# אלגו' בביולוגיה חישובית (76558)
# תרגיל 2: זיהוי איי CpG לפי רצף
תאריך הגשה: 19/12/2024

בתרגיל זה אתם מתבקשים לכתוב תוכנה שתזהה איי CpG בתוך רצף דנ"א ארוך.

**תיאור הדאטה**

מצורפים לתרגיל שני קבצים.

הקובץ CpG-islands.2K.seq.fa.gz הוא קובץ fasta, מכווץ בעזרת תוכנת gzip. הוא מכיל כ-1,103 רצפים באורך 2,000 בסיסים כל אחד. כל רצף מכיל את הדנ"א הגנומי של אי CpG בודד, וכן את הרצפים הגנומיים המקיפים אותו משני הצדדים.

למשל שתי השורות הראשונות בקובץ:

```
>chr1:134858-136857 (266,1294)
TAAAAAATTCGGGCTTGGCGCAGAAACTCACTCCAAATAAATTACCTACCAAAACATTTACATAATGGTGGAAATATTCCAAAATTCAATATTTTGGGATTTATACACAAAAGATAAACAAATTA
GAGGCCAAGAGGCTGCCGGAAGGGAAAAACAGGGCCTGGAATGGCCGACGTGAGGAATGAGCTGGGCCTAAAGAGGCCACTGGCAGGCAGGAGCTGGACCTGCCGAAGTGGCCGAAAGGCAGGAG
CTTTGGACTGGGGAGGCCGCAGTGAGGCGAGAGCTAGCTGGGCGTGGAGAGTCCGCTGTGAGGCCGAGGCCGAGGCTGGGCCCGTGCAGGCCTTCGAGACGCAGGAGGCCGGGCCTGCAAAGGCC
GACTGGAGATCAAGTTCTGCGCCTGAAGAGGCTGCCAAAAGTCAAAAGCGGGGCCTGGGAAGGCCGCCGAGAGCCATGAGCTGGGCTGGGCCGAAAGAGGCCACTGGAGGCAGGAGGAGCTGGG
CCTGGAGAGGCTGACTCGAGGAAGTTTTGCACCTGGAGAGGCCGTCGAGAGGACGGAGCTGGGCCCAGGGAGGCCGACTTGCTGCTCTTCCAGGCCCACTTCCAGGCCGACTTGAGGACGACTTG
GGCCTGCAGAGGCCGCCGGGAGGCTGGAGCTAAGCCTGGAGAGACTGACTTCGGGACGATTTGGCCTGCGGAGGCCGCCGGGAGGCCCAAGCTGGGCCTAGAGGAGCCCACCGACCGGAGGCCA
TTTGGGGCCTGCAGATGTCATCGGAGGGCCAGGAGCTGAGCCTGGAGAGGCCACCGCGAGGCCTGAGCTGGGCCTGGGGAGCTTGGCTTAGGGAAGTTGTGGGCCTACCAGGGCCGCTGGGAGCT
GGGCAGGAGCTGAGTCCAAAGACGTTGTTGGGACCTGGAGTCGGGCCAGAGTCCGGCCTGGAGATGCAGCCGGGAGGAAGAGCTGGGCCCGGAGGGGGCGCCGGGAGGCTGCAAGTGGGTCTGAG
AGGCCAACTTGAGGAGGCCTGGCCTCTGCCTCCCGCATTGCCCAGCTGTTCCTCCTGGCTGCATCTCCCACCTCCCAGCAAACAAGCTCTTTTGGCTCAGCTCCCGCCTGCGTTTGTAGACCCCA
AAGTTTCTGCAACCAAGCTCTTCAGACCCACATCCCTTCTCCCAGTGACTGAACAGTCCCAGCTCCGGCTGGAGAAGGGTGTCTGCAGACCCCGCTGTTGCCTCCCAGGGGAGTCTCCAGGCCCA
GCTCTCGCCCCACCGCGACCTCCCAGGCCCAAGTCCCTGCCTACCTCCCAGCAGCCCGAGTGCGATCCTGTTCCTCCCTCACGGTGGCCTGTTGAGGCAGGGGGTCACGCTGACCTCTGTCCGCG
TGGGAGGGGCCGGTGTGAGGCAAGGGCTCACACTGACCTCTCTCAGCGTGGGAGGGGCCGGTGTGAGGCAAGGGGCTCACGCTGACCTCTGTCCGCGTGGGAGGGGCCGGTGTGAGGCAAGGGCT
CACACTGACCTCTCTCAGCGTGGGAGGGGCCGGTGTGAGGCAAGGGGCTCACGCTGACCTCTGTCCGCGTGGGAGGGGCTGGTGTGAGGCAAGGGCTCAGGCTGACCTCTCTCAGCGTGGGAGGG
GCCGGTGTGAGGCAAGGGGCTCACGCTGACCTCTGTCCGCGTGGGAGGGGCCGGTGTGAGACAAGGGGCTCACACTGACCTCTCTCAGCGTGGGAGGGGCCGGTGTGAGGCAAGGGGCTCAGGCT
GACCTCTGTCCGCGTGGGAGGGGCCGGTGTGAGGCAAGGGGCTCAGGCTGACCTCTGTCCGCGTGGGAGGGGCCGGTGTGAGGCAAGGGGCTCAGGCTGACCTCTGTCCGCGTGGGAGGGGCCGG
GGTGAGGCAAGGGCTCACACTGACCTCTCTCAGCGTGGGAGGGGCCGGTGTGAGGCAAGGGGCTCGGGCTGACCTCTCTCAGCGTGGGAGGGGCCGGTGTGAGGCAAGGGGCTCGGGCTGACCTC
```

מציגה רצף כזה לדוגמא, הלקוח מגנום האדם, כרומוזום chr1, בקואורדינטות 134858-136857.

עוד מופיעים בשם הרצף אורכי השוליים מימין ומשמאל לאי ה-CpG (במקרה שלנו, 266 בסיסים לפני, ו-1294 בסיסים אחרי). לנוחיותכם, סימנתי את רצף האי עצמו בכחול.

הקובץ CpG-islands.2K.lbl.fa.gz הוא קובץ fasta דומה, ובו מספר זהה של רצפים באותם האורכים והשמות, המקודדים את מיקום השוליים והאיים בכל אחד מהרצפים המקוריים. כלומר שתי השורות הראשונות בקובץ זה, יראו כך:

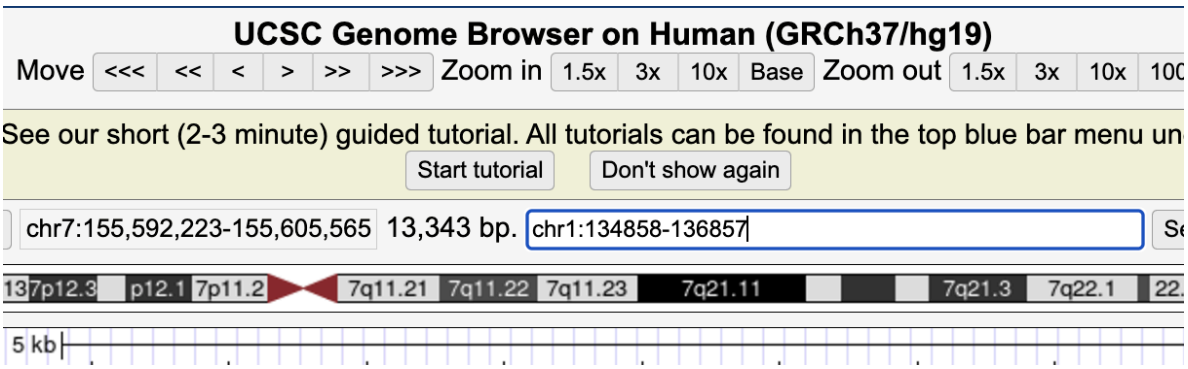
```
>chr1:134858-136857 (266,1294)
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

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

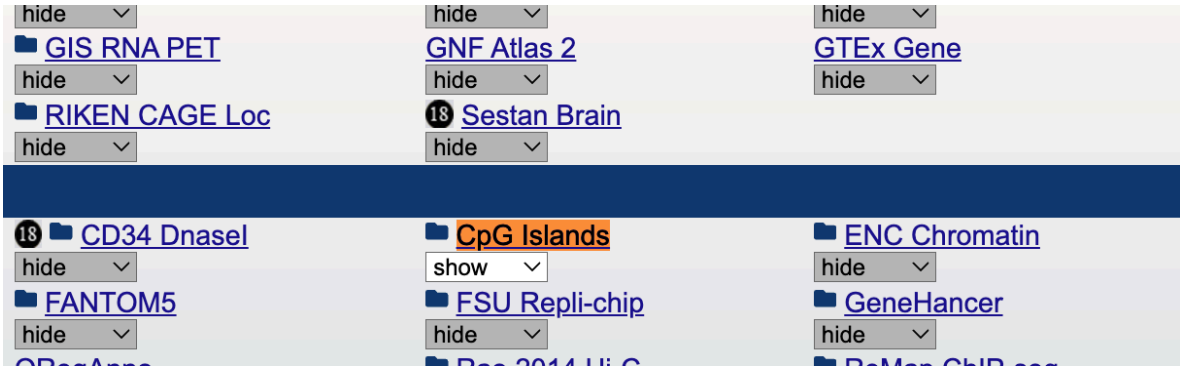
באופן כללי, כדי למצוא את הרצפים ואת האנוטציות, הורדתי את רצף גנום האדם מהאתר של אוניברסיטת סנטה קרוז בקליפורניה, וכן את מיקומי האיים: https://hgdownload.soe.ucsc.edu לפי גירסא hg19 של גנום האדם.

תוכלו למשל לגלוש לדפדפן הגנומי: https://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hg19

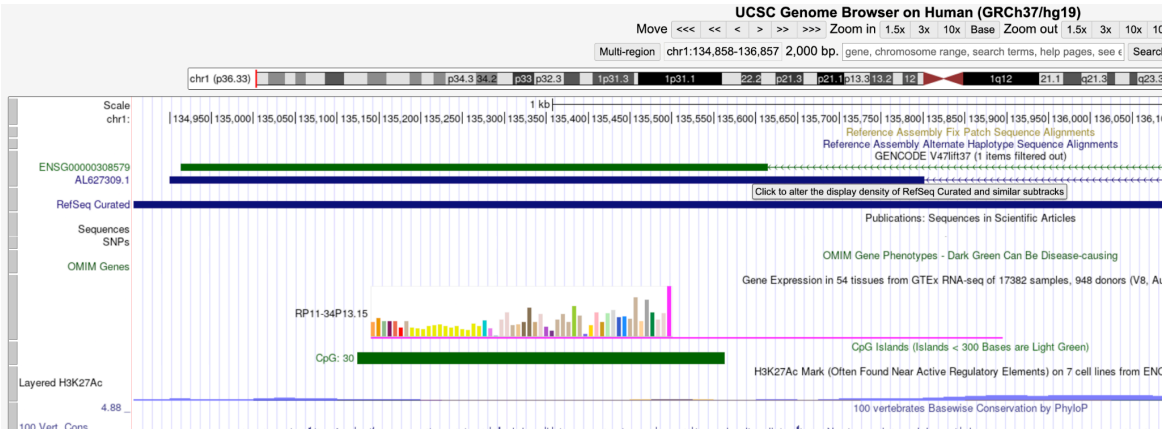
למיקום הגנומי ממנו לקחתי את הרצף הנ"ל




להוסיף אנוטציות של איי CpGs (וללחוץ Represh בצד ימין)




ולראות את האיזור הגנומי, כולל מיקום האי כמלבן ירוק. אם תקליקו עליו תראו עליו מידע, כגון אורכו, מספר אתרי המתילציה, וכן הלאה.




בקובץ הנ"ל בחרתי את כל האיים מכרומוזום chr1, באורכים שבין 150 ל-500 בסיסים.

