

אלגוריתמים בביולוגיה חישובית (76558)

תרגיל 2: EM וזיהוי איו CpG לפי רצך

תאריך הגשה: 07/12/2025

2.....	חלק 1: מימוש EM למציאת מטבע מותה
2.....	סיפור המסגרת (בדי לתת מוטיבציה)
2.....	תזכורת לגבי נוסחאות חשיבות
3.....	תיאור הדאטה – חלק 1
3.....	שאלות ומטרות חלק 1
4.....	חלק 2: מימוש מודל לזיהוי איו CpG מדאטה קיימ
4.....	תיאור הדאטה – חלק 2
7.....	מטרות חלק 2
7.....	המלצות (לא באמת המלצות)
7.....	דרישות חלק 2
8.....	ויש גם תחרות (!)
9.....	מה להגיש?
9.....	הוראות בליליות למצגת
9.....	חלק 1
9.....	חלק 2
9.....	וביחד
10.....	שימוש בבבלי AI
10.....	קריטריוני ציון

חלק 1: מימוש EM למציאת מטבע מותה

בחלק זה נתרגל את השימוש באלגוריתם ה-EM.

סיפור המסגרת (בדי לחת מוטיבציה)

Q (ולא ג'יימס בונד 😞), נשלח על ידי W לקדינו רויאל בМОונטנגוRo במשימה ללימוד את השימוש באלגוריתם ה-EM בחלק ממשימה לנצח משחק מזל של הטלת מטבעות.

מהלך המשחק: הדילר מטיל מטבע 55 פעמים, בכל פעם מהמרים מוחשים על איזה צד המטבע ייפול ומהמרים בהתאם. המטבע המוטל יכול להיות מותה או לא מותה. למטבע יש במובן שני צדדים (עץ או פלי).

על Q ללימוד 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 לזכות את אירופי הchallenge בזמן המתאים ולהמר בהתאם כדי למסם את בספי הזכיה לטובת הבתר הבריטי 🎉.

המשימה כולמת בחלק זה תריהה ללמידה פרמטרים **למודל HMM** באשר ורק התצפויות ידועות. בלומר, עליכם למצוא תצרכו למצוא את ערבי המטריצות π , A . בעזרת ערבים אלו ניתן יהיה לסייע לג'יימס לנצח את הקדינו.

זיכרון לגבי נוסחים חשובות

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

אלגוריתם באומ-וילש

- אתחול θ [רנדומית]
- בכל איטרציה t
- לבל רצף X חשבו את הפורוורד והבקוורד (בעזרת θ הנוכחי)

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

$$\tilde{\pi}_{k,l} = \frac{N_{k,l}}{\sum_m N_{k,m}} \quad \tilde{e}_k(x) = \frac{N_{k,x}}{\sum_y N_{k,y}}$$

- עיצרו אם השיפור בלוג הנראות [או אם השינוי ב- θ] קטן מספיק

תיאור הדאטה – חלק 1

במסגרת המשימה Q ישב בקזינו זמן רב וערך תצפיות על סדרות הטילות. כאמור, כל סדרה שבצדאת מכילה 50 הטילות. הוא דיווח את הסדרות בקובץ, שנשמר בפורמט **fasta**. הקובץ מביל שורות של סדרות הטילות. צוות הקורס והה שיפ (בעל הקזינו) הגדרו את ההתיה של כל מטיבע ואת ההסתברות להחלף בין המטבעות. שורה לדוגמה בקובץ:

>seq1

THHHHTTTTTHTHTTTHHHHTTTTHTTHHHHHHHHHHTTHHTHHTTTTTTTHT

שאלות ומטרות חלק 1

עליכם לכתוב תבנית שתקבל קובץ **fasta** עם רצפים שיוצרו באופן מלאכותי (הרצפים יוצאו לפי הסתברויות מעבר ופליטה מסוימות), ולדוח על הסתברויות שאתם הגעתם אליהן. בלומר, למש את אלגוריתם-EM של אום-וולש (במתואר בפסאודו-קוד). לנוחותכם מצורף לצד חלק זה (אין חובה להשתמש, אבל מומלץ להשתכל טרם העבודה). אסור להשתמש במימוש קיים ממספריה של האלגוריתם, **עליכם למש אותו בעצמכם**.

