

עבודת סיום תשפ"ג, 2022-2023 – קורס ביולוגיה חישובית (10554)

תאריך הגשה: 15.1.23

בשיעור האחרון (19.1.23) יערכו מבחנים בע"פ עבור כל צוות בנפרד על הפרויקט ועל הנושאים שנלמדו במהלך הסמסטר.

יש להרשם צוותים עד **27.12.22** [בקישור זה](#).

חלק א': איסוף ועיבוד מידע אודוט גנום החידק בצלוס סבטיליס

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

1. הכרת וספרת האלמנטים בגנום

גנום החידק מכיל אזורים מסוימים שונים כגון מקודדים לחלבון, גנים המקודדים לרנ"א מסוימים וכן אזורים רגולטוריים. דוווחו כמה אלמנטים יש מכל סוג בקובץ המדובר. צרו מילון שהמפתחות שלו הם סוג האזור (למשל 'gene', 'CDS', 'gene' ועוד, בהתאם לתוכן הקובץ) והעריכים הם מספר המופיעים.

2. אפיון אורכי גנים

- עבור כל גן, חשבו את אורכו (הכונה היא לאורך הגן ברכף בדנ"א).
- חלקו את הגנים לשתי קבוצות: גנים אשר מקודדים לחלבון, וכל השאר.
- עבור כל קבוצה גנים דוווחו סטטיסטיות אודוט האורך: מינימום, מקסימום, ממוצע וסטיית תקן.
- כדי להבין כיצד מתפלגים אורכי הגנים, צירו שלוש הסטוגרמות (histogram): הסטוגרמה של אורכי כל הגנים, הסטוגרמה של אורכי הגנים המקודדים לחלבון והסטוגרמה של אורכי הגנים שאינם מקודדים לחלבון.
מה תוכלו לומר על הגرافים שהתקבלו?

3. חישוב אחוז AT בגנים

- דוווחו מה הוא אחוז ה-TAT הממוצע בגנים החידק (ברצף הגנים כולו).
- לכל גן אשר מקודד לחלבון, חשבו AT%, ודוווחו מה הוא הממוצע על פני כל הגנים אשר מקודדים לחלבון.

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

ד. ציירו הסטוגרמה של AT% עבור הגנים המקודדים לחלבון.
ה. דוחו: מהם חמשת הגנים העשירים ביותר ב-AT%, מהם חמשת הגנים עם הרכיב ה-AT הנמוך ביותר. צינו במידוח פרטים כגון שם הגן, התחלת, סוף, סטרנד ו-.%AT.

4. גנים המכילים בתיאור wall cell

אחד השדות בקובץ GenBank הוא תיאור הפונקציה של הגן/ החלבון (ליתר דיוק, החלבון שמקודד ע"י הגן).

א. מצאו את הגנים המכילים בתיאור את צמד המיללים "wall cell" (ביחד). כמה גנים מצאתם?

ב. עבור הגנים שמצאתם בסעיף א' חשבו את אורכי הגנים, דוחו סטטיסטיות (מינימום, מקסIMUM, ממוצע וסטיית תקן) וציירו הסטוגרמה.

ג. עבור הגנים שמצאתם בסעיף א' חשבו את AT% ברכפי הגנים, דוחו סטטיסטיות (מינימום, מקסIMUM, ממוצע וסטיית תקן) וציירו הסטוגרמה.

הערה: העזרו בתוצאות שהיחסתם בשאלות הקודמות, אין צורך לחשב את סעיפים ב' ו-ג' מחדש.

5. בדיקות עקביות בקובץ הדטה

מטבע הדברים, כאשר עובדים עם data שלא אנחנו יוצרים, ובפרט עם מאגר מידע ביולוגי שחקקו מיוצר באופן אוטומטי, יתכונו מצבים של מידע חסר או מידע סותר. למשל, יתכן מצב שבו רשומה של גן מסוים, רצף הדנ"א לא מתאים לרצף החלבון. במידה ומצאתן רשומות שגויות (מגנון שיקולים שעליין להגדר), דוחו:

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

הערות:

- א. עבור יצירת/grafim:
- יש להשתמש בספרית Matplotlib או Seaborn של פית'ון.
 - ניתן להשתמש ב subplot כדי להציג באותו figureグラפים שמתיחסים באותו סעיף.
 - הקפידו על איחיות של הסקלאות (של ציר x וציר y) עבור גרפים שמתיחסים לאותו מודד (למשל גרפים שהתבקשו להציג באותו סעיף).
 - הוסיפו labels לצירים.
- ב. השתמשו בחבילת pandas כדי לאוסף את המידע הנדרש בסעיפים הבאים ב-`dataframe`. עבור כל גן שמרו מידע אודות פרטיה הגן (למשל מיקום, strand, שם), סוג הגן (מקודד לחלבון, רנ"א וכו') וכל מידע נוסף שהישבთם (למשל הרכב AT ווחישובים נוספים במקורה הצורך לשיקולכם). לבסוף מינו לפי קואורדינטת ההתחלה ושמרו לקובץ csv בשם "part_a.csv".
- ג. עלייכם לכתוב קוד גנרי ומודולרי. למשל, לאפשר תמייה בכל קובץ genBank. כלומר, יש להשיקיע מחשבה בכתיבה קוד נכון לפי עקרונות של הנדסת תוכנה.

חלק ב': אналיזת חלבונים בעזרת אתר UniProt

בחלק זה נעבד עם מידע אודות חלבונים שנוריד מהאתר UniProt. ראשית, יש לשלוּף מהאתר את הטבלה המתאימה לחידק שניתנו בחלק א'. תוכלו למצוא את הטבלה המתאימה בעזרת החיפוש הבא:

The screenshot shows the UniProtKB search interface. At the top, there's a navigation bar with links for BLAST, Align, Peptide search, ID mapping, SPARC, UniProtKB (selected), and Bacillus subtilis. Below the navigation bar, a status section shows 'Reviewed (Swiss-Prot) (4,191)' and 'Unreviewed (TrEMBL) (4,347)'. The main title 'UniProtKB 8,538 results' is displayed. Underneath it, there are several buttons: BLAST, Align, Map IDs, Download, Add, View: Cards, Table, Customize columns, and Share. A 'Popular organisms' dropdown menu is open, with 'B. subtilis (8,538)' highlighted by a red box. Other options in the dropdown include Escherichia coli, Saccharomyces cerevisiae, and Homo sapiens. At the bottom left, there's a 'Taxonomy' link.

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

א. הצלבו בין החלבונים מקובץ GeneBank ובין החלבונים מקובץ UniProt. קלומר, האם יש חלבונים שנמצאים בקובץ הראשון אך לא בשני (ולהפר?) כמוთ את ההפרש, הדגימו עם ייזואלייזציה מתאימה. מאיפה נובעים ההבדלים (אם יש). את ההצלבה יש לבצע על סמך שמות הגנים. שימו לב שב UniProt מופיעים לעתים מספר שמות עבור כל שורה. ציינו מה שם העמודה בUniProt שהשתמשה בה עבור ההצלבה (UniProt מציע עמודות שונות עם שמות גנים, אין תשובה אחת בלבד נכון, זה חלק מהמחקר של שאלה זו, ציינו איזו עמודה/עמודות בחרתן ולמה).

ב. שלפו את הרצפים הטרנסמברנליים (הידוע על קר נמצא בעמודה Transmembrane) מתוך רצפי החלבונים. שימו לב, לא לכל חלבון יש אזור טרנסמברנלי, ולחלק מהחלבונים יש יותר מאשר אחד צזה. מדובר באזורי קצרים יחסית. אפיינו את הרצפים הללו:

- מה התפלגות האורכים שלהם (צייר הסטוגרמה), מה האורך הממוצע, המינימלי, והמקסימלי.
- מה התפלגות אחוז חומצות האmino הhidroforeיות ברכפים האלה, מה הערך הממוצע על פני כל הרצפים הללו? האם זה توأم לציפויו של כן מאזוריים כאלה? הסבירו וציינו באילו מקורות מודיע נעזרת (חפשו מידע אודות המשמעות של Transmembrane וצטטו את המקורות הרלוונטיים).

ג. נסמן את קבוצת הגנים שהם CDS באוט A. עבור רצפי הגנים שנמצאו בחיתוך בין ה- ProtiofUn ובין ה- GenBank, אתרו את הגנים שמקילים לפחות אוצר טנסומברני אחד. נסמן קבוצת גנים זו ב-B.

- מה התפלגות AT% ברצפי הגנים בקבוצה B?

• סכמו בטבלה את הסטטיסטיות עבור קבוצת גנים אלא בהשוואה לקבוצה A. (שחישבתם בחלק א'): מינימום, מקסימום, ממוצע וסטיית תקן. בנוסף, צירו את ההתפלגות של AT% (הסתוגרמה) בשתי קבוצות הגנים על אותו גרף עם ארבעה חלקים (השתמשו ב- subplot של ספרית Matplotlib של פית'ון) עבור קבוצות הגנים הבאות:

i. קבוצה A

ii. קבוצה B

iii. הגנים ב-A שאינם ב-B

iv. באותו גרף הקבוצות מסעיף ii לעומת זאת iii (בשני צבעים שונים)

הקפידו על איחדות של הסקלולות (של ציר x וציר y) עבור כל הגרפים, וכן על שמות אינפורטטיביים לצירים ולכותרות.

חלק ג': אנליזה מנוקדת מבט אבולוציונית - וירוסים

1. עבור הקוד הגנטי המתאים, חשבו עבור כל קודון כמה עמדות הן סינונימיות. דוווחו את התוצאה בעזרת מילון שהמפתחות שלו הם הקודונים השונים והעריכו כמה המספר המתאים.

2. הורידו את קובץ ה- GenBank של וירוס הקורונה מינואר 2021 (accession number: MZ383039.1) והשו אותו לוירוס הקורונה שבודד בדצמבר 2022 (accession number: OQ065689.1).

א. כמה גנים יש בכל אחד מהם? מתוכם כמה גנים מקודדים לחלבונים?
ב. כמה גנים משותפים יש ביניהם (הסתמכו על שמות גנים)? האם יש גנים שיש באחד ולא באחר? אם כן, פרטו את רישימת השמות.

ג. בחרו חמישה גנים משותפים וחשבו עבור כל גן את מדרד-hspsd. דוווחו בטבלה את פרטי הגנים שבחרתם (למשל שם, תפקיד ופרטים נוספים), את תוצאות ה- hspsd וכן האם התרחשה בגן זה סלקציה חיובית, ניטרלית או שלילית.

הוראות הגשה:

- את התשובות לשאלות המילולית כתבו בקובץ word בשם **final_project.docx** בשם **final_project.docx** בשם **final_project.zip** המכיל:
 - השתמשו בפונט Arial, גודל 11, רווח 1.5 שורות, **עד 3 עמודים של מיל** (לא כולל גרפים וטבלאות). את הגрафים ואת הטבלאות יש לצרף בתוך כל שאלה ולא בנספחיהם.
 - בתחילת העמוד הראשון כתבו שמות + מספרי ת"ז של המתאים.
 - בנוסף, הסבירו בקצרה לגבי שיקולים שעשיתם בכתיבת הקוד (הסביר קצר על המחלקות/סקרייפטים שככבות ודברים אחרים שחשוב לדעתכם לציין)
- הגיעו קובץ压缩文件名为 final_project.zip
- תיקייה עם הקוד שמיימשתם
- קבצי דата שיצרתם (למשל **a.csv** וקבצים נוספים)
- **final_project.docx**
- קובץ **README.docx** עם הוראות הריצה
- קובץ **data** הרלוונטי לחלק ב' (הטבלה מאתר UniProt)
- לאחר ההגשה יערוך מבחון בע"פ על הפרויקט שהगשתם ועל נושאים נוספים שנלמדו לאורך הקורס.
- העבודה היא בצדדים, אך יש להגיש רק משתמש אחד באתר הקורס. הקפideo על עבודה עצמאית בצדדים, עבודות דומות של צוותים שונים יפסלו.
- הקפideo על הגהה לעבודה שאתם מגישים, ירדו נקודות עבור משפטים לא ברורים (אם לא אבין למה ה提到了你们的项目，我将扣分).
- בהצלחה!

נספח: חלוקה של חומצות אמינו לשוגים 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 |