מבוא לתכנות מערכות 10010 סמסטר א' – תש"פ

תרגיל בית מס' 4

נושא התרגיל: קבצים, מחרוזות, מצביעים לפונקציות, הידור מותנה, סיביות

יש להגיש אך ורק דרך תפריט המטלות שבאתר הקורס, כפי שהוסבר בתרגול.

**ניתן לעבוד בזוגות (שני הסטודנטים צריכים לעלות את העבודה למודל)**

הנחיות הגשה כלליות:

* התרגיל ייבדק בסביבת MS Visual Studio 17
* הקוד חייב לעבור קומפילציה, **קוד שאינו מתקמפל לא ייבדק**
* יש להגיש את כל תיקיית הפרויקט, מקובצת לקובץ zip/rar/tar ששמו כשם הסטודנט. שם פרטי ומשפחה.

**הוראות כלליות:**

1. יש להקפיד על כללי הנדסת התוכנה:
   1. מבנה התכנית (הזחות) ותיעוד במידת הצורך.
   2. חובה להשתמש בקבועים במקומות המתאימים.
   3. יש להשתמש בפונקציות קצרות, כלליות, קריאות ושימושיות.
   4. יש להקפיד על בדיקת תקינות קלט.
   5. הפלט צריך להיות כפי שניתן בתרגיל.
   6. קוד קצר, לא מסורבל ויעיל הן מבחינת כתיבתו והן מבחינת ריצת התוכנית.
2. יש לחלק את הפרויקט לקבצים. חשוב איך לארגן את הקוד בקבצים השונים.
3. יש לקרוא את ההוראות במדויק.

**פירוט התרגיל**

**עימוד רצפים** היא שיטה לסידור המידע על אודות רצפי [DNA](https://he.wikipedia.org/wiki/DNA) ,[RNA](https://he.wikipedia.org/wiki/RNA) או [חלבונים](https://he.wikipedia.org/wiki/%D7%97%D7%9C%D7%91%D7%95%D7%A0%D7%99%D7%9D), בצורה כזו שיהיה ניתן לזהות אזורים דומים, אשר ייתכן כי הם תוצאה של קשרים פונקציונאלים, מבניים או אבולוציוניים.

<https://he.wikipedia.org/wiki/%D7%A2%D7%99%D7%9E%D7%95%D7%93_%D7%A8%D7%A6%D7%A4%D7%99%D7%9D>

<https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/Bioinformatics+Tools+FAQ>

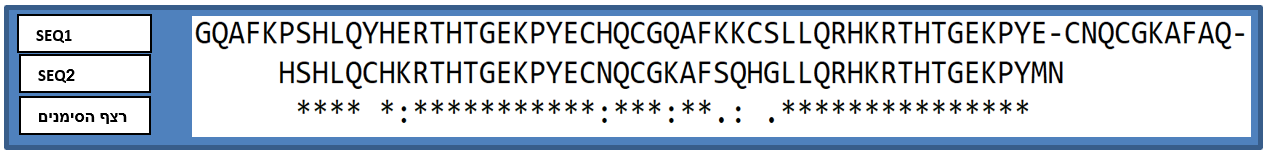
בתרגיל הזה ניישם את הגישה הנאיבית של סידור המידע.

את רצפי ה-[DNA](https://he.wikipedia.org/wiki/DNA) ,ה-[RNA](https://he.wikipedia.org/wiki/RNA) או ה[חלבונים](https://he.wikipedia.org/wiki/%D7%97%D7%9C%D7%91%D7%95%D7%A0%D7%99%D7%9D) אפשר לקודד בעזרת סדרת אותיות, כולל רווחים.

ישנה משמעות רבה לזהות מלאה של שני רצפים או לדרגת הקירבה ביניהם. אחת מהבעיות הקלאסיות בביו-אינפורמטיקה היא לזהות האם ברצף הנתון SEQ1 נמצאת כתת-סידרה סדרה SEQ2 אחרת. ואם לא – מה המקום ב-SEQ1 שהכי דומה ל-SEQ2 מבחינת קריטריון קירבה מסוים.

באיור מוצגים שני רצפים באותו גודל, SEQ1 ומתחתיה SEQ2. הרצף SEQ2 מתחיל ממקום 5 ביחס לתחילת SEQ1 . הרצף של הסימנים הוא בגודל של SEQ2.

* מקף מסמל חוסר של חומצה אמינית במקום מסויים ברצף.
* בשורה התחתונה מופיעים סימני השוואה בין האותיות ברצפים:
* כוכבית - האותיות זהות (אותה חומצה אמינית)
* רווח – האותיות לא זהות (חומצות אמיניות שונות)
* נקודתיים – האותיות לא זהות (חומצות אמיניות שונות אך עם תכונות קרובות)
* נקודה – האותיות לא זהות (חומצות אמיניות שונות עם תכונות דומות)



מהם אותיות עם תכונות **קרובות** (סימן השוואה **נקודתיים**)?

נתונה רשימה של קבוצות חומצות אמיניות שונות.

אותיות הנמצאות באותה קבוצה מוגדרות כאותיות עם תכונות קרובות.

|  |  |  |
| --- | --- | --- |
| NDEQ | **NEQK** | STA |
| MILV | QHRK | NHQK |
| FYW | HY | MILF |

לדוגמא: באיור הזוג השמיני מצד שמאל **EK** סומן עם נקודתיים כי הוא מופיע בקבוצה **NEQK**

מהם אותיות עם תכונות **דומות** (סימן השוואה **נקודה**)?

הזיהוי של התכונות **הדומות** מתבצע רק אם נכשל הזיהוי של תכונות **הקרובות**.

נתונה רשימה של קבוצות חומצות אמיניות שונות.

אותיות הנמצאות באותה קבוצה מוגדרות כאותיות עם תכונות קרובות.

|  |  |  |
| --- | --- | --- |
| SAG | ATV | CSA |
| SGND | STPA | STNK |
| NEQHRK | NDEQHK | **SNDEQK** |
|  | HFY | FVLIM |

לדוגמא: באיור הזוג **KS** סומן עם נקודה כי הוא מופיע בקבוצה **SNDEQK** (ישנה קבוצה נוספת אבל מספיק למצאו שייכות לקבוצה אחת)

**שים לב: הסימן נקודתיים (: ) עדיף על הסימן נקודה (.)**

**משימות התרגיל**

בכול המשימות הרצף SEQ2 לא ארוך יותר מהרצף SEQ1

1. כתוב פונקציה **createSequence** שמקבלת שם של קובץ טקסט ובונה רצף אותיות. בקובץ רשומות אותיות הרצף אחת אחרי השנייה ללא רווחים ביניהן.
2. כתוב פונקציות **createSigns**שמקבלת שני רצפים של אותיות SEQ1 ו- SEQ2בגדלים שונים ומיקום **n**של SEQ2 מתחילת SEQ1 . הפונקציה בונה רצף סימני השוואה בצורה דחוסה.

ההשוואה מתבצעת בין כל הזוגות SEQ1[n +i] ו-SEQ2[i] כל עוד SEQ2 לא "גולשת" מ- SEQ1. מספור המקום מתחיל מ-0.

יש להחזיק את רצף הסימנים בצורה דחוסה, תוך שימוש בכמות זיכרון מינימאלית.

רמז: הוגדרו רק ארבעה סימני השוואה שונים – רווח, כוכבית, נקודתיים ונקודה.

1. כתוב פונקציה **getCount**להערכת קירבה בין שני הרצפים. הפונקציהמקבלת רצף סימני השוואה בצורה דחוסה ומחזירה הפרש בין כמות הכוכביות ובין כמות הנקודתיים ברצף.

1. כתוב פונקציות **evaluateDifference**שמקבלת שני רצפים של אותיות SEQ1 ו- SEQ2בגדלים שונים ומיקום **n .** הפונקציה מחזירה את ערך הקירבה בין שני הרצפים. ההשוואה מתבצעת בין כל הזוגות SEQ1[n +i] ו-SEQ2[i] כל עוד SEQ2 לא "גולשת" מ- SEQ1. מספור המקום מתחיל מ-0.
2. (משימת רשות) עבור כל **n** כך ש- SEQ2 לא "גולשת" מ- SEQ1 בנה מערך זוגות **(n, p)** כאשר **n** הוא מיקום של SEQ2 מתחילת SEQ1 ו-**p** הוא ערך הקירבה בין שני הרצפים.
3. (משימת רשות) מיין את מערך הזוגות מהסעיף הקודם לפי ערך הקירבה **p**. יש להשתמש בפונקציה **qsort** לצורך המיון.
4. (משימת רשות) כתוב פונקציה **findLocation** שמקבלת את מערך הזוגות מסעיף 5 ומחפשת את הזוג **(n, p)** לפי **p** הנתון. יש להשתמש בפונקציה **bsearch** לצורך החיפוש.
5. כתוב פונקציה **main** שמדגימה את כל הפונקציונאליות שפיתחת בסעיפים הקודמים.
   * + ה-main מקבלת דרך הפרמטרים את שמות כל הקבצים הנדרשים.
     + בתוך ה-main ספק שתי אפשרויות הרצה: הצגת הרצפים והסימנים למסך או אי הצגת הרצפים והסימנים למסך. אפשרות ההרצה תיקבע על פי המקרו PRINT\_SEQUENCES שיוגדר או לא יוגדר.