

**המחלקה להנדסת תוכנה**

**Paralleling BWA-Aligner**

# מקבול חיפוש מיקומם של מקטעי DNA על פני הגנום האנושי.

**חיבור זה מהווה חלק מהדרישה לקבלת**

**תואר ראשון בהנדסה**

**מאת**

**אבי טרנר**

תמוז תשע"ה יולי 2015



**המחלקה להנדסת תוכנה**

**Paralleling BWA-Aligner**

**מקבול חיפוש מיקומם של מקטעי DNA על פני הגנום האנושי.**

**חיבור זה מהווה חלק מהדרישה לקבלת**

**תואר ראשון בהנדסה**

**מאת**

**אבי טרנר**

**מנחה: ד"ר חסין יהודה**

**אישור המנחה: תאריך:**

# תקציר

הפרויקט הוא פרויקט מחקרי, על מנת לאפשר הבנה טובה יותר של בעית החיפוש בDNA ולהציע פתרון. נפתח בתיאור קצר של הרקע ושל מונחים הקשורים לפרויקט.

**DNA**: הDNA היא מולקולת ענק שמצויה בכל אחד ואחד מתאי הגוף שלנו ובה מצוי כל המידע התורשתי לבניית החלבונים בתא אצל כל האורגניזמים הידועים, מחיידקים ועד לבני אדם.

תמונה 1

המבנה של הDNA בנוי כמעיין "סולם" שמסתלסל סביב עצמו, כאשר ה"שלבים בסולם" מורכבים, כל אחד, מזוג בסיסים המתחברים זה לזה ומסומנים באותיות הלטיניות A, G, T, C.

בכל "שלב" מתחברים הבסיסים עם בן זוג קבוע – A עם T ו C עם G, כך שאם ידוע לנו רק צד אחד של ה"סולם" אנו יכולים לשחזר ממנו במדויק גם את הצד השני.

המדהים הוא שכ99.9% מהDNA של כל בני האדם משותף למרות אבני הבניין המועטות והפשוטות– וזוהי תכונה קרדינלית לפרויקט זה.

אם כן, את המידע המגולם בDNA ניתן לייצג כמחרוזת של הבסיסים המרכיבים אותו, וכך נתייחס אליו בפרויקט זה (לדוג': TGACCGTCAG....)

הDNA מורכב מכ שלבים. במונחים דיגיטלים, המידע הגלום בDNA שווה ערך למעט יותר מ 1.6 GB (או כ2 דיסקים...)



תמונה 2

כמו כן, DNA יכול לעבור מוטציה, שינוי. רוב המוטציות אינן מזיקות אך אם הן מופיעות במקומות מסוימים על גבי רצף הDNA הן יכולות לגרום לבעיות גנטיות וביניהן לנטיה למחלות גנטיות ובפרט לסרטן.

כיום כאשר חולה מגיע לאבחון לראות האם הוא חולה בסרטן, תהליך הבדיקה הוא ארוך ומסורבל:

1. לקיחת דגימת DNA.
2. השוואת הDNA לDNA של אדם בריא.
3. איתור מוטציות בDNA.
4. חיפוש ידני במאגרי מידע האם המוטציות הן במקום שידוע כגורם לסרטן.
5. חיפוש ידני האם קיימת תרופה שעוזרת לסוג המסוים של הסרטן הנ"ל.

בין הבעיות בתהליך זה ניתן לציין את תהליך ההשוואה (2) שלוקח זמן ארוך במיוחד. כיום, בתי החולים שוכרים חוות שרתים ע"מ ליעל את החישוב הארוך של השוואה.

עם זאת, גם כיום שלב זה לוקח כיום שלם – פרויקט זה יתמקד בבחינת ייעול שלב ההשוואה.

על מנת לבחון את ייעול שלב זה התמקדנו בבחינת:

1. השפעות אורך הדגימות על זמן הריצה.
2. השפעות מספר הדגימות על זמן הריצה.
3. השפעות של מיון הדגימות על זמן הריצה.
4. השפעת מקבול התהליך על זמן הריצה.

לאחר בדיקה והשוואות הממצאים נראה שהתוצאות הטובות ביותר מתקבלות ע"י מקבול התהליך, באופן שבו כל דגימת DNA מקבלת תהליכון (בניגוד למקבול פנימי של התהליך).

יעול מסוג זה נשאר יציב ללא תלות באורך / מספר / מיון הדגימות.

# הצהרה

**ההעבודה נעשתה בהנחיית ד"ר חסין יהודה, המכללה האקדמית להנדסה ירושלים – עזריאלי המחלקה להנדסת תוכנה.**

**החיבור מציג את עבודתי האישית ומהווה חלק מהדרישות לקבלת תואר ראשון בהנדסה.**

# תודות

**ברצוני להביע תודות לאנשים שבהם נעזרתי בפרויקט.**

**תודה מיוחדת לאשתי ומשפחתי, שעזרו ותמכו בי לאורך כל מהלך הלימודים בכלל והפרויקט בפרט.**

**תודה לעמיתי עז ידגר, מבצע הפריקט המקביל, שעמו התייעצתי במהלך הפרויקט.**

**ולבסוף תודה מיוחדת למנחה הפרויקט ד"ר יהודה חסין על העזרה המקצועית, יחסו האישי, העצמאות והתמיכה שניתנה לי במהלך ביצוע עבודת הגמר.**

**תודה!**

# תוכן העניינים

[מקבול חיפוש מיקומם של מקטעי DNA על פני הגנום האנושי. 1](#_Toc423703731)

[1. תקציר 3](#_Toc423703732)

[2. הצהרה 5](#_Toc423703733)

[3. תודות 6](#_Toc423703734)

[4. תוכן העניינים 7](#_Toc423703735)

[5. מילון מונחים 8](#_Toc423703736)

[6. תיאור מסגרת הפרויקט 8](#_Toc423703737)

[7. תיאור הבעיה 9](#_Toc423703738)

[8. תיאור הפתרון 11](#_Toc423703739)

[המחשה של שלבי האלגוריתם 12](#_Toc423703740)

[תיכון 13](#_Toc423703741)

[שימוש ברכיבים קיימים 13](#_Toc423703742)

[9. תיאור המערכת 13](#_Toc423703743)

[מהי המערכת 13](#_Toc423703744)

[תהליכים ונתוני המערכת 13](#_Toc423703745)

[תיאור המימוש 14](#_Toc423703746)

[10. בדיקות 18](#_Toc423703747)

[11. מסקנות 19](#_Toc423703748)

[12. השוואה לפתרונות בספרות 19](#_Toc423703749)

[טבלת סיכונים 21](#_Toc423703750)

[13. רשימת ספרות 21](#_Toc423703751)

[14. נספחים 22](#_Toc423703752)

[15. Abstract 23](#_Toc423703753)

# מילון מונחים

– קריאה – קטע מהDNA של חולה שנדגם לצורך בדיקה.

– רפרנס - הגנום שנשתמש בו להשוואה לכל הקריאות.

– base pair - זוג בסיסים המהווים שלב ב"סולם הDNA". משמש כמידת אורך של קריאה.

- אורך של קריאה. באופן טיפוסי הקריאות הן באורך של bp 100- bp 200 אך ה"לקוח" הספציפי של תוצר הפרויקט, משתמש בקריאות באורך bp35.

– המחרוזת על גבי באורך , המתחילה באינדקס .

– הקריאה שהחלה במקום ה ב X

– אורכו של X ביחידות bp -

Illumina – מכונה הדוגמת קריאות.

מרחק עריכה – מרחק עריכה בין שתי מחרוזות מוגדר כמספר המינימלי של פעולות עריכה שיש לבצע על מחרוזת אחת כדי להגיע למחרוזת השנייה, כאשר פעולות העריכה המותרות הן: הוספת אות, מחיקת אות או שינוי אות לאות אחרת.

שגיאה בקריאה – תו שנקרא כתו אחר, תו שהוסף, או תו שהושמט בקריאה שנקראה ע"י Illumina,

# 

# תיאור מסגרת הפרויקט

לאחר שב2003 מופה הגנום האנושי, תהליכים רפואיים הקשורים לגנום הפכו לאפשריים.

אבחון מוטציות היה אחד היישומים המבטיחים, בפרט בתחום אבחון הסרטן.

האתגר שבא לפתחם של המהנדסים היה היכולת לעבד כמויות מידע בסדר הגודל של הגנום האנושי, תוך שמירה על זמן ריצה מעשי.

פרוייקט זה מבוצע תחת הנחייתו של דר' חסין יהודה.

פרוייקט זה מהווה פרוייקט גמר לתואר "הנדסת תוכנה" במכללת "עזריאלי – המכללה האקדמית להנדסה" בירושלים.

סטודנטים נוספים העובדים על פרוייקטים הקשואים לפרויקט זה: עז ידגר ומוטי מונסונגו.

# תיאור הבעיה

***הרחבה: NGS – Next Generation Sequencing – ריצוף מהדור החדש***

*באופן כללי, Sequencing היא שיטה לקביעת הסדר המדויק של נוקלאוטידים במולקולת DNA נתונה, או בהפשטה – בהינתן מולקולת DNA, מהו סדר אבני הבסיס שבה.*

*בעשור האחרון השימוש בSequencing גדל בצורה אקספוננציאלית בעיקר בשירות מעבדות קליניות ומחקר בכל רחבי העולם, כשהמתשמש הבולט מביניהם היה פרויקט מיפוי הגנום האנושי שהסתיים ב2003.*

*בפרויקט מיפוי הגנום האנושי, נעשה שימוש בSanger Sequencing שפותח ב1977 והיה שייך לדור הראשון של הSequencing (ריצוף).*

*מאז השלמת רצף הגנום האנושי הראשון, חל גידול בביקוש לשיטות Sequencing זולות ומהירות יותר. דרישה זו האיצה את פיתוח Next Generation Sequencing (NGS). פלטפורמות NGS מבצעת ריצוף מסיבי באופן מקבילי, שבמהלכו מיליוני שברי DNA ממדגם בודד מרוצפים בתיאום. כחלק מהתהליך, שברי הDNA נחתכים ל"אורך" בגלל המבנה של הDNA , ניתן לשחזר כל צד מהDNA ע"י בן זוגו שנחתף. מקטעי DNA אלו הם הקלט של פרויקט זה.*

*טכנולוגיית ה NGS מאפשרת תפוקה גבוהה מהדור הקודם, וריצוף של כל הגנום בפחות מיום אחד.*

*בעשור האחרון, מספר פלטפורמות NGS פותחו והן מספקות ריצוף תוך עלות נמוכה, וקצב תפוקה גבוה.פיתוח פלטפורמות אלו הנגיש את הNGS עבור מספר גדל והולך של מעבדות, ובעקבות כך גדל המחקר והאבחון קליני המבוצעים ע"י ריצוף חומצות גרעין.*

*בין שתי הפלטפורמות הנפוצות קיום עבור NGS, נתין לציין את Illumina שנמצא בשימוש בביה"ח הדסה, ועבורו מתבצע פרויקט זה.*

**הקדמה:**

כחלק מתהליך אבחון סרטן אצל חולים, נדגם הDNA שלהם ונבדק על מנת למצוא מוטציות שעלולות לגרום לסרטן.

תהליך הדגימה מתבצע באמצעות מכונה הנקראת Illuminaהחותכת מספר גדול של מקטעי DNA באורך קצר יחסית (כbp 35-200) במקומות אקראיים, קוראת אותם ומיצאת את תוכן הקריאות לקובץ בפורמט של הבסיסים החנקניים של המקטעים שנדגמו (לדוגמא: ATCGCA…).

הIllumina עובדת במתודולוגיה הנקראת NGS (ראה מסגרת להרחבה).

**תיאור פורמאלי**:

בהינתן:

– רצף של תווים באורך .

– אוסף של תווים בעלי אורך , .

– המחרוזת על גבי באורך , המתחילה באינדקס .

– מרחק העריכה המותר בהשוואת 2 מחרוזות, כך שהן יחשבו עדיין "דומות".

אזי עלינו למצוא עבור כל את אוסף כל ה הנמצאים במרחק עריכה ממנו.

**בפרויקט שלנו**:

– רצף הגנום האנושי, באורך של ~.

– אוסף של מקטעי ה DNA שנדגמו ע"י הIllumina, כל דגימה היא באורך של ,

.

כלומר סה"כ הדגימות מכסים את כל הגנום כ30 פעמים (בלי לקחת בחשבון חפיפות).

– פרמטר למרחק עריכה מותר – באופן טיפוסי - 2.

נתאר מעט את תהליך ההשוואה של מחרוזות שלא נפלה בהן טעות כדי לעמוד על הקושי הכרוך ביישום מעשי של אלגוריתם שכזה:

1. **האלגוריתם הנאיבי**:

עבור כל נבדוק עבור כל האם

יעילות האלגוריתם: .

מכיוון ש גדול מאוד, אלגוריתם זה אינו מעשי במקרה שלנו.

1. **KMP**

אלגוריתם המנצל את מבנה התבנית על מנת ליעל את החיפוש.

יעילות האלגוריתם:

למרות שאלגוריתם זה הוא יעיל למדי, עקב העובדה ש שלנו גדול מאוד, ומפני שיש צורך לבצע אלגוריתם זה עבור כל אחת מהקריאות (סדר גודל של כמה מליארדים), גם הוא לא טוב מספיק.

כפי שציינו, באלגוריתמים שסקרנו הנחנו שבמחרוזת שמחפשים לא נפלו טעויות. אם נניח שייתכן שנפלו טעויות במחרוזת שמחפשים, זמן הריצה בפועל גדל משמעותית.

נעמוד על הקושי שבהשוואה שכזו על ידי ניתוח של אלגוריתם השוואה נאיבי של קריאה (מחרוזת) שיתכן שנפלו בה 0-2 שגיאות במקום לא ידוע:

נניח שאורך של קריאה הוא 100.

Table 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| המקרה | מספר השמחרוזות להשוואה |  | מיקום השגיאות | שגיאות אפשריות |
| אם לא נפלה אף שגיאה | ישנה מחרוזת 1 להשוואה |  |  |  |
| אם נפלה שגיאה אחת | ישנן 300 מחרוזות להשוואה |  |  |  |
| אם נפלו 2 שגיאות | ישנם 44,550מחרוזות להשוואה |  |  |  |

סה"כ, בהנחה של עד 2 שגיאות, עברנו מקריאה אחת באורך 100 ל44,550 מחרוזות באורך 100 שנצטרך להשוות. מכיוון שמלכתחילה יש לנו כ3,000,000,000 מחרוזות כאלו, ברור שחיפוש שכזה אינו ישים עבור מידע מסדר גודל של הגנום האנושי.

**הערה:** בתיאור זה הנחנו, לשם הפשטות, שטעויות באות לידי ביטוי בהחלפת אות אחת באות אחרת בעוד שלמעשה יתכנו טעויות של החלפת מיקומים של אותיות \ קטעים, וכן טעויות של השמטת / הוספת אות.

על מנת להתגבר על בעיות כגון אלו, בשלב הדגימה לוקחים המון דגימות – בכמות כזו שסטטיסיטית כל מקטע של הDNA נדגם מספר פעמים. דבר זה עוזר כדי לוודא שאכן כל הDNA נדגם (בסבירות גבוהה) וגם מחפה על טעויות בקריאה (לא סביר שתהיה טעות קריאה של המכונה באותו המקום בכל הדגימות).

# תיאור הפתרון

על מנת להתגבר על בעית זמן הריצה, פותח האלגוריתם BWA.

בבסיסו, האלגוריתם מחולק לשלושה שלבים:

1. *Index* - אינדוקס של הגנום – שמירת הגנום בצורה שניתן לחפש עליו מחרוזות בצורה יעילה. זו פעולה שיש לעשות אותה פעם אחת בלבד (ולא בכל בדיקה של חולה...). האינדקס שמור כעץ סיפות (ראה הרחבה בהמשך).
2. *Alignment* – מציאת המיקום של הקריאות על פני הגנום.
3. *Pairing* – כחלק מקריאת הDNA של החולה, הדגימות נחתכות ל- 2 ויש צורך למצוא התאמה בין 2 חלקי הדגימה.

שלב ה *Alignment* (2), שבו יתמקד פרויקט זה, מיועל ע"י ישום של "עץ סיפות" השומר את כל הסיומות האפשריות של הגנום (זהו האינדקס משלב 1), ומימוש של אלגוריתם המחפש ב"צורה חכמה" עבור כל קריאה לאיזה סיומת היא מתאימה בהינתן מספר טעויות כלשהו (זהו פרמטר הניתן לשינוי).

יעילות האלגוריתם: (לא תלוי באורך הגנום!).

בפועל, יש לאלגוריתם זה עלויות נוספות:

* Pre Processing :
* את הפעולה מבצעים על מספר גדול מאוד של קריאות, .

ולכן, סה"כ יעילות הלגוריתם היא

כיום, למרות השימוש באלגוריתם זה, התהליך לוקח זמן רב (כיום שלם) עקב ריבוי הקריאות ().

כאמור, אלגוריתם זה ודומיו הם המובילים בעולם ביו אינפורמטיקה בתחום התאמת קריאות על פני הגנום, וכבר נמצא בשימוש ע"י בתי החולים ששוכרים חוות שרתים לביצוע חישוב זה, ועדיין התהליך לוקח כיום שלם.

בפרויקט זה נחקור את יעול האלגוריתם של BWA-Alignment באמצעות מיקבולו על גבי מעבד מרובה ליבות, כך ש הפעולות לא יתבצעו באופן סדרתי.

לשם הבנת יתרון המקבול בצורה פשוטה, אם היה ברשותנו כלי למקבול בקנה מידה אינסופי (או לפחות גדול מ).

אזי, עבור כל בדיקת DNA הזמן שידרש לריצת האלגוריתם היה .

מכייון שאין כלי שכזה בנמצא, אזי עבור "פקטור מקבול" , זמן הריצה בפועל יהיה:

במעבד מרובע ליבות i7, בו השתמשנו, פקטור המקבול הוא מעט יותר מ4 בזכות טכנולוגיית hyper-threading.

נדגיש כי ניתן ליישם את המחקר שהמתבצע בפרויקט זה גם במקבול על גבי כרטיסי מסך, שבהם פקטור המקבול יכול להגיע לכמה אלפים בודדים.

## המחשה של שלבי האלגוריתם

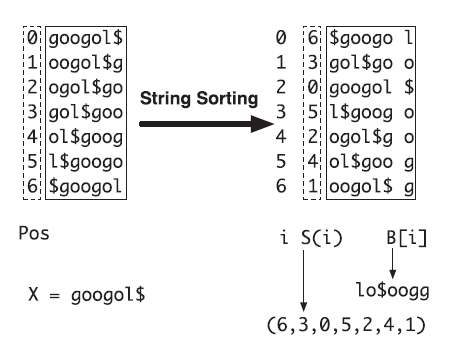
בניית האינדקס בעזרת מערך סייפות:

Figure 1 - בניית האינדקס - עץ סייפות

יהי X = googol$

נבצע הזזה מחזורית של X, ונשמור תוצאה של כל הזזה כרשומה.

לאחר מכן נמיין את הרשומות מיון לקסוגרפי.

לאחר המיון, אוסף התווים הראשונים מכל רשומה מהווים את מערך הסיומת S(i) (6,3,0,5,2,4,1)

השרשור של התוים האחרונים של הרשומות נותן לנו את "מחרוזות BWT" b[i]- lo$oogg.

וזהו למעשה האינדקס שנשתמש בו.

(ניתן לעיין בנוגע ל טרנספורם BWT – זהו נושא שלם בפני עצמו).

**נמחיש את צורת העבודה של האלגוריתם באמצעות עץ רישות.**

**הערה**: מבלי להוכיח זאת, עץ רישות שקול למערך סיפות (לצורך אינטואיציה – עץ הרישות של זהה לעץ הסיפות של , ולכן כל הדגמה על עץ רישות נכונה גם עבור עץ סיפות.

להלן **עץ רישות** של המחרוזת "Googol". הסמל ∧ מסמן את תחילתה של המחרוזת.

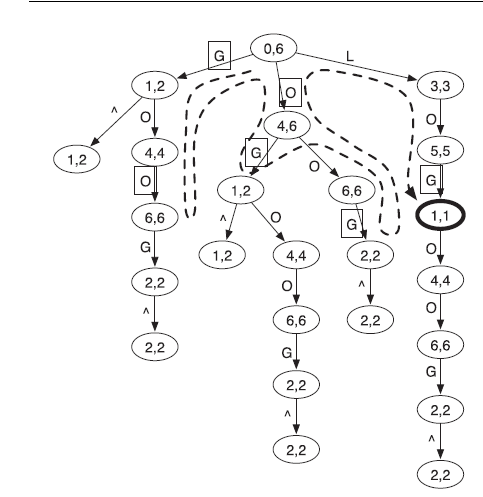
****שני המספרים בכל צומת הם האינטרוול ב SA (מערך הסיפות) של המחרוזת המיוצגת על ידי

Figure 2 - חיפוש על גבי עץ רישות

הצומת.

החלק המעניין בתרשים זה הוא הקו המקווקו המראה את המסלול של האלגוריתם בחיפוש אחר המחרוזת 'LOL', תוך אפשור חוסר התאמה אחד. נשים לב **לחידוש הגדול שבאלגוריתם** זה כפי שהוא מצוין בנקודות הבאות:

1. הקו לא יורד לכל עומק העץ – האלגוריתם יודע להתמודד עם שגיאות, **וממשיך לרדת** במורד העץ גם לאחר שגיאה אחת. ומאידך, ברגע שישנן יותר מדי שגיאות (2 במקרה הזה) **החיפוש נעצר** והקו עולה בחזרה במעלה העץ.
2. רישות שלהן יש רישה משותפת, **נמצאים על אותו מסלול של האלגוריתם!** (חסכון בזמן ריצה).

האותיות על הצלעות המוקפות ריבוע,מסמנות חוסר התאמה ("שגיאה") לשאילתא בחיפוש. ההתאמה היחידה לחיפוש היא הצומת המודגשת [1,1] המייצגת את המחרוזת 'GOL'.

**הסבר**: מדוע הצומת המודגשת [1,1] המייצגת את המחרוזת 'GOL'?

[1,1] מייצג טווח של אינדקסים, ובמקרה שלנו – אינקדס יחיד: 1.

נלך לטבלת הBWT, ואכן באינדקס 1, מופיעה המחרוזת 'GOL'.

## תיכון

ביישום הפרויקט נשתמש בשפות הבאות:

1. C# - לבניית אב טיפוס, הוכחת התכנות, בדיקות יחידה וניסויי שיפור באלגוריתם, השוואת זמני ריצה.

במהלך הפיתוח נשתמש במערכות ההפעלה הבאות:

1. Windows – לבניית מוצר עם ממשק גרפי להדגמת זמני הריצה של האלגוריתם.

פלטפורמות אלו נבחרו בעיקר על מנת לאפשר פיתוח מהיר יותר תוך שימוש בכלים קיימים המוכרים למפתח.

### שימוש ברכיבים קיימים

* הפרויקט מבוסס על יעול של אלגוריתם קיים. לכן, אנו השתמשנו במודל המבוסס על המסמך המדעי של BWA[[1]](#endnote-1).
* עבור כלי בדיקה, נשתמש בnunit.

# תיאור המערכת

## מהי המערכת

המערכת היא תכנה שמבצעת חיפוש של מחרוזת קטנה בתוך מחרוזת ארוכה (מאוד), כאשר החיפוש יניב תוצאות גם אם נפלו שגיאות במחרוזת הקצרה, כך שהיא לא זהה במדויק למקטע במחרוזת האורכה.

ערב

## תהליכים ונתוני המערכת

במערכת ישנם שלושה תהליכים:

1. Index - אינדוקס של הגנום – שמירת הגנום בצורה שניתן לגשת אליו בצורה יעילה. זו פעולה שיש לעשות אותה פעם אחת בלבד (ולא בכל בדיקה של חולה...).
2. Alignment – מציאת המיקום של הקריאות על פני הגנום.
3. Pairing – כחלק מקריאת הDNA של החולה, הדגימות נחתכות ל- 2 ויש צורך למצוא התאמה בין 2 חלקי הדגימה.

בפרויקט זה מימשנו את שלבי הindex וה Alignment.

המידע שקיים במערכת הוא:

1. הגנום האנושי המאונדקס (מידע סטטי).
2. קטעי דגימות DNA (הקלט של המערכת).

תיאור הכלים המשמשים לפתרון

ע"מ לייעל את מימוש האלגוריתם BWA השתמשנו במעבד מרובה ליבות (כולל טכנולוגית Hyper threading) בסביבת עבודה של Visual Studio, וספציפית נשתמש בספריית Parallel של מייקרוסופט לשם מקבול מיטבי.

## תיאור המימוש

(במאגר הקוד זמינים:[**קבצים בינארים**](https://github.com/turner11/BWA-Final_Project/blob/master/Files/BWA-Aligner.zip?raw=true)**[[2]](#endnote-2)** [**וקוד**](https://github.com/turner11/BWA-Final_Project/tree/master/Code)**[[3]](#endnote-3)**):

המערכת מכילה את הרכיבים הבאים:

1. ביצוע טרנספורם BWT (חלק מתהליך האינדוקס)
2. הכנת אינדקס של טקסט המדמה גנום אנושי (מבוסס על עץ סיפות).
3. חיפוש של מחרוזת בודדת באינדקס.
4. יצירת מדגם דגימות המדמות דגימות מחולה, כולל פרמטרים כגון:
   * הסתברות לטעות בכל תו.
   * אורך הדגימה.
   * מספר הדגימות.
5. חיפוש של כל אחת מהדימות מסעיף (d) באינדקס בחיפוש ממוקבל / סידרתי.
6. ביצוע חיפושים עוקבים ממוקבלים / סדרתיים עם אורך דגימה משתנה ושרטוט גרף תוצאות.
7. ביצוע חיפושים עוקבים ממוקבלים / סדרתיים עם מספר דגימות משתנה ושרטוט גרף תוצאות.
8. ביצוע חיפושים עוקבים ממוקבלים עם נתונים ממוינים / לא ממוינים ועם מספר דגימות משתנה ושרטוט גרף תוצאות.

החלק העיקרי של אב הטיפוס מבצע את החיפוש של הדגימות מסעיף (d) פעמיים, פעם בתהליכון אחד, ופעם עם ריבוי תהליכונים (מספר התהליכונים נקבע באופן אוטומטי ע"י ספריית .Net).

לשם השוואה, על מחשב עם הנתונים הבאים:

* זכרון:
  + 8 GB DDR3 SDRAM
* מעבד:
  + 2nd Gen Intel(r) Core™ i7-2670QM2
  + 2.2 GHz processor speed
  + Turbo Boost up to 3.1GHz
  + 8-way processing

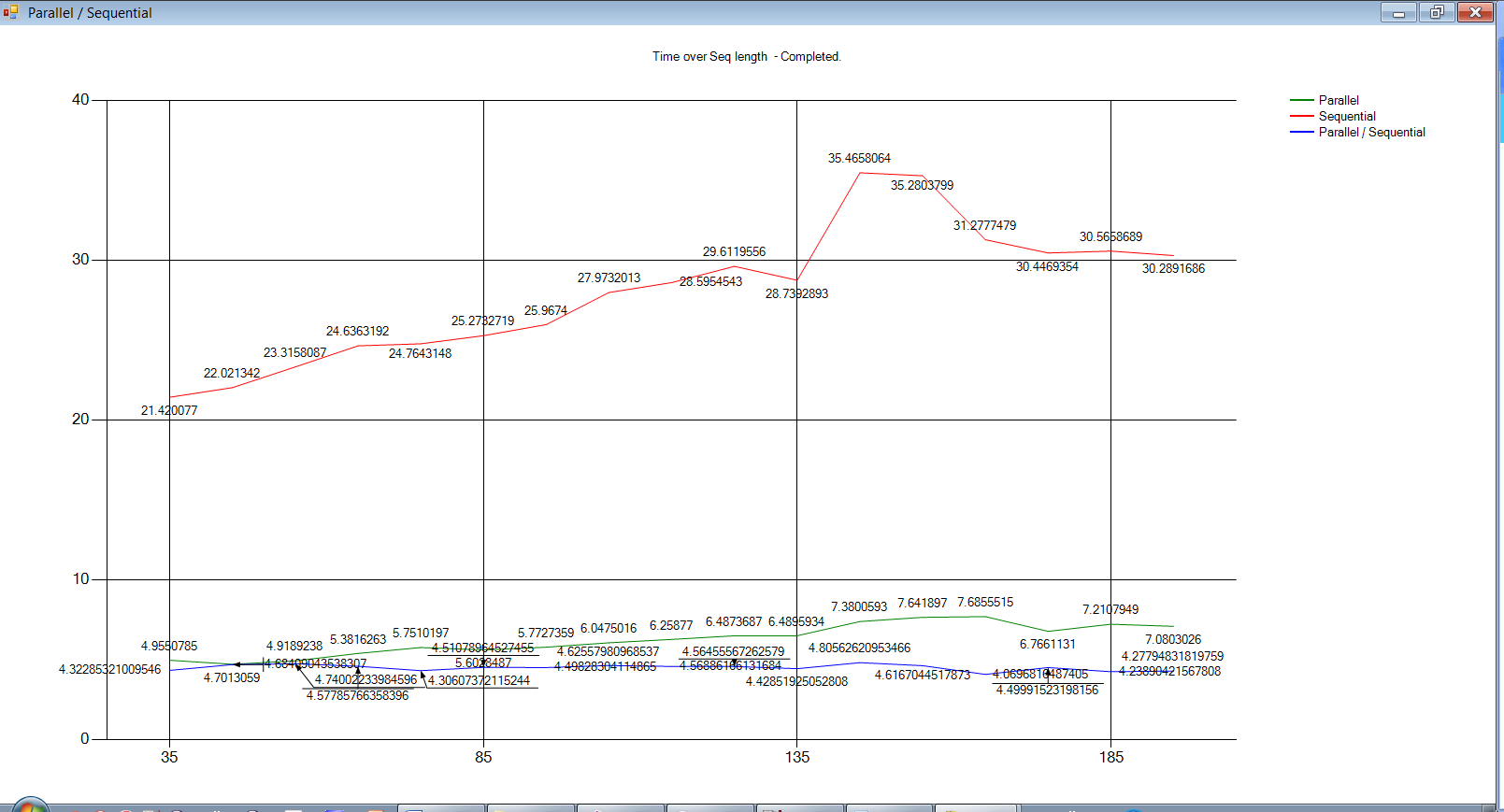
עבור קלט של:

* גנום באורך 2000 (זה לא אמור לשנות לזמן ריצה שתלוי אורך הקריאה...)
* אפשור ל 2 טעויות.
* סיכוי של 3% לטעות בכל תו בקריאה.
* קיבלנו את התוצאות הבאות:

ציר X: אורך הדגימה

ציר Y: סה"כ שניות שנדרשו להשוואת 100 דגימות (ממוקבל / לא ממוקבל)

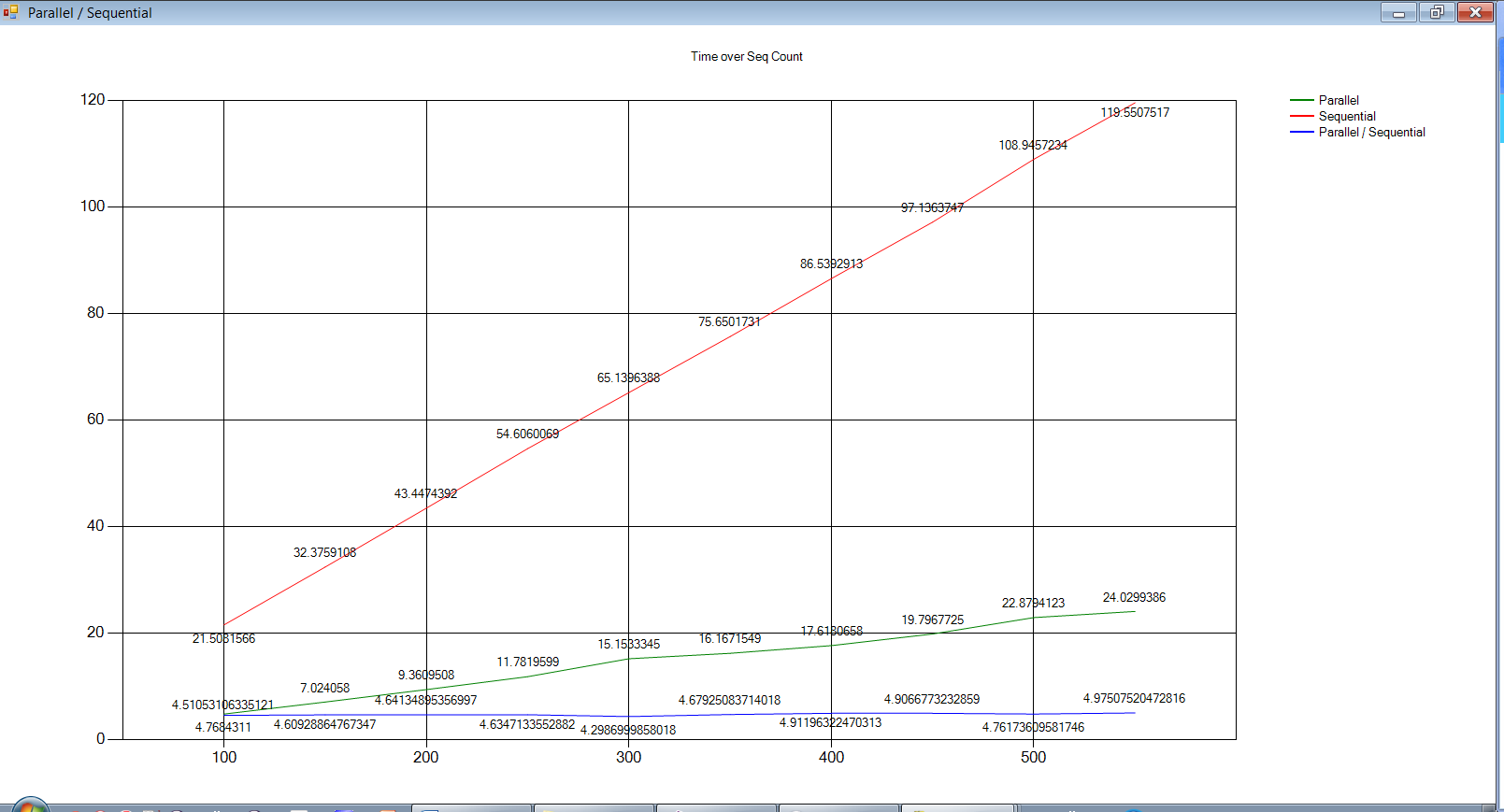
Chart 1



ציר X: מספר הדגימות

ציר Y: סה"כ שניות שנדרשו להשוואת X דגימות באורך 35bp (ממוקבל / לא ממוקבל)

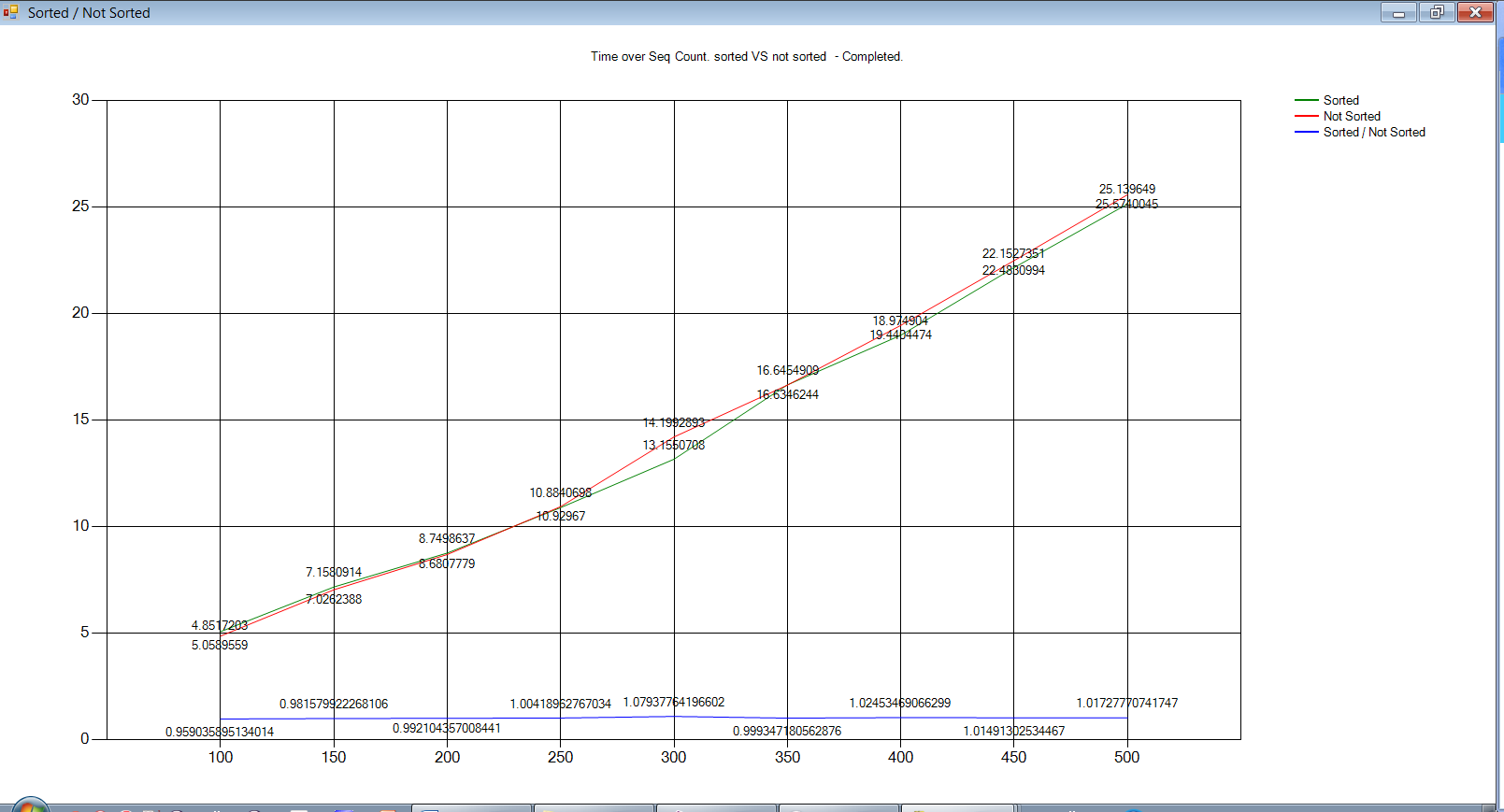
Chart 2



ציר X: מספר הדגימות

ציר Y: סה"כ שניות שנדרשו להשוואת X דגימות באורך 35bp בתהליך ממוקבל (ממויין / לא ממויין)

Chart 3



# בדיקות

תכנון וכתיבת בדיקות הן חלק בלתי נפרד מכתיבת תוכנה. הן עוזרות לנו לגלות תקלות בשלב מוקדם ומאפשרות עדכון בסיס קוד קיים תוך סיכון מינימלי בפגיעה במערכת.

בתכנון הבדיקות ניסיתי להקיף את הנקודות שעלולות לגרום לבאגים בקוד, ולוודא שאחרי המקבול:

* + - 1. הפלט של הפונקציה עם משתנה בעקבות המקבול.
      2. זמן הריצה משתפר משמעותית (מטרת הפרויקט).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | סוג הבדיקה | שם הבדיקה | אופן הבדיקה | פרמטרים | תוצאה מצופה |
| 1 | יחידה | זיהוי מחרוזת זהה | נקח קטע באורך bp35 מתוך הגנום עצמו וניתן לאלגוריתם לחפש אותו. | קטע באורך bp35, 0 טעויות מותרות. | Match באינדקס שממנו נלקח הקטע |
| 2 | יחידה | אי זיהוי מחרוזת עם n טעויות של תווים שהתחלפו. | נקח קטע באורך bp35 מתוך הגנום עצמו, נשנה n תווים בקטע, וניתן לאגוריתם לחפש אותו. | קטע באורך bp35 n-1 טעויות מותרות. | אין Match באינדקס שממנו נלקח הקטע |
| 3 | יחידה | אי זיהוי מחרוזת עם n טעויות של תווים שנמחקו. | נקח קטע באורך bp35 מתוך הגנום עצמו, נמחק n תווים בקטע, וניתן לאגוריתם לחפש אותו. | קטע באורך bp35 n-1 טעויות מותרות. | אין Match באינדקס שממנו נלקח הקטע |
| 4 | יחידה | זיהוי של מחרוזת עם n טעויות של תווים שהתחלפו. | נקח קטע באורך bp35 מתוך הגנום עצמו, נשנה n תווים בקטע, וניתן לאגוריתם לחפש אותו. | קטע באורך bp35 n טעויות מותרות. | Match באינדקס שממנו נלקח הקטע |
| 5 | יחידה | זיהוי של מחרוזת עם n טעויות של תווים שנמחקו. | נקח קטע באורך bp35 מתוך הגנום עצמו, נמחק n תווים בקטע, וניתן לאגוריתם לחפש אותו. | קטע באורך bp35 n טעויות מותרות. | Match באינדקס שממנו נלקח הקטע |
| 6 | אינטגרציה | השוואת פלט פונקציה ממוקבלת לפונקציה לא ממוקבלת | בהינתן גנום וסט של קריאות נבדוק שעבור אותם הפרמטרים הפונקציה הממוקבלת והלא ממוקבלת מוציאות את אותה התוצאה. |  | אותן תוצאות ב2 ההרצות. |
| 7 | קבלה | השוואת זמני ריצה של פונקציה ממוקבלת לפונקציה לא ממוקבלת | בהינתן גנום, פרמטרים, וסט של קריאות נבדוק שעבור שזמני הריצה של הפונקציה הממוקבלת טובה באופן משמעותי מהפונקציה הלא ממוקבלת. |  | אותן תוצאות ב2 ההרצות. עם זמן ריצה טוב יותר באופן משמעותי עם המקבול (משמעותי = פי 3) |

* הבדיקות מתוארות במונחים כללים של "n" טעויות בגלל הקלות של שכפול בדיקות עם פרמטרים שונים באמצעות הספרייה nunit ושימוש ב TestCase.
* כל אחת מהבדיקות הנ"ל מייצגת למעשה מספר בדיקות שונות:
  + עבור קטע מתחילת הגנום.
  + עבור קטע מאמצע הגנום.
  + עבור קטע מסוף הגנום.

באותו אופן קל מקטע טומן בחובו מספר בדיקות :

* עבור טעויות שנפלו בתחילת הקטע
* עבור טעויות שנפלו באמצע הקטע
* עבור טעויות שנפלו בסוף הקטע

ושוב, כל אחד מאלו מחזיק בתוכו בדיקות עבור:

* אות שהוחלפה.
* אות שהוחסרה

בסופו של דבר, עם כל ה Test Cases קיימות 66 בדיקות למערכת.

# מסקנות

גורמים שלא משפיעים על זמן הריצה של האלגוריתם הממוקבל ביחס לאלגוריתם הסדרתי:

1. אורך הדגימות.
2. מיון / אי מיון הדגימות.

גורמים המשפיעים על זמן הריצה:

1. לאחר בדיקה והשוואות נראה שהתוצאות הטובות ביותר מתקבלות ע"י מקבול התהליך, באופן שבו כל דגימת DNA מקבלת תהליכון (בניגוד למקבול פנימי של התהליך כך שכל תהליכון יתמקד בתת רקורסיה של האלגוריתם).

לסיכום נראה שהאלגוריתם הממוקבל נשאר יציב, מבחינת סדרי הגודל של היעול, גם כאשר מעמיסים עליו דגימות ארוכות.

מאידך, נראה שמיון הדגימות, לא עוזר לשפר את זמן הריצה.

במחשב מרובע ליבות i7 קיבלנו זמן ריצה טוב פי 4-5 בעזרת האלגוריתם הממוקבל (הערה: זמן ריצה טוב ביותר מפי 4 על מחשב עם 4 ליבות התאפשר בזכות טכנולוגיית Hyper Threading הממומשת בחומרה של מעבדי i7)

נציין, שאם האלגוריתם ימוקבל ע"ג כרטיס מסך, המיון צפוי לעזור באופן משמעותי בעקבות הורדה דרסטית של הLock Step (סנכרון תהליכונים) המאפיין תכנות מקבילי שכזה.

קיווינו למקבול זה במסגרת הפרויקט, אך החומרה הנדרשת סופקה רק ביום מועד ההגשה.

# השוואה לפתרונות בספרות

בנוסף לאלגוריתמים שתוארו בסעיף "תיאור הבעיה" קיימים הפרויקטים הבאים:

* + - 1. פרויקט הקוד הפתוח [BarraCUDA](http://seqbarracuda.sourceforge.net/)[[4]](#endnote-4), המיישם את אלגוריתם BWA בעזרת תכנות מקבילי על גבי כרטיס מסך. הפרויקט בתהליכי פיתוח כ 6 שנים, כחלק מהמעבר לתכנות על גבי כרטיס מסך, הוסרה האפשרות לחפש מחרוזות עם תווים שהוסרו.
      2. פרויקט [BWA](http://sourceforge.net/projects/bio-bwa/?source=navbar) – הפרויקט המוביל בתחום ביואינפורמטיקה שמימש את שלגוריתם הBWA – עם תאימות ללינוקס בלבד. זמני הריצה של BWA טובים משלנו, אך הקוד נכתב בצורה של תיקון על גבי תיקון ויעול על גבי יעול במהלך שנים, וללא תיעוד – כך שקשה מאוד להכנס אליו.
      3. [BowTie](http://bowtie-bio.sourceforge.net/index.shtml)[[5]](#footnote-1) – אלגוריתם מקביל המתבסס על אותם עקרונות כמו BWA

**הערה:** בגלל אופיו המחקרי של הפרויקט, נפח העבודה במונחים של כמות קוד היא קטנה יחסית ולכן הן Waterfall והן agile לא נמצאו מתאימים כמתודלוגיית עבודה. צווארי הבקבוק שאובחנו הם:

1. הכנת אבטיפוס.
2. כתיבת בדיקות מקיפות.
3. מקבול הקוד.

## טבלת סיכונים

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | הסיכון | חומרה | סבירות | מדד סיכון (סבירות X חומרה) | מענה אפשרי |
| 1 | המיקבול יתברר כלא אפשרי. | 5 | 1 | 5 | ננסה ליעל את האלגוריתם עצמו |
| 2 | המיקבול יהיה פחות יעיל ממקבול קיים. | 3 | 3 | 9 | נמקבל מה שאפשרי, וננסה ליעל את האלגוריתם. |
| 3 | חוסר הבנה של האלגוריתם. | 3 | 3 | 9 | נבצע בדיקות של זמן ריצה ותוצאות בסיוע של בדיקות יחידה לוודא נכונות גם בלי הבנה מעמיקה. אם יהיה צורך, נממש את הפונקציה בצורה פחות יעילה – אך הפיצוי לכך יהיה במיקבול. |
| 4 | מבני הנתונים שבשימוש הפונקציה הממוקבלת גדולים מדי בשביל המקבול על גבי כרטיס מסך | 5 | 2 | 10 | נכנס לעובי הקורה ונראה היכן אפשר להשמיט / לצמצם אוביקיטים (לדוגמא ע"י השמטת שדות שלא בשימוש |
|  |  |  |  |  |  |

# רשימת ספרות

1. להכנת מסמך זה נעזרתי ב:
   1. בויקיפדיה:
      1. Cuda: <http://he.wikipedia.org/wiki/CUDA>
      2. Short-Read Sequence Alignment: [http://en.wikipedia.org/wiki/List\_of\_sequence\_alignment\_software#Short-Read\_Sequence\_Alignment](http://en.wikipedia.org/wiki/List_of_sequence_alignment_software" \l "Short-Read_Sequence_Alignment)
   2. האיציקלופדיה הקטנה של המדעים מאת רן לוי (מהדורה ראשונה 2011).
   3. ספרות עם חומר שרלוונטי לפרויקט:
      1. הרצאות בביואינפורמטיקה פרקים 17-21. <https://www.cs.cmu.edu/~ckingsf/bioinfo-lectures/>.
      2. שאלות נפוצות בקשר ל BWA: <http://bio-bwa.sourceforge.net/>
      3. מסמך המתאר את צורת העבודה של האלגוריתם: <http://www.math.pku.edu.cn/teachers/xirb/Courses/biostatistics2013/Bioinformatics-2009-Li-1754-60.pdf>.
      4. השוואה בין שיטות NGS: [http://en.wikipedia.org/wiki/DNA\_sequencing#Next-generation\_methods](http://en.wikipedia.org/wiki/DNA_sequencing" \l "Next-generation_methods)

# נספחים

**יומן הפרויקט:** <http://projects.jce.ac.il/moodle/mod/wiki/view.php?wid=1&title=%D7%99%D7%95%D7%9E%D7%9F+%D7%A4%D7%A8%D7%95%D7%99%D7%A7%D7%98&groupanduser=9-26>

**מאגר קוד:** <https://github.com/turner11/BWA-Final_Project>

**רשימת משימות פתוחות:** [**https://github.com/turner11/BWA-Final\_Project/issues?q=is%3Aopen+is%3Aissue**](https://github.com/turner11/BWA-Final_Project/issues?q=is%3Aopen+is%3Aissue)

**רשימת משימות סגורות:**  <https://github.com/turner11/BWA-Final_Project/issues?q=is%3Aissue+is%3Aclosedhttps://github.com/turner11/BWA-Final_Project/issues?q=is%3Aissue+is%3Aclosed>

# Abstract

This is a research project. In order to have a better grasp of the issue that this project is facing and better understand the suggested solution we will start with presenting some background:

**DNA**: The DNA is a molecule that exists in each and every of the cells in our body, and contains most of the [genetic](https://en.wikipedia.org/wiki/Genetics) instructions used in the development and functioning of all known living [organisms](https://en.wikipedia.org/wiki/Organism), from the simplest bacteria to human bei

The DNA consists of two strands coiled around each other. Each of the strands is composed from a sequence of Nitrogenous bases (labeled as A, T, G & C).

The two strands are connected to each other in a common manner: "A" and "T" will always be connected, and "C" and "G" will always be connected. This feature allows us to reconstruct a strand, given the other strand it was connected to. Every connected pare is referred to as a "base pair"

The surprising fact about DNA is that 99.9% human DNA is common to all human beings. And this is a feature this project heavily relies on.

Figure 1 - The DNA consists of two strands coiled around each other

The data encapsulated within the genome can be represented as string of the bases it is composed of, and in this project we will treat it as such (e.g. TGACCGTCAG...).

The DNA is composed of – digitally wise it is equivalent to a little more the 1.6 GB (or about to CDs…).

Another worth mentioning "feature" of the DNA is that it could mutate. The vast majority of mutations are harmless but is the mutation is located in certain location in the DNA it might cause genetic diseases, and cancer in particular.

These days, when a patient is getting diagnosed for cancer the review process is long and cumbersome:

Figure 2 – A connects to T and G connects to C

1. Sampling the patient DNA.
2. Comparing the samples to a DNA from a healthy source.
3. Locating mutations.
4. Manually searching in medicine databases for information about the mutations and their correlation with cancer.
5. Manually searching for an existing drug that is affective with the specific diagnosed type of cancer.

Among other challenges in the above process the comparing (#2) stands out as a time consuming step. In order to accomplish this task at a feasible time line, hospitals rent server farms for doing the comparisons. And still, this process takes about a full day.

This project will focus on the optimization of this step.

For the optimization we have tested and benchmarked few items:

1. The impact of length of samples on calculation runs time.
2. The impact of number of samples on calculation runs time.
3. The impact of sorting the samples on calculation runs time.
4. The impact of making the process parallel on multiple cores.

After testing and benchmarking, it seems we get best results by making the process parallel in a manner that every sampled DNA sequence is being aligned using a separate thread (i.e. rather than having the alignment algorithm aligned internally).

This type of optimization stays stable with no dependency on samples length/ number / sort.

1. <http://sourceforge.net/projects/bio-bwa/?source=navbar> [↑](#endnote-ref-1)
2. <https://github.com/turner11/BWA-Final_Project/blob/master/Files/BWA-Aligner.zip?raw=true> [↑](#endnote-ref-2)
3. <https://github.com/turner11/BWA-Final_Project/tree/master/Code> [↑](#endnote-ref-3)
4. <http://seqbarracuda.sourceforge.net/> [↑](#endnote-ref-4)
5. [http://bowtie-bio.sourceforge.net/index.shtml](%20http:/bowtie-bio.sourceforge.net/index.shtml) [↑](#footnote-ref-1)