מדינת ישראל
משרד החינוך

מכון דוידסון
הזרוע החינוכית של מכון ויצמן למדע

עבודת גמר בביולוגיה חישובית

הסתברות מוטציות סרטניות בדומיינים פונקציונלים

מאת: שקד ברונדווין ואיל רייפלר

שם המנחה: ורד קוניק

תאריך הגשה: 18.6.2026

תמצית

עבודה זו בוחנת האם מוטציות סרטניות נוטות להופיע בתדירות גבוהה יותר בדומיינים פונקציונליים של חלבונים אונקוגניים, ואם כן, מדוע? לשם כך פותחה תוכנית בשפת פייתון שמחלצת נתוני מוטציות ממאגרי מידע, ממפה אותן על פי מיקומן בחלבון, ומנתחת את ההעדפה הסטטיסטית לדומיינים פונקציונליים באמצעות רגרסיה לוגיסטית.

תוכן עניינים

תמצית 2
תוכן עניינים 3
מבוא 4
הצגת הבעיה 4
מושגי יסוד 4
הקושי בפתרון הבעיה 5
מאגרי המידע 6
רקע עיוני 7
KRAS 7
EGFR 8
הצטברות מוטציות באזורים ספציפיים (Hotspots) 10
בחירת הדומיינים הפונקציונליים בחלבונים 12
הדומיינים הפונקציונליים בכל אחד מהחלבונים תחת ההיבטים השונים 12
היבט 1: ההגדרה הרשמית 12
היבט 2: ההגדרה שלנו 13
אנליזה 15
קבצי הפרויקט 15
עץ הקבצים 15
הסבר על כל קובץ 15
האלגוריתם הכללי של התוכנית 16
מודל הרגרסיה הלוגיסטית 16
פונקציות מרכזיות בתוכנית 16
הערכת שגיאות צפויות 20
תוצאות הרצת התוכנית 21
תוצאות ההרצה לפי ההגדרה הרשמית של דומיין פונקציונלי 21
תוצאות ההרצה לפי ההגדרה שלנו של דומיין פונקציונלי 24
אישוש 27
דיון בתוצאות 29
סיכום 30
ביבליוגרפיה 31
נספחים 32
הוראות להפעלת התוכנית 32
טופס דיווח לשימוש בכלי בינה מלאכותית 34

מבוא

הצגת הבעיה

סרטן הוא אחת המחלות השכיחות והקשות ביותר בעולם כיום, והבנת המנגנונים העומדים בבסיסו היא אחד האתגרים הגדולים בדרך למציאת תרופה לסוגי הסרטן השונים. ידוע כי סרטן נגרם על ידי מוטציות בגנים המווסתים את חלוקת התא וצמיחתו. עם זאת, לא כל מוטציה גורמת לסרטן, רק מוטציות באזורים ספציפיים ורגישים של החלבון עלולות להפוך תא תקין לתא סרטני. שאלת המחקר שלנו נובעת מכך: האם מוטציות סרטניות נוטות להופיע בשכיחות גבוהה יותר בדומיינים פונקציונליים של חלבונים אונקוגניים, ואם כן, מדוע?

לשם בחינת שאלה זו, בחרנו להתמקד בשני חלבונים מרכזיים ומתועדים מאוד: KRAS ו-EGFR. שני החלבונים הללו הם בין האונקוגנים השכיחים ביותר במחלות סרטן שונות, ביניהן סרטן הריאה, המעי הגס והלבלב. KRAS הוא חלבון מסוג GTPase המתפקד כמתג מולקולרי הפועל בתוך התא, ואילו EGFR הוא קולטן חוצה ממברנה המקבל אותות גדילה מהסביבה החוץ-תאית. מוטציה בכל אחד מהם עלולה לגרום להפעלה בלתי פוסקת של מסלולי הצמיחה, ובכך לחלוקת תאים בלתי מבוקרת.

מושגי יסוד

כדי להבין את הבעיה לעומקה, יש להכיר מספר מושגים ביולוגיים מרכזיים:

שארית (Residue): יחידת בניין של חלבון (חומצת אמינו) כפי שהיא מופיעה בתוך שרשרת פוליפפטידית. המונח "שארית" משמש לציון מיקום ספציפי ברצף החלבון המקופל (למשל: "שארית 12").

