דוח ביולוגיה חישובית

חלק 1

שאלה 1 :

```
Number of each type of region:
{'gene': 4536, 'CDS': 4237, 'misc_RNA': 93, 'misc_feature': 89, 'tRNA': 86, 'rRNA': 30, 'ncRNA': 2}
```

שאלה 2 :

"אנחנו יצרנו מילון עבור כל הגנים, והוספנו לו את כל הפרטים הנדרשים כגון: סוג הגן, תאריכי התחלת וסיום הגן, סטרנד, אורך הגן ושם הגן, ובסוף המרנו את מילון ל- data frame

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

```
      type    start      end  strand  length   name
1     gene      409     1750       1    1341   dnaA
2      CDS      409     1750       1    1341   dnaA
3     gene     1938     3075       1    1137   dnaN
4      CDS     1938     3075       1    1137   dnaN
5     gene     3205     3421       1     216   rlbA
...    ...      ...      ...     ...     ...    ...
9069   CDS  4213822  4214608      -1     786  oxaAA
9070  gene  4214752  4215103      -1     351   rnpA
9071   CDS  4214752  4215103      -1     351   rnpA
9072  gene  4215254  4215389      -1     135   rpmH
9073   CDS  4215254  4215389      -1     135   rpmH

[9073 rows x 6 columns]
```

א. אני חשבתי את האורך של כל גנים וכעת יש לי את האורך של כל גן מוכן

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

```
      type    start      end  strand  length   name
1     gene      409     1750       1    1341   dnaA
3     gene     1938     3075       1    1137   dnaN
5     gene     3205     3421       1     216   rlbA
7     gene     3436     4549       1    1113   recF
9     gene     4566     4812       1     246   remB
...    ...      ...      ...     ...     ...    ...
9064  gene  4211509  4212889      -1    1380   mnmE
9066  gene  4213199  4213826      -1     627    jag
9068  gene  4213822  4214608      -1     786  oxaAA
9070  gene  4214752  4215103      -1     351   rnpA
9072  gene  4215254  4215389      -1     135   rpmH

[4536 rows x 6 columns]
```

ב. בנינו פונקציה שמחלקת את כל הגנים שמקודדים לחלבון וגנים שלא מוקדדים לחלבון בשם divide_genes :

הגנים המקודדים לחלבון

```
      type    start      end  strand  length   name
2      CDS      409     1750       1    1341   dnaA
4      CDS     1938     3075       1    1137   dnaN
6      CDS     3205     3421       1     216   rlbA
8      CDS     3436     4549       1    1113   recF
10     CDS     4566     4812       1     246   remB
...    ...      ...      ...     ...     ...    ...
9065   CDS  4211509  4212889      -1    1380   mnmE
9067   CDS  4213199  4213826      -1     627    jag
9069   CDS  4213822  4214608      -1     786  oxaAA
9071   CDS  4214752  4215103      -1     351   rnpA
9073   CDS  4215254  4215389      -1     135   rpmH

[4237 rows x 6 columns]
```

הגנים הלא מקודדים לחלבון

```
              type    start      end  strand  length      name
16            rRNA     9809    11364       1    1555  rrnO-16S
18            tRNA    11463    11540       1      77  trnO-Ile
20            tRNA    11551    11627       1      76  trnO-Ala
22            rRNA    11708    14636       1    2928  rrnO-23S
24            rRNA    14691    14810       1     119   rrnO-5S
...            ...      ...      ...     ...     ...       ...
8961      misc_RNA  4169801  4169919      -1     118      mswM
8965  misc_feature  4171395  4171635      -1     240      yyzI
8967  misc_feature  4171624  4171789      -1     165      yyzJ
8971  misc_feature  4172258  4172405      -1     147      yyzK
8973  misc_feature  4172386  4172536      -1     150      yyzL

[300 rows x 6 columns]
```

ג.

