פיתוח תרופות בעידן המידע

תרגיל 1

מועד הגשה: 04.11.2021

- **תרגיל הבית מיועד להגשה עצמאית. כל העתקה שתימצא תגרור פסילה של התרגיל כולו, גם עבור הסטודנט המעתיק, וגם עבור הסטודנט שהתרגיל הועתק ממנו. אנא הימנעו מאי נעימות.**
- **כל תכנית צריכה להכיל תיעוד מסודר של מטרת התכנית, ושל פקודות חשובות בגוף התוכנית.**
- **יש לרשום בראש כל קובץ פייתון הערה עם שם הסטודנט ות"ז.**
- **אין להשתמש בחומר שטרם נלמד. מתוך הנושאים שנלמדו, מותר להיעזר בכל תיעוד רלוונטי.**
- **על כל ההערות ותיעוד התכנית להכתב בשפה האנגלית בלבד!**
  - מומלץ לתעד בפורמט הרלוונטי לפייתון, כולל שימוש ב-docstrings
- **יש להקפיד על כתיבת קוד תקין ונכון בהתאם לכללים שנלמדו. קוד שאינו תואם את העקרונות שנלמדו יקבל ניקוד חלקי בלבד.**
  - יש לתת שמות משמעותיים למשתנים.
  - ניתן להעזר בפונקציות משנה כרצונכם, בהתאם לעקרונות תכנות נכון ומודולרי.
- **יש להקפיד לעבוד על פי הפורמט המבוקש, רווחים מיותרים ופיסוק שגוי יגררו הורדת ניקוד.**

התרגיל יוגש בקובץ יחיד בשם exercise1_id.py דרך תיבת ההגשה הרלוונטית במודל. ניתן להגיש מספר פעמים עד למועד ההגשה הסופי. הגרסה האחרונה היא שתיבדק.

- לדוגמה exercise1_123456789.py

על התרגיל להיות מסוגל לרוץ באופן הבא:

```
python exercise1_partX_id.py argument1 argument2
```

- כדי לקבל משתנים לתוכנית, ניתן להשתמש בתחביר המפורט בקישור הבא:

```
import sys
# print first argument
print(sys.argv[1])
```

חלק 1 (35 נק'): זיהוי רצף חוזרני מסוג Simple sequence repeat (SSR) (מיקרוסטליט)

רצף חוזרני מסוג microsatellite מוגדר בדנ"א כאשר תבנית של נוקלאוטיד אחד או יותר חוזרת על עצמה מספר פעמים בזו אחר זו. הגדרה מפורטת יותר ניתן למצוא בקישור לעיל. לצורך התרגיל, רצף חוזרני יוגדר כרצף באורך 1-6 נוקלאוטידים, החוזר ברציפות שלוש פעמים או יותר.

בהינתן רצף דנ"א, כתבו פונקציה ***find_srr(dna_seq)*** המוצאת את כל החזרות הפשוטות ברצף לפי ההגדרה הנ"ל וכן כמה פעמים הן מופיעות ברצף. רצף דנ"א יחיד עשוי להכיל מספר SRR שונים, יש להחזיר את כל האפשרויות וכל האורכים (תתי-מיקרוסטליטים). באם אין חזרות יש להחזיר None. ניתן להניח תקינות קלט.

החזרות יוחזרו כמחרוזת בפורמט הבא, בסדר עולה של אורך הרצף, וכאשר רצף שמופיע ראשון קודם:

repeat_sequence1,num_repeats1;repeat_sequence2,num_repeats2

דוגמה:

עבור הרצף `ATCAAATCAAATCAAGAGAGAG` הפונקציה תחזיר

A,3;A,3;AG,4;GA,3;AG,3;ATCAA,3

חלק 2 (25 נק'): שעתוק דנ"א

רצף הדנ"א מורכב מארבעה נוקלאוטידים: אדנין (A), גואנין (G), תימין (T) וציטוזין (C), כאשר זיווג הבסיסים ביניהם מתבצע באופן הבא: A-T, G-C.

על מנת לייצר חלבון, רצף הדנ"א מתורגם לרצף רנ"א המשלים את הרצף של הדנ"א. רצף הרנ"א מורכב מהבסיסים הבאים: A, G, T, U וזיווגם הוא A-U, G-C.

בהינתן רצף דנ"א, נרצה להמירו לרצף רנ"א. כתבו פונקציה בשם ***transcribe(dna_seq)*** המבצעת זאת ומחזירה את הרצף המשועתק לפי הכיוונים המתאימה (5' ל-3'). יש להקפיד על אותיות uppercase בפלט המוחזר. ניתן להניח תקינות קלט.

דוגמה:

עבור הרצף `ATCaaG` הפונקציה תחזיר `CUUGAU`

חלק 3 (25 נק'): תרגום רנ"א

רצף רנ"א נקרא על ידי הריבוזום, שמתאים לכל 3 נוקלאוטידים חומצה אמינית מתאימה. על מנת להתחיל את התרגום לחומצות אמינו, הריבוזום זקוק לקודון התחלה (מתיונין). לאחר שמצא מתיונין ברצף, הריבוזום קורא את הנוקלאוטידים בשלשות, כדי להתאים חומצה אמינית כנדרש.

כתבו פונקציה בשם ***translate(rna_seq)*** המתרגמת את הרנ"א החל מקודון המתיונין הראשון. ניתן להניח שהתרגום נמשך עד השלשה המלאה האחרונה או עד לקודון העצירה הראשון (המוקדם מביניהם). יש לבחור במסגרת הקריאה הארוכה ביותר. באם מדובר ברנ"א שאינו מקודד, יש להחזיר None. באם לשתי מסגרות קריאה אותו אורך חלבון יש לבחור בראשונה מביניהן. ניתן להניח תקינות קלט. ניתן להניח קודונים סטנדרטיים.

הקודונים יוחזרו כמחרוזת בפורמט הבא:

codon1;codon2;codon3;codon4

דוגמה:

עבור הרצף `AUCAUGAACAUGCAGAUCAA` הפונקציה תחזיר

AUG;AAC;AUG;CAG;AUC

תוכנית ראשית (15 נק'):

ניתן להיעזר בתיעוד כאן בנוגע לשימוש ב-__name__ למעוניינים אך אין חובה.

על התוכנית הראשית לקבל רצף דנ"א.

1. התוכנית תבדוק אילו רצפים חוזרניים מסוג SRR קיימים ברצף (5 נק')
2. התוכנית תשעתק את הרצף לרנ"א. (5 נק')
3. התוכנית תנסה לתרגם את הרצף, באם אין אפשרות תודיע שמדובר ברצף שאינו מקודד. (5 נק')

דוגמת הרצה 1:

```
python exercise1_123456789.py TTGATCTGCATGTTCATGAT
```

פלט:

```
No simple repeats in DNA sequence

RNA sequence: AUCAUGAACAUGCAGAUCAA

Translation: AUG;AAC;AUG;CAG;AUC
```

דוגמת הרצה 2:

```
python exercise1_123456789.py ATCAAATCAAATCAAATCAA
```

פלט:

```
A,3;A,3;A,3;ATCAA,4;TCAAA,3;CAAAT,3;AAATC,3;AATCA,3;ATCAA,3

RNA sequence: UUGAUUUGAUUUGAUUUGAU

Non-coding RNA
```