# תרגיל 1 – אלגוריתמים בביולוגיה חישובית

**שאלה 1 – מספר העימודים האפשריים**

יהיו  *שני רצפי DNA באורך n כל אחד.*

*נרצה להראות כי מספר העימודים האפשריים של מול 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 לבין .*

*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 אפשרויות (עם שמירה על סדר רצף ה-DNA).*

*נראה כי כבר הביטוי לבחירת מיקומי בסיסים של s מהרצפים הוא אקספוננציאלי ב- n.*

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

*בנוסף, נצין כי עבור כל עמודה נפרדת (שלב V\_i) ניתן לבחור בשלוש אפשרויות לעימוד והן -*

*תו מ-s ומתחתיו תו מt (si,tj)*

*תו מ-t ומתחתיו רווח (si,-)*

*רווח ומתחתיו תו מt- (-,tj)*

***שאלה מספר 2 - זמן ריצה***

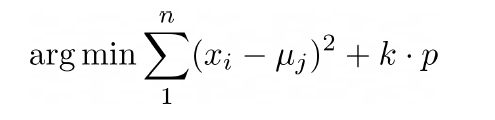
נגדיר (על פי ההוראות)

אורך הסדרה -N

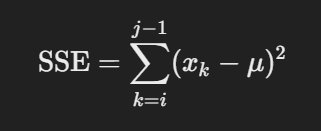
קנס - P

אורך מקסימלי של סגמנט - Q

נשים לב כי השתמש באלגוריתם תכנון דינאמי על מנת להגיע למזעור ריבועי המרחקים של כל נקודה (איבר בסדרה X שנקבל כקלט) לערך הממוצע בסגמנט אליה היא שייכת, בתוספת קנס רגוליזציה P בבלוק. כלומר מזעור הנוסחא:



לשם כך - עבור כל נקודת סיום אפשרית לסגמנט (קיימות N כאלה כמספר האיברים) חישבנו עבור כל נקודת התחלה אפשרית לסימנט (קיימות Q כאלה כגודל הסיגמנט המקסימאלי) את ערך ה sum of squared errors



כאשר i נקודת ההתחלה של הסיגמנט j נקודת הסוף וu הערך הממוצע של הסיגמנט.

עבור כל נקודת סיום עידכנו את נקודת ההתחלה ואת מערך הCOST במידה ומצאנו מינימום חדש.

לכן N\*Q פעולות בהן בכל פעם חישבנו את ה-SSE:

לחישוב הSSE של הסיגמנט בצורה יעילה השתמשנו בשני מערכי עזר-

האחד cumulative\_sum

לחישוב סכום מצטבר - לקח OׂN

השני cumulative\_sum\_squared

לחישוב סכום ריבועי מצטבר - לקח גם O(N)

כך שבכל פעם יכולנו לחשב את הממוצע של הסיגמנט ואת הנוסחא של SSE במספר פעולות קבוע (חמש - כלומ O של אחד).

סך הכל לאלגוריתם הדינאמי - O(N⋅Qׁ

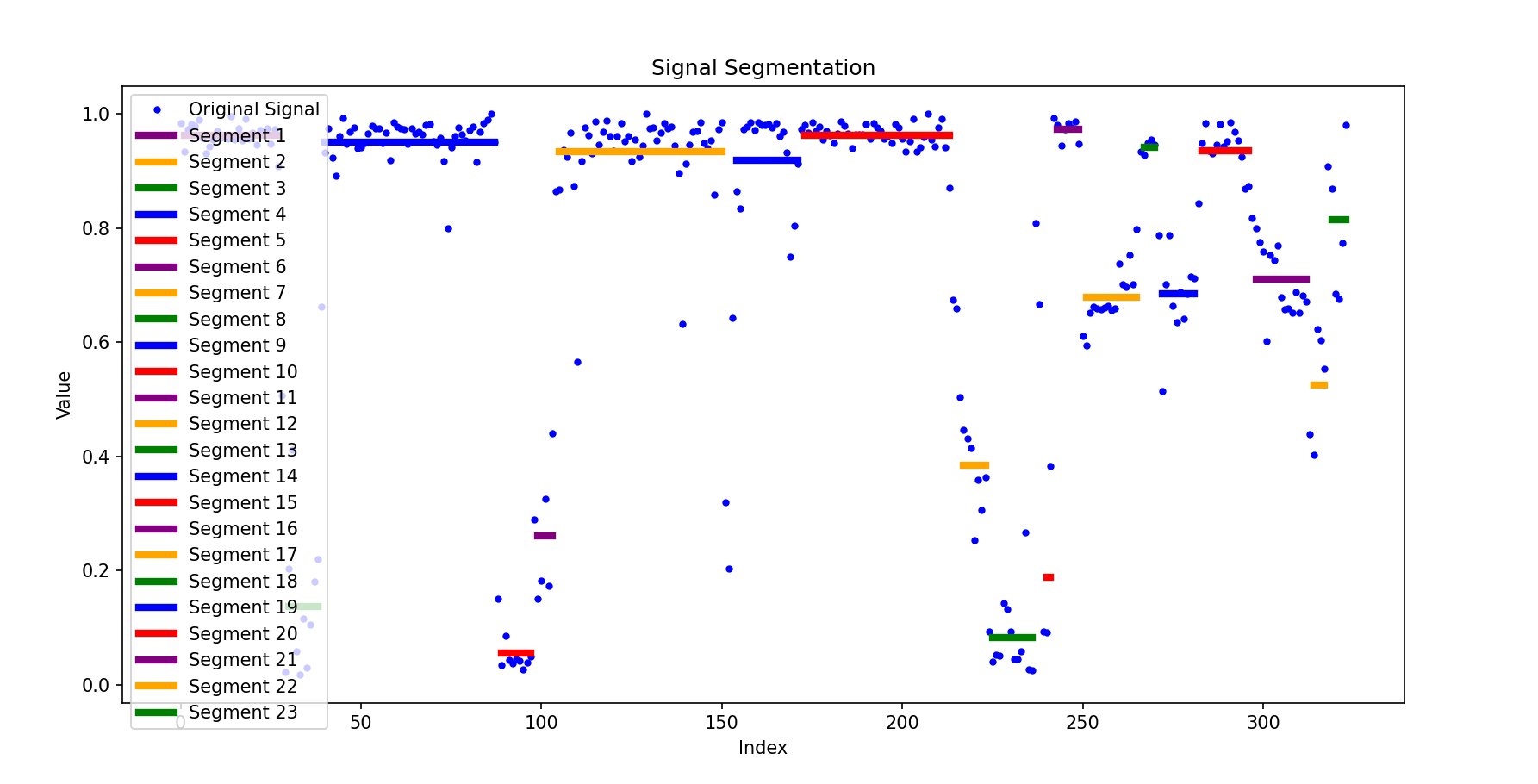
בנוסף לטבות בקטרקינג לשחזור הסיגמנטים - נתחיל מN ונלך לאחור - מאחר ויש K סגמנטים למצוא נקודת התחלה וסוף וחישוב הממוצע (מספר קבוע של פעולות) יקח לנו OK

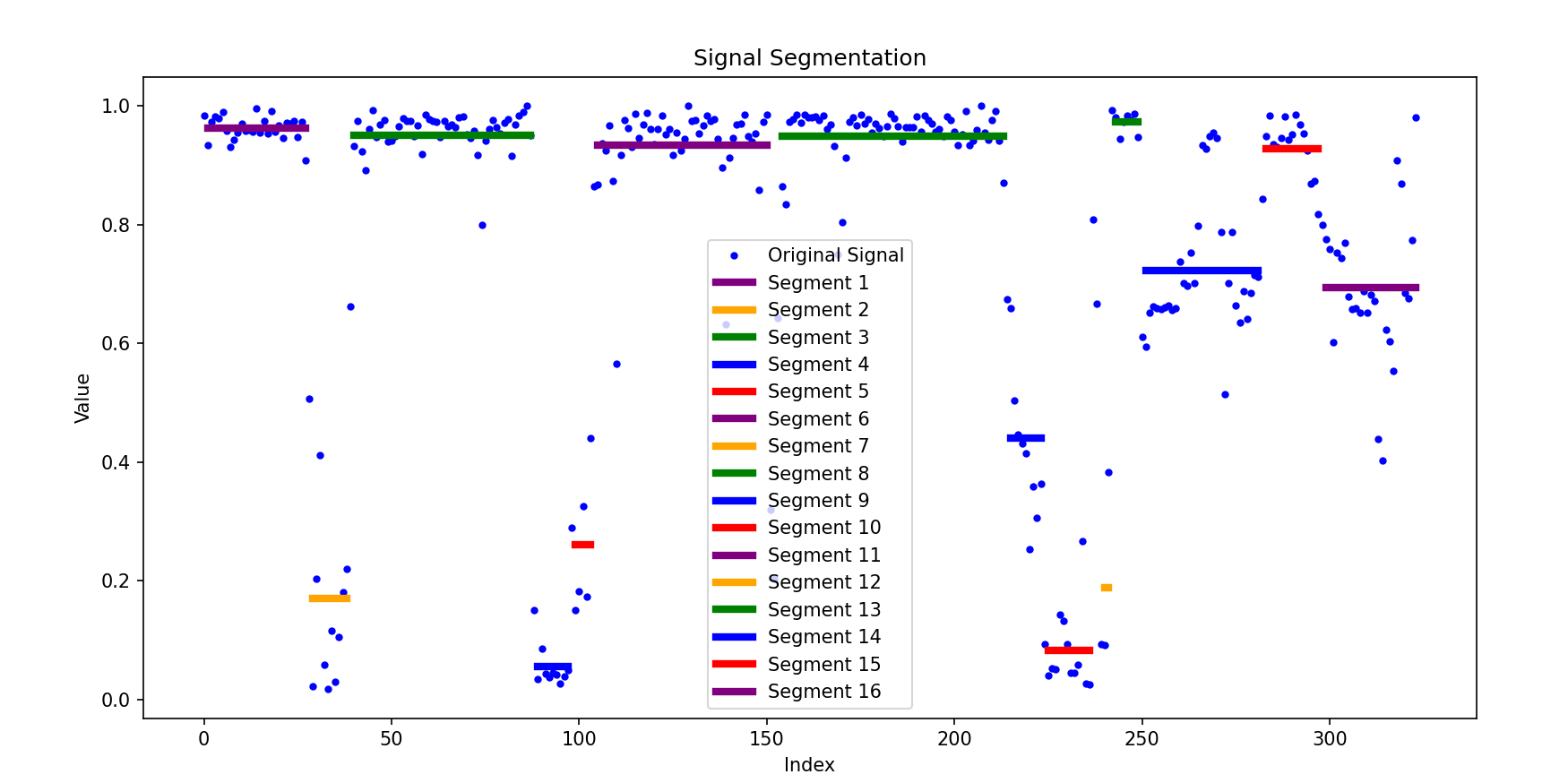
סך הכל קיבלנו שזמן הריצה הכולל הינו:

O(N)+O(N⋅Q)+O(K)

דוגמאות לתוצרי גרפים

--penalty 0.1 --max\_len 50



--penalty 0.15 --max\_len 100

מקרים מיוחדים -

Q=O(1)

דוגמא לגרף עם עונש נמוך ו-Q=1

מההצבה בנסוחת זמן ריצה שחישבנו - (O(N+K

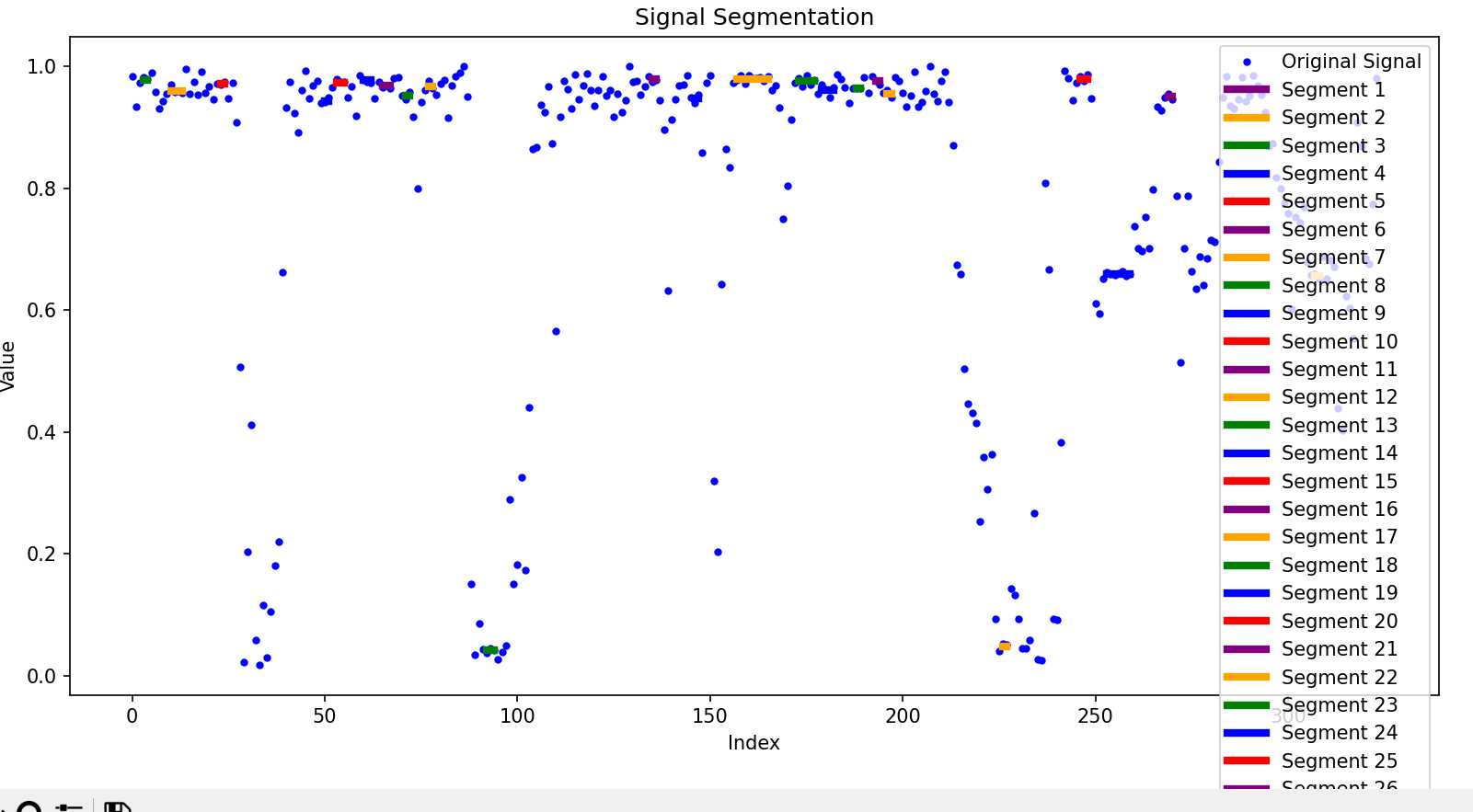
(כל נקודה סיגמנט נפרד)

Worst case (Q=O(N), K=O(N)

O(N^2)

דוגמא לגרף עם עונש נמוך וQ=N

הרבה סיגמנטים כי העונש נמוך וגודל הסיגמנט אינו מוגבל



דפוסים נוספים - ככל שהpanelty בכל שלב עולה - פחות סיגמנטים - כי כל שלב עולה יותר