הנה מספר שאלות מוחות שעלייכם לענות עליהם (ויסיעו להבין):

- מה הסתברות הפליטה של כל מטבע בכל מצב חבוי? האם ההסתברות הגיונית? לאיזה כיוון מוטה המטבע המוטה? (שים לב, האיתחולים הרנדומליים משפיעים איז הריצזו את האלגוריתם מספר פעמים לפני שת ביבריעו |)
- מה הסתברות המעבר בין המצבים החבויים? תוכלו להסביר מדוע היא הגיונית?
- עלייכם להראות גраф המראה את תהליך המקסימיזציה, בלומר את השינוי ב- $(\theta|X^j)$ (הציגו את מספר האיטרציה בציר ה-X ואת ערך ה- $(\theta|X^j)$ log (בסכימה על כל הרצפים) בציר ה-Y) , שבנוו אותנו שהאלגוריתם מתכנס. האם הגרף שקיבלתם הגיוני להתנהגות האלגוריתם?
- הריצזו את האלגוריתם עם שינוי במספר האיטרציות המקסימליות, והסתבלו על הנראות הסופית שהתקבלה, איך הגרף מתנהג? האם בכלל שמספר האיטרציות המקסימליות עולה בהכרח הנראות משתפרת?
- נסו את האלגוריתם שלכם עם אתחולים שונים (למשל, אתחול אחד לסתברויות הפליטה, אתחול שונה להסתברות הבחירה הראשית ועוד) איזה שינוי אתם צופים בכל אתחול? תוכלו להסביר אותו לפי הנוסחה שפותחה בביתה? גם בסעיף זה עלייכם להראות גраф המראה את תהליך המקסימיזציה, בלומר את השינוי ב- $(\theta|X^j)$ log (הציגו את מספר האיטרציה בציר ה-X ואת ערך ה- $(\theta|X^j)$ log (בסכימה על כל הרצפים) בציר ה-Y) אולם בעת כלל אתחול שבחורתם.

חלק 2: מימוש מודל לזיהוי איי Cp מדאטה קיימים

בחלק זה אתם מתבקשים לบทוב תוכנה שתזהה אiei CpG בתחום רצף דג"א ארוך. בנויגוד לחלק הקודם, הפעם למידת הפרמטרים למודל ה-HMM שלנו קלה יותר. למה? בגלל שהדעתה שלנו מוביל תוצאות לאירועי ה-CpG-נובל למדוד את המטריצות π , ζ ישירות ממנה. הפעם, במקרים סיפור מסגרת יש סיפורו מודיע.

רקע ביולוגי (קצר ביותר

בגנום, ישם אתרים בהם מופיע הנקלאוטיד C ולאחריו הנקלאוטיד G, אתרים שבאליה נקראים אתרים CpG. CpG אתרים שבהם מופיע ריבוץ גבוה של אתרים CpG. נקראים איי CpG.

Left: CpG sites at 1/10 nucleotides, constituting a CpG island. The sample is of a gene-promoter, the highlighted ATG constitutes the start codon.

Right: CpG sites present at every 1/100 nucleotides, constituting a more normal example of the genome, or a region of the genome that is commonly methylated.

adapted from wikipedia

اذורים שבם מופיע ריבוז גבוה של אתרי CpG, נקראים אי CpG ולهم יש משמעות רבה בבראה וביפוי של גנים. אתרי CpG בעלי חשיבות ביולוגיות בגלל שהם יכולים לעبور תחלוף הנקרא: **מתילציה**. במסגרת תחלוף זה הנוקלאוטיד C עובר שינוי בימי (מקבל קבוצת מתיל), שינוי זה משפיע על היכולת של מבוגנות שעתוק לגשת כדי לשעתק את האתר הממותל (בשים אתרים רבים באלו באזור מרובץ, בלומר, בא). כאמור, אי CpG מבילים ריבוז גבוה של אתרים בהם ניתן לבצע מתילציה ולבן יש להם השפעה נרחבת בבראה וביפוי גנים (ניתן למשל לחשוב שאתרים שבם מתחילה רצף של גן, יטו יותר להסביר אי CpG ובערך ניתן יהיה לבקר את הביטוי של הגן בתאים מסוימים שונים, על ידי הוספת או הורדת מתילציה).

