

# אלגוריתמים בביולוגיה חישובית (76558)

## תרגיל 2: EM וזיהוי איי CpG לפי רצף

תאריך הגשה: 07/12/2025

- חלק 1: מימוש EM למציאת מטבע מוטה..... 2
- סיפור המסגרת (כדי לתת מוטיבציה)..... 2
- תזכורת לגבי נוסחאות חשובות..... 2
- תיאור הדאטה - חלק 1..... 3
- שאלות ומטרות חלק 1..... 3
- חלק 2: מימוש מודל לזיהוי איי CpG מדאטה קיים..... 4
- תיאור הדאטה - חלק 2..... 4
- מטרות חלק 2..... 7
- המלצות (לא באמת המלצות)..... 7
- דרישות חלק 2..... 7
- ויש גם תחרות (!)..... 8
- מה להגיש?..... 9
- הוראות כלליות למצגת..... 9
- חלק 1..... 9
- חלק 2..... 9
- וביחד..... 9
- שימוש בכלי AI..... 10
- קריטריוני ציון..... 10

# חלק 1: מימוש EM למציאת מטבע מוטה

בחלק זה נתרגל את השימוש באלגוריתם ה-EM.

## סיפור המסגרת (כדי לתת מוטיבציה)

Q (ולא ג'יימס בונד 😞), נשלח על ידי M לקזינו רויאל במונטנגרו במשימה ללמוד את השימוש באלגוריתם ה-EM כחלק ממשימה לנצח משחק מזל של הטלת מטבעות.

**מהלך המשחק:** הדילר מטיל מטבע 50 פעמים, בכל פעם המהמרים מנחשים על איזה צד המטבע יפול ומהמרים בהתאם. המטבע המוטל יכול להיות מוטה או לא מוטה. למטבע יש כמובן שני צדדים (עץ או פלי).

על Q ללמוד את ההסתברויות להחלפה בין המטבעות, ואת ההסתברות של כל מטבע ליפול על עץ או פלי. הוא יעשה זאת בעזרת אלגוריתם ה-EM. אם Q יצליח במשימתו הוא יסייע לג'יימס בונד לזהות את אירועי ההחלפה בזמן המתאים ולהמר בהתאם כדי למקסם את כספי הזכייה לטובת הכתר הבריטי 🏰.

המשימה שלכם בחלק זה תהיה ללמוד פרמטרים למודל HMM כאשר רק התצפיות ידועות. כלומר, עליכם למצוא תצטרכו למצוא את ערכי המטריצות  $\pi$ ,  $\tau$ . בעזרת ערכים אלו ניתן יהיה לסייע לג'יימס לנצח את הקזינו.

## תזכורת לגבי נוסחאות חשובות

\* בתרגיל זה, נשנה את הפסאודו-קוד כך שהאלגוריתם יעצור אחרי מספר איטרציות מקסימלי ( $>= 150$ ) (לא לפי שינוי הנראות או ה- $\theta$ ) כדי להקל על המימוש

# אלגוריתם באום-וולש

- אתחול  $\theta$  [רנדומית]
- בכל איטרציה  $t$
- לכל רצף  $X^i$  חשבו את הפורוורד והבקוורד (בעזרת  $\theta$  הנוכחית)

• שלב E: 
$$\tilde{N}_{k,l} = \sum_{j=1}^N \sum_{i=2}^{n_i} [F_k(i-1) \cdot \tau_{k,l} \cdot e_l(X_i) \cdot B_l(i)] \quad \tilde{N}_{k,x} = \sum_{j=1}^N \sum_{i: X_i^j = x} \frac{F_k(i) \cdot B_k(i)}{P(\vec{X}^j)}$$

• שלב M: 
$$\tilde{\tau}_{k,l} = \frac{N_{k,l}}{\sum_m N_{k,m}} \quad \tilde{e}_k(x) = \frac{N_{k,x}}{\sum_y N_{k,y}}$$

