# 234122 – מבוא לתכנות מערכות

# תרגיל בית 1

# סמסטר אביב תשע"ח

**מתרגל אחראי:** אור אייזקס

תאריך הגשה: 22.4.18 23:30
אופן ההגשה: הגשה בזוגות **מומלצת !!** הגשה אלקטרונית בלבד, באמצעות האתר של הקורס במערכת GR. פרטים נוספים הרלוונטיים להגשה רשומים בסוף תרגיל זה.
משקל התרגיל: 5% מהציון הסופי (תקף)

## הערה

שאלות בנושא התרגיל יש לשלוח לאתר הפורום במערכת Moodle (לא GR ) בלבד.
יש להיכנס ל- https://moodle.technion.ac.il - חשבון האישי - אתר הקורס - פורום של תרגיל בית 1. בבקשה לא לשלוח שאלות בעניין התרגיל למייל האישי של המתרגל האחראי על התרגיל.

## מבוא

סטודנט חרוץ בשם בוב סיים הסמסטר את הקורס "מבוא לתכנות מערכות" 234122 ומחליט שהוא רוצה לנסות לעבוד בתחום פיתוח תוכנה. בוב עשה באותו הסמסטר גם את הקורס "גנטיקה כללית" 134020, שם הוא הכיר את אליס, סטודנטית בפקולטה לביולוגיה.
אליס סיפרה לבוב שהיא עובדת במעבדה אשר מחפשת מתכנת לפרוייקט מאתגר ויוצא דופן.
בוב מבין שזו הזדמנות טובה בשבילו גם להתנסות בעבודה וגם להמשיך ולפתח את הקשר עם אליס, אותה הוא מחבב עד מאוד.

כעבור מספר שבועות בוב מתקבל לעבודה (אחרי שעבר שיחת הכרות עם מנהלת המעבדה, אותה הרשים הן בציונים, הן ביכולות שלו, הן במוטיבציה) ומקבל לידיו את הפרוייקט.
הוא נדרש לממש מערכת אשר מבצעת אנליזה סטטיסטית עבור מערכת ביולוגית נתונה.
בוב מקבל לידיו מאמר המציג את הרעיון מאחורי השיטה, אותו הוא קורא מספר פעמים עד שלבסוף מבין את השלבים בבניית המערכת. הוא מחלק את המימוש לשלושה חלקים, כאשר כל חלק ממומש ע"י סקריפט בשפת BASH.

בתרגיל זה, נממש את המערכת יחד עם בוב.

# סעיף א'

בוב קיבל לידיו קבצים המכילים רצפי דנ"א בפורמט הנקרא sanger fastq (נקראים גם "raw" files):

- כל רצף מאופיין ע"י 4 שורות רציפות בקובץ.
- שורה ראשונה הינה השם של הרצף (מתחיל בתו '@').
- שורה שניה היא רצף הדנ"א עצמו (קומבינציה כלשהי של האותיות ATCG).
- שורה שלישית ריקה או מכילה תו בודד.
- שורה רביעית מכילה רצף תווים ללא שום לוגיקה מוגדרת.

לצערינו, בשל תקלה בתוכנית יכול להיות שחלק מהרצפים מכילים אותיות לא חוקיות (הן חייבות להכיל רק A, C, G, T כאותיות uppercase בלבד).

**דוגמא**

```
@seq_1
GATVTTGGGGTTCAAAGCAGTATCGATCAAAhTAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
@seq_2
ADcTCGTAGTCTAGTCTATGCTAGTGCGATGCTAGTGCTAGTCGTATG2CATGGCTATGTGTG

208DA8308AD8SF83FH0SD8F08APFIDJFN34JW830UDS8UFDSADPFIJ3N8DAA
```

כל השורות פרט לשורה של הרצף של הדנ"א הן תופעות לוואי של המכונה שכתבה את הרצף לקובץ, ולכן בשלב הראשון יש לסנן את השורות המכילות רצפי דנ"א משאר הדברים הפחות מעניינים.

כתבו סקריפט בשם rawFileFilter אשר מקבל כקלט:

1. infile - שם של "raw" file (ניתן להניח שנמצא באותה תיקייה כמו הסקריפט rawFileFilter). ניתן להניח שהוא מילה יחידה ללא רווחים.
2. num - מספר בין 1 ל-49 (כולל).
3. outfile – שם של קובץ פלט. ניתן להניח שהוא מילה יחידה ללא רווחים.

rawFileFilter יאסוף מהקובץ infile את num האותיות הראשונות **<u>החוקיות</u>** מכל רצפי הדנ"א, ויוציא כל אחד בשורה נפרדת לקובץ הפלט outfile, ניתן להניח שקובץ outfile לא קיים במערכת ויש ליצור אותו לתוך תיקייה שהשם שלה הוא המספר num (כל תיקייה num תכיל קבצים אשר בהם הרצפים הם באורך num).

**המחשה עבור הדוגמא מלעיל(כאשר num=5):**

```
Rawfilefilter infile 5 outfile
GATTT
ATCGT
```

**הערות:**

- אם infile לא קיים יש לסיים את התוכנית.
- אם num אינו בעל ערכים בתחום שנאמר יש לסיים את התוכנית (ניתן להניח שזה מספר).
- אם התיקייה של num אינה קיימת יש ליצור אותה.
- ניתן להניח שאורך רצפי הדנ"א ב- infile הוא אחיד.
- ניתן להניח שאורך רצפי הדנ"א ב- infile גדול מ-num פלוס האותיות הלא חוקיות.
- בכדי לבודד רצף תווים בשורה, השתמשו בדגל המתאים של הפקודה cut.

## סעיף ב'

בוב מעוניין לסדר את קבצי הרצפים כך שיהיה קל למנות את המופעים של כל רצף.

כתבו סקריפט בשם count אשר מקבלת כקלט:

1. num – שם התיקייה המכילה את הקבצים שנוצרו מהסעיף הקודם.

count יאגד את כל הקבצים שבתיקייה num ואז ימיין את הרצפים ע"פ מספר הופעתם בסדר יורד כאשר במקרה תיקו ההופעה לפי סדר לקסיקוגרפי יורד ויוציא אותם כפלט סטנדרטי בפורמט הבא:

count**TAB**sequence**newLine**

כל שורה מכילה את כמות ההופעות של הרצף (count) ואחריו רווח מסוג tab, אחריו יש הרצף (sequence) ואחריו תו ירידת שורה ללא רווח. הרצפים יכתבו בסדר יורד (ראשית יופיע הרצף בעל מספר מופעים מירבי, בסוף יופיע רצף בעל מספר מופעים מינימלי).

**המחשה עבור הדוגמא מסעיף א'(כאשר num=5):**

```
count 5
1     GATTT
1     ATCGT
```

**הערות**

- אם התיקייה num לא קיימת, יש לסיים את התוכנית.
- ניתן להניח כי כל הקבצים בתיקייה חוקיים.
- אין להשתמש בקבצים זמניים, תעבדו תחת ההנחה שאפשר להכניס את כל המידע למשתנה יחיד.

## סעיף ג'

עתה נתבקשנו לבצע את המשימה הבאה: יש לנו קלט של קובץ שמכיל רצפי דנ"א כמו שמקבלים מסעיף א'. לכל רצף דנ"א בקובץ יש להפרידו לכמה שורות כך שבשורה הראשונה הרצף מופיע מתחילתו עד ה-T הראשון שיש בו, בשורה השנייה הרצף מופיע מתחילתו עד ה-T השני בו, בשורה השלישית הרצף מופיע מתחילתו עד ה-T השלישי בו וחוזר חלילה עד הופעת ה-T האחרון (לא חייב שהרצף המקורי יופיע).

כתבו סקריפט בשם seperateSeq אשר מקבלת כקלט:

1. infile - שם של קובץ (ניתן להניח שנמצא באותה תיקייה כמו הסקריפט seperateSeq). **שם הקובץ עלול להכיל רווחים.**

seperateSeq יבצע את מה שנאמר בתחילת הסעיף לכל רצף בתוך infile. בצורה:

Sequence**<u>newLine</u>**

כל שורה מכילה רצף יחיד (sequence) ואחריו תו ירידת שורה ללא רווח.

### המחשה עבור הדוגמא מסעיף א'(כאשר num=5):

```
seperateSeq infile
GAT
GATT
GATTT
AT
ATCGT
```

**הסבר:**
יש לנו רצף של כל מילה לפי הסדר מסעיף א' בסדר יורד לקסיקוגרפית.

### הערות

- אם infile לא קיים יש לסיים את התוכנית.

## הערות חשובות לפני הגשת התרגיל

לפני שאתם מגישים את התוכנית, הקפידו לבדוק אותה כאשר היא רצה בחשבון שלכם במחשב csl2.
וודאו שאתם עובדים ב – BASH ולא ב-shell אחר ע"י הפקודה echo $0
במקרה הצורך, עברו ל-BASH ע"י הפקודה bash
ניתן להניח שהתיקייה הנוכחית (.) תמצא ב-PATH. זה אומר שאתם רשאים לקרוא לתוכניות script שלכם מתוך התוכניות שכתבתם ללא שימוש בסימון "./" לפני כל סקריפט.
אסור להריץ תוכניות script ע"י הפקודה source.
הסעיפים נבדקים עצמאית אחד מהשני.

## דוגמאות לבדיקת התוכנית

התיקייה ~mtmchk/public/1718b/ex1/examples מכילה דוגמאות הרצה, כמו גם הפלט המצופה המתקבל ע"י כל אחד מהתוכניות שתכתבו בתרגיל זה. התיקייה מורכבת מ:

- תיקייה בשם input ובה שתי דוגמאות לקבצי fastq.
- תיקייה result המכילה את הפלט של התוכניות מהסעיפים א' - ג' על קבצי הקלט בתיקייה input.
- קובץ commands.txt המכיל הוראות השוואה בין הפלט של התוכניות שאתם כתבתם עם הפלט בתיקייה res.

בדקו, בין היתר, שהתוכנית שלכם תומכת בדוגמאות אלה.

## הגשת התרגיל

התרגיל יוגש אלקטרונית בלבד, באמצעות האתר של הקורס.
ההגשה האלקטרונית כוללת:
קובץ ZIP שמכיל:

1. כל קבצי הסקריפט הנדרשים. ניתן להגיש גם קבצי עזר נוספים אם כתבתם קבצים כאלה.
2. קובץ student.txt שמכיל פרטים אישיים של המגישים.הקובץ נמצא תחת תווית שעורי בית באתר הקורס. בבקשה לא לשנות את מבנה הקובץ ולא לשנות את הכותרות המצויות בו וגם לא למחוק תווי רווח ו/או להוסיף שורות. אך ורק להשלים את הפרטים החסרים במקומות הרלוונטיים. בבקשה לכתוב באנגלית בלבד.

**מומלץ להשתמש בפקודה sed בעבודה.**
**במידה וכתבתם סקריפט בווינדוס תזכרו לשנות לנוסח UNIX-** ניתן ומומלץ להשתמש בפקודה dos2unix.
**הקפידו בבקשה על שמות מדויקים של הקבצים**
**הקפידו בבקשה על קובץ zip בלבד, ולא שום פורמט אחר**
**הקפידו בבקשה על כך שבקובץ zip אין שום תת ספריות**
**הקפידו שכל הקבצים נמצאים כנדרש**

# בהצלחה!