

**Acceleration of Local Sensitivity Hashing for Genome Assembly**

**­**

**פרויקט #6392  
 סמסטר חורף 2022  
בהנחיית לאוניד יביץ**

גל אלבז

תוכן עניינים

[הקדמה 3](#_Toc93614387)

[מטרות הפרויקט 4](#_Toc93614388)

[אלגוריתם הMinHash 5](#_Toc93614389)

[השלב הראשון של האלגוריתם 5](#_Toc93614390)

[השלב השני של האלגוריתם 6](#_Toc93614391)

[השלב השלישי של האלגוריתם 8](#_Toc93614392)

[השלב הרביעי של האלגוריתם 8](#_Toc93614393)

[השלב החמישי של האלגוריתם 9](#_Toc93614394)

[הארכיטקטורה 10](#_Toc93614395)

[Top-Down Design 10](#_Toc93614396)

[Hash-Instance 13](#_Toc93614397)

[Min-Instance 14](#_Toc93614398)

[Counter-Instance 15](#_Toc93614399)

[תיאור סכמתי של כלי הסינתזה 16](#_Toc93614400)

[Top Design 16](#_Toc93614401)

[Hash Module 17](#_Toc93614402)

[Min Module 18](#_Toc93614403)

[Counter Module 18](#_Toc93614404)

[דוחות מהסינתזה 19](#_Toc93614405)

[Area Report 19](#_Toc93614406)

[Power Report 20](#_Toc93614407)

[Layout 21](#_Toc93614408)

[Adding Power 21](#_Toc93614409)

[Placing Cells 22](#_Toc93614410)

[Clock Tree Synthesis 23](#_Toc93614411)

[Amoeba View 25](#_Toc93614412)

[סימולטור תוכנתי 26](#_Toc93614413)

[סימולטור חומרתי 28](#_Toc93614414)

[סימולציות 29](#_Toc93614415)

[סימולציה מספר 1 – זהות מוחלטת של הרצפים 29](#_Toc93614416)

[סימולציה מספר 2 – שוני מוחלט של הרצפים 31](#_Toc93614417)

[סימולציה מספר 3 – רצפים כללים 1 33](#_Toc93614418)

[סימולציה מספר 4 – רצפים כללים 2 35](#_Toc93614419)

[סימולציה מספר 5 – רצפים כללים 3 37](#_Toc93614420)

[סימולציה מספר 6 – רצפים כללים 4 39](#_Toc93614421)

[סיכום ומסקנות 41](#_Toc93614422)

[רפרנס 42](#_Toc93614423)

[תודות 42](#_Toc93614424)

# הקדמה

הרכבת הגנום הוא מונח המתאר תהליך בו מתבצע חיבור של מספר רב של רצפים בכדי ליצור את הגנום הכולל.  
קיימות שתי גישות עיקריות של הרכבת הגנום:

1. Mapping assembly.
2. De novo assembly.

בפרויקט זה עסקנו בגישה מספר 2 אשר מבצעת את תהליך ההרכבה של גנום מאפס, ללא רפרנס על המידע הגנטי.  
זאת בניגוד לגישה מספר 1 אשר לפיה הרכבת הגנום מתבצעת בהתבסס על מידע גנטי ידוע מראש.

הרכבת הגנום בשיטת de novo דורשת עלות חישובית גובהה אשר מהווה צוואר בקבוק משמעותי כאשר ניגשים לבעיות הכוללות גנומים ארוכים.  
על מנת להצדיק שימוש בהרכבת גנום מבוססת de novo יש צורך בהאצה של המשימות הגוזלות זמן רב בתהליך.

בפרויקט זה, נחקור גישה הסתברותית המנסה לפתור את צוואר הבקבוק בעזרת אלגוריתם MHAP (MinHash Alignment Process), אשר משתמש בטכניקה של שימוש בפונקציות hash לצורך ייצוג קומפקטי יותר של קריאות הרצף (reads) והשוואתם על מנת לדעת האם ניתן לאחד את הreads לייצוג חופף (contigs) שישמשו ליצירת מפה פיזיקלית המשחזרת את רצף הDNA המקורי.

Diagram

Description automatically generated with low confidence

איור 1  
 תיאור גנום כולל, פירוקו לreads והרכבת הגנום מחדש על פי אזורים חופפים לcontigs.

# מטרות הפרויקט

כאמור, בפרויקט זה תכננו מאיץ חומרתי אשר יפתור את צוואר הבקבוק הנובע מהעלות החישובית של הרכבת הגנום בשיטת de novo.  
בפרויקט זה:

* למדנו על אלגוריתם הMinHash שתפקידו להאיץ את תהליך הרכבת הגנום.
* בנינו תכנון חומרתי היררכי של האלגוריתם במטרה להאיץ את חישוב הרכבת הגנום ככל הניתן.
* מימשנו סימולטור תוכנתי בPython אשר מחקה את התנהגות האלגוריתם.
* מימשנו סימולטור חומרתי בSystemVerilog אשר מחקה את התנהגות האלגוריתם.
* בצענו מספר רב של סימולציות ובדקנו התאמה בתוצאות בין שני הסימולטורים המתוארים.
* בצענו סינתזה לתכנון הלוגי שרשמנו, הסינתזה כללה קומפילציה של התכנון והפקת דוחות על המסלול הקריטי, שטח וצריכת ההספק של הרכיב.
* תכנון ומימוש מעגל VLSI.

# אלגוריתם הMinHash

האלגוריתם ידוע באופן פורמלי בשם אשר תוכנן בראשיתו לצורך זיהוי של דפי אינטרנט כפולים.  
אם נחשוב על דף אינטרנט כאחד מקריאות הרצף (reads) והמטרה שלנו היא למצוא אשר דומים זה לזה, קיים הצדקה והיגיון בשימוש באלגוריתם הנ"ל למטרה שלנו.

אלגוריתם MHAP (**MinHash Alignment Process**) מהווה משערך מהיר של זהות ג'קארד , אשר מוגדרת עבור שני סטים של רצפים בקשר:

על פי שערוך זה, אנו מחלקים את הרצף הגנטי שלנו לסט של תתי רצפים אשר מכונים כ שעליהם מתבצעים מספר פעולות מתמטיות. לבסוף, תוצאת השערוך מחושבת לפי **נוסחה מספר .**

ניתן לראות שעל פי **נוסחה מספר** , זהות ג'קארד שווה לאפס כאשר שני הסטים זרים זה לזה.  
באופן דומה, זהות ג'קארד שווה ל1 כאשר שני הסטים זהים זה לזה לחלוטין, אחרת התוצאה היא בין 0 ל1.

# השלב הראשון של האלגוריתם

בשלב הראשון של האלגוריתם, מתבצעת חלוקה של הרצף הגנטי הנתון לתתי מחרוזות, .  
מספר ה שכל רצף תורם נתון על ידיי הנוסחה:  
כאשר מתקיים:

* הוא אורך הרצף המקורי.
* הוא אורך כל תת רצף (פרמטר נתון לבחירה).

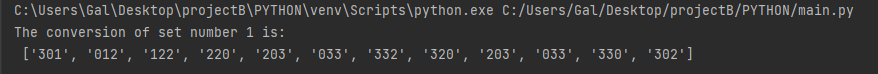
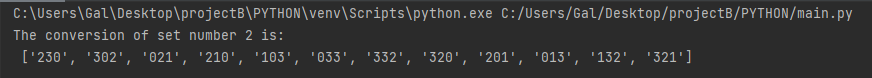
בסיום שלב זה, מקבלים סט של ה המרכיבים את הרצף המקורי.

לדוגמא, נבצע את השלב הראשון של האלגוריתם על הרצפים הגנטיים הבאים, נקבע שאורך כל הוא 3.  
עם סיום השלב הראשון, יתקבלו הסטים הבאים:  
לפי **נוסחה מספר 2**, אכן התקבלו 12 סה"כ.

# השלב השני של האלגוריתם

בשלב השני של האלגוריתם, כל עובר תרגום לחתימה אשר מיוצגת על ידיי מספר שלם.  
בפרויקט שלנו, בחרנו לבצע את התרגום הבא:  
כידוע, הנוקלאוטידים המרכיבים את חומצת הגרעין הם .

תחילה, בחרנו לייצג כל אות-אות באמצעות ספרה בודדת בבסיס עשרוני:  
לאחר מכן, בצענו המרה של ספרה-ספרה לייצוג בינארי ברוחב 2 ביט:

 בהמשך לדוגמה הקודמת, נבצע מעבר לייצוג הנ"ל עבור הסטים של ה שקיבלנו:

איור 4  
מעבר לתצוגה בינארית אות-אות עבור הסט .

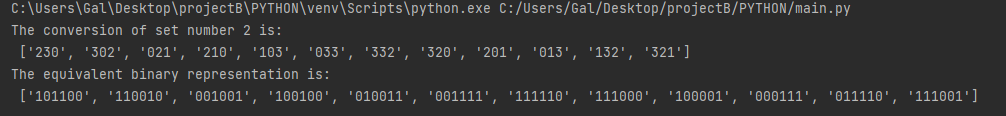
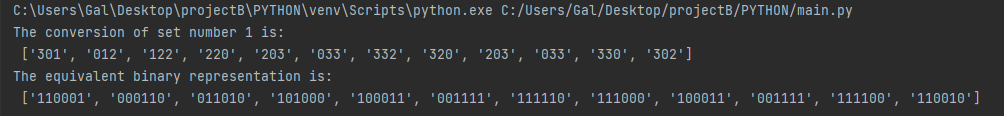
איור 4  
מעבר לתצוגה בינארית אות-אות עבור הסט .

איור 3  
פירוק הסט לייצוג אות-אות בבסיס עשרוני

איור 3  
פירוק הסט לייצוג אות-אות בבסיס עשרוני

איור 2  
פירוק הסט לייצוג אות-אות בבסיס עשרוני

איור 2  
פירוק הסט לייצוג אות-אות בבסיס עשרוני

כעת נעבור לייצוג הבינארי:

איור 5  
מעבר לתצוגה בינארית אות-אות עבור הסט .

איור 5  
מעבר לתצוגה בינארית אות-אות עבור הסט .

הסיבה שבגללה בצענו מעבר בין הייצוג הגנטי דרך שני בסיסים היא שראשית רצינו לקבל ייצוג קומפקטי וקריא של הרצף הגנטי באמצעות ספרות בבסיס עשרוני.  
לאחר מכן, בצענו המרה של **ספרה-ספרה** לייצוג בינארי וכך קיבלנו את החתימה שאיתה אנו עובדים לצורך החישובים.

נשים לב שעבור איור מספר 4, לא מתקיים השוויון הבא עבור ה הראשון:  
שכן ההמרה מתבצעת ספרה-ספרה כך שלמעשה מתקיים:  
בעזרת ייצוג זה, הצלחנו לבטא את הערך המספרי של החתימה על ידיי מספר שבהכרח קטן יותר מאשר הייצוג הספרתי-דצימלי.  
כלומר:

יתרון נוסף של אופן ייצוג זה הוא שהצלחנו לבצע מיפוי של ארוכים לחתימה שמתורגמת למספר עשרוני נמוך שניתן לייצג באמצעות פחות ביטים.

נדגים יתרון זה על הרצף הבא ובחירה של באורך 16.  
ה ראשון של רצף זה נתון על ידיי:  
נעביר לייצוג כפי שהסברנו:  
במידה ונעבוד עם ייצוג זה כ שלנו, נקבל ייצוג בינארי באורך 43 ביט!

לעומת זאת, על פי השיטה שלנו, תחילה נעבור לייצוג כפי שהסברנו:  
במידה ונעבוד עם חתימה זו בתור המספר שלנו, מתקבל כי החתימה מיוצגת בבינארי כמספר ברוחב 32 ביט בלבד.

# השלב השלישי של האלגוריתם

בשלב השלישי של האלגוריתם, מעבירים כל אחת מהחתימות שיצרנו דרך פונקציות ידועות מראש, את התוצאה המתקבלת מהפעולה:  
כאשר:  
שומרים ביחד עם אותו ה עבור השלב הרביעי באלגוריתם.

A picture containing text, receipt

Description automatically generatedלדוגמה, עבור ה של הרצפים הגנטיים שניתחנו קודם לכן ובחירה של פונקציות מסוימות, נקבל:

איור 6  
הפעלת פונקציות ה על כל .  
לקוח מתוך [המאמר](../מסמכים/חומר%20רקע/Assembling%20large%20genomes%20with%20single-molecule%20sequencing%20and%20locality-se....pdf)

איור 6  
הפעלת פונקציות ה על כל .  
לקוח מתוך [המאמר](../מסמכים/חומר%20רקע/Assembling%20large%20genomes%20with%20single-molecule%20sequencing%20and%20locality-se....pdf)

כל עמודה מתארת את תוצאת פונקציית ה המתקבלת על הפעלת הפונקציה על של כל אחד מה אשר מתוארים בשורות.

# השלב הרביעי של האלגוריתם

בשלב הרביעי של האלגוריתם נלקחת התוצאה המינימאלית מכל פונקציית , כלומר עבור כל עמודה לוקחים את הערך המינימאלי.

**Text

Description automatically generated with medium confidence**בהמשך לדוגמה המוצגת בשלבים הקודמים, נקבל את תוצאת הביניים הבאה:  
התוצאות המינימאליות מסומנות באדום עבור כל עמודה, ביחד איתם מסומן ה המתאים אשר הוביל לתוצאה המינימאלית בחישוב.

איור 7  
שמירה של הערך המינימאלי מכל עמודה, כלומר תוצאות החישוב של פונקציית ה ה.  
לקוח מתוך [המאמר](../מסמכים/חומר%20רקע/Assembling%20large%20genomes%20with%20single-molecule%20sequencing%20and%20locality-se....pdf)

איור 7  
שמירה של הערך המינימאלי מכל עמודה, כלומר תוצאות החישוב של פונקציית ה ה.  
לקוח מתוך [המאמר](../מסמכים/חומר%20רקע/Assembling%20large%20genomes%20with%20single-molecule%20sequencing%20and%20locality-se....pdf)

התוצאות המינימאליות עבור על סט נשמרות במערך חד-מימדי באופן הבא:

# השלב החמישי של האלגוריתם

השלב החמישי והאחרון באלגורית הוא חישוב זהות ג'קארד בהתבסס על תוצאות המינימום שהתקבלו מהשלב הקודם.  
על פי **נוסחה מספר 1** מתקיים כי מקדם הזהות מחושב לפי:

כלומר יש לקחת את החיתוך של הסטים ולחלק בגודל התוצאות שהתקבלו.

בדוגמה שהרצנו עד כה, נקבל כי זהות ג'קארד היא:

על פי מקדם הזהות, נקבל אינדיקציה עד כמה הרצפים הגנטיים זהים.

A picture containing chart

Description automatically generatedניתן לראות את כל השלבים אחד אחריי השני באיור הבא:

איור 8  
פעולת האלגוריתם באופן מלא  
לקוח מתוך [המאמר](../מסמכים/חומר%20רקע/Assembling%20large%20genomes%20with%20single-molecule%20sequencing%20and%20locality-se....pdf)

איור 8  
פעולת האלגוריתם באופן מלא  
לקוח מתוך [המאמר](../מסמכים/חומר%20רקע/Assembling%20large%20genomes%20with%20single-molecule%20sequencing%20and%20locality-se....pdf)

# הארכיטקטורה

מטרתנו הייתה לדמות את התנהגות האלגוריתם הנ"ל באמצעות שפת תיאור החומרה .  
התכנון שלנו כלל חלוקה של האלגוריתם למספר מודולים המסודרים בצורה היררכית כך שכל מודול מטפל בשלב שונה באלגוריתם ומסתמך על הפלט של המודול הקודם לו.

הסתמכנו על עקרונות אופטימיזציה של תכנון באמצעות לצורך האצת החישובים והשגת התוצאה הסופית במספר בודד של מחזורי שעון.

בתכנון שלנו, פעלנו על פי האפיון הבא של כניסות ודרישות המערכת:

1. אורך רצף גנטי מורכב מ64 נוקלאוטידים.
2. אורך הוא 16 נוקלאוטידים.
3. שימוש ב8 פונקציות הפועלות על פי הכלל הבא:

* הוא מקדם נרמול ושווה ל255.
* הינם מספרים רנדומליים המתאימים לפונקציית ה ה.

1. מקדם ג'קארד אינו מנורמל במספר פונקציות ה שבחרנו לייצוג.  
   כלומר, מקדם ג'קארד נע בין 0 ל8 כאשר 0 מייצג חוסר זהות ו8 מייצג זהות מוחלטת.

# Top-Down Design

נתאר את המערכת שלנו מלמעלה-למטה כך שככל שנתקדם למטה בהיררכיה נתבונן בחלקים קטנים יותר על מנת להבין את המערכת הכוללת.

התכנון ברמה העליונה ביותר מודגם על ידיי הסכימה הבאה:

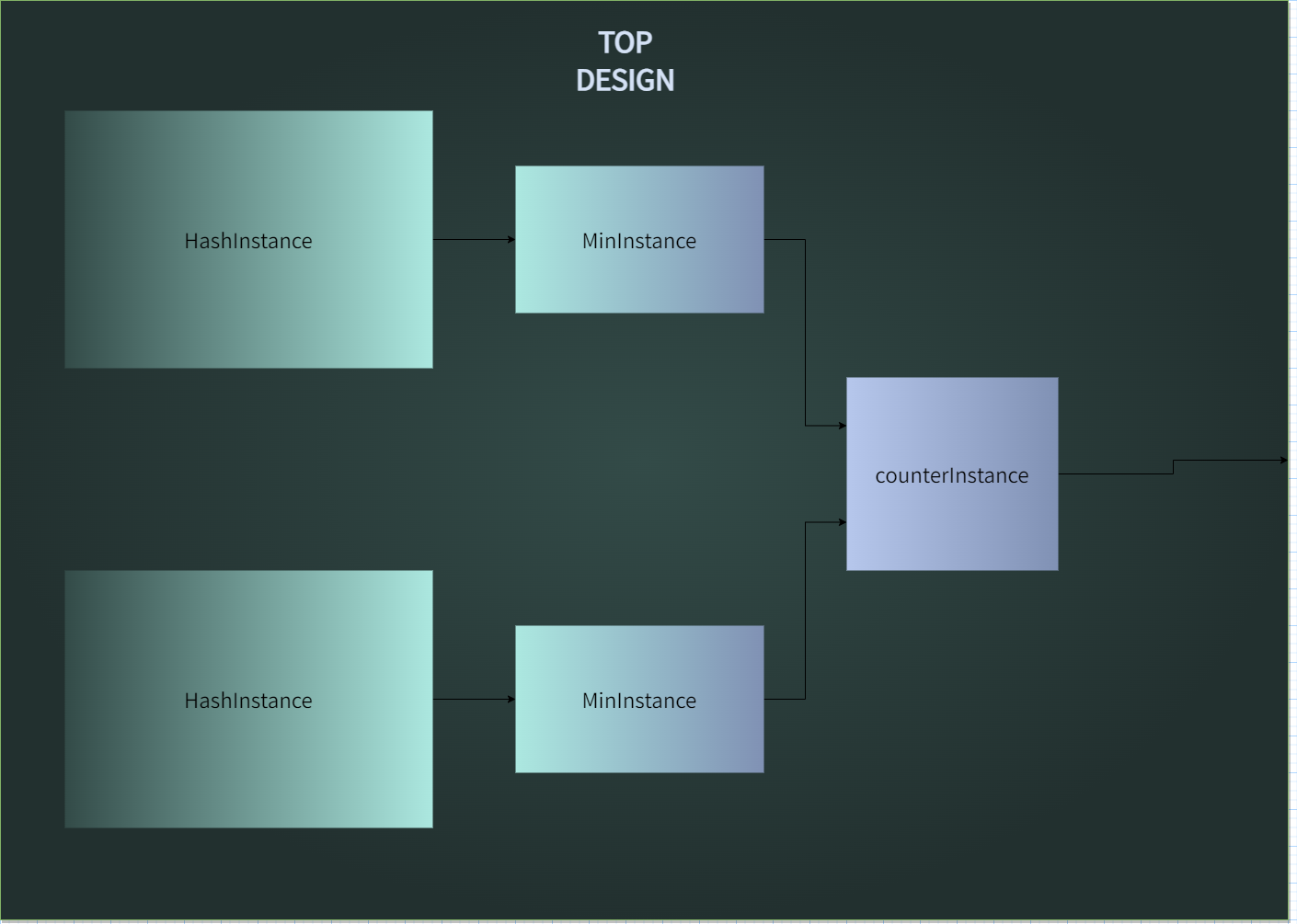
איור 9  
סכמת בלוקים מופשטת של ה.

איור 9  
סכמת בלוקים מופשטת של ה.



התכנון הנ"ל מכיל את הכניסות והיציאות של המערכת ברמת הרכיב הכולל.  
ניתן לראות שהכניסות הם:

* שעון, .
* הדק איפוס, .
* זוג מספרים רנדומליים המשמשים לחישוב פונקציות ה.
* זוג מערכים המכילים את ה של הרצפים. המערכים המכילים את ה מוחזקים קבועים לאורך כל האלגוריתם, המספרים הרנדומליים משתנים כתלות בעליית שעון ופונקציית ה הבאה שיש להריץ.

נבצע למערכת ונראה את התיאור המופשט הבא:  
  
במבט פנימי יותר ניתן לראות שהמודל הכולל מכיל סך הכל חמישה מודלים:

איור 10  
close up על סכמת בלוקים מופשטת של ה.

איור 10  
close up על סכמת בלוקים מופשטת של ה.

1. , מודל אשר מבצע את פעולת ה על כל רצף גנטי.  
   כיוון שיש לנו שני רצפים גנטיים, שכפלנו מודל זה.
2. , מודל אשר מוצא את התוצאה המינימאלית מבין התוצאות המתקבלות מפעולת ה.   
   כיוון שיש לנו שני רצפים גנטיים שמועברים דרך פונקציית ה, שכפלנו מודל זה.
3. , מודל אשר מבצע חישוב זהות בין שני התוצאות המינימאליות המתקבלות.

# Hash-Instance

נתאר את מודל ה.  
מודל זה אחראי על הפעלת פונקציית ה על ה של הרצף הגנטי, בכל עליית שעון משתנה פונקציית ה שמפעילים בכך שהמשתנים מתחלפים.  
כזכור על פי האפיון שלנו, פונקציית ה שלנו מתוארת על פי הכלל הבא:

תיאור הכניסות והיציאות של המודל באופן מופשט נתון על ידיי:

איור 11  
סכמת בלוקים מופשטת של מודל ה

איור 11  
סכמת בלוקים מופשטת של מודל ה

ניתן לראות שהכניסות הן:

* שעון, .
* הדק איפוס, .
* זוג מספרים רנדומליים המשמשים לחישוב פונקציות ה.
* מערך המכיל את ה של הרצף. המערך מכיל את ה כקבועים לאורך כל האלגוריתם.

היציאה היא:

* מערך התוצאות, .

בכל עליית שעון הערכים של משתנים על מנת לבצע את החישוב על החתימות עם פונקציית ה הבא.  
סה"כ עבור 8 פונקציות שונות יש צורך ב8 עליות שעון.

# Min-Instance

נתאר את מודל ה.  
מודל זה אחראי על קבלת תוצאות החישוב מפונקציית ה ומציאת הערך המינימאלי.  
נשים לב שמודל זה יכול לפעול כראוי רק בעליית השעון השנייה, בה מוכנות תוצאות החישוב של הפעלת פונקציית על ה ועל כן יש להוסיף למודל זה השהיה.

תיאור הכניסות והיציאות של המודל באופן מופשט נתון על ידיי:

ניתן לראות שהכניסות הן:

איור 12  
סכמת בלוקים מופשטת של מודל MinInstance

איור 12  
סכמת בלוקים מופשטת של מודל MinInstance

* שעון, .
* הדק איפוס, .
* מערך התוצאות, .

היציאה היא:

* הערך המינימאלי מבין מערך התוצאות, .

בכל עליית שעון, החל מעליית השעון בפעם השנייה בה המידע תקף, מערך התוצאות מכיל את תוצאות החישוב של הפעלת פונקציית ה על כל החתימות.  
על ידיי מעבר בלולאה על ערכי המערך, מוצאים את המינימום ומעבירים אותו לשלב הבא בצינור.

# Counter-Instance

נתאר את מודל ה.  
מודל זה אחראי על חישוב מקדם ג'קארד כך שבכל עליית שעון, החל מהעלייה השלישית בו המידע תקף בכניסת המודל, מתרחשת בדיקת זהות בין ערכי המינימום מהשלב הקודם עבור הרצפים .  
גם במודל זה הוספנו השהיה על מנת לדגום מידע תקף בלבד ולשמור על נכונות המערכת.

תיאור הכניסות והיציאות של המודל באופן מופשט נתון על ידיי:

 ניתן לראות שהכניסות הן:

איור 13  
סכמת בלוקים מופשטת של מודל CounterInstance

איור 13  
סכמת בלוקים מופשטת של מודל CounterInstance

* שעון, .
* הדק איפוס, .
* זוג הערכים המינימליים, .
* מקדם ג'קארד שחושב עד כה.

היציאה היא:

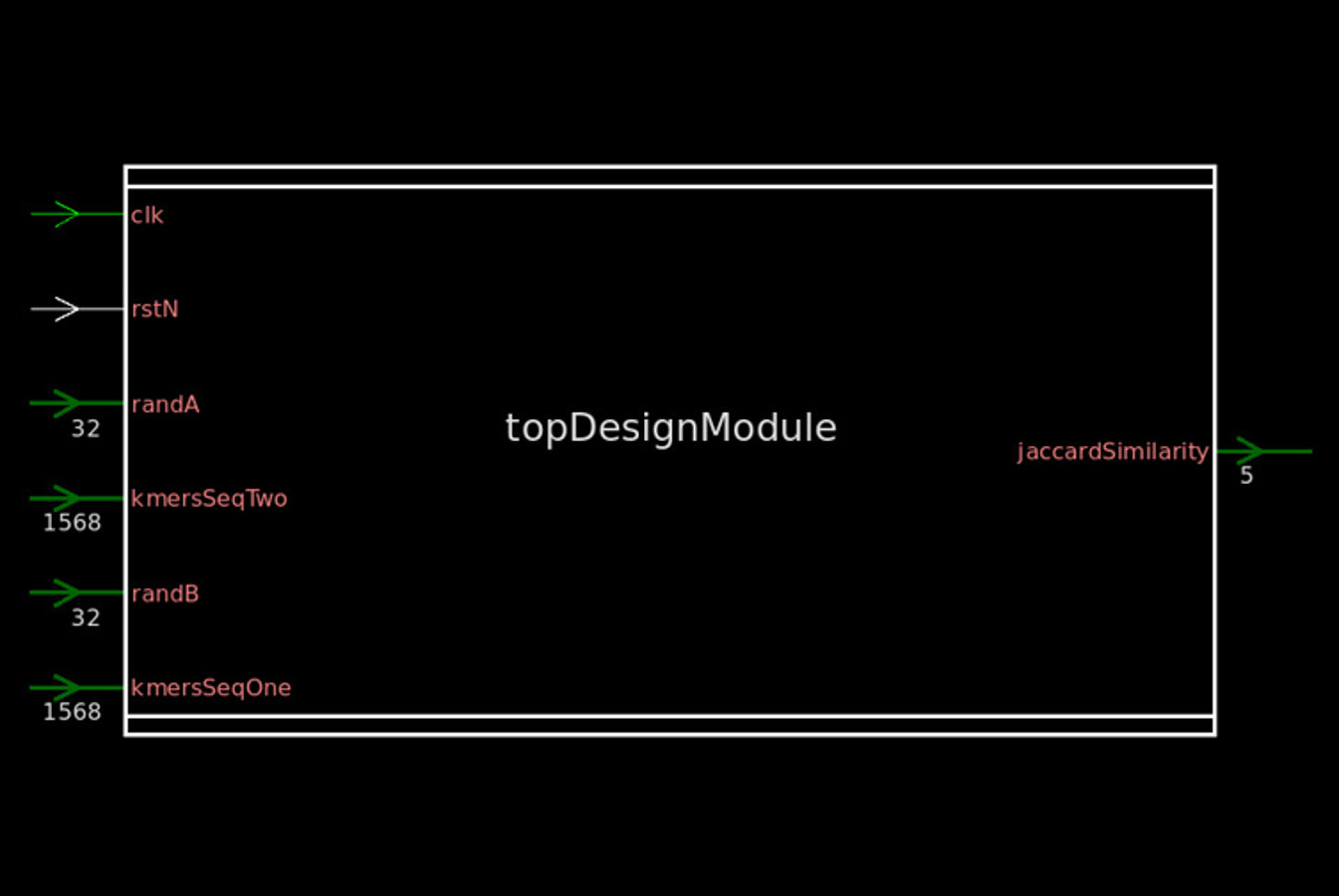
* מקדם ג'קארד, .

בכל עליית שעון, החל מעליית השעון בפעם השלישית בה המידע תקף, הכניסות בעלות הערך המינימאלי שהתקבל עבור הפעלת פונקציית ה על החתימות.  
המוצא גדל באחד בכל פעם בו מתקבל כי אכן ערכי המינימום זהים, לבסוף נפלטת תוצאת מקדם ג'קארד הלא מנורמלת.

# תיאור סכמתי של כלי הסינתזה

נציג את תיאור מבנה הבלוקים של המודל שתכננו שהתקבל לאחר ביצוע סינתזה על התכנון שלנו.

# Top Design

****

איור 14  
סכמת בלוקים לאחר סינתזה של מודל ה

# Hash Module

איור 16  
סכמת בלוקים לאחר סינתזה של מודל ה השני.

**Graphical user interface, application

Description automatically generated**

איור 15  
סכמת בלוקים לאחר סינתזה של מודל ה הראשון.

# Min Module

איור 18  
סכמת בלוקים לאחר סינתזה של מודל ה הראשון.

איור 17  
סכמת בלוקים לאחר סינתזה של מודל ה הראשון.

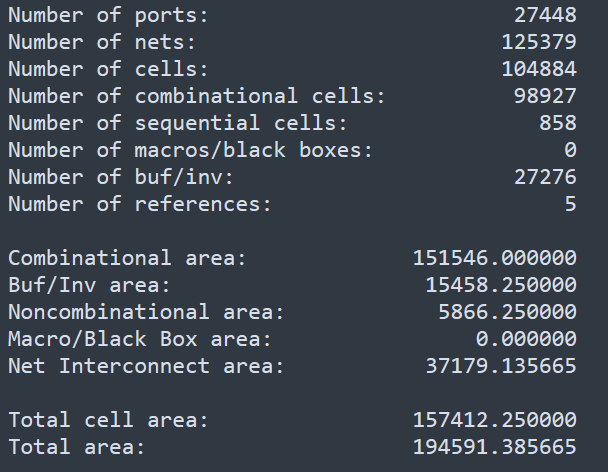
# Counter Module

איור 19  
סכמת בלוקים לאחר סינתזה של מודל ה.

****

# דוחות מהסינתזה

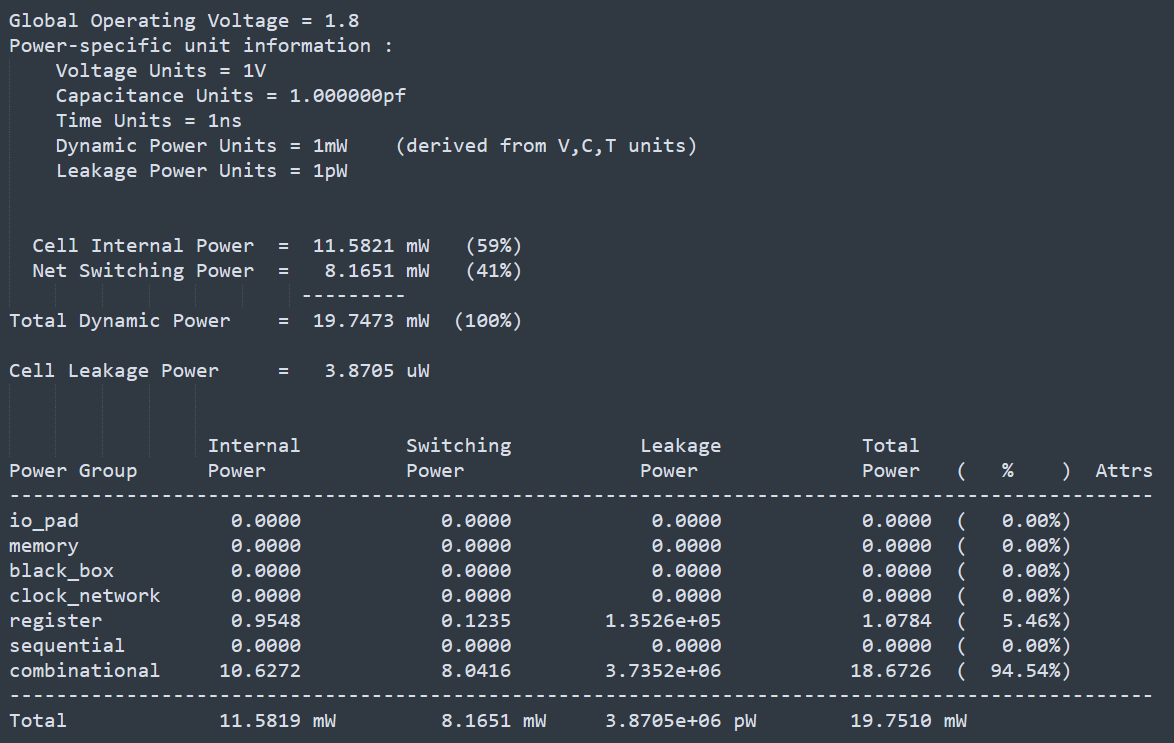
# Area Report

  
  
Timing Report

איור 21  
דו"ח מסלול קריטי

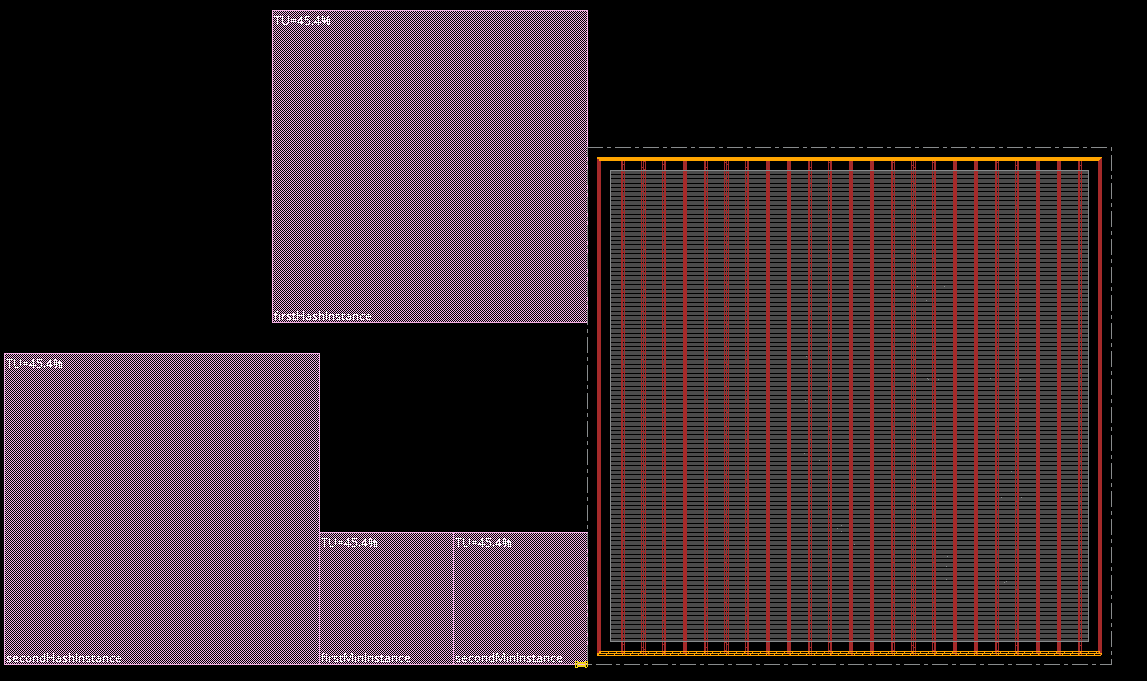
# Power Report

איור 22  
דו"ח צריכת הספק



# Layout

# Adding Power

****

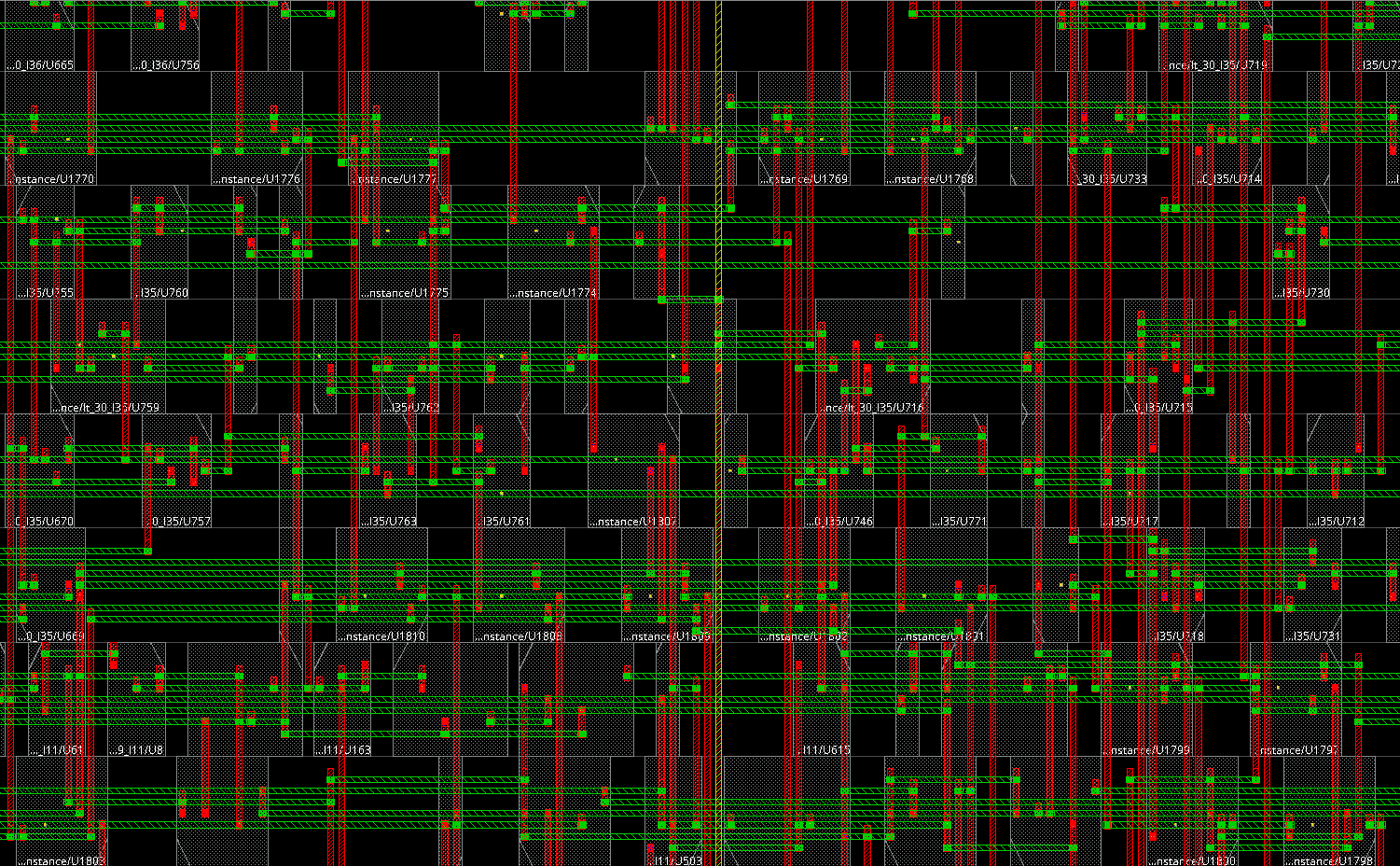
איור 24  
הגדרת רשת האספקה.

איור 23  
הגדרת רשת האספקה.

# Placing Cells

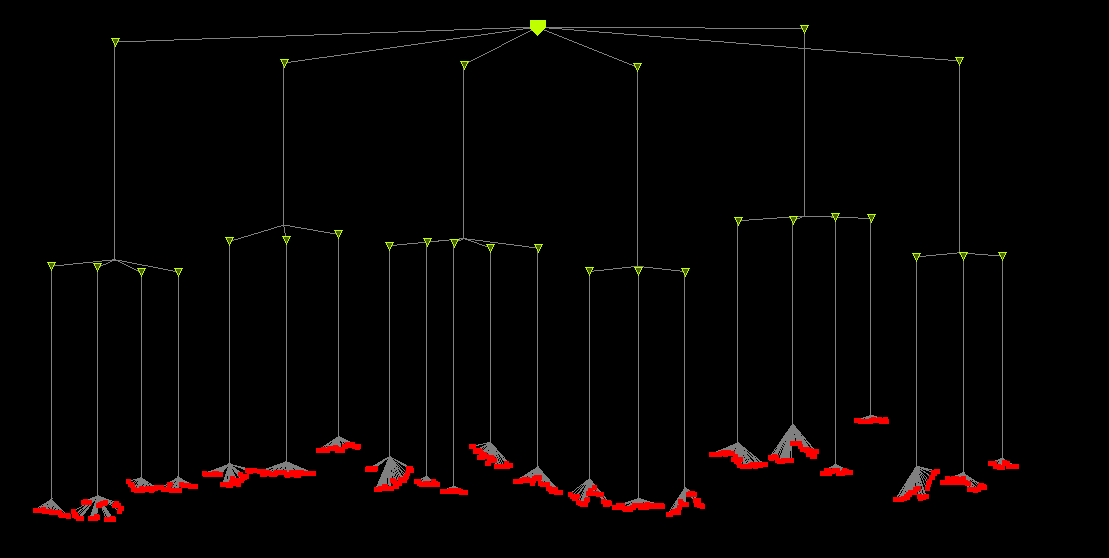
איור 26  
זום על מיקום התאים.

איור 25  
מיקום התאים.



איור 27  
זום נוסף על מיקום התאים.

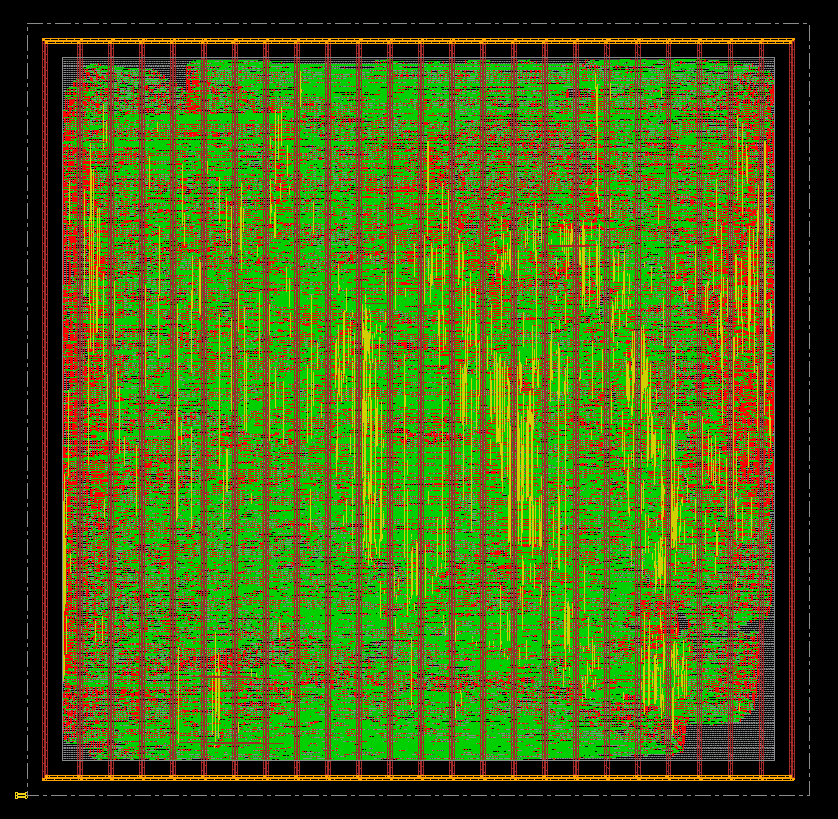
# Clock Tree Synthesis

****

איור 28  
עץ שעון מאוזן

**A picture containing text, display

Description automatically generated**

****Fill Spaces

איור 30  
מילוי רווחים בתאי המילוי.

איור 29  
 לאחר הוספת עץ שעון מאוזן.

****Amoeba View

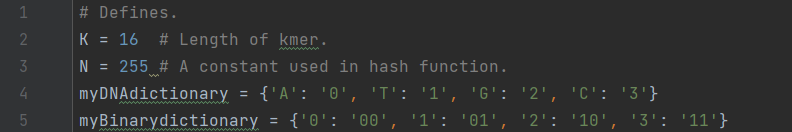
איור 31  
תצוגת Amoeba של הרכיב.

# סימולטור תוכנתי

בנינו סימולטור תוכנתי אשר מומש ב על מנת לקבל השוואה לתוצאות הסימולציה החומרתית.  
הסימולטור הנ"ל מממש את אלגוריתם אותו הצגנו בסביבת עבודה אשר נתמכת על ידי שפת תכנות עילית.  
בנוסף להשוואת התוצאות בין שני הסימולטורים, הסימולטור התוכנתי מקל בכתיבת .

נסביר את השימוש בסימולטור התוכנתי.

1. *תחילה יש להגדיר את הקבועים הרצויים להמשך פעולות הסימולטור.  
   ניתן להגדיר את הקבועים הבאים:*

* *אורך ה, .*
* *מקדם הנרמול עבור פונקציית ה,* .
* הגדרת הייצוג הדצימלי של כל נוקלאוטיד, מוגדר על ידיי מילון.
* ******הגדרת הייצוג הבינארי של כל ספרה, מוגדר על ידיי מילון.

איור 32  
הגדרת קבועים בסימולטור תוכנתי.

1. יש להגדיר כמחרוזות את הרצפים הגנטיים שעליהם מעוניינים לבצע את החישוב.

איור 33  
הגדרת הרצפים הגנטיים.

1. יש להפעיל את הפונקציה אשר מבצעת המרה של הרצפים הגנטיים לייצוג דצימלי ספרה-ספרה (כפי שהסברנו בעמוד 5), לאחר מכן להפעיל את הפונקציה אשר מוצאת את סט תתי – הרצפים של הרצף הגנטי ולבסוף להמיר כל תת – רצף זה לחתימה בינארית באמצעות פונקציה ייעודית.

איור 34  
ייצוג החתימה של כל .

1. יש להגדיר את מערך המספרים הרנדומליים אשר לפיהם מתבצע החישוב של פונקציות ה.

איור 35  
הגדרת המספרים הרנדומליים.

1. יש להריץ את הסימולציה, לטרמינל תודפס התוצאה של מקדם ג'קארד.

# סימולטור חומרתי

רשמנו מספר פונקציות אשר מקלות על מימוש הטסטבנצ'ים והכתיבה החומרתית.

* Text

  Description automatically generatedהפונקציה מפרקת את ה לתצוגה בינארית ומדפיסה את מערך ה בפורמט המתאים לקוד ה.  
  הפונקציה מקבלת את הרצף הגנטי בייצוג נוקלאוטידי ומשתנה בוליאני אשר מציין האם ההשמה מתבצעת לרצף גנטי מספר 1 או 2.

איור 36  
שימוש בפונקציה אשר מדפיסה פורמט מוכן של הkmers ל.

* Text

  Description automatically generatedהפונקציה מבצעת הדפסה של המספרים הרנדומליים לפורמט מתאים לקוד ה.

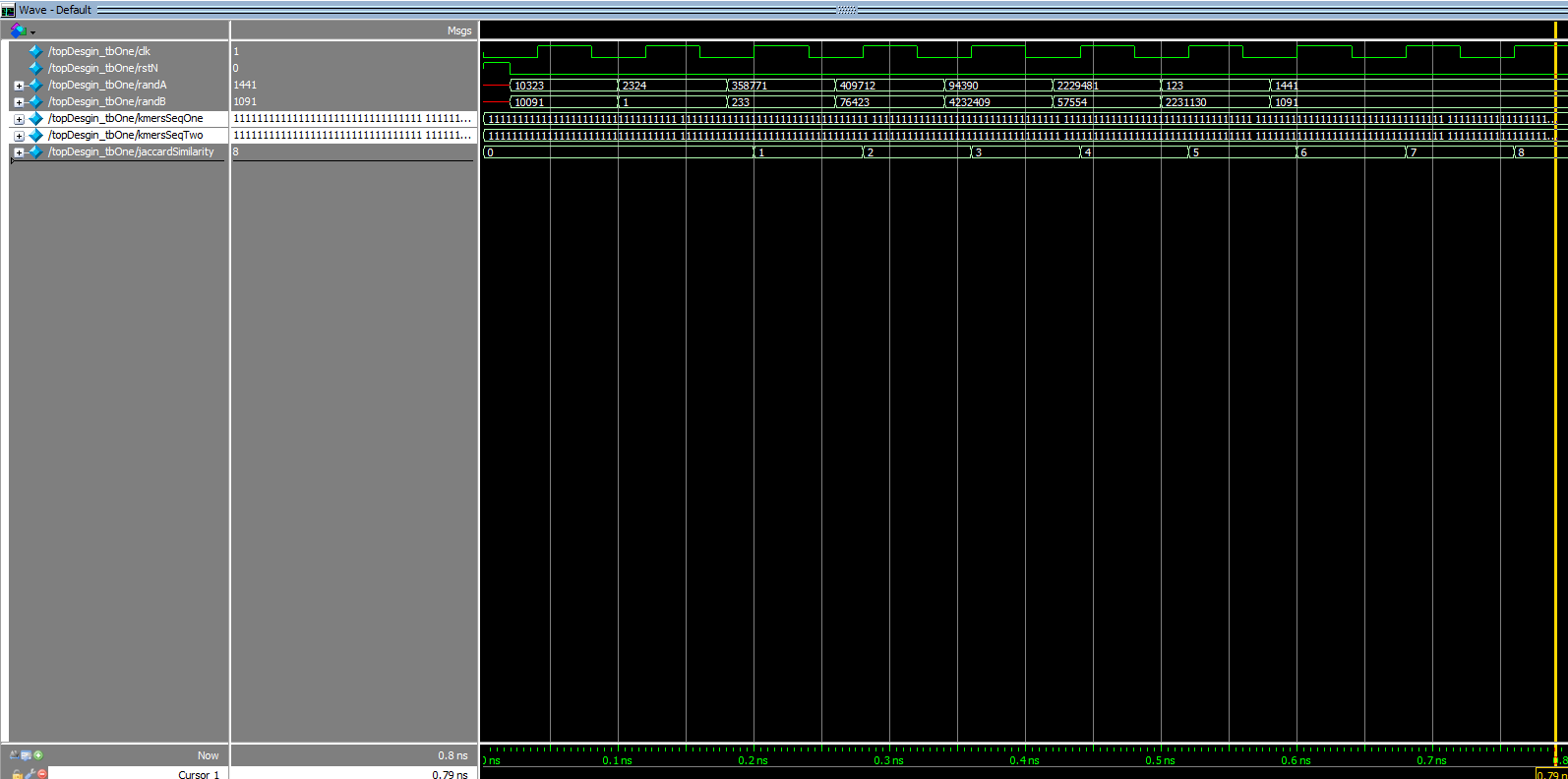
איור 37  
שימוש בפונקציה אשר מדפיסה פורמט מוכן של המספרים הרנדומליים לשימוש ב

# סימולציות

כעת נדגים את פעולת האלגוריתם על מספר סימולציות שהתבצעו על המימוש החומרתי שבנינו ונשווה את התוצאות לסימולטור התוכנתי.  
הסימולטור החומרתי שלנו מגיע לתוצאה סופית בעליית שעון העשירית.

# סימולציה מספר 1 – זהות מוחלטת של הרצפים

בסימולציה זו בדקנו את התוצאה המתקבלת עבור מקדם ג'קארד כאשר הרצפים זהים לחלוטין.  
ברור שכיוון שהרצפים זהים לחלוטין, מתקיים כי גם ה ובפרט ה שלהם זהים ועל כן נצפה לקבל:  
*כלומר, מקדם דמיון מקסימלי.*  
*(התוצאה היא 8 ולא 1 שכן ציינו כי לא בצענו נרמול במספר פונקציות ה*)  
*בחרנו ברצפים הגנטיים הבאים:*



איור 38  
תוצאת הסימולציה החומרתית.

Text

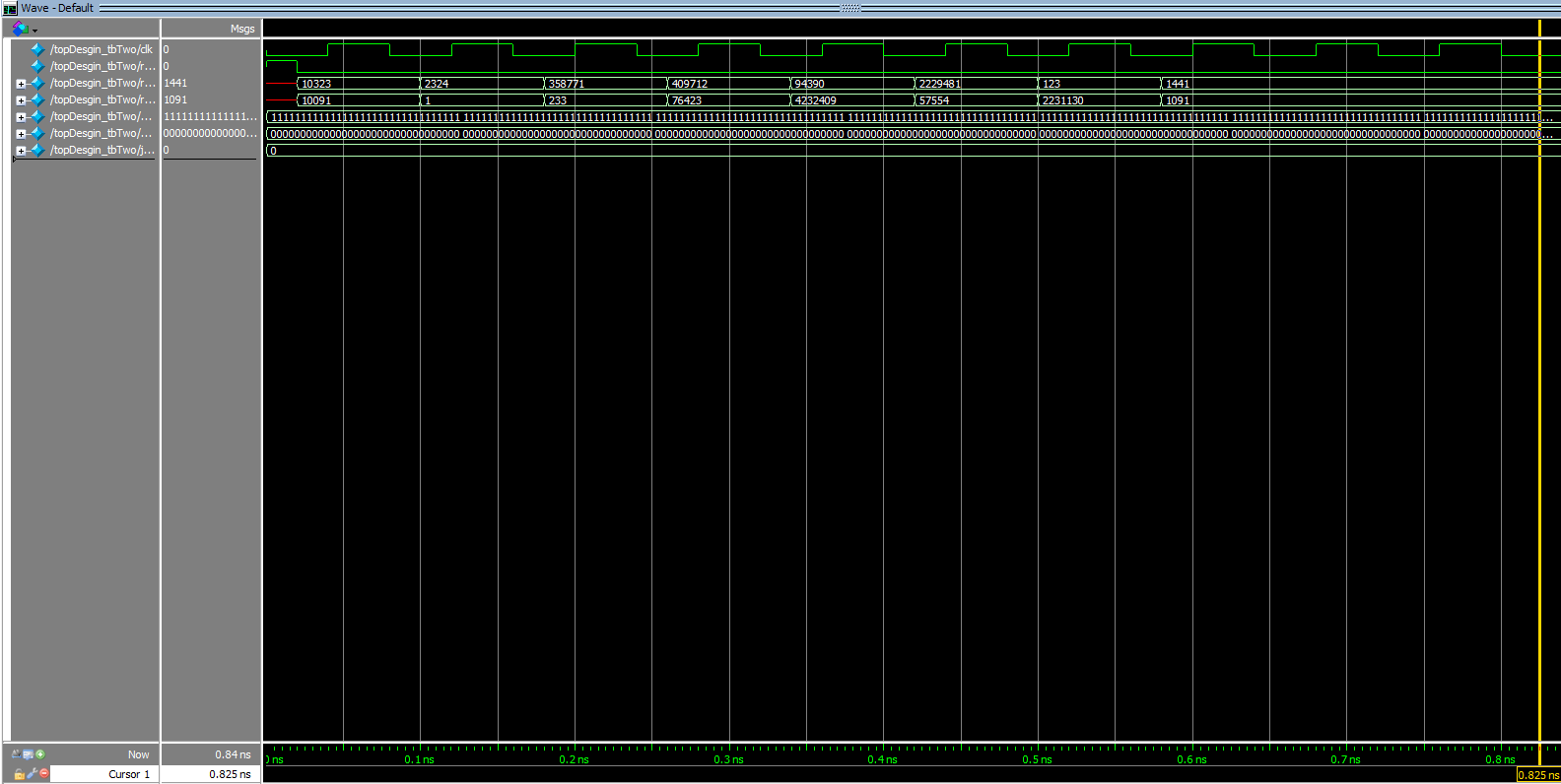
Description automatically generated  
*אכן קיבלנו בשני המקרים כי מקדם ג'קארד הוא 8, בהתאם לציפיות שלנו.*

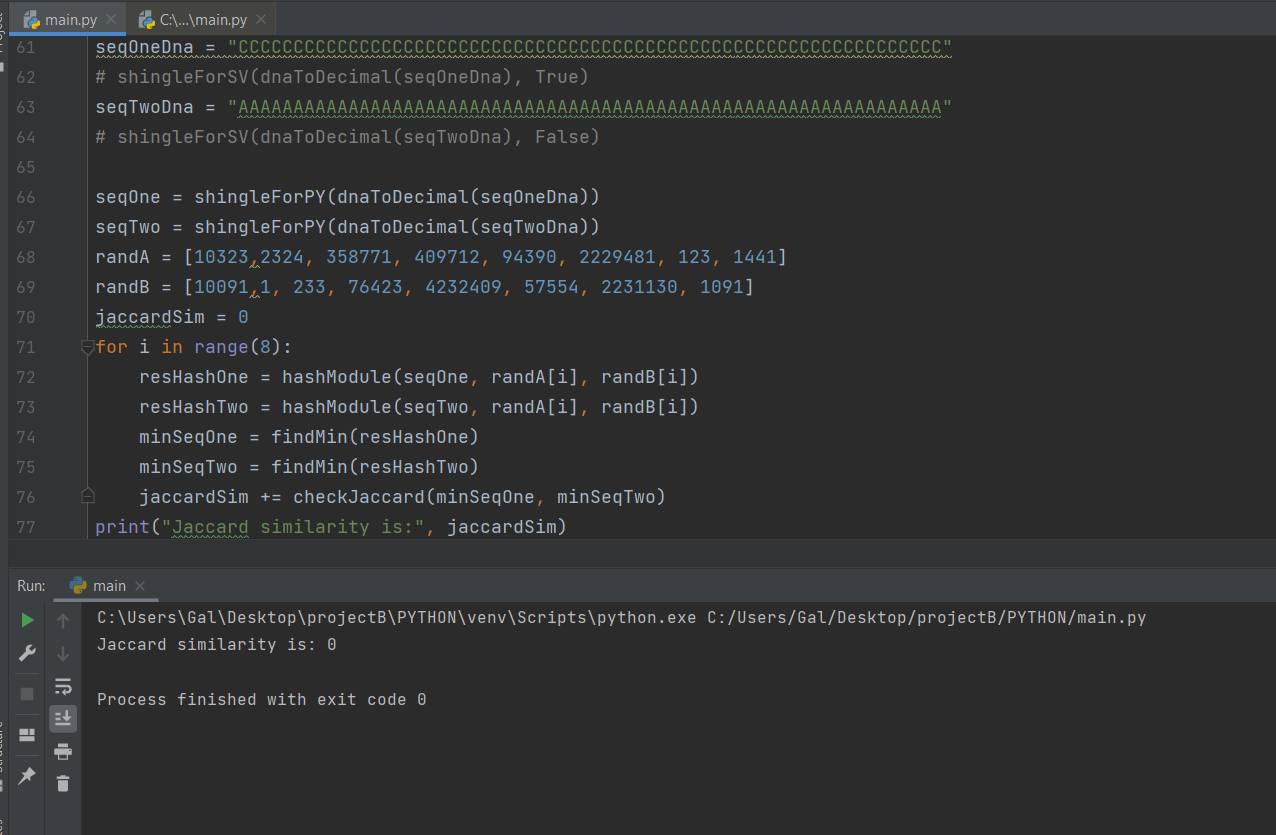
איור39   
תוצאת הסימולציה התוכנתית.

# סימולציה מספר 2 – שוני מוחלט של הרצפים

בסימולציה זו בדקנו את התוצאה המתקבלת עבור מקדם ג'קארד כאשר הרצפים שונים לחלוטין.  
ברור שכיוון שהרצפים שונים לחלוטין, מתקיים כי גם ה ובפרט ה שלהם שונים ועל כן נצפה לקבל:  
*בחרנו בשני הרצפים הגנטיים הבאים:*

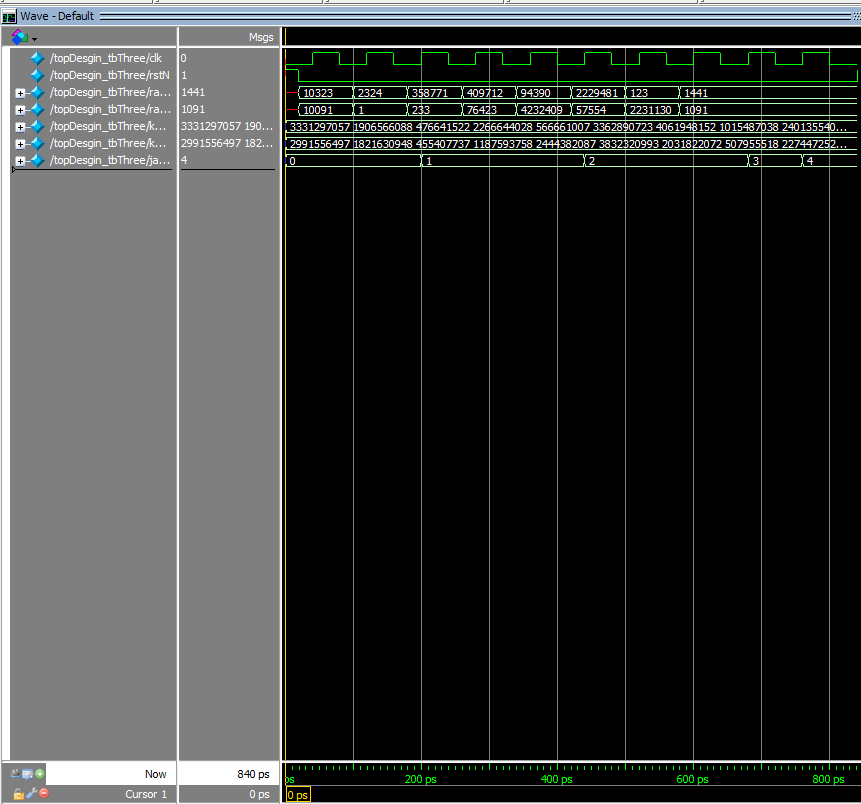
איור40   
תוצאת הסימולציה החומרתית.

**

*אכן קיבלנו בשני המקרים כי מקדם ג'קארד הוא 0, בהתאם לציפיות שלנו.*

איור 41   
תוצאת הסימולציה התוכנתית.

# סימולציה מספר 3 – רצפים כללים 1

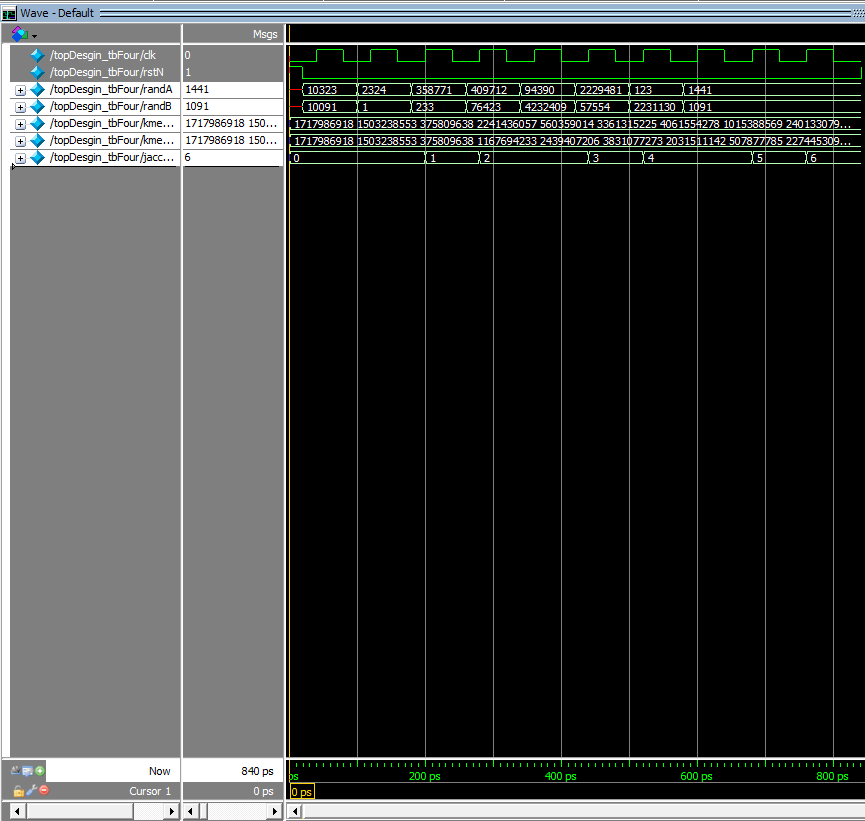
בסימולציה זו בדקנו את התוצאה המתקבלת עבור מקדם ג'קארד כאשר הרצפים בעלי דמיון.  
השתמשנו ברצפים גנטיים מורחבים ל64 אותיות מהרצפים שהצגנו בדוגמה בעמוד 5.  
עבור הרצפים הנ"ל באורך 14 ו באורך 3 ו4 פונקציות , [מהמאמר](../מסמכים/חומר%20רקע/Assembling%20large%20genomes%20with%20single-molecule%20sequencing%20and%20locality-se....pdf) שהתבססנו עליו מתקיים כי מקדם הזהות של שני רצפים אלו הוא 2 ללא נרמול.  
לכן, נצפה שעבור הגדלה של מספר פונקציות ה מקדם הזהות יהיה לפחות 2.

איור 42   
תוצאת הסימולציה החומרתי.

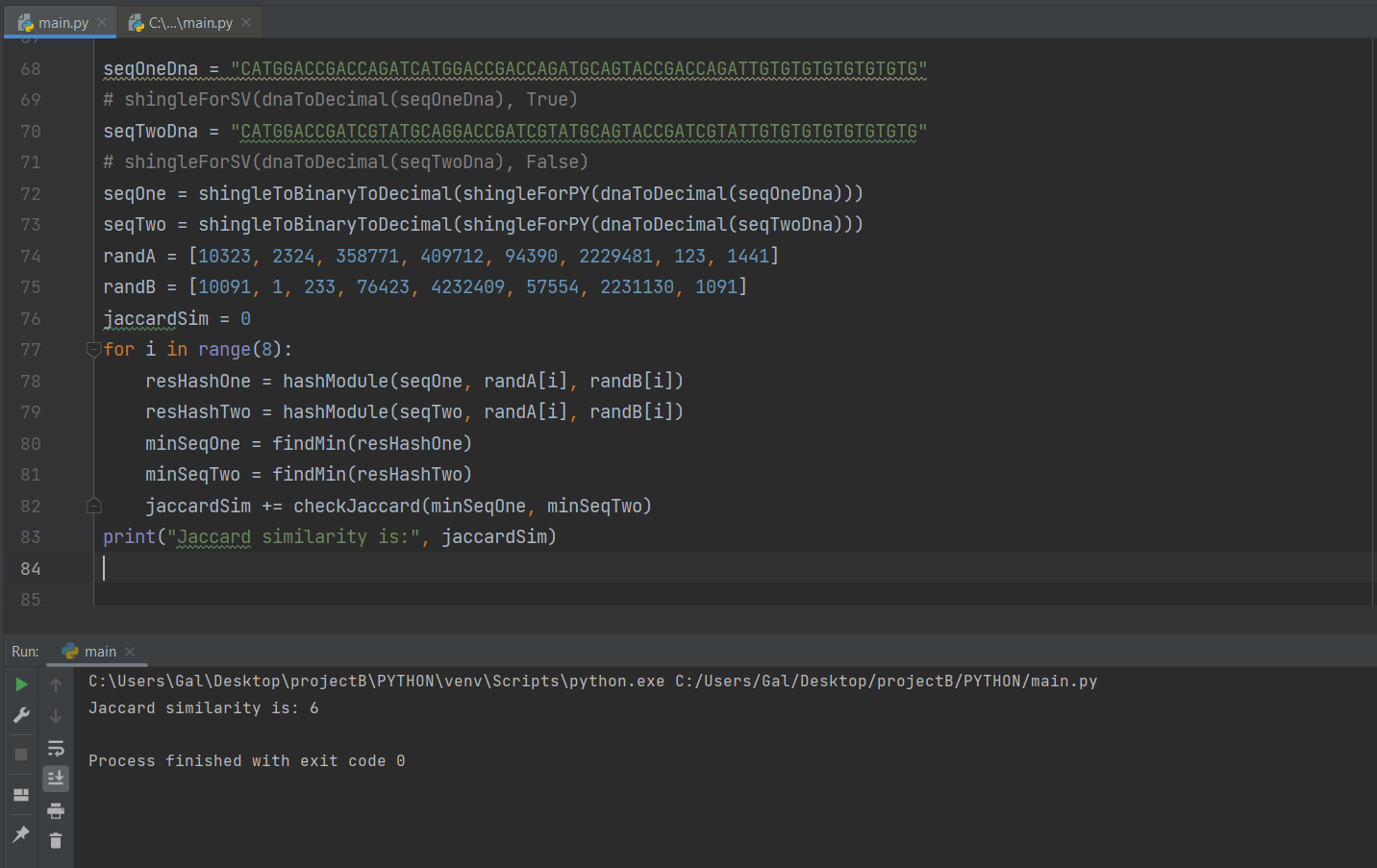
****אכן קיבלנו מקדם זהות זהה עבור שני הסימולציות ובפרט גדול מ2.  
ניתן לראות את הרצפים שבחרנו בצורה נוחה בקוד הפיתון.

איור 43   
תוצאת הסימולציה התוכנתי.

# סימולציה מספר 4 – רצפים כללים 2

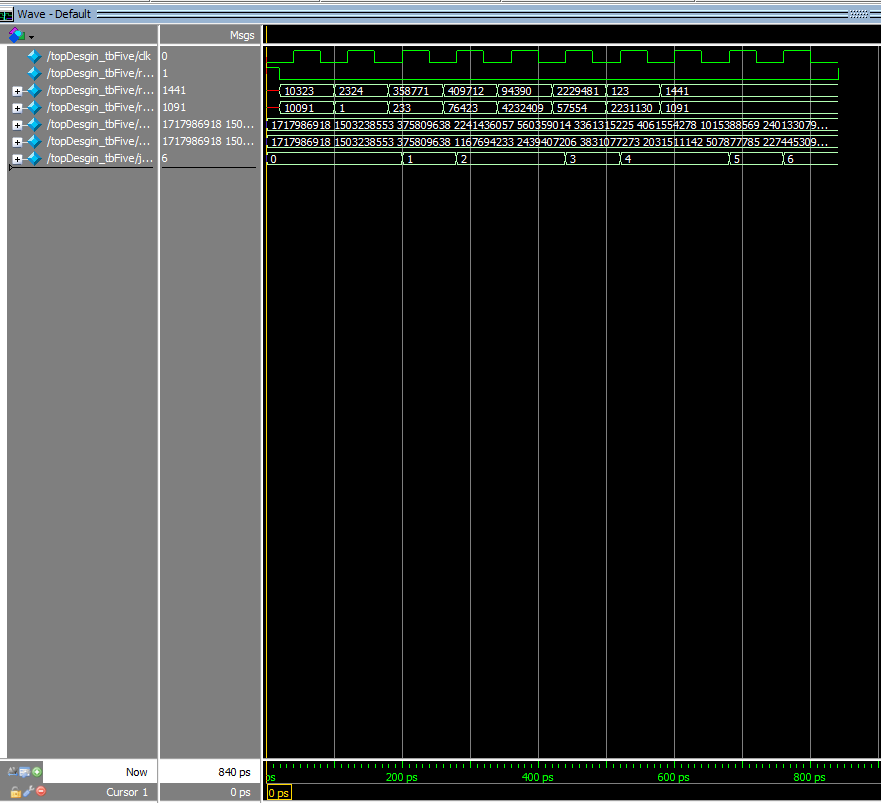
בסימולציה זו בדקנו את התוצאה המתקבלת עבור מקדם ג'קארד כאשר הרצפים הוקשו רנדומלית על ידנו כך שהם יהיו באורך 64 אותיות והשתדלנו להתאים בין הלחיצות עבור הרצפים כך שיתקבל מקדם זהות גבוהה.

איור 44   
תוצאת הסימולציה החומרתי.

אכן קיבלנו מקדם זהות זהה עבור שני הסימולציות ובפרט כיוון שציינו כי השתדלנו לבצע הקלדה זהה של הרצפים, קיבלנו מקדם זהות גבוה.  
ניתן לראות את הרצפים שבחרנו בצורה נוחה בקוד הפיתון.

איור 45   
תוצאת הסימולציה התוכנתי.

# סימולציה מספר 5 – רצפים כללים 3

גם בסימולציה זו בדקנו את התוצאה המתקבלת עבור מקדם ג'קארד כאשר הרצפים הוקשו רנדומלית על ידנו כך שהם יהיו באורך 64 אותיות והשתדלנו להתאים בין הלחיצות עבור הרצפים כך שיתקבל מקדם זהות גבוהה.  


איור 46   
תוצאת הסימולציה החומרתי.



איור 47   
תוצאת הסימולציה התוכנתי.

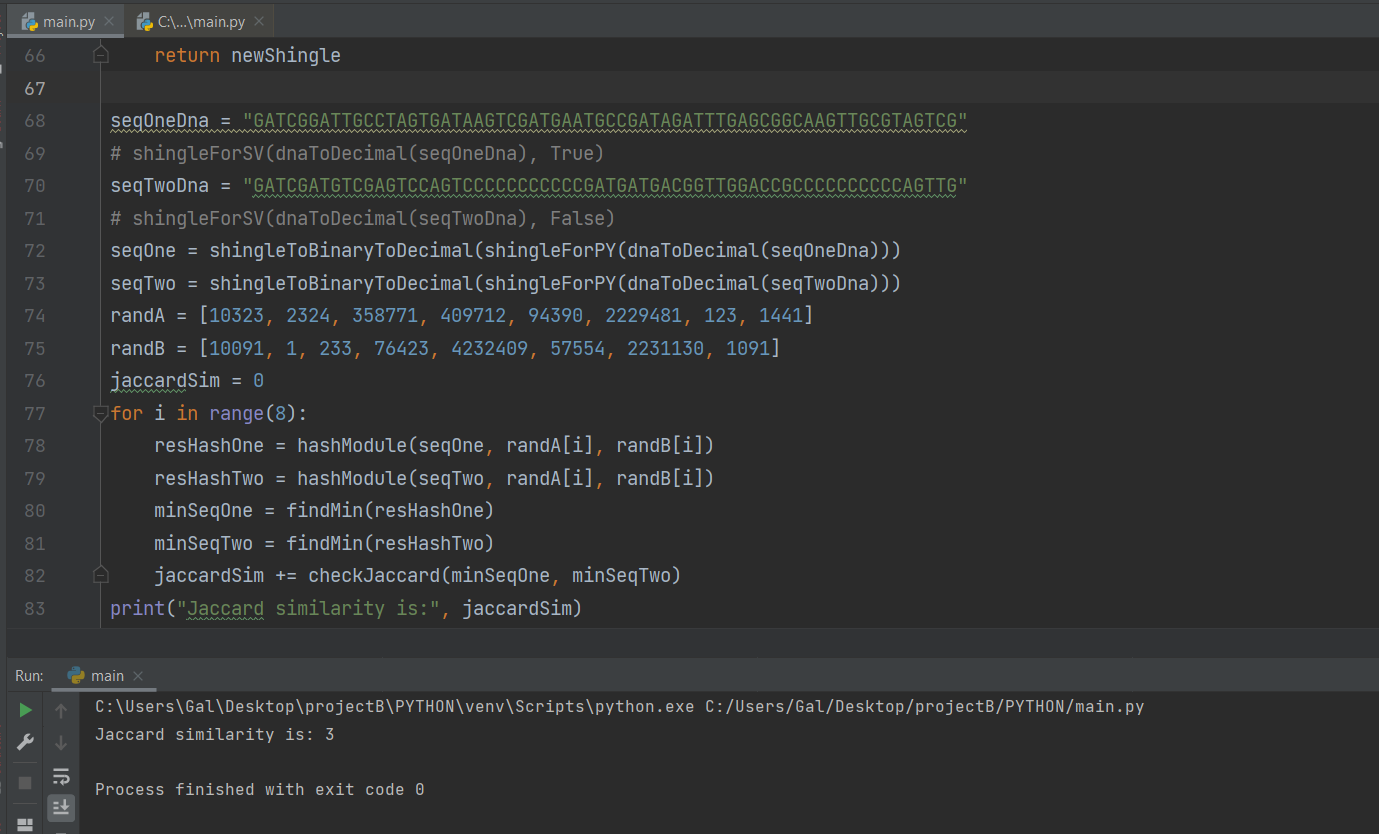
אכן קיבלנו מקדם זהות זהה עבור שני הסימולציות ובפרט כיוון שציינו כי השתדלנו לבצע הקלדה זהה של הרצפים, קיבלנו מקדם זהות גבוה.  
ניתן לראות את הרצפים שבחרנו בצורה נוחה בקוד הפיתון.

# סימולציה מספר 6 – רצפים כללים 4

בסימולציה זו בדקנו את התוצאה המתקבלת עבור מקדם ג'קארד כאשר הרצפים הוקשו רנדומלית על ידנו כך שהם יהיו באורך 64 אותיות והשתדלנו לא להתאים בין הלחיצות עבור הרצפים כך שיתקבל מקדם זהות נמוך. A picture containing chart

Description automatically generated

איור 48   
תוצאת הסימולציה התוכנתי.



איור 49   
תוצאת הסימולציה התוכנתי.

אכן קיבלנו מקדם זהות זהה עבור שני הסימולציות ובפרט כיוון שציינו כי השתדלנו לבצע הקלדה שונה של רצפים, קיבלנו מקדם זהות נמוך, שכן מידת הדמיון בין הרצפים נמוכה.  
ניתן לראות את הרצפים שבחרנו בצורה נוחה בקוד הפיתון.

# סיכום ומסקנות

בפרויקט זה חקרנו את אלגוריתם לצורך שיפור ביצועי הרכבת גנום בשיטת שכאמור בעל עלות חישובית גבוהה.

למדנו את האלגוריתם ויצרנו ייצוג מספרי של רצף גנטי אשר מטיב עם החישובים שלנו ועומד באילוצי אורך של הרכיב שלנו.  
הצלחנו לתכנן ארכיטקטורה אשר מוכיחה את יעילות ומהירות האלגוריתם לצורך זיהוי והרכבתם במידה והם זהים.

בתכנון שלנו נעזרנו בתכנון אשר משלב היררכיות לצורך פירוק הארכיטקטורה לחתיכות כך שכל אחת אחראית על תפקיד משלה, בכך יצרנו מנגנון מצונר אשר בו כל שלב מסתמך על השלב הקודם.  
בנוסף, בנינו סימולטור תוכנתי אשר מהווה אינדיקציה לבדיקה מהירה של תוצאות הסימולטור החומרתי וגילינו התאמה מלאה עבור הסימולציות שבצענו.  
יתרה מזאת, הסימולטור התוכנתי כולל בתוכו ממשק נוח לכתיבה קלה של לצורך בדיקת הרכיב החומרתי עבור בדיקות עתידיות, במידת הצורך.

ניתן לראות כי הביצועים המהירים שהשגנו באים על חשבון שטח הרכיב, מדובר בתכנון עם המון כניסות עבור תתי הרצפים ותוצאות החישוב המתמטיות שמועברות בין השלבים השונים.

# רפרנס

1. <https://www.nature.com/articles/nbt.3238>
2. <https://en.wikipedia.org/wiki/MinHash>
3. <https://www.youtube.com/watch?v=5wvGapmA5zM&t=96s&ab_channel=BioinformaticsDotCa>
4. <https://www.youtube.com/watch?v=l0lZCSU-Gyo&t=1080s&ab_channel=ISCB>

# תודות

בהזדמנות זו נרצה להודות למנחה לאוניד יביץ ולאחראי המעבדה ל גואל סמואל על העזרה, ההכוונה וההנחיה לאורך הפרויקט וההזדמנות להכיר ולהתמודד עם תחום חדש.