בס"ד

קמפוס טל

החוג לביואינפורמטיקה

פרוייקט גמר

"זיהוי וריאנטים חדשים של וירוס הקורונה בזמן אמת"

מגישות: הדס סמואלס 323027441

נטע סולברג 322217225

מנחה: אוריה יעקב

אב תשפ"ב

בס"ד

תקציר:

בסוף שנת 2019 פרצה בעולם מגפה של נגיף הקורונה, שנקראה בשם COVID-19. צבירת מוטציות זו תכונה נפוצה בנגיפים רבים, ובעקבות התפשטותו הרחבה לאורך זמן של נגיף הקורונה, הוא צבר מוטציות רבות. לעיתים צבר מוטציות מסוים השפיע על תכונות פונקציונליות של הנגיף – שינוי רמת ההדבקה שלו, חומרת המחלה או אינטראקציות עם מערכת החיסון של המארח[1]. כך נוצרים וריאנטים שונים (וריאנט– זהו תת זן של אותו נגיף, השונה ממנו מבחינה גנטית). סט המוטציות בבסיסו של הווריאנט עלול להקנות לו יתרון משמעותי ביחס לנגיף המקורי, ולכן קריטי לעקוב אחר התפתחות הווריאנטים בזמן אמת. כמובן שקיימות מוטציות שמקנות חסרונות לנגיף ומחלישות אותו, ובמקרה זה אותה "וריאציה" של הנגיף תחדל מלהתקיים לאחר זמן קצר. במחקר זה שיתפנו פעולה עם צוות חיצוני והתמקדנו בווריאנטים שנוצרים ע"י "קפיצה" – צבירת הרבה מוטציות בבת אחת. באמצעות תכנות בשפת Python ניתחנו רצפי RNA של נגיפי קורונה שרוצפו מרחבי העולם. בתחילה למדנו על הרקע להיווצרותם של הווריאנטים, ועל בסיס זה כתבנו קוד שמוצא את הסטטיסטיקות להימצאות כל מוטציה בריצופים הנתונים. לאחר מכן התחלנו לכתוב את הקוד למציאת קפיצות של מוטציות ברצפים. כתבנו 3 גרסאות לקוד זה: 1. על בסיס וריאנט ידוע (תוספת של קובץ referernce) 2. התחשבות בסוג המוטציות שמגדירות קפיצה 3. ללא קובץ reference של וריאנט ידוע. לאחר שמצאנו את הקפיצות ברצפי הנגיפים, יכולנו למצוא ולהתריע על היווצרותם של וריאנטים חדשים.

בס"ד

מבוא:

הפרויקט שלנו עסק בווריאנטים של וירוס הקורונה. לווירוסים יש יכולת ליצור וריאנטים חדשים בקלות יחסית ובפרקי זמן קצרים. יכולת זו נובעת מכך שהגנום שלהם מורכב במקרים רבים מsingle strand, שצובר מוטציות רבות יותר מdouble strand, ובנוסף, הגנום שלהם קצר יחסית ולכן המוטציות משפיעות עליו באופן משמעותי[2]. וריאנטים חדשים יכולים להיווצר ב2 דרכים אפשריות: 1) צבירת מוטציות באיטיות, בזו אחר זו. רוב המוטציות יהיו ניטרליות, אך חלקן יצרו יתרון לנגיף או יגרמו לו נזק. 2) קפיצה- צבירת הרבה מוטציות בבת אחת. קפיצה יכולה להיווצר בגופם של חולים עם דיכוי חיסוני שחלו במשך זמן רב (חודשים) ובמשך הזמן הזה הנגיף צבר בגופם עוד ועוד מוטציות. בפרויקט שלנו התמקדנו בסוג זה של וריאנטים.

בדצמבר 2019, מחלה חדשה הנגרמת על ידי וירוס ממשפחת הקורונה התגלתה בWuhan שבסין. מחלה זו, שהפכה למגפה, קיבלה על ידי ארגון הבריאות העולמי את השם COVID-19 (coronavirus disease 2019) . מגפה זו הינה גורם התמותה העולמי המשמעותי ביותר מאז מלחמת העולם השנייה[3]. הגנום של וירוס הקורונה הוא מסוג ssRNA, ולכן נוטה לעבור מוטציות רבות בהשוואה לוירוסי DNA (למרות שמכיל גם מנגנון proofreading). מטבע הדברים, כאשר נגיף מתפשט בצורה רחבה כל כך, הוא צובר מוטציות רבות ונוצרים וריאנטים. וריאנט (variant) הוא מונח בביולוגיה המתאר תת סוג של מיקרואורגניזמים. הווריאנט שונה מבחינה גנטית מהזן הראשי שלו, אך אינו שונה מספיק כדי להיחשב זן בפני עצמו. נכון להיום, שלושה וריאנטים של וירוס הקורונה התפרסמו והתפשטו במיוחד: אלפא α, דלתא Δ ואומיקרון O[4].

וריאנטים חדשים עלולים ליצור לנגיף יתרונות שלא היו קיימים ב"גרסאות" הקודמות שלו. לדוגמא: יכולת הדבקה טובה יותר, מחלה קשה יותר, עמידות לחיסון ועוד. ולכן קיימת חשיבות רבה לזיהוי מוקדם של וריאנטים חדשים העלולים להיות מסוכנים, כדי לעצור את ההתפשטות שלהם. לצורך כך, הוקם על ידי אמ"ן (אגף המודיעין של צה"ל) ומשרד הבריאות צוות שעוסק בפיתוח "רדאר וריאנטים"- מערכת שיכולה לנתח דגימות RNA של וירוס הקורונה שרוצפו בכל העולם ולהתריע בשלב מוקדם מאוד כאשר קיים חשש לווריאנט חדש ובעייתי. הפיתוח הזה הינו רעיון חדשני ומקורי של מדינת ישראל, ואינו קיים במקומות אחרים (שנעזרים ומתייעצים במומחים הישראלים). מטרת הפרויקט שלנו הייתה להשתלב בפיתוח רדאר הווריאנטים על ידי כתיבת אלגוריתם שיוכל לזהות מקרה של קפיצה מתוך מאגר רצפי RNA של וירוס הקורונה.

מגפת הקורונה הינה המגפה הראשונה עימה מתמודד העולם המודרני, מעולם לא נצרך העולם הממוחשב להתמודד עם נגיף שמתפשט וצובר מוטציות במהירות רבה כל כך כמו וירוס הקורונה. מסיבה זו, לא קיימת מערכת זיהוי וריאנטים טובה מספיק עליה ניתן להתבסס בפיתוח רדאר הווריאנטים. קיימים עצים פילוגנטיים המתארים התפתחות וריאנטים בווירוסים שונים, ביניהם גם הקורונה, אך העצים הללו אינם מדויקים ואמינים מספיק לצורך חיזוי והתרעה על וריאנט חדש בזמן אמת. יתר על כן, הוספה של וריאנט חדש לעץ פילוגנטי יכולה להתבצע רק לאחר שהווריאנט כבר התפשט וזוהה, ואילו מטרת רדאר הווריאנטים היא לזהות את הווריאנט החדש לפני שהוא מתפשט. בפרויקט נעזרנו בעצים פילוגנטיים על מנת להעריך ולבדוק את התוצאות שלנו, אך לא התבססנו עליהם על מנת לזהות קפיצות. (עצים פילוגנטיים ניתן לראות באתר nextrain[5])

בס"ד

מטרת הפרויקט שלנו הייתה למצוא קפיצות ברצפי RNA של וירוס הקורונה שנדגמו מחולים ברחבי העולם, על מנת לזהות ולהתריע על וריאנטים חדשים בזמן אמת.

שיטות:

- **לימוד רקע תיאורטי:**

רקע ביולוגי – קורסים אינטרנטיים העוסקים באפידמיולוגיה גנומית ופילוגנטיקה[6,7]. בקורס האפידמיולוגיה הגנומית למדנו על הנושאים שקשורים ליישום הגנומי לחקר האפידמיולוגי של הנגיף. בתוכו נכלל גם מבוא לקורס בפילוגנטיקה והשימוש ב-Nextstrain. בקורס בפילוגנטיקה קיבלנו ידע בסיסי על פילוגנטיקה כללית, עם דגש על קריאה וניתוח של עצים פילוגנטיים.
רקע חישובי – קורס אינטרנטי של חבילת PANDAS ב-python (התמודדות עם data frames בפרויקט).

- **משימה ראשונה – הגדרת סט מוטציות בבסיסו של הוריאנט:**

רקע – סטטיסטיקה להימצאות של כל מוטציה בריצופים הנתונים. מוטציות הבסיס יימצאו באחוז גבוה של הרצפים, מוטציות הוריאנט באחוז נמוך.
קלט – קובץ CSV המכיל את פירוט המוטציות לכל רצף.
פלט – קובץ CSV המכיל את המוטציות מסודרות לפי סדר הימצאותן ברצפים, בסדר יורד. לכל מוטציה מוצג גם אחוז הרצפים בהם נמצאה, ובטור נוסף את מספר הרצפים הנ"ל.
האלגוריתם –
*קריאת קובץ הקלט
*חישוב מספר הרצפים בקובץ
*חישוב הסטטיסטיקות הנדרשות לכל מוטציה
*עריכת קובץ ה-CSV בהתאם לפלט הדרוש (הוספת הסטטיסטיקות, מיון בסדר יורד)

כאן התמודדנו לראשונה עם ממשק משתמש, ע"י שימוש בחבילת tkinter ב-python.

כעת התחלנו לעבוד על אלגוריתם ממשי בעזרתו נמצא קפיצה שקיימת ברצפים ממשיים שרוצפו, שמהווה פוטנציאל לוריאנט חדש.

- **אלגוריתם לזיהוי קפיצה:**

גרסה א:

מטרה – זיהוי קפיצה ברצפים של וריאנט ידוע
קלט – 2 קבצי CSV: 1. reference (מוטציות של הוריאנט הידוע) 2. Samples (רצפים עם פירוט המוטציות שלהם).
פלט – הקפיצות (לכל קפיצה המוטציות שמכילה + הרצפים שמכילים אותה).
האלגוריתם –
*קבועים: MIN_JUMP (אורך הקפיצה המינימלי)
*קריאת קבצי הקלט
*יצירת מילון dict_seq שיכיל את המוטציות הייחודיות (שלא מופיעות בקובץ ה-reference) לכל רצף.
*הגדרת קומבינציות המוטציות בין כל 2 רצפים ב-dict_seq. לשם כך נגדיר 2 רשימות: mut, seq.
seq מכיל את שמות הרצפים באותה קומבינציה, ו-mut את המוטציות המשותפות להם בהתאמה.
*מעבר על כל קומבינציה עם מצבים אפשריים:
1. גודל הקומבינציה קטן מ-MIN_JUMP: הקומבינציה בין אותם 2 רצפים קטנה מידי, ולא מהווה חלק מקפיצה.
כעת עוברים על כל רצף ב-dict_seq בהשוואה לאותה קומבינציה, יוצרים את combine (קומבינציה של הרצף עם אותה קומבינציה), ובודקים:
2. הגודל של combine גדול/שווה ל-MIN_JUMP: קפיצה אפשרית.
3. הגודל של combine גדול מ-0 אך קטן מ-MIN_JUMP:

3.1. גודל ההפרש בין combine לקומבינציה קטן מ-MIN_JUMP: פסילת קפיצה.
3.2. גודל אותו הפרש גדול/שווה ל-MIN_JUMP: הקומבינציה מהווה קפיצה אפשרית, שאינה מכילה את מוטציות ה-combine.
4. הגודל של combine הוא 0: קיימת קפיצה אפשרית באותה קומבינציה, ללא הרצף השלישי (combine מודגר להיות הקומבינציה ההתחלתית).

לאחר הבדיקות הללו, ובדיקה נוספת למניעת דופליקטים, ניתן להסיק שאותו combine (שעבר את כל הסינונים) מהווה קפיצה.
*הגדרת 2 רשימות: jump_seq, jump_mut.
ל-jump_seq מוסיפים את שמות הרצפים שמכילים את הקפיצה, ול-jump_mut בהתאמה את מוטציות הקפיצה.
*ההדפסות הנחוצות מ-jump_seq, jump_mut.

גרסה ב:

זו וריאציה נוספת של האלגוריתם (מיושנת יותר), כעת עם **התחשבות בסוג המוטציות** – הקפיצה צריכה להכיל לפחות MIN_MUT מוטציות שמוגדרות כ-**NS** (Non Synonymous). אלו מוטציות שלא מוגדרת כ-extragenic (לא מוטציות שאינן במסגרת הקריאה של חלבונים) ובנוסף לכך הן לא מוטציות "שקטות" (לא מוטציות שאינן מביאות לשינוי ברצף חומצות האמינו).
המסקנה היא שלקפיצה מחפשים מוטציות שנמצאות במסגרת הקריאה ומביאות לשינוי ברצף חומצות האמינו. (נשים לב שההגדרה של הקפיצה עדיין מכילה את כל המוטציות, גם אלו שאינן NS). באופן כללי האלגוריתם דומה מאוד לגרסה א, אבל כן מסתמך על קוד מיושן יותר שכתבנו. כקלט מקבלים קובץ אחד שמכיל גם את הרצפים וגם את מוטציות הווריאנט בעמודה נפרדת (לכן נחשב דומה לגרסה א – מסתמך על reference). במהלך האלגוריתם, בנוסף למציאת הקפיצה, נעשה גם סיווג לכל מוטציה לסוג שלה (N – Non Synonymous, S – כל השאר). בנוסף לכך, נעשה שימוש במילוני עזר long & short – הסבר בקוד עצמו.

גרסה ג:

השלב האחרון היה שיפור וייעול האלגוריתם (גרסה א), ויצירת גרסה שלו **ללא קובץ reference** של וריאנט נתון. השינוי היחידי בגרסה זו הוא בהתחלה, ביצירת dict_seq (כעת אין צורך להוריד את מוטציות הווריאנט שבקובץ ה-reference).

בס"ד

תוצאות:

התוצאות שהתקבלו מהרצה האלגוריתם שלנו הן קפיצות אפשריות בתוך אוסף הרצפים שהתקבלו כפלט. לעיתים התקבלה קפיצה אחת אפשרית ולעיתים יותר, אך כולן עמדו בתנאים שאותם הגדרנו באלגוריתם: מספר הרצפים בקפיצה הוא 2 או יותר, מספר המוטציות בקפיצה הוא 3 או יותר ומספר המוטציות מתוך המוטציות המשתתפות בקפיצה שיכולות להופיע גם ברצפים אחרים הוא מקסימום 2. את התוצאות הדפסנו לקובץ טקסט, כאשר עבור כל קפיצה שעמדה בקריטריונים הנ"ל הודפסו המוטציות המשתתפות בקפיצה והרצפים המשתתפים בקפיצה.

דוגמאות הרצה – קבצי הפלט:

- <u>משימה ראשונה – הגדרת סט מוטציות בבסיסו של הווריאנט:</u>

J	I	H	G	F	E	D	C	B	A	
perce	freq	varname	annotation	varclass	variant	protein	qvar	refvar	refpos	1
100	554	S:D614G	Spike	SNP	D614G	S	G	A	23403	2
100	554	NSP12b:P	RNA-depe	SNP	P314L	NSP12b	T	C	14408	3
99.639	552	S:S982A	Spike	SNP	S982A	S	G	T	24506	4
99.639	552	S:D1118H	Spike	SNP	D1118H	S	C	G	24914	5
99.639	552	S:N501Y	Spike	SNP	N501Y	S	T	A	23063	6
99.639	552	NSP3:F10	Predicted	SNP_silen	F1089F	NSP3	T	C	5986	7
99.458	551	ORF8:Q27	ORF8 prot	SNP_stop	Q27*	ORF8	T	C	27972	8
99.458	551	ORF8:Y73	ORF8 prot	SNP	Y73C	ORF8	G	A	28111	9
99.458	551	NSP3:F10	Predicted	SNP_silen	F106F	NSP3	T	C	3037	10

- <u>אלגוריתם לזיהוי קפיצה:</u>

1. זיהוי קפיצה ברצפים נתונים (ללא קובץ reference – גרסה ג):

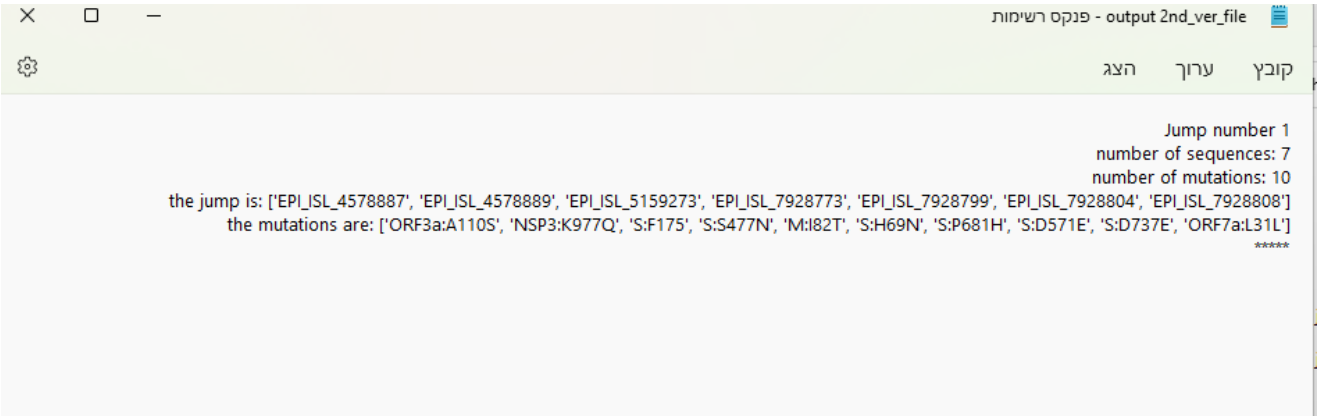
output 3rd_ver_file.txt - פנקס רשימות

קובץ ערוך הצג

```
Jump number 1
number of sequences: 11
number of mutations: 13
the jump is: ['EPI_ISL_10079203', 'EPI_ISL_10079216', 'EPI_ISL_10079231', 'EPI_ISL_11221226', 'EPI_ISL_11776163', 'EPI_ISL_11776165', 'EPI_ISL_11776191', 'EPI_ISL_11776192', 'EPI_ISL_12305044', 'EPI_ISL_7900788', 'EPI_ISL_8528732']
the mutations are: ['A11206G', 'C23638T', 'C12756T', 'C25585G', 'G29742T', 'C24707A', 'G19936A', 'C14110A', 'G2747T', 'C2902T', 'G20433T', 'C11036A', 'C27752T']
*****
Jump number 2
number of sequences: 3
number of mutations: 9
the jump is: ['EPI_ISL_10883114', 'EPI_ISL_9416740', 'EPI_ISL_9416753']
the mutations are: ['A13533G', 'C24370T', 'C25300T', 'C28311T', 'G29742T', 'T25627C', 'G25644T', 'C12890T', 'C29870A']
*****
Jump number 3
number of sequences: 4
number of mutations: 7
the jump is: ['EPI_ISL_9670643', 'EPI_ISL_9670644', 'EPI_ISL_9670645', 'EPI_ISL_9670647']
the mutations are: ['G12613A', 'C12513T', 'A19359G', 'G29742T', 'C20930T', 'C25585G', 'C27752T']
*****
```

שורה 1, עמודה 1 100% Windows (CRLF) UTF-8

2. זיהוי קפיצה + התחשבות בסוג המוטציות (גרסה ב):



תוצאות האלגוריתם יכולות לשמש לזיהוי קפיצות מתוך מאגרי רצפים של וירוס הקורונה ובאמצעות הקפיצות הללו- לאתר וריאנטים חדשים של הווירוס בזמן אמת.

דיון:

התוצאות שקיבלנו על ידי הרצה של האלגוריתם הן מוטציות חדשות שהופיעו בגנום של וירוס הקורונה העלולות להוות וריאנט חדש, ובנוסף- הרצפים בהם הופיעו מוטציות אלה. ניתן להשתמש בתוצאות אלו על מנת לזהות בזמן אמת וריאנטים חדשים לפני שהם התפשטו בצורה רחבה ואת המקור שבו הם הופיעו, ועל ידי שימוש באמצעים כגון בידוד, סגר ובדיקות רבות- למנוע את התפשטותם.

האלגוריתם שכתבנו אינו עומד בפני עצמו, הוא נועד להיות חלק ממערכת גדולה יותר הנקראת "רדאר וריאנטים". על מנת שיהיה ניתן לזהות וריאנטים חדשים ולהתריע עליהם בצורה מיטבית יש לבצע מספר שלבים הקודמים לאלגוריתם : צריך לבצע ריצוף לדגימות RNA של קורונה מרחבי העולם ולבצע עיבוד ראשוני של התוצאות המתקבלות כקובץ fastq (בעזרת pipeline של משרד הבריאות). לאחר מכן, השוואה של קובץ הfastq לווריאנטים ידועים, סינון של מוטציות החוזרות על עצמן בחלק גדול מהרצפים (הן לא יהוו חלק מקפיצה) וארגון של התוצאות בקובץ csv עליו ניתן לעבוד באלגוריתם שלנו.

שלבים נוספים שרצוי לבצע לאחר שמקבלים את הפלט מהאלגוריתם הם אימות של התוצאות- בדיקה שהמוטציות שהתקבלו אכן מהוות קפיצה משמעותית ואינן מופיעות בהדרגה ברצפים אחרים (ייתכן שהן מופיעות ברצפים שלא בדקנו בהרצה שבה נמצאה הקפיצה). וכן בדיקה שהמוטציות האלו אכן יכולות לגרום ליתרון לנגיף- להפוך אותו למדבק יותר, מסוכן יותר או עמיד לחיסון.

ביבליוגרפיה(REFERENCES) :

1. Harvey, W.T., Carabelli, A.M., Jackson, B. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* **19,** 409–424 (2021).
2. Sanjuán, R. & Domingo-Calap, P. Mechanisms of viral mutation. *Cell Mol Life Sci* **73**, 4433–4448 (2016).
3. Chakraborty, I. & Maity, P. COVID-19 outbreak: Migration, effects on society, global environment and prevention. *Sci Total Environ* **728**, (2020).
4. AID GENOMICS. מה זה וריאנט ואילו סוגי וריאנטים קיימים? *https://aidg.co.il/variant/*.
5. Trevor Bedford & Richard Neher. nextrain. *https://nextstrain.org/*.
6. Laura Emery. Phylogenetics. *https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics/#vf-tabs__section--overview* (2022).
7. Centers for Disease Control and Prevention. COVID-19 Genomic Epidemiology Toolkit. *https://www.cdc.gov/amd/training/covid-19-gen-epi-toolkit.html*