

תרגיל 3, מבוא לתכנות מערכות, חורף 2020-2021

הגשה **בזוגות או ביחידים** דרך המודל
התרגיל הוא להגשה עד ליום חמישי, 10/12/2020 בשעה 23:59

תיאור התרגיל

בתרגיל זה תממשו תוכנית אשר קוראת קבצים המכילים רצפים של חלבונים ומדפיסה מספר סטטיסטיקות פשוטות עליהן. שימו לב לפרטים במהלך ובסוף התרגיל הנוגעים לפלט הנדרש והבדיקות. חלק מבדיקת התרגיל היא אוטומטית, הימנעו מאיבוד נקודות על ידי כך שתודאו שהפלט של התוכנית שלכם זהה לדוגמאות המובאות במודל.

מהם חלבונים?

חלבונים הם מולקולות אורגניות המורכבות מרצפים באורך משתנה של 20 חומצות אמינו (ח"א) אותן מסמנים באמצעות אותיות מהשפה האנגלית. למשל: P עבור Proline, A עבור Alanine ועוד. לחלבונים חלק מרכזי בכל החיים עפ"י כדה"א והם משמשים במגוון תפקידים בתא החי (בפרט כאנזימים – זרזים לריאקציות חשובות בתא). לפרטים נוספים אפשר לקרוא כאן (ובעוד המון מקומות באינטרנט): <https://he.wikipedia.org/wiki/%D7%97%D7%9C%D7%91%D7%95%D7%9F>

פורמט מקובל לקבצים המתארים רצפי חלבונים הוא fasta. אנחנו נשתמש בגרסא פשוטה של הפורמט אשר 'יגדר באופן הבא: כל חלבון מתואר בשתי שורות. כאשר

- השורה הראשונה מתחילה ב- ">" ולאחריה, בצמוד, נתון שם החלבון. שם החלבון לא מכיל רווחים מכל סוג שהוא והשורה מסתיימת מייד לאחריו בתו ירידת שורה ('ח'). הניחו ששם הרצף לא יכול להיות ארוך מ-32 תווים.
- השורה השנייה מתארת את רצף חומצות האמינו של החלבון ללא רווחים. בסוף הרצף חייב להופיע תו ירידת שורה. אין הגבלה על אורך הרצף.
- קובץ fasta יכול להכיל מספר לא מוגבל של חלבונים אחד אחרי השני (ללא רווחים ביניהם).

למשל: בדוגמא הבאה מתוארים שני חלבונים, prot1 ו-prot2:

```
>prot1
MEIVCNQNEFNIAIQLVSKAVASRPHTHPILANLLLTADQGT
>prot2
MCGIVGIIVSCDDVNOO
```

אורך הרצף prot1 הוא 41 חומצות אמינו ואורך הרצף prot2 הוא 16 חומצות אמינו.

משימות

עליכם לממש תוכנית שתקרא קובץ fasta המתאר חלבונים ותדפיס את הסטטיסטיקות הבאות על החלבונים:

- מספר החלבון בקובץ (מתחיל מ-1)
- שם החלבון
- אורך החלבון
- אחוז ח"א בחלבון שהן הידרופוביות. ח"א הבאות הן הידרופוביות (hydrophobic): A, I, L, F, V, P, G
- אחוז ח"א בחלבון שהן טעונות (charged). ח"א הבאות הן טעונות: R, K, D, E
- אחוז ח"א בחלבון שהן פולריות (polar). ח"א הבאות הן פולריות: Q, N, H, S, T, Y, C, M, W

למשל: עבור קובץ המכיל את שני החלבונים בדוגמא למעלה פלט התוכנית צריך להראות כך:

```
~/Assignments/Matam.ex1 $ ./protstats example.fasta
```

| N | Name | Len | Hydro | Charged | Polar |
|---|-------|-----|-------|---------|-------|
| 0 | prot1 | 41 | 48.8 | 12.2 | 39.0 |
| 1 | prot2 | 16 | 43.8 | 12.5 | 43.8 |

השורה הראשונה מכילה כותרות עבור כל העמודות וחייבת להיות זהה לשורה המופיעה כאן. השורות הבאות מתארות את כל החלבונים לפי סדר הופעתם בקובץ. כל הרווחים המופיעים בקובץ הפלט הם טאבים ('t').

את המימוש עליכם לעשות באופן הבא:

1. צרו קובץ `ProtStats.h` אשר יכיל את הפרטים הבאים:

- מבנה (struct) בשם `ProtStats` אשר יכיל נתונים עבור חלבון בודד בשדות הבאים:

- `name` – שם הרצף

- `length` – אורך הרצף

- `aa_freq` – מערך של 3 שדות אשר יכיל את אחוז ח"א ההידרופוביות, פולריות וטעונות.

- טיפוס בשם `aa_type` אשר יוגדר על ידי enum ויכיל שלושה ערכים אפשריים: `Hydrophobic`, `Polar` ו-`Charged`.

- הצהרה על פונקציה בשם `num_prots` אשר מקבלת שם של קובץ `fasta` של חלבונים ומחזירה את מספר החלבונים בו.

- הצהרה על פונקציה בשם `ProtStatsCreate` אשר מקבלת שם של חלבון ואת הרצף שלו ומחזירה מצביע ל-`ProtStats` שמכיל את כל המידע הרלוונטי עבור החלבון.

- הצהרה על פונקציה בשם `ProtStatsInit` אשר מקבלת שם של חלבון, את הרצף שלו וכן כתובת של `ProtStats` ומכניסה את כל המידע הרלוונטי עבור החלבון לאובייקט שהועבר.

- הצהרה על פונקציה בשם `ProtStatsCopy` אשר מקבלת מצביעים לשני משתנים מטיפוס `ProtStats` ומעתיקה את התוכן של השני לראשון.

- הצהרה על פונקציה בשם `ProtStatsSwap` אשר מקבלת מצביעים לשני משתנים מטיפוס `ProtStats` ומחליפה את תוכנם.

- הצהרה על פונקציה בשם `read_fasta_file` אשר מקבלת שם של קובץ ומצביע למשתנה מטיפוס `unsigned int` ומחזירה כתובת של מערך של `ProtStats` עם הנתונים של כל החלבונים. הפונקציה תקרא את כל החלבונים מהקובץ ותייצר מערך של `ProtStats` עם הנתונים של כל החלבונים. הפונקציה תכניס את מספר החלבונים שנקראו (גודל המערך) למשתנה שכתובתו הועברה.

- הצהרה על פונקציה בשם `find_type` אשר מקבלת תו המייצג אחת מ-20 חומצות האמינו הקיימות ומחזירה ערך מטיפוס `aa_type` אשר מייצג את הסוג שלה (הידרופובית, פולרית או טעונה). ניתן להניח שהקלט תקין ומכיל רק אותיות המייצגות את אחת מ-20 חומצות האמינו.

2. צרו קובץ `ProtStat.c` אשר יכיל את מימוש הפונקציות שהגדרתם ב-`ProtStat.h`.

3. צרו קובץ `main.c` אשר יעשה שימוש בפונקציות שהוגדרו על מנת לממש את התוכנית. התוכנית צריכה לקבל את שם הקובץ דרך ה-command line ולהדפיס את הקלט למסך (standard output).

דגשים

- הקפידו על תכנות נכון! הכניסו הערות היכן שצריך, עמדו את הקובץ כך שהקוד יראה אסטטי וברור והקפידו על קוד קריא וברור עם שמות משתנים הגיוניים.
- הקפידו על שמות הפונקציות והטיפוסים כך שיהיו כפי שהוגדר.
- באתר המודל יש דוגמאות למספר קבצי קלט והפלט המצופה מהם. הקפידו שהפלט שלכם **זהה לחלוטין** לפלט בקבצים אלה! ניתן לבדוק זאת בעזרת פקודת הלינוקס diff אשר מקבלת שני קבצים כקלט ומציגה את ההבדלים ביניהם.
- יש לבדוק תקינות קלטים לכל הפונקציות, ולהציג הודעות שגיאה בהתאם אלא אם כן צויין אחרת.
- במקרים של שימוש בזיכרון דינאמי, יש לוודא כי ההקצאות אכן ניתנו ע"י מערכת ההפעלה, וכן יש לנהל בקפידה את הזיכרון ולדאוג שבתום התכנית כל זיכרון דינאמי אכן משוחרר. אותם כללים תקפים לשימוש בקבצים.
- יש לוודא כי התכנית עוברת קומפילציית gcc על שרת החוג ללא כל שגיאות או אזהרות כלשהן, ורצה בהצלחה.

הגשה ומימוש

- כתבו את השמות ומספרי ת.ז. של המגישים בתחילת קובץ ה-main בהערה
- הגישו במערכת Moodle קובץ tar בשם `ex3.tar` המכיל את כל קבצי המנשק והמימוש

אנא הקפידו על ההנחיות על מנת להמנע מהורדת נקודות על שטויות.

בהצלחה!