קורס ביולוגיה חישובית - עבודת סיום תשפ"ב

מאת: יעל ליברמן (318376449), טהר צביטמן (318192838) ואילת ג'יבלי (208691675)

חלק א': איסוף ועיבוד מידע אודות גנום החיידק בצילוס סבטיליס

1. הכרת וספירת האלמנטים בגנום:

דיווח על מספר האלמנטים מכל אזור בגנום החיידק:

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

2. אפיון אורכי הגנים:

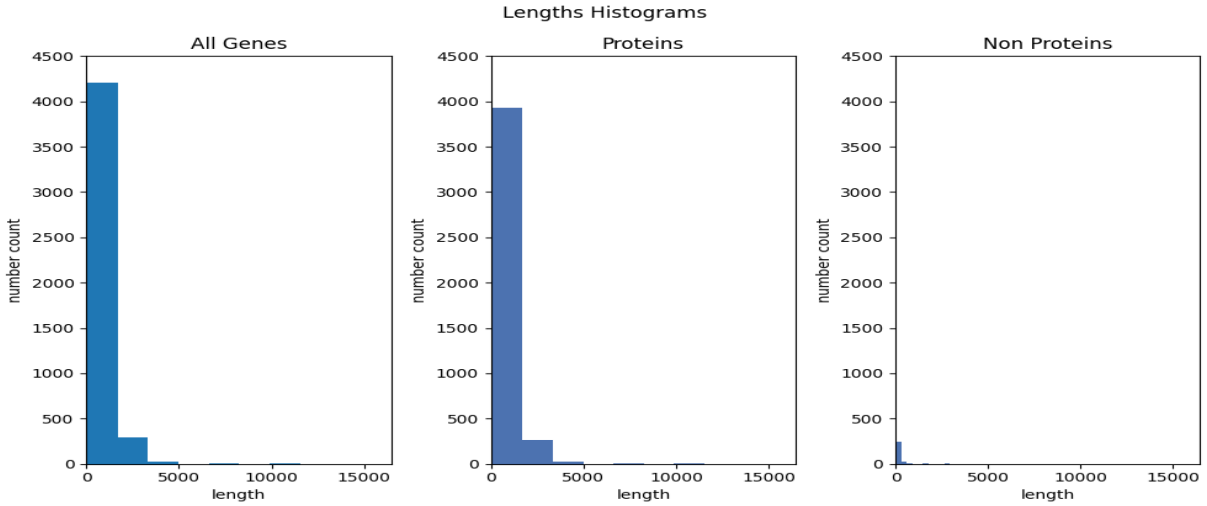
א. עבור כל גן חישבנו את אורכו על ידי end - start (לאחר בדיקה שהsequence לא כולל את הend). הוספנו את עמודת האורכים לתוך הdataframe של קובץ הGenBank.

ב. חילקנו את הdataframe לשניים: השורות של הגנים שמקודדים לחלבון וכל השאר.

ג. הגנים אשר מקודדים לחלבון: ממוצע: 874.57, מינימום: 63, מקסימום: 16467.

שאר הגנים: ממוצע: 324.12, מינימום: 33, מקסימום: 2928.

ד. ההתפלגות עבור כל קבוצה:



המסקנות שלנו מהגרף: ניתן לראות כי הגנים המקודדים לחלבון ארוכים יותר באופן משמעותי מהגנים שאינם מקודדים לחלבון. זה אכן תואם לציפיות מאחר וגנים המקודדים לחלבון צריכים להיות בעלי קידוד אחיד שיתורגם בהמשך לחומצות אמינו שמתקפלות לחלבון, ולכן הם בד"כ יותר ארוכים. לעומת זאת, גנים שאינם מקודדים לחלבון יהיו בד"כ רצפי בקרה, שמקשרים בין חלבונים או מובילים לחלבונים אחרים, ולכן גם האורך שלהם יהיה קצר יותר.

3. חישוב אחוז GC בגנים:

א. ממוצע בגנום החיידק: 43.51%.

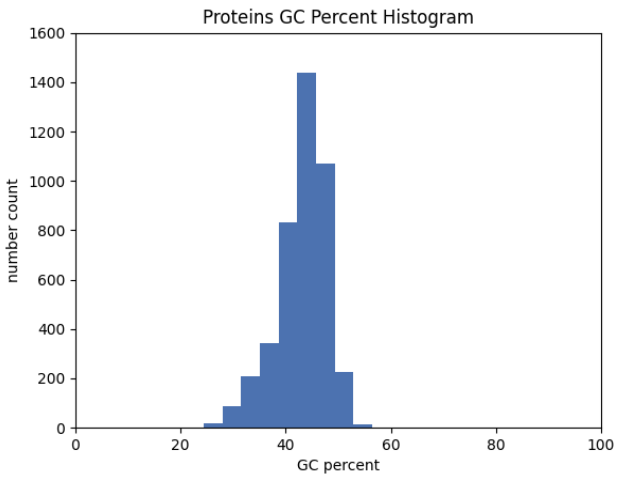
ב. עבור כל גן חישבנו את GC%. הוספנו את עמודת הGC% לתוך הdataframe של קובץ הGenBank.

ממוצע בגנים המקודדים לחלבון: 43.12%.

ג. האם התוצאה תואמת לציפיות? אזורים עם אחוז GC גבוהה הם לרוב אזורי בקרה. התכונה של הנוקלאוטידים האלו היא שהם יוצרים קשר כימי חזק וזוהי תכונה חשובה לרצפי בקרה שמטרתם להיצמד

לאתרים וליצור קשרים. ניתן לראות כי אחוז הGC קטן מחצי בגנום כולו, ונמוך יותר בגנים המקודדים לחלבון. זה אכן הגיוני מאחר ורוב הגנום מורכב מגנים המקודדים לחלבון שבהם אחוז הGC לא גבוהה במיוחד.

ד. התפלגות GC% בחלבונים:



ה. חמשת הגנים שבהם יש את אחוז הGC הגבוה ביותר:

מקום	שם הגן (locus tag)	התחלה	סוף	סטרנד	סוג	אורך	GC%
.1	BSU tRNA 86	3194454	3194527	-1	tRNA	73	67.12
.2	BSU tRNA 6	11551	11627	1	tRNA	76	65.79
.3	BSU tRNA 9	32019	32095	1	tRNA	76	65.79
.4	BSU tRNA 20	96145	96221	1	tRNA	76	65.79
.5	BSU tRNA 28	166252	166328	1	tRNA	76	65.79

חמשת הגנים שבהם יש את אחוז הGC הנמוך ביותר:

מקום	שם הגן (locus tag)	התחלה	סוף	סטרנד	סוג	אורך	GC%
.1	BSU 26360	2699509	2699677	-1	CDS	168	20.83
.2	BSU 17700	1904994	1905195	-1	CDS	201	23.38
.3	BSU 17670	1901116	1901377	-1	CDS	261	24.52
.4	BSU 39290	4036343	4036787	-1	CDS	444	25.45
.5	BSU 40210	4132337	4132736	-1	CDS	399	25.81

4. בדיקת עקביות בקובץ הדאטה:

כדי לבדוק את עקביות קובץ הדאטה, הגדרנו את השיקולים הבאים עבור גנים המתורגם לחלבון:

1. בדיקה האם רצף הגן מתחיל בstart codon.
2. בדיקה האם אורך רצף הגן מתחלק ב 3.
3. בדיקה האם התרגום הנתון נכון.
4. בדיקה האם רצף הגן מסתיים בstop codon ולא קיים stop codon באמצע הרצף (לפי מסגרת הקריאה).

הגנים שעבורם נמצאה סתירה:

מס'	שם הגן (locus tag)	התחלה	סוף	סטרנד	סוג	טבלת תרגום	סיבה
.1	BSU 20040	2159980	2161778	-1	CDS	11	Sequence length 1798 is not a multiple of three
.2	BSU 20060	2162107	2165614	-1	CDS	11	Extra in frame stop codon found
.3	BSU 35290	3627138	3628240	-1	CDS	11	Sequence length 1102 is not a multiple of three

חלק ב': אנליזת חלבונים בעזרת אתר ה-UniProt

1. הצלבה בין החלבונים בקובץ ה-GenBank לבין החלבונים מקובץ ה-UniProt:

לאחר חקירה של קובץ GenBank, בחרנו את השדה "locus_tag" כמזהה של הגן, מאחר והוא מופיע בכל הרשומות וייחודי לכל רשומה. לאחר מכן, הוספנו את העמודה המתאימה לקובץ הUniProt - שם העמודה באתר הוא "Gene names (ordered locus)". על סמך מזהה זה, הצלבנו בין החלבונים הנמצאים בין 2 הקבצים. היה צורך בעיבוד מקדים, מאחר ובקובץ הGenBank היה מקף תחתון בערך המזהה, ובקובץ הUniProt היו המון רשומות שלא היה להם מזהה או רשומות שבהם הופיעו כמה מזהים לאותו חלבון.

סיכום ההפרשים:

In Genbank but not in Uniport:

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

Missing:80, from total: 4237

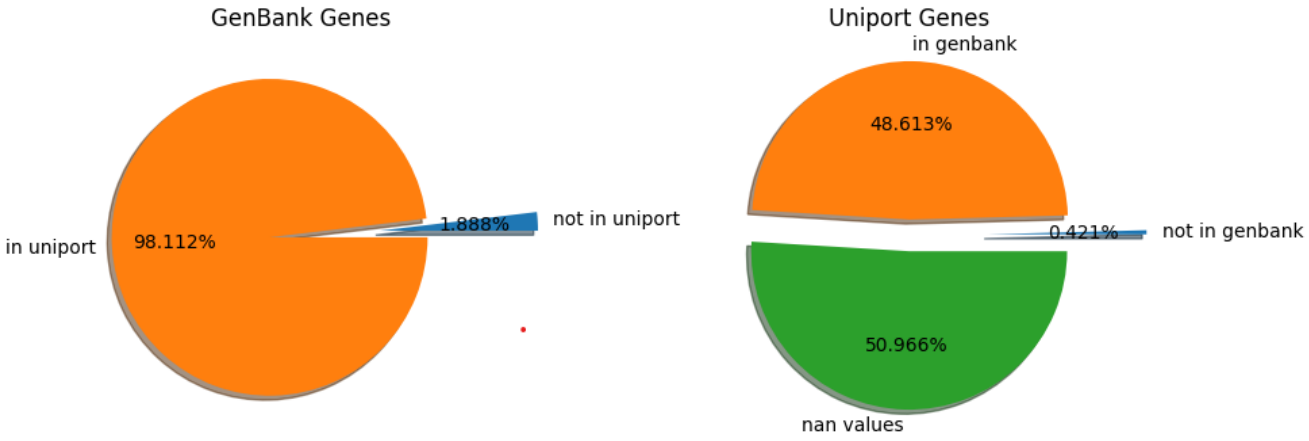
In Uniport but not in Genbank (after removing nan):

['BSU16840', 'BSU33220', 'BSU18930', 'BSU16890', 'BSU16900', 'BSU06810', 'BSU13790', 'BSU20030', 'BSU26080', 'BSU25760', 'BSU26390', 'BSU03570', 'BSU06050', 'BSU34410', 'BSU35610', 'BSU26040', 'BSU02180', 'BSU34420', 'BSU35230', 'BSU16640', 'BSU35150', 'BSU13799', 'BSU18110', 'BSU18940', 'BSU39030', 'BSU23290', 'BSU12670', 'BSU07740',

'BSU22660', 'BSU07180', 'BSU40020', 'BSU11381', 'BSU11382', 'BSU28480', 'BSU01840', 'BSU35690']

Missing: 36, Nan values: 4353, from total: 8541

בצורה ויזואלית:

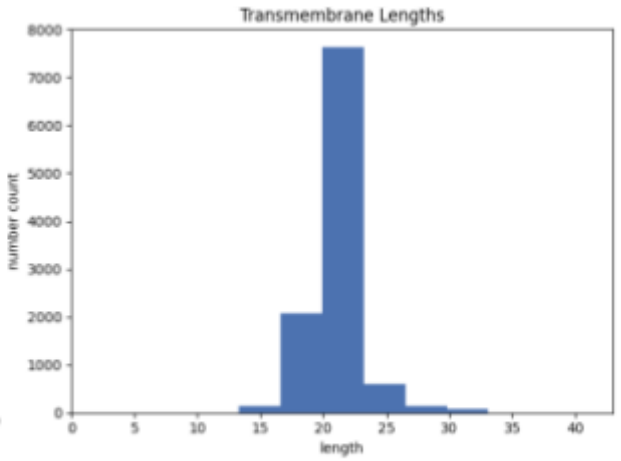


מאיפה נובעים ההבדלים? ההבדלים נובעים מהשוני בין סוגי הקבצים. בקבצי genBank מוצגים כל הרצפים בגן שמקודדים לחלבון. לעומת זאת uniprot מציג רק גנים שמקודדים לחלבונים פונקציונליים, זאת אומרת- גנים שנחקרו ויש עליהם מידע. לכן אם מופיע קודון אתחול של גן, הגן שמתחיל אחריו יופיע בקובץ genbank ללא קשר למידע הנוסף שיש עליו, לעומת זאת ב uniprot הוא יופיע רק אם הוא נחקר וידוע לנו מה הפונקציונליות שלו.

2. אפיון הרצפים הטרנסממברנליים:

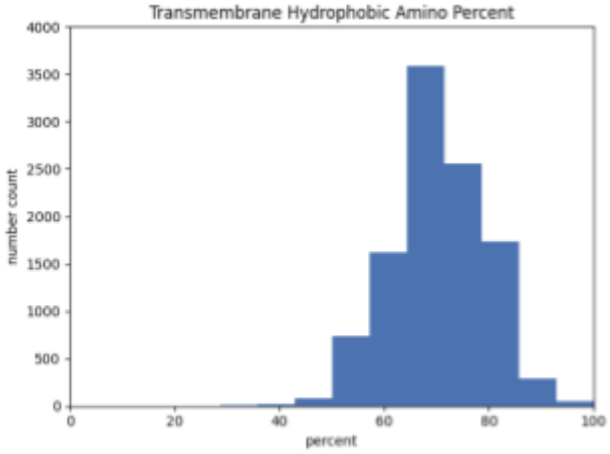
על מנת לבודד את הרצפים הטרנסממברנליים בנינו dataframe נפרד עבורם. לכל תת רצף שורה נפרדת. לאחר מכן הוספנו לdataframe הזה את העמודה של האורכים.

א. אפיון אורכי הרצפים - ההתפלגות:



ממוצע: 20.40, מינימלי: 10, מקסימלי: 43.

ב. אפיון אחוז חומצות האמינו ההידרופוביות - ההתפלגות:

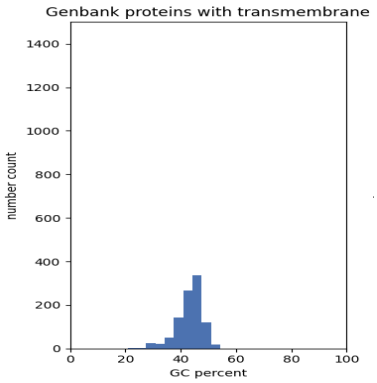


ממוצע: 70.09, מינימלי: 28.57, מקסימלי: 100.

האם התוצאה של אחוז חומצות האמינו ההידרופוביות ברצפי הטרנסממברנליים תואמת לציפיות? רצפים טרנסממברנליים אלו רציפים המקודדים לחומצות אמינו הנמצאות בתוך הממברנה (קרום התא). הממברנה היא הידרופובית מאחר והיא עשויה מליפידים בעלי ראש הידרופילי, ושני זנבות של חומצות שומן הידרופוביות. המבנה הזה מאפשר לה להוות חיץ הידרופובי בין שני תמיסות מימיות- תמיסה חיצונית לתא ותמיסה פנימית לתא. בגלל המבנה הזה, על מנת לשמור על יציבות החלבון, הרצפים הטרנסממברנליים צריכים לקודד לחומצות אמינו הידרופוביות כדי להתאים למבנה הממברנה. מכאן ניתן להסיק שהתוצאות אכן תואמות לציפיות מאחר והתוצאות מראות אחוזי חומצות אמינו הידרופיביות גבוהה ברצפים אלו.

אפיון התפלגות אחוז GC בקבוצות השונות:

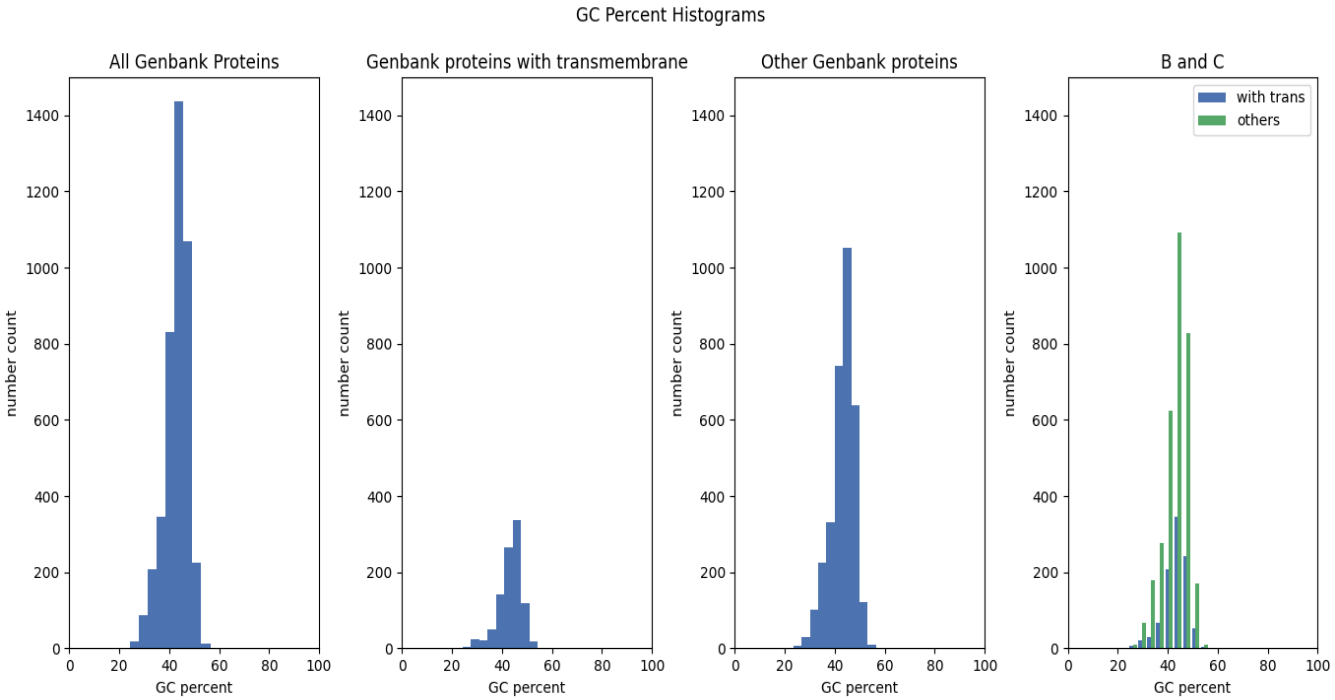
א. התפלגות GC% ברצפי הגנים שהם CDS, שנמצאים בחיתוך בין הUniProt לבין הGenBank ויש להם לפחות אזור טרנסממברנלי אחד:



ב. טבלה המסכמת את הסטטיקות עבור קבוצות הגנים השונות:

קבוצה	ממוצע	חציון	מינימום	מקסימום
A: All proteins	43.12	43.89	20.83	56.44
B: Proteins with transmembrane	43.28	44.08	20.83	54.26
C: Proteins without transmembrane	43.07	43.87	23.38	56.44

התפלגות ה GC% בכל הקבוצות על אותו הגרף עם ארבעה חלקים:



חלק ג': אנליזה מנקודת מבט אבולוציונית - וירוסים

1. חישוב העמדות הסינונימיות לכל קודון עבור הקוד הגנטי של וירוס הקורונה:

הקוד הגנטי המתאים לוירוס הקורונה הוא הקוד מטבלה 1, מכיוון שבקובץ לא מופיע השדה המציין את מספר הטבלה מדובר בקוד הגנטי הסטנדרטי מטבלה 1.

דיווח על מספר עמדות סינונימיות עבור כל קודון (לפי השיטה של (Nei-Gojobori (NG86 :

{'TTT': 0.33, 'TCT': 1.0, 'TAT': 0.43, 'TGT': 0.38, 'TTC': 0.33, 'TCC': 1.0, 'TAC': 0.43, 'TGC': 0.38, 'TTA': 0.86, 'TCA': 1.29, 'TAA': 0.0, 'TGA': 0.0, 'TTG': 0.75, 'TCG': 1.12, 'TAG': 0.0, 'TGG': 0.0, 'CTT': 1.0, 'CCT': 1.0, 'CAT': 0.33, 'CGT': 1.0, 'CTC': 1.0, 'CCC': 1.0, 'CAC': 0.33, 'CGC': 1.0, 'CTA': 1.33, 'CCA': 1.0, 'CAA': 0.38, 'CGA': 1.5, 'CTG': 1.33, 'CCG': 1.0, 'CAG': 0.38, 'CGG': 1.33, 'ATT': 0.67, 'ACT': 1.0, 'AAT': 0.33, 'AGT': 0.33, 'ATC': 0.67, 'ACC': 1.0, 'AAC': 0.33, 'AGC': 0.33, 'ATA': 0.67, 'ACA': 1.0, 'AAA': 0.38, 'AGA': 0.75, 'ATG': 0.0, 'ACG': 1.0, 'AAG': 0.38, 'AGG': 0.67, 'GTT': 1.0, 'GCT': 1.0, 'GAT': 0.33, 'GGT': 1.0, 'GTC': 1.0, 'GCC': 1.0, 'GAC': 0.33, 'GGC': 1.0, 'GTA': 1.0, 'GCA': 1.0, 'GAA': 0.38, 'GGA': 1.12, 'GTG': 1.0, 'GCG': 1.0, 'GAG': 0.38, 'GGG': 1.0}

2. השוואה בין וירוס הקורונה שבודד ביולי 2020 לבין וירוס הקורונה שבודד בינואר 2022:

א. בתור מזהי הגנים לקחנו את שדה "gene" המופיע בכולם (בניגוד לlocus_tag שבקובץ זה אינו מופיע בכולם). לאחר ההשוואה ראינו שכל 11 מזהי הגנים משני הקבצים משותפים.

שמות הגנים המשותפים:

['ORF1ab', 'S', 'ORF3a', 'E', 'M', 'ORF6', 'ORF7a', 'ORF7b', 'ORF8', 'N', 'ORF10']

ב. חישוב dnds - אופן החישוב:

1. ביצוע עימוד לפי התרגום לחלבונים, מיון העימודים האפשריים ולקיחת האחרון מביניהם. את העימוד נבצע על סמך מטריצת "BLOSUM62". נשים לב שרק בעימוד האחרון ברשימה הממוינת לא נכנס גאפ במקום של התחלפות חלבונים (כי בצורה זאת הגאפ "מעלים" מוטציה אסינונמית) אלא יש גאפים רק במקרה ובאמת חסרים חלבונים.

דוגמה הממחישה למה רק העימוד האחרון הוא המתאים לחישוב המדד:

```
aligner = Align.PairwiseAligner()
alignments = aligner.align("MKIILFL", "MKIAILFG")
for alignment in sorted(alignments):
    print("Score = %.1f:" % alignment.score)
    print(alignment)
```

```
Score = 6.0:
MKI-ILF-L
|||-|||--
MKIAILFG-

Score = 6.0:
MKI-ILFL-
|||-|||--
MKIAILF-G

Score = 6.0:
MKI-ILFL
|||-|||.
MKIAILFG
```

2. תרגום חזרה לdna לפי הreading frame והdna המקורי, כאשר גאפ נתרגם חזרה כשלושה גאפים.
3. חישוב מדדי dn,ds בעזרת הפונקציה cal_dn_ds של ספרית Bio.

התוצאות עבור חמישה גנים מהגנים המשותפים:

מקום	שם הגן	תפקיד	התחלה	סוף	dn	ds	סוג סלקציה
.1	S	CDS	21562	25384	0.000338	0.001194	Negative
.2	M	CDS	26522	27191	0.0	0.0	Natural
.3	ORF6	CDS	27201	27387	0.0	0.0	Natural
.4	ORF7a	CDS	27393	27759	0.003614	0.0	Positive
.5	N	CDS	28273	29533	0.0	0.0	Natural

הסבר קצר על תיכון הקוד:

מאחר ויש קריאה של קובץ GenBank בכל חלק בפרויקט, ועדכון תמידי של תכונות נוספות על סמך המידע בקובץ, יצרנו מחלקה המייצגת אובייקט של GenBank עם התכונות הבאות:

- dataframe של pandas השומר לכל פירוט של גן את העמודות (לא שומר את הרשומות שהופיעו עם type 'gene' מאחר ומיד אחריהם מופיעה רשומה עם פירוט על אותו הגן) :
  'id', 'start', 'end', 'strand', 'type', 'table', 'translation', 'codon_start'
  (ואפשר להוסיף לו עמודות לפי הצורך)
- משתנה sequence השומר את רצף הגנום.
- משתנה genes השומר את שמות הגנים שהופיעו עם type 'gene'.
- משתנה cds ששומר את השורות הרלוונטיות מהdataframe של הגנים שהופיעו עם type 'cds'.
- משתנה not_cds ששומר את השורות הרלוונטיות מהdataframe של הגנים שלא הופיעו עם type 'cds'.

בנוסף, פונקציות גלובליות השימושיות לכל החלקים של הפרויקט, שמרנו במודול נפרד בשם 'generic_functions'.