אלגוריתמים בביולוגיה חישובית / תרגיל 1 – עימוד רצפים

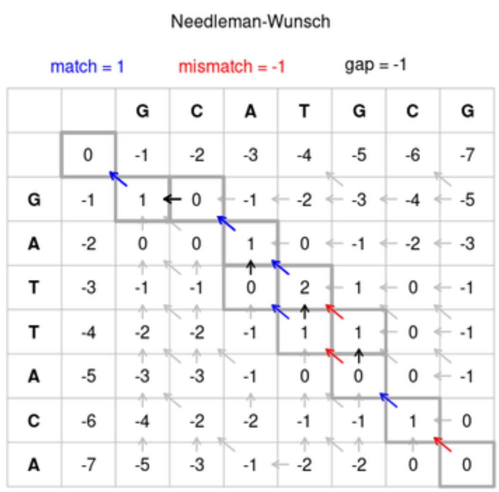
אבי קופינסקי 318336070

עדי רבינוביץ 314952540

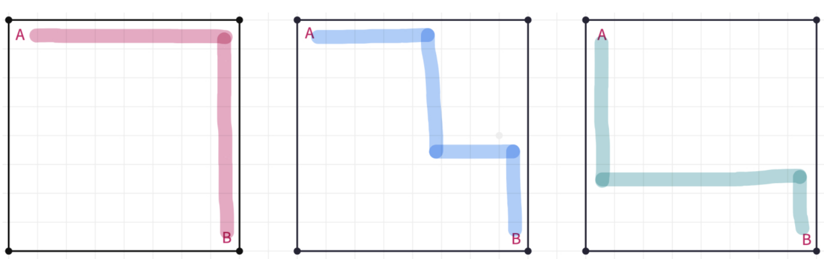
חלק I – חלק תיאורטי

1. מספר העימודים האפשריים

- יהיו שני רצפים S, T, באורך n כל אחד. נוכיח כי מספר העימודים האפשריים אקספוננציאלי ב־n.
- נמצא את העימודים האפשריים תחת הנחות אלו בעזרת האלגוריתם Needleman-Wunch שבעזרנו ניתן לעמד רצפים גלובלים.
- כזכור, האלגוריתם מבוסס על מילוי מטריצת V_{ij} המגדירה את ציון העימוד המיטבי עבור הרצפים S ו־T, כאשר את העימוד המתאים לציון זה (העימוד המיטבי) ניתן למצוא באמצעות מילוי טבלה נוספת המגדירה מצביעים בהתאם לבחירה (לה קראנו Trace matrix) ובסיום לפעול בעקיבה לאחור החל מהתא המתאים V_{nm}.
- כלומר, סך העימודים האפשריים הוא סך הדרכים החוקיים למסלול מהתא n, n לתא $0, 0$, כאשר הצעדים החוקיים הם ימינה, שמאלה, או אלכסונית למעלה ושמאלה. באופן ויזואלי:



- נוכיח שכמות המסלולים האפשריים של בעיה זו היא אקספוננציאלית ב־n, ובכך נראה כי סך העימודים האפשריים גם הוא אקספוננציאלי ב־n.
- על מנת לעשות זאת, נסתכל על תת בעיה לבעיית כמות המסלולים האפשריים כפי שתיארנו, שהיא כמות המסלולים האפשריים תחת ההנחה שהמהלכים החוקיים הם צעד ימינה ושמאלה בלבד, לדוגמה:



- היות וזו תת בעיה של הבעיה המקורית, מספר המסלולים האפשריים שניתן לתאר הוא קטן ממספר המסלולים מהבעיה המקורית.
- נוכיח שמספר המסלולים בבעיה זו אקספוננציאלי ב־n (באורך הרצפים, שהוא למעשה מימדי המטריצה), ובכך נסיק בפרט שכמות המסלולים בבעיה המקורית אקספוננציאלי ב־n, ולכן כאמור כמות העימודים האפשריים אקספוננציאלי ב־n.

- כל הדרכים האפשריות במטריצה יהיו באורך $2n$ צעדים, כאשר n צעדים מתוכם יהיו כלפי מעלה, ו־n צעדים מתוכם שמאלה.
- נייצג את הצעדים של הדרך על ידי סדרת מספרים $\{1, \ldots, 2n\}$.
- נבחר n צעדים שמאלה מתוך $2n$ מספרים, (באופן סימטרי מתקיים הדבר אם היינו בוחרים n צעדים למעלה). מעקרונות קומבינטורים, מתקיים שמספר המסלולים החוקי הוא $\binom{2n}{n}$.
- כעת, נותר להוכיח ש־ $\binom{2n}{n}$ אקספוננציאלי ב־n.
- נוכיח טענת עזר: מתקיים $0 \overset{n\to\infty}{\to} \frac{\binom{2n}{n}}{4^n}$.
- הוכחה: מקירוב סטירלינג נובע

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{\sqrt{4\pi n}\left(\frac{2n}{e}\right)^{2n}}{\left(\sqrt{2\pi n}\left(\frac{n}{e}\right)^n\right)^2} = \frac{\sqrt{4\pi n}\left(\frac{2n}{e}\right)^{2n}}{2\pi n\left(\frac{n}{e}\right)^{2n}} = \frac{\sqrt{4\pi n}\cdot 4\cdot\left(\frac{n}{e}\right)^{2n}}{2\pi n\left(\frac{n}{e}\right)^{2n}} = \frac{\sqrt{(\pi n)}\cdot 4^n}{\pi n} = \frac{4^n}{\sqrt{\pi n}}$$

לכן

$$\frac{\binom{2n}{n}}{4^n} \approx \frac{\left(\frac{4^n}{\sqrt{\pi n}}\right)}{4^n} = \frac{4^n}{4^n\sqrt{\pi n}} = \frac{1}{\sqrt{\pi n}} \overset{n\to\infty}{\to} 0$$

- מכך ניתן להסיק כי הביטוי $\binom{2n}{n}$ אקספוננציאלי ב־n, כלומר כמות המסלולים האפשריים בתת הבעיה שהגדרנו היא אקספוננציאלית ב־n, ולכן בפרט כמות המסלולים בבעיה המקורית גם הם אקספוננציאלים ב־n.
- לכן, סה"כ מתקיים שמספר העימודים האפשריים אקספוננציאלים ב־n, כנדרש.

2. **קנס אפיני על רווחים ־ AffineGapPanelty**

(א) נגדיר טבלה V בה נשמור את ציוני עימוד הרישאות, כלומר הציון האופטימלי המייצג את העימוד יהיה בתא $V(n,m)$.

(ב) נגדיר טבלה P להיות העימוד האופטימלי (כלומר מטריצת Trace).

(ג) אתחול:

$$V(0,0) = 0$$

$$V(i,0) = d + (i-1)\cdot e \qquad \text{for } i = 1,\ldots,n$$

$$V(0,j) = d + (j-1)\cdot e \qquad \text{for } j = 1,\ldots,m$$

$$P(i,0) = \text{"up"} \qquad \text{for } i = 1,\ldots,n$$

$$P(0,j) = \text{"left"} \qquad \text{for } j = 1,\ldots,m$$

(ד) נגדיר את הפונקציות הבאות:

$$V(i,j) = \max\left(D(i,j), L(i,j), U(i,j)\right)$$

$$D(i,j) = V(i-1,j-1) + \sigma(s_i,t_i) \qquad \text{match } S \text{ with } T$$

$$L(i,j) = \max\begin{cases} V(i,j-1)+d & \text{open a gap in } S \\ L(i,j-1)+e & \text{extend a gap in } S \end{cases}$$

$$U(i,j) = \max\begin{cases} V(i-1,j)+d & \text{open a gap in } T \\ U(i-1,j)+e & \text{extend a gap in } T \end{cases}$$

$$P(i,j) = \begin{cases} \text{"diagonal"} & \text{argmax}([D(i,j), L(i,j), U(i,j)])\text{=0} \\ \text{"left"} & \text{argmax}([D(i,j), L(i,j), U(i,j)])\text{=1} \\ \text{"up"} & \text{argmax}([D(i,j), L(i,j), U(i,j)])\text{=2} \end{cases}$$

(ה) את פונקציות אלו נריץ עבור $i \leq i \leq n, 1 \leq j \leq m$, ובכך נמלא את שתי המטריצות כמתואר.

(ו) העקיבה לאחור תיהיה החל מ־$V(m,n)$, עד שנגיע ל־$V(0,0)$, על ידי עקיבה במטריצה P.

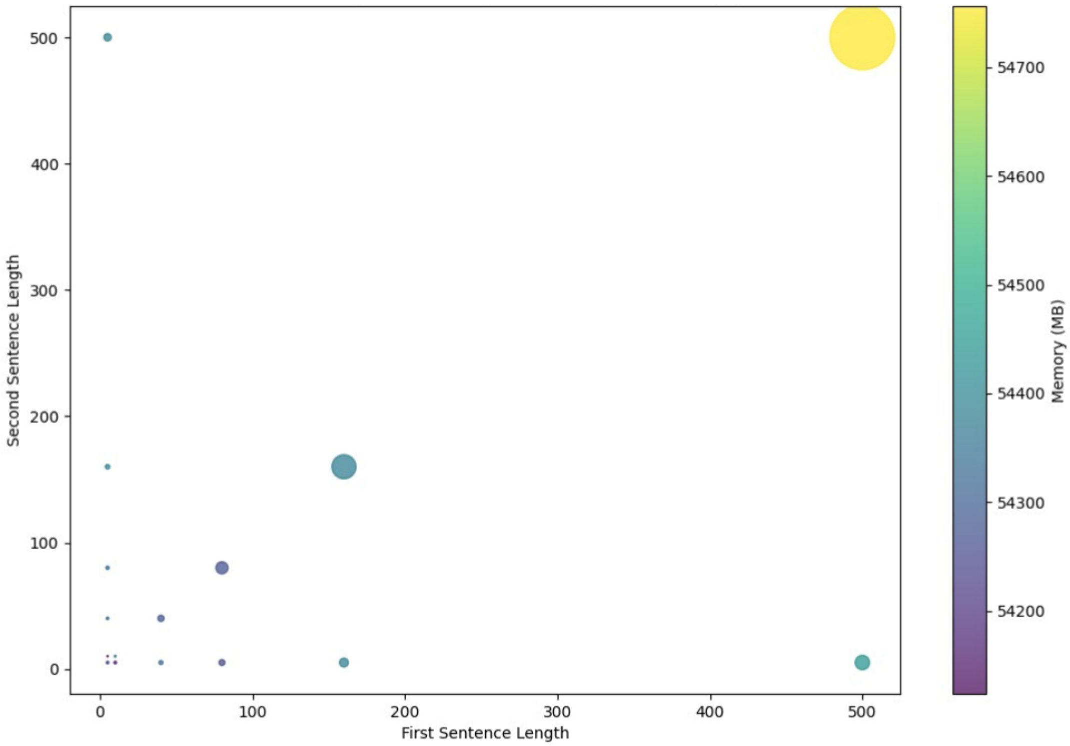
(זוהי למעשה הדרך בה מתבצעת העקיבה לאחור באלגוריתם Needleman-Wunch המקורי, כפי שראינו בהרצאה)

3. **ניתוח זמן ריצה**

(א) בניתוח של אלגוריתם עימוד הרצפים הגלובלי של Needleman-Wunsch, הגדלת האורך של כל אחד מהרצפים הביאה לעלייה ליניארית בזמן הריצה ובזמן הזיכרון. כאשר שני הרצפים הוגדלו בו זמנית, נצפתה השפעה בולטת יותר, לא ליניארית, הן על זמן הריצה והן על השימוש בזיכרון.

הניתוח הזה מתיישר עם העובדה שסיבוכיות זמן הריצה של האלגוריתם וסיבוכיות המקום ריבועי באורך הרצפים.

גרף 1: אורך מילה נגד זיכרון (גודל וצבע הבועה מייצגים את הזיכרון ב־MB)



גרף 2: אורך מילה נגד זמן ריצה (גודל וצבע הבועה מייצגים את זמן הריצה בשניות)