גל סזנה 318510633
נועה מרגוליס 208635334

תרגיל 1 – אלגוריתמים בביולוגיה חישובית

שאלה 1 – מספר העימודים האפשריים

יהיו s,t שני רצפי DNA באורך n כל אחד.

נרצה להראות כי מספר העימודים האפשריים של s מול t הוא אקספוננציאלי ב-n.

נבחין כי הגודל המקסימלי של כלל הסידור יהיה $2n$.

נפשט את הבעיה (נצמצם אפשרויות סידור) ונגיד בלי הגבלת הכלליות שאנחנו בוחרים מיקומים של הרצף s ומיקומי הבסיסים של הרצף t יוגדרו ככה שהם בוחרים מיקום אקראי בודד בין s_1 לבין s_n.

לכן, הבעיה שקולה לבחירת n מקומות להעמדה של הבסיסים של s מתוך $2n$. וזו בעיה מוכרת בקומבינטוריקה – לבחירת n מקומות להעמדת הבסיסים של s מתוך $2n$ מיקומים קיימות $\binom{2n}{n}$ אפשרויות (עם שמירה על סדר רצף ה-DNA).

נראה כי כבר הביטוי $\binom{2n}{n}$ לבחירת מיקומי בסיסים של s מהרצפים הוא אקספוננציאלי ב-n.

לכן, נקבל כי מספר העימודים האפשריים של s מול t הוא לפחות אקספוננציאלי ב-n.

שאלה 2:

ניתוח זמן ריצה

נסמן את אורך סדרת הנקודות – n, קנס - p, אורך מקסימלי של סגמנט - q.

באלגוריתם אנחנו מוצאים את הארגומנט המינימלי של נוסחת ה-$SSE\ Cost$ עם מוכפל במספר הסגמנטים:

$$\arg\min \sum_{1}^{n} (x_i - \mu_j)^2 + k \cdot p$$

אנחנו עושים זאת על ידי מעבר על כל הנקודות בסדרה, ועבור כל נקודה בודקים מה הסגמנט האידיאלי (שייתן $cost$ מינימלי) שנגמר בה – בודקים עבור כל נקודה את ה-$cost$ עבור כל מקטע שמתחיל עד q נקודות לפניה:

$$SSE\ cost = \sum_{k=i}^{j-1} (x_k - \mu)^2$$

כאשר i היא נקודת ההתחלה של הסגמנט, j היא נקודת הסוף ו- μ הערך הממוצע של הסגמנט.

עבור כל נקודה עדכנו את נקודת ההתחלה ואת מערך ה-$cost$ במידה ומצאנו מינימום חדש.

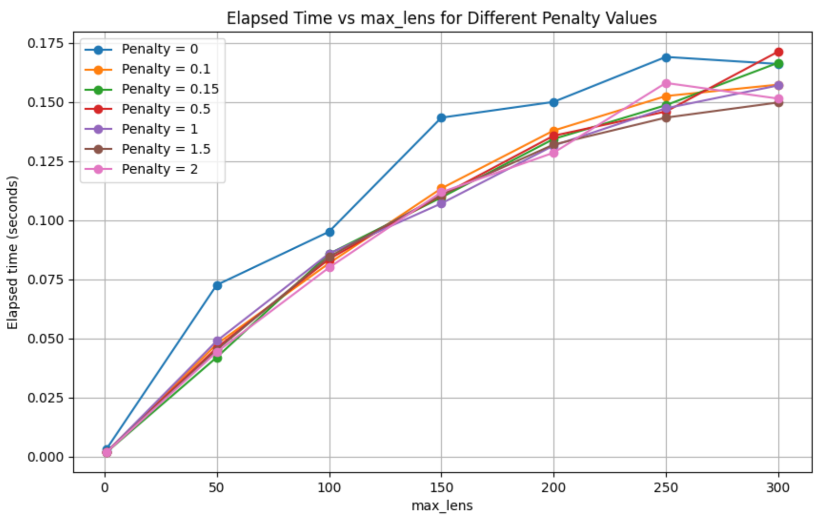
לכן, נקבל כי אנחנו מבצעים $q * n$ חישובים של ה-$cost$, כל חישוב על פי הנוסחה הוא $O(1)$.

לאחר סיום החלק הזה באלגוריתם אנחנו משחזרים את הפתרון האופטימלי שנמצא על ידי מעבר על איברים ממערך t שיצרנו לשם הפתרון – נעבור על כל איבר ב-t שייצג עבור כל $1 \le i \le n$ *t[i]* יציין איפה הסגמנט מתחיל.

במקרה הגרוע ביותר נעבור על n נקודות לכן שלב זה מתבצע ב-$O(n)$ כי לא ייתכנו יותר סגמנטים ממספר הנקודות ($n > k$).

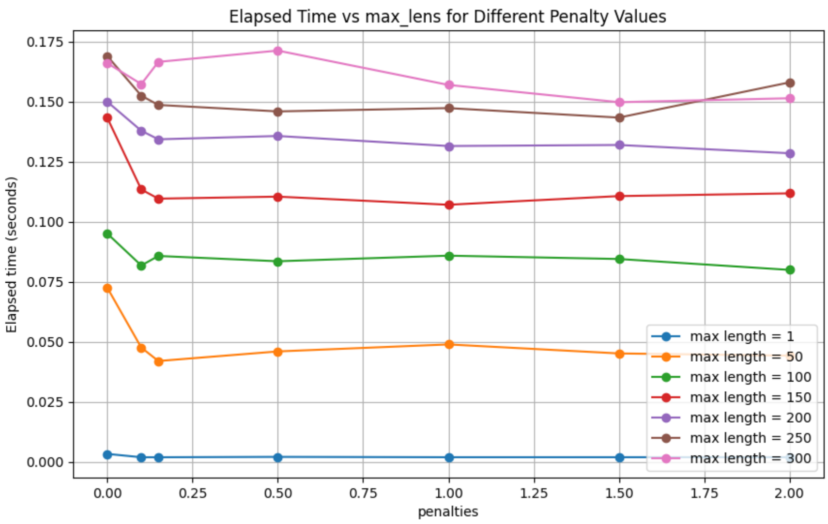
לכן נקבל סה"כ שסיבוכיות זמן הריצה של האלגוריתם הדינאמי הוא $O(n * q)$.

נרצה להראות את השינוי בזמן הריצה כתלות ב-q (אורך הסגמנט המקסימלי):



יצרנו 7 גרפים (גרף לכל ערך *penalty* שונה) ומדדנו את זמן הריצה כתלות בערכי q שונים. ניתן לראות כי אכן רואים קשר כמעט לינארי בין הגבלת אורך הסגמנט המקסימלי לבין זמן ריצה של התוכנית.

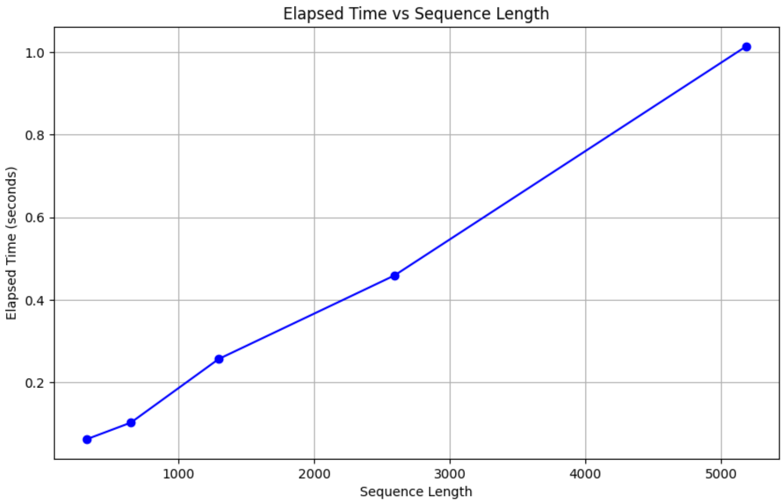
נרצה לוודא כי אכן גודל הקנס לא משפיע על זמן הריצה:



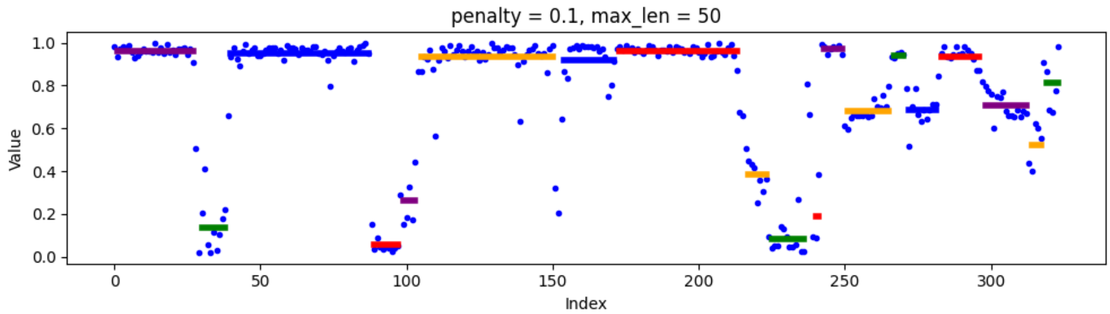
כפי שציפינו, ניתן לראות כי גודל הקנס לא משפיע באופן כלשהו על זמן ריצת התוכנית.

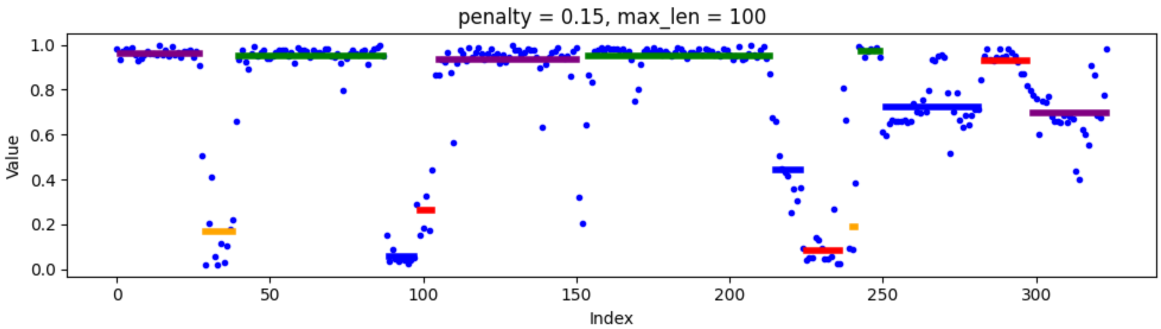
גל סזנה 318510633
נועה מרגוליס 208635334

נראה כי אכן קיים קשר לינארי בין אורך סדרת הנקודות לבין זמן הריצה של התוכנית כאשר q, p קבועים :



כעת, נביט בגרפים שונים על ידי שינוי הפרמטרים q , k עם קובץ ה- *input.txt* שסופק :



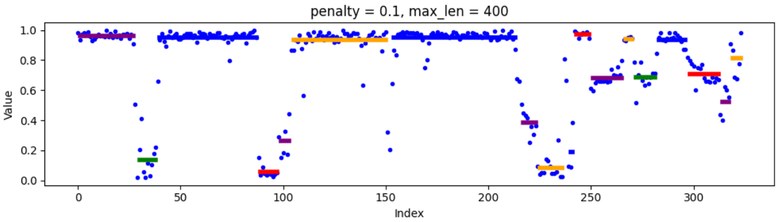


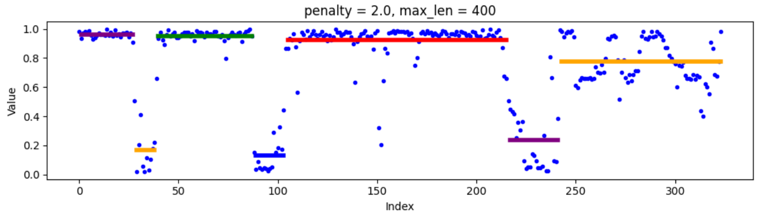
מהשוואה בין הגרפים ניתן לראות כי על בשל העלאת ערך ה- *penalty* בריצה השנייה נוצרו פחות סגמנטים = נוצרו סגמנטים ארוכים יותר, וכן נוצר סגמנט ארוך שכולל יותר מ-50 נקודות (154 עד 214).

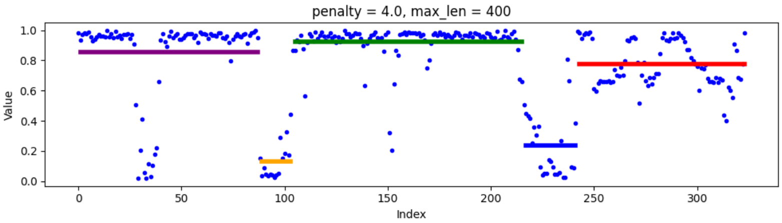
גל סזונה 318510633
נועה מרגוליס 208635334

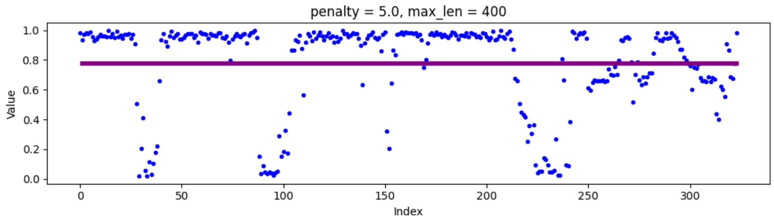
בחינת מקרי קיצון:

ללא הגבלת *max_len* עם *penalty* הולך וגדל:



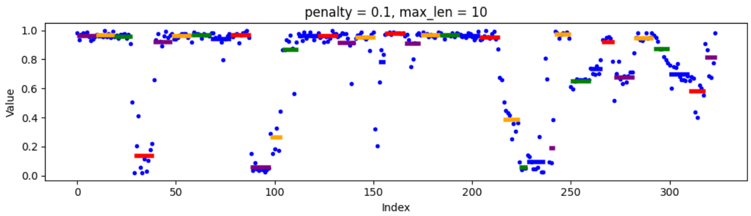


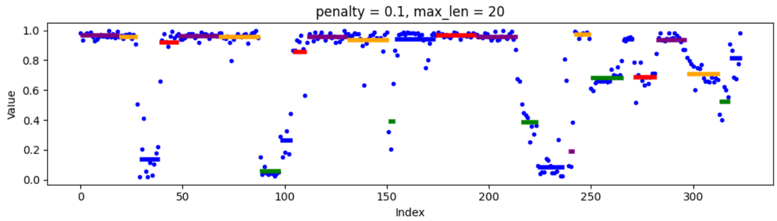


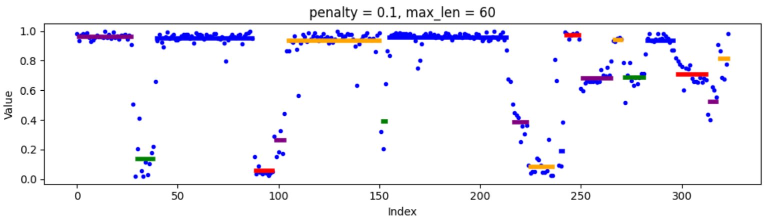


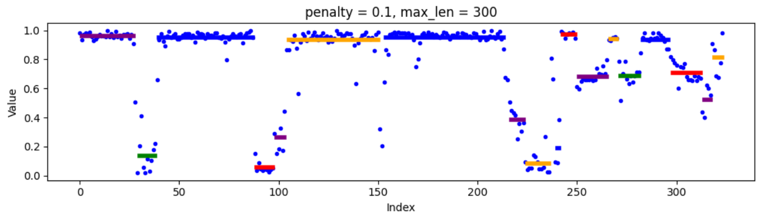
כפי שניתן לצפות, רואים שככל שנעלה את הקנס על הסגמנטים נקבל פחות ופחות סגמנטים.

penalty קבוע עם *max_len* משתנה:







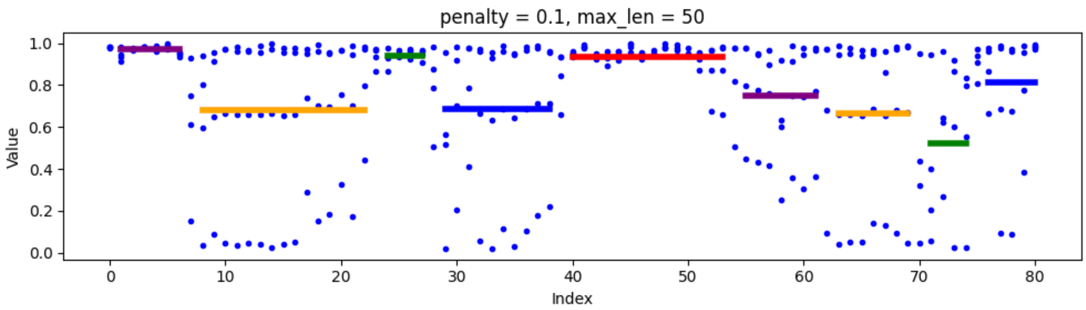


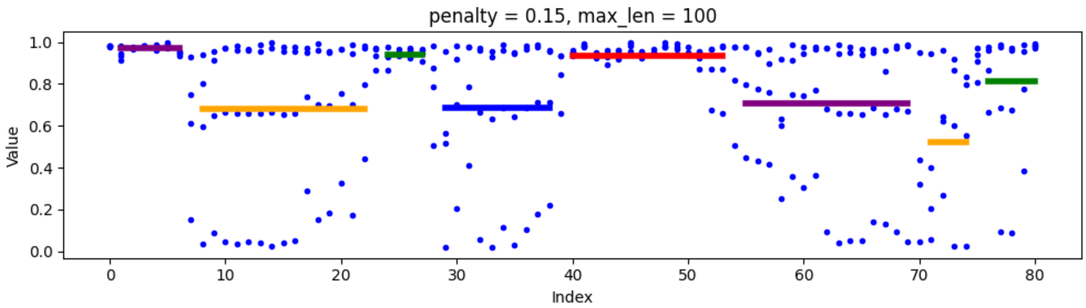
באופן מעניין, ניתן לראות מרצף הגרפים כי ככל שאנחנו מאפשרים אורך מקסימלי גדול יותר (10,20,60) נראה כי אכן נקבל סגמנטים באורכים גדלים, אבל זה עד לשלב מסוים שבו התוכנית תגיע למיקסום למינימום של ערך ה-*cost* עבור חלוקה לסגמנטים מסוימת והחל משלב זה העלאת אורך הסגמנט המקסימלי לא תשפיע על החלוקה לסגמנטים (כפי שרואים בהשוואה בין *max_len=60* ו-*max_len=300*).

גל סזנה 318510633
נועה מרגוליס 208635334

חלק בונוס – סגמנטציה למידע רב ערוצי:

נריץ את קובץ הקלט של חלק הבונוס עם ערכים משתנים של *penalty* ו- *max_len* :





בגרפים שמנו את כל הנקודות מכל הערוצים על הגרף וניתן לראות כי אכן הסגמנטים הנבחרים מצאו איזון כלשהו בין כל הערוצים השונים. ואף ניתן לראות כי גם ערך ה- *cost* גבוה יותר מערך ה– *cost* כאשר חישבנו בערוץ בודד בחלק הקודם.