פיתוח תרופות בעידן המידע

תרגיל 4

מועד הגשה: 21.01.2022

- **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 exercise4_id.py דרך תיבת ההגשה הרלוונטית במודל. ניתן להגיש מספר פעמים עד למועד ההגשה הסופי. הגרסה האחרונה היא שתיבדק.

- לדוגמה exercise4_987654321_123456789.py

על התרגיל להיות מסוגל לרוץ באופן הבא:

```
python exercise4_id1_id2.py argument1 argument2 …
```

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

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

הסבר כללי:

- בתרגיל זה נבצע שינויים מינוריים בתוכנית הראשית, ונעזר בכל מה שכתבנו עד כה כדי לבצע סימולציות שונות ולענות על שאלות ביולוגיות.
- המימוש לבחירתכם המלאה למעט היכן שמצויין במפורש.
- יש לספק את קובץ התוכנית וכן קובץ **PDF** עם מענה לשאלות.

מומלץ לקרוא את התרגיל והשאלות עד הסוף לפני תחילת המימוש כדי להבין היטב את הנדרש.

את כל הקבצים יש להגיש יחד, **לא מכווצים** (תתאפשר העלאה של מספר קבצים במודל).

כפי שמפורט בהמשך, על כל הקבצים שנוצרים מהתוכנית להיווצר באותה תיקיה כמו התוכנית ולהתחיל בפורמט `exercise4_id1_id2` ולאחר מכן המשך שם הקובץ וסיומת.

חלק 1: שדרוג התוכנית הקודמת

1. במקום קובץ עם רשימה של רצפים גנומיים, תקבל התוכנית קובץ FASTA עם רצפים גנומיים. יש לעשות שימוש בחבילה Biopython שנלמדה בתרגול לצורך קריאת הקובץ. פרטי מימוש מעבר לכך לבחירתכם. (10 נק' מתוך 40 נק' על התוכנית)

2. שנו את התא המוטנטי כך שבעת האתחול יקבל גם קצב טעויות שכפול באופן הבא:

```
def __init__(self, genome, num_mutations=0, error_rate=0.05):
```

ברירת המחדל לקצב הטעויות היא 1 ל-20, כפי שנעשה בתרגיל הקודם.

חלק 2 (40 נק'): סימולציה (הניקוד יינתן במלואו על ריצה ללא שגיאות ולפי הדרישות המודגשות)

1. בתוכנית זו נבצע סימולציה של תרבית תאים מוטנטיים עבור קצבי טעויות שונים, מספר מחזורי חלוקה שונים ולכל אחד מהרצפים שניתן בקובץ ה-FASTA (כל רצף יהווה גנום בפני עצמו). אין הגבלה על מספר התאים (כלומר הוא אינסופי).
מכיוון שישנו מרכיב רנדומי בסימולציה שלנו, נרצה לחזור על כל ניסוי **שלוש פעמים לכל קומבינציית פרמטרים** (5 נק' מתוך 40 נק' על התוכנית). ניתן לבחור לשמור את המידע מכל שלושת החזרות או רק את ממוצע הניסוי.

2. נרצה לדעת בסוף כל סימולציה את הפרטים הבאים (או ממוצע של שלוש חזרות שלהם):
   - מספר התאים הסרטניים
   - מספר התאים המוטנטיים שאינם סרטניים
   - מספר החלבונים השונים שנוצר בתרבית (באופן דומה לסעיף 3 בתוכנית הראשית מהתרגיל הקודם).

3. קצב הטעויות שנרצה לבדוק הוא בין 0.05 (1 ל-100 נוקלאוטידים) לבין 0.5 (1 ל-2 נוקלאוטידים), במקפצות של 0.05 ובנוסף את 0.01 (כלומר [0.01, 0.05, 0.1, 0.15, 0.2, …])
מספר מחזורי החלוקה שנרצה לבדוק הוא בין 1 ל-5 במקפצות של 1 (כלומר [1, 2, 3, 4, 5])

4. את המידע הנ"ל **יש לשמור בקובץ CSV**, ולהוסיף עמודות המתארות את קצב הטעות, מספר מחזורי החלוקה וכן שם הרצף (ע"פ הנתון בקובץ ה-FASTA) המתאים לאותה איטרציה בסימולציה. ניתן להוסיף מידע לבחירתכם. (10 נק' מתוך 40 נק' על התוכנית)
שם הקובץ יהיה בפורמט דומה לזה של התרגיל ויישמר באותו מיקום כמו התוכנית (על הנתיב להיות רלטיבי ולא hardcoded – כדי שיוכל לרוץ מכל מיקום ומכל מחשב). באם קיים קובץ כזה, יש לדרוס אותו.

```
exercise4_id1_id2.csv
```

5. לבסוף, יש להציג את המידע באופן גרפי.

   a. עבור מקסימום מחזורי חלוקה שנבדקו בניסוי (5), נרצה ליצור לכל אחד מהמדדים ולכל אחד מהרצפים גרף לפי הפירוט הבא:
      i. ציר ה-X הוא קצב הטעויות.
      ii. ציר ה-Y הוא אותו המדד.
      iii. שם הגרף הוא שם הרצף הנבדק.

   b. עבור כל הדטא (כל מחזורי החלוקה), נרצה לייצג את הקשר בין מספר מחזורי החלוקה וקצב הטעויות לבין מספר החלבונים הכולל הנוצר בתרבית. יש לבחור ייצוג גרפי מתאים שייתן מענה לשאלות בעניין (ראו בהמשך), ניתן לייצר מספר גרפים בהתאם לשיקול דעתכם.

יש לשמור את הגרפים באותו מיקום של התוכנית ולכלול את תעודות הזהות של המגישים בשמות הקבצים (באופן דומה לקובץ ה-CSV). ניתן לאחר מכן להוסיף את שם הרצף והמדד או כל דבר אחר מעבר לכך כדי לשייך את הגרף, לבחירתכם. (10 נק' מתוך 40 נק' על התוכנית)

יש להגיש את התוכנית של חלק זה, כולל ייצור הגרפים, בשם `exercise4_id1_id2.py`

<u>שאלות לחלק 2 (60 נק'):</u>

1. מה החשיבות של שלוש חזרות לכל קצב טעויות והאם היא זהה עבור כל המדדים? (5 נק')

צרפו את הגרפים מסעיף **5a**.

2. באיזה סוג גרף בחרתם לייצוג המידע בסעיף **5a**, ומדוע? (5 נק')
3. האם ייתכן ייצוג מוצלח יותר של המידע הנ"ל עבור כל הרצפים, כל הנתונים וכל החזרות? (5 נק')
4. תארו את השפעת קצב המוטציות על מספר החלבונים הסופי. מדוע לדעתכם זה המצב? (10 נק')
5. תארו את השפעת קצב המוטציות על מספר התאים הסרטניים ומספר התאים המוטנטיים שאינם סרטניים. האם היה ניתן להסתפק באחד מהמדדים? מדוע? (5 נק')
6. האם יש הבדל באופן ההתנהגות של מדד מספר החלבונים לכל אחד מהרצפים? הסבירו. (10 נק')

צרפו את הגרפים מסעיף **5b**.

7. באיזה סוג גרף בחרתם לייצוג המידע בסעיף **5b**, ומדוע? (5 נק')
8. תארו את השפעת מספר האיטרציות ביחס להשפעת קצב המוטציות על מספר החלבונים שנוצר. למי השפעה גדולה יותר? מדוע אלו התוצאות שהתקבלו? (15 נק')

<u>יש להגיש את התשובות של חלק זה, כולל גרפים, בקובץ PDF בשם</u> `exercise4_id1_id2.pdf`