בס"ד

המרכז האקדמי לב

קמפוס טל

החוג לביואינפורמטיקה

פרוייקט גמר

שם הפרויקט

מציאת מערכות טוקסין-אנטי טוקסין מסוג II בגנומים חידקיים

מגישות:

דינה כהן ונעה בבצ'יק

מנחה: ד"ר אוריה ורדי

תאריך: י"ט אב תשפ"ג

מוצאי שבת פרשת עקב

בס"ד

תקציר:

מטרתנו בפרויקט זה הייתה יצירת כלי שמאפשר מציאת חלבוני TA מסוג II בגנום חיידקי ולאחר מכן שימוש בו לצורך אנליזה של גנומים חיידקיים ולמידה על הTA שלהם. על מנת לייצרו פיתחנו אלגוריתם שמשתמש בתוכנת prokka ,שאפשרה לנו למצוא את החלבונים הקיימים בגנום של החיידק. לאחר מכן ביצענו הרצת blastp בבת אחת על כל החלבונים שנמצאו על ידי prokka לעומת מאגר חלבוני TA מסוג II כדי למצוא את חלבוני הTA ולבודדם משאר החלבונים שקיימים בחיידק. אפשרנו למשתמש לבחור cutoff לevalue של התוצאות , על מנת לסנן תוצאות שלא מספיק קרובות לשאילתות של החלבונים שprokka מצאה וכנראה לא סביר שהם אכן חלבונים שקיימים בחיידק. ביצענו סינונים נוספים כגון הורדת תוצאות שהתאימו לאותה שאילתה, והשארת התוצאה שהכי הייתה דומה לשאילתה. לאחר הסינונים, קיבלנו את רשימת הטוקסינים והאנטי טוקסינים שחשודה להיות בגנום החיידקי. בהמשך הפרויקט, הורדנו גנומים של זנים של Pseudomonas aeruginosa ובאמצעות שימוש בכלי , בדקנו אילו TA נמצאים בגנומים ובנינו גרף של סך החלבונים הכללי שמצאנו לכל זן. בהמשך, כתבנו קוד שמבצע אנליזות שונות על קבצי הכלי. בנינו גרפים המתארים את חלבוני ה TA המשותפים לגנומים של הזנים עליהם ביצענו את האנליזות, ואת חלבוני ה TAיחודיים לכל זן. בנוסף, בדקנו את המרחקים בין מיקומי ה TA שנמצאו על מנת לבדוק את הימצאותם של TA island בגנומים הנבדקים ובנינו גרף המתאר את מספר TA island שנמצאו בכל זן. לסיום, אנו מציעות לקבוצות מחקר שמתעסקות בTA מסוג II להשתמש באלגוריתם שלנו על מנת להקל על עצמן במציאת החלבונים ולאפשר אנליזה נוחה וטובה שלהם. וכן אנו מציעות לבדוק את נכונות האלגוריתם שלנו על ידי ניסויים רטובים שיאששו את הימצאות החלבונים שמצא בגנום של הזנים שייבדקו.

הקדמה:

[1]מערכות טוקסין- אנטי טוקסין הן מערכות המורכבות "מטוקסין" -רעלן יציב ומ"אנטי טוקסין"- רעלן לא יציב. מערכות אלו שכיחות בגנומים של חיידקים וארכיאות .הטוקסין הוא בדרך כלל חלבון בעוד שהאנטיטוקסין יכול להיות חלבון או RNA.

מנגנון הפעולה של המערכות הללו הוא כזה -האנטי-טוקסין מונע מהטוקסין שלו לגרום לרעילות. במצבים מסוימים , כמו אובדן פלסמיד , האנטי טוקסין כלה או שכמותו קטנה, ובכך הוא משחרר את הטוקסין לעשות את פעילותו.

כאשר מערכות אלו נמצאות בפלסמידים -אז ניתן להבטיח שרק תאי הבת שירשו את הפלסמיד ישרדו לאחר חלוקת התא. אם הפלסמיד לא קיים בתא בת, האנטי-טוקסין הלא יציב מתפורר והחלבון הרעיל היציב הורג את התא החדש.

מערכות טוקסין-אנטי-טוקסין מסווגות בדרך כלל לפי האופן שבו האנטי-טוקסין מנטרל את הטוקסין. במערכת טוקסין-אנטי-טוקסין מסוג I, mRNA המקודד לטוקסין מעוכב על ידי קשירה של RNA קטן ולא מקודד של אנטי טוקסין . במערכת מסוג II , הטוקסין מעוכב לאחר תרגום על ידי קשירה של חלבון אנטי טוקסין. מערכות טוקסין-אנטי-טוקסין מסוג III מורכבות

[1] ויקיפדיה

מ-RNA קטן הנקשר ישירות לחלבון של הטוקסין ומעכב את פעילותו. ישנם גם סוגים IV-VI, שהם פחות נפוצים.

הפרויקט שלנו התמקד במערכת מסוג II. מטרתו הינה לבנות אלגוריתם שיאפשר מציאה של חלבוני TA מהסוג הנ"ל בתוך גנום חיידקי שיתקבל מהמשתמש (הקלט הינו רצף נוקלאוטידי של החיידק).

השיטה המקובלת עד היום לצורך זה הייתה להשתמש בprokka -תוכנה המאפשרת מציאת חלבונים בגנומים חיידקיים. וברירת החלבונים השייכים דווקא למערכות אלו על ידי הרצת blastp על התוצאות הללו למול מאגר חלבוני TA שידועים במחקר.

החיסרון בשיטה זו הינו שהוא לא מאפשר סינון של התוצאות עם התייחסות לבעיה הביולוגית של המערכות הללו – והיא שבהרצת הblastp לדוגמא , מתקבלות תוצאות שיש להן אמנם מובהקות גבוהה יחסית, אבל הן לא רלוונטיות, משום שהתקבל רק טוקסין בתוצאות אבל לא האנטי שלו.

כמו כן, יכולות להתקבל מספר אפשרויות של טוקסינים או אנטי-טוקסינים לחלבון שprokka הביא, ועל כן יש צורך בדרך לברור את התוצאות באופן שיאפשר את בחירת התוצאה שהכי תתאים מבין שאר התוצאות לאותו חלבון.

הפתרון שלנו מתמודד עם הבעיות הללו, על ידי סינונים שונים שהוספנו- כגון הורדת FP , בחירת התוצאה המתאימה ביותר לחלבון prokka כלשהו על ידי בחירה בחלבון שיש לו את evalue הכי נמוך (ואם יש כמה חלבונים עם אותו ציון evalue התייחסנו לפרמטרים נוספים על מנת לקבוע איזה מהתוצאות הכי טובה ,ופרטנו עליהם בשיטות.)

על כן מטרת המחקר שלנו הינה בניית אלגוריתם שיאפשר מציאה של חלבוני TA מהסוג הנ"ל בתוך גנום חיידקי שיתקבל מהמשתמש. לאחר מכן, הוצבה מטרה נוספת והיא שימוש באלגוריתם זה לצורך אנליזה של גנומים חיידקיים מזנים שונים של pseudomonas aeruginosa, והסקת מסקנות לגבי מערכות הTA שהם מכילים- האם הם מכילים TA ייחודיים וכן בדיקה האם קיימים בתוכם TA Islands.

שיטות:

ראשית, על מנת להשיג את המטרה הראשונה שהוצגה – בנינו את האלגוריתם. האלגוריתם משתמש בprokka שהיא תוכנה שמקבלת כקלט גנום חיידקי ומחזירה כפלט קבצים שונים המתארים את החלבונים שקיימים בגנום זה.

לאחר שprokka מצאה את החלבונים, הרצנו blastp בבת אחת על כולם מול מאגר של TA שהורדנו מהאתר המצורף. המאגר הינו קובץ fasta גדול של נוקלאוטידים , שמכיל את הטוקסינים והאנטי טוקסינים מסוג II שקיימים באתר[2]. ניתן לצוות טוקסין לאנטי שלו לפי מספר זהות משותף שלהם- כל התחלה של שורה שמתארת את הרצף החלבוני של טוקסין או אנטי בקובץ מתחילה בשם המאגר, pipe ולאחר מכן סימן אם זה טוקסין או אנטי טוקסין

[2] All the in silico predicted and experimentally validated Type II TA Nucleotide

בס"ד

ומספר הזהות של טוקסין או אנטי זה. דוגמא: אנטי טוקסין- TADB|AT1 והטוקסין שלו- TADB|T1.

נתנו אופציה למשתמש להגדיר את סף הevalue של תוצאות הblastp , שמעליו לא נקבל תוצאות מהתוכנה.

