עבודת סיום תשפ"ד, 2024 – קורס ביולוגיה חישובית (10554)

תאריך הגשה: 28.03.24

יש להרשם כצוותים עד 07.03.24 בקישור זה.

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

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

1. **הכרת וספירת האלמנטים בגנום**

גנום החיידק מכיל אזורים מסוגים שונים (למשל, גנים המקודדים לחלבון, גנים המקודדים לרנ"א tRNA, גנים המקודדים ל rRNA, ועוד). דווחו כמה אלמנטים יש מכל סוג בקובץ (מצאו את כל הסוגים הקיימים). צרו מילון שהמפתחות שלו הם סוג האזור (למשל 'CDS', 'gene', ועוד, בהתאם לתוכן הקובץ) והערכים הם מספר ההופעות.

2. **אפיון אורכי הגנים בין שני ה strands**

א. עבור כל גן, חשבו את אורכו (הכוונה היא לאורך הגן ברצף הדנ"א) וציירו הסטוגרמה (histogram) של אורכי הגנים.

ב. חלקו את הגנים לשתי קבוצות: (1) גנים אשר נמצאים על ה plus strand ו-(2) גנים אשר נמצאים על ה minus strand. כמה גנים יש בכל קבוצה? כמה גנים בכל קבוצה אשר מקודדים לחלבון? (type='CDS')

ג. עבור הגנים שמקודדים לחלבון, ציירו 2 היסטוגרמות עבור אורכי הגנים, אחת לכל קבוצה (כלומר, עבור הגנים שעל הסטרנד החיובי ועבור הגנים שעל הסטרנד השלילי). האם ההתפלגויות דומות?

ד. הציגו בטבלה ודווחו לכל קבוצה סטטיסטיקות אודות האורך: מינימום, מקסימום, ממוצע וסטיית תקן.

ה. חזרו על סעיפים ג' ו-ד' עבור הגנים שאינם מקודדים לחלבון.

ו. מה תוכלו לומר על ההבדלים בין הסטרנדים? מה תוכלו לומר על ההבדלים בין הגנים שמקודדים לחלבון ואילו שאינם מקודדים לחלבון?

3. **חישוב אחוז CT בגנים**

א. דווחו מה הוא אחוז ה-CT הממוצע בגנום החיידק (<u>ברצף הגנום כולו</u>) – כלומר כמה אותיות הן C או T מתוך אורך הגנום כולו.

ב. לכל גן אשר מקודד לחלבון, חשבו CT%, 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 מבחינת מתמטית? הסבירו.

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

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

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. דוגמאות עבור רשומה של גן מסוים, ייתכן שרצף הדנ"א לא מתאים לרצף החלבון, או שחסר מידע אודות רצף החלבון. חשבו על מקרים אפשריים נוספים. במידה ומצאתם רשומות שגויות (ממגוון שיקולים שעליכן להגדיר), דווחו:

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

הערות:

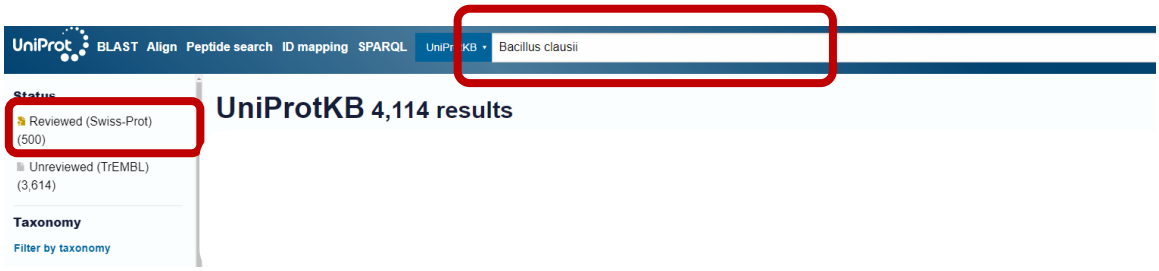
א. עבור יצירת הגרפים:
- יש להשתמש בספרית Matplotlib או Seaborn של פיית'ון.
- ניתן להשתמש ב-subplot כדי להציג באותו figure גרפים שמתייחסים לאותו סעיף.
- הקפידו על אחידות של הסקאלות (של ציר x וציר y) עבור גרפים שמתייחסים לאותו מדד (למשל גרפים שהתבקשתם להציג באותו סעיף).
- הוסיפו labels לצירים.

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 CT 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 genBank. כלומר, יש להשקיע מחשבה בכתיבת קוד נכון לפי עקרונות של הנדסת תוכנה. הסבירו את הדיזיין של התוכנה שלכם ואת ההגיון שעומד מאחוריו בקובץ README.docx.

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 GenBank 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. שימו לב שב UniProt מופיעים לעיתים מספר שמות עבור כל שורה. ציינו מה שם העמודה ב-UniProt שהשתמשתם בה עבור ההצלבה (UniProt מציע עמודות שונות עם שמות גנים, אין תשובה אחת בלבד נכונה, זה חלק מהמחקר של שאלה זו, ציינו איזו עמודה/עמודות בחרתן ולמה).

ב. שלפו את הרצפים הטרנסממברנליים (המידע על כך נמצא בעמודה Transmembrane) מתוך רצפי החלבונים. שימו לב, לא לכל חלבון יש אזור 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 (ראו נספח בסוף המסמך הזה)? מה הערך הממוצע על פני כל הרצפים הללו? האם זה תואם לציפיות שלכן מאזורים כאלה? הסבירו וציינו באילו מקורות מידע נעזרתן (חפשו מידע אודות המשמעות של Transmembrane וצטטו את המקורות הרלוונטיים).

חלק ג': אנליזה מנקודת מבט אבולוציונית - וירוסים

1. עבור הקוד הגנטי המתאים, 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 מאפריל 2021 (accession number: MZ054892.1) והשוו אותו לוירוס הקורונה שבודד בפברואר 2024 (accession number: PP348372.1).
   א. מה אורך כל גנום?
   ב. כמה גנים יש בכל אחד מהם? מתוכם כמה גנים מקודדים לחלבונים?
   ג. כמה גנים משותפים יש ביניהם (הסתמכו על שמות גנים)? האם יש גנים שיש באחד ולא באחר? אם כן, פרטו את רשימת השמות.
   ד. 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, רווח 1.5 שורות, **עד 3 עמודים של מלל** (לא כולל גרפים וטבלאות). את הגרפים ואת הטבלאות יש לצרף **בתוך כל שאלה** ולא בנספחים.
  - בתחילת 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 והסברים נוספים
  - קובץ ה data הרלוונטי לחלק ב' (הטבלה מאתר UniProt)
- 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.
- הקפידו על הגהה לעבודה שאתם מגישים, ירדו נקודות עבור משפטים לא ברורים (אם לא אבין למה התכוונתם, אניח שזו טעות).
- בהצלחה!

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