**יומן הפרויקט**

ביומן זה אעדכן את התקדמות הפרויקט פעם בשבוע/יים, אדווח על התקדמות בעבודה וכן את סיכומי הפגישות עם מנחה הפרויקט, ד"ר יהודה חסין.

29/10

בפגישה הראשונה לשנה עם ד"ר חסין דיברנו על הפרויקט, מה המוטיבציה ומה המטרה הסופית שלנו.

קיבלתי מד"ר חסין מאמר על אלגוריתם בשם Burrows-Wheeler Aligner אותו עלי לקרוא, להבין ולממש את הקוד הבסיסי שלו על מנת שאוכל להתחיל בעבודה עם ידע בסיסי על האלגוריתם לפני שאתחיל לעבוד עם הקוד הקיים ואנסה למקבל אותו.

5/11

סיכום השבוע הראשון לעבודה הוא שהעבודה מתקדמת לאט מאד, המאמר שקיבלתי מד"ר חסין על אלגוריתם BWA הוא באנגלית וכן ברמה גבוהה של תכנות ולכן לוקח לי זמן להבין על מה מדובר וכן לתכנת את הדברים.

התחלתי בתכנות של הגרסה עם הכלל הרקורסיבי תוך ידיעה שזה יהיה לא שימושי מכיוון שאפילו האלגוריתם הבסיסי לא משתמש בחישובים הבזבזניים הללו. החשיבות בעבודה היתה להבין את מה שעובד מאחורי הקלעים על מנת שאוכל להבין את החלקים היותר מתקדמים.

בפגישה עם ד"ר חסין הצגתי לו את העבודה וכן את הטעות בקוד שמצאתי, הצגתי לו את הפתרון ההגיוני עבורי (טעות בתנאי הקיצון של החישובים שנמצאים במאמר) ויחד עברנו על שתי דוגמאות והוא אישר את התיקון שהצעתי. לשבוע הבא מוטל עלי להתקדם עם הקריאה וכן לתכנת את אלגוריתם BWA עצמו.

12/11

כתבתי את קוד המימוש של אלגוריתם BWA, מהשלב המקדים של חישוב Burrows-Wheeler Transform עבור מחרוזת התייחסות (מחרוזת ארוכה בה נחפש המון מחרוזות קטנות) ועד הקוד הממשי של חישוב כלל ה- Inexact matching עבור מחרוזת קטנה (כלומר מציאת כל ההופעות של המחרוזת הקטנה במחרוזת הגדולה כולל התחשבות במרווח טעות של תווים שיתבטאו במחיקה, הוספה ושינוי של תווים).

הבנתי את עיקרי האלגוריתם אך הקוד אינו עובד כמצופה ואינני מקבל תוצאות נכונות כדי שאני אמור לקבל. את הבעיות שהיו לי הצגתי לד"ר חסין בפגישה השבועית והוא הראה לי היכן טעיתי בפירוש אחת השורות במאמר (חישוב שאני ביצעתי הוא סיבוב המחרוזת המתקבלת מ- BWT בעוד שהמאמר מדבר על חישוב BWT עבור סיבוב מחרוזת ההתייחסות). בנוסף בפגישה קבענו להיפגש בעוד שבוע על מנת לראות את התקדמות העבודה וכן להחליט על ההתקדמות אליה נשאף הלאה.

19/11

בשבוע האחרון פתרתי תקלות שנתקלתי בהן עם מימוש אלגוריתם BWA בשפת C. בנוסף העברתי את הקוד לשפת JAVA לשם נוחות הבדיקה ואכן הקוד עובד וניתן למצוא את כלל המיקומים בהם המחרוזת אותה אנו מחפשים מופיע במחרוזת ההתייחסות (התאמות מדויקות של המחרוזת וכן התאמות בהן יש מספר קטן של חוסר התאמות המתבטאות בהשמטה או בהוספה של תווים).

בפגישה עם ד"ר חסין הראתי לו את התקדמות העבודה בשבוע האחרון, הוא הטיל עלי ללמוד על pair end mapping בתוך next generation sequencing. האלגוריתם העומד בבסיסו של הנושא הינו החלק אותו אצטרך למקבל במהלך הפרויקט. בשבוע הקרוב אצטרך להתרכז גם בכתיבת דוח ההצעה. כתיבת הצעת הפרויקט מעט בעייתית עבורי מכיוון שעד כה עבדנו על השלב המקדים לאלגוריתם אותו אני ממקבל נכנס לפעולה ואינני יודע מה אמלא בדוח במידה ואגיש אותו בתאריך המבוקש. ד"ר חסין הציע שנדחה את הגשת ההצעה עד לאחר שאהיה מודע לחלק שלי בצורה מספיקה.

25/11

לאחר זמן רב בו לא היתה לי גישה למערכת הקורס (בעיות עם ספק אינטרנט hotnet), הבעיה סודרה ויש לי גישה למערכת הקורס. קישורים לדף היומן ולמאגר הקוד של הפרויקט נמצאים עתה במערכת הקורס. המימושים עליהם עבדתי בחודש האחרון נמצאים כעת במאגר הקוד כך שיהיו זמינים לכל דורש.

בפגישה השבועית עם ד"ר חסין הוא החליט שלא נבקש דחייה עבור הגשת ההצעה ושאגיש את ההצעה יחד עם כלל הסטודנטים עד מחר. כמו כן הוחלט להתמקד ב-pair end mapping ולסיים לחלוטין את העבודה על בסיס אלגוריתם ה-BWA. את המחקר על האלגוריתם בבסיס ה-pair end mapping נתחיל ע"י חיפוש מאמרים המדברים על הנושא.

2/12

בשבועיים האחרונים מאז שהוחלט להתמקד יותר בנושא ה-pair end mapping חיפשתי מאמרים באינטרנט על הנושא אך למרות שמספר המאמרים וההסברים על אלגוריתם ה-BWA הוא רב, כמעט ואין מאמרים המסבירים מה עומד בבסיסו של ה-pair end mapping.

בפגישה עם ד"ר חסין הצגתי לו את הבעיה ולאחר שהוא ניסה למצוא מאמרים בצורה בלתי תלויה ולא הצליח, הוא הודיע לי שמכיוון שאינו רוצה שנבזבז זמן יקר ללא תוצאות, נוריד את קוד המקור של האלגוריתם ונבצע הנדסה הפוכה על הקוד כדי להבין ממנו מהם הפעולות היקרות וצורכות הזמן ולמקבל אותן בקוד ללא ידע מוקדם על האלגוריתם שעומד בבסיסן.

17/12

קוד המקור של האלגוריתם זמין להורדה באתר: <http://bio-bwa.sourceforge.net/>

קוד המקור שהורדתי איננו מתועד כמעט בכלל וכן אינו קריא מחמת חוסר התיעוד, הקיצורים השונים וכן השמות חסרי המשמעות שניתנו לפונקציות ולמבני נתונים בקוד.

פניתי לד"ר חסין לעזרה ובפגישה השבועית איתו יצא שברגסה אותה אני הורדתי חסר ה-main. לאחר הורדה של קובץ הכולל את ה-main, אוכל להתחקות אחרי הקריאות ממנו לקבצים אחרים וכך אוכל לאתר את הקוד המבצע את ה-pairing.

בזמן הקרוב אתחקה אחרי הקריאות לפונקציות השונות וכן אחקור את מבנה הנתונים שהתוכנה משתמשת בו, את ההתקדמות שלי אתעד בקובץ word כדי לשמור על פשטות הגישה אליו.

24/12

אחרי הקריאות השונות משיטת ה-main הוביל אותי לפונקציה שמבצעת את ה-pairing. אופן ההתחקות אחרי הקוד מוסבר בדף התיעוד ולכן לא ארחיב פה עוד.

בפגישה השבועית הצגתי לד"ר חסין את ממצאי, עברנו יחד על שתי השיטות והוחלט להתמקד קודם באחת שמבצעת אלגוריתם שנקרא Smith-Waterman ולאחר מכן בשיטה השנייה. בשבוע הקרוב אתמקד בקריאה של הקוד וכן במבני הנתונים בו הוא משתמש.

31/12

השבוע התמקדתי בהבנת הקוד הממומש של אלגוריתם Smith-Waterman וכן בהבנה של עיקרי האלגוריתם כדי לראות כיצד המימוש עובד. הקוד אינו קריא (כרגיל) ומלא בשמות משתנים ללא משמעות, שימוש במאקרו, היעדרות [תיעוד](http://projects.jce.ac.il/moodle/mod/assign/view.php?id=160) וכן קוד רב בתוך מימוש האלגוריתם שאיננו נראה כקשור לאלגוריתם כפי שהוא מופיע בפשטותו וכן גם בהתייחסות להנחת קלט של מחרוזת מכווצת. בפגישה השבועית עם ד"ר חסין הוחלט שמכיוון ומימוש הקוד המופיע בתוך כלל אלגוריתם BWA אינו קריא, אממש בעצמי את האלגוריתם ואנסה לצמצם את צריכת הזיכרון שלו מכיוון שהזיכרון שבכרטיס המסך הוא מוגבל (בגלל העובדה שהזיכרון מוגבל, נרצה שכל מעבד יצרוך כמה שפחות זיכרון עבור פעולתו כי במידה וכל מעבד יקח את מינימום הזיכרון האפשרי אזי מספר המעבדים שנוכל לנצל עם הזיכרון המוגבל יגדל). ד"ר חסין העלה השערה שניתן לצמצם את הזיכרון הדרוש לאלגוריתם ל- 2n, אני אצטרך לבדוק האם הדבר אפשרי ובמידה שכן אממש את הקוד עם יעילות זיכרון שכזו.

7/1

הרחבתי בשבוע האחרון את ידיעתי בנושאים העיקריים של אלגוריתם Smith-Waterma, חיפשתי הסברים על האלגוריתם וכן הצעות מימוש וייעול שונות. בהתחלה לא יכולתי לממש את הצעתו של ד"ר חסין מכיוון שכאשר האלגוריתם מחשב ערך של תא הוא דורש מעבר על כל התאים מעל הצא המדובר וכן כל התאים משמאלו. בגלל סיבה זו לא יכולתי להיפטר מהזיכרון של כל השורות שסיימתי איתן. במהלך החיפושים נתקלתי בהצעה שמצמצמת את זמן הריצה של האלגוריתם מ- n^3 ל- n^2 תוך הקרבה של עוד זיכרון (פי 3 למעשה). לאחר הרהורים רבים אודות עקרונות ייעול האלגוריתם עלה בדעתי שניתן להשתמש בהצעה זו כדי לפתור את הבעיה שנתקלתי בה בניסיון לממש את הצעתו של ד"ר חסין וכך אם אשלב בין שתי ההצעות אוכל להוריד את צריכת הזיכרון ל- 6n. מחמת השלג בירושלים וביטול הלימודים במכללה, לא התקיימה השבוע פגישה שבועית עם ד"ר חסין. במקום זאת דיווחתי לו על התקדמותי במייל והמשכתי לחשוב על ייעולים נוספים עבור הקוד.

14/1

בזמן שהתפנה לי עקב ביטול הלימודים בעקבות השלג (למרות שהייתי חולה מאד ברוב הימים) יצא לי לחשוב על ייעולים נוספים למימוש. הצלחתי לצמצם עוד יותר את צריכת הזיכרון ולכן צריכת הזיכרון של הגרסה הרביעית הוא 3n. ייעול זה משמעותי גם אם הוא מקטין רק בחצי את צריכת הזיכרון מכיוון שבפוטנציאל אם יש לנו זיכרון נתון ומספר רב של מעבדים, אם הזיכרון שדורש מעבד אחד קטן בחצי אזי כמות המעבדים שנוכל להפעיל גדלה פי 2.

לאחר מימוש הגרסה הרביעית עלה בדעתי ייעול נוסף שיביא לצריכת זיכרון של 2n, כרגע השארתי את המימוש תיאורטי בלבד מחמת חוסר זמן בשבוע החדש עקב השלמות רבות שאני צריך לבצע של תרגולים שהתבטלו בגלל השלג. אציג לד"ר חסין את הצעת הייעול ואראה האם לממש אותה או שהייעול אליו הגעתי הינו מספיק.

בפגישה השבועית הצגתי לד"ר חסין את ההתקדמות בפרויקט עד כה, את הגרסאות השונות והייעולים אליהם הגעתי. הוא ביקש שאתחיל ללמוד להשתמש ב - CUDA parallel programming model, שזהו כלי בו נתכנת בשפת C ונוכל להריץ חישובים על גבי כרטיס מסך.

21/1

בשבוע שעבר מאז העדכון האחרון התחלתי להיערך להגשת דוח האמצע. בין היתר ביצעתי את הפעולות הבאות: סידרתי את גרסאות הקוד השונות כך שיהיו ברורות וקריאות, תיעדתי את העבודה שעשיתי עבור הגרסאות השונות, עשיתי בדיקות נכונות (ותיקנתי את הבעיות שנתקלתי בהן) וכן בדיקות זמן ריצה כדי לראות את ההבדלים בזמני הריצה בין הגרסאות השונות.

נפגשתי עם ד"ר חסין ודיברנו על הגשת אב הטיפוס שייערך בעוד פחות משבוע, קבענו את הסדר ההצעה והוחלט שאתרכז בזמן הקרוב בהגשה של הדוח וכן הצגת הפרויקט שבוע אחרי ואשים בעדיפות משנית את המשך העבודה על הקוד.

25/1

נפגשתי עם ד"ר חסין והצגתי לו את טיוטת דו"ח אב הטיפוס שהכנתי, הוא נתן לי מספר הערות וכן שינויים אותם הוא רוצה שאכניס לדו"ח. הוא ביקש ממני להוסיף גרפים המתארים את השפעת אורך המחרוזת/מספר המחרוזות אותן מעבדים על זמן הריצה וכן לפרט היטב את הדוגמאות שכללתי בדו"ח.

יש לי עד הערב לתקן את כל הדברים אותם הוא וביקש ולהגיש את דו"ח אב הטיפוס ולאחר מכן אתחיל לעבוד על המצגת אותה אציג במהלך ההצגה שנקבעה לתאריך 2/2.

28/1

נפגשתי עם ד"ר חסין והצגתי לו את המצגת שהכנתי והוא העיר לי שהיא מלאה ביותר מידי מלל ולא ניתן לעקוב אחרי ההתקדמות שלה בקלות, הוא הטיל עלי לפשט את המצגת ככל הניתן למרות סיבוכיות אלגוריתם BWA על שלביו השונים. ד"ר חסין היה מרוצה מהדוגמאות שהכללתי בתוך המצגת ואמר לי להבליט אותן בזמן שאסביר מילולית מה רואים בהן במהלך ההצגה.

2/2

הצגתי היום את אב הטיפוס, בכללי היה נראה שההצגה עברה בטוב למרות שאני לא מרגיש טוב ביומיים האחרונים ושאני צרוד בימים האחרונים. הקול שלי נשבר מספר פעמים במהלך ההצגה ואחרי שיצאתי לא יכולתי כבר לדבר בכלל אך מלבד זאת הפסקתי להעביר את החומר שתכננתי ולענות לשאלות הבוחנים.

4/3

עקב בעיות בריאות קשות במהלך החודש האחרון (מצבי הבריאותי התדרדר מיד לאחר הצגת אב הטיפוס) לא הגעתי לאחד ממבחני סוף הסמסטר שלי (מתוך שניים) וכן לא התקדמתי כלל עם פרויקט הגמר בחודש שעבר.

עכשיו כשאני מרגיש יותר טוב אעבור שוב על מצב הפרויקט וכן אמשיך מהמקום שבו הפסקתי, מימוש גרסה מספר חמש של אלגוריתם SW כך שזמן הריצה יישאר אותו הדבר (כלומר n^2) אך הזיכרון הנצרך יקטן ויעמוד על 2n בלבד.

10/3

השבוע נפגשתי עם ד"ר חסין לאחר שלא נפגשנו חודש בעקבות מצבי הבריאותי, עדכנתי אותו בסיבות לחוסר ההתקדמות, בצורת פעולת הגרסה החדשה ביותר של הקוד וכן דננו בהמשך הפרויקט.

ד"ר חסין הטיל עלי להריץ את קוד ה-BWA האמיתי על מנת שנוכל לנתח את מבני הנתונים, נדע כיצד למדוד את זמני הריצה וכן כדי שנוכל לבצע אינטגרציה של הקוד עליו אנו עובדים לתוך המערכת הקיימת.

בשבוע הקרוב אתחיל לעבוד יותר עם מחשבי המכללה מכיוון שאיני יכול להריץ BWA במחשב הפרטי מחמת חוסר התאמה של מערכת הפעלה (דרישה ללינוקס 64 ביט) ומכיוון שלא ניתן לעבוד במחשבי המכללה עם Github (לא ניתן להתקין תוכנות נוספות במחשבי המכללה) אשתמש מעתה ב-google drive כדי לסנכרן קוד במהלך הסמסטר ואעלה את הקוד ל-github רק בסוף העבודה.

17/3

לא הצלחתי לבצע את המטלה שד"ר חסין הטיל עלי בשבוע האחרון (הרצת קוד BWA האמיתי) מכיוון שאין במחשבי המכללה את הזיכרון הנדרש להרצת התהליך, קובץ ה-DNA שוקל 3.2GB וכל קובץ דגימות שוקל 750MB וכן יש להוסיף את גדלי קבצי הביניים וקבצי התוצאות.

פנייה למחלקת המחשוב לשם הגדלת הזיכרון המוקצה עבור המשתמש שלי גרמה לכך שאקבל 500MB במקום 250MB אותם יש לכל סטודנט רגיל אך כאשר פנינו אליהם פעם נוספת הם השיבו שלא ניתן להקצות עבורי יותר זיכרון.

במהלך הפגישה השבועית עדכנתי את ד"ר חסין בבעיה עם מחשבי המכללה והוא פנה אל מחלקת המחשוב על מנת שיסדרו לנו מחשב פרויקט ייעודי, כלומר מחשב ללא מגבלות על הזיכרון המוקצה מלבד מגבלות החומרה, הרשאת גישה של admin על מנת שנוכל להתקין כל תוכנה שנרצה וכן שיהיה נעול לכל משתמש אחר מלבד לפרויקטים של ד"ר חסין. בנוסף הוחלט שכל עוד לא ניתן לעבוד עם מחשבי המכללה ועם קוד BWA האמיתי, נכתוב את שלבי ה-index וה-alignment שיעבדו בצורה נכונה ויעילה כדי שנוכל לממש את שלב ה-pairing בעצמנו ולאחר מכן לעשות התאמה עבור מבני הנתונים של קוד ה-BWA.

בנוסף בשבוע הבא יש לי מועד ב עבור המבחן אותו לא עשיתי בעקבות המחלה, אי לכך לא אוכל להתקדם הרבה במהלך השבועיים הקרובים עד הפגישה הבאה עם ד"ר חסין.

31/3

בפגישה עם ד"ר חסין דיווחתי לו שאני נתקל בבעיות אי נכונות של הקוד (מספק תוצאות לא נכונות) עבור המימוש בשפת C. מכיוון שהקוד איתו עבדתי הוא הקוד שמימשתי בעבר וגם הוא לא עבר ולכן עברתי למימוש בשפת JAVA ד"ר חסין הציע לכתוב את הקוד מחדש תוך שימת דגש על המימוש של שפת JAVA שעיקרו עבד נכון, בנוסף הוא הציע שמאחר ושלושת השלבים לא קשורים אחד בשני אפילו ניתן להשלים את המימוש בשפת JAVA למימוש מלא ונכון ורק לבצע את השלב האחרון בשפת C.

במהלך השבוע הקרוב אכתוב את הקוד מחדש בשפת C החל מהבסיס ועד ממשק המשתמש ואם עדיין לא אצליח אשלים את מימוש שפת JAVA שנוח יותר לדיבוג ובדיקה.

14/4

במהלך חופשת הפסח סיימתי את מימוש שפת C, עדיין היתה בעיה עם הקצאת הזיכרון ולכן עברתי חזרה לשפת JAVA והשלמתי את המימוש לגמרי החל מתחילת שלב ה-index ועד סיום שלב ה-alignment עבור כלל המחרוזות.

בדקתי את נכונות הקוד שמימשתי בשפת JAVA בשני שלבים:

בשלב הראשון בדקתי את נכונות שלב ה-index, הכנסתי מספר דוגמאות קצרות מהאינטרנט ובדקתי שאכן הפלט (המחרוזת B ומערך המיקומים SA) תואם את הדוגמאות שהבאתי. בנוסף בדקתי עבור מחרוזת ארוכה רנדומאלית את פעולת האלגוריתם ווידאתי מול הרצה ידנית של התהליך שאכן התוצאות נכונות.

בשלב השני בדקתי את נכונות שלב ה-alignment ע"י הכנסה של מספר מחרוזות התייחסות וכן מחרוזות דגימה ממספר דוגמאות על קוד BWA הקיימות באינטרנט, את התוצאות שקיבלתי מריצת התהליך שלי השוואתי לתוצאות שפורטו בדוגמאות הנ"ל. השתמשתי בדוגמאות מהאינטרנט בלבד מכיוון שלא קל להריץ באופן ידני את שלב ה-alignment עבור מספר רב של מחרוזות.

בפגישה עם ד"ר חסין דיווחתי לו על כל העבודה במהלך השבועיים האחרונים והוא נתן אור ירוק לכתיבת קוד הלוקח את מחרוזות הדגימה וכן את כלל ההתאמות שנמצאו ונותן עבור כל מחרוזת דגימה את ההתאמה הטובה ביותר מבין ההתאמות שנמצאו עבורה ע"י שימוש בקוד ה-SW. בשבוע הקרוב אכתוב את הקוד שמפעיל את אלגוריתם SW עבור כל התאמה שנמצאה ובוחר את ההתאמה הטובה ביותר, בנוסף אצטרך לכתוב קוד המשחזר את הפיתרון של ההתאמה הטובה ביותר מכיוון שהמידע הזה לא נשמר בפונקציות שכתבנו אלא רק הציון של ההתאמה.

21/4

כתבתי את קוד שחזור הפתרון עבור של אלגוריתם SW בהתבסס על גרסה מספר 2 שמימשתי בעבר, לא ניתן להתבסס על גרסאות מתקדמות יותר מכיוון שהחיסכון בזיכרון הקיים בגרסאות מתקדמות יותר מונע את בניית טבלת שחזור הפתרון.

בנוסף כתבתי קוד הלוקח דגימה ועבור כל התאמה שנמצאה מחשב את ציון ההתאמה, עבור ההתאמה המקסימאלית מחושב שחזור הפיתרון והתוצאה מודפסת לתוך קובץ תוצאות.

במהלך הבדיקות שביצעתי עם דוגמאות מהאינטרנט גיליתי שיש לי טעות ולמרות שההתאמה הטובה ביותר נמצאה תמיד הודפסה ההתאמה האחרונה שנמצאה יחד עם ציון ההתאמה המקסימאלית, הקוד תוקן לשמירת ההיסט של ההתאמה המקסימאלית. בנוסף גיליתי שלא בהכרח ציון ההתאמה שוכן בתא האחרון בטבלת SW אלא הוא הציון הגבוהה ביותר הקיים במהלך חישובי SW, תיקנתי את קוד ה-SW וכן את קוד שחזור הפתרון כדי שיתייחסו לעניין זה, ערכתי את הגרסאות הקיימות בתוך קוד השלב הנוכחי בלבד מכיוון שלחזור ולשנות את כלל הגרסאות שאין בהן שימוש יהיה בזבוז זמן.

בפגישה השבועית עם ד"ר חסין דיווחתי לו על ההתקדמות והוחלט שכעת שהקוד עובד ננסה למקבל אותו בכמה אופנים:

1. מקבול הרצת אלגוריתם SW עבור כל התאמה (מקבול thread per hit).
2. מקבול הקוד האחראי לבצע SW עבור כל התאמה שנמצאה עבור מחרוזת הדגימה והחזרת התוצאה (מקבול thread per read).
3. מקבול הקוד האחראי לבצע SW עבור כל התאמה וכן מקבול הרצת אלגוריתם SW עבור כל התאמה (מקבול thread per hit inside thread per read)

28/4

בשבוע האחרון עבדתי על מקבול הפונקציה המבצעת את אלגוריתם SW כך שנוכל להפעיל את האלגוריתם עבור כל התאמה קיימת בצורה מקבילית. בדיקת הקוד העלתה שהוא עובד בצורה תקנית ונכונה ואכן מוצא את ההתאמות הטובות ביותר בדיוק כמו הקוד העובד סדרתית. בדיקת זמני ריצה העלתה שהריצה המקבילית לוקחת הרבה יותר זמן מאשר הריצה הסדרתית, בחינה מעמיקה של הסיבה לכך העלתה שהתקורה של התהליך הממוקבל היא גבוהה מידי, הזמן המתבזבז על יצירת כמות ה-threads, ההפעלה שלהם וההמתנה שיסתיימו כדי לבחור את ההתאמה המקסימאלית עולה הרבה יותר מזמן הריצה של קוד ה-SW עצמו. בנוסף בגלל שמספר כמות ה-threads הוא גבוהה, זמן ה-context switch כבר אינו זניח והוא גורם משפיע במערכת.

בפגישה עם ד"ר חסין הוחלט לזנוח את המקבול המדובר (מקבול thread per hit) וכן לוותר גם על אופציית המקבול השלישית שכן היא מכילה בתוכה את אותו עיקרון ולכן את אותה הבעיה בה נתקלנו עם אופציית המקבול הראשונה. ד"ר חסין הנחה אותי לפנות לאופציית המקבול השנייה (מקבול thread per read) ולכן לחפש פתרון חסכוני במספר ה-threads העובדים בו זמנית (מספר ה-threads יהיה כמספר הליבות במערכת בלבד ואז לא יהיו context switch מיותרים).

בנוסף קבענו שלא ניפגש שבוע הבא שכן יש לי להציג פרויקט בקורס "נושאים מתקדמים במערכות אחסון" בתאריך 11/4 וזמן העבודה על הפרויקט ייקח כשבוע.

12/5

כדי לממש את אופציית המקבול השנייה שהצענו (מקבול thread per read) החלטתי להשתמש בקוד של threadpool אותו מימשתי בשפת C. הגעתי להחלטה זו לאחר התייעצות עם גב' שמרית צור דוד שאף מימשה נתנה לנו תרגיל המשתמש בקוד threadpool במהלך הקורס "יישומי תקשורת מחשבים" בלמדתי בשנתי השלישית לתואר.

מעיקרי ה-threadpool: פתיחתו של ה-threadpool קובעים את מספר ה-threads שירוצו על המערכת ומספר זה אינו משתנה במשך ריצת התוכנית, את מספר זה נקבע לפי כמות הליבות השונות במחשב עליו מריצים את התוכנית, ל-threadpool שולחים עבודות והוא דואג לקחת את העבודות מהתור, לבצע אותן ולמשיך לקחת את העבודה הבאה מהתור עד שאין עוד עבודות בתור וכן נשלחה הפקודה לסגור את ה-threadpool.

התאמתי את הקוד כדי שהטיפול בכל מחרוזת דגימה (read) ייעשה בתוך פונקציה המקבלת את המחרוזת ואת ההתאמות שנמצאו עבורו. בדיקות של הקוד העלו שהוא עובד כמו שצריך וההוצאה של התהליך לפונקציה נפרדת לא פגעה בנכונותו של הקוד. בדיקות מתקדמות עם ה-threadpool העלו שיש בו שגיאה שגורמת לתוכנית לקרוס, הבעיה עדיין לא נפתרה.

בפגישה עם ד"ר חסין הוא דיווח לי שהוא קיבל הודעה ביום שלפני שמחשב פרויקט ייעודי הוכן עבורנו במעבדת המחשבים ולכן הוא הנחה אותי לזנוח את הקוד עליו אני עובד עכשיו ולהקדיש את השבוע הקרוב להרצת אלגוריתם BWA ולהתחיל לנתח את זמני הריצה ומבני הנתונים העומדים בבסיסו.

19/5

בשבוע שעבר מאז קבלת המחשב הייעודי עסקתי בהתקנות שונות של תוכנות שישמשו אותנו בעבודה כגון סביבת עבודה וירטואלית של לינוקס לשם הרצת הקוד, תוכנות המסוגלות להתמודד עם קבצי טקסט גדולים, התקנת תוכנות שליטה מרחוק (לא עובד עדיין מכיוון שהרשת של המכללה חוסמת את הפורטים המיועדים, בקשה הוגשה למחלקת מחשוב למען טיפול בעניין) וכן github desktop לשם מעבר לעבודה במאגר קוד מסוגר ונוח להמשך הפרויקט.

בנוסף למדתי שלא ניתן לקמפל את קוד ה-BWA על המחשב הייעודי ולכן אצטרך לקמפל כל פעם על מחשב לינוקס הפתוח לכולם ולהעביר את התוכנית המקומפלת למחשב הייעודי לשם ההרצה (במחשב אחד לא ניתן לקמפל כי זו אינה מערכת לינוקס ובמחשב אחר לא ניתן להריץ כי אין את הזיכרון הדרוש) ושוב לא ניתן להשתמש ב-github כדי לסנכרן קוד מכיוון שלא ניתן להתקין אותו על המחשב בעל יכולת הקימפול ולכן נישאר עם google drive כדי להעביר מידע בין המחשבים השונים.

את הרצת קוד ה-bwa עושים באופן הבא:

* קימפול הקוד ע"י הפקודה

make bwa

* הרצת שלב ה-index על קובץ הגנום

./bwa index hg19.fa

* הרצת שלב ה-alignment על הגנום וקריאה ראשונה

./bwa aln hg19.fa read1.fastq -f read1.sai

* הרצת שלב ה-alignment על הגנום וקריאה שנייה

./bwa aln hg19.fa read2.fastq -f read2.sai

* הרצת שלב ה-pair end על הגנום, תוצאי שלב ה-alignment והקריאות

./bwa sampe hg19.fa read1.sai read2.sai read1.fastq read2.fastq > read.sam

ובמקרה שרוצים לשנות דברים בקוד הקיים, קימפול נעשה עם הפקודה:

gcc -g -Wall -Wno-unused-function -O2 -DHAVE\_PTHREAD -DUSE\_MALLOC\_WRAPPERS QSufSort.c bwt\_gen.c bwase.c bwaseqio.c bwtgap.c bwtaln.c bamlite.c is.c bwtindex.c bwape.c kopen.c pemerge.c bwtsw2\_core.c bwtsw2\_main.c bwtsw2\_aux.c bwt\_lite.c bwtsw2\_chain.c fastmap.c bwtsw2\_pair.c main.c -o bwa -L. -lbwa -lm -lz -lpthread

זמן רב במהלך השבוע שעבר התבזבז על המתנה להתקנות השונות, הפעלת שלב האינדקס מספר פעמים (לא היה מספיק זיכרון פנוי לחישובי הביניים והייתי צריך לאנדקס את ה-DNA מספר פעמים שכל פעם התהליך קורס ואני מגדיל את הזיכרון במערכת הוירטואלית, כל הרצה לוקחת כשעה וחצי) וכן הרצות ממשיות של הקוד שלקחו זמן רב עקב המספר הגבוהה של מחרוזות דגימה שעבדתי איתן.

השבוע לא התקיימה פגישה עם ד"ר חסין שכן הוא היה צריך לצאת לפני הזמן הרגיל של הפגישה שלנו מהמכללה ולא הצלחנו לתאם מועד אחר לפגישה באותו שבוע. החלטתי לסיים עם הקוד עליו עבדתי לפני שבוע שכן קוד ה-BWA הינו מסובך וקשה להבנה ומכיוון שאין לי מושג איך להמשיך בעבודה כרגע אני אסיים על מה שהתחלתי ואתקן את קוד ה-threadpool כדי שיהיה לי את לב העבודה של הפרויקט כבר מוכנה.

26/5

מאז העדכון האחרון ישבתי לנתח את קוד ה-threadpool כדי לאתר את בעיית הסגמנטציה שבו. ביצעתי בדיקות של מקרים רבים של שליחות של עבודות לתור והוצאה וביצוע שלהן וגיליתי שהבעיה מופיעה כאשר התור מתרוקן ולאחר מכן מתמלא חזרה בעבודות נוספות, תיקנתי את הקוד וכעת ניתן להריץ את הקוד בצורה ממוקבלת או בצורה סידרתית וכן לקבוע את מספר ה-threads הקיימים בתוך ה-threadpool ולראות כיצד המשתנים הללו משפיעים על זמן הריצה.

בפגישה עם ד"ר חסין דיווחתי לו על ההתקדמות וכן הוא הנחה אותי לדבג את קוד ה-BWA כדי לאתר בצורה מדויקת את אופן העבודה של הקוד (איזה פונקציות עובדות, כמה פעמים כל פונקציה נקראת וכו') על מנת שנוכל לבצע את החיבור בין הקוד הממוקבל שכתבנו עבור השלב השלישי ובין הקוד הקיים לשני השלבים הראשונים מכיוון שבניגוד לקוד שאני מימשתי עבורם הקוד הקיים ב-BWA הוא יעיל במיוחד ומהיר הרבה יותר מהקוד שלי.

2/6

בפגישה השבועית עם ד"ר חסין הצגתי לו את התוצאות אותן קיבלתי במהלך השבוע בו עבדתי על הקוד, הראתי לו עד היכן הגעתי מבחינת דיבוג וכן את התוצאות עד חלק זה. ד"ר חסין הטיל עלי לחזור אחורה ולהוסיף מדידת זמנים עבור כל החלקים עליהם עברתי עד כה בניסיון לזהות בדיוק את הפונקציות הבעייתיות וכן לגבות זאת על ידי נתוני אמת. יידעתי את ד"ר חסין שיש התקדמות אך היא איטית כי כל ריצה לוקחת לפחות רבע שעה וזאת מכיוון שאם אני מכניס מספר קטן של מחרוזות לא ניתן לזהות בוודאות את קטעי הקוד שלוקחים זמן וכן מספר הקריאות לפונקציות השונות תלוי בצורה ישירה במספר מחרוזות הדגימה בהן אני משתמש.

בנוסף בשבוע הקרוב לא יהיה לנו זמן לפגישה נוספת שכן יש לי פרויקט נוסף להציג (בתאריך ה-15/6) בקורס "נושאים מתקדמים במערכות תקשורת" ואת ההצגה שלו אני צריך להכין במהלך השבוע הבא.

16/6

בפגישה עם ד"ר חסין הצגתי לו את מדידת הזמנים השונים אליהם הגעתי שהם עדיין לא ניתוח מלא של כל הפונקציה ולאחר שהוא בחן את הקוד בעצמו והתחקה אחרי הנתיב שנתתי לו הוא אמר לי שהזמן לעבוד על הפרויקט מתקצר מאחר שמועד ההגשה מתקרב בצורה מהירה ולכן הוא הנחה אותי לעזוב בינתיים את ההתעסקות עם מדידת הזמנים של BWA ושלח אותי להפעיל את קוד הפרויקט barraCUDA על מחשב הפרויקט הייעודי (בעל כרטיס מסך gtx980) שהוא פרויקט נוסף שממקבל את אלגוריתם BWA על שלביו השונים.

23/6

במהלך השבוע האחרון ניסיתי להתקין את הדרייברים המתאימים לכרטיס המסך ללא הצלחה כלל, לאחר שדיברתי עם אבי טרנר (הסטודנט העובד על מקבל השלב השני ב-BWA) הוא אמר לי שלא ניתן להתקין את הדרייברים המתאימים על מכונה וירטואלית וכן יש לשנות את מערכת ההפעלה עצמה מ-windows ל-linux.

יידעתי את ד"ר חסין על הבעיה במייל והוא נתן לי שתי משימות, הראשונה היתה לבחון את הקוד של פרויקט barraCUDA לחיפוש קטעי הקוד הממוקבלים לכרטיס מסך והשנייה היתה לפנות למחלקת מחשוב על מנת שישנו את מערכת ההפעלה בקיימת על המחשב. במחלקת המחשוב אמרו שהם יטפלו בבעיה כאשר הם יתפנו לכך, לאחר שאבי סיפר לי שהוא לוחץ עליהם ולא הולך לו במיוחד סיפרתי לד"ר חסין על כך במהלך פגישה מקרית במסדרון והוא לקח אותי למחלקת מחשוב ודיבר איתם על הנושא כבר וביקש שיטפלו בכך במהירות האפשרית.

בחינת קוד פרויקט barraCUDA העלתה שהם לא מיקבלו עדיין את השלב האחרון שתהליך שעליו אנו עובדים אלא רק את השלב הראשון (index) והשני (alignment). בפגישה השבועית עם ד"ר חסין הוא הציע שננסה לחבר את הקוד אותו מיקבלתי לקוד ה-barraCUDA כדי להציג תוצאות אפילו יותר טובות מבחינת זמן הריצה מכיוון שאם המיקבול עובד על מספר הליבות המוגבל הקיים במעבד שבמחשב, נשיג תוצאות אפילו יותר מרשימות כשהוא יפעל על מספר גבוה של מעבדים (בכרטיס המסך המחובר למחשב הפרויקט, gtx980, יש 2048 מעבדים).

ד"ר חסין יידע אותי בפגישה השבועית שמחלקת מחשוב חזרו אליו כמה דקות לפני הפגישה שלנו ויידעו אותי שבאותו הערב הם הולכים להתקין את מערכת ההפעלה החדשה, הוא אמר לי לחכות שהם יסיימו את ההתקנה ולהריץ את קוד ה-barraCUDA כמה שיותר מהר. בנוסף ד"ר חסין פנה לד"ר יגל בבקשת דחייה עבור תאריך הצגת פרויקט הגמר וכן מסירת הדו"ח והנספחים המפורטים באתר פרויקט הגמר.

28/6

עד כה לא הותקנה מערכת ההפעלה החדשה, שלחתי לד"ר חסין מייל בנוגע לעניין והוא הנחה אותי להתחיל להכין את הפרויקט למסירה.

בין המטלות השונות אותן עלי לבצע לקראת שלב המסירה:

* 1. העברת הקוד מ-google drive אל מאגר הקוד ב-github (בגלל שלא הייתי יכול להשתמש ב-github במחשבי המכללה הייתי צריך פתרון אחר להעברת קוד בין מחשבים שונים).
  2. הכנת הסקריפט לסרטון והצגה לד"ר חסין לאישור.
  3. העלאת יומן הפרויקט מהמחשב האישי לאינטרנט (במהלך הצגת אב הטיפוס התברר שירדו לי נקודות רבות על כך שהטקסט היה מעומד בצורה מוזרה, כשבחנתי זאת ראיתי שהשימוש ב-google docs לא התאים והיומן כולו התעוות).
  4. כתיבת הדו"ח הסופי.

ד"ר חסין הודיע לי שאני מציג את הפרויקט בתאריך 13/7 ולכן עלי להגיש את כלל הדברים המפורטים למעלה עד התאריך 6/7.

עדכון: בשעות הערב קיבלתי מייל שמערכת ההפעלה החדשה הותקנה כמו שצריך, מחמת חוסר זמן עד מועד ההגשה הנושא כבר לא עומד על הפרק ולא אשתמש בזה כלל. בנוסף אבי דיווח לי יומיים אחר כך שהמחשב לא עובד כמו שצריך והחזירו אותו למערכת ההפעלה הישנה שלו.

5/7

נפגשתי עם ד"ר