- עיצרו אם השיפור בלוג הניראות [או אם השינוי ב- $\theta$ ] קטן מספיק

## תיאור הדאטה – חלק 1

במסגרת המשימה Q ישב בקזינו זמן רב וערך תצפיות על סדרות ההטלות. כאמור, כל סדרה שכזאת מכילה 50 הטלות. הוא דיווח את הסדרות בקובץ, שנשמר בפורמט fasta. הקובץ מכיל שורות של סדרות הטלות. צוות הקורס ולה שיף (בעל הקזינו) הגדירו את ההטיה של כל מטבע ואת ההסתברות להחליף בין המטבעות. שורה לדוגמה בקובץ:

>seq1

TTHHHTTTTTHTHTTTHHHTTTTTHTTTHHHHHHHHHHTTHTHHHTTTTTTHT

## שאלות ומטרות חלק 1

עליכם לכתוב תכנית שתקבל קובץ fasta עם רצפים שיוצרו באופן מלאכותי (הרצפים יוצרו לפי הסתברויות מעבר ופליטה מסוימות), ולדווח על הסתברויות שאתם הגעתם אליהן. כלומר, לממש את אלגוריתם ה-EM של באום-וולש (כמתואר בפסאודו-קוד). לנוחותכם מצורף שלד לחלק זה (אין חובה להשתמש, אבל מומלץ להסתכל טרם העבודה). **אסור** להשתמש במימוש קיים מספרייה של האלגוריתם, **עליכם לממש אותו בעצמכם**.

הנה מספר שאלות מנחות שעליכם לענות עליהם (ויסייעו להבין):

- מה הסתברות הפליטה של כל מטבע בכל מצב חבוי? האם ההסתברויות הגיוניות? לאיזה כיוון מוטה המטבע המוטה? (שימו לב, האיתחולים הרנדומליים משפיעים אז הריצו את האלגוריתם מספר פעמים לפני שתכריעו)
- מה הסתברות המעבר בין המצבים החבויים? תוכלו להסביר מדוע היא הגיונית?
- עליכם להראות גרף המראה את תהליך המקסימיזציה, כלומר את השינוי ב- $\log(P(X^j|\theta))$  (הציגו את מספר האיטרציה בציר ה-x ואת ערך ה- $\log(P(X^j|\theta))$  (בסכימה על כל הרצפים) בציר ה-y), שכנעו אותנו שהאלגוריתם מתכנס. האם הגרף שקיבלתם הגיוני להתנהגות האלגוריתם?
- הריצו את האלגוריתם עם שינוי במספר האיטרציות המקסימליות, והסתכלו על הנראות הסופית שהתקבלה, איך הגרף מתנהג? האם ככל שמספר האיטרציות המקסימליות עולה בהכרח הנראות משתפרת?
- נסו את האלגוריתם שלכם עם אתחולים שונים (למשל, אתחול אחיד להסתברויות הפליטה, אתחול שונה להסתברות הבחירה ההתחלתית ועוד) איזה שינוי אתם צופים בכל אתחול? תוכלו להסביר אותו לפי הנוסחה שפותחה בכיתה? גם בסעיף זה עליכם להראות גרף המראה את תהליך המקסימיזציה, כלומר את השינוי ב- $\log(P(X^j|\theta))$  (הציגו את מספר האיטרציה בציר ה-x ואת ערך ה- $\log(P(X^j|\theta))$  (בסכימה על כל הרצפים) בציר ה-y) אולם כעת לכל אתחול שבחרתם.

## חלק 2: מימוש מודל לזיהוי איי CpG מדאטה קיים

בחלק זה אתם מתבקשים לכתוב תוכנה שתזהה איי CpG בתוך רצף דנ"א ארוך. בניגוד לחלק הקודם, הפעם למידת הפרמטרים למודל ה-HMM שלנו קלה יותר. למה? בגלל שהדאטה שלנו מכיל תיוגים לאתרי ה-CpG נוכל ללמוד את המטריצות  $\pi$ ,  $\tau$  ישירות ממנו. הפעם, במקום סיפור מסגרת יש סיפור מדעי.

רקע ביולוגי (קצר ביותר 😊)

בגנום, ישנם אתרים בהם מופיע הנוקלאוטיד C ולאחריו הנוקלאוטיד G, אתרים שכאלה נקראים אתרי CpG. אזורים שבהם מופיע ריכוז גבוה של אתרי CpG, נקראים איי CpG

ולהם יש משמעות רבה בבקרה וביטוי של גנים. אתרי CpG בעלי חשיבות ביולוגית בגלל שהם יכולים לעבור תהליך הנקרא: **מתילציה**. במסגרת תהליך זה הנוקלאוטיד C עובר שינוי כימי (מקבל קבוצת מתיל), שינוי זה משפיע על היכולת של מכונות שעתוק לגשת כדי לשעתק את האתר הממותל (כשיש אתרים רבים כאלו באזור מרוכז, כלומר, באי). כאמור, איי CpG מכילים ריכוז גבוה של אתרים בהם ניתן לבצע מתילציה ולכן יש להם השפעה נרחבת בבקרה וביטוי גנים (ניתן למשל לחשוב שאתרים שבהם מתחיל רצף של גן, יטו יותר להכיל איי CpG ובכך ניתן יהיה לבקר את הביטוי של הגן בתאים מסוגים שונים, על ידי הוספת או הורדת מתילציה).

## תיאור הדאטה – חלק 2

מצורפים לחלק זה שני קבצים.

1. **CpG-islands.2K.seq.fasta**, הוא קובץ fasta, מכוון בעזרת תוכנת gzip. הוא מכיל כ-1,103 רצפים באורך 2,000 בסיסים כל אחד, המכילים את רצף הדנ"א הגנומי של אי CpG בודד, וכן את הרצף הגנומי המקיף אותו משני הצדדים.

למשל, הרצף הראשון בקובץ:

>chr1:134,858-136,857 (266,1294)

```
TAAAAAATTCGGGCTTGGCGCAGAAACTCACTCCAAATAAATTACCTACCAAAACATTTACATAATGGTGGAATATTCCAAAATTCATATTTTGGGATTATACACAAAAGATAAA
CAAAATTAGAGGCCAAAGAGGCTGCCGGAAGGGAAGGAGGCGCTGGAATGGCCGACGTGAGGAATGAGCTGGGCTAAAGAGGCCACTGGCAGGCAGGAGCTGGACCTG
CCGAAGTGGCCGAAAGGCAGGAGCTTTGGACTGGGGAGGCCGCGAGTGAGGCGAGAGCTAGCTGGGCGTGGAGAGTCCGCTGTGAGGCGCGAGGCCGAGGCTGGGCCC
GTGACAGGCCCTTCGAGACGACGAGGAGGCCCGGGCCCTGCAGAGGCCGACTGGAGATCAAGTTCTGCGCCCTGAAGAGGGCTGCCAAAAGTCAAAAAGCGGGGGCCCTGGGAAGGCC
CGCCGAGAGCCATGAGCTGCGGCTGGGCGGAAAGAGGCCCACTGGGAGGCGAGGAGGCTGGGCGCTGGAGAGGCTGACTCGAGGAAGTTTTCACCTGGAGAGGCCGTC
GAGAGGAGCAGGCTGGGCGCCAGGGAGGCCGACTTGTGCTCTTCCAGGCCCACTTCCAGGCCGAGCTTGGAGACGACTTGGGCGCTGCAGAGGCCGCCCGGGAGGCTGGGA
GCTAAGCCTGGAGAGACTGACTTTCGGGACGATTGGGCGCTGCGGAGGCCGCCGCGGAGGCCCAAGCTGGGCGCTAGAGGAGGCCACCGACCGAGGCCATTTGGGCGCTG
CAGATGTGTCATCGGAGGCCAGGAGCTGAGCCTGGAGAGGCCACCGCAGGCCGAGCCTGAGCTGGGCGCTGGGAGCTTGGCTTAGGGAAGTTGTGGGCGCTACCGGCCGCTG
GGAGCTGGGCGAGGAGCTGAGTCCAAAGACGTTGTTGGGACCTGGAGTGGGCCAGAGTCCGGCCTGGAGATGCAGCCGGGAGGAAGAGCTGGGCCCGGAGGGGGCGC
```

הדוגמא מציגה רצף כזה, הלקוח מגנום האדם, כרומוזום אחד, בקואורדינטות 136857-134858. עוד מופיעים בשם הרצף אורך השוליים מימין ומשמאל לאי ה-CpG (במקרה זה, 266 בסיסים מימין, 1294 בסיסים משמאל). **לנוחיותכם**, רצף האי עצמו מסומן בبولד (בדוגמה זו).