**מטרת התרגיל**

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

**המלצות**

אתם מוזמנים להשתמש בכל אלגוריתם שתרצו. למשל, תוכלו לאמן שני מודלים מרקוביים של רצפים גנומיים (בתוך אי, ומחוץ לאי), ולחשב את לוג יחס הניראות של לאורך רצפי הקלט.

או שאתם יכולים ללמוד מודל מרקובי חבוי, עם שני מצבים (C ו-N), אשר פולט רצפי דנ"א.

או שאתם יכולים ללמוד מסווג מבוסס על פיצ'רים של bag-of-words על מילים קצרות - עם פרספטרון, או רגרסיה לוגיסטית, או SVM או רשת עמוקה.

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**

עליכם לתאר את המודל על פיו אתם עובדים בצורה ברורה ופורמלית, ברמה שתאפשר לנו ליישם מחדש את המודל שלכם, אם נרצה. אנא הקפידו על שרטוטים, תיאור החלקים השונים במודל, הסתברויות המעבר והפליטה, וכן הלאה. אנו נשים על כך דגש בבואנו לתת ציון לתרגיל.

כמו כן, אנא הקפידו לציין בצורה ברורה מיהם הפרמטרים השונים של המודל וכיצד אתם לומדים אותם. ככל שתשתמשו באומדנים כאלה או אחרים (למשל אומד ניראות מירבית, MLE), אנא הקפידו לתאר מדוע, וכיצד, והאם הם הפרמטרים האופטימליים. מה ההנחות שהנחתם? שוב, התיאור צריך להיות ברמה שתאפשר לחזור על צעדיכם בצורה מלאה.

אנא הקפידו על זמן ריצה סביר - הן בלמידה, והן בניתוח רצפים חדשים, והשתדלו להשתמש בחבילות פיית'ון נפוצות.

אנא הקפידו על תיאור ותיעוד ברורים של הקוד שלכם, שיאפשרו לנו לעיין בו ולהבין מה אתם חושבים שעשיתם בקוד.

**ניתן להגיש בזוגות.**

**דרישות טכניות**

- את הקובץ annotate_cpg.py יהיה ניתן להפעיל מהטרמינל עם הפקודה

```
Python3 annotate_cpg.py --fasta_path input.fa.gz --output_file output.fa.gz
```

- מומלץ, אך לא חובה, להיעזר בשלד התרגיל שמכיל פונקציות לקריאת קבצי האימון וכתיבת התוצאות לקובץ.
- קבצי האימון הם hard coded בשלד התרגיל, ונמצאים בתיקייה data המצורפת.
- על מנת להקל את המימוש, שלד התרגיל כולל אימון מחדש של המודל על קבצי האימון המצורפים בכל קריאה לקובץ, באופן שמתאים לשערוך פרמטרים במודלים מרקוביים וכדומה.
- אם בחרתם להשתמש במודל עם זמן אימון ארוך באופן יחסי (מעל דקה-שתיים), אתם רשאים להגיש את המודל המאומן בקובץ pickle. במקרה כזה יש להקפיד שהקובץ annotate_cpg.py ירוץ כהלכה ללא שגיאות.

**מה להגיש?**

קובץ tar המכיל:

- קובץ PDF ובו תיאור מפורט (בעברית או באנגלית) של הפתרון שלכם. כולל תיאור המודל, תיאור ההנחות עליהן נשענתם, חבילות התוכנה בהן השתמשתם, הסברים על אימון המודל והזמן/מספר הרצפים שזה לקח, דוגמאות ריצה מתוזמנות - כולל התייחסות לטעויות מסוג 1 ומסוג 2, וכו'. אנא ציינו את שמותיכם.
- קובץ python אחד או יותר, עם התכנית. לנוחיותכם, הכנו לכם קבצים מוכנים עם מספר פונקציות עזר.

השימוש בכלי עזר לתיכנות מבוססי בינה מלאכותית (דוגמת copilot) מותר, אולם עליכם **להצהיר על כך בקובץ ה-PDF**, לפרט באיזה כלים השתמשתם, מה הפרומפטים שהכנסתם, ולתאר במילים באופן מפורט את תהליך העבודה על התרגיל.

הציון ינתן תוך שקלול דיוק הניבוי של התכנית שתגישו, תיאור המודל בצורה ברורה ופורמלית, תיאור האופן בו הוא נלמד, זמן הריצה, אלגנטיות הקוד, וכו'.