אוניברסיטת תל אביב – ביה"ס למדעי המחשב

גנומיקה חישובית 0382.3102.01 -

מועד א' תשע"א 26.1.11

מרצים: פרופ' רון שמיר, פרופ' רודד שרן

מתרגל: גיא קרליבך

משך הבחינה: שלוש שעות ללא אפשרות הארכה

חומר עזר: שני דפים כתובים בכתב יד (משני הצדדים)

יש לענות על שלוש מתוך ארבע השאלות. לכל השאלות ניקוד שווה.

יש לענות בצורה קצרה, ברורה ומדויקת ולנמק כל טיעון. תשובה לא מנומקת לא תזכה בנקודות.

בשאלות בהן נדרש לתאר אלגוריתם יש לציין במפורש את סיבוכיות הזמן.

בשאלות בהן השתמשת באלגוריתם שתואר בשיעור או בתרגיל אין צורך לחזור עליו לפרטיו אלא רק לפרט במדויק את השינויים לעומת מה שנלמד.

שאלה 1

נתונים שני רצפים שאורכו של כל אחד מהם n וסכימת ניקוד כך שפעולת match מקבלת ניקוד 1+ ופעולת mismatch/indel מקבלת ניקוד 1-.

העמדה ***בלעדית*** בין זוג רצפים היא העמדה גלובאלית שמקבלת ציון S כך שהציון של כל העמדה אחרת בין אותו זוג רצפים שונה מ-S. העמדה בלעדית אופטימלית היא העמדה שמבין כל ההעמדות הבלעדיות מקבלת את הניקוד הגבוה ביותר. כלומר ייתכנו העמדות בעלות ציון גבוה יותר שאינן בלעדיות. תאר אלגוריתם יעיל ככל שתוכל למציאת העמדה בלעדית אופטימאלית בין זוג הרצפים.

שאלה 2

בהינתן מחרוזת S (למשל רצף גנומי), ***זוג מקסימלי*** הוא שתי תתי מחרוזות α ו-β שמתחילות באינדקסים שונים של S, מכילות את אותו רצף בדיוק, והתו שמופיע מימין (משמאל) ל- α ב-S שונה מהתו שמופיע מימין (משמאל) ל- β. לדוגמא, עבור הרצף AGTCGCTAGTCCTAGT המחרוזות AGTC שמתחילות באינדקסים 1 ו-8 הן זוג מקסימלי. במקרה כזה α (המחרוזת α הזהה ל-β) נקראת ***חזרה מקסימלית.***

תן חסם עליון הדוק ככל שתוכל על מספר החזרות המקסימליות השונות במחרוזת.

שאלה 3

נתונה דגימה אקראית של ערכי הדמיון בפטרן הביטוי של זוגות גנים. הנח שהגנים מתחלקים לשני צבירים (clusters) כך שערכי הדמיון בתוך צביר מתפלגים נורמלית עם תוחלת w וסטיית תקן s ואילו ערכי הדמיון בין הצבירים מתפלגים נורמלית עם תוחלת b ואותה סטיית תקן.

WWW-9

א. (13 נק') רשום ביטוי מפורט עבור נראות הדגימה הנתונה.

ב. (20 נק') לשם שימוש באלגוריתם EM לחישוב פרמטרים שמביאים למקסימום (מקומי) את פונקצית הנראות: חשב במפורש את הפונקציה Q והסק את נוסחת העדכון של s (נוסחה זו יכולה להיות תלויה ב-b ו-w; אין צורך לפרט את יתר נוסחאות העדכון).

<u>שאלה 4</u>

כל גן באדם (למעט גנים על כרומוזומי המין) מופיע על שני כרומוזומים הומולוגיים בשני עותקים כמעט זהים. ההבדלים בין שני העותקים הן בנקודות בודדות לאורך הגן. בכל נקודת הבדל מופיעים בסיסים (אללים) שונים בשני העותקים. בריצוף של הגן מתקבלים מקטעים קצרים (reads) משני העותקים ללא ידיעה מאיזה עותק נוצר כל מקטע. מטרתנו לזהות את הרצף המדוייק של כל אחד משני העותקים של הגן ע"י זיהוי האלל המופיע בו בכל נקודת הבדל (רצף אללים כזה נקרא הפלוטיפ). המפתח לפתרון הבעיה הוא בכך שאללים על אותו מקטע מקורם בהכרח באותו עותק.

נניח שביצענו ריצוף עמוק של הגן באדם מסויים, המקטעים שהתקבלו בריצוף מופו לגנום וזוהו נקודות ההבדל, שסומנו 1,2,...n בסדר זה לאורכו של הגן. בה"כ נסמן את שני האללים בכל נקודת הבדל ב- 0 ו-1. שים לב ששני ההפלוטיפים הם בייצוג זה סדרות בינאריות משלימות זו לזו. מספר המקטעים גדול, וכל מקטע מכיל שתי נקודות הבדל בדיוק.הנתונים סוכמו בטבלה N בה N(i,s,t) הוא מספר המקטעים עבורם בנקודות ההבדל i,i+1 מופיעים האללים s,t בהתאמה ($s,t,\in\{0,1\}$).

א. (20 נק') תאר אלגוריתם יעיל לפתרון הבעיה בהנחה שאין שגיאות ריצוף.

ב. (13 נק') שגיאת ריצוף במקטע מסויים גורמת לכך שהאלל הלא נכון מופיע באחת משתי נקודות ההבדל שבו. תאר אלגוריתם יעיל לפתרון הבעיה המביא למינימום את מספר שגיאות הריצוף הכולל .

בהצלחה!