תשובות לשאלות מילוליות:

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

1. 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: בנספחים

ג': סטטיסטיקה של הגנים המקודדים למול אלה שלא:

תמונה שמכילה טקסט, צג, צילום מסך, מקורה

התיאור נוצר באופן אוטומטי

את ההיסטוגרמות ניתן לראות בנספחים. 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.

1. 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%

הערה: 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

1. 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.

נספחים:

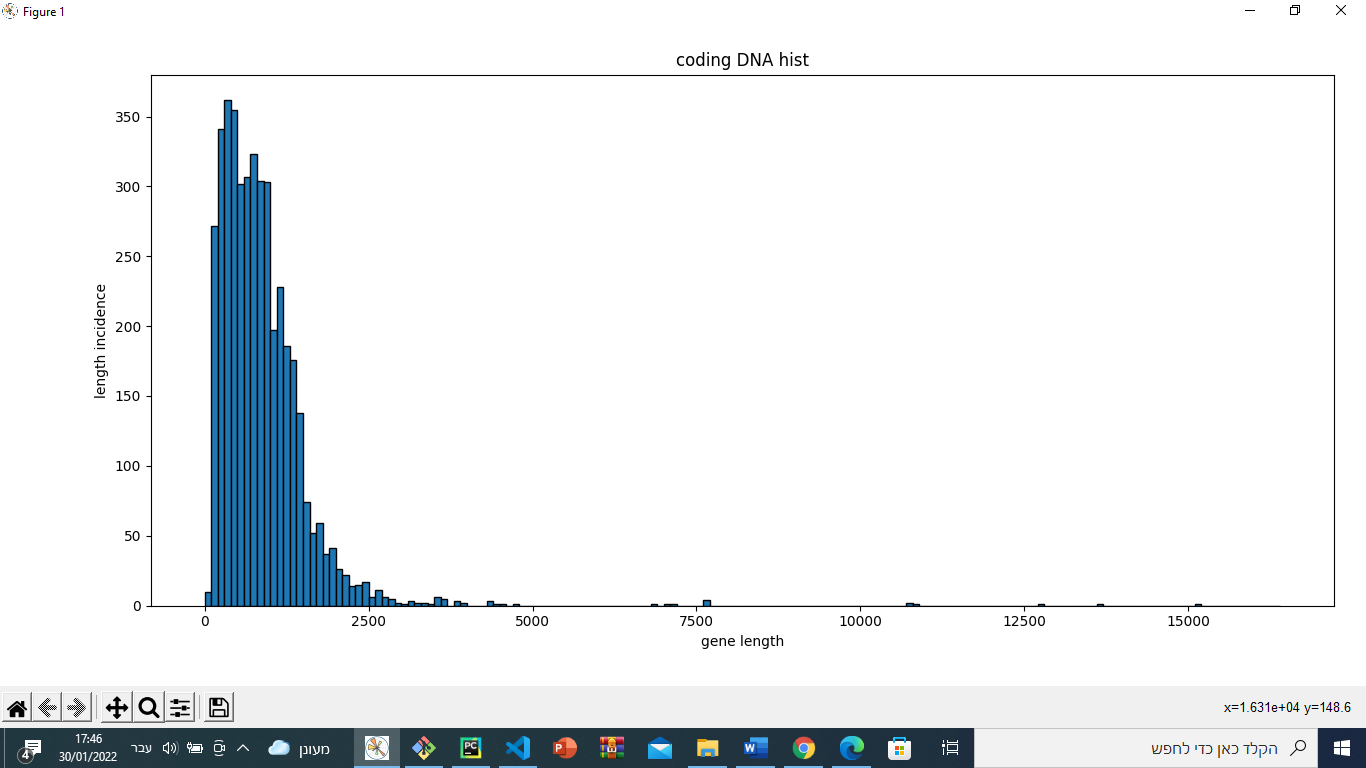
הדפסה של אורכי כל הגנים 5 ראשונים ואחרונים

תמונה שמכילה טקסט, צג, צילום מסך, מקורה

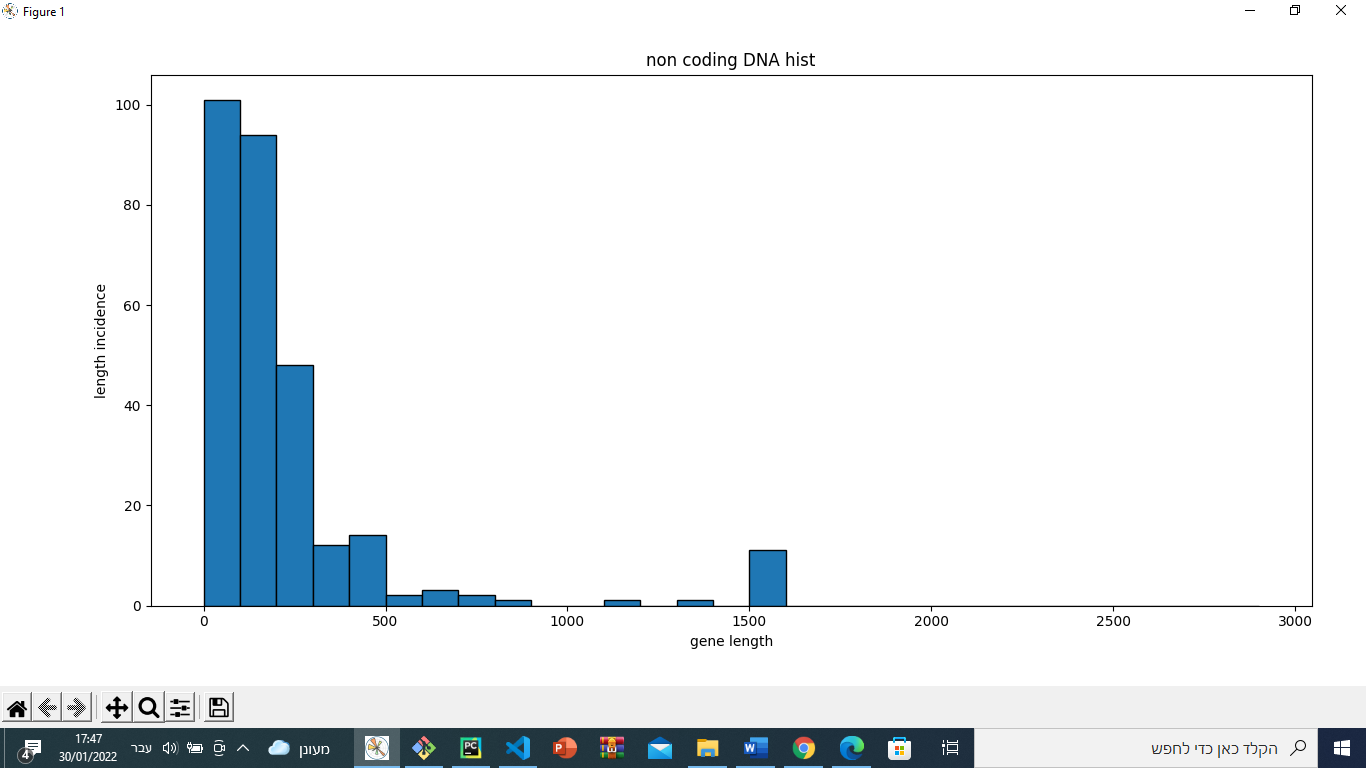
התיאור נוצר באופן אוטומטי

ההיסטוגרמות משאלה 2 בחלק הראשון:

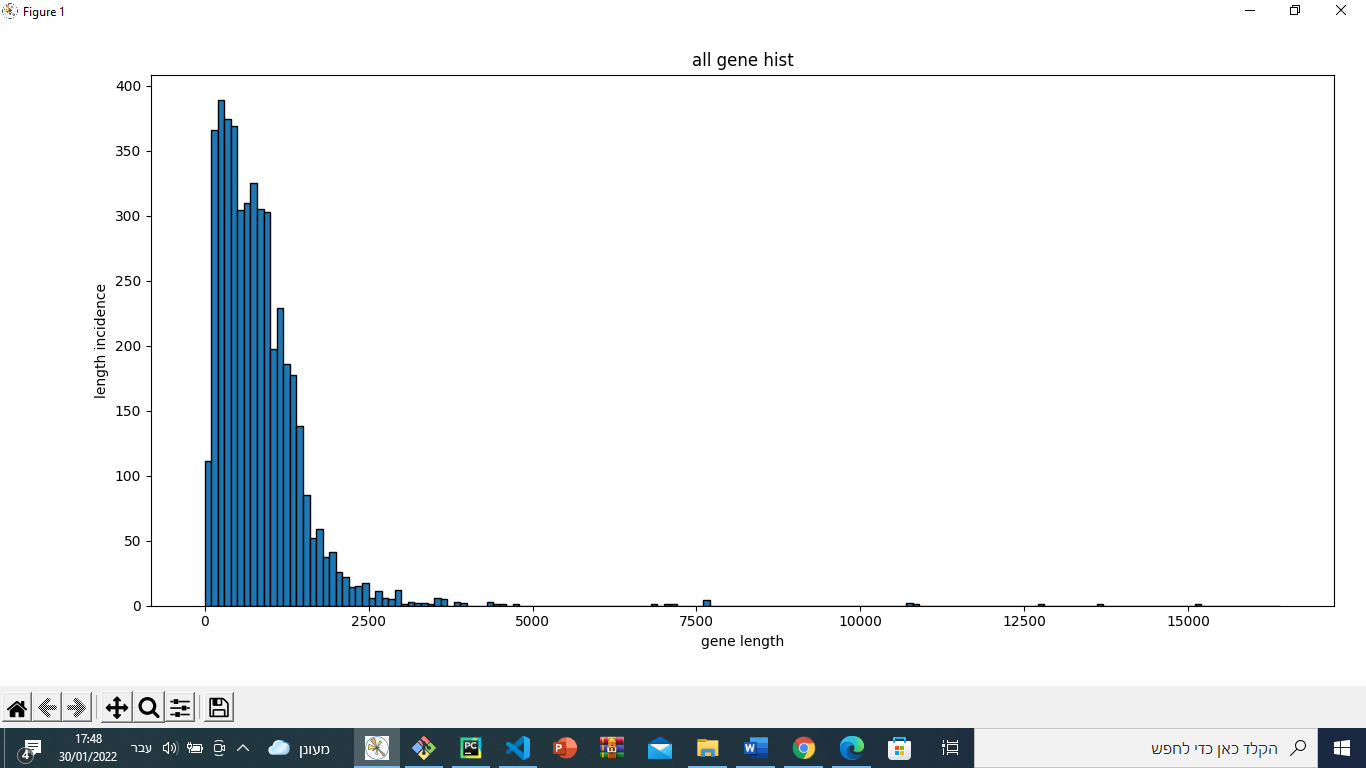
גרף של אורכי הגנים המקודדים:



גרף של אורכי הגנים הלא מקודדים:



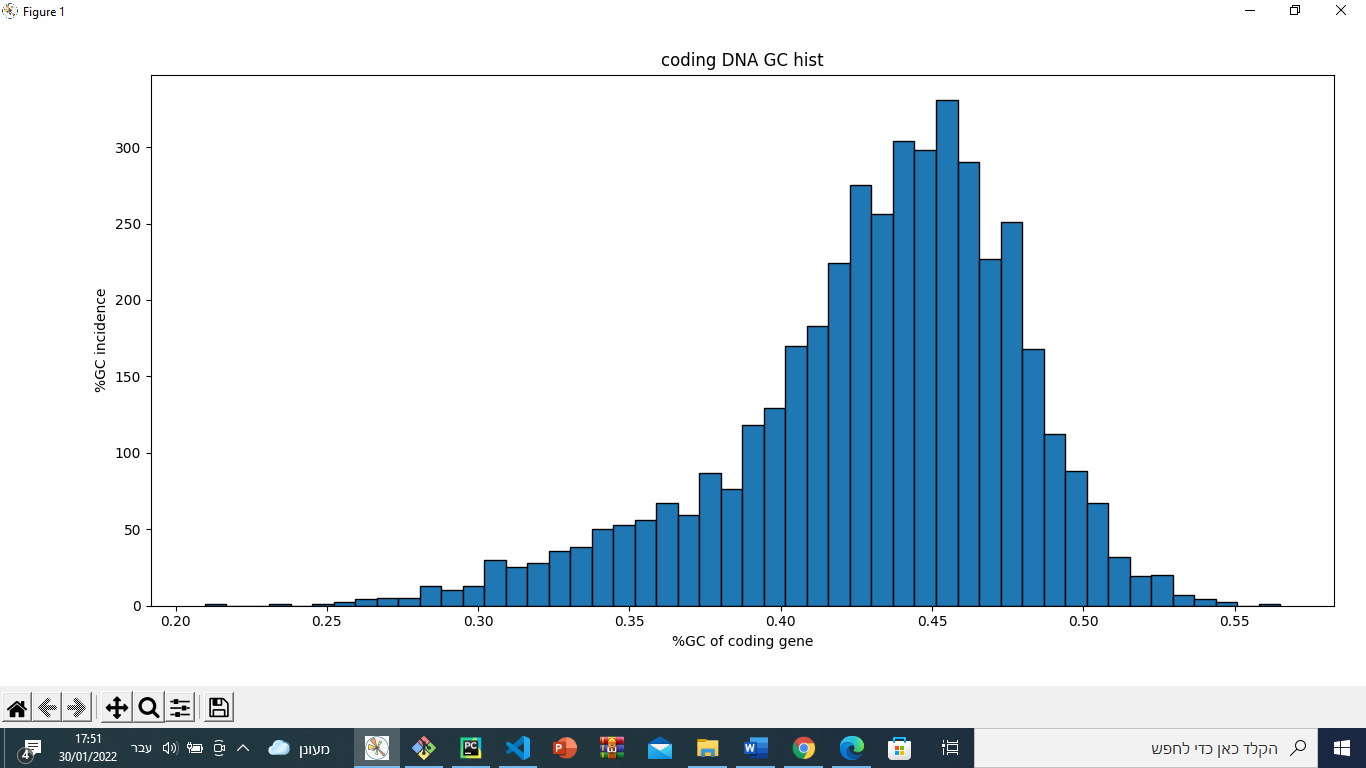
אורכי כל הגנים:



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

שאלה 3:

היסטוגרמת %GC



5 הגנים העשירים\דלילים בGC:

5 הראשונים הכי דלילים בGC

והאחרונים הכי עשירים.

תמונה שמכילה טקסט, צילום מסך, צג, אלקטרוניקה

התיאור נוצר באופן אוטומטי