אלעד דוד 206760274
דנה ברודו 318379914
ענבר שמייה 208774026
פברואר, 2022

עבודת סיום תשפ"ב , 2022 – קורס ביולוגיה חישובית (10554)

שיקולים שבוצעו בעת כתיבת הקוד

הקוד נכתב בפלטפורמת Google Colab:

בעת הריצה דרך ה-Google Colab יש לטעון שני קבצים רלוונטיים לאזור התוכן של האפליקציה Google Drive, והם: UniProt.csv ו-BS168.gb.

הקוד חולק לשלושה חלקים (בלתי תלויים ככל הניתן): חלק א', חלק ב' וחלק ג' (כמתואר בהוראות התרגיל). בעת כתיבת הקוד הייתה השתדלות רבה לכתוב באופן גנרי ככל הניתן, ועם זאת ליצור פונקציה אחת לכל שאלה/ סעיף בתרגיל שתמוקד בהוראות העבודה.

דאגנו כי לכל פונקציה יופיע תיאור בתחילתה בו מפורטים: תפקיד הפונקציה, פרמטרים אותם הפונקציה מקבלת וערכים אותם היא מחזירה. הושקעה מחשבה בהדפסות ובתוצרי הריצה המופיעים למשתמש המריץ את הקוד, תוך דגש על אסתטיקה והבנה.

חלק א'

שאלה 1: הכרת וספירת האלמנטים בגנום

נספרו האלמנטים השונים בחיידק Bacillus subtilis והודפסו מספר המופעים שלהם. להלן תוצאת הרצת הקוד:

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

שאלה 2: אפיון אורכי גנים

א. עבור קובץ ה-GeneBank שנטען, חושבו האורכים של כל הגנים של החיידק Bacillus subtilis.

ב. הגנים חולקו לשתי קבוצות: גנים המקודדים לחלבון וכל השאר.

ג. דווחו סטטיסטיקות עבור כל אחת מקבוצות הגנים מהסעיף הקודם. להלן תוצאת הרצת הקוד:

```
Average length of the protein-coding genes:  874.5702147746047
Maximum length of the protein-coding genes:  16467
Minimum length of the protein-coding genes:  63

Average length of the genes that aren't coding to protein:  324.12
Maximum length of the genes that aren't coding to protein:  2928
Minimum length of the genes that aren't coding to protein:  33
```

ד. צוירו שלוש היסטוגרמות:

1) היסטוגרמה של אורכי כל הגנים.
2) היסטוגרמה של אורכי כל הגנים שמקודדים לחלבון.
3) היסטוגרמה של אורכי כל יתר הגנים (גנים שאינם מקודדים לחלבון).

אלעד דוד 206760274
דנה ברודו 318379914
ענבר שמייה 208774026
פברואר, 2022

ניתן לצפות בהן בפרק הנספחים: *נספח א': שאלה 2: אפיון אורכי גנים – סעיף ד'.*

ניתן לראות כי הגנים שאינם מקודדים לחלבון קצרים באופן משמעותי מהגנים שמקודדים לחלבון. הסבר הגיוני לכך הוא כי הגנים שאינם מקודדים לחלבון מהווים מעין אמצעי "עזר" לתהליך התרגום לחלבון. כך למשל הריבוזום (rRNA) מעניק את ה"צינור" דרכו התרגום מתבצע. הפלט המרכזי של ה-DNA הוא החלבון אליו הוא מתורגם, וזו הסיבה העיקרית לשמירה של הקוד (לצורכי שכפול ובקרה). אם המצב היה הפוך, וגנים שאינם מקודדים לחלבון היו ארוכים יותר, היה "מתבזבז" זיכרון רב בקוד ה-DNA עבור רכיבים רגולטוריים ועזרים למיניהם, ולא לעיקר.

שאלה 3: חישוב אחוז GC בגנים

א. אחוז ה-GC הממוצע בגנום החיידק (ברצף הגנום כולו) הוא: 43.51440813017155%

ב. לכל גן אשר מקודד לחלבון חושב GC%.

הממוצע על פני כל הגנים אשר מקודדים לחלבון הוא: 43.119097854054765%

ג. נשים לב כי GC% עבור הגנים שמקודדים לחלבון הוא נמוך יותר ב-0.4% מהגנום כולו. נתון זה הגיוני מכיוון שמבחינה כימית, המולקולות שמייצגות את הנוקלאוטיד G והנוקלאוטיד C נוטות להיצמד אחת לשנייה, דבר הגורם לסלילי ה-DNA להיצמד גם כן ולהקשות על המעבר עליו צורך התרגום לחלבון. נשים לב כי 0.4% זהו הפרש לא גדול במיוחד, אך חשוב לזכור כי בסעיף א' חושב הממוצע של כלל הגנים- כולל הגנים שמקודדים לחלבון, דבר שהעלה את הממוצע ככל הנראה באופן משמעותי.

ד. צוירה היסטוגרמה המציגה את התפלגות GC% עבור הגנים המקודדים לחלבון.

תוצאת הרצת הקוד נמצאת בנספחים: *נספח ב': שאלה 3: חישוב אחוז **GC** בגנים – סעיף ד'*

ה. חושבו מהם חמשת הגנים העשירים ביותר ב-GC ומהם חמשת הגנים עם הרכב ה-GC הנמוך ביותר מבין כלל הגנים של החיידק.

תוצאת הרצת הקוד נמצאת בנספחים: *נספח ג': שאלה 3: חישוב אחוז **GC** בגנים – סעיף ה'*

שאלה 4: בדיקות עקביות בקובץ הדאטה

בשאלה זו נבדק הרצף הגנטי במספר אופנים: בדיקת תקינות בהיבט של כלל הרצף (בדיקה שאכן כלל הנוקלאוטידים קבילים), בדיקה כי אכן כל גן שמתורגם לחלבון מתחיל בקודון התחלה ('M') ו בדיקה כי אכן כל גן שמתורגם לחלבון מסתיים בקודון עצירה ('_'). הבדיקה הראשונה לא הניבה כשלים וחריגות, אך הבדיקות האחרות איתרו לא מעט רצפי CDS שאינם מתחילים ומסתיימים בקודונים המתאימים. התוצאות דווחו לקובץ gene_exceptions.csv.

סקטור הערות: הערה ב'

עבור כל גן נשמר מידע אודותיו (למשל מיקום, strand, שם), סוג הגן (מקודד לחלבון, רנ"א וכו') ומידע נוסף שחושב (למשל הרכב GC). כלל הגנים מוינו לפי קואורדינטת ההתחלה ונשמרו לקובץ מסוג csv בשם part_a.csv.

אלעד דוד 206760274
דנה ברודו 318379914
ענבר שמייה 208774026
פברואר, 2022

חלק ב'

א. הצלבנו בין החלבונים מקובץ ה-GeneBank ובין החלבונים מקובץ ה-UniProt. יש חלבונים שנמצאים בקובץ הראשון אך לא בשני. כמתנו את ההפרשים, והדגמנו עם ויזואליזציה מתאימה. ההבדלים נובעים מכך שמדובר במקורות שונים עם עדכונים המתוארכים שונה ומחקר אודות הגנום שאינו משותף ככל הנראה לשני המאגרים. פרמטר נוסף שעשוי להשפיע על ההבדלים בין המאגרים הוא המזהה שבחרנו לכל גן- שם הגן עשוי להיות מזהה לא מדויק ואחיד בין שני המאגרים, ולפיכך התוצאות יהיו שונות ולא חופפות בהכרח.
ההצלבה בוצעה באמצעות שמות הגנים, מכיוון ששמנו לב כי תכונה זו משותפת לגנים רבים (למשל לעומת protein_id שנמצא רק בגנים מסוג מסוים, או type שאינו מזהה ייחודי לכל הגנים). ניתן לצפות בוויזואליזציה המתאימה בנספחים: *<u>נספח ד':</u> חלק ב' – סעיף א'*

