עבודת סיום תשפ"ב, 2022 – קורס ביולוגיה חישובית (10554)

חלק א': איסוף ועיבוד מידע אודות גנום החיידק בצילוס סבטיליס

בחלק זה עליכם לעבוד עם קובץ GenBank של החיידק Bacillus subtilis. יש לעבוד עם קובץ ה GenBank הנמצא באתר הקורס (ראו חומר למידה מסוג "הנחיות והסברים על הפרויקט").

1. **הכרת וספירת האלמנטים בגנום**

גנום החיידק מכיל אזורים מסוגים שונים כגון גנים המקודדים לחלבון, גנים המקודדים לרנ"א מסוגים שונים וכן אזורים רגולטורים. דווחו כמה אלמנטים יש מכל סוג בקובץ המדובר. צרו מילון שהמפתחות שלו הם סוג האזור (למשל 'CDS', 'gene' ועוד, בהתאם לתוכן הקובץ) והערכים הם מספר המופעים.

2. **אפיון אורכי גנים**

א. עבור כל גן, חשבו את אורכו (הכוונה היא לאורך הגן ברצף בדנ"א)

ב. חלקו את הגנים לשתי קבוצות: גנים אשר מקודדים לחלבון, וכל השאר.

ג. עבור כל קבוצת גנים דווחו סטטיסטיקות אודות האורך: ממוצע, מינימום ומקסימום.

ד. כדי להבין כיצד מתפלגים אורכי הגנים, ציירו שלוש הסטוגרמות: הסטוגרמה (histogram) של אורכי כל הגנים, הסטוגרמה של אורכי הגנים המקודדים לחלבון והסטוגרמה של אורכי הגנים שאינם מקודדים לחלבון.

מה תוכלו לומר על הגרפים שהתקבלו?

3. **חישוב אחוז GC בגנים**

א. דווחו מה הוא אחוז ה-GC הממוצע בגנום החיידק (ברצף הגנום כולו).

ב. לכל גן אשר מקודד לחלבון, חשבו GC%, ודווחו מה הוא הממוצע על פני כל הגנים אשר מקודדים לחלבון.

ג. השוו את הממוצע בסעיף ב' לתוצאה מסעיף א'. האם התוצאה תואמת לציפיות שלכם? הסבירו.

ד. ציירו הסטוגרמה של GC% עבור הגנים המקודדים לחלבון.

ה. דווחו: מהם חמשת הגנים העשירים ביותר ב-GC, מהם חמשת הגנים עם הרכב ה-GC הנמוך ביותר. ציינו בדיווח פרטים כגון שם הגן, התחלה, סוף, סטרנד ו-GC%.

4. **בדיקות עקביות בקובץ הדאטה**

מטבע הדברים, כאשר עובדים עם data שלא אנחנו יצרנו, ובפרט עם מאגר מידע ביולוגי שחלקו מיוצר באופן אוטומטי, ייתכנו מצבים של מידע חסר או מידע סותר. למשל, ייתכן מצב שבו עבור רשומה של גן מסוים, רצף הדנ"א לא מתאים לרצף החלבון. במידה ומצאתן רשומות שגויות (ממגוון שיקולים שעליכן להגדיר), דווחו:

- עבור אילו גנים נמצאה סתירה ומה הסתירה. את הדיווח שמרו לקובץ gene_exceptions.csv.

הערות:

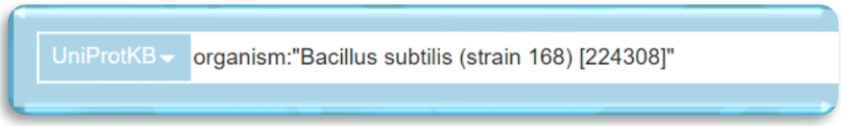
א. עבור יצירת הגרפים:
   - יש להשתמש בספרית Matplotlib או Seaborn של פיית'ון.
   - ניתן להשתמש ב-subplot כדי להציג באותו figure גרפים שמתייחסים לאותו סעיף.
   - הקפידו על אחידות של הסקאלות (של ציר x וציר y) עבור גרפים שמתייחסים לאותו מדד (למשל גרפים שהתבקשתם להציג באותו סעיף).

ב. השתמשו בחבילת pandas כדי לאסוף את המידע הנדרש בסעיפים הבאים ב-dataframe. עבור כל גן שמרו מידע אודות פרטי הגן (למשל מיקום, strand, שם), סוג הגן (מקודד לחלבון, רנ"א וכו') וכל מידע נוסף שחישבתם (למשל הרכב GC וחישובים נוספים במקרה הצורך לשיקולכם). לבסוף מיינו לפי קואורדינטת ההתחלה ושמרו לקובץ csv בשם "part_a.csv".

ג. עליכם לכתוב קוד גנרי ומודולרי. למשל, לאפשר תמיכה בכל קובץ geneBank. כלומר, יש להשקיע מחשבה בכתיבת קוד נכון לפי עקרונות של הנדסת תוכנה.

חלק ב': אנליזת חלבונים בעזרת אתר ה-UniProt

בחלק זה נעבוד עם מידע אודות חלבונים שנוריד מהאתר UniProt. ראשית, יש לשלוף מהאתר את הטבלה המתאימה לחיידק שניתחנו בחלק א'. תוכלו למצוא את הטבלה המתאימה בעזרת החיפוש הבא:



האתר מציע אוסף רחב של עמודות שניתן להוסיף לטבלת הנתונים. יש להוסיף את עמודת ה-"Transmembrane" (תוכלו למצוא אותה תחת הקטיגוריה "Subcellular location") ועמודות נוספות לפי שיקול דעתכן.

א. הצליבו בין החלבונים מקובץ ה GeneBank ובין החלבונים מקובץ ה-UniProt. כלומר, האם יש חלבונים שנמצאים בקובץ הראשון אך לא בשני (ולהפך)? כמתו את ההפרשים, הדגימו עם ויזואליזציה מתאימה. מאיפה נובעים ההבדלים (אם יש). את ההצלבה יש לבצע על סמך שמות הגנים. ציינו מה שם העמודה ב-UniProt שהשתמשתם בה עבור ההצלבה.

ב. שלפו את הרצפים הטרנסממברנליים (המידע על כך נמצא בעמודה Transmembrane) מתוך רצפי החלבונים. שימו לב, לא לכל חלבון יש איזור טרנסממברנלי, ולחלק מהחלבונים יש יותר מאזור אחד כזה. מדובר באזורים קצרים יחסית. אפיינו את הרצפים הללו:

- מה התפלגות האורכים שלהם (ציירו הסטוגרמה), מה האורך הממוצע, המינימלי, והמקסימלי.
- מה התפלגות אחוז חומצות האמינו ההידרופוביות ברצפים האלה, מה הערך הממוצע על פני כל הרצפים הללו? האם זה תואם לציפיות שלכן מאזורים כאלה?

ג. נסמן את קבוצת הגנים שהם CDS באות A. עבור רצפי הגנים שנמצאו בחיתוך בין ה UniProt ובין ה GenBank, אתרו את הגנים שמכילים לפחות אזור טרנסממברנלי אחד. נסמן קבוצת גנים זו ב-B.

- מה התפלגות GC% ברצפי הגנים בקבוצה B?
- סכמו בטבלה את הסטיסטיקות עבור קבוצת גנים אלא בהשוואה לקבוצה A. (שחישבתם בחלק א'): ממוצע, חציון, מינימום ומקסימום. בנוסף, ציירו את ההתפלגות של GC% (היסטוגרמה) בשתי קבוצות הגנים על אותו גרף עם ארבעה

חלקים (השתמשו ב subplot של ספרית Matplolib של פיית'ון) עבור קבוצות הגנים הבאות:

i. קבוצה A
ii. קבוצה B
iii. הגנים ב-A שאינם ב-B
iv. באותו גרף הקבוצוצ מסעיף ii לעומת iii.

הקפידו על אחידות של הסקאלות (של ציר x וציר y) עבור כל הגרפים, וכן על שמות אינפורמטיביים לצירים ולכותרות.

חלק ג': אנליזה מנקודת מבט אבולוציונית - וירוסים

1. עבור הקוד הגנטי המתאים לוירוס הקורונה, חשבו עבור כל קודון כמה עמדות הן סינונימיות. דווחו את התוצאה בעזרת מילון שהמפתחות שלו הם הקודונים השונים והערכים הם המספר המתאים.
2. הורידו את קובץ ה GenBank של וירוס הקורונה מיולי 2020 (accession number: **NC_045512.2**) והשוו אותו לוירוס הקורונה שבודד בינואר 2022 (accession number: **LC666924.1**).
   א. כמה גנים משותפים יש ביניהם? האם יש גנים שיש באחד ולא באחר?
   ב. בחרו חמישה גנים משותפים וחשבו עבור כל גן את מדד ה-dnds. דווחו בטבלה את פרטי הגנים שבחרתם (למשל שם, תפקיד ופרטים נוספים), את תוצאות ה- dnds וכן האם התרחשה בגן זה סלקציה חיובית, ניטרלית או שלילית.

הוראות הגשה:

- את התשובות לשאלות המילולית כתבו בקובץ word בשם final_project.docx
  - השתמשו בפונט אריאל, גודל 11, שורה וחצי, עד 3 עמודים של מלל (לא כולל גרפים וטבלאות)
  - בראש העמוד הראשון כתבו שמות + מספרי ת"ז של המגישים
  - בנוסף, הסבירו בקצרה לגבי שיקולים שעשיתם בכתיבת הקוד (הסבר קצר על המחלקות/סקריפטים שכתבת ודברים אחרים שחשוב לדעתכם לציין)
- הגישו קובץ zip בשם: final_project.zip המכיל:
  - תיקייה עם הקוד שמימשתם
  - קבצי דאטה שיצרתם (למשל part_a.csv וקבצים נוספים)
  - final_project.docx
  - קובץ README.docx עם הוראות הרצה
- לאחר ההגשה יערך מבחן בע"פ על הפרויקט שהגשתם ועל נושאים נוספים שנלמדו לאורך הקורס.
- העבודה היא בצוותים, אך יש להגיש רק ממשתמש אחד באתר הקורס. הקפידו על עבודה עצמאית בצוותים, עבודות דומות של צוותים שונים יפסלו.

נספח: חלוקה של חומצות אמינו לסוגים שונים

Amino Acids

Hydrophobic amino acids:

Name	Code	Name	Code
Alanine	Ala	Valine	Val
Phenylalanine	Phe	Methionine	Met
Leucine	Leu	Proline	Pro
Isoleucine	Ile	Tryptophane	Trp

Hydrophilic amino acids:

Name	Code	Name	Code
Glycine	Gly	Threonine	Thr
Serine	Ser	Cysteine	Cys
Tyrosine	Tyr	Asparagine	Asn
Glutamine	Gln	Arginine	Arg
Lysine	Lys	Histidine	His
Aspartic acid	Asp	Glutamic acid	Glu