לאחר שהתקבלו תוצאות הblastp , ציירנו גרפים המראים את ההתפלגות של התוצאות לפי הcoverage וכן לפי הprecent identity שלהם. על פיהם , המשתמש יבחר בcutoff הולם על מנת להסיר תוצאות עם מדדים נמוכים מידי. (נתנו זאת כאופציה על מנת שלא נחליט על סף כללי מראש ויותר מידי נחמיר את התוצאות שיצאו לנו.)

כמו כן, עבור שאילתות מסוימות עלולות להתקבל כמה תוצאות מהגרף- אז השארנו את זו עם הevalue המינימלי. ואם יצאו כמה תוצאות עם evalue מינימלי, אז השארנו את זו עם הcoverage המקסימלי. ואם גם בזה היה שוויון בין כמה תוצאות, נבחר מביניהן את זו עם ה precent identity המקסימלי. ואם אפילו בזה יש שוויון תיבחר התוצאה שהופיעה ראשונה.

לאחר מכן, אם אותו אנטי/טוקסין היה מתאים לכמה שאילתות שונות, השארנו את התוצאה עם המדדים הטובים ביותר(באופן דומה למה שתואר בסינון הקודם, מבחינת קדימות המדדים).

לאחר סינון זה, הורדנו תוצאות שהן FP. כלומר, לכל טוקסין צריך להיות אנטי טוקסין ,(וכן הפוך) ולכן הורדנו חלבונים שלא נמצא להם זוג.

לסיום, חילקנו את הקבצים שמרכזים את התוצאות הסופיות כך שבאחד יש את הטוקסינים, ובשני את האנטי טוקסינים.

לאחר שהתקבלו התוצאות מהכלי, התחלנו בשלב האנליזה. בחרנו חמישה זנים של Pseudomonas aeruginosa והורדנו את הגנומים שלהם מהncbi . (חמשת הזנים הינם:PAO1 PAC1 NCTC10332 F30658 DSM50071).

לאחר מכן, התחלנו באנליזת הנתונים.

על מנת למצוא את מספר הTA הייחודיים בכל זן, יצרנו רשימת TA משותפים בכל הזנים ואת מספר הTA שנמצאו בכל זן מרשימת הTA המשותפים והחסרנו בין המשותפים שלו לרשימה הכוללת של הTA שנמצאה לאותו זן כך שקיבלנו רק את הייחודיים לו.

לאחר מכן, חיפשנו את מספר איי ה TA בכל זן (TA island) . הכוונה היא באיי TA היא ל"צברים" של TA בגנום. כלומר קבוצה של TA שנמצאים זה ליד זה בגנום.

כדי למצוא קבוצות אלו, חיפשנו בכל זן בנפרד את המרחק בין כל TA שיש בו, לפי המיקומים שבהם חלבוני הTA שלו נמצאים בגנום שלו. הסרנו את חלבוני הTA שהמרחק ביניהם לבין ה TA שאחריהם היה גדול משמעותית ביחס למרחקים שנמצאו בין ה TA שבשאר הזנים וקיבלנו את קבוצות הTA החשודות להיות islands .

תוצאות:

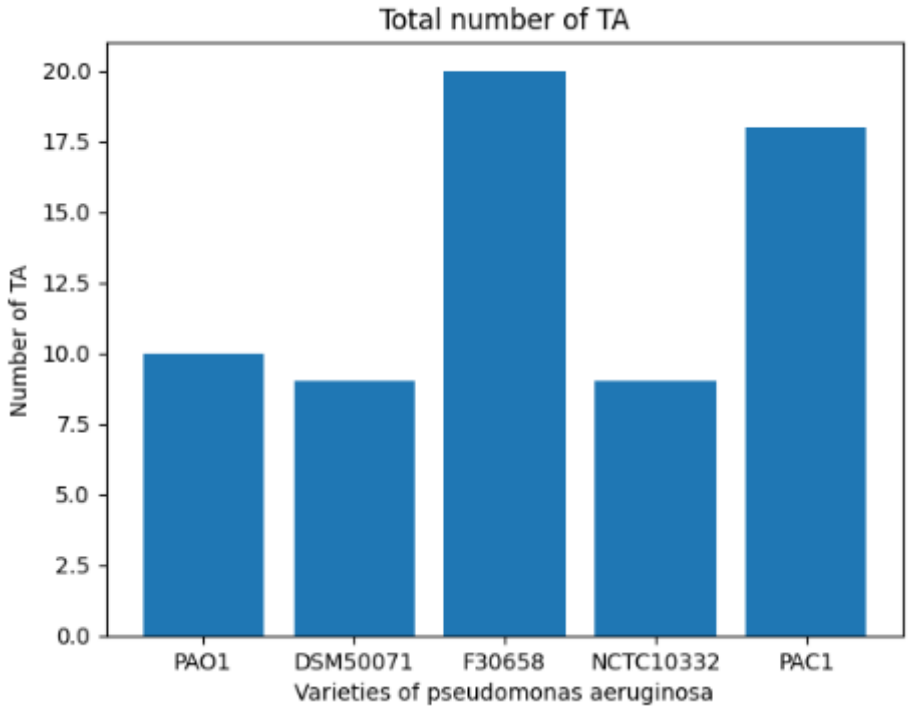
בס"ד

ראשית, נציין כי קבענו בשביל האנליזה שלנו שהתוצאות שיתקבלו מblastp יהיו בעלות מובהקות סטטיסטית של evalue השווה ל0.0001. וסיננו אותן כך שישארו תוצאות שיש להן coverage ו precent identity של למעלה מ45 אחוזים.

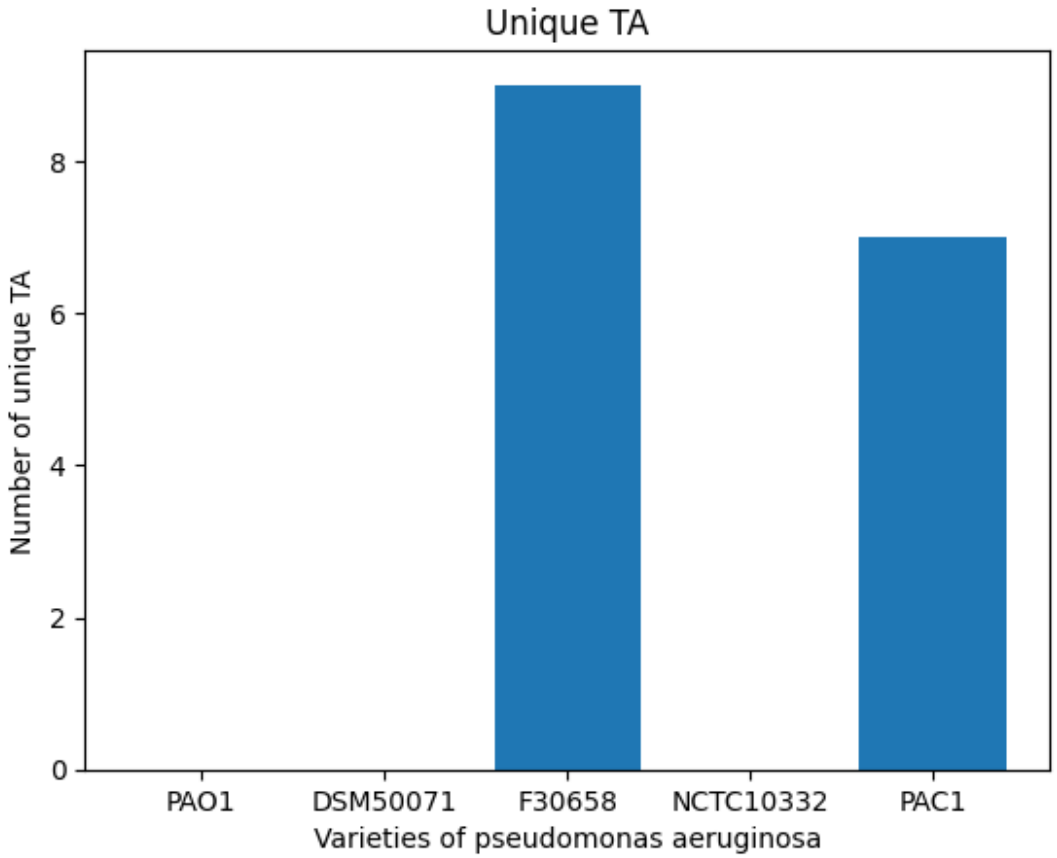
על כן, עם ההגבלות הנ"ל ,הרצנו את הכלי שבנינו על חמישה זני האורגניזם - pseudomonas aeruginosa:

- PAO1
- DSM50071
- F30658
- NCTC10332
- PAC1

יצרנו גרף המתאר את כל הTA שקיימים בזנים הנ"ל:



וכן, יצרנו גרף של מספר החלבונים היחודיים שנמצאו בכל זן:



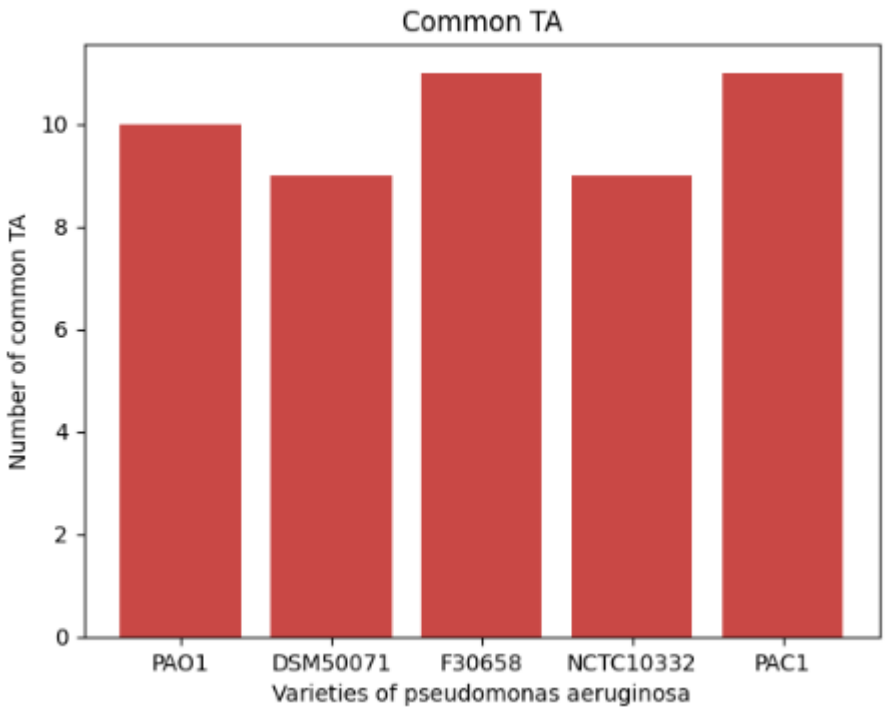
ניתן לראות שרק בשני זנים נמצאו חלבוני ייחודיים – שנמצאו רק אצלם ולא אצל הארבעה הזנים האחרים. בזן F30658 נמצאו תשעה חלבוני TA ייחודיים, ובזן PAC1 נמצאו 7 חלבונים ייחודיים.

לאחר מכן יצרנו רשימת חלבוני טוקסין שנמצאו ביותר מזן אחד:

```
'T6299', 'T6308', 'T6307', 'T4036', 'T6306', 'T6305', 'T6302',
'T6300', 'T6314', 'T6304', 'T3407'
```

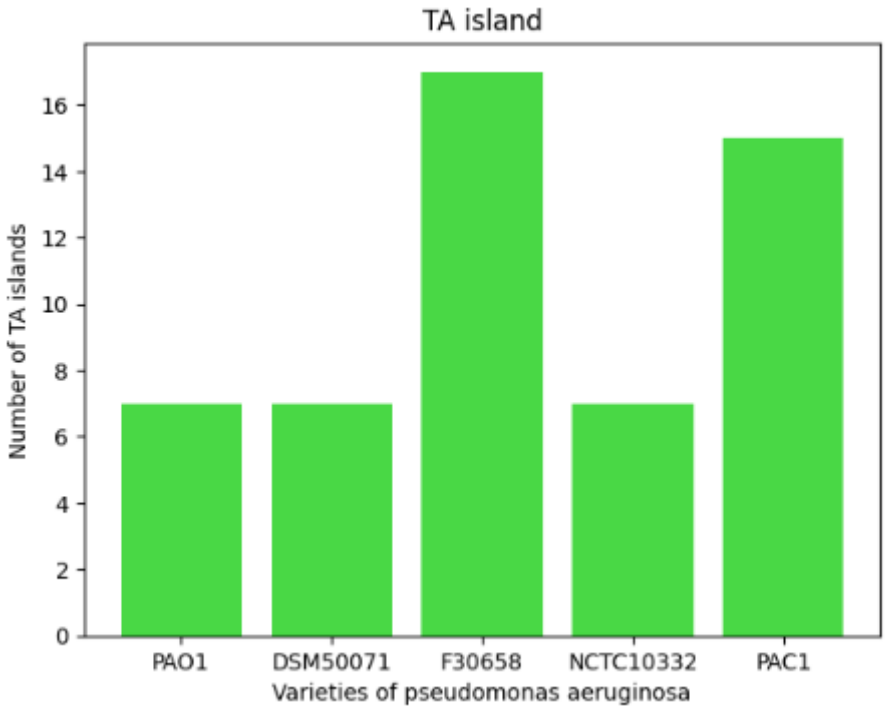
ויצרנו גרף המתאר את מספר החלבונים שנמצאו בכל זן מרשימת החלבונים המשותפים:

בס"ד



ניתן לראות שרוב חלבוני הTA שנמצאו בכל זן היו חלבונים שהופיעו גם בשאר הזנים.

בנוסף לאחר מציאת ערך מינימלי של בסיסים שבו ניתן למצוא איי טוקסינים (הערך הנבחר הוא מיליון בסיסים בין חלבון לחלבון), יצרנו גרף של מספר האיים שנמצאו בכל זן:



ניתן לראות שבשלושה זנים היו 7 איים, לעומת כ15 איים בשני הזנים האחרים.

ועל כן, בכך הצלחנו לענות ולמלא את מטרת המחקר שהצבנו.

בס"ד

דיון:

חשיבותן של תוצאות הפרויקט היא בכך שכעת ניתן לראות שחלבוני TA מאורגנים הרבה פעמים בצברים, ומבחינה אבולוציונית היה נראה כי הם משותפים בין זנים של אותו אורגניזם.

על מנת לאשש ההשערה הזו יש לבדוק זאת על יותר זנים ובקרב יותר אורגניזמים.

כמו כן, ניתן להשתמש בכלי של הפרויקט להמשך מחקר כללי על TA. כדי להמשיך בהסקת מסקנות נוספות על חלבוני הטוקסין והאנטי טוקסין ולייישמם במחקר, כדאי להוסיף אנליזות נוספות המותאמות למחקר הספציפי של משתמשי הכלי .ובנוסף, כדאי להוסיף גרפים ולשנות במידת הצורך את ערכי הסינונים, על מנת שניתן יהיה למצות ככל האפשר את התוצאות שהכלי מפיק .בנוסף לכל זאת, כיוון נוסף להמשכת הפרויקט הוא בדיקת נכונותו ודיוקו של הכלי על ידי ביצוע "ניסויים רטובים" אשר בודקים את הימצאותם של הטוקסינים והאנטי טוקסינים שהכלי מצא.

ביבליוגרפיה:

מאמר שעליו התבסס הידע הביולוגי שלנו על מערכות הAT.

(*Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology*, n.d.)

נספחים

נתיב לכלי שיצרנו בסביבת Linux:

"azhome/2023/babchick/mini/blast_all_TA_T.py/"

דוגמא להרצה:

```
python3.6 blast_all_TA_T.py /azhome/2023/babchick/mini/output/
/azhome/2023/babchick/mini_project_Dina_Noa/NC_013037.fasta
/azhome/ovardi/prokka/bin/prokka 0.0001
/azhome/2023/babchick/mini/db_FINAL.txt
```

הארגומנטים שהכלי מקבל(לפי הסדר):

1.שם התיקייה אליה המשתמש רוצה שהפלט יכנס

(בדוגמא: /azhome/2023/babchick/mini/output/)

2. הנתיב לגנום החיידקי

(בדוגמא: /azhome/2023/babchick/mini_project_Dina_Noa/NC_013037.fasta)

3. הנתיב להיכן שprokka מותקנת על השרת

בס"ד

(בדוגמא: /azhome/ovardi/prokka/bin/prokka)

4. הcutoff של הevalue של תוצאות הblastp

(בדוגמא: 0.0001)

5. הנתיב לdb של blastp

(בדוגמא: /azhome/2023/babchick/mini/db_FINAL.txt)

קישור למחברת colab בה בוצעו האנליזות:

https://colab.research.google.com/drive/1V3lf4G8ngBwnhwmlkAyFUoFmFBrZkaWi#scrollTo=lZb10Br59o7W&uniqifier=1

קישורים נוספים:

1. ויקיפדיה -מידע על מערכות TA
2. prokka- **קישור להדרכה של שימוש בprokka**
3. blastp – **קישור להדרכה לשימוש בblastp**
4. קישור לאתר ממנו הורדנו את הdb לצורך ה blastp
5. **קישור לאתר ממנו הורדנו את הגנומים החיידקיים**