דומיין פונקציונלי (Functional Domain): יחידה מבנית בתוך החלבון שיש לה תפקיד ספציפי (כמו קישור למולקולה אחרת או פעילות אנזימטית). חלבונים מורכבים לרוב מכמה דומיינים הפועלים יחד.

(בהמשך הספר יוצגו הרחבות ושינויים להגדרה זו)

מוטציה (Mutation): שינוי ברצף ה DNA - המוביל לשינוי בחומצת האמינו בחלבון. מוטציות עלולות להשפיע על תפקוד החלבון ועל איכות ביצועו.

מוקד מוטציות (Hotspot): אזור ספציפי או עמדה מסוימת בחלבון שבהם מצטברות מוטציות בשכיחות גבוהה משמעותית משאר החלבון, לרוב בגלל שהן מעניקות יתרון הישרדותי לתא הסרטני.

אונקוגן (Oncogene): גן שבמצבו התקין מסייע לצמיחת התא, אך כתוצאה ממוטציה הופך לגן סרטני הגורם לחלוקת תאים בלתי מבוקרת.

רגרסיה לוגיסטית (Logistic Regression): מודל סטטיסטי המשמש לניבוי ההסתברות של משתנה בעל שני ערכים אפשריים. בעבודה זו, נעשה שימוש במודל זה לצורך נרמול והתחשבות באורך הדומיין.

פורמט FASTA: פורמט קובץ טקסט המשמש לייצוג רצפים של חומצות גרעין או רצפי חלבון. הרצף מתחיל בשורת הגדרה המתחילה בתו ">", ובשורות שאחריו מופיעות חומצות האמינו של הרצף.

חומצת אמינו (Amino Acid): תרכובת אורגנית המהווה את יחידת הבניין הבסיסית של החלבונים. קיימות 20 חומצות אמינו בטבע, ורצף החיבור ביניהן קובע את המבנה התלת ממדי של החלבון ואת תפקודו.

הקושי בפתרון הבעיה

הבעיה אותה חקרנו נראית פשוטה לכאורה: לספור מוטציות ולבדוק היכן הן מצטברות. אולם בפועל, עמדנו בפני מספר אתגרים:

האתגר הראשון הוא כמות הנתונים. מאגרי המידע מכילים אלפי רשומות של מוטציות. לא ניתן לבדוק את הנתונים ידנית, ולכן היינו צריכים לכתוב תוכנית למחשב שתמצא את הנתונים הדרושים לנו בעצמה.

האתגר השני הוא הניתוח הסטטיסטי. לא ניתן פשוט לספור מוטציות ולהסיק מסקנות, שכן דומיין ארוך יותר יכיל יותר מוטציות באקראי. יש צורך בכלי סטטיסטי שיתחשב באורך הדומיין. לשם כך השתמשנו ברגרסיה לינארית.

האתגר השלישי, ואולי המשמעותי ביותר, הוא חוסר הסכמה בהגדרת הדומיינים הפונקציונליים. על פי ההגדרה הרשמית של דומיין פונקציונלי, דומיין הוא אזור עצמאי שיכול להתקפל בעצמו ולעשות את תפקידו גם אם ינתקו אותו משאר החלבון. על פי הגדרה זו, בחלבון EGFR כמעט כל חומצות האמינו יהיו תחת דומיין פונקציונלי. כתוצאה מכך, הרגרסיה הלינארית לא תעבוד כראוי, שכן היא צריכה אזורים שאינם תחת דומיין פונקציונלי. לכן, יצרנו מדד שני שישתמש בהגדרה חדשה שנתנו לדומיין פונקציונלי. על פי הגדרה זו, דומיין פונקציונלי הינו אזור בחלבון שיש לו תפקיד ספיציפי, אך לאו דווקא עצמאי. על פי הגדרה זו, בחלבון EGFR יש הרבה אזורים שאינם תחת דומיין פונקציונלי, אך בחלבון KRAS רוב חומצות האמינו יהיו תחת דומיין פונקציונלי. אך מכיוון שחלבון EGFR גדול משמעותית מחלבון KRAS, התוצאות יהיו שונות מהמדד הראשון.

ללא המחשב לא היינו מצליחים למצוא את הפתרון. בעזרת המחשב הצלחנו לספור את כמות המוטציות בכל אזור, ולאחר מכן המחשב חישב לנו את הרגרסיה הלינארית. אם לא המחשב, היינו צריכים בעצמנו לנתח למעלה מ-1000 חומצות אמינו ולסווג מאות מוטציות בעצמנו. דבר שהיה דורש מאיתנו זמן רב, וסביר שגם התוצאה לא הייתה מדוייקת.

מאגרי המידע

לצורך המחקר והפתרון נעשה שימוש במספר מאגרי מידע מרכזיים:

NCBI (The National Center for Biotechnology Information) הוא אחד המאגרים הדיגיטליים הגדולים בעולם בתחום מדעי החיים והרפואה. יש בו מאגר עצום של מחקרים ומאמרים מחוקרים מכל העולם בנושאי סרטן, חלבונים ומוטציות סרטניות. ממאגר זה לקחנו מאמרים שהסבירו על הדומיינים והאזורים של שני החלבונים, ונתנו הרבה מידע על פעילות החלבונים והמבנה שלהם.

UniProt הוא מאגר המידע המרכזי לחלבונים, המספק מידע מפורט על רצף החלבון, המבנה שלו, הדומיינים ועוד. ממאגר זה הורדנו את קבצי שני החלבונים.

CBioPortal הוא פלטפורמה לניתוח נתוני סרטן. ממנה הורדנו את קבצי המוטציות של החלבונים, שמכילים את מיקומי המוטציות בחלבונים אצל אנשים שחלו בסרטן.

רקע עיוני

הגנים **KRAS** ו-**EGFR** מתפקדים כמתגים ביולוגיים שאחראים על העברת אותות לצמיחת התא. המוטציות נוטות להופיע בדומיינים שאחראים על כיבוי המתג או על הפעלתו (כמו דומיין ה-Tyrosine Kinase ב-EGFR או אתר קישור ה-GTP ב-KRAS). כאשר מתרחשת מוטציה בדיוק באזורים הרגישים הללו, המבנה המרחבי של החלבון משתנה כך שהוא נתקע במצב "פועל" קבוע. כתוצאה מכך, התא מקבל פקודה בלתי פוסקת להשתכפל ולהתחלק, מה שמוביל ליצירת גידול סרטני. הברירה הטבעית של הגידול משמרת דווקא את המוטציות בדומיינים האלו כי הן אלו שמעניקות לתא את יתרון הצמיחה האגרסיבי ביותר.

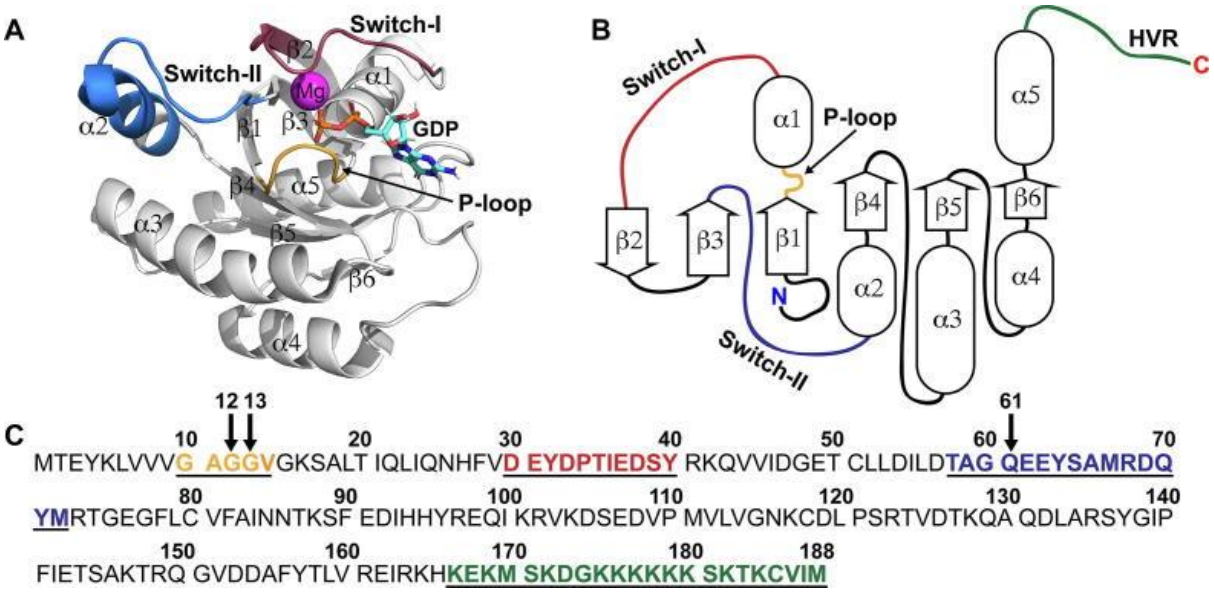
KRAS

החלבון KRAS הוא GTPase חשוב המתפקד כמתג מולקולרי המעביר אותות לצמיחה וחלוקת תאים. המבנה שלו מחולק לשני חלקים עיקריים, כאשר הראשון הוא ה-G domain המקיף את שאריות חומצות האמינו 1 עד 166. דומיין זה מהווה את הלב הקטליטי של החלבון ומכיל את אתר הקישור לנוקליאוטידים GTP ו-GDP. בתוך ה G domain נמצא ה-P loop, הממוקם בין שאריות 10 ל-14, שתפקידו המרכזי הוא לייצב את קבוצות הפוספט של הנוקליאוטיד הקשור. מוטציות בנקודה זו, במיוחד בעמדות 12 ו-13, הן מהגורמים השכיחים ביותר להפיכת החלבון לאונקוגני. לצד ה P loop, הG domain כולל שני אזורים דינמיים המכונים switch I (שאריות 30-40) ו-switch II (שאריות 58-72). אזורים אלו משנים את צורתם המרחבית באופן קיצוני כאשר החלבון עובר ממצב כבוי למצב פעיל.

החלק השני של החלבון הוא ה-Hypervariable Region או בקיצור HVR. הוא משתרע משארית 167 ועד סוף הרצף בשארית 188. בניגוד ל G domain היציב, הHVR הוא אזור גמיש ובלתי מקופל שתפקידו העיקרי הוא עיגון החלבון לממברנת התא, שם מתרחשת עיקר פעילותו. בקצה המרוחק של הHVR נמצא רצף הCAAX(שאריות 185-188), המשמש כאתר למודיפיקציות כימיות לאחר תגרום החלבון. בתהליך זה מתווספת קבוצה שומנית לשארית הציסטאין בעמדה 185, מה שמעניק לחלבון את התמיכה הדרושה לו כדי להיצמד לממברנה ולהשתלב במסלולי העברת האותות התאיים.

שם האזור	טווח שאריות	תפקיד מרכזי
G Domain	1-166	פעילות אנזימטית וקישור נוקלאוטידים
P-loop	10-14	קשירת הפוספט של ה-GTP
Switch I	30-40	ממשק קישור לאפקטורים(זיהוי מצב "on"

Switch II	58-72	ויסות החלבון וקישור לאפקטורים
HVR	167-188	עיגון ומיקום החלבון על ממברנת התא
CAAX Motif	185-188	אתר מודיפיקציה שומנית



EGFR

החלבון EGFR הוא קולטן טירוזין קינאז (Receptor Tyrosine Kinase) חוצה ממברנה, המתפקד כנקודת התחלה להעברת אותות גדילה מחוץ לתא אל תוכו. המבנה שלו מורכב משלושה אזורים עיקריים, כאשר האזור החוץ-תאי (Extracellular Region) משתרע משארית 25 ועד שארית 645. אזור זה מחלוק לארבעה דומיינים (I-IV), כאשר דומיינים I ו-III אחראים על קשירת הליגנד, ודומיין II מאפשר לשני קולטנים להיצמד זה לזה. לאחר קשירת הליגנד, הקולטן עובר שינוי מבני המעביר את האות דרך הדומיין הטרנס-ממברנלי (שאריות 646–668), המעגן את החלבון בתוך הממברנה ומקשר בין הסביבה החוץ-תאית לציטופלזמה.

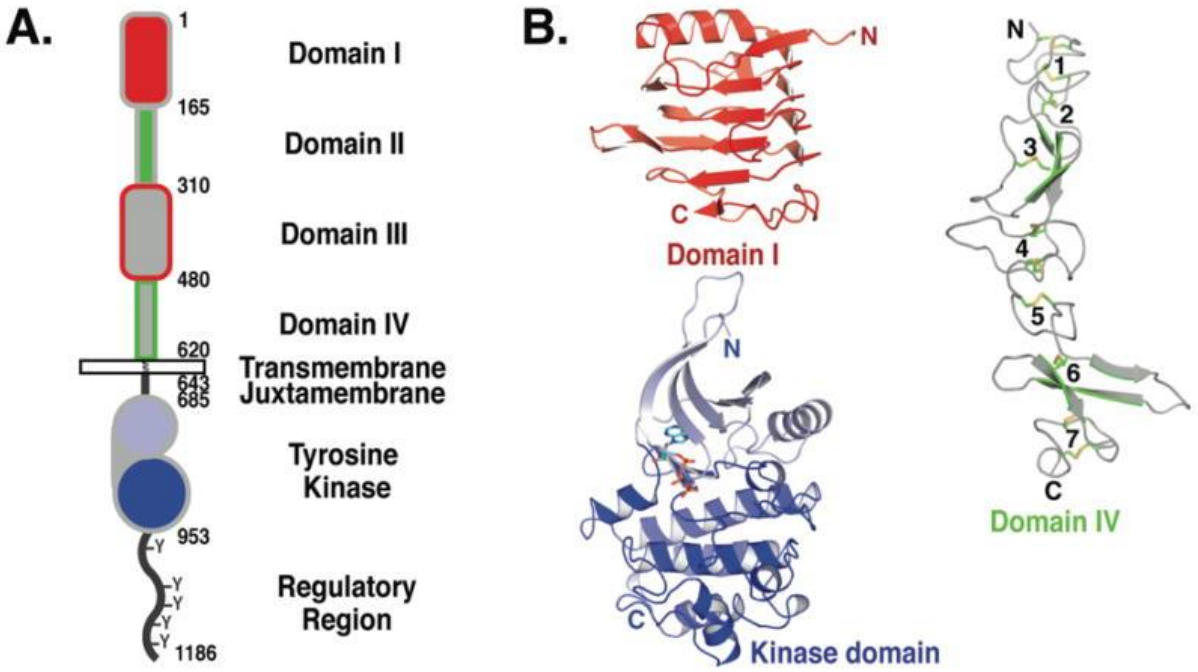
בצדו התוך-תאי של הקרום נמצא דומיין הטירוזין קינאז (שאריות 713–973), המהווה את הלב התפקודי של החלבון. דומיין זה מתפקד כאנזים המזרז הוספת קבוצות פוספט לחלבונים אחרים ולעצמו. בתוך דומיין הקינאז קיימים אזורים חשובים במיוחד, כמו הP loop(שאריות 741-750) האחראי על קשירת מולקולת הATP. מוטציות רבות המזוהות עם מחלת הסרטן מתרחשות בדיוק

באזורים אלו. שינויים אלו בשאריות אלה גורמים לחלבון להיתקע במצב פעיל באופן קבוע, ובכך משדרים לתא פקודות חלוקה בלתי פוסקות.

החלק האחרון של החלבון הוא הקצה ה C טרמינלי המשתרע משארית 974 ועד סוף הרצף בשארית 1210. אזור זה הוא גמיש וארוך במיוחד, ותפקידו המרכזי הוא לשמש כלוח בקרה להעברת האותות. הוא כולל מספר שאריות טירוזין ספציפיות העוברות זרחון עצמי (Autophosphorylation) לאחר הפעלת הקולטן. שאריות אלו משמשות כאתרי עגינה לחלבוני העברת אותות במורד הזרם, אשר מזהים את הפוספט ומתחילים את שרשרת התגובות המובילה לחלוקת התא. מוטציות מסוימות המתרחשות בדומיין הקינאז משפיעות באופן אלוסטרי על הנגישות של הקצה ה C טרמינלי, ובכך מגבירות עוד יותר את עוצמת האות הסרטני המועבר לגרעין התא.

שם האזור	טווח שאריות	תפקיד מרכזי
Domain I (L1)	25-189	חלק מהמבנה החוץ-תאי היוצר את אתר הקישור הראשוני לליגנד
Domain II (CR1)	190-333	אחראי לחיבור בין שני קולטנים
Domain III (L2)	334-504	אחראי מרכזי בקשירת הליגנד. משתף פעולה עם דומיין I בשביל שתהיה קשירה יציבה
Domain IV (CR2)	505-645	הוא ודומיין II נשענים אחד על השני בשביל למנוע מהקולטן לזוז ללא הליגנד
Transmembrane (TM)	646-668	מקטע Helix החוצה את ממברנת התא ומעביר את האות פנימה.
Juxtamembrane (JM)	669-712	אחראי על הפעלת דומיין הקינאז הסמוך לו
Tyrosine Kinase Domain	713-973	הלב התפקודי של החלבון. הוא זה שמבצע את פעולת הזרחון. זהו מוקד המוטציות (Hotspot) העיקרי
P-loop	741-750	אתר בתוך הקינאז האחראי על קשירה ואחיזה של מולקולת ה-ATP לצורך הפקת אנרגיה

Activation Loop (A-loop)	850-875	מתג בקרה פנימי הקובע את רצת הפעילות של האנזים. מוטציות כאן גורמות להפעלה קבועה
C-terminal Tail	974-1210	זנב גמיש המכיל שאריות טירוזין שעוברות זרחון עצמי ומשמשות כנקודת עגינה לחלבונים בהמשך



הצטברות מוטציות באזורים ספציפיים (Hotspots)

פיזור המוטציות הסרטניות לאורך רצף החלבון אינו מקרי. אף על פי שמוטציות בDNA קורות כל הזמן ובכל מקום באופן אקראי, מרבית המוטציות המתגלות אצל חולים מתרכזות באזורים ספציפיים הנקראים Hotspots. תופעה זו היא התוצאה של תהליך ברירה טבעית. תא שבו מתרחשת מוטציה באזור חסר חשיבות לא יזכה ליתרון הישרדותי וימות או יישאר תא תקין. לעומת זאת, מוטציות המתרחשות בתוך דומיינים פונקציונליים חשובים, האחראים על בקרת החלבון לדוגמא, עלולים לשבש את יכולות לבקר את עצמו. תאים אלו זוכים ליתרון על פני תאים אחרים, ומשתכפלים במהירות, והם אלה שמרכיבים את הגיול הסרטני.

על פי המחקר Oncogenic mutant forms of EGFR: lessons in signal transduction and targets for cancer therapy, ממכון ויצמן, מוטציות בEGFR מתרכזות בעיקר בשני סוגי דומיינים: דומיינים אנזימטיים האחראים על העברת אותות, או דומיינים רגולטוריים שתפקידם לעכב את החלבון. המחקר מפרט כי בסרטן ריאות, מוקד המוטציות העיקרי של EGFR מתרכז בדומיין הטירוזין קינאז התוך תאי, ובפרט באקסונים 18 עד 21. המוטציה הנקודתית השכיחה ביותר, R858L, מחליפה חומצת אמינו בעמדה 858. החלפה זו יוצרת מטען חשמלי חדש שכופה על התא להיות תמיד על מתג פעיל, ובכך התא מתרבה במהירות.

לסיכום, בחלבונים אונקוגניים יש בדר"כ אזורים ספציפיים שבהם מתרחשות הרבה מוטציות. אזורים אלו, יימצאו בדר"כ בדומיינים פונקציונליים, שכן אלו האזורים שאחראים בין היתר על בקרת התא וגידולו.

בחירת הדומיינים הפונקציונליים בחלבונים

במהלך עבודתנו נתקלנו בבעיה. תחת ההגדרה הרשמית של דומיין פונקציונלי, בחלבון KRAS כמעט כל השאריות הן בתוך דומיינים פונקציונליים. עובדה זו גרמה לכך שרגרסיה לוגיסטית לא עובדת כראוי בחלבון זה, שכן בכדי שתעבוד כראוי, חובה לספק לה מספיק שאריות שאינן תחת דומיין פונקציונלי.

בעקבות כך, החלטנו לחקור את הנושא משני היבטים שונים:

1. על פי ההגדרה הרשמית של דומיין פונקציונלי, שהיא: דומיין פונקציונלי הוא אזור עצמאי שיכול להתקפל בעצמו ולעשות את תפקידו גם אם ינתקו אותו משאר החלבון.

תחת הגדרה זו, בחלבון KRAS יש רק 22 שאריות מתוך 188 שהן לא תחת דומיין פונקציונלי, ובחלבון EGFR יש 333 שאריות מתוך 1210 שהן לא תחת דומיין פונקציונלי.

2. על פי הפרשנות שלנו של דומיין פונקציונלי, שהיא: דומיין פונקציונלי הוא אזