מבוא לגנומיקה חישובית ומערכתית - challenge:

מטרה - לבנות פרדיקטור שחוזה את רמות חלבון של גנים על סמך הרצף של הרנ"א שליח שלהם (האיזור המקודד וה UTR'5) על סמך known SET. ולדרג את הגנים לפי רמות החלבון שלהם ב unknown SET.

תיאור המשימה -

מצורפים לכם 2 קובצי אקסל:
קובץ ראשון (known SET) -- מכיל את הנתונים הבאים עבור 80% מהגנים (3258 גנים) של החיידק E.coli: אינדקס של הגן, רצף ה UTR'5 , רצף האזור מקודד, מדידות של רמות חלבון.

קובץ שני (unknown SET) -- מכיל את הנתונים הבאים עבור 20% מהגנים (815 גנים) של החיידק E.coli: אינדקס של הגן, רצף ה UTR'5 , רצף האזור מקודד.

הקבצים מכילים לכל גן את הפיצ'רים שיצרנו עבורכם:
1. CAI - מידה אשר משערכת את רמת ה-codon bias בגן. גנים עם ערך CAI גבוה נוטים להיות גנים שבאים לידי ביטוי בצורה רבה יותר.
קישורים : https://en.wikipedia.org/wiki/Codon_Adaptation_Index, נספחים.
2. ציון shine-dalgarno - אנרגיית היברידיזציה בין רצף ה- SD ברנ"א שליח והרצף המשלים ברנ"א הריבוזומלי.
קישורים: https://en.wikipedia.org/wiki/Shine-Dalgarno_sequence, נספחים.
3. אנרגיית קיפול של הרנ"א שליח בחלונות של 30 נוקליאוטידים עבור ה-100 חלונות הראשונים באזור המקודד.
קישורים:https://www.nature.com/articles/nrg3681, נספחים.
4. התדירות היחסית של הופעת החומצה אמינית ארגנין (R).
5. התדירות היחסית של הופעת הקודון AGG.
6. אורך הרצף המקודד.

א. סעיף חימום: צרו את הפיצ'רים הבאים בעצמם - התדירות היחסית של הופעת החומצה אמינית ארגנין (R), התדירות היחסית של הופעת הקודון AGG, אורך הרצף המקודד. ודאו כי הפיצ'רים שיצרתם זהים/קורלטיבים לפיצ'רים אשר נתונים לכם.

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

ג. על סמך קובץ ה- known set צרו רגרסור (או כל פרדיקטור אחר) המשתמש בחלק מהפיצ'רים השונים על מנת לחזות רמות חלבון של גנים ללא OVER FITTING. אפשר להשתמש בשיטה שהוצגה בכיתה לבחירת פיצ'רים או בכל שיטה אחרת.

ד. חשבו את קורלציית ספירמן בין רמות הביטוי של הגנים של דאטא האימון עם תוצאות הרצת הרגרסור על דאטא האימון.

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

ו. דרגו את הגנים לפי רמות החלבון שלהם.

תיאור הגשה -
1. צרפו מסמך המסביר בצורה מפורשת על הרגרסור/פרדיקטור שלכם ועל אופן בחירת הפיצ'רים שהכנסתם לרגרסור. (ראו קובץ הגשה לדוגמא שמכיל את הפרמטרים שבהם יש לדון).
2. צרפו קובץ מטלב המכיל את דירוג הגנים.
3. צרפו את קוד המטלב של הרגרסור/פרדיקטור.

*שימו לב כי ניתן להגיש בקבוצות של עד 3 סטודנטים.
* כאשר דירוג = 1 - גן שבא לידי ביטוי בצורה הנמוכה ביותר (רמת חלבון נמוכה).

<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 לmRNA בחיידקים וארכיאות, נמצא בערך 8-12 נוקליאוטידים לפני קודון ההתחלה. מהיחידה הקטנה של הריבוזום מציץ זנב של רנ"א ריבוזומלי אשר נקשר בהיברידיזציה לmRNA באזור ה-SD. תפקיד ה- SD 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.

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 - מספר מקסימלי 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. הmRNA 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.

נספח -- מסמך הגשה לדוגמא:

1. 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?

2. תרשים זרימה שמתאר את פעולת הרגרסור/פרדיקטור.

3. פיצ'רים - האם יצרתם פיצרים חדשים ? אם כן -- איזה פיצרים יצרתם ? (מדוע ?) איזה פיצ'רים נבחרו בסוף ? (מדוע לדעתכם אילו הפיצ'רים שנבחרו ולא אחרים ?) מהי מידת ההצלחה שהפיצ'רים שלכם עם הרגרסור/פרדיקטור חוזים את רמות הביטוי?

4. 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?