מבוא לגנומיקה חישובית ומערכתית - challenge:

מטרה - לבנות פרדיקטור שחוזה את רמות חלבון של ספריית וריאנטים של אופרונים סינתטיים (12900 וריאנטים). את רמות החלבון נחזה על סמך רצף הוריאנטים וכן חלוקתם לבינים.

תיאור המשימה–

1. הוריאנטים בנויים בצורה הבאה:

CTGCTGGTTCTGGCGAATAGACTAGTXXXXXXXXXXXXXXXXXXXXXXXXAAGGGCGAGGAGCTCTTTAC

חלק מגן ה-RFP, רצף קבוע ACTAGT, רצף רנדומי באורך של 24 נוקליאוטידים וחלק מגן ה-GFP.

הוריאנטים הם למעשה ספרייה סינתטית של אופרונים שמטרתה הייתה לבחון re-initiation של תהליך התרגום באופרונים בפרוקריוטיים. להסברים נוספים ניתן לפנות לשיעור 5 וכן לתרגולים 4 ו-5.

2. הוריאנטים חולקו לפי רמות החלבון לבינים. כאשר bin_num=~500 ומתוך כמות הבינים נגזרה כמות הוריאנטים בכל בין.

3. לאחר מכן, הבינים חולקו בצורה רנדומית לסט נתונים ידוע וסט נתונים לא ידוע.

מצורפים לכם הקבצים הבאים:

קובץ ראשון – מכיל את הנתונים הבאים:
- אינדקס של כל וריאנט.
- הרצף של כל וריאנט.

קובץ שני (known SET) - מכיל את הנתונים הבאים עבור 80% מהבינים:
- אינדקס הבין.
- אינדקס של כל הוריאנטים שנמצאים בבין (האינדקסים תואמים לאקסל של רצפי הוריאנטים).
- רמות חלבון ממוצעות של הבין (לאחר החישוב שהוסבר בכיתה).
- פיצ'רים ממוצעים מסוימים (פירוט למטה).

קובץ שלישי (unknown SET) - מכיל את הנתונים הבאים עבור 20% מהבינים:
- אינדקס הבין.
- אינדקס של כל הוריאנטים שנמצאים בבין (האינדקסים תואמים לאקסל של רצפי הוריאנטים).
- פיצ'רים ממוצעים מסוימים (פירוט למטה).

הקבצים מכילים לכל בין את הפיצ'רים הבאים שיצרנו עבורכם:

1. **הפיצ'ר max RBS** – לכל וריאנט חושבה אנרגיית ההיברידיזציה בין רצף ה-SD באזור הרנדומי והרצף המשלים ברנ"א הריבוזומלי. לאחר מכן נלקחה האנרגיה המקסימלית (הכי שלילית) לכל וריאנט וממוצע האנרגיה המקסימלית חושבה לכל בין.
קישורים: https://en.wikipedia.org/wiki/Shine-Dalgarno_sequence ונספחים.

2. **הפיצ'ר Average RBS** – לכל וריאנט חושבה אנרגיית ההיברידיזציה בין רצף ה-SD באזור הרנדומי והרצף המשלים ברנ"א הריבוזומלי. לאחר מכן נלקחה האנרגיה הממוצעת (בכל החלונות) לכל וריאנט וממוצע האנרגיה הממוצעת חושבה לכל בין.
קישורים: https://en.wikipedia.org/wiki/Shine-Dalgarno_sequence ונספחים.
3. **הפיצ'ר Total fold** - אנרגיית הקיפול הכוללת של כל וריאנט
קישורים:https://www.nature.com/articles/nrg3681 ונספחים.

4. **הפיצ'ר CAI** - מידה אשר משערכת את רמת - codon bias של כל וריאנט בחלק הרנדומי. גנים עם ערך CAI גבוה נוטים להיות גנים שבאים לידי ביטוי בצורה רבה יותר.
קישורים : https://en.wikipedia.org/wiki/Codon_Adaptation_Index ונספחים.

5. **הפיצ'ר AAA codon count** - כמות ההופעות הכללית של הקודון AAA בחלק הרנדומי.

6. **הפיצ'ר GCA codon count** - כמות ההופעות הכללית של הקודון GCA בחלק הרנדומי.

7. **הפיצ'ר TTT codon count** - כמות ההופעות הכללית של הקודון TTT בחלק הרנדומי.

8. **הפיצ'ר FE window 1** - אנרגיית קיפול של הוריאנט בחלון הראשון ברצף (30 נוקליאוטידים ראשונים).
קישורים:https://www.nature.com/articles/nrg3681 ונספחים.

9. **הפיצ'ר FE window 2** - אנרגיית קיפול של הוריאנט בחלון הראשון ברצף (30 נוקליאוטידים הבאים, חלון נע של נוקליאוטיד 1).
קישורים:https://www.nature.com/articles/nrg3681 ונספחים.

10. **הפיצ'ר FE window 3** - אנרגיית קיפול של הוריאנט בחלון הראשון ברצף (30 נוקליאוטידים הבאים, חלון נע של נוקליאוטיד 1).
קישורים:https://www.nature.com/articles/nrg3681 ונספחים.

משימות -

א. סעיף חימום: צרו את הפיצ'רים הבאים בעצמם -
הפיצ'ר AAA codon count, GCA codon count, TTT codon count.
ודאו כי הפיצ'רים זהים לפיצ'רים אשר נתונים לכם.

ב. סעיף חימום: צרו את הפיצ'רים הבאים בעצמם -
הפיצ'ר FE window 1, FE window 2, FE window3.
ודאו כי הפיצ'רים קורלטיביים לפיצ'רים אשר נתונים לכם (נסתכל על קורלציה כיוון שישנן גרסאות שונות לתוכנת חישוב הקיפול).

על מנת לחשב פיצ'רים אלו אנחנו נשתמש בתוכנה שנקראת ויאנה (אלגוריתם לתכנות דינמי שמוצא את האנרגיית קיפול על בסיס אפטום האנרגיה החופשית של המבנה השניוני של ה-mRNA) (http://rna.tbi.univie.ac.at) הורידו את הקובץ RNAfold ה-executable שלמעשה מאפשר הרצה של הקוד על המחשב שלכם.

- זכרו לחלק את הרצף לחלונות כפי שנלמד בכיתה.
- המשיכו לחשב את האנרגיה עבור כל החלונות האפשריים.
- על מנת להריץ את האלגוריתם יש להשתמש בפקודה הבאה:

```
for j=1:size(seq,1)
   system(['echo "' seq{j} '" >> /Users/shirelitzur/vien/sequences']);
end
```

חלק זה למעשה יוצר קובץ שמכיל את הרצפים שלהם תרצו לחשב את האנרגיה, ניתן לחשב יחד את האנרגיה לכמה רצפים שתרצו – שימו לב לשנות את הנתיב לפי המחשב שלכם.

```
[status,cmdout] = system(['/Users/shirelitzur/vien/RNAfold < /Users/shirelitzur/vien/sequences']);
```

זוהי הפקודה אשר מריצה את הפונקציה, התשובה תהיה במשתנה cmdout - שימו לב לשנות את הנתיב לפי המחשב שלכם.

- דגש חשוב הוא שיש להוציא את ערך האנרגיה מהמשתנה cmdout (כלומר לפרסר את מה שמתקבל).
- דגש חשוב שהפקודה echo שיוצרת את הקובץ מוסיפה רצפים לקובץ לכן אם הייתה לכם טעות, יש למחוק את הקובץ שנוצר ולייצר קובץ חדש (אחרת הרצפים החדשים יצטרפו לקודמים) , או להשתמש בפקודה הבאה:

```
[status1,cmdout1] = system(['rm "/Users/shirelitzur/vien/sequences""']);
```

ג. צרו פיצ'רים נוספים כראות עיניכם (כפי שהוסבר והודגם בכיתה).

ד. על סמך קובץ ה- known set צרו רגרסור (או כל פרדיקטור אחר) המשתמש בחלק מהפיצ'רים השונים על מנת לחזות רמות חלבון ממוצעות של כל בין. שימו לב ל-over fitting.
אפשר להשתמש בשיטה שהוצגה בכיתה לבחירת פיצ'רים או בכל שיטה אחרת.

ה. חשבו את קורלציית ספירמן בין רמות הביטוי של הבינים של דאטא האימון של סט הולדיציה עם תוצאות הרצת הרגרסור על דאטא האימון של סט הולידציה והציגו תוצאה זו במסמך ההגשה.

ו. הריצו את הרגרסור על הדאטא של ה-unknown set על מנת לחזות את רמות החלבון של כל אחד מהבינים.

תיאור הגשה -

1. צרו מסמך PDF אשר יכיל:

- הסבר על הרגרסור/פרדיקטור - איזה רגרסור/פרדיקטור בחרתם? למה בחרתם אותו? האם ניסיתם רגרסורים/פרדיקטורים אחרים קודם לכן, ואם כן מדוע לדעתם הם הצליחו פחות?
- תרשים זרימה שמתאר את פעולת הרגרסור/פרדיקטור.
- פיצ'רים -האם יצרתם פיצ'רים חדשים?(אם כן, איזה פיצרים יצרתם) מדוע?, איזה פיצ'רים נבחרו בסוף?, מדוע לדעתכם אילו הפיצ'רים שנבחרו ולא אחרים?
- דיון - מהי מידת ההצלחה שהפיצ'רים שלכם עם הרגרסור/פרדיקטור חוזים את רמות הביטוי. האם ואיך ניתן לשפר את הרגרסור/פרדיקטור שבניתם? מה הן נקודות החולשה של הרגרסור/פרדיקטור שלכם? מה היתרונות של הרגרסור/פרדיקטור שלכם? האם ישנם פיצ'רים שידעתם שחשובים לשים ברגרסור/פרדיקטור, אם כן, מדוע?

2. צרפו קובץ מטלב/אקסל שמכיל את ערכי החלבון שחזיתם עבור כל הבינים בחלוקות השונות של ה-unknown set.

3. צרפו את הקוד של הרגרסור/פרדיקטור כולל מסמך READ ME שמסביר את אופן ההרצה בצורה מפורטת וברורה. קוד שלא ירוץ לא ייבדק!

<u>נספח --על הפיצ'רים שיצרנו:</u>

תהליך התרגום בפרוקריוטים:

ההבדל המהותי בתהליך התרגום של פרוקריוטים ואאוקריוטים הוא שלב האתחול (initiation). רצף ה-(SD) shine dalgarno הוא אזור קשירה של רנ"א ריבוזומלי (RBS) ל mRNA בחיידקים וארכיאות, נמצא בערך 8-12 נוקליאוטידים לפני קודון ההתחלה. מהיחידה הקטנה של הריבוזום מציץ זנב של רנ"א ריבוזומלי אשר נקשר בהיברידיזציה ל mRNA באזור ה-SD . תפקיד ה-SD (RBS) הוא לאפשר ליחידה הקטנה של הריבוזום למצוא את קודון ההתחלה ולהתחיל את תהליך התרגום. כלומר, צפויה אנרגיית היברידיזציה חזקה (שלילית) בין ה- mRNA לרנא הריבוזומלי לפני קודון ההתחלה.

CAI-codon adaptation index:

זוהי מידה שבוחנת את האופטימליות של הקודונים. ה-CAI מודד את הדרגה שבה הגנים משתמשים בקודונים מועדפים.
ראשית מחשבים משקל עבור כל קודון בסט רפרנס -גנים שבאים לידי ביטוי בצורה חזקה (highly expressed genes). כיוון שאלו הם גנים שמבוטאים יותר כנראה שהם משתמשים בקודונים יותר אופטימליים. משקל גבוה משמע שהקודון יותר אופטימלי,ערך המשקולת המקסימלית הוא 1.מערכי המשקל של הקודונים ניתן לזהות את הקודונים שמשתמשים בהם יותר עבור כל חומצה אמינית. למעשה המשקולות מחושבות על ידי התדירות של כל קודון עבור חומצה אמינית מסויימת חלקי התדירות המקסימלית של הקודונים הסינונימים שמקודדים לאותה חומצה אמינית. נשים לב כי צריך לבחון מהו סט הרנפרנס, בבחירות שונות של סט זה ניתן לקבל תוצאות קצת שונות. ערך יותר גבוה של משקל משמע שקודון זה נוטה להיות יותר תדיר לכן במובן מסויים של ביטוי גנים כנראה שהוא יותר אופטימלי. לאחר שחישבנו את המשקולות ניתן לחשב את ערך ה-CAI עבור כל גן. למעשה, מחשבים ממוצע גיאומטרי עבור המשקולות, כיוון שכל המשקולות הן בין 0 ל-1 נקבל ערך גם כן בין 0 ל-1. נבחין כי בגנים שהם highly expressed יש יותר bias לקבוצה קטנה יותר של קודונים ופילוג השימוש בקודונים הוא פחות יוניפורמי. בגנים שהם lowly expressed קיים bias אך בתבניות יותר חלשות, פילוג הקודונים יותר יוניפורמי.

יציבות הקיפול של mRNA:

קיפול של mRNA באזורים שונים משפיע על יעילות של ביטוי גנים. ה mRNA עובר מבחינה תרמודינמית לקיפול בו רמת האנרגיה היא מינימלית. רנ"א מסונתז כמולקולה חד גדילית. זיווג בסיסים מפיק מבנה שניוני. את היציבות של המבנה השניוני ניתן לכמת באמצעות האנרגיה החופשית שמשתחררת או שמשתמשים בה ליצירתו. אנרגיה חופשית חיובית משמע שיש צורך להשקיע עבודה על מנת ליצור את המבנה. אנרגיה חופשית שלילית משחררת מבנה קיים. ככל שערך האנרגיה החופשית של המבנה יותר שלילי יש סיכוי גבוה יותר למבנה זה להיווצר, כיוון שיותר אנרגיה אגורה משתחררת. ניתן לחשב למולקולה או לאזור במולקולת mRNA מה הקיפול הכי חזק שמשחרר את האנרגיה הרבה ביותר -אנרגיה חופשית. ערך שלילי משמע שצריך להשקיע אנרגיה כדי לפתוח את הקיפול. נצפה לקבל ערך שלילי. ניתן לעשות פרדיקציה לוקלית של קיפול mRNA, הפרדיקציה נותנת גם מהי האנרגיה החופשית שצריך להשקיע כדי לפתוח את הקיפול, יותר שלילית משמע קיפול יותר חזק. ניתן לחשב את האנרגיה החופשית בחלון מסויים ולהשוות אותו לרנדום -למשל אם עשינו פרמוטציות לקודונים. כאשר הקיפול של ה- mRNA חזק יותר - אנרגיה שלילית יותר כך ייתכן ויהיה קושי לריבוזום לעבור על ה- mRNA ולבצע את תהליך התרגום וכך נפגע ברמת הביטוי של הגן.