	Minimum length	Minimum length	Average length	Standard deviation
All gene	33	16467	838.1029541446209	796.2591499885929
Protein-coding genes	63	16467	874.5702147746047	797.2605368609218
Non-protein-coding genes	33	2928	324.12	572.5823382246635

ד. ציירנו את הגרפים. מספר הגנים יצא 4536, מספר הגנים שמקודדים לחלבון יצא 4237, ומספר הגנים שלא מקודדים לחלבון היה 299 (4536 - 4237)

השתמשתי ב-bins ביצרת הסטוגרמות :

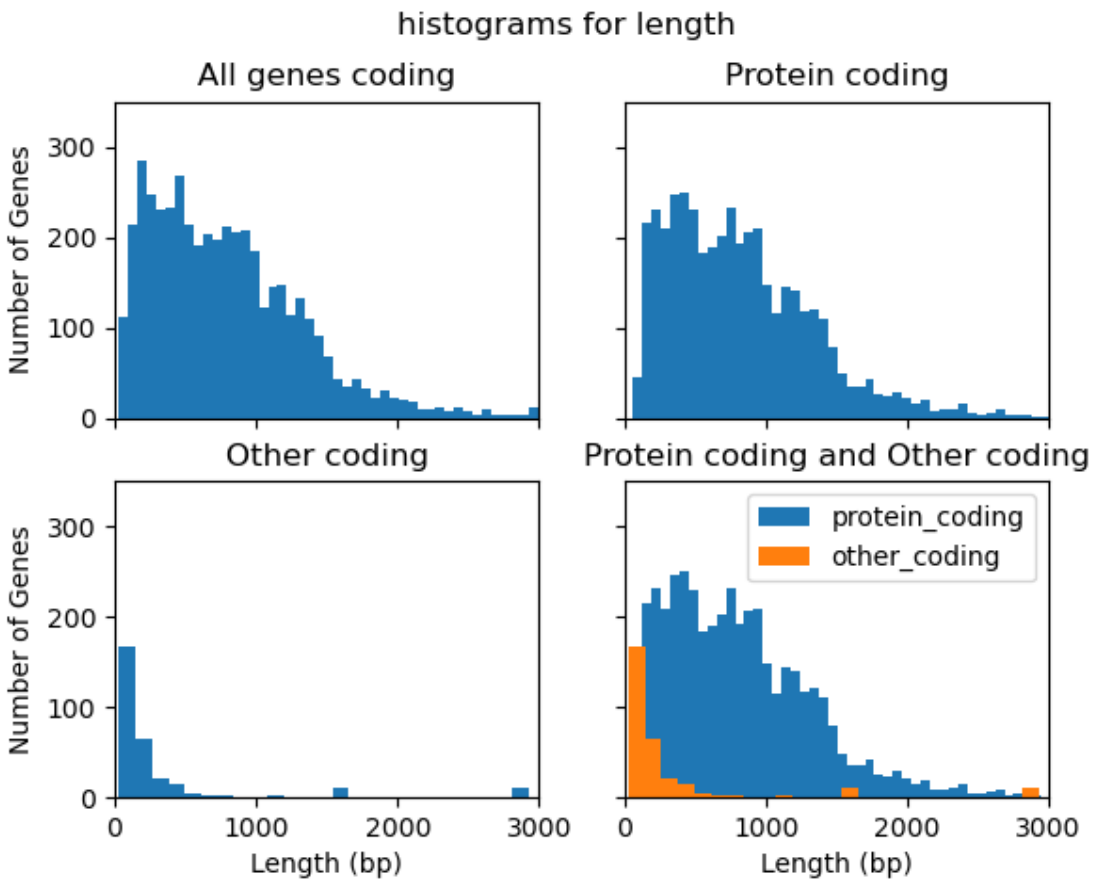
עבור כל הגנים השתמשנו ב- bins שווה ל 250

ועבור קבוצת הגנים שמקודדים לחלבון השתמשנו ב- bins שווה ל 250

ועבור קבוצת הגנים שלא מקודדים לחלבון השתמשנו ב- bins שווה ל 25

ואז כתוצאה מהבנת הגרפים ראינו עבור קבוצת הגנים שלא מקודדים לחלבון מרוכזים בין ארוכים 33 עד 250 ואחר כן יורדים ירידה משמעותית, ועבור האורכים 63 עד 250 ועבור ה-bins שבחרנו יצא שמספר קבוצות הגנים שמקודדים לחלבון הגיעו למספרים כמטע שווים מה שגורם למסקנה :

עבור האורכים בין 63 עד 250 מספר קבוצות הגנים שמקודדים לחלבון גדול פי 10 ממספר הגנים שלא מקודדים לחלבון



שאלה 3 סעיף א: Genome AT percentage: 56.48559186982845

סעיף ב : Protein-coding gene AT percentage: 56.880833819814235

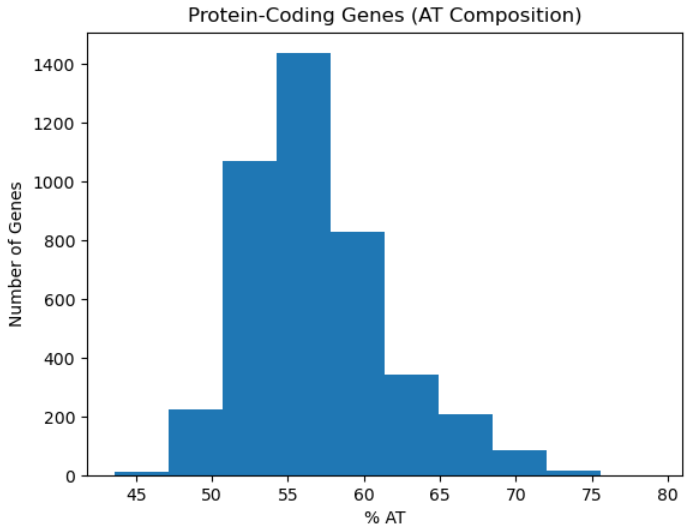
סעיף ג : The average AT composition for the protein-coding genes is higher than the average for the entire genome

ציפיות מבחינת מתמטית כן תואמות מכיוון ש-

Non-protein-coding gene AT percentage: 51.376298424305666

מכיוון שאחוז ה-AT בגנים שאינם מקודדים לחלבון נמוך מאחוז ה-AT בגנים שמקודדים לחלבון אז זה גורם שאחוז ה-AT הממוצע בגנום החיידק יהיה נמוך.

סעיף ד:



סעיף ה:

```
The 5 top AT-rich genes are:
        start      end  strand  length  name  AT_composition
5856  2699509  2699677      -1     168  yqaD       79.166667
3953  1904994  1905195      -1     201  cotC       76.616915
3939  1901116  1901377      -1     261  cotU       75.478927
8678  4036343  4036787      -1     444  rtbE       74.549550
8870  4132337  4132736      -1     399  yydC       74.185464

The 5 top AT-poor genes are:
        start      end  strand  length        name  AT_composition
6944  3194454  3194527      -1      73  trnSL-Ala1       32.876712
19      11551    11627       1      76    trnO-Ala       34.210526
397    166252   166328       1      76    trnI-Ala       34.210526
67      32019    32095       1      76    trnA-Ala       34.210526
217     96145    96221       1      76    trnJ-Ala       34.210526
```

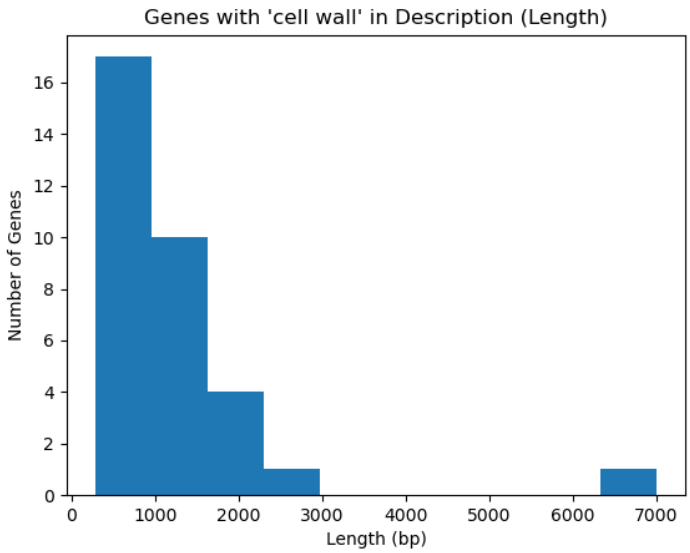
שאלה 4 סעיף א: מצאנו 33 גנים מכילים בתיאור את צמד המילים "cell wall" (ביחד)

	name	AT_composition	cell wall
106	yabE	58.447489	putative cell wall shaping enzyme
632	cwlJ	47.785548	spore cortex cell wall hydrolase
638	phoD	51.312785	secreted phosphodiesterase (endo-hydrolysis at...
1476	walM	60.215054	protein involved in cell wall metabolism
2162	lytE	58.507463	cell wall dl-endopeptidase; phosphatase-associ...
2432	wprA	55.083799	cell wall-associated protease
2930	ykfD	49.696970	putative cell wall oligopeptide ABC transporte...
3124	ykvT	60.127592	cell wall hydrolase related to spore cortex-ly...
3158	pbpH	59.763593	penicillin-binding enzyme for formation of rod...
3178	ykuG	64.413634	putative cell wall-binding protein
3346	ftsW	57.178218	cell-division protein; transporter of lipid-li...
3866	ymaG	52.536232	inner spore coat protein; cell wall associated...
4188	yoaR	56.578947	putative factor for cell wall maintenance or s...
4306	walL	58.333333	exported cell wall lytic enzyme
4350	cwlS	57.590361	peptidoglycan hydrolase (cell wall-binding d,l...
5135	ypbE	58.367911	putative enzyme possibly involved in cell wall...
5251	spoIIM	59.379845	autolysin component for dissolution of the sep...
5569	yqgA	62.937063	cell wall protein
6741	ytrF	59.572845	metabolite permease involved in resistance to ...
6743	ytrE	53.879310	ABC transporter (ATP-binding protein) involved...
6745	ytrD	59.406953	ABC transporter, permease component involved i...
6747	ytrC	59.979737	ABC transporter, permease component involved i...
6749	ytrB	55.858931	ABC transporter (ATP-binding protein) involved...
6751	ytrA	57.251908	transcriptional regulator (GntR family, cell w...
7363	liaR	51.257862	two-component response regulator [YvqE] respon...
7365	liaS	52.539243	two-component sensor histidine kinase [LiaR(Yv...
7369	liaG	53.150057	sensor of antibiotic stress on the cell wall
7391	yvrG	57.659208	two-component sensor histidine kinase YvrG [ce...
7393	yvrH	54.201681	two-component sensor histidine kinase YvrG [ce...
7723	cwlO	59.212377	secreted cell wall DL-endopeptidase
7971	ywsB	58.659218	putative cell wall binding enzyme
8443	ywcD	59.114583	putative cell wall glycosylation protein
8667	wapA	57.230550	cell wall-associated tRNA nuclease precursor; ...

סעיף ב :

Minimum length: 276, Maximum length: 7005, Average length: 1180.8181818181818

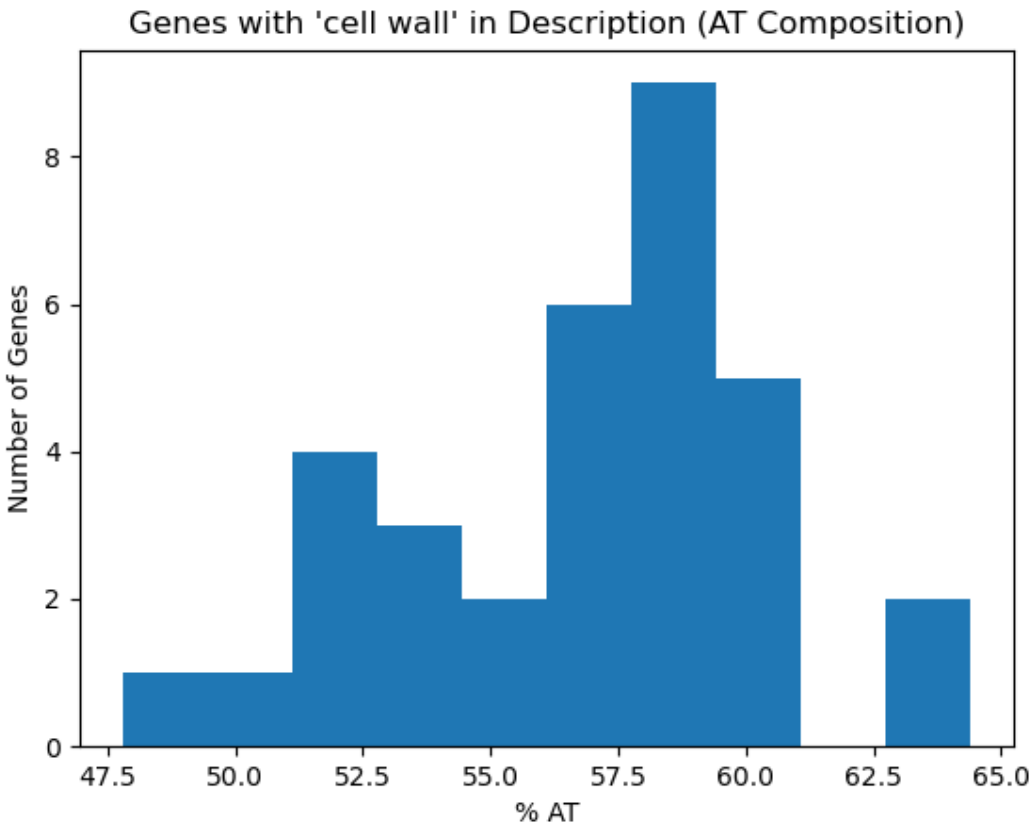
Standard deviation: 1178.4523869928266



סעיף ג :

Minimum: 47.785547785547784, Maximum: 64.41363373772387

Average: 56.8857666462032, Standard deviation: 3.7353109862626073



שאלה 5 : מצאנו שתי סתירות, בגן nrdFB ובגן prfB הסתירה שהרצף לא מתחלק ב 3 ובגן nrdEB יש לנו שגיאה בקודון העצירה.

```
{'nrdFB': 'Cant be divided into 3', 'nrdEB': TranslationError("Extra in frame stop codon 'TGA' found."), 'prfB': 'Cant be divided into 3'}
```

חלק 2

סעיף 1:

קיים 737 חלבונים בקובץ Genbank שלא נמצאים בקובץ UniProt, וקיים 761 חלבונים בקובץ UniProt שלא נמצאים בקובץ Genbank.

```
comparison:
common proteins number between Genbank and Uniport is: 3719

737 proteins  are in Genbank file but not in Uniport file

761 proteins  are in Uniport file but not in Genbank file
```

השוואה שנעשתה על ידי שימוש בשורה **"Gene Names (primary)"** , בחרנו העמודה הזו מכיוון שהיא אחת העמודות הטובות ביותר ויש לה כמה יתרונות וחסרונות:

היתרון: שמות החלבונים ב-uniport ו- Genbank הם זהים (gene field).

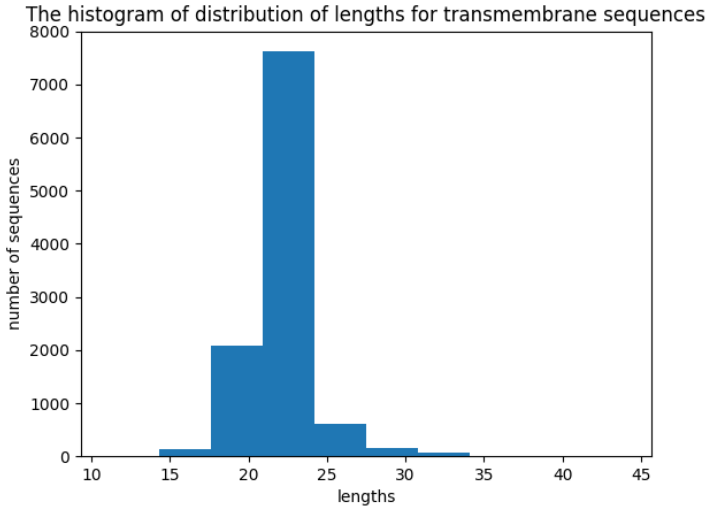
חסרונן: יש חלק מהשורות ריקות ב-uniport , וב-GenBank לחלק מהגנים אין שם.

למה לא בחרנו עמודה אחרת? למשל, אם אנחנו בחרנו להשתמש בתג לוקוס ב-GenBank, וב- UniProt משתמשים בעמודה - Genes Name (ordered locus), אבל יש הרבה שורות בעמודה הזו עם תג לוקוס ישן (old locus tag) ולא תג לוקוס רגיל, ולגבי GenBank לא לכל הגנים יש תג לוקוס ישן, וזה מקשה להשוואה וגם לא נותן מספרים הגיונים, מלבד שחלק מהשורות בעמודה - Genes Name (ordered locus) ריקות.

אחד היתרון להשתמש בתג לוקוס, הוא בקובץ בכל החלבונים שנמצאים ב GenBank יש תג לוקוס.

סעיף 2:

1- זוהי ההיסטוגרמה של אחוז AT ברצפים טרנסממברניים שקיבלנו מ-UniProt:



הסטטיסטיקות:

ערך המקסימלי של אורכי הרצפים הטרנסממברניים : **44**

ערך המינימלי של אורכי הרצים הטרנסממברניים: **11**

ממוצע אורכי הרצים הטרנסממברניים: **21.39673760194994**

2- אחוז חומצת אמינו הידרופובי בכל הרצפים הטרנסממברניים הוא:

70.44370155845408

הערך הממוצע של אחוז חומצת אמינו הידרופובי ברצפים טרנסממברניים הוא:

70.48859100969469

מצאנו שהממוצע אורך הרצפים בטרנסממברניים הוא **21.39673760194994** והערך הממוצע של אחוז חומצת אמינו הידרופובי ברצפים טרנסממברניים הוא **70.48859100969469**, זה פחות מצפוי שלנו (21.39673760194994/ 18 = 84.124974259) אבל זה קרוב, ועל סמך המקורות האלא:

"The most genera description of the transmembrane helical regions (TMs) is that they comprise a region of 18 or more amino acids with a largely hydrophobic character." National Library of Medicine

Other source say : "**A length of helix of 18–21 amino acid residues is sufficient to span the usual width of a lipid bilaye."** ScienceDirect .

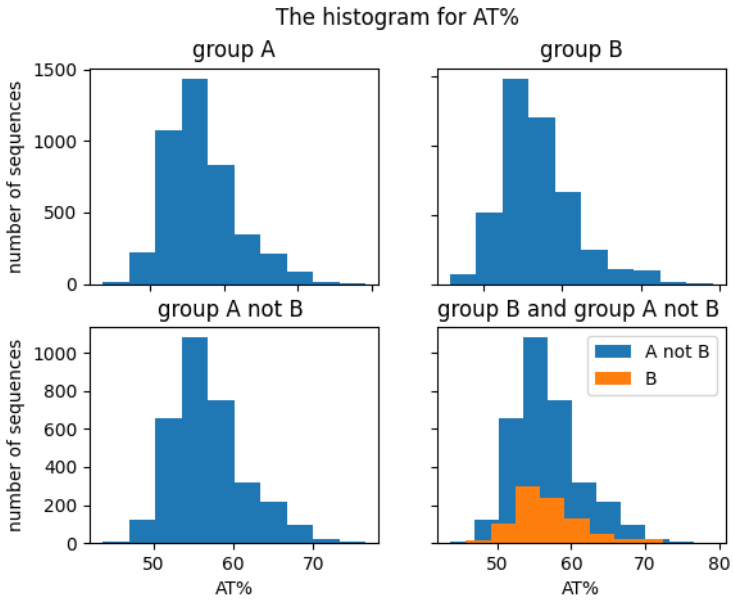
סעיף 3:

התפלגות של AT% ברצפי גינים של קבוצה B: **55.980773873645326**

סטטיסטיקה של גנים מקבוצה B בהשוואה לקבוצה A:

Standards/Groups	GroupA	GroupB
Maximum	16467	3606
Minimum	63	78
Average	874.319802	923.046380
Standard deviation	796.855018	543.837848

זוהי ההיסטוגרמה של אחוז AT ברצפים בקבוצות: קבוצה A, קבוצה B, הגנים ב-A שאינם ב-B, קבוצה B



והגנים ב-A שאינם ב-B:

חלק 3

1- בנינו פונקציה שמחשבת עבור כל קודון כמה עמדות הן סינונימיות הפונקציה מבצעת את החישוב שלמדנו בהרצאות (הסבר מפורט על איך הפונקציה עובדת כתוב כהערות בקוד).

תוצאות הריצה:

{
"ATA":0.6666666666666666,
"ATC":0.6666666666666666,
"ATT":0.6666666666666666,
"ATG":0.0,
"ACA":1.0,
"ACC":1.0,

"ACG":1.0,
"ACT":1.0,
"AAC":0.3333333333333333,
"AAT":0.3333333333333333,
"AAA":0.375,
"AAG":0.375,
"AGC":0.3333333333333333,
"AGT":0.3333333333333333,
"AGA":0.75,
"AGG":0.6666666666666666,
"CTA":1.3333333333333333,
"CTC":1.0,
"CTG":1.3333333333333333,
"CTT":1.0,
"CCA":1.0,
"CCC":1.0,
"CCG":1.0,
"CCT":1.0,
"CAC":0.3333333333333333,
"CAT":0.3333333333333333,
"CAA":0.375,
"CAG":0.375,
"CGA":1.5,
"CGC":1.0,
"CGG":1.3333333333333333,
"CGT":1.0,
"GTA":1.0,
"GTC":1.0,
"GTG":1.0,
"GTT":1.0,
"GCA":1.0,
"GCC":1.0,
"GCG":1.0,
"GCT":1.0,
"GAC":0.3333333333333333,
"GAT":0.3333333333333333,
"GAA":0.375,

"GAG":0.375,
"GGA":1.125,
"GGC":1.0,
"GGG":1.0,
"GGT":1.0,
"TCA":1.2857142857142858,
"TCC":1.0,
"TCG":1.125,
"TCT":1.0,
"TTC":0.3333333333333333,
"TTT":0.3333333333333333,
"TTA":0.8571428571428571,
"TTG":0.75,
"TAC":0.42857142857142855,
"TAT":0.42857142857142855,
"TAA":0.8571428571428571,
"TAG":0.375,
"TGC":0.375,
"TGT":0.375,
"TGA":0.375,
"TGG":0.0
}

2- השוואה בין וירוס Corona_2021 לבין Corona_2022:

א- כמה גנים יש בכל אחד מהם? מתוכם כמה גנים מקודדים לחלבונים?

הפונקציה שבנינו מחזירה את החישובים הנדרשים בשתי צורות:

1- כמספרים וזה על ידי לשלוח לה as_sets=False או לא לשלוח את המשתנה הזה בכלל כי ברירת המחדל היא False.

2- כמבני נתונים וזה על ידי לשלוח לה as_sets=True מבני הנתונים האלה מכילים את שמות הגנים אבל אחד מבני הנתונים הוא מילון שמכיל את שם הגן וכמות החלבונים הנוספים שהוא מתורגם להם למשל {'ORF1ab': 1} המשמעות היא שהגן 'ORF1ab' מתורגם לשתי חלבונים (ה- 1 במילון אמור שיש עוד תרגום [תרגום נוסף]).

Corona 2021	
Statistics As Numbers	
Total number of genes	11
Number of genes that could be converted to protein	11
Number of genes that couldn’t be converted to protein	0
Number of genes that could be converted to more than one protein	1
Number of CDS [The Function Doesn’t return this value but it could be calculated easily by the returned values] = total number of genes that could be converted to protein + the sum of the values of the returned dictionary in the case of as_sets=True as I explained above.	12

Corona 2021	
Statistics As Data Structures	
All genes	{'ORF10', 'ORF7b', 'ORF3a', 'ORF8', 'S', 'M', 'ORF7a', 'N', 'E', 'ORF1ab', 'ORF6'}
genes that could be converted to protein	{'ORF10', 'ORF7b', 'ORF3a', 'ORF8', 'S', 'M', 'ORF7a', 'N', 'E', 'ORF1ab', 'ORF6'}
genes that couldn’t be converted to protein	None
genes that could be converted to more than one protein	{'ORF1ab': 1}

Corona 2022	
Statistics As Numbers	
Total number of genes	11
Number of genes that could be converted to protein	11
Number of genes that couldn’t be converted to protein	0
Number of genes that could be converted to more than one protein	1
Number of CDS [The Function Doesn’t return this value but it could be calculated easily by the returned values] = total	12

number of genes that could be converted to protein + the sum of the values of the returned dictionary in the case of as_sets=True as I explained above.	

Corona 2022	
Statistics As Data Structures	
All genes	{'ORF10', 'ORF7b', 'ORF3a', 'ORF8', 'S', 'M', 'ORF7a', 'N', 'E', 'ORF1ab', 'ORF6'}
genes that could be converted to protein	{'ORF10', 'ORF7b', 'ORF3a', 'ORF8', 'S', 'M', 'ORF7a', 'N', 'E', 'ORF1ab', 'ORF6'}
genes that couldn't be converted to protein	None
genes that could be converted to more than one protein	{'ORF1ab': 1}

מסקנה: שתי הווירוסים Corona_2021 ו- Corona_2022 יש להם את אותן הסטטיסטיקות.

ב- כמה גנים משותפים יש ביניהם (הסתמכו על שמות גנים)?האם יש גנים שיש באחד ולא באחר? אם כן, פרטו את רשימת השמות.

הפונקציה שבנינו מחזירה את החישובים הנדרשים בשתי צורות:

1- כמספרים וזה על ידי לשלוח לה as_sets=False או לא לשלוח את המשתנה הזה בכלל כי ברירת המחדל היא False.

2- כמבני נתונים וזה על ידי לשלוח לה as_sets=True מבני הנתונים האלה מכילים את שמות הגנים.

Corona 2021 – Corona 2022	
Statistics As Numbers	
Total number of shared genes	11
Number of genes that exists just in first virus (Corona 2021) and not in the second (Corona 2022)	0
Number of genes that exists just in second virus (Corona 2022) and not in first second (Corona 2021)	0

Corona 2021 – Corona 2022	
Statistics As Data Structures	
Shared genes	{'ORF1ab', 'ORF8', 'ORF7b', 'ORF6', 'ORF3a', 'ORF10', 'M', 'S', 'E', 'N', 'ORF7a'}
Genes that exist just in first virus (Corona 2021) and not in the second (Corona 2022)	None
Genes that exist just in second virus (Corona 2022) and not in first second (Corona 2021)	None

מסקנה: כל הגנים שיש לווירוסים הם גנים משותפים עם אותם השמות של הגנים זאת אומרת שאין גנים נוספים או חסרים אצל אחד הווירוסים.

ג- בחרו חמישה גנים משותפים וחשבו עבור כל גן את מדד ה- dnds:

Shared Gene Name	dN	dS	dN_dS_ratio	Selection Type
E	0	0	Zero or can't be calculated	Special case it could be negative or neutral
M	0.00394997621	0	Zero or can't be calculated	Special case it could be negative or neutral
ORF8	0	0	Zero or can't be calculated	Special case it could be

				negative or neutral
N	0.00105078826	0.00353357544	0.2973725283	negative
ORF7b	0	0.04000948552	0	negative

הסבר התוצאות: dN=dS=0 או שה- dS=0 זה מקרה פרטי שפירושו שהרצף לא השתנה וזה מעיד על רך שיש סלקציה שלילית עבור הגנים האלה (לשימור הרצף) או שאין מספיק מדע בשלב הזה כדי שנוכל לקבוע ואולי זו סלקציה ניטרלית במילות אחרות לא עבר מספיק זמן כדי שיצטברו מוטציות כדי שנוכל לקבוע את זה.

גם במקרים האחרים ה- dN_dS_ratio שקיבלנו הוא או קטן מ-1 וקרוב ל-0 ולפי מה שלמדנו זה מעיד על סלקציה שלילית כלומר: *״בעקבות המוטציה, הפיטנס של האלל החדש יורד - הסיכויים של היצור המוטנט לשרוד קטנים יותר בעקבותיה (במילות אחרות סלקציה "עובדת נגד" המוטציה)״.*