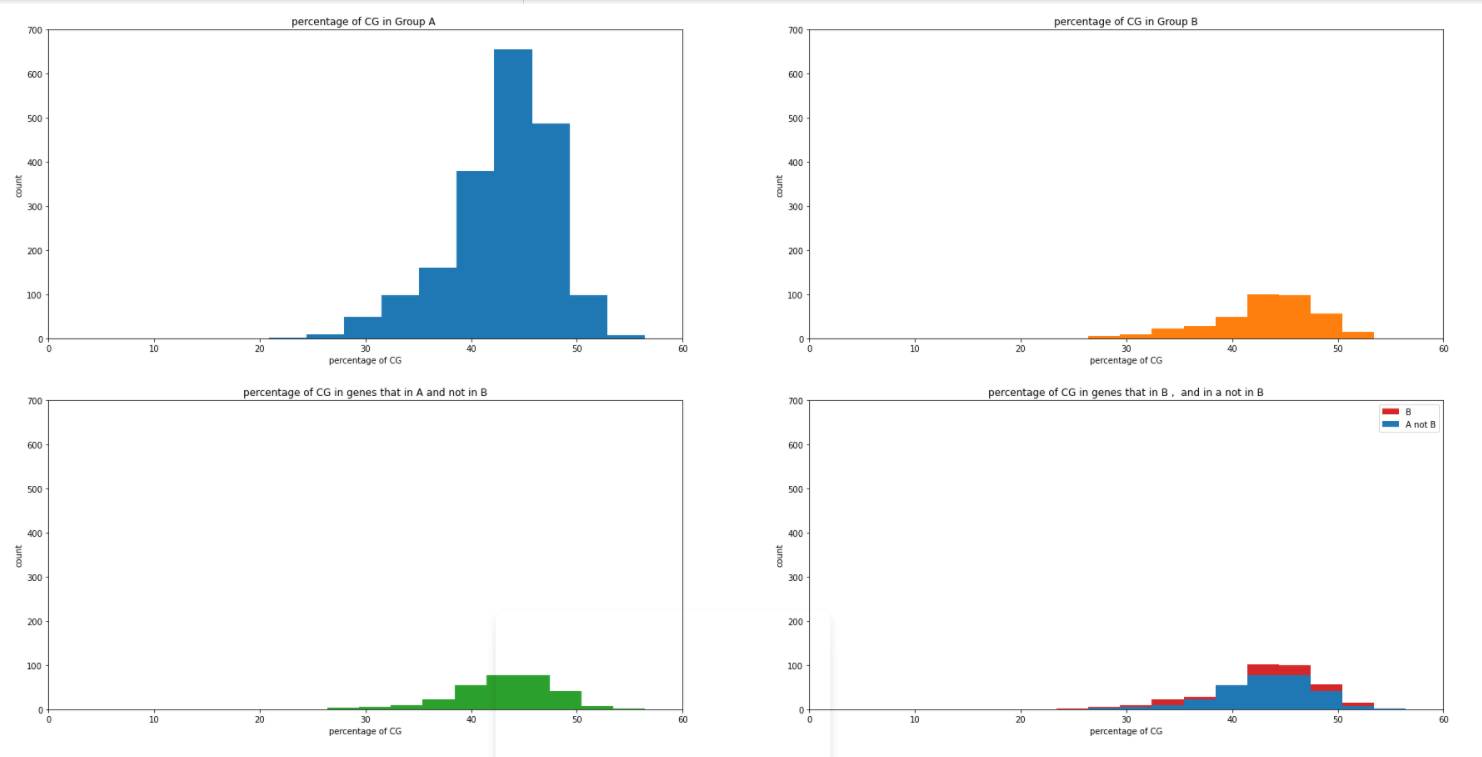
2. הערך הממוצע על פני כל הרצפים: 0.04611083306678117

ג.

1.

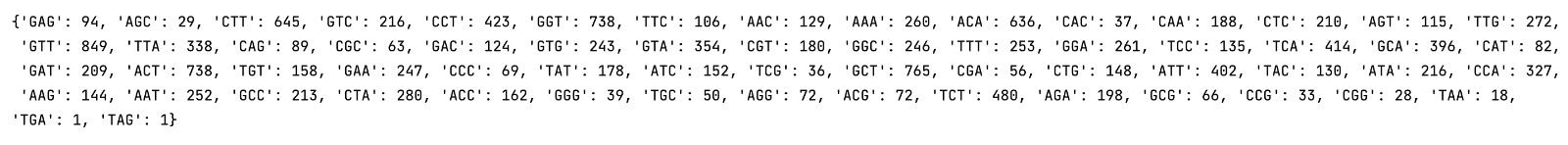


2.



◄חלק ג' – אנליזה מנקודת מבט אבולוציונית וירוסים

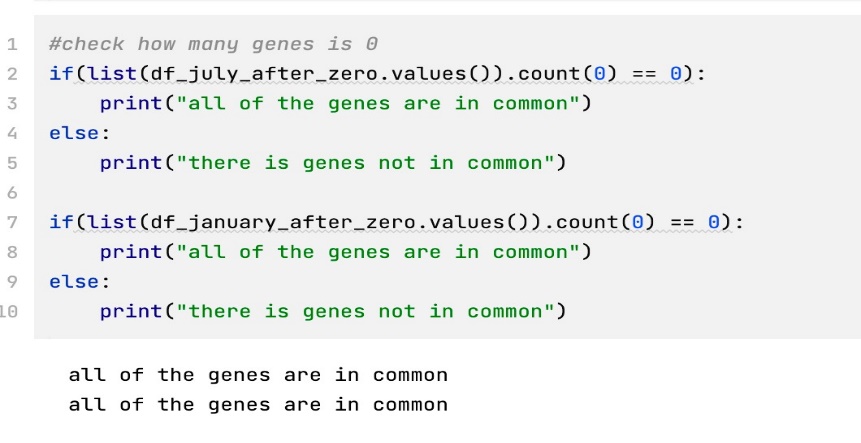
1.

יצרנו מילון, עברנו על כל קודון בגנום וספרנו כמה עמדות סינונימיות.

2.

א. לכל גינום יצרנו מילון, הוספנו את פרטי הגנים (סוג הגן, התחלת הגן, סוף הגן, סטרנד, שם הגן, אורך הגן)

אחר כך המרנו את המילון לדאתה פראם ומצאנו את הגנים המשותפים לפי השם, כל הגנים שנמצאים בשני הקבצים יצאו לנו משותפים.



ב. בחרנו את הגנים המשותפים האלה lst = [[3, 31] , [9, 37] , [7, 35] ,[5, 33] , [19, 47]]

כדי לחשב את הdnds עשינו תרגום לרצף כדי לקבל את החלבון, אחר כך עשינו עימוד ואחר כך החזרנו את הרצף לרצף ה DNA עם הגאפים ובסוף הכנסנו את הרצפים האלה וחישבנו את הdnds.

