תרגיל 2 בביולוגיה חישובית

מגישים: רוני חפץ(212355093 מביולוגיה) יוסף זומר(318808573 ממדעי המחשב)

האם מגיע לנו בונוס? : כן!

איך להריץ ומה צריך להתקין? : הכל בREADME.

קישור לגיט:https://github.com/hroni1234/exercise2-computational-biology

כתבנו אלגוריתם גנטי לפתירת המשחק פוטושיקי, בשלוש דרכים: אלגוריתם גנטי סטנדרטי, דרוויני ולמארקי.

התהליך בקצרה: אתחול דור חדש(בפעם הראשונה) -> ביצוע הכלאות בין פרטי הדור הקודם והוספת הילדים לדור החדש-> ביצוע מוטציה על 30% מהאוכלוסייה -> להעתיק לדור החדש את האחוזון העליון מהדור הקודם -> ביצוע אופטימיזציה על האוכלוסייה החדשה (במידה ואנו במצב הדרוויני או הלמארקי) -> חוזר חלילה עד מספר מסוים של צעדים (הסבר בהמשך תחת הכותרת "התכנסות מוקדמת") או הגעה לפתרון עם אפס שגיאות (בסיום יודפסו כל הנתונים והפתרון).

אתחול אוכלוסייה ראשוני: האוכלוסייה בנויה מפרטים שנבנו בצורה אקראית.

גודל אוכלוסייה: גודל האוכלוסייה יהיה גודל הלוח בחזקת 3 אך חסום בין 100 ל500, בכדי לאפשר מצד אחד אוכלוסיות גדולות יותר ללוחות גדולים יותר(ככול שהלוח גדול יותר, מרחב האפשרויות גדול יותר ואז נרצה להחזיק יותר פתרונות כי קיימים יותר פתרונות אפשריים ללוח גדול), אך מצד שני לא נרצה אוכלוסיות גדולות מידי מטעמי משאבי זמן ומקום. בנוסף לא נרצה אוכלוסיות קטנות מידי כי זה יגרום לאלגוריתם לרוץ זמן רב יותר עד למציאת פתרון חוקי.

ייצוג פרטים (פתרונות): אנו מחזיקים מטריצה עם ערכים מספריים.

ייצוג הבעיה: נחזיק מיפוי של מיקום לערכים, בכדי לשמור היכן יש ערכים מקובעים. נחזיק את המספר שמייצג את הרוחב והגבוה של מטריצת פתרון (N). נחזיק גם זוגות של מיקומים כאשר המיקום השמאלי בזוג מייצג את העובדה שהערך במיקום הנ"ל, גדול מהערך שנמצא במיקום הימני (כמו בקובץ הקלט).

פונקציית הערכה (Fitness): פונקציית ההערכה בנויה על חישוב שנותן "ציון שגיאות". עבור פרט מהאוכלוסייה תחילה נחשב את ציון השגיאה שלו. ציון השגיאה בנוי מסך כמות הכפילויות בשורות בחזקת משקל1 (ערכו 1), בחיבור עם סך כמות הכפילויות בעמודות בחזקת משקל1, בחיבור עם מספר האילוצים שלא מסופקים בפתרון בחזקת משקל2 (ערכו 2), כל הסכום הזה בחזקת משקל3 (ערכו 0.5). **משקל2 גדול ממשקל1** זאת מפני שאנו רוצים לבטא את העובדה שפגיעה באילוצים היא חמורה יותר מכפילות בשורה/עמודה. השימוש הכפול במשקל1 גם עבור השורות וגם עבור העמודות, מבטא את הסימטריה שבין השורות לעמודות. **משקל3 יהיה קטן מ1.0**, זאת בשביל לתת "פקטור" לפרטים עם הרבה שגיאות ולאפשר להם סיכוי טוב יותר להתרבות בעתיד עם שאר הפרטים, זאת בשביל ליצור **גיוון גנטי** ולהימנע ממקסימום מקומי.

לאחר חישוב ציון השגיאה, נחסר זאת מציון השגיאה הכי גרוע שקיים (פשוט ציון השגיאה שהיינו נותנים לפתרון שבו כל התאים בעלי ערכים זהים) ואז את התוצאה נחלק בציון השגיאה הכי גרוע שקיים ונכפיל ב100, בשביל לנרמל את ערכי הפונקציה להיות בין 0 ל100. ביצוע החיסור הוא מטעמי נוחות - **אנו רוצים הערכה מקסימלית כהערכה של הפתרון הכי טוב (אם פשוט היינו מחזירים את ציון השגיאה ללא חיסור, היינו צריכים לדרוש הערכה מינימלית כהערכה של הפתרון הכי טוב).**

ביצוע הכלאה (cross over) : עבור שני פרטים, נבחר תא שעד אליו ניקח את התאים מהפרט הראשון, וממנו והילך ניקח את התאים מהפרט השני, נחבר את התאים ונקבל מטריצה חדשה שמייצגת פתרון. את בחירת התא אנו עושים באופן מסוים. בהסתברות של 50% המיקום יהיה אקראי, ובהסתברות של 50% נבצע את ההכלאה על תא ששייך לאילוץ (חוצים את המטריצות במקום שבו יש אילוץ) **. כך אנו מאפשרים ניסיון של פתירת אילוצים בעזרת הכלאה אך עדיין מאפשרים אקראיות שתאפשר שיפור של מצב העמודות, השורות והאילוצים** (אך כמובן שהשיפור במקרה של בחירה אקראיות הוא נדיר יותר).

בחירת זוג פרטים להכלאה: הפרטים להכלאה נבחרים הסתברותית בהתאם להערכה שלהם (נתנו פקטור שורש לחלשים, מתואר בחלק של ההסבר על פונקציית ההערכה). הסיכוי הוא ההערכה שלך חלקי סך כל ההערכות של כל הפרטים באוכלוסייה. יתכן שאותו פרט יבחר כמה פעמים, אך לא בהכרח עם אותו בן זוג.

מוטציה (mutation) : נחבר בצורה אקראית 30% מהאוכלוסייה ונבצע עליה מוטציה. פרט שנבחר עובר תהליך שבו כל תא שלו יכול לשנות את ערכו לערך אקראי, אך זה קורה בהסתברות של 1 חלק N בריבוע (N הוא רוחב של מטריצת פתרון). **המוטציה עוזרת לנו לצאת ממקסימום מקומי.**

התכנסות מוקדמת : בשביל למנועה הגעה למקסימום מקומי, לאחר 2N איטרציות נבצע את התהליך הבא: נשמור את האחוזון העליון מהדור הנוכחי, נגריל דור חדש ונמזג 99% ממנו עם האחוזון העליון ששמרנו בצד. כך אנו בעצם עושים סוג של התחלה מחדש אך עדיין לא מותרים לגמרי על התוצרים שנוצרו עד עכשיו. **עבור מספר מסוים גדול מאוד של איטרציות (N בחזקת 3 כפול 2), התוכנית תעצור ותדפיס את התוצרים הכי טובים שהשיגה.**

אופטימיזציה: עבור פרט ננסה תחילה לשפר את האילוצים. לכל היותר נמצא 3 אילוצים לא חוקיים ופשוט נבצע פעולות החלפה בין הערכים. אם הדבר שיפר את ההערכה אז נסיים. אחרת לא נשמור את השינוי וננסה לבצע אופטימיזציה על השורות, פשוט נמצא בכל שורה כפילות אחת לכל היותר ונחליף את אחד הערכים שגורמים לבעיה, בערך שלא מופיע בשורה. אם הדבר שיפר את ההערכה אז נסיים. אחרת לא נשמור את השינוי וננסה לבצע אופטימיזציה על העמודות, שהיא סימטרית לזאת של השורות. צעדי האופטימיזציה חסומים מלעיל בכמות השורות (N) שיש בפתרון. גם הדרוונית וגם הלאמרקית יעשו שימוש בפעולה.

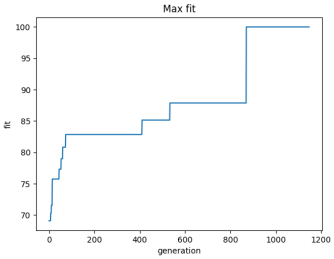
אופטימיזציה דרוונית: בחישוב פונקציית ההערכה, נעתיק את הפרט, נבצע על ההעתק אופטימיזציה ונחזיר את **ההערכה** של ההעתק שעבר אופטימיזציה. בכך נבטא את **הפוטנציאל הגנטי** של אותו פרט.

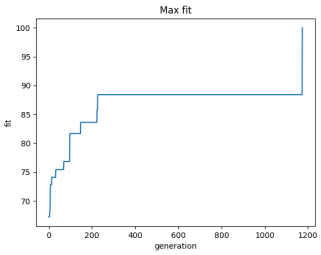
אופטימיזציה למארקית: פשוט בסוף איטרציה נבצע על כל הפרטים אופטימיזציה.

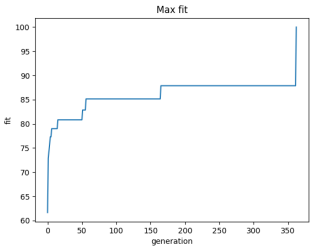
השוואה בין האלגוריתמים : (ציון 100 משמע מציאת פתרון חוקי)

1) **על פי הציון הטוב ביותר :**

5*5 :



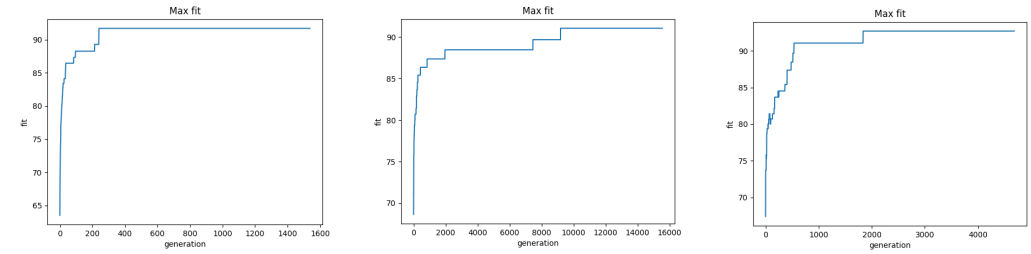




למארקי -> רגיל –> דארויני

ניתן לשים לב , שהאלגוריתם הרגיל נתקע במקסימום מקומי הכי הרבה זמן .האלגוריתם הדארוויני משתפר בהדרגתיות ולעומתו הלמארקי משתפר בקפיצות גדולות יותר ,ומגיע לפתרון מהר יותר .

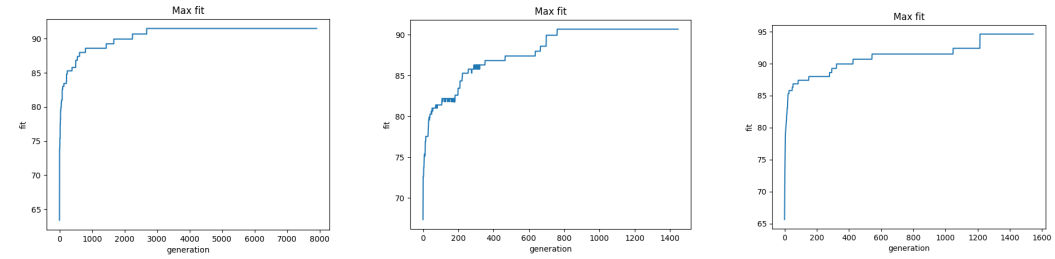
6*6 :



דארוויני -> רגיל –> למארקי

כאן ניתן לראות שהציון הטוב ביותר נמצא אצל הדארוויני . כמו כן ,הלמארקי מגיע מהר למקסימום מקומי ונשאר שם ואילו הדארוויני הצליח להשתפר .

7*7 :



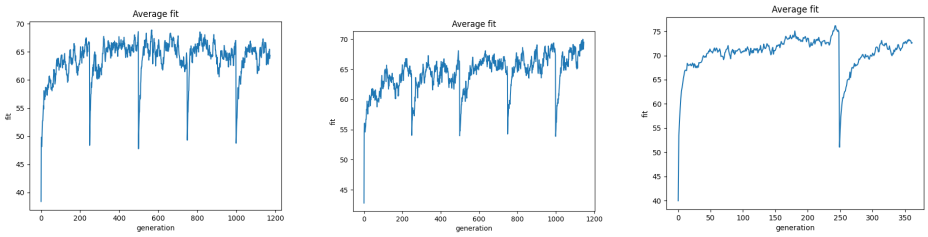
למארקי -> דארוויני –> רגיל

נשים לב שהלמארקי מגיע לציון הכי גבוה , ולאחריו הדארוויני ולבסוף שוב הרגיל .כמו כן ,הם מגיעים לכך במספר דורות נמוך יותר.

2) **על פי ממוצע האוכלוסיה :**

הערה : הירידות בממוצע נובעות מהפעולה שתיארנו תחת הסעיף התכנסות מוקדמת.

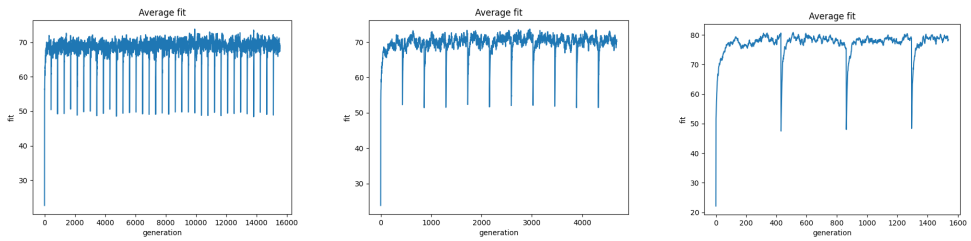
5*5 :



למארקי -> דארוויני –> רגיל

נשים לב שיש עלייה חדה יותר בלמארקי , והוא מגיע גם לממוצע גבוה יותר מהר יותר.

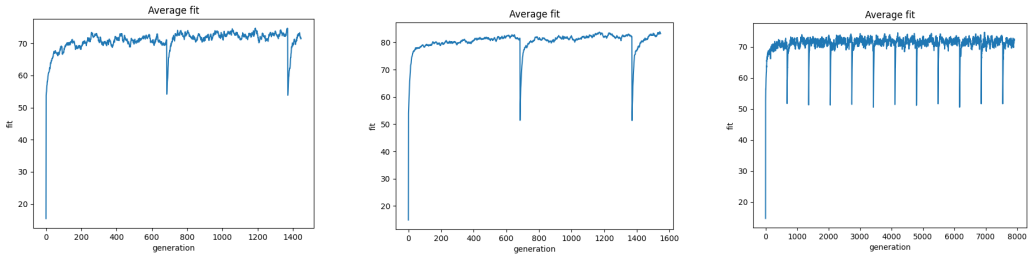
6*6 :



למארקי -> דארוויני –> רגיל

ניתן לראות שהממוצע הכי גבוה נמצא אצל הלמארקי וברגיל ודארוויני הוא דומה.

:7*7



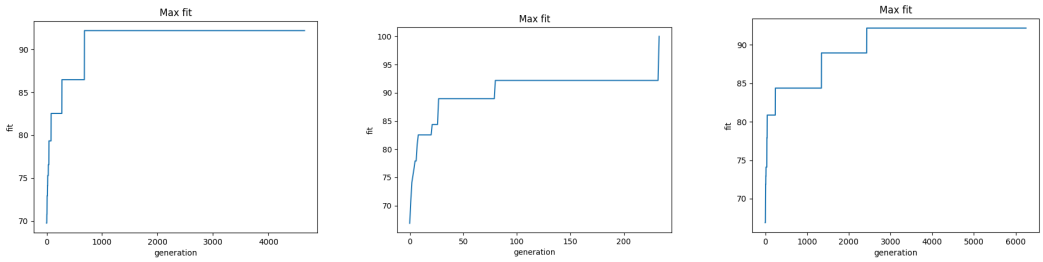
רגיל -> למארקי –> דארוויני

הרגיל והלמארקי נשארים במקסימום מקומי של 70 ואילו הלמארקי מצליח לעלות מהר מאוד למקסימוום של 80 .

השוואה לפי רמות קושי :

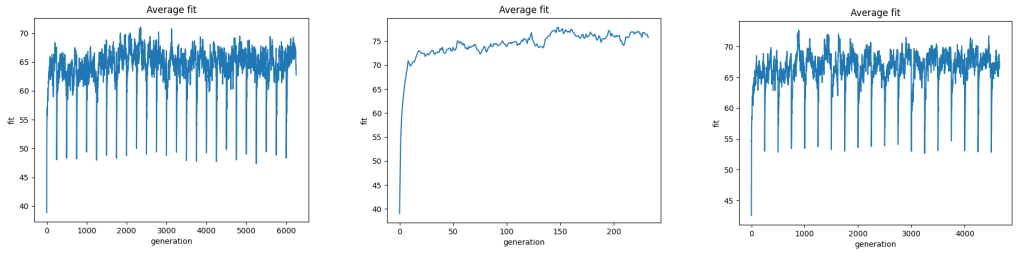
נשווה למשל את הלוח 5*5 ברמות קושי tricky ו easy .

tricky :



רגיל -> למארקי –> דארוויני

אם נשווה לגרפים שצירפנו למעלה שהם ברמת קושי easy , הלמארקי מתכנס מאוד מהר, ואילו השניים האחרים מגיעים לציון גבוה אבל לא מתכנסים. כמו כן, הדארוויני מגיע מהר יותר לציון גבוה יותר.



דארוויני -> למארקי –> רגיל

נשים לב שהממוצע הכי גבוה נמצא אצל הלמארקי , והדארוויני מתקרב אליו בחלק מהנקודות , ובהשוואה לeasy הם נמוכים יותר.

מסקנות :

1) גודל הלוח ורמת הקושי משפיעים על ביצועי האלגוריתם.
2) הלמארקי נוטה להגיע למקסימום מקומי עקב השיפור הכוללני של האוכלוסיה(תכונות טובות שירשו מועברות לדור הבא), ואילו הדארוויני שמעביר "פוטנציאל גנטי" מגיע יותר מאוחר למקסימום מקומי.