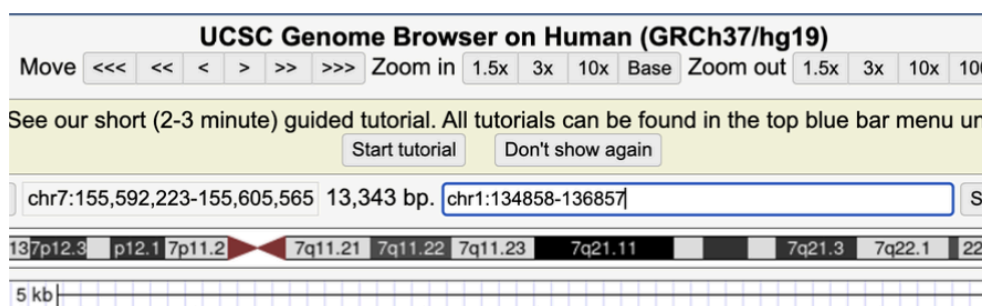
למשל, הרצף הראשון בקובץ:

**שימו לב** שבקובץ זה, N מסמן בסיס (נוקלאוטיד) כלשהו, ורצף ה-C מסמן את מיקום האי.

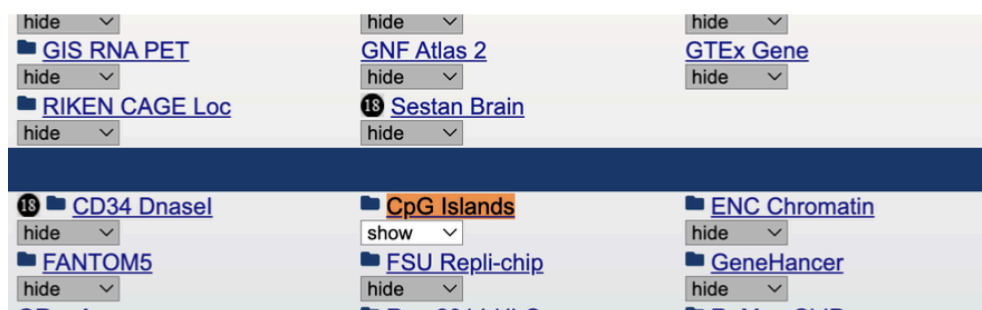
## מקור הנתונים:

באופן כללי, כדי למצוא את הרצפים ואת האנוטציות, הרצפים נלקחו מגנום האדם מאתר של אוניברסיטת סנטה קרוז בקליפורניה (קישור לאתר), וכך גם מיקומי האיים, לפי גירסא 19 של גנום האדם (hg19) (הדפדפן הגנומי).

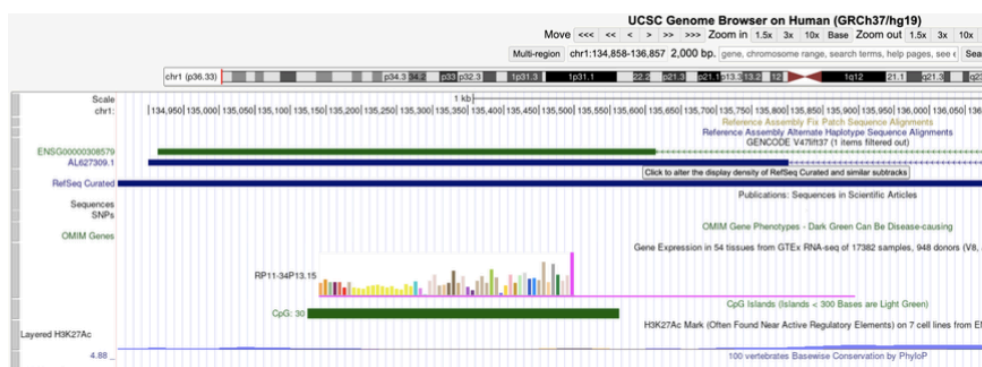
ניתן לגלוש למיקום הגנומי ממנו נלקח הרצף הנ"ל (chr1:134858-136857 2,000 bp)



להוסיף אנוטציות של איי CpG (על ידי לחיצה על CpG Islands ואז Refresh בצד ימין) ולראות את האזור הגנומי, כולל מיקום האי במלבן ירוק.



הקלקה על המלבן תציג מידע, כגון אורכו, מספר אתרי המתילציה, וכן הלאה.



**הערה חשובה:** בקובץ הנ"ל נבחרו כל האיים מכרומוזום אחד, באורכים שבין 150 ל-500 בסיסים.

## מטרות חלק 2

עליכם לכתוב תכנית שתקבל קובץ fasta מכווץ, בפורמט זהה לקובץ הרצפים הגנומיים (CpG-islands.2K.seq.fasta), וליצור כפלט קובץ בפורמט זהה לקובץ (CpG-islands.2K.lbl.fasta), בו יהיה ניבוי טוב ככל האפשר של מיקומי האיים.

## המלצות (לא באמת המלצות)

אתם **חייבים** לבנות מודל HMM. אבל, מוזמנים לגלות יצירתיות במבנה שלו.

למשל:

- תוכלו לאמן שני מודלים מרקוביים של רצפים גנומיים, ולחשב את לוג יחס הנראות של תתי-רצפים בקלט.
- או שאתם יכולים ללמוד מודל מרקובי חבוי, עם שני מצבים (N-C), אשר פולט רצפי דנ"א. או כזה עם יותר מצבים חבויים.

**דגש:** זיכרו שהדנ"א הוא דו-גדילי, וכדי להגדיל את סט האימון שלכם, אתם יכולים גם להפוך את הרצפים reverse complement (וכמובן לחשב מחדש את מיקום האי).

## דרישות חלק 2

### • תיאור המודל (פורמלי ומפורט):

- עליכם לתאר את המודל על פיו אתם עובדים בצורה ברורה ופורמלית, ברמה שתאפשר לנו ליישם מחדש את המודל שלכם, אם נרצה (אם הוא יהיה ממש טוב אז כנראה שממש נרצה).
- אנא הקפידו על שרטוטים, תיאור החלקים השונים במודל, הסתברויות המעבר והפליטה, וכן הלאה.
- אנו נשים על כך **דגש** בבואנו לתת ציון לתרגיל.

### • לימוד הפרמטרים:

- אנא הקפידו לציין בצורה ברורה כיצד למדתם את הפרמטרים השונים של המודל. האם ניחשתם אותם?
- השתמשתם באומדנים כאלה או אחרים (למשל אומד ניראות מירבית, MLE)? כיצד חישובתם את ערכו? מה ההנחות שהנחתם? וכו'.
- שוב, התיאור צריך להיות ברמה שתאפשר לחזור על צעדיכם בצורה מלאה.

### • זמן ריצה וספריות:

- אנא הקפידו על זמן ריצה סביר הן בלמידה והן בניתוח רצפים חדשים.
- השתמשו רק בחבילות פיית'ון נפוצות. (מומלץ להשתמש ב-hmmlearn)

## ויש גם תחרות (!)

המודל הטוב ביותר (שיביס גם את המודל של צוות הקורס, וגם כל מודל אחר של כל משתתפי הקורס) יזכה בפוסט יפה בפורום ההודעות ובבונוס לציון התרגיל. על מנת להשתתף בתחרות עליכם לוודא שהשלד ממומש כנדרש ללא שינוי. **שימו לב**, על המודל שלכם להצליח מול דאטה שעליו הוא לא אומן (test data). לכן, מומלץ שתחלקו בעבודתכם את הדטאה ל-train/validation/test כך שתוכלו לבחון את ביצועי המודל שלכם על דאטה שעליו המודל לא אומן.



# מה להגיש?

על מנת לעודד למידה פעילה עליכם להקליט מצגת בה תציגו את פתרון התרגיל שלכם. לעניין הגשה בזוגות, ניתן (ומומלץ מאוד) להגיש בזוגות.

## הוראות כלליות למצגת

על הקלטת הסבר המצגת להיות באורך של 5-8 דקות. ניתן להציג באנגלית או בעברית. ההקלטה (וודיאו) צריכה לכלול את המצגת. למשל, הקלטה של zoom של ההצגה כאשר המצגת משותפת בשיתוף מסך. המצגת צריכה להיות ברורה ומסודרת (ראו מצגות מצורפות עם המלצות בעניין כיצד כדאי לבנות מצגת, תחת לשונית הסקרייב במודל, **הקפידו על מצגת מסודרת**). יש לציין במצגת את המקור של חומרים (גרפיקה, איורים וכו').

### חלק 1

- חלק 1 במצגת ובו תיאור מפורט (בעברית או באנגלית) של הפתרון שלכם, הגרפים הנדרשים והמענה לכל השאלות בחלק 1 (שאלות ומטרות חלק 1).
- **קובץ python אחד או יותר:** עם התכנית. לנוחיותכם, הכנו לכם קבצים מוכנים עם מספר פונקציות עזר. אנא הקפידו על תיאור ותיעוד ברורים של הקוד שלכם, שיאפשרו לנו לעיין בו ולהבין מה אתם חושבים שעשיתם בקוד.

### חלק 2

- חלק 2 במצגת ובו תיאור מפורט (בעברית או באנגלית) של הפתרון שלכם.
  - עליכם לכלול את תיאור המודל, תיאור ההנחות עליהן נשענתם, חבילות התוכנה בהן השתמשתם, הסברים על אימון המודל והזמן/מספר הרצפים שזה לקח. שימו לב ל-דרישות חלק 2
- **קובץ python אחד או יותר:** עם התכנית. לנוחיותכם, הכנו לכם קבצים מוכנים עם מספר פונקציות עזר. אנא הקפידו על תיאור ותיעוד ברורים של הקוד שלכם, שיאפשרו לנו לעיין בו ולהבין מה אתם חושבים שעשיתם בקוד. כדי להשתתף בתחרות, שמרו על המבנה בקובץ השלד שסיפקנו לכם.

### וביחד

**תיקייה בפורמט tar/zip, המורכבת משתי תתי תיקיות, לקוד של חלק 1 ולקוד של חלק 2, על התיקיות להכיל את הנדרש בכל חלק. בנוסף, pdf של המצגת אותה הקלטתם (אשר מכילה את הגרפים והתרשימים הנדרשים לכל חלק) בתוך התיקייה הגדולה.** את הקלטת המצגת יש להגיש **בקישור** לסרטון פרטי (עדיף יוטיוב)

בתיבה המתאימה תחת לשונית הגשת התרגיל במודל. נא להוסיף בשם התיקיה את שמות/ת.ז. המגשים  
ex2-1, כלומר: ex2\_name1\_name2.tar.

## שימוש בכלי AI

השימוש בכלי עזר לתיכנות מבוססי בינה מלאכותית (דוגמת chatgpt) מותר, אולם עליכם:

- להצהיר על כך במסגרת שקופית במצגת (במידה ולא השתמשתם, יש להצהיר גם על כך)
- לפרט באיזה כלים השתמשתם
- מה הפרומפטים שהכנסתם (בקווים כלליים)
- לתאר במילים באופן מפורט את תהליך העבודה על התרגיל עם הכלי/ם

## קריטריוני ציון

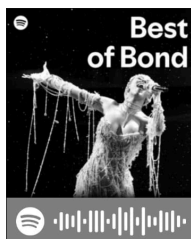
ככל, הקפידו על הדברים הבאים:

**בחלק 1:** הקפידו על מענה לכל השאלות המנחות, הגישו את הגרפים הדרושים, ושכנעו אותנו במצגת שהבנתם את האלגוריתם.

**בחלק 2:** הציון יינתן תוך שקלול דיוק הניבוי של התכנית שתגישו, תיאור המודל בצורה ברורה ופורמלית (עם תרשימים ברורים), תיאור האופן בו הוא נלמד, זמן הריצה, אלגנטיות הקוד, וכו'.

בנוסף, הקפידו על מצגת ברורה ומסודרת, וגרפים ברורים ומסודרים. ניתן **דגש** גם לאיכות המצגת והגרפים בבואנו לתת ציון.

## בהצלחה!



פלייליסט מומלץ לחלק 1: