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 על כל סגמנט - p, אורך מקסימלי של סגמנט - q.

מטרת האלגוריתם למצוא את הארגומנט המינימלי של נוסחת ה-$Cost\ SSE$ (סכום הפרשי הריבועים) על מנת להגיע למזעור המרחקים של כל נקודה איבר בסדרה שנקבל כקלט) לערך הממוצע בסגמנט אליה נשייך, בתוספת קנס p:

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

אופן פעולת האלגוריתם הינו מעבר על כל הנקודות בסדרה (קיימות n כאלה כמספר האיברים), ועבור כל נקודה בודקים מה הסגמנט האידיאלי (שייתן $cost$ מינימלי) שנגמר באותה נקודה – כלומר נקודת התחלה אידיאלית, 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.

גל סזנה 318510633
נועה מרגוליס 208635334

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

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

בנוסף, לחישוב הSSE של הסגמנט בצורה יעילה השתמשנו בשני מערכי עזר:

האחד $comulative\ sum$ לחישוב סכום מצטבר – $O(n)$

השני $comulative\ sum\ squared$ לחישוב סכום ריבועי מצטבר – $O(n)$

כך שבכל פעם יכולנו לחשב את הממוצע של הסגמנט ואת הנוסחה של SSE במספר פעולות קבוע (חמש).

סך הכל לאלגוריתם הדינאמי - $O(n) + O(n \cdot q)$

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 $1 \leq i \leq n$ $t[i]$ יציין איפה הסגמנט מתחיל. $O(K)$ למצוא K סגמנטים כאלו.

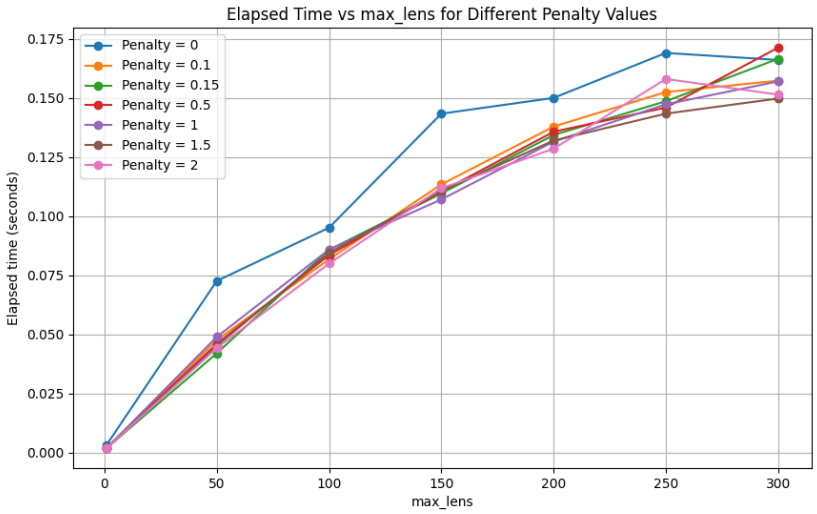
זמן הריצה הכולל:

$$O(n) + O(n \cdot q) + O(n)$$

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 $O(n * q)$.

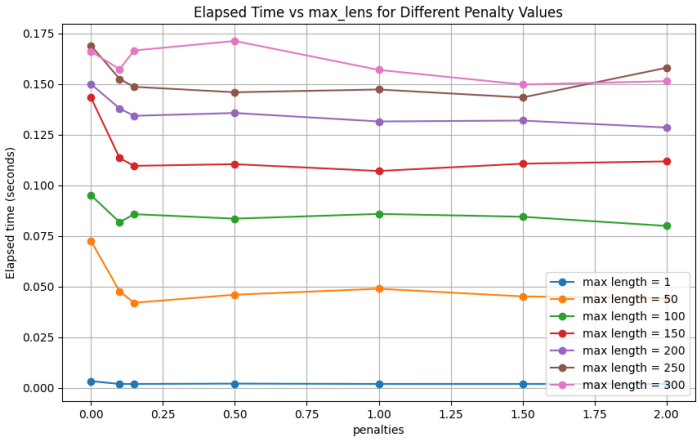
נרצה להראות את השינוי בזמן הריצה כתלות ב-q (אורך הסגמנט המקסימלי):



יצרנו 7 קווים (לכל ערך $penalty$ שונה) המציגים 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 – ככל שהגבלת הגודל גדולה יותר זמן הריצה עולה.

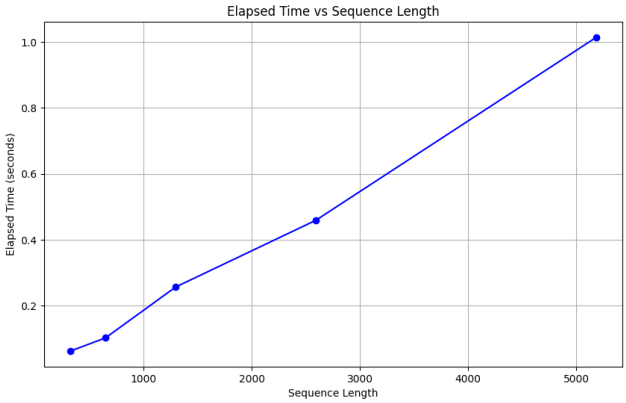
גל סזנה 318510633
נועה מרגוליס 208635334

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

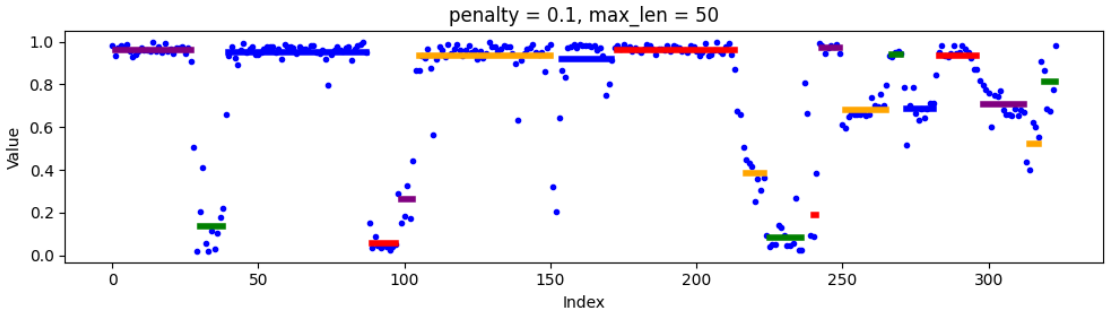


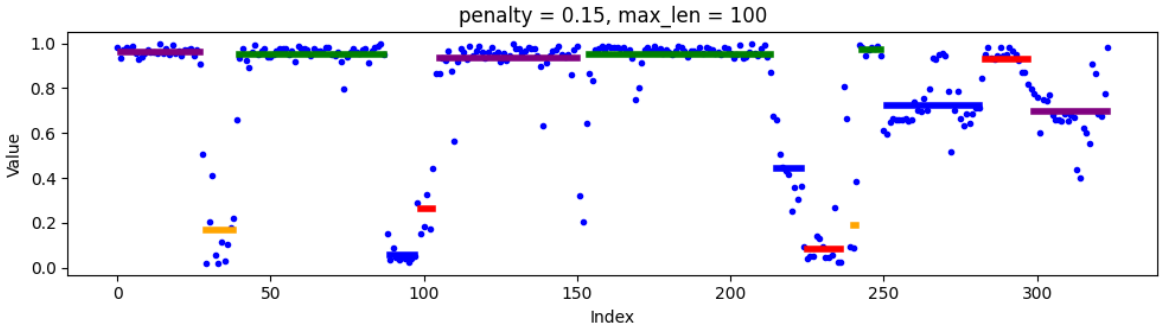
כפי שציפינו, ניתן לראות כי לא ניתן להצביע על קשר בין גודל הקנס לבין זמן ריצת התוכנית.

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:



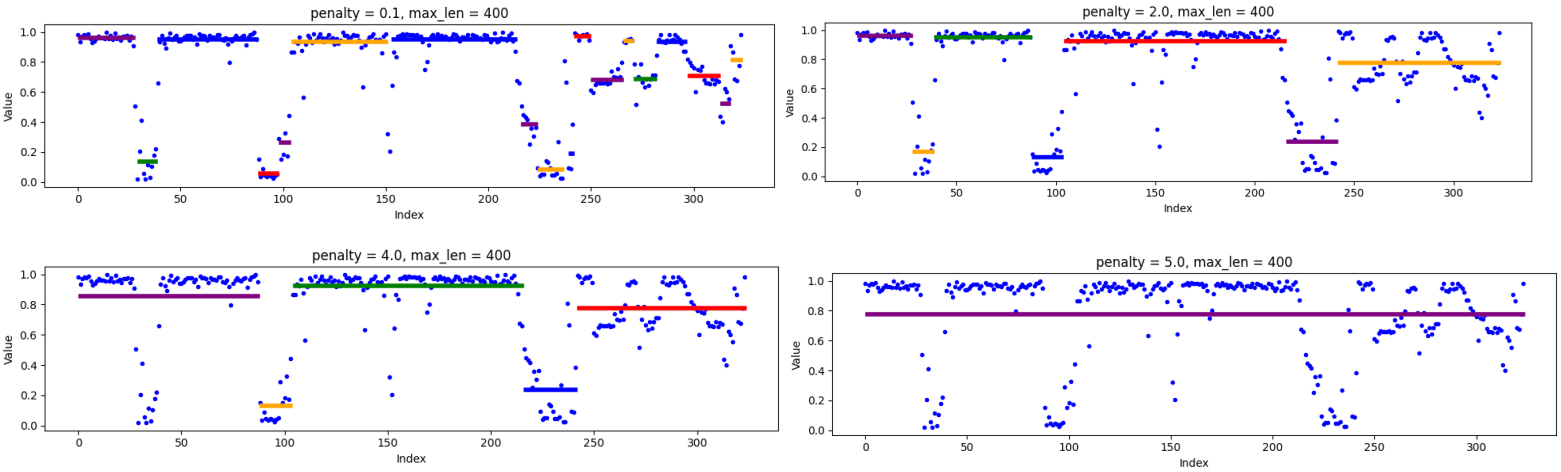


גל סזנה 318510633
נועה מרגוליס 208635334

מהשוואה בין הגרפים ניתן לראות כי בשל העלאת ערך ה- $penalty$ בריצה השנייה נוצרו פחות סגמנטים וכתוצאה מכך הסגמנטים ארוכים יותר (וכן כי אפשרנו אורך ארוך יותר לכל סגמנט), וכן נוצר סגמנט ארוך שכולל יותר מ-50 נקודות (154 עד 214).

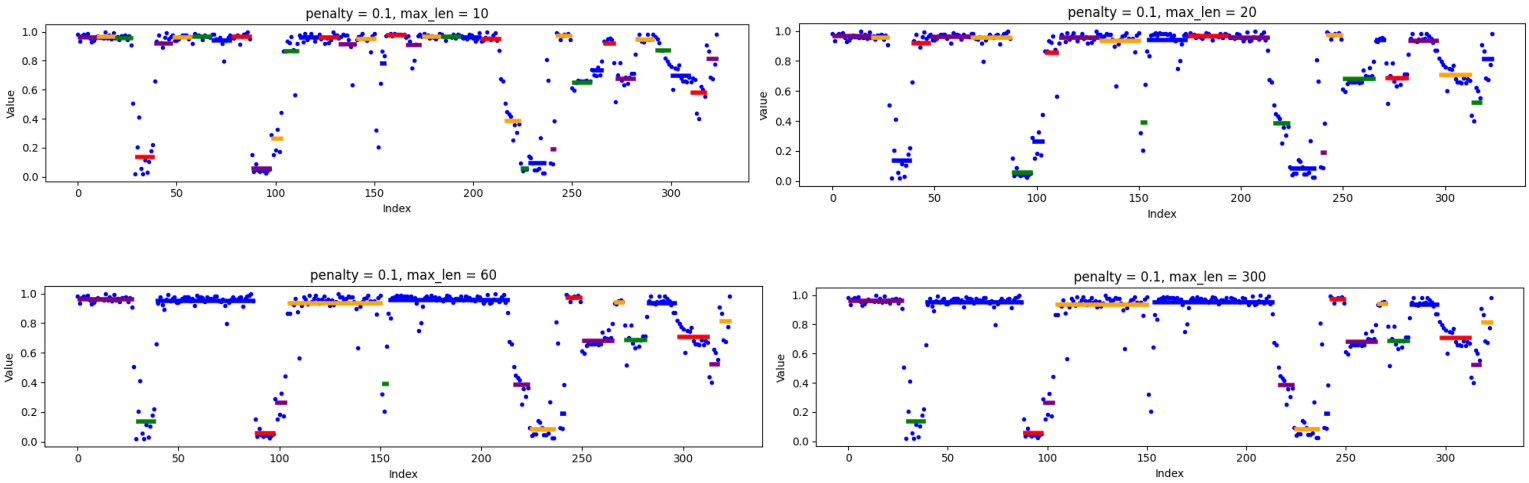
בחינת מקרי קיצון:

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:



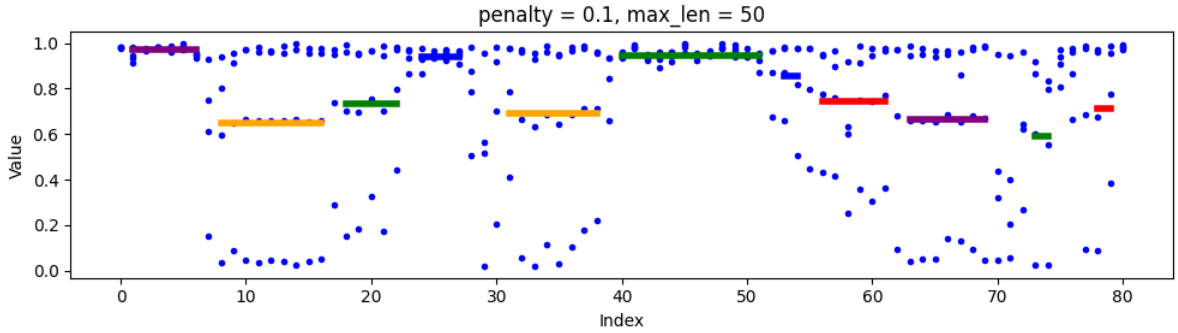
מעניין לראות מרצף הגרפים כי ככל שאנחנו מאפשרים אורך מקסימלי גדול יותר (10,20,60) נראה כי אכן נקבל סגמנטים באורכים גדלים, אומנם עד לשלב מסוים שבו התוכנית תגיע למינימום של ערך ה-$cost$ עבור חלוקה לסגמנטים מסוימת והחל משלב זה העלאת אורך הסגמנט המקסימלי לא תשפיע על החלוקה לסגמנטים (כפי שרואים בהשוואה בין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

גל סזנה 318510633
נועה מרגוליס 208635334

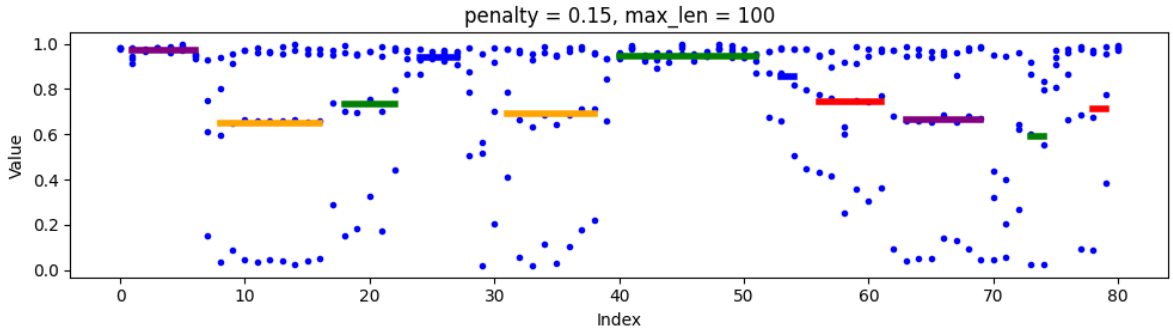
חלק בונוס – סגמנטציה למידע רב ערוצי:

נריץ את קובץ הקלט של חלק הבונוס עם ערכים שונים של $penalty$ ו- max_len:

```
1 7 3.88
8 17 2.588
18 23 2.924
24 28 3.768
29 29 3.152
30 30 2.06
31 39 2.768
40 52 3.784
53 55 3.412
56 62 2.972
63 70 2.668
71 72 1.684
73 75 2.372
76 77 3.552
78 80 2.844
81 81 3.924
2.298
```




```
1 7 3.88
8 17 2.588
18 23 2.924
24 28 3.768
29 29 3.152
30 30 2.06
31 39 2.768
40 52 3.784
53 55 3.412
56 62 2.972
63 70 2.668
71 72 1.684
73 75 2.372
76 77 3.552
78 80 2.844
81 81 3.924
3.048
```



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.