ב. שלפנו את הרצפים הטרנסממברנליים ואפיינו אותם.
האורך הממוצע הוא: 20.395294338207723, המינימלי הוא: 10 והמקסימלי הוא: 43.
התפלגות אחוז חומצות האמינו ההידרופוביות ברצפים האלה מתוארת בנספחים: *<u>נספח ה':</u> חלק ב' – סעיף ב'*. הערך הממוצע על פני כל הרצפים הללו: 70.03451651599204%
זה תואם לציפיות שלנו מאזורים כאלה, שכן אזורים אלה נוטים להיות הידרופוביים בכדי ליצור משקעים במים ולאפשר הובלה של חומרים ספציפיים על פני הממברנה.

ג. נסמן את קבוצת הגנים שהם CDS באות A. עבור רצפי הגנים שנמצאו בחיתוך בין ה-UniProt ובין ה-GeneBank, הגנים שמכילים לפחות אזור טרנסממברנלי אחד סומנה ב-B.
התפלגות GC% ברצפי הגנים בקבוצה B וכמו כן גם טבלה המסכמת את הסטטיסטיקות עבור קבוצת גנים, שני אלה מופיעים בנספחים: *<u>נספח ו':</u> חלק ב' – סעיף ג'*
צוירו 4 היסטוגרמות המסכמות את אחוזי ה-GC של קבוצות שונות, תוצאת ההרצה מפורטת בנספחים: *<u>נספח ז':</u> חלק ב' – סעיף ג' (4 גרפים)*

חלק ג'

שאלה 1

עבור הקוד הגנטי המתאים, חושב עבור כל קודון את מספר העמדות הסינונימיות.

שאלה 2

א. כלל הגנים משותפים לשני הקבצים- לקובץ המתאר את הווירוס מיולי 2020 ולקובץ המתאר את הווירוס מינואר 2022.

ב. נבחרו חמישה גנים משותפים וחושב עבור כל אחד מהם את יחס ה- d_N/d_S.
הטבלה מפורטת בנספחים: *<u>נספח ח':</u> חלק ג' – שאלה 2*

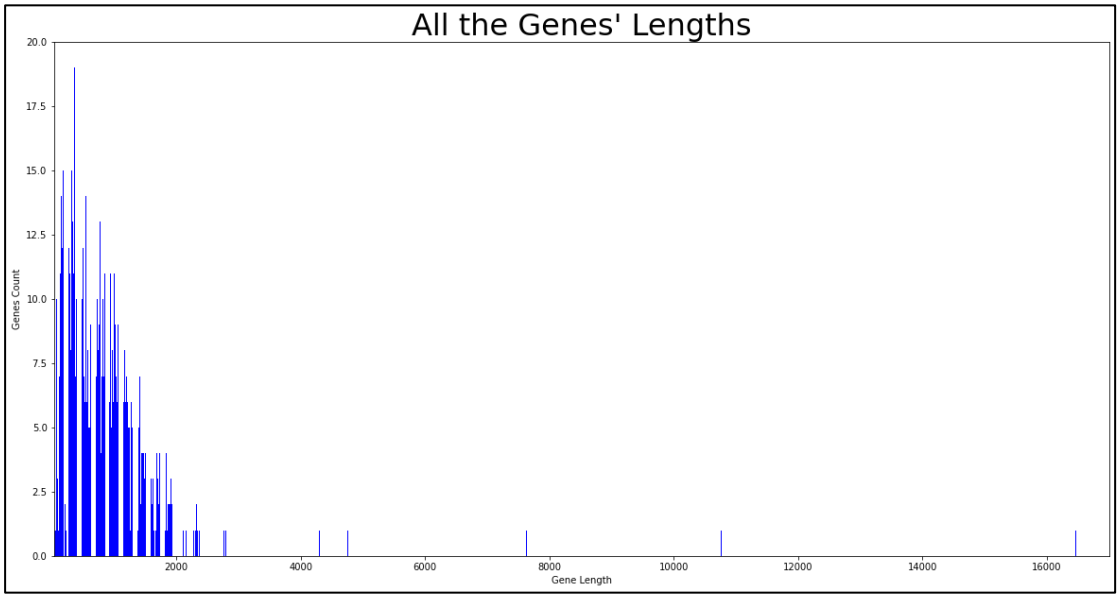
אלעד דוד 206760274
דנה ברודו 318379914
ענבר שמייה 208774026

פברואר, 2022

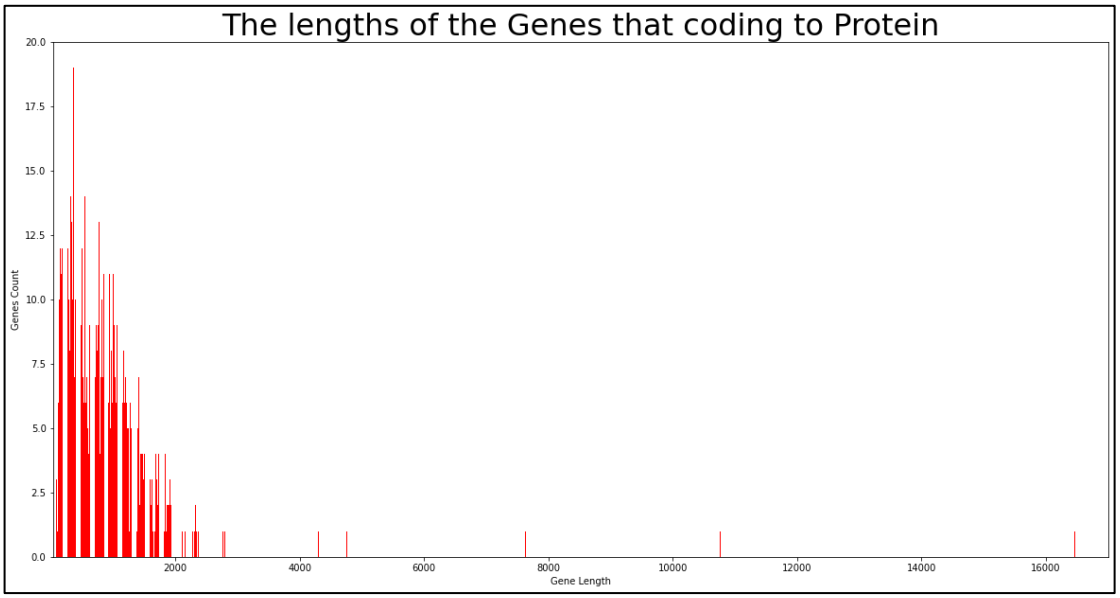
נספחים

נספח א': שאלה 2: אפיון אורכי גנים – סעיף ד'

גרף 1: התפלגות אורכי כל הגנים של החיידק Bacillus subtilis.



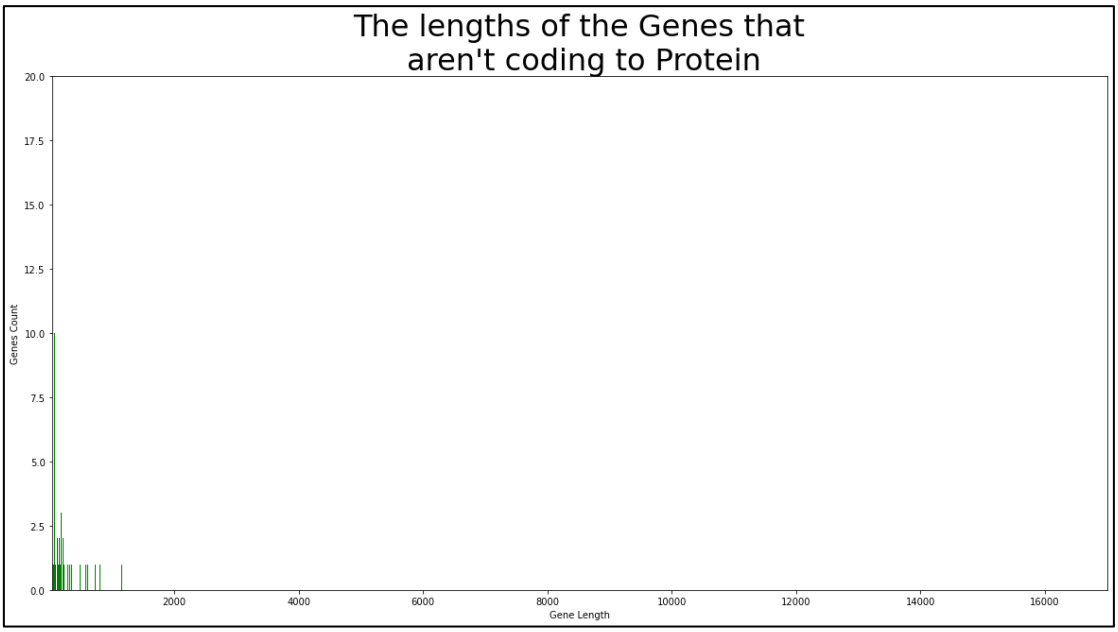
גרף 2: התפלגות אורכי כל הגנים שמתוגרמים לחלבון של החיידק Bacillus subtilis.



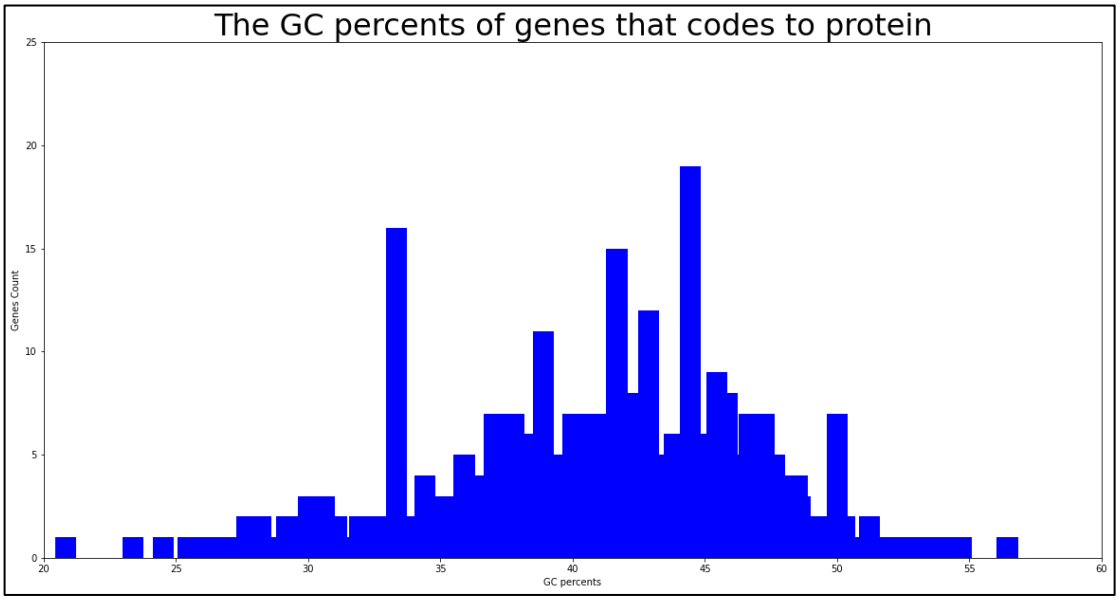
אלעד דוד 206760274
דנה ברודו 318379914
ענבר שמייה 208774026

פברואר, 2022

גרף 1: התפלגות אורכי כל הגנים שאינם מתורגמים לחלבון של החיידק Bacillus subtilis.



נספח ב': שאלה 3: חישוב אחוז GC בגנים – סעיף ד'



אלעד דוד 206760274
דנה ברודו 318379914
ענבר שמייה 208774026
פברואר, 2022

נספח ג': שאלה 3: חישוב אחוז GC בגנים – סעיף ה'

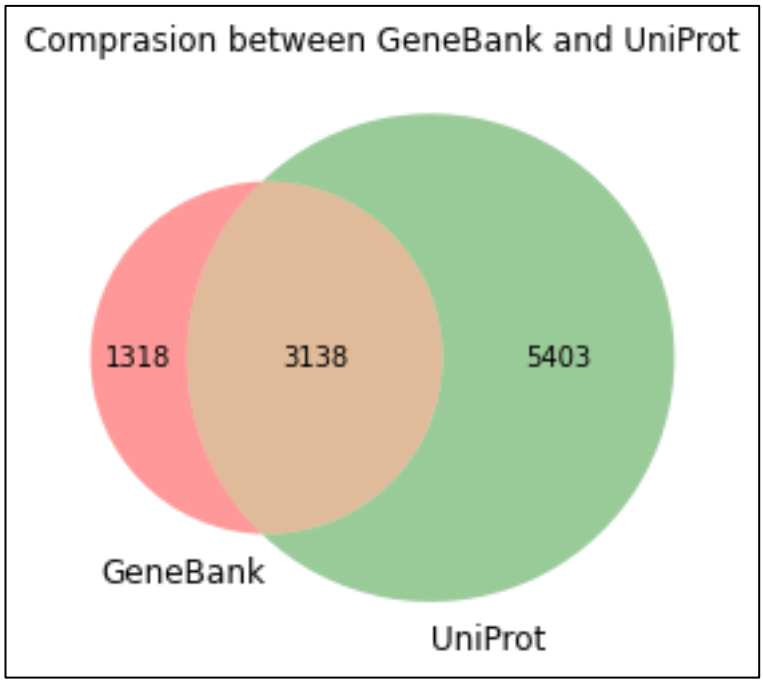
Five lowest genes:

1 | name: yqaD | gene details: [2699509:2699677](-) | GC percents: 20.833333333333336 %
2 | name: cotC | gene details: [1904994:1905195](-) | GC percents: 23.383084577114428 %
3 | name: cotU | gene details: [1901116:1901377](-) | GC percents: 24.521072796934863 %
4 | name: rtbE | gene details: [4036343:4036787](-) | GC percents: 25.45045045045045 %
5 | name: yydC | gene details: [4132337:4132736](-) | GC percents: 25.81453634085213 %

Five highest genes:

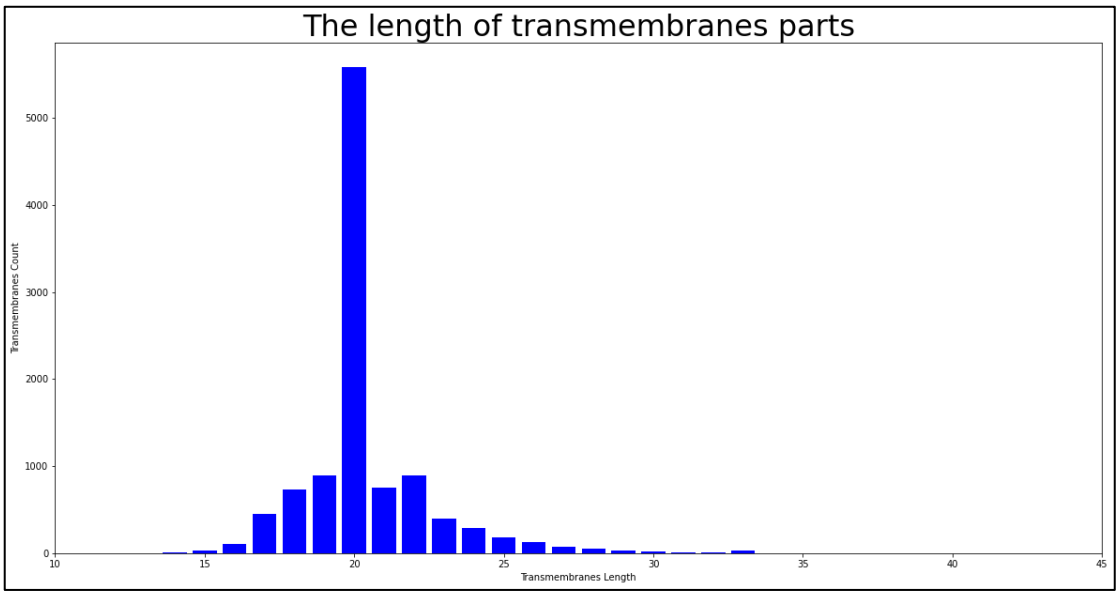
1 | name: rhgX | gene details: [772141:773980](+) | GC percents: 53.833605220228385 %
2 | name: ydaS | gene details: [492653:492911](-) | GC percents: 54.263565891472865 %
3 | name: aag | gene details: [3964277:3964868](-) | GC percents: 54.314720812182735 %
4 | name: epsM | gene details: [3516232:3516883](-) | GC percents: 54.68509984639017 %
5 | name: nnrA | gene details: [3972447:3973278](-) | GC percents: 56.438026474127554 %

נספח ד': חלק ב' – סעיף א'

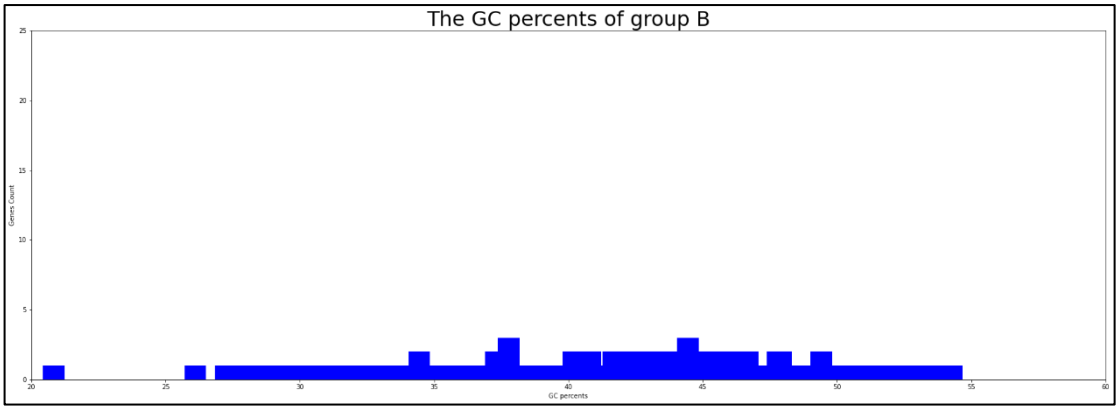


אלעד דוד 206760274
דנה ברודו 318379914
ענבר שמייה 208774026
פברואר, 2022

נספח ה': חלק ב' – סעיף ב'



נספח ו': חלק ב' – סעיף ג'



A group statistics:

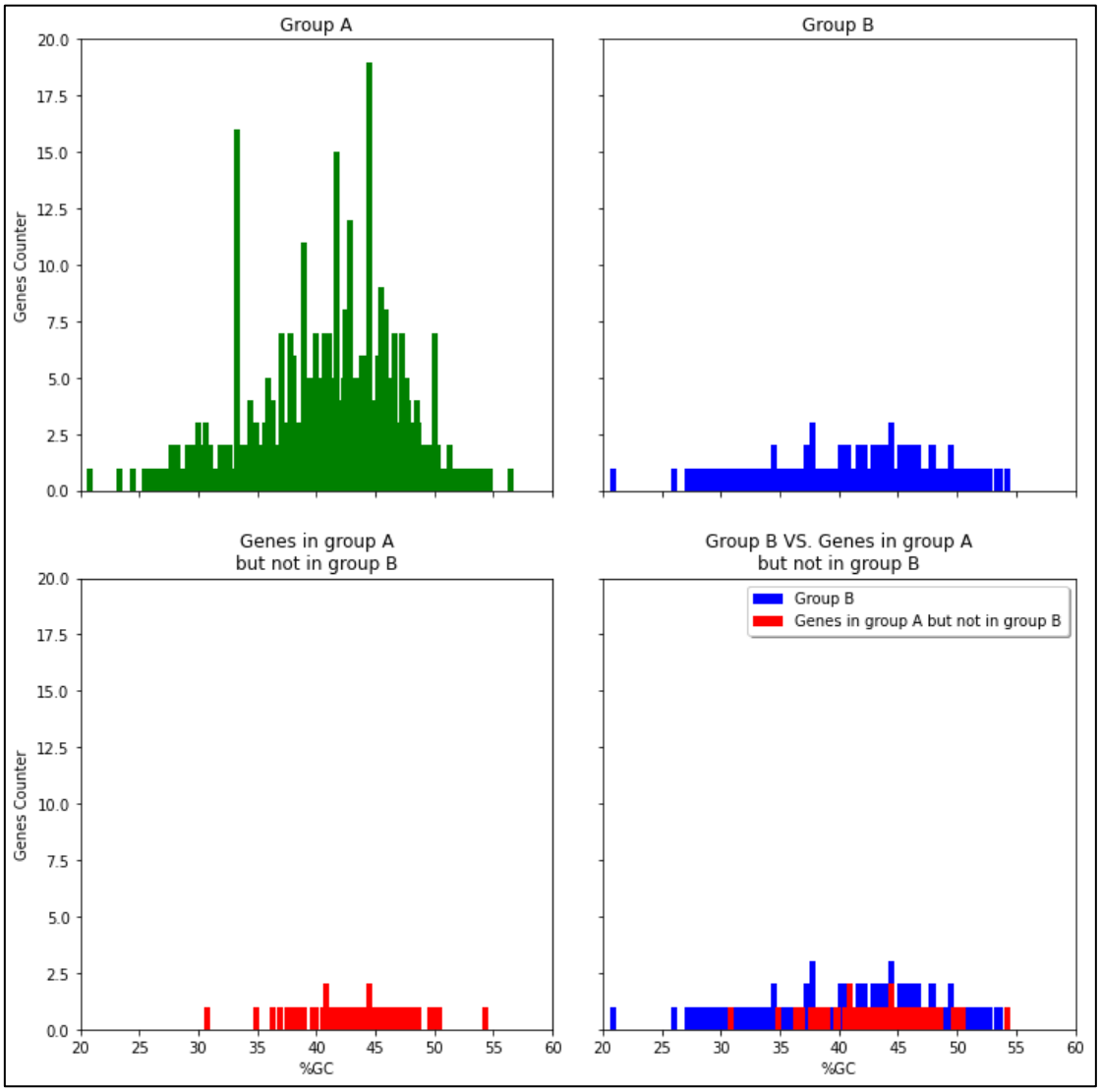
Min: 20.833333333333336 , Max: 56.438026474127554 , Median: 43.890865954922894 , Average: 43.11909785405482

B group statistics:

Min: 20.833333333333336 , Max: 54.263565891472865 , Median: 43.82716049382716 , Average: 43.040212766534076

אלעד דוד 206760274
דנה ברודו 318379914
ענבר שמייה 208774026
פברואר, 2022

נספח ז': חלק ב' – סעיף ג' (4 גרפים)



נספח ח': חלק ג' – שאלה 2

Name	Product	Protein ID	Strand	Start Loc.	End Loc.	dN/dS Ratio	Selection
ORF6	ORF6 protein	YP_009724394.1	1	27201	27387	1	Neutral
ORF7a	ORF7a protein	YP_009724395.1	1	27393	27759	1	Neutral
ORF7b	ORF7b	YP_009725318.1	1	27755	27887	1	Neutral
ORF8	ORF8 protein	YP_009724396.1	1	27893	28259	1	Neutral
N	nucleocapsid phosphoprotein	YP_009724397.2	1	28273	29533	1	Neutral