נכתב על ידי עידו בן צבי (208627232) שגיא בן שושן (209351147) ועדיאל רוזנפלד (208349019)

**חלק א'**

1.

Number of gene: 4203

Number of CDS: 4108

Number of rRNA: 22

Number of tRNA: 73

Number of misc\_feature: 27

2.  
א. תמונה שמכילה צילום מסך, טקסט, עלילה, קו

התיאור נוצר באופן אוטומטי

ב.  
 Number of genes on plus strand: 2074

Number of genes on minus strand: 2129

Number of protein coding genes on plus strand: 1998

תמונה שמכילה טקסט, צילום מסך, מספר, עלילה

התיאור נוצר באופן אוטומטיNumber of protein coding genes on minus strand: 2110

ג. תמונה שמכילה צילום מסך, תרשים, טקסט, עלילה

התיאור נוצר באופן אוטומטי התפלגויות דומות הממוצע באזור ה 900 בסיסים לגן.

ד.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Std | Mean | Max | Min | Strand | CDS type |
| 565.016 | 900.657 | 8613 | 102 | Plus(1998) |
| 535.217 | 908.709 | 4896 | 81 | Minus(2110) |

ה.

תמונה שמכילה טקסט, צילום מסך, קו, תרשים

התיאור נוצר באופן אוטומטי

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Std | Mean | Max | Min | Strand | tRNA type |
| 4.909 | 77.5 | 94 | 72 | Plus(54) |
| 5.983 | 16.684 | 93 | 71 | Minus(19) |

תמונה שמכילה טקסט, צילום מסך, קו, מספר

התיאור נוצר באופן אוטומטי

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Std | Mean | Max | Min | Strand | rRNA type |
| 1110.503 | 1427.682 | 28004 | 112 | Plus(22) |
| 0 | 0 | 0 | 0 | Minus(0) |

ו.

* ההבדלים בין הסטרנדים:
  + ניתן לראות לגבי גנים שמקודדים לחלבון כי אין הבדל בין הסטרנדים, הסטטיסטיקות מצביעות על כך שאורכי הגנים שמקודדים לחלבון דיי זהים בין סטרנד חיובי לשלילי כמו גם הממוצע וסטיית התקן.
  + לגבי tRNA ניתן לראות כי גם כאן אין הבדל משמעותי בין הסטרנדים, הסטטיסטיקות מצביעות על כך שאורכי הגנים שמקודדים לtRNA דיי זהים בין הסטרנד החיובי לשלילי כמו גם הממוצע וסטיית התקן.
  + לגבי rRNA רואים בבירור שהגנים נמצאים אך ורק על הסטרנד החיובי קיימים 22 גנים כאלו והאורכים שלהם דיי מפוזרים על גבי הגרף מה שמצביע על שונות גדולה (סטיית התקן גדולה מאוד). סיבות מגוונות יכולות להיות לעובדה שרק על הסטרנד החיובי מופיעים הקידודים לrRNA, סיבה אבולוציונית שנתנה איזשהו ייתרון לחיידקים שלהם יש על הסטרנד החיובי את הגן, מבנה הכרומוזום בחיידק אולי משפיע וייתכן גם שהסביבה בה חיי החיידק השפיע.
* ההבדלי בין גנים שמקודדים לחלבון לבין כאלו שלא מקודדים לחלבון:

ישנם המון גנים שמקודדים לחלבון, למעשה הרוב המוחלט מהגנים של החיידק מקודדים לחלבון, חלק קטן מהגנים מקודדים לRNA ריבוזומאלי ולRNA שליח שהם משתתפים בתהליך ייצור החלבונים בתא. גוף החיידק צריך כמות גדולה של חלבונים וגם מסוגים שונים אשר מתפקדים במגוון של תפקידים ומקומות בגוף החיידק, וזה מסביר את הכמות הגדולה של החלבונים. לגבי tRAN rRNA אלו אברונים שמתפקדים בתהליך התרגום לחלבון ולכן לא נדרש מגוון רחב מאוד שלהם.

3. א. מספר נוקלאוטידים מסוג C 965102, מספר נוקלאוטידים מסוג T 1189975, אחוז בסיסים אלו מכול הגנום: 81%.

ב. ממוצע כמות ה CT בגנים המקודדים לחלבון הינו 50% מבסיסי הגן.

ג. בסעיף הקודם מצאנו כי הגנים המקודדים לחלבון נמצאים על שני הגדילים באופן זהה (כמעט חצי חצי) כלומר ישנה כמות דומה של גנים שמקודדים לחלבון על כול סטרנד, כעט מצאנו כי כמות ה C והT של כול הגנים שמקודדים לחלבון היא 50% כלומר חצי מהבסיסים של כול גן הם בסיסים מסוג C T, מכאן שהחצי השני של הבסיסים שלהם מסוגים תואמים G A. ידוע לנו שDNA הוא דו גדילי (ישנם 2 סטרנדים) וגם ידוע לנו שנוקלאוטיד מסוג T בגדיל מסויים תמיד עומד מנגד לנוקלאוטיד מסוג A כמו שG עומד מול C בסטרנד הנגדי.  
לכן אם חצי מהגנים יושבים על סטרנד אחד וחצי מהגנים יושבים על הסטרנד האחר היינו מצפים ליחס בסיסים שכזה.   
המחשה: תמונה שמכילה קו, צילום מסך, צבעוני, טקסט

התיאור נוצר באופן אוטומטי

ד. תמונה שמכילה צילום מסך, עלילה, תרשים

התיאור נוצר באופן אוטומטי

ה.

5 richest CDS genes with the highest C and T nucleotide content:

'name': 'Gene name not available, locus tag: ABC3680', 'start': 3836971, 'end': 3837355, 'strand': -1, 'C\_T\_content\_by\_percent': 0.67  
'name': 'Gene name not available, locus tag: ABC3803', 'start': 3963886, 'end': 3964210, 'strand': -1, 'C\_T\_content\_by\_percent': 0.67  
'name': 'Gene name not available, locus tag: ABC2100', 'start': 2221887, 'end': 2222022, 'strand': -1, 'C\_T\_content\_by\_percent': 0.7  
'name': 'Gene name not available, locus tag: ABC3192', 'start': 3313898, 'end': 3314048, 'strand': -1, 'C\_T\_content\_by\_percent': 0.72  
'name': 'Gene name not available, locus tag: ABC3162', 'start': 3285891, 'end': 3286071, 'strand': -1, 'C\_T\_content\_by\_percent': 0.77

5 poorest CDS genes with the lowest C and T nucleotide content

'name': 'Gene name not available, locus tag: ABC1762', 'start': 1882854, 'end': 1883031, 'strand': 1, 'C\_T\_content\_by\_percent': 0.29  
'name': 'Gene name not available, locus tag: ABC1314', 'start': 1431863, 'end': 1432016, 'strand': 1, 'C\_T\_content\_by\_percent': 0.32  
'name': 'Gene name not available, locus tag: ABC0946', 'start': 1027483, 'end': 1027762, 'strand': 1, 'C\_T\_content\_by\_percent': 0.33  
'name': 'gsiB', 'start': 3644227, 'end': 3644629, 'strand': 1, 'C\_T\_content\_by\_percent': 0.33  
'name': 'Gene name not available, locus tag: ABC0090', 'start': 108960, 'end': 109257, 'strand': 1, 'C\_T\_content\_by\_percent': 0.34

**חלק ב'**

1. נתחיל ונאמר שכדי להשוות היינו צריכים להתאים את הצורה שבה המידע מונגש לנו מכול קובץ שכן לא כל המידע מונגש באותה צורה ויש להתחשב בכל קובץ בנפרד.  
   הכנו DataFrame שהכתרות שלו הם "Identifier", "Gene Name", "Description", "Protein Name", "Length", "Sequence". כך קראנו את קובץ ה FASTA תמונה שמכילה טקסט, צילום מסך, תוכנה

   התיאור נוצר באופן אוטומטידגש חשוב הוא שלא תמיד מופיע לכול גן את שם הגן (היו 3 חלבונים כאלו לפחות מתוך קובץ של 500).  
   כך קראנו את קובץ הGenBank:תמונה שמכילה טקסט, צילום מסך, תוכנה

   התיאור נוצר באופן אוטומטיהדגש הוא שגם כאן היה חסר השם של הגן ואם מצאנו שם שחסר שמנו את ה Loucus\_tag שלו.  
   לאחר השוואה על פי שם הגן בלבד מצאנו 338 שורות עם שמות זהים. ההפרש ייתכן מאוד שכן בשני הקבצים היו חוסרים של שמות הגנים. כאן X מייצג קובץ Fasta Y מייצג קובץ GenBank.  
   תמונה שמכילה טקסט, תפריט, צילום מסך

   התיאור נוצר באופן אוטומטיהחלטנו שעקב המידע החסר נשווה על פי עמודת "חלבון מתורגם" ומצאנו 490 שורות זהות שאותן ניתן לראות בקובץ שיצאנו שנקרא "part\_b.csv" כמוכן בקובץ ניתן לראות את 10 החלבונים מהFasta שאותם לא מצאנו בGenBank.
2. תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

   התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

התיאור נוצר באופן אוטומטיהממוצע של חומצות האמינו שנמצאות בחלבונים הטרנס-ממברנליים יצא לנו 0.55 אחוז.  
כלומר שקצת יותר מחצי מהחלבון דוחה מים. באופן כללי, חלבונים טראנסממבניליים הם חלבונים שחוצים לרוחב את קרום הממברנה של התא. קרום התא עשוי 2 שכבות של פוספוליפידים שלהם ראש "נמשך" למים או "הידרופילי" וזנבות שדוחות מים או צד "הידרופובי", הפוספוליפידים יושבים בשכבות כך שהראשים מופנים כלפי פנים וחוץ התא והזנבות יוצרות חיץ לנוזלים – כך נוצר חיץ בין הסביבה שבפנים התא לבין הסביבה שמחוץ לתא. תמונה שמכילה טקסט, צילום מסך, אומנות ילדים, אומנות

התיאור נוצר באופן אוטומטי

<http://www.annefrankschool.org.il/wp-content/uploads/2014/02/%D7%A7%D7%A8%D7%95%D7%9D-%D7%94%D7%AA%D7%90-%D7%95%D7%93%D7%A8%D7%9B%D7%99-%D7%9E%D7%A2%D7%91%D7%A8-%D7%97%D7%95%D7%9E%D7%A8%D7%99%D7%9D.pdf>  
מה שמצאנו הוא שיש כ55% חומצות אמינו שהן "דוחות" מים ברוב החלבונים טראנסממבניליים מה שהגיוני בסך הכול מכיוון שחלבונים אלו חוצים את ממברנת קרום התא ונמצאים גם בסביבה מיימית וגם בסביבה שדוחה נוזלים. <https://he.wikipedia.org/wiki/%D7%97%D7%9C%D7%91%D7%95%D7%9F_%D7%98%D7%A8%D7%90%D7%A0%D7%A1%D7%9E%D7%9E%D7%91%D7%A8%D7%A0%D7%9C%D7%99>  
  
**חלק ג'**

חשוב להדגיש שהטבלה עליה עשינו את החישוב של עמדות סינונימיות לכול קודון על הטבלה הסטנדרטית כפי שהובהר לנו בשיעור עם המרצה!

'TTT': 1, 'TTC': 1, 'TTA': 2, 'TTG': 2, 'TCT': 3, 'TCC': 3, 'TCA': 3, 'TCG': 3, 'TAT': 1, 'TAC': 1, 'TGT': 1, 'TGC': 1, 'CTT': 3, 'CTC': 3, 'CTA': 4, 'CTG': 4, 'CCT': 3, 'CCC': 3, 'CCA': 3, 'CCG': 3, 'CAT': 1, 'CAC': 1, 'CAA': 1, 'CAG': 1, 'CGT': 3, 'CGC': 3, 'CGA': 4, 'CGG': 4, 'ATT': 2, 'ATC': 2, 'ATA': 2, 'ACT': 3, 'ACC': 3, 'ACA': 3, 'ACG': 3, 'AAT': 1, 'AAC': 1, 'AAA': 1, 'AAG': 1, 'AGT': 1, 'AGC': 1, 'AGA': 2, 'AGG': 2, 'GTT': 3, 'GTC': 3, 'GTA': 3, 'GTG': 3, 'GCT': 3, 'GCC': 3, 'GCA': 3, 'GCG': 3, 'GAT': 1, 'GAC': 1, 'GAA': 1, 'GAG': 1, 'GGT': 3, 'GGC': 3, 'GGA': 3, 'GGG': 3, 'TAA': 2, 'TAG': 1, 'TGA': 1

א.

(SARS-CoV-2) 29-APR-2021 has 29740 nucleotides.

(SARS-CoV-2) 20-FEB-2024 has 29661 nucleotides.

ב.   
בכול אחד מקבצי הCovid נמצאים 11 גנים באופן כללי וכולם מקודדים לחלבון, ישנו (גם כן) בכול אחד חלבון אחד שמופיע פעמיים במקומות שונים ברצף אך מדובר באותו חלבון (בערך - חסרים כמה חומצות אמינו בסוף התרגום).

ג.  
ישנם 2 גנים שזהים אחד לאחד (ORF7a, ORF10) מבחינת התרגום.  
כל השאר זהים אחד לשני

ד.

Gens N dN: 0.0011, dS: 0.014, dN/dS: 0.07360495318165225 Negative selection

Gens ORF3a dN: 0.0, dS: 0.0055, dN/dS: 0.0 Negative selection

Gens ORF7a dN: 0.0, dS: 0.029, dN/dS: 0.0 Negative selection

Gens M dN: 0.002, dS: 0.013, dN/dS: 0.15160257596741933 Neutral selection

Gens E dN: 0.0, dS: 0.083, dN/dS: 0.0 Negative selection