תיאור הדאטה - חלק 2

מצורפים לחלק זה שני קבצים.

1. CpG-islands.2K.seq.fa הוא קובץ fasta, מכובץ בעדרת תוכנת gzip. הוא מכיל ב-1,103 רצפים באורך 2,000 בסיסים כל אחד, המכילים את רצף הדג"א הגנומי של אי CpG בודד, וכן את הרצף הגנומי המקורי ממנו הגדירם.

למשל, הרץ הראשון בקונבנציון:

Volume 28 • Number 1 • January 2003

הדגמא מציג רצח בדה, הלקו מגנום האדם, ברומזות אחד, בקואורדינטות 136857-134858. עוד מופיעים בשם הרצח אורך השולדים מימין ומשמאלו לאי-h-Gc (במקרה זה, 266 בסיסים מימין, 1294 בסיסים משמאלי). **לנוחיותכם**, רצח האי עצמו מסומן בבולוד (בדוגמה זו).

2. CpG-islands.2K.lbl.fa הוא קובץ fasta דומה, שבו מספר זהה של רצפים באוטם האורכים והשמות, המקודדים את מיקומי השולטים והאים בכל אחד מהרצפים המקוריים.

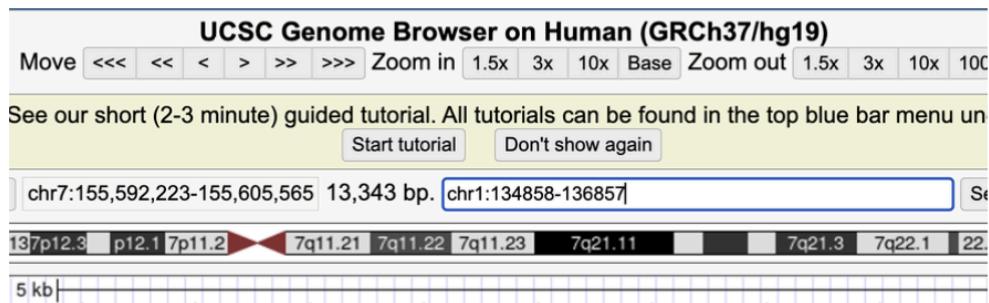
למשל, הרץ' הראשון בקובץ:

שיםו לב שבקובץ זה, A מסמן בסיס (בוקלאוטיד) בלהשו, ורכף ה-C מסמן את מיקום האי.

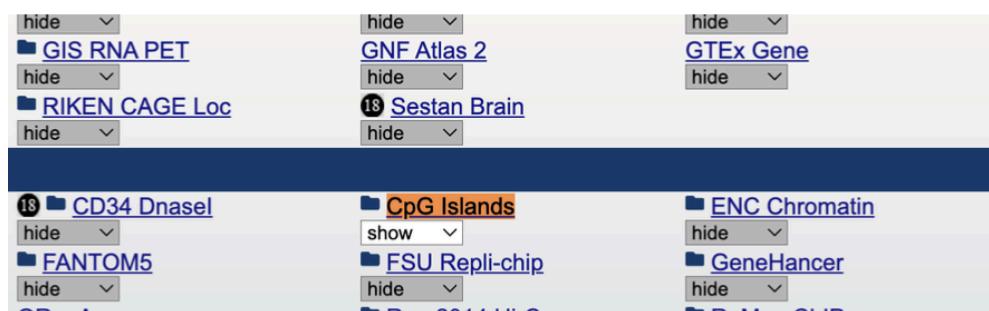
מקור הנתונים:

באופן כללי, כדי למצוא את הרצפים ואת האנוטציות, הרצפים נלקחו מגנים האדם מאתר של אוניברסיטה סנטה קרוז בקליפורניה (קישור לאתר), וכך גם מיקומי האיים, לפי גורסה 19 של גנים האדם (hg19) (הדף הגנומי).

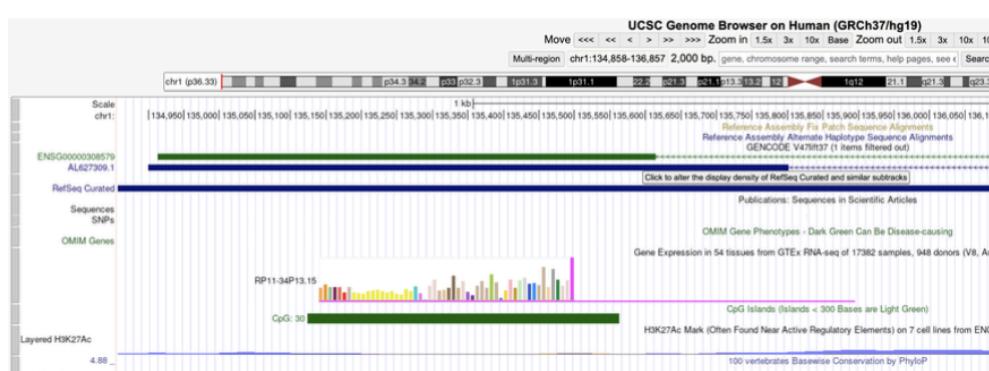
ניתן לגלוש למיקום הגנומי ממנה נלקח הרץ הנ"ל (chr1:134858-136857 2,000 bp)



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



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



הערה חשובה: בקובץ הנ"ל נבחרו כל האיים מכromosome אחד, באורכיהם שבין 150 ל-500 בסיסים.

מטרות חלק 2

עליכם לבתוב תכנית שתקבל קובץ `fasta` מכווץ, בפורמט זהה לקובץ הרצפים הגנומיים (`CpG-islands.2K.seq.fa.gz`), וילצר בפלט קובץ בפורמט זהה לקובץ (`CpG-islands.2K.lbl.fa.gz`), בו יהיה ניבוי טוב ככל האפשר של מיקומי האימים.

המלצות (לא באמת המלצות)

אתם **חייבים** לבנות מודל WMM. אבל, מוזמנים לגלוות יצירתיות במבנה שלו.

למשל:

- תוכלו לאמן שני מודלים מركוביים של רצפים גנומיים, ולהחשב את לוג יחס הניראות של תתי-רצפים בקלט.
- או שאתם יכולים למודד מركובי חבוי, עם שני מצבים (C ו-N), אשר פולט רצפי DN'A. או בזה עם יותר מצבים חבויים.

דgesh: זיברו שהdn'A הוא דו-גדייל, ובди להגדיל את סט האימון שלכם, אתם יכולים גם להפוך את הרצפים **reverse complement** (ובמובן 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 לישם מחדש את המודל שלכם, אם נדרש (אם הוא יהיה ממש טוב אז בגראה שמשש נרצה).
 - אני הקפידו על שרטוטים, תיאור החלקים השונים במודל, הסתברויות המעבר והפליטה, וכן הלאה.
 - אנו נשים על כך **דgesh** בבוינו מתחת ציון לתרגיל.
- **לומוד הפרמטרים:**
 - אני הקפידו לציין בצורה ברורה כיצד למדתם את הפרמטרים השונים של המודל. האם ניחשתם אותם?
 - השתמשתם באומדנים באלה או אחרים (למשל אומד ניראות מירבית, MLE)? כיצד חישבתם את ערכו? מה ההנחה שהנחתם? וכו'.
 - שוב, התיאור צריך להיות ברמה שתאפשר לחזור על צעדייכם בצורה מלאה.
- **זמן ריצה וספריות:**
 - אני הקפידו על זמן ריצה סביר הן בלמידה והן בניתוח רצפים חדשים.
 - השתמשו רק בחבילות פיתוח'ון נפוצות. (מומלץ להשתמש ב-hmmlearn)

ויש גם תחרות (!)

המודל הטוב ביותר (шибvais גם את המודל של צוות הקורס, וגם כל מודל אחר של **בל** משתתפי הקורס) יזכה בפרס מיוחד בפורום ההודעות ובבונוס לציון התרגיל. על מנת להשתתף בתחרות עליהם לודא שהשכלד ממומש ונדרש ללא שינוי. **שים לב**, על המודל שלכם להצליח מול דאטה שעליו הוא לא אומן (test data). לכן, מומלץ שתחלקו בעבודתכם את הדטה ל-**train/validation/test** כך שתוכלו לבדוק את ביצועי המודל שלכם על דאטה שעליו המודל לא אומן.

מה להגיש?

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

הוראות כלליות למציגת

על הקלטת הסבר המציגת להיות באורך של 5–8 דקות. ניתן להציג באנגלית או בעברית. הקלטה (וודיאו) צריכה לכלול את המציגת. למשל, הקלטה של zoom של הציגה באשר המציגת משותפת בשיתוף מסך. המציגת צריכה להיות ברורה ומוסדרת (ראו ממצגות מצורפות עם המלצות בעניין כיצד כדאי לבנות מצגת, תחת לשונית הסקריב במודול, **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 ותשובות חלק 1).
- **קובץ `python` אחד או יותר:** עם התכנית. לנוחיותכם, הבנו לכם קבצים מוכנים עם מספר פונקציות עזר. אני הקפידו על תיאור ותיעוד ברורים של הקוד שלכם, שיאפשרו לנו להבין 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 עליהם נשענתם, חビルות התבונה בהן השתמשתם, הסברים על אימון המודל והזמן/מספר הריצפים שזה לאה. שימוש לב-דרישות חלק 2
- **קובץ `python` אחד או יותר:** עם התכנית. לנוחיותכם, הבנו לכם קבצים מוכנים עם מספר פונקציות עזר. אני הקפידו על תיאור ותיעוד ברורים של הקוד שלכם, שיאפשרו לנו להבין בו ולהבין מה אתם חושבים שעשיתם בקוד. כדי להשתתף בתחרות, שמרו על המבנה בקובץ השלים שסיפקנו לכם.

וביחד

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

בתיבה המתאימה תחת לשוניתו הגדת התרגיל במודל. נא להוסיף שם התקייה את שמות/ת.ז. המגישים .ex2-name1_name2.tar. למשל:

שימוש ברכי AI

השימוש ברכי עדיף לתיכנות מובססי בין מלאותיות (דוגמת chatgpt) מותר, אולם עליכם:

- להציגו על בך במסגרת שקופה במצגת (**במידה ולא השתמשתם, יש להציגו גם על בך**)
- לפרט באיזה כלים השתמשתם (**בקווים בלבדים**)
- מה הפרומפטים שהבנשתם (**בקווים בלבדים**)
- לתאר במילים באופן מפורט את תהליך העבודה על התרגיל עם הכלויים

קריטריוני ציון

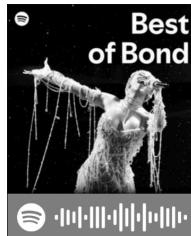
בכל, הקפידו על הדברים הבאים:

בחילק 1: הקפידו על מענה לכל השאלות המנוחות, הגישו את הגרפים הדרושים, ושכנעו אותנו במצב שהבנתם את האלגוריתם.

בחילק 2: הציגו יינטן תוך שכולל דיקן הניבו של התבנית שתגישיו, 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 **דges** גם לאיכות המצגת והגרפים בבעונו לחת ציון.

בהצלחה!



פליאוליסט מומלץ לחילק 1: