בס"ד

ביולוגיה חישובית- תרגיל 4

חישוב מולקולרי

מגישות: שיר בן אהרון, תמר סעד

בעיית מציאת מעגל ארוך ביותר בגרף:

<u>הבעיה:</u> נתון גרף לא מכוון עם E קשתות ללא משקלות ו V-צמתים. המטרה היא למצוא את המעגל הארוך ביותר בגרף, כלומר, המסלול הארוך ביותר המתחיל ומסתיים באותה צומת.

<u>פתרון הבעיה:</u>
בפתרון הבעיה השתמשנו בעקרון דומה לפתרון בעיית המסלול ההמילטוני של אדלמן, וכן בכלים שלמדנו בהרצאה.

<u>סעיף א':</u> בהנחה שיש טכניקה מעבדתית שמאפשרת להפריד בין DNA מעגלי לבין DNA פתוח שאינו נסגר על עצמו.

<u>רעיון כללי:</u>
האלגוריתם הגנרי-
נניח שקיים מעגל כלשהו בגרף G הנתון.

1. באופן רנדומלי צור את כל המעגלים האפשריים בגרף.
2. שמור רק את המעגלים שמייצרים מעגל חוקי כלשהו עד אורך V.
3. שמור רק את המעגלים בהם אין חזרה של קודקודים.
4. מצא את המעגל הארוך ביותר.

ייצוג הבעיה בכלים גנטיים:
בדומה לפתרון של אדלמן, בחרנו לייצג כל קודקוד בגרף באמצעות רצף ייחודי של 20 נוקלאוטידים, כך שבכל קודקוד קיים רצף זיהוי של אנזים רסטריקציה שונה. כלומר, רצינו **לבחור V אנזימי רסטריקציה** שונים כך שכל אנזים רסטריקציה מזהה וחותך קודקוד יחיד.

גם את הקשתות בחרנו לייצג כמו בפתרון של אדלמן, כך שכל קשת מורכבת מה-10 נוקלאוטידים הראשונים של קודקוד אחד ומ-10 נוקלאוטידים האחרונים של קודקוד שני.
כיוון שהגרף אינו מכוון, כל קשת תיוצג באמצעות שתי האפשרויות: קשת ראשונה של 10 הנוקלאוטידים הראשונים של קודקוד א' ו-10 הנוקלאוטידים האחרונים של קודקוד ב', וקשת שניה של 10 הנוקלאוטידים הראשונים של קודקוד ב' ו-10 הנוקלאוטידים האחרונים של קודקוד א'.

בנוסף, נרצה ליצור את הרצף המשלים לכל קודקוד, כלומר הרצף ההופכי שיכול להיקשר לקודקוד.

יצירת מעגלים בגרף:
נערבב את כל הרצפים שהזכרנו לעיל **במבחנה** (ייצוג הקשתות + משלימים של הקודקודים) לתערובת אחת יחד עם האנזים **ליגאז.**

מה שיקרה כעת הוא שייווצרו רצפים דו-גדיליים ארוכים שמסמלים מסלול כלשהו בגרף, כפי שמפורט בפתרון של אדלמן.
הבדל מהותי הוא שבפתרון של אדלמן היה רצף ייחודי שסימן התחלה או סיום, וכאן לא קיים רצף כזה. כלומר, הדרך היחידה לעצור את התארכות הגדיל היא **באמצעות סגירה שלו לדנ"א מעגלי**. סגירה כזאת יכולה להתרחש כיוון שנותרים 10 נוקלאוטידים חופשיים בכל צד של הגדיל, ואם הרצף של תחילת הגדיל מגיע מאותו קודקוד של סוף הרצף, אזי קיים לרצף הזה גדיל משלים (הקודקודים המשלימים שיצרנו), והוא יכול לעבור היברידיזציה לרצף ולסגור אותו לדנ"א מעגלי. כמובן שברגע שנוצר מעגל הגדיל נחסם לגדילה.

בצורה כזו אנו יוצרים מספר גדול של גדילים מעגליים באורכים ובהרכבים שונים.

על מנת לשפר את התהליך ניתן להעלות את ריכז המלחים בתמיסה, מה שממסך את המטען השלילי שעל הדנ"א ומאפשר לו להתקרב לעצמו בקלות רבה יותר, וכך מעגלים יכולים להיווצר בקלות רבה יותר.

הפרדה של דנ"א מעגלי מלינארי:
בשלב זה התמיסה מכילה דנ"א מעגלי ודנ"א לינארי, ונרצה להמשיך עם דנ"א מעגלי בלבד.

כיוון שאנו מניחים בסעיף זה שקיימת טכניקה מעבדתית שמאפשרת לנו להפריד את הדנ"א המעגלי מהלינארי- נשתמש בה ונישאר עם הדנ"א המעגלי.

בידוד מעגלים לפי גודל:
בשלב הראשון נריץ את כל הדנ"א המעגלי **בג'ל אלקטרופורזה**, כך שיעבור מיון לפי גודל. נריץ במקביל גם מרקר שייתן לנו אינדיקציה של גודל המקטעים שייווצרו.
נשים לב כי **אורך המעגל הארוך ביותר שיכול להיות רלוונטי לפתרון- חסום ב-V,** כיוון שלא ייתכן מעגל ארוך יותר ללא חזרה של קודקודים. כלומר, תיאורטית יש לנו עד V אורכים של מעגלים שיכולים להיות מעגלים פוטנציאליים. מהשוואה למרקר שהרצנו, וכיוון שאנו יודעים את מספר הקודקודים בגרף ומה המשקל המולקולרי של כל קודקוד, נוכל לקחת את המקטעים הרלוונטיים אלינו מהג'ל ולהמשיך איתם לשלבים הבאים.

ניקח כל מקטע בנפרד ונעבור איתו לשלב הבא. כלומר, בשלבים הבאים כל המקטעים יעברו את אותה הפרוצדורה אבל בצורה נפרדת ובמקביל, כך שיהיו קיימים לכל היותר V תהליכים שונים, כשכל תהליך מכיל מעגלים בגודל זהה. (כלומר, את השלבים הבאים נבצע עבור כל המעגלים באורך V בקבוצת ניסוי אחת, עבור כל המעגלים באורך V-1 בקבוצת ניסוי נוספת וכן הלאה).

מציאת מעגלים ללא חזרה של קודקודים:
כעת נרצה לבודד מעגלים בהם כל קודקוד מופיע פעם אחת בלבד. לצורך כך נעבור מספר שלבים, וכל שלב יוריד לנו את כמות הדנ"א שתעבור לשלב הבא, ולכן נשתמש **ב-PCR** להגברת החומר הגנטי הקיים.

ניקח אנזים רסטריקציה ספציפי לקודקוד מסוים (נקרא לו קודקוד i) ונוסיף אותו לדנ"א כך שיחתוך את המעגל בנקודת קודקוד זה, ואז נריץ את התוצרים בג'ל.

מקטעי דנ"א שלא מכילים את i לא יעברו חיתוך כלל ויישארו מעגליים, ומקטעים שמכילים את i פעם אחת יעברו חיתוך אחד ויהפכו ממקטע מעגלי ללינארי. מקטעים שמכילים את i יותר מפעם אחת, לעומת זאת, יתפרקו למספר קטעים קטנים יותר. כשנריץ את כל המקטעים הללו בג'ל, נוכל לראות בבירור את המקטעים שנותרו באותו הגודל, לעומת המקטעים שעברו חיתוך ואיבדו מגודלם וממשקלם.

ניקח את המקטעים שנותרו באותו הגודל לשלב הבא- מעגליים/לינאריים שלמים, ונחבר אותם בחזרה למעגל עם האנזים ליגאז.

כעת נחזור על השלבים לעיל שוב ושוב עבור כל אחד מהקודקודים.

הדנ"א שנישאר איתו בסוף התהליך יהיה דנ"א מעגלי שלא מכיל את אותו הקודקוד יותר מפעם אחת.

<u>זמן ריצה של התהליך:</u>

נשים לב כי שלב זה מצריך שימוש באנזים רסטריקציה עבור כל קודקוד, כלומר הוא O(V), וכן שהוא נעשה עבור V קבוצות שונות. עם זאת, זמן הריצה של התהליך אינו $O(V^2)$ כיוון שהקבוצות השונות אינן תלויות זו בזו וניתן לעבוד עליהן בנפרד, ועל כן זמן הריצה פה נותר O(V).

מציאת מעגל ארוך ביותר:
כיוון שאנו יודעים את אורך הדנ"א שבכל קבוצת ניסוי, נבחר לקחת את הדנ"א מקבוצת הניסוי שמכילה את המקטעים הארוכים ביותר שהתקבלו בה תוצאות בשלב הקודם.

את הדנ"א הזה נשלח **לריצוף** ונבדוק התאמה לקודקודים שתכננו, וכך נגלה את המסלול שמרכיב את המעגל הארוך ביותר.

בס"ד

<u>סעיף ב':</u> בהנחה שאין טכניקה מעבדתית שמאפשרת להפריד בין DNA מעגלי לבין DNA פתוח שאינו נסגר על עצמו.

<u>רעיון כללי:</u>

האלגוריתם הגנרי-

נניח שקיים מעגל כלשהו בגרף G הנתון.

עבור קודקוד i מ-i-V, חזור על הסעיפים הבאים:

1. באופן רנדומלי צור את כל המסלולים האפשריים שמתחילים ונגמרים באותו קודקוד בגרף.
2. שמור רק את המסלולים הלינאריים שמייצגים מעגל חוקי כלשהו עד אורך V.
3. שמור רק את המסלולים בהם אין חזרה של קודקודים.
4. מצא את המעגל הארוך ביותר.

לאחר שהאלגוריתם מבוצע על כל הקודקודים:

5. מצא את המעגל הארוך ביותר מבין המעגלים הארוכים ביותר.

ייצוג הבעיה בכלים גנטיים:

בדומה לפתרון שהצגנו לעיל, גם בסעיף זה נייצג כל קודקוד ע"י 20 נוקליאוטידים ייחודיים עם אתרי רסטריקציה לאנזימים ייעודיים. גם את הקשתות נייצג באופן דומה.

השוני בין הפתרון לעיל לפתרון המוצג כאן- לא ננסה ליצור מולקולות מעגליות אלא רק מולקולות לינאריות דו-גדיליות בדומה לתוצר בפתרון של אדלמן.

השוני בין הפתרון המוצג לפתרון של אדלמן הוא שבמקרה שלנו אנו כן רוצים שהמסלול יתחיל ויסתיים באותו קודקוד ולכן נבחר באופן סדרתי בכל פעם קודקוד כלשהו מבין קודקודי הגרף, את הקשתות היוצאות/נכנסות לאותו קודקוד נייצג כמו בפתרון של אדלמן כארוכות יותר משאר הקשתות ע"י 30 נוקליאוטידים- 10 נוקליאוטידים מהקשת שיוצאת/מגיעה + 20 נוקליאוטידים של הקודקוד שבחרנו כקודקוד ההתחלה/סיום.

שוב, נרצה ליצור את הרצף המשלים לכל קודקוד, כלומר הרצף ההופכי שיכול להיקשר לקודקוד.

יצירת מסלולים בגרף:

נערבב את כל הרצפים שהזכרנו לעיל **במבחנה** (ייצוג הקשתות + משלימים של הקודקודים) לתערובת אחת יחד עם האנזים **ליגאז.**

מה שיקרה כעת הוא שייווצרו רצפים דו-גדיליים ארוכים שמסמלים מסלול כלשהו בגרף, כפי שמפורט בפתרון של אדלמן.

נבצע **חימום** של התוצרים במבחנה כדי להפריד את הגדילים שנוצרו ובעזרת **PCR** נגביר את הביטוי של מולקולות שמתחילות ומסתיימות בקודקוד i שבחרנו.

בידוד מסלולים לפי גודל:

בשלב זה, נריץ את כל הדנ"א שהגברנו **בג'ל אלקטרופורזה**, כך שיעבור מיון לפי גודל. נריץ במקביל גם מרקר שייתן לנו אינדיקציה של גודל המקטעים שייווצרו.

נשים לב כי **אורך המסלול הארוך ביותר (המייצג מעגל ארוך ביותר ללא חזרה של קודקודים) שיכול להיות רלוונטי לפתרון- חסום ב-V+1,** כיוון שלא ייתכן מעגל ארוך יותר ללא חזרה של קודקודים וכן יש לנו קודקוד שחוזר פעמיים- בהתחלה ובסוף. מהשוואה למרקר שהרצנו, וכיוון שאנו יודעים את מספר הקודקודים בגרף ומה המשקל המולקולרי של כל קודקוד, נוכל לקחת את המקטעים הרלוונטיים אלינו מהג'ל ולהמשיך איתם לשלבים הבאים.

ניקח כל מקטע בנפרד ונעבור איתו לשלב הבא.

מציאת מסלולים ללא חזרה של קודקודים:

כעת נרצה לבודד מסלולים בהם כל קודקוד מופיע פעם אחת בלבד- מלבד הקודקוד שבחרנו להיות קודקוד ההתחלה והסיום.

נבחר בכל פעם קודקוד אחד ונוסיף למבחנה את הרצף המשלים לאותו קודקוד, כך שתתבצע היברידיזציה של הקודקוד המשלים למקטעים שמכילים את אותו קודקוד.

נריץ את המקטעים לאחר החיבור בג'ל כך שתתבצע הפרדה לפי גודל. קישור לקודקוד משלים גורמת למקטע להיות כבד יותר, כך שנוכל לראות בבירור אילו מקטעים לא עברו כלל קישור לקודקוד, אילו מקטעים עברו קישור אחד ואילו עברו יותר. ניקח לשלבים הבאים רק את המקטעים שעברו קישור אחד או שלא עברו קישור כלל. יתר המקטעים עברו יותר מקישור אחד ועל כן יש בהם חזרה על הקודקוד הנבדק.

לאחר לקיחה של הדנ"א הרלוונטי, נחמם את המבחנה ונפריד את הדנ"א למקטעים חד גדיליים.

כעת נחזור על השלבים לעיל שוב ושוב עבור כל אחד מהקודקודים.

כמובן שנחריג מהכלל את קודקוד ההתחלה והסיום, בו נאפשר שני קישורים של קודקוד משלים.

הדנ"א שנישאר איתו בסוף התהליך יהיה דנ"א לינארי שמתחיל ומסתיים באותו קודקוד ולא מכיל קודקודים נוספים יותר מפעם אחת.

מציאת מסלול ארוך ביותר:

כיוון שאנו יודעים את אורך הדנ"א שבכל קבוצת ניסוי, נבחר לקחת את הדנ"א מקבוצת הניסוי שמכילה את המקטעים הארוכים ביותר שהתקבלו בה תוצאות בשלב הקודם.

חזרה על שלבי הניסוי:

על הפעולה המתוארת מעלה נחזור עבור כל קודקוד מקודקודי הגרף כך שנקבל עבור כל אחד מהם את המסלול האפשרי הארוך ביותר המתאר מעגל אפשרי ארוך ביותר בגרף.

נשווה בין כל המסלולים האפשריים שקיבלנו וניקח את הארוך ביותר המתקבל. נמצא את הקודקודים שמרכיבים אותו באמצעות ריצוף.

<u>סיכום שיטות בהן השתמשנו:</u>

- סינתזה מלאכותית של גדילי דנ"א
- היברידיזציה של דנ"א
- חיבור באמצעות ליגאז
- חיתוך באנזימי רסטריקציה
- הרצה בג'ל
- הגברת דנ"א עם PCR
- ריצוף דנ"א

<u>טעויות או בעיות אפשריות בפתרון:</u>

- כמות חומרים:
  - ככל שהגרף גדול יותר יש צורך ביותר נוקלאוטידים לייצוג של קודקודים וקשתות. בנוסף, כל קודקוד מצריך הרצות נפרדות בג'ל וכד', מה שמצריך משאבים נוספים.
  - יש צורך באנזים רסטריקציה ייחודי עבור כל קודקוד. עם זאת, קיים מספר מוגבל של אנזימי רסטריקציה כך שגודל הקלט האפשרי מצומצם מאוד.
- חוסר דיוק ביולוגי:
  - היברידיזציה לא נכונה של מקטעי דנ"א- אם שני מקטעים דומים מספיק, הם יכולים להיצמד אחד לשני למרות שהם לא מתאימים בדיוק. אפשר לנסות לצמצם את הטעויות הללו בעזרת אמצעים כמו טמפרטורה גבוהה שמקשה על קישור לא ספציפי, אבל אי אפשר למנוע אותן.
  - שימוש מרובה באנזימי רסטריקציה- בכל שימוש כזה לא כל הגדילים עוברים חיתוך, כך שייתכן שבסוף נישאר עם גדילים שיש בהם חזרות והמידע שנקבל לא ייתן את התוצאה המבוקשת.
  - טעויות שכפול של PCR- יכולות לשנות אתרי הכרה של אנזימי רסטריקציה, כך שייווצרו חיתוכים באזורים שלא היו אמורים להיחתך במקור, או שאזורי חיתוך עברו שינוי וכעת לא מזוהים על ידי אנזימי הרסטריקציה.
  - שלבים מרובים של הדבקה והרצה בג'ל- בכל שלב כזה לא כל הגדילים שעברו חיתוך עוברים הדבקה מחדש, וכן לא כל הגדילים מצליחים לרוץ למקום הנכון בג'ל או עוברים את הניקיון

מהג'ל. כל שלב כזה מוריד את כמות החומר הגנטי שממשיך לשלבים הבאים. בנוסף, ריבוי השלבים מעלה את הסיכוי לטעויות אנוש, batch effects, זיהומים וכדומה.

- אורך המעגל הארוך ביותר:
  - במקרה שהמעגל הארוך ביותר קצר מדיי, יהיה לשני קצוות הגדיל סיכוי נמוך יותר להגיע למצב מקופל בו הם יהיו סמוכים זה לזה ויוכל להיווצר מעגל.
  - במקרה שהמעגל הארוך ביותר ארוך מאוד, ישנה הסתברות נמוכה שהוא יצליח ליצור את הרצף הארוך ביותר, וסביר יותר שהוא ייסגר על עצמו בשלב מוקדם יותר.

לסיכום, כפי שראינו בפתרון של אדלמן, פתרון מולקולרי כזה יכול לפתור בעיית NP-Complete בזמן לינארי, אבל בצורה מוגבלת: פתרון כזה יוצר הרבה בעיות שלא קיימות בפתרון מתמטי-תכנותי, וכן הוא מוגבל בגודל הקלט שהוא יכול לפתור.