מבוא לגנומיקה חישובית ומערכתית – Challenge

מטרה - לבנות predictor החוזה את התגובה של אופרון ה-LUX (שמקודד למספר חלבונים שיוצרים לומינציה) ב-*E. coli* לרמות ה-DNT באוויר על פעל סמך הרצפים (וריאנטים) סביב הפרומוטור של האופרון. ה-predictor מקבל כקלט את רצפי הווריאנטים, מאומן על סמך נתוני ההארה ב- training set ונבדק על הווריאנטים ב- test set.

נתונים מצורפים:

1. הקובץ Train_data.xlsx, הכולל טבלה עם הווריאנטים ב-training set עם הגליונות הבאים הבאים:
   - גליון 1 – Variants data:
   - Variant number – מספר הווריאנט (וריאנט 1 הוא ה-control, ושאר 290 הווריאנטים הנם מוטנטים).
   - Variant sequence – רצף הווריאנט.
   - גליון 2 – luminescence without DNT: שיעור הלומינציה של הווריאנט ללא נוכחות DNT, כפי שנדגם ב-1321 נקודות זמן.
   - גליון 3 – luminescence with DNT: שיעור הלומינציה של הווריאנט ללא נוכחות DNT, כפי שנדגם ב-1321 נקודות זמן.
   - גליון 4 – Features: הכולל את חמישה מהפיצ'רים המפורטים בנספח 3 להלן עבור הוריאנטים. תוכלו להשתמש בהם כדי לוודא שחישבתם את הפיצ'רים נכון.

2. הקובץ Test_data.xlsx, הכולל את כל הנתונים המפורטים בסעיף 1 עבור סט ה-Test.

3. הקובץ asd_hyb. xlsx, הכולל את אנרגיית ההיברידיזציה של רצף ה-Anti Shine-Dalgarno עם כל שישיית נוקלאוטידים אפשרית (ראו נספח 1).

4. הקובץ pssm. xlsx, הכולל מטריצות PSSM (ראו נספח 2).

יצירת המודל:

1. לכל וריאנט נגדיר את d_t להיות ההפרש בין luminescence with DNT לבין luminescence with DNT בנקודת הזמן t. חשבו לכל וריאנט את:
   - ההפרש הממוצע $\overline{d_t}$
   - ההפרש המקסימלי D_T

2. חשבו פיצ'רים רלוונטיים לוריאנטים. בנספח 3 תוכלו למצוא הצעה לרשימת פיצ'רים, אך אתם מוזמנים להוסיף עוד כראות עיניכם.
3. השתמשו בפיצ'רים אלו כדי לאמן 2 מודלים (לבחירתכם ע"י רגרסיה לינארית/ stepwise regression/LASSO/ridge/elastic net/אחרים), החוזים את 2 הפרמטרים שחישבתם בסעיף 1.
4. הפעילו את שני המודלים שאימנתם על הוריאנטים ב-Test set, כדי לחזות את הפרמטרים שלהם (ההפרש והממוצע).

<u>הגשת ה-challenge:</u>

1. צרפו מסמך המסביר בצורה מפורטת את ה-predictor שלכם:
   - איזה סוג מודל בחרתם? למה דווקא אותו? איזה מודל עבד בצורה טובה יותר?
   - האם הוספתם פיצ'רים משלכם? אם כן – אילו (ומדוע?)? אילו פיצ'רים נבחרו בסוף? מדוע לדעתכם אלו הפיצ'רים שנבחרו ולא אחרים?

2. צרפו את **כל** הקבצים הדרושים להרצת המודל מאפס (קוד ונתונים, אם קיימים), למעט קבצים אשר צורפו למסמך זה.

3. צרפו את החיזוי של שני המודלים שלכם לפרמטרים של ה-Tesr set; קובץ אקסל הכולל את העמודות הבאות:
   - name – שם הוריאנט.
   - dt_average – ההפרש הממוצע החזוי עבור הוריאנט בין luminescence with DNT לבין luminescence without DNT ($\overline{d_t}$).
   - DT_max - ההפרש המקסימלי החזוי עבור הוריאנט בין luminescence with DNT לבין luminescence without DNT (D_T).

אנחנו נעשה קורלציית ספירמן בין הפרדיקציות שלכם והערכים האמיתיים כך שאפשר גם לתת דירוג לכל וריאנט.

בהצלחה!

נספח 1: Shine-Dalgarno

בבקטריות וארכיאות רבות מופיע הרצף AGGAGG מעט לפני קודון ההתחלה. רצף זה משלים לרצף המצוי בתת-היחידה הקטנה של הריבוזום. הרצף AGGAGG קרוי רצף Shine-Dalgarno (על שם שני החוקרים שגילו אותו), והרצף המשלים לו קרוי Anti Shine-Dalgarno. רצף ה-SD מאפשר לריבוזום זיהוי נכון של קודון ההתחלה ובכך מסייע לשלב ה-initiation. מכיוון שבפרויקט זה אנו עוסקים בבקטריה (*E. coli*), וספציפית ב-promoter שלה, יש חשיבות לאפיניות בין רצף הAnti-SD לבין ה-mRNA; אפיניות זו מיוצגת ע"י אנרגיית ההיברידיזציה בין השניים, המופיעה בקובץ asd_hyb.xlsx.

נספח 2: ציוני PSSM

עבור אוסף של רצפים, ניתן לחפש מוטיבים (motifs) – תתי-רצפים קצרים המופיעים באוסף יותר מהמצופה (enriched). נצפה שלמוטיבים אלו תהיה משמעות ביולוגית כלשהי, אינפורמציה המקודדת ברצף הנוקלאוטידים. המוטיבים מיוצגים ע"י PSSM, שהיא מטריצת הסתברות ההופעות של כל נוקלאוטיד באותו תת-רצף. ע"י שימוש במטריצה זו, ניתן להעריך את הדמיון בין רצף נתון לבין המוטיב המתאים ל-PSSM.

לדוגמה, באוסף הרצפים הבא:

G	T	C	C	G	A	T	A	T
G	G	T	G	A	G	C	C	C
G	T	T	A	T	G	G	C	A
G	C	C	C	T	G	C	T	C

מטריצת ה-PSSM של מיקומים 1:5 היא:

		1	2	3	4	5
A		0	0	0	0.25	0.25
C		0	0.25	0.5	0.5	0
G		1	0.25	0	0.25	0.25
T		0	0.5	0.5	0	0.5

(מכיוון שזו תדירות ההופעה של כל בסיס בכל מיקום)

בעזרת מטריצה זו ניתן לחשב את הציון, או הדמיון, של הרצף GGCCA (או כל רצף אחר) ביחס למודל שביטאנו במטריצת ה-PSSM:

$$P(GGCCA) = P_{G1} \cdot P_{G2} \cdot P_{C3} \cdot P_{C4} \cdot P_{A5} = 1 \cdot 0.25 \cdot 0.5 \cdot 0.5 \cdot 0.25 = 0.015$$

(כלומר הסיכוי לקבל את הרצף GGCCA בהינתן תדירות ההופעה של כל בסיס בכל מיקום)

בקובץ pssm.xlsx מצורפות 12 מטריצות PSSM של מוטיבים שנמצאו כ-enriched. מוטיבים אלו נמצאו בעזרת תוכנה לחיפוש מוטיבים שנקראת HOMER באופן הבא:

לכל וריאנט חישבנו את ה-D_T הממוצע שלו (על פני שלושה ניסויים), כלומר את ההפרש המקסימלי ברמת ההארה (עם לעומת בלי DNT) אליה הגיע. לאחר מכן זיהינו וריאנטים שה-D_T הממוצע שלהם היה ב-20% העליונים, והשווינו את קבוצת וריאנטים אלו לקבוצת כלל הווריאנטים; כלומר חיפשנו מוטיבים שהופיעו יותר מהמצופה בווריאנטים בעלי רמת הארה גבוהה, בהשוואה לסט הווריאנטים הכולל. זאת מתוך הנחה שלמוטיבים בווריאנטים אלו יש השפעה חיובית על תהליך הביטוי. תוכלו להשתמש במטריצות אלו כדי לחשב ציוני PSSM של תת-רצפים שונים בווריאנטים, כמתואר בפסקאות הקודמות.

<u>נספח 3: פיצ'רים, הצעת הגשה</u>

(אין חובה להשתמש בכל הפיצ'רים המוצעים)

- ה-GC content של הווריאנט
- אנרגיות הקיפול של sliding window באורך 40 *(עוד על אנרגיות קיפול – בהמשך הקורס)*
- אנרגיית הקיפול של הווריאנט
- ההפרש בין אנרגיית הקיפול של הווריאנט לאנרגיית הקיפול של ה-Control
- אנרגיות ההיברידיזציה של sliding window באורך 6 עם ה-Anti-SD
- מספר ומיקום המוטציות בווריאנט ביחס ל-Control
- מספר הקודונים שהשתנו בווריאנט ביחס ל-Control
- ציון ה-ChimeraARS של הווריאנט, בהתבסס על הגנום של E. coli *(עוד על כימרה – בהמשך הקורס)*
- ה-CAI וה-tAI של ה-ORF, כאשר סט הרפרנס הוא כל הגנום של E. coli
- ציוני ה-PSSM של תתי-רצפים בווריאנט, על סמך 12 המוטיבים שצורפו