כפיר ענבל

אור ש. נעים

ביולוגיה חישובית - תרגיל 4

בעיית הסוכן הנוסע מוגדרת כך:

סוכן נוסע רוצה לבקר ב- **N** ערים.

הוא מתחיל בעיר מספר **1**, מסיים בעיר מספר **n** ומבקר בכל עיר בדיוק פעם אחת.

ניתן להניח שנתונה טבלה המכילה את מרחק ההגעה מכל עיר לכל עיר.

המטרה היא למצוא את המסלול הקצר ביותר שבה הסוכן יכול לבצע את משימתו.

הסבירו כיצד ניתן לפתור את בעיית הסוכן הנוסע תוך שימוש בחישוב מולקולרי .

שימו לב לפרטים הבאים בתשובה:

- מה הרעיון הכללי של הפתרון
- מהם רכיבי ה**DNA** בהם אתם משתמשים? מה מייצגים חלקיהם?
- מהן שיטות המעבדה בהם תצטרכו להשתמש לאורך החישוב?
- היכן, לפי הבנתכם, עלולה להיווצר טעות בחישוב?

על הפתרון המוצע להיות מקורי וברור ותמציתי (לא יותר משלושה עמודים) מוגש כקובץ **WORD** או **PDF**, ניתן ורצוי לשלב איורים להבהרת השיטה.

תאריך הגשה: **30.6** אך למי שירצה ניתן להגיש עד תאריך המבחן ב **6.7**
ניתן להגיש בזוגות אבל על כל זוג לעשות עבודה עצמאית

ב ה צ ל ח ה !

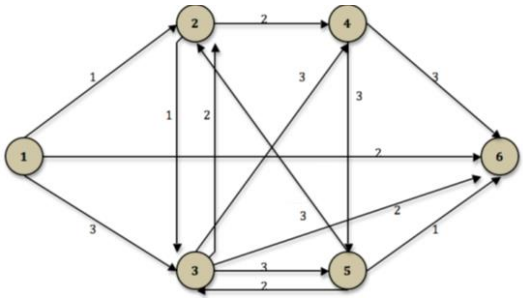
פתרון:

הבעיה שלפנינו שקולה לבעיית מסלול המילטון קצר ביותר.

הדרך הנאיבית לפתור את הבעיה היא כדלקמן; נבנה גרף קליק (בהנחה שקיים מסלול **ישיר** (כזה שלא עובר באף עיר אחרת) מכל עיר, לכל עיר אחרת. אם לא קיים, נבנה גרף שאינו קליק) ממושקל, שהקודקודים בו הן הערים בהן הסוכן מעוניין לבקר ומשקל כל קשת u, v, הנו מרחק ההגעה מעיר **u** אל עיר **v**מודגש כי מכיוון ועסקינן במרחקים, המשקלים **אי שליליים**. ולמצוא את כל מסלולי ההמילטון האפשריים בגרף, לסכום את המרחקים מכל עיר לעיר במסלול שכזה, ולהחזיר את המסלול שסכום המרחקים בו הוא הקצר ביותר. בהנתן שהמסלול חייב להתחיל בעיר מספר **1** ולהסתיים בעיר מספר **n** (זו תהיה ההנחה שלנו) זמן הריצה של האלגוריתם הנו: $O(n-2)! \cdot O(n)$. אם עיר מספר **1** ועיר מספר **n** יכולה להיות כל עיר, אז סיבוכיות הזמן של האלגוריתם היא: $O(n!) \cdot O(n)$. מדובר כאמור בבעיה **NP** קשה, אך ננסה לפתור אותה בזמן מהיר, באמצעות **חישוב מולקולרי**.

מקור בו השתמשנו ללמידת הפתרון המבוסס חישוב מולקולרי:

"Solving Shortest Hamiltonian Path Problem Using DNA Computing" (2012),
King Saud University Saudi Arabia, Hala Mohammed Alshamlan et. Al.



תרשים 1: גרף ממושקל ומכוון שאינו קליק.

באדיבות **Hala Mohammed Alshamlan et. Al.**

פתרון מבוסס חישוב מולקולרי:

בקצרה:

1. קידוד כל הקודקודים, הקשתות והמשקלים שתוארו **בפתרון הנאיבי**
2. יצירת כל המסלולים האקראיים מקודקוד המקור (נשתמש ב**DNA Ligase**).
3. **נשמור רק** את המסלולים שמתחילים בקודקוד המייצג את עיר **1** (נקרא לקודקוד זה, **s** מעתה והלאה), ומסתיימים בקודקוד המייצג את עיר **n** (נקרא לקודקוד זה, **t**, מעתה והלאה).
4. מהמסלולים הנותרים, **נשמור** רק את המסלולים שמבקרים ב**n** קודקודים בדיוק.
5. מהמסלולים הנותרים **נשמור** רק את המסלולים שביקרו בכל קודקוד **פעם אחת** (באופן סימטרי, **נסנן** את המסלולים בהם ביקרנו **יותר** מפעם **אחת** בקודקוד כלשהו).
6. אם לאחר הסינונים המתוארים בסעיפים 3 עד 5, נותרנו עם מסלול אחד לפחות, **נשמור רק** את המסלול הקל ביותר (בהכרח נותר, בהנתן שקיים מסלול ישיר בין כל שני קודקודים (גרף קליק)).
7. החזר את המסלול (אם קיים) או $\emptyset$ אחרת.

שלב ראשון: קידוד באמצעות DNA:

נסביר כיצד לקדד את הקודקודים, הקשתות והמשקלים בגרף:

קידוד קודקודים: נשתמש בגדיל **DNA** אחד. כאמור אורכו של כל גדיל הוא **20** נוקלואוטידים (או **mer 20**). נקדד כל קודקוד בגרף באופן יחודי. כך למשל, נקודד את קודקוד **1**, **2** ו**6** כך:

$$VE(v_1) = ATGCAAGGTC\ TGACGGTTCA$$
$$VE(v_2) = GATCGGTAAC\ GACTGGTTAC$$
$$VE(v_6) = TACGGTTACA\ TCGATTGAAA$$

קידוד משקלים: לפני שנסביר כיצד אנו מקודדים קודקודים, נציג כמה הגדרות:

הגדרה **1 אוסף משקלים A**: בהנתן גרף ממושקל **G**, **אוסף המשקלים A** שלו מכיל את כל המשקלים הקיימים בגרף. כך למשל, בגרף בתרשים **1**, משקל חלק מהקשתות הוא **1** חלק מהן, משקלן **2** והנותרים, משקלן הוא **3**. לפיכך, $A = \{1,2,3\}$.

הגדרה **2 פונקציית מינימום MIN**: פונקציה זו, מחזירה את **ערך המינימום** של **אוסף המשקלים A**. כך למשל, בדוגמא שבהגדרה **1**, מתקיים: $MIN(A) = 1$.

הגדרה 3 **פונקציית מיפוי גדיל משלים** H: בהשראת עקרון המשלים של וטסון-קריק, פונקציה זו תחזיר את סדרת הנוקלאוטידים המשלימה לקלט שלה. כך למשל: $H(ATTGCA) = TAACGT$.

הגדרה 4 **יחס חלוקה**: יחס זה מייצג את הפער בין המשקלים והממוצע שלהם בגרף נתון. כל הערכים של הבדל הממוצעים של אוסף המשקלים ***A*** לוקח חלק בגרף. הנוסחא שלה היא: $\frac{\Sigma_{j=1}^{k}\Sigma_{i=j+1}^{k}(W_i-W_j)}{C_n^2+n}$ כאשר W_i וW_j הם המשקלים של קשתות ***i*** ושל ***j*** והם כאמור שייכים שאוסף המשקלים ***A***. ***n*** היא הגודל של האוסף ***A***. כך למשל, **יחס החלוקה** של הגרף באיור 1 ניתן לחישוב כדלקמן:

$$\frac{(2-1)+(3-1)+(3-2)}{C_3^2+3} = \frac{(2-1)+(3-1)+(3-2)}{1+3} = \frac{1+2+1}{4} = \frac{4}{4} = 1$$

עתה, ניתן לתאר את שיטת הקידוד של המשקלים. המשקלים מקודדים כ***DNA*** דו גדילי בעל גודל משתנה. כאמור, במבנים ביולוגיים, זוגות ***A*** ו***T*** יוצרים שני קשרי מימן, וזוגות ***G*** ו***C*** יוצרים שלושה קשרי מימן. בעזרת הבחנה זו, ניתן להיעזר בזוג ***G*** ו***C*** בשביל לבטא 1 ובזוג ***A*** ו***T*** בשביל לייצג 0. באמצעות הגדרה זו, ניתן לתרגם את המשקלים למחרוזות בינאריות. כאשר, הקידוד הבינארי של 1 הנו "01" ושל 0 הוא "00". אנו מתייחסים לאורך של המחרוזת הבינארית כמשתנה. עתה, השאלה היא: האם ניתן להשתמש באורכו של גדיל ***DNA*** בקידוד זה, על מנת להבחין באורכים השונים של מסלול בגרף ממושקל? ננסה להשתמש בדוגמא בשביל להדגים. נניח שניתן להגיע מקודקוד אחד אל קודקוד אחר, באמצעות קשת שמשקלה 5 (*). ולחילופין, ניתן להגיע מאותו קודקוד אל אותו קודקוד אחר, באמצעות מסע על שתי קשתות שמשקלן 2 כל אחת (**), כמובן ש $5 > 2 + 2 = 4$. לעומת זאת, הקידוד הבינארי של 2 הנו: "10" ולכן על מנת לבטא את המסלול המתואר ב(**) נעזר במחרוזת באורך 4:

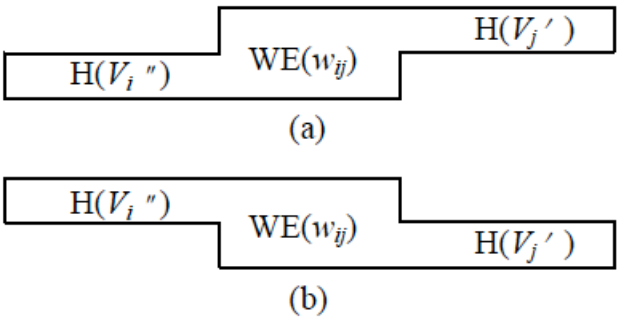
"10 10" (כאן אנו לא מתייחסים לאורך הקידוד של הקודקוד). לעומת זאת, הקידוד הבינארי של 5 הנו "101" ולכן, המסלול המתואר ב(*) מקודד כ: "101".

מתעוררת כאן בעיה, שכן 4 קטן מ5, אך מיוצג בינארית על ידי קידוד **שאינו** קטן מהקידוד של 5: $\underbrace{101}_{Rep.5} \nleq \underbrace{1010}_{Rep.2,2}$. לפיכך, לא ניתן להשתמש באורך הקידוד לפתרון הבעיה של מסלול המילטוני קצר ביותר. על מנת לפתור את הבעיה, הצגנו את הגדרה 4, יחס חלוקה. בעבור הגרף המוצג באיור 1, אוסף המשקלים $A = \{1,2,3\}$ ו $MIN(A) = 1$, בהתאם להגדרה לעיל, יחס החלוקה של ***A*** הוא כאמור 1. ניתן לקודד בהתאם לערך מינימום שהוגדר על ידי ***MIN*** והחלוקה. אז בדוגמא שלנו, הקידוד של 1 הוא "01", ולכן, מולקולות ה***DNA*** יכולות להתבטא באופן הבא:

$$\begin{array}{c} CT \\ WE(1) = GA \\ 2 = 1 + 1 \rightarrow The\ binary\ encoding\ is\ \mathbf{0101}, \\ GAGA \\ WE(2) = CTCT \\ 3 = 1 + 1 + 1. The\ binary\ encoding\ is: 010101, \\ CTGAGA \\ WE(3) = GACTCT \end{array}$$

קידוד קשתות:

קידוד קשתות תלוי בקידוד הקודקודים וכן בקידוד המשקלים שהצגנו קודם לכן. אנו מפעילים צורה מולקולרית לא שלמה בשיטת קידוד הקשתות המוצעת. לייצוג הקשת, ראשית אנו שמים שני חד גדילים (***single strand ss***), המייצגים את הקודקודים v_i, v_j למבחנת הבדיקה. כל ***ss*** מחולק לשני חצאים. כך למשל, v_i הופך ל v_i' המייצג את החצי הראשון וv_i'' המייצג את החצי השני. לאחר מכן, את הדו גדילי הלא מוגמר (***double strand ds***) באותה מבחנה, אשר מייצג את קידוד המשקל של הקשת בין v_i לv_j שנקרא: w_{ij}. תרשים 2 מייצג את המבנה של הקשת.



תרשים 2, מבנה קידוד הקשת, באדיבות Hala Mohammed Alshamlan et. Al.

בתרשים, המבנה של מולקולות לא שלמות, מכיל שלושה חלקים. החלק הראשון $H(v_i'')$ (ראה הגדרה 3) הוא הסדרה המשלימה של v_i'', שהוא החצי השני של החד גדילי המייצג את קודקוד v_i כאמור. החלק הנוסף הוא $WE(w_{ij})$ שהוא דו גדילי המייצג את משקל הקשת בין v_i, v_j. והחלק השלישי הוא $H(v_j')$ שהוא הסדרה המשלימה של v_j'- החצי הראשון של החד גדילי המייצג את קודקוד v_j. בנוסף, כל קשת בגרף המוצג בתרשים 1, מקודדת לשני דו גדילים לא מושלמים כמתואר בתרשים 2. למשל, הקידוד של הקשת מקודקוד 1 ל2 ששוקלת 1 מורכב משני דו גדיליים לא מושלמים כדלקמן:

$$EE(v_1 - v_2) = H(v_1'') + WE(1) + H(v_2') =$$

$$\begin{array}{l} CTAACTAGCCATTG \\ ACTGCCAAGTGA \end{array}$$

וכן:

$$\begin{array}{l} ACTGCCAAGTCT \\ GAAACTAGCCATTG \end{array}$$

על ידי שימוש בדרך הנ"ל, ניתן לקודד **כל קודקוד** בגרף המתואר בתרשים 1. במיוחד, ש$v_i = v_i' + v_i''$. וכן, ניתן לבטאות אותו על ידי 20 זוגות בסיסים של קטעי אוליגונוקלואוטידים. כאשר הגודל של v_i', v_i'' הוא 10 זוגות בסיסים כל אחד. וכעת ניתן לחשב את אוסף המשקלים ***A*** של הגרף ואת יחס החלוקה שלו. לאחר מכן, ניתן לקודד את המשקלים בשימוש בדרך שפירטנו קודם לכן, ואז לקודד את מבני המולקולות הלא מושלמות של כל קשת.

עתה אנו מוכנים לפרט על כל אחד משלבי האלגוריתם:

1. קידוד הקודקודים, הקשתות והמשקלים: בשלב זה דנו עד כה.
2. יצירת המסלולים האקראיים: בשלב זה, נערבב מבחנה ראשונה המכילה את קידודי הקודקודים, ומבחנה שנייה (***T***) המכילה את קידודי הקשתות וקידודי המשקלים גם כן. ב***T*** יווצרו תהליכי ליגאז רבים.
3. שמירת המסלולים המתחילים בעיר 1 ומסתיימים בעיר ***n***: נעשה זאת באמצעות פעולת ***PCR*** ושימוש בפרימרים של הקודקודים המייצגים את העיר הראשונה והעיר האחרונה. נניח והעיר הראשונה מיוצגת כv_1 והעיר האחרונה כv_6, אנו נגדיל את התגובה בעבור קודקודים אלו, ומספר הגדילים המתחילים בv_1 ומסתיימים בv_6 יגדלו באופן חד, אך מספר הגדילים שאינם עונים לתנאי זה לא ישתנו.
4. סינון כל המסלולים שאינם מבקרים ב***n*** קודקודים בדיוק: התוצאה של שלב זה הסתברותית, שכן בבעיה זו הגדרנו כבר את קידוד הקודקודים על ידי 20 זוגות בסיסים של קטעי אוליגונוקלואוטידים. ולא ניתן לקבוע את המשקל של כל קשת במסלול. נניח כי ***N*** זה מספר הקודקודים בגרף, ***MDR*** הוא ערך המינימום שנמצא על ידי ***MIN***, ויחס החלוקה ***AVG*** הוא ממוצע המשקלים באוסף המשקלים ***A***. בשלב זה משתמשים בג'ל אלקטרופורסיס, והאורך ***L*** של גדילים מתקבלים הוא כדלקמן:

$$\big((N+1)\cdot 20\big) + \big((N+1)\cdot |AVG|\big) < L < (N\cdot 20) + (N\cdot |MDR|)$$

5. סינון כל המסלולים שלא מבקרים בכל הקודקודים: זה נעשה על ידי הפרדה של רצפים באמצעות שימוש במשלים של כל קודקוד (ראה הגדרה 3) הרבה פעמים, עד לכדי מצב שכל קודקוד מופיע **בדיוק** בעם אחת במבחנה.
6. שמירת המסלול הקצר ביותר: גדילי ה***DNA*** הנוצרים בשלב 5, ניתנים לפיצול על ידי שימוש בג'ל אלקטרופורסיס. המולקולות מחולקות בהתאם למשקלן אשר הוא כמעט תואם לאורכן. מכיוון שכל קשת $e(i,j)$ היא בעלת שלושה חלקים כפי שהסברנו, $\big(H(v_i'') + WE(w_{ij}) + H(v_j')\big)$ ואורכן של $H(v_i''), H(v_j')$ זהה, אנו מסיקים שההבדל בין הקשתות נובע מהחלק האמצעי: $WE(w_{ij})$ העוסק במשקל כאמור. לכן, ככל שהמסלול ארוך יותר, כך האורך של הגדיל.
7. קבלת התשובה: בשלב זה, עלינו להשתמש בפעולת ***PCR*** ובפרימרים לקביעת הסדר של כל קודקוד בבעיית מסלול המילטון קצר ביותר כדלקמן: נבצע אמפליפיקציה באמצעות סדרה של ***PCR*** . נשתמש בפרימרים בעבור קודקוד ההתחלה ובעבור הקודקוד הסיום (האיבר ה***n*** במסלול). אז על מנת למצוא היכן הקודקוד ***x*** נמצא במסלול, נבצע ***PCR*** על הפרימרים מקודקוד ההתחלה עד לקודקוד ***x***. ועל ידי שימוש באלגוריתם הבא: נניח ש***L*** הוא האורך מהתחלת המסלול ועד לקודקוד ***x***.
   1. נגדיר $\bar{L}$ להיות $\left\lceil \frac{L}{20} \right\rceil$.
   2. $E = (\bar{L} - 1)\cdot |AVG|$. כאשר ***E*** הוא תוחלת המשקל ***AVG*** הוא **הממוצע** של אוסף המשקלים ***A*** של הגרף (ראה הגדרה ראשונה).
   3. נכה את תוחלת המשקל מהאורך ***L***: $L = L - E$.
   4. הסדר של קודקוד ***x*** הוא: $x = \left\lceil \frac{L}{20} \right\rceil$.

עתה ננתח את התוצאות על דוגמא ישימה:

נניח ויש לנו גרף ***G*** שמכילה שני מסלולי המילטון, הראשון, עלותו 5, השני עלותו 3. משקל המסלול הראשון מיוצג כך כאמור:

$$GTGTGTGTGT$$
$$CACACACACA$$

כלומר, אנו נדרשים ל 10 נוקלואוטידים. מצד שני, משקל המסלול השני מיוצג כך:

$$GTGTGT$$
$$CACACA$$

ואנו נדרשים רק ל6 נוקלואוטידים כאמור. ולכן, כאשר נשים אותם בג'ל אלקטרופורסיס, נבחין שהמסלול השני עובר מהר יותר. טכניקה זו מבוססת על העובדה שמולקולות ***DNA*** טעונות שלילית. לכן אם נשים אותם בשדה חשמלי, הם יזוזו לעבר האלקטרודה החיובית **במהירויות שונות**. וככל שהמולקולה ארוכה יותר היא תנוע בקצב איטי יותר, כפי שראינו בתרגול. וכך הבעיה נפתרה.

סיכום: לפתרון הבעיה של הסוכן הנוסע, ייצגנו את הערים שהסוכן מעוניין לבקר בהם ואת המרחקים בניהם, באמצעות גרף ממושקל. באמצעות ייצוג זה, יכולנו לפתור את הבעיה הנתונה על ידי פתרון בעיית מציאת מסלול המילטון קצר ביותר. בשיטה שתיארנו, השתמשנו בחלקים שונים של ה***DNA*** לקידוד הקודקודים, הקשתות והמשקלים, פירטנו בהרחבה במהלך הפתרון באילו חלקים השתמשנו ובעבור אילו צרכים. בנוסף, השתמשנו במספר שיטות מעבדה כאלטרנטיבה לחישוב פתרון לבעיה על גבי מחשב. פירטנו את שיטות המעבדה השונות, גם כן תוך כדיי הפתרון. לעניות דעתנו, הטעות בחישוב עשוייה לנבוע בין היתר משלב 4 באלגוריתם שתיארנו. כמו כן, הצלחת שלב 2 גם היא תלוייה בכך שיווצרו גדילים במסגרת תהליך הליגאז בעבור כל מסלול אפשרי.

...