**ביולוגיה חישובית – פרויקט סוף קורס 2023**

אביב אלדד (311551477), אלחנן ששון (208272625),  
רוני פנדריך (208502849), שירן גלסר (208608174)

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

1. **הכרת וספירת האלמנטים בגנום:**

מספר האלמנטים מכל אזור בגנום החיידק:

gene: **4,536**, CDS: **4,237**, misc\_RNA: **93**, misc\_feature: **89**,  
tRNA: **86**, rRNA: **30**, ncRNA: **2**

1. **אפיון אורכי הגנים:**

עבור כל גן חישבנו את אורכו והוספנו עמודה של האורך לDataFrame.

סטטיסטיקות על הגנים אשר מקודדים לחלבון:  
ממוצע – **874.5**7, מינימום – **63**, מקסימום – **16,467**, סטיית תקן – **797.17**.

סטטיסטיקות על הגנים אשר אינם מקודדים לחלבון:  
ממוצע – **324.12**, מינימום – **33**, מקסימום – **2,928**, סטיית תקן – **571.63**.

Chart

Description automatically generated**ההתפלגות עבור כל קבוצה:**

**מסקנות מהגרף:**

הגרף אכן תואם את מה שציפינו לקבל. ניתן לראות שאורך הגנים המקודדים לחלבון ארוכים יותר באופן משמעותי מהגנים אשר אינם מקודדים לחלבון. גנים המקודדים לחלבון צריכים להיות בעלי קידוד אחיד על מנת שיתורגמו לחומצות אמינו ולאחר מכן לחלבון, לכן אורכם לרוב יהיה ארוך יותר מגנים אשר אינם מקודדים לחלבון. לעומת זאת, גנים אשר אינם מקודדים לחלבון בד"כ יהיו רצפי בקרה המקשרים בין חלבונים, ולכן אורכם יהיה בד"כ קצר יותר.

1. **חישוב אחוז AT בגנים:**

ממוצע אחוז ה-AT בגנום החיידק (ברצף הגנום כולו): **56.49%**

עבור כל גן חישבנו את אחוז ה-AT שלו והוספנו עמודה המציגה אחוז זה לכל גן לתוך הdataframe של קובץ הGenBank.  
ממוצע אחוז ה-AT בגנים אשר מקודדים לחלבון: **56.88%**

מבחינה מתמטית התוצאות אכן תואמות את מה שצפינו מכיוון ש-93.4% מהגנום כולו מקודד לחלבונים. כלומר עבור 93.4% יש 56.88% AT ולכן ה6.6% הנותרים לא השפיעו הרבה על ממוצע הרצף כולו.

**Chart, histogram

Description automatically generatedהתפלגות AT% בחלבונים:**

**חמשת הגנים עם אחוז הAT הגבוה ביותר:**

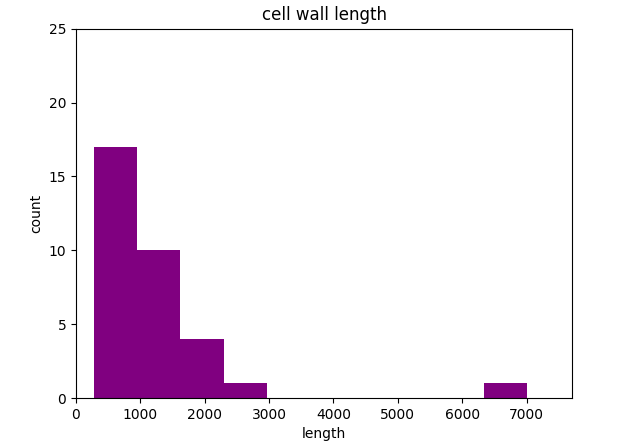
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **%AT** | **Length** | **Type** | **Strand** | **End** | | **Start** | | **Name** | | **No** | |
| 79.16667 | 168 | CDS | -1 | | 2699677 | | 2699509 | | BSU\_26360 | | **1** |
| 76.61692 | 201 | CDS | -1 | | 1905195 | | 1904994 | | BSU\_17700 | | **2** |
| 75.47893 | 261 | CDS | -1 | | 1901377 | | 1901116 | | BSU\_17670 | | **3** |
| 74.54955 | 444 | CDS | -1 | | 4036787 | | 4036343 | | BSU\_39290 | | **4** |
| 74.18546 | 399 | CDS | -1 | | 4132736 | | 4132337 | | BSU\_40210 | | **5** |

**חמשת הגנים עם אחוז הAT הנמוך ביותר:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **%AT** | **Length** | **Type** | **Strand** | **End** | **Start** | **Name** | **No** |
| 32.87671 | 73 | tRNA | -1 | 3194527 | 3194454 | BSU\_tRNA\_86 | **1** |
| 34.21053 | 76 | tRNA | 1 | 11627 | 11551 | BSU\_tRNA\_6 | **2** |
| 34.21053 | 76 | tRNA | 1 | 32095 | 32019 | BSU\_tRNA\_9 | **3** |
| 34.21053 | 76 | tRNA | 1 | 96221 | 96145 | BSU\_tRNA\_20 | **4** |
| 34.21053 | 76 | tRNA | 1 | 166328 | 166252 | BSU\_tRNA\_28 | **5** |

1. **גנים המכילים בתיאור cell wall:**

מספר הגנים אשר מכילים בתיאור את צמד המילים cell wall: 33

**סטטיסטיקות עבור אורך אותם 33 גנים:**

ממוצע: **1,180.82**  
מינימום: **276**  
מקסימום: **7,005**  
סטיית תקן: **1,160.46**

Chart, histogram

Description automatically generated**סטטיסטיקות עבור אחוז AT באותם 33 גנים**:

ממוצע: **56.89**  
מינימום: **47.79**  
מקסימום: **64.41**  
סטיית תקן: **3.68**

1. **בדיקת עקביות בקובץ הדאטה:**

על מנת לבדוק את עקביות קובץ הדאטה השתמשנו בספריית Bio של פייתון.

**השיקולים לבדיקה שהגדרנו:**  
1. לבדוק האם רצף הגן מתחיל ב-start codon.

2. לבדוק האם רצף הגן מסתיים ב-stop codon ולא קיים כזה באמצע הרצף.

3. לבדוק האם אורך הרצף מתחלק ב-3

4. לבדוק האם התרגום נכון.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Description** | **Table** | **Type** | **Strand** | **End** | **Start** | **Name** | **No** |
| Sequence length 1798 is not a multiple of three | 11 | CDS | -1 | 2161778 | 2159980 | BSU\_20040 | **1** |
| Extra in frame stop codon 'TGA' found. | 11 | CDS | -1 | 2165614 | 2162107 | BSU\_20060 | **2** |
| Sequence length 1102 is not a multiple of three | 11 | CDS | -1 | 3628240 | 3627138 | BSU\_35290 | **3** |

**חלק ב' – אנליזת חלבונים בעזרת אתר ה-UniProt**

1. **הצלבה בין החלבונים מGenBank לחלבונים בUniProt:**

את ההצלבה ביצענו על סמך העמודה (Gene Names (ordered locus)) באתר הUniProt. בחרנו בעמודה לאחר בדיקה של כמה עמודות אחרות של שמות הגנים. הגענו למסקנה שזו העמודה שהכי דומה לעמודת השמות בקובץ הGenBank, ובנוסף נראה שזו העמודה המלאה ביותר מבחינת ערכים. לאחר שבחרנו בעמודה, היה צורך לעשות הכנה של המידע לפני ההצלבה. ב UniPort היו כמה רשומות עם כמה מזהים לאותו חלבון והיה צורך לפצל, וב GenBank השם הופיע עם מקף תחתון.

**להלן סיכום ההבדלים:**

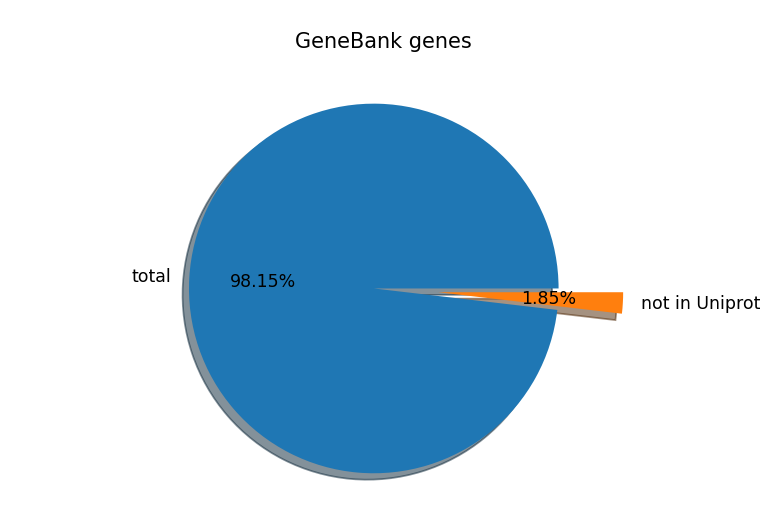
In UniProt and not in GenBank:

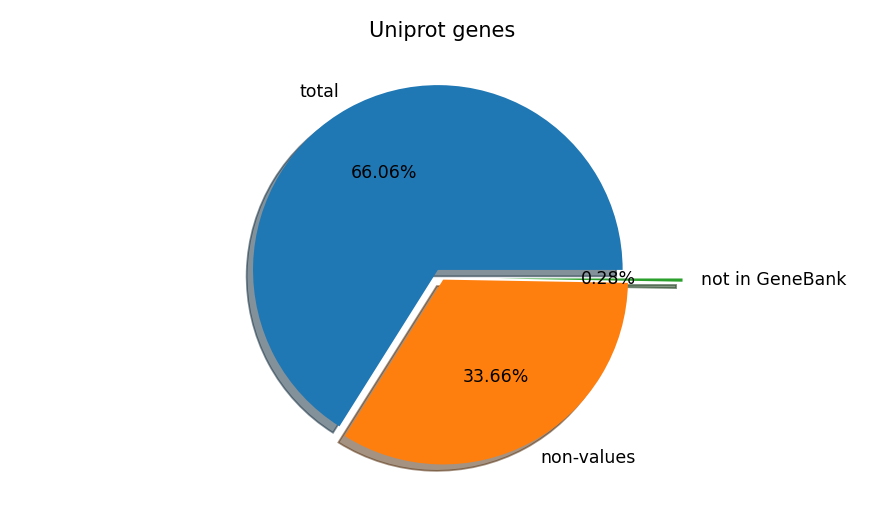
]'BSU12670', 'BSU28480', 'BSU25760', 'BSU06810', 'BSU22660', 'BSU11381', 'BSU39030', 'BSU06050', 'BSU16640', 'BSU13799', 'BSU07740', 'BSU20030', 'BSU16900', 'BSU35690', 'BSU23290', 'BSU34410', 'BSU18940', 'BSU33220', 'BSU35230', 'BSU03570', 'BSU26080', 'BSU02180', 'BSU35150', 'BSU13790', 'BSU35610', 'BSU01840', 'BSU26390', 'BSU34420', 'BSU16890', 'BSU40020', 'BSU18930', 'BSU11382', 'BSU16840', 'BSU18110', 'BSU07180', 'BSU26040'] **missing: 36, nan values: 4350, total: 8538**

In GenBank and not in UniProt:

]'BSU11525', 'BSU22036', 'BSU35678', 'BSU12671', 'BSU17099', 'BSU02585', 'BSU32539', 'BSU26826', 'BSU04345', 'BSU37089', 'BSU26827', 'BSU04745', 'BSU27085', 'BSU28645', 'BSU19745', 'BSU27035', 'BSU17715', 'BSU24205', 'BSU29479', 'BSU19915', 'BSU13545', 'BSU17679', 'BSU31725', 'BSU06812', 'BSU26449', 'BSU25565', 'BSU26399', 'BSU18978', 'BSU26569', 'BSU18595', 'BSU26935', 'BSU11800', 'BSU21925', 'BSU07735', 'BSU22205', 'BSU40358', 'BSU12815', 'BSU27009', 'BSU26305', 'BSU40022', 'BSU18275', 'BSU16845', 'BSU36739', 'BSU36215', 'BSU36668', 'BSU04536', 'BSU21058', 'BSU17845', 'BSU01790', 'BSU21638', 'BSU17689', 'BSU28709', 'BSU03385', 'BSU36575', 'BSU21639', 'BSU25875', 'BSU27185', 'BSU33221', 'BSU11515', 'BSU30466', 'BSU40576', 'BSU29845', 'BSU27786', 'BSU21409', 'BSU26075', 'BSU31289', 'BSU02785', 'BSU21546', 'BSU36079', 'BSU12875', 'BSU18596', 'BSU04849', 'BSU27935', 'BSU14568', 'BSU26055', 'BSU18689', 'BSU34399', 'BSU28475', 'BSU37569', 'BSU38495'] **missing: 80, nan values: 0, total: 4237**

**ובצורה ויזואלית:**



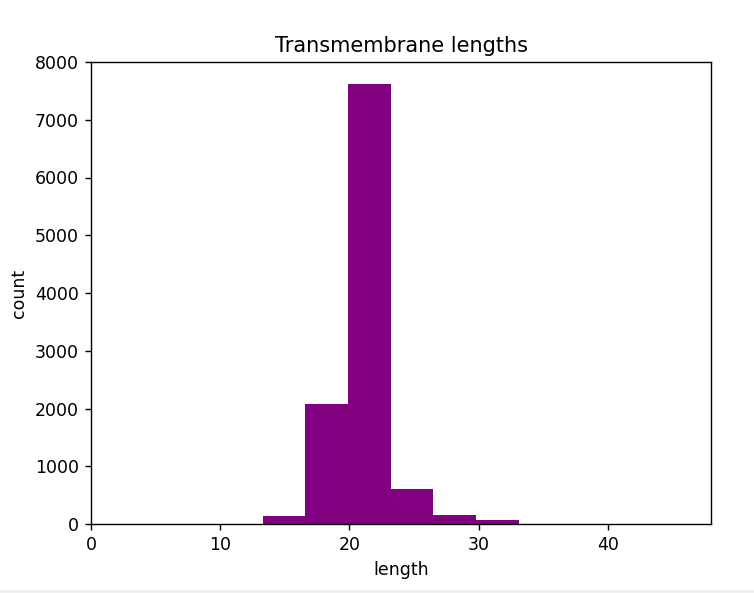


ההבדלים נובעים מהשוני בין האתרים (בעצם בין סוגי הקבצים והמידע). בעוד באתר UniPort מוצגים רק גנים שנחקרו ויש עליהם מידע, בGenBank מוצגים כל הרצפים שמקודדים לחלבון בגן.

1. **אפיון של רצפים** **טרנסממברנליים:**

בודדנו את הרשומות בהן העמודה של הרצף הטרנסממברנלי לא ריק, ובנינו DF חדש עבור הרצפים הטרנסממברנליים, כאשר לכל שורה פירקנו את הרצפים שהיו בה (היו כמה רצפים בכל שורה שכזו). לאחר מכן הוספנו את עמודת האורכים.

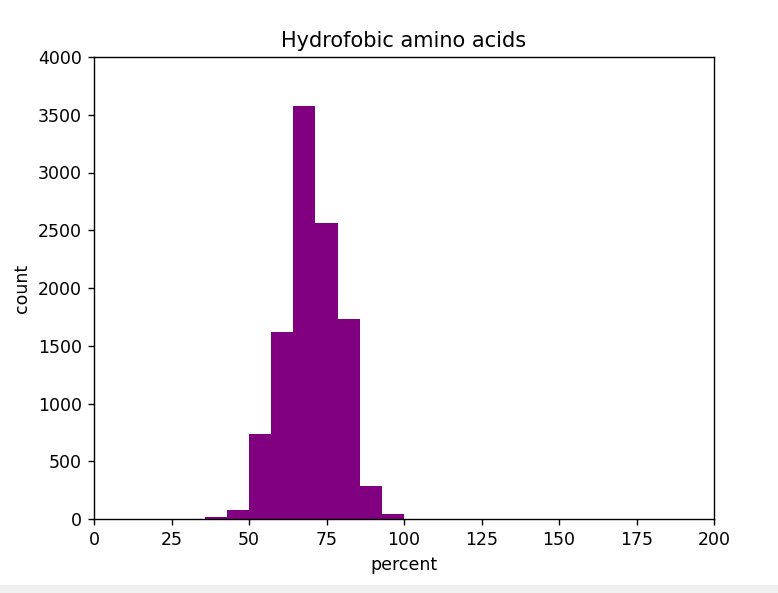
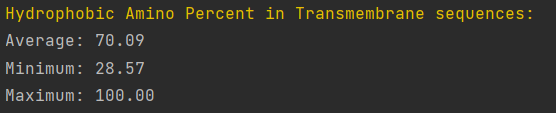
**התפלגות האורכים:**



**תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי**

**התפלגות אחוז חומצות האמינו ההידרופוביות:**



הממברנה של התא עשויה מפוספוליפידים, שבנויים מבחוץ מחומצות אמינו הידרופוליות, ומבפנים מחומצות אמינו הידרופוביות. על מנת שהחלבון יוכל לעבור בממברנה של התא, נצפה שיהיו אחוזים גבוהים של חומצות אמינו הידרופוביות ברצפים טרנסממברנליים, כמו שאכן קיבלנו, כלומר התוצאה אכן תואמת לציפייה. (https://en.wikipedia.org/wiki/Transmembrane\_domain)

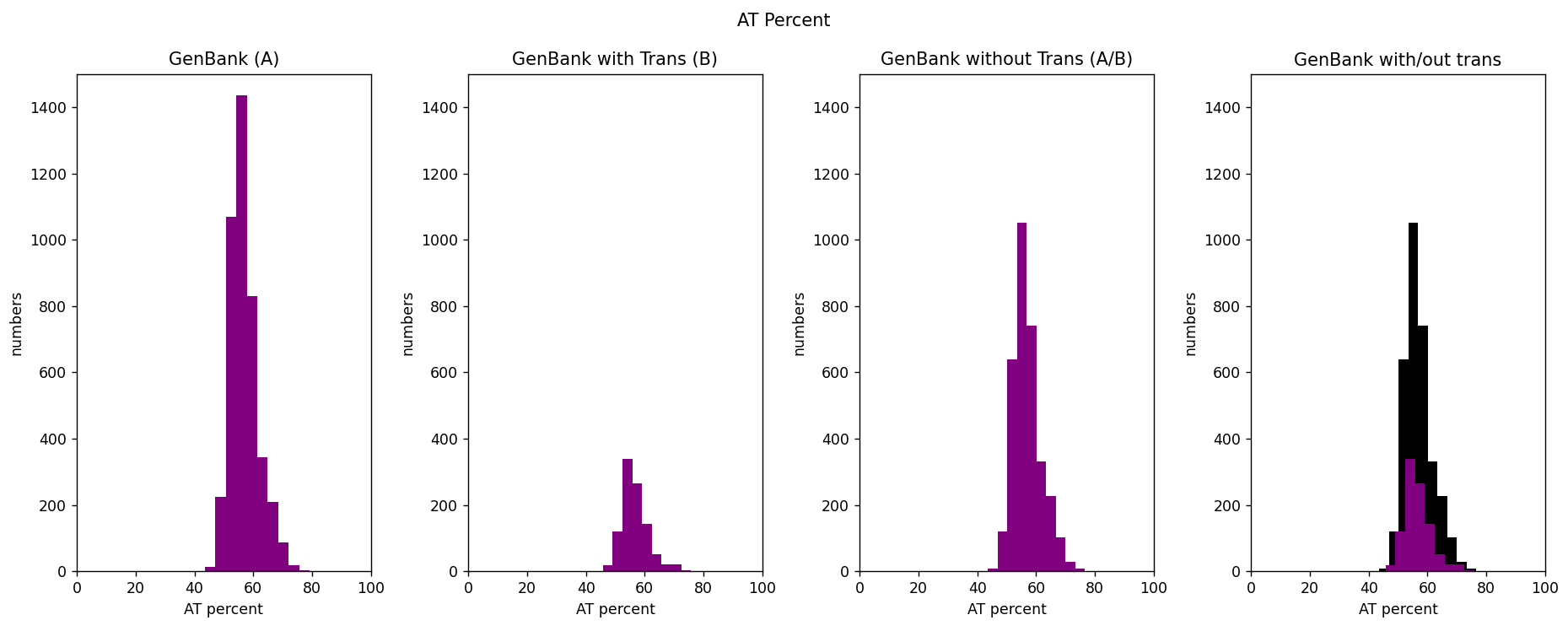
1. **התפלגות AT:**

**Chart, histogram

Description automatically generated**התפלגות אחוזי AT ברצפי CDS שנמצאו בחיתוך בין uniport לבין GenBank, שיש להם לפחות אזור טרנסממברנלי אחד:

**סטטיסטיקות עבור קבוצת גנים:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Standard deviation** | **Maximum** | **Minimum** | **Average** | **Name** |
| 4.74 | 79.17 | 43.56 | 56.88 | GenBank AT percent |
| 4.59 | 79.17 | 45.74 | 56.71 | GenBank with Transmembrane intersection AT percent |
| 4.78 | 76.62 | 43.56 | 56.93 | GenBank without Transmembrane intersection AT percent |

**התפלגות AT בכל הקבוצות:**

**חלק ג' – אנליזה מנקודת מבט אבולוציונית - וירוסים**