Final project : Deep Learning in Computational Biology

מגישות: רותם כהן 316220607 רעות לב 207385741

מטרת הפרויקט שלנו היא לנבא את עוצמת קשירת הRNA בכל prob של RNA compete.

רקע ביולוגי:

בגופנו ישנם חלבונים רבים שנוצרים ע"י תהליך התרגום שבו המידע הגנטי שלנו מתורגם לחלבונים. בתהליך זה משתתף הRNA והוא נקשר לחלבונים שמשתתפים בתהליך. חוזק הקישור בין הRNA לחלבונים משפיע על תהליכים חיוניים שונים בגוף.
במסגרת הפרויקט, אנו מתכוונות לחקור ולחזות את הכמות הקשרים בין כ-240,000 מולקולות RNA לחלבון מסוים. קיבלנו מידע מניסויים שבהם בדקו את רמת ההיקשרות של מולקולות RNA לחלבונים בריכוזים שונים. בין מולקולת הRNA לאתרי הקישור על החלבונים ישנה תחרות, ככל שריכוז החלבון גבוהה יותר כך הסיכוי לקשירה של מולקולות הRNA גבוהה יותר. כמו כן, אם רואים קשירה של מולקולת RNA לחלבון שריכוזו נמוך נסיק כי עוצמת הקשירה היא גובה מאחר ויש תחרות ומולקולת הRNA צלחה בכל זאת להתחבר לחלבון.
הניסוי שהתבצע:
מכניסים למבחנה הרבה מולקולות RNA. מוסיפים חלבון בריכוז מסוים , למשל בריכוז 5nM ובודקים כמה מהמולקולות נקשרו ואם בכלל. את אותן המולקולות מרצפים. חוזרים על הניסוי עם ריכוזים שונים.

שורת ההרצה:

```
python3 MLBProject.py RBP1_output.txt RNACMPT_seq.txt RBP1_input.seq
RBP1_5nM.seq RBP1_20nM.seq RBP1_80nM.seq RBP1_320nM.seq
RBP1_1300nM.seq
```

זוהי דוגמה עבור החלבון הראשון RBP1. ביצענו את שורת ההרצה עבור כל החלבונים שנתנו.
הסבר שורת ההרצה:

- MLBProject.py – שם קובץ ההרצה.
- RBP1_output.txt – שם קובץ הפלט, אליו נכתוב את עוצמות הקשירה שקיבלנו מהמודל.
- RNACMPT_seq.txt – (הקובץ שהורדנו מהמודל RNAcompete_sequences) שמכיל שורות של רצפי RNA שונים. המטרה שלנו היא לחזות עבורם את עוצמת הקשירה של החלבון אליהם.
- RBP1_input.seq – קובץ שמכיל את כל הרצפים בניסוי, גם רצפים שנקשרו וגם כאלה שלא.
- RBP1_5nM/20nM/80nM/320nM/1300nM.seq – הקבצים שקיבלנו מהניסוי בהתאם לריכוז של החלבון.

בכל ניסוי, יש מספר רב של מולקולות RNA במבחנה, לאחר הוספת חלבון בריכוז מסוים בודקים אילו מהרצפים נקשרו לחלבון ומרצפים אותם. הקבצים של הריכוזים השונים מכילים את הרצפים שנקשרו בהתאם לריכוז.

<u>תיאור מבנה המודל:</u>
המודל מורכב מ5 שכבות:

- <u>שכבת הקלט -</u>
  השכבה הראשונה היא שכבת הקלט והיא מיוצגת ע"י שני פרמטרים :
  לוקח רצפי קלט באורך 'sequence_length' - אורך הרצף הארוך ביותר בקובץ שבו עושים את החיזוי. עם וקטור בגודל 4 לייצוג הבסיסים (אחד לכל תו ברצף ה- U\ DNA\RNA: A, C, G, T) .
  ה- 'input_shape' מוגדר ל-(sequence_length, 4) .
- <u>שכבת קונבולוציה 1D -</u>
  שכבת Conv1D היא השכבה השנייה במודל ומטרתה היא לעבד את הרצפים. בשכבה זו ישנם 64 פילטרים, גודל הקרנל הוא 3 ופונקצית האקטיבציה של השכבה היא relu.
- <u>שכבת Global Max Pooling 1D -</u>
  שכבת Global Max Pooling 1D היא השכבה השלישית במודל ומטרתה היא לקחת את הערך הכי חשוב ,הערך המקסימלי, בשכבה הקודמת ולהעביר אותו הלאה. (מקטין את הממדים המרחביים של הקלט)
- <u>שכבת Fully Connected-</u>
  שכבת Fully Connected הינו השכבה הרביעית במודל ומטרתה היא לשלב את המידע מהשכבות הקודמות ולייצר פיצ'רים.
  גודל שכבה זו הוא 128 ופונקציית האקטיבציה היא relu.
- <u>שכבת פלט -</u>
  שכבת הפלט היא השכבה החמישית במודל. שכבה זו כל רכיב הוא ההסתברות של רצף הRNA להקשר לחלבון מסוים בריכוז מסוים.
  גודל הפלט הוא כמספר הקבצים הניתנים בשורת ההרצה – קובץ ה input וקבצי הריכוזים.

אימנו את המודל שלנו באמצעות 80% מהמידע ו20% מהווים מבחן.
פונקציית הloss במודל שלנו היא MSE עם adam כאופטימייזר. התחלנו את ההרצות עם 50 epochs אך הגענו למסקנה ש10 epochs מביא את התוצאה הטובה ביותר בזמן הקצר ביותר.
כמו כן, גודל כל batch של 32.

שלב העיבוד המקדים:

על מנת להתחיל לאמן את המודל שלנו ביצענו עיבוד מקדים על המידע שקיבלנו. המידע שלנו מחולק לשני חלקים – קובץ אחד שמכיל את הקלט וקבצים שמכילים את הריכוזים. בדקנו האם הרצף נמצא באחד מקבצי הריכוזים ואם כן ספרנו אותו והחשבנו אותו כנקשר לחלבון. בגלל התחרות על אתרי הקשירה שציינו קודם, ככל שהרצף נקשר בריכוז נמוך יותר כך עוצמת הקשירה היא גבוהה יותר. עברנו על כל הרצפים שנמצאים בקבצים וסימנו באיזה ריכוז הרצף נקשר.
כמו כן, כדי שמידע באימון יתאים למידע שעליו נבצע את החיזוי נרצה שהרצפים יהיו באותו האורך. לכן ביצענו padding עבור הרצפים שעליהם עושים חיזוי. הpadding נעשה ע"י לקיחת האורך המקסימלי של הרצף מתוך קובץ הrncmpt או מתוך הdata של האימון.

עיבוד מקדים של הרצף:

ביצענו עיבוד מקדים גם לרצף שאותו בודקים , כלומר הפכנו אותו לone-hot vector כדי שיהיה יותר נוח לעבוד איתו. עשינו זאת ע"י המרת הרצף לוקטור בגודל 4 כך : האות A היא במקום 0, האות C היא במקום 1, האות G היא במקום 2 והאות T היא במקום 3.
. DNA_VOCAB = {'A': 0, 'C': 1, 'G': 2, 'T': 3}
כך לדוגמה, הרצף TC יוצג בצורה הבאה: [[0,1,0,0],[0,0,0,1]].
מקרה מיוחד מטופל עבור התו 'N', שבו מוקצית התפלגות הסתברות ברירת מחדל DEFAULT_CHAR_PROBABILITY שכן 'N' מייצג בסיס לא ידוע.
את גודל הלייבל שלנו בנינו כך שספרנו את כמות קבצי הריכוזים לפי שורת הפקודה. ברשימת הקבצים שיצרנו קובץ הinput הוא הראשון ואחריו מופיעים הקבצים לפי הריכוזים שלהם. ככל שהריכוז גבוהה יותר כך הוא יהיה נמוך יותר ברשימה.

קבצים :

מיון שמות קבצים - רשימת שמות הקבצים (input_and_concentrations_file_names) ממוינת על סמך ערכי הריכוז שחולצו משמות הקבצים. זה חשוב לארגון הקבצים בסדר ריכוז עולה.
חילוץ מספרים משמות הקבצים - הפונקציה 'extract_number' מחלצת ערך ריכוז משם קובץ. זה מניח שהמספר מגיע אחרי מפריד שצוין (קו תחתון `_` במקרה זה)

עיבוד מקדים לליבל:

הקוד משתמש במבנה מילוני לניהול רצפי DNA. מפתחות במילון תואמים לרצפי DNA ייחודיים, המיוצגים בצורה one-hot encoded . ערכים המשויכים לכל מפתח מייצגים את התווית עבור הרצף המתאים. במהלך האיטרציה על קבצי הרצף, המסומנים על ידי המשתנה 'i' כאינדקס של הקובץ הנוכחי ברשימה, מתבצע הליך התיוג הבא:
אם כבר קיים רצף במילון, התווית מתעדכנת על ידי הגדרת הערך באינדקס ' i' ל-1.
אם רצף אינו נמצא במילון, הוא נוסף עם וקטור תווית ראשוני `[0, 0, 0, ..., 0]`. אורכו של וקטור זה מתאים למספר הכולל של קבצים.
התוויות מייצגות למעשה את הסבירות של רצף להיות קשור לחלבון בריכוז מסוים 'i' האינדקס של האלמנט שאינו אפס בווקטור התווית מציין את רמת הריכוז.
גישת תיוג זו דומה למשימת סיווג, הקובעת לאיזו רמת ריכוז חלבון שייך רצף נתון.
תהליך התיוג נועד ללכוד את הקשר בין רצפי DNA וריכוזי חלבון, לספק ייצוג מובנה של התפלגות ההסתברות עבור כל רצף על פני רמות ריכוז שונות

אחרי שאימנו את המודל וביצענו חיזוי , עיבדנו את הלייבל שחזינו עבור כל רצף.
עיבוד התוצאות:

הורדנו את הסיכוי להיות בinput מההסתברות להיקשר לסכום שני האיברים האחרונים בוקטור שהם בעצם 2 הריכוזים הגבוהים ביותר שיש.
אם יש רצף שעוצמת הקשירה שלו נמוכה, ההסתברות שלו להקשר לחלבון נמוכה גם אם ריכוזו של החלבון גבוה. לכן, ההסתברות שנקבל את הרצף בinput גבוה. על כן, נקבל שעוצמת הקשירה של הרצף נמוכה.
מנגד, אם יש רצף שעוצמת הקשירה שלו גבוהה, ההסתברות שלו להקשר לחלבון גם בריכוז נמוך גובה. לכן, ההסתברות שנקבל את הרצף בinput נמוכה. על כן, נקבל שעוצמת הקשירה של הרצף גבוהה.
עבור רצפים שאינם נקשרים לחלבון , ההסתברות להקשר היא 0 ולכן נקבל מספר שלילי או קטן מאוד.
לבסוף, נקבל את התוצאות ונדע איזה רצפים נקשרים טוב יותר, כלומר בעלי עוצמת קשירה טובה יותר לחלבון משאר הרצפים.

עבור 16 החלבונים הראשונים (RBP1,RBP2,....,RBP16) חישבנו את קורלציית פירסון עם תוצאות האמת שנתונות לנו.

התוצאות של החלבונים בסט האימון (קורלציית פירסון ומספר חלבון) :

Protien	Correlation
RBP1	0.250479996710097
RBP2	0.27299279408691657
RBP3	0.03037346737920467
RBP4	0.0965302536549499
RBP5	0.24875309528019252
RBP6	0.23159054839371268
RBP7	0.024395979015385396
RBP8	0.0806884481826018
RBP9	-0.07275908931171494
RBP10	0.06747835264163475
RBP11	0.1496448510563671
RBP12	0.3347752445087757
RBP13	0.3536284155986621
RBP14	-0.006099763360174129
RBP15	0.0728158170200024
RBP16	0.33500791961577164
	0.1543754949079016

אחרי שסיימנו לאמן על 16 הקבצים ולחשב את קורולציית פירסון לכל ריצה ביצענו ממוצע : 0.1543.

נשים לב שיש קורלציה אך היא הינה גבוהה.

קבצי החיזויים נמצאים בנתיב הבא:

https://drive.google.com/drive/folders/1h6rgi_cHQNEHVYTB_Yxuiv1_mju7tizh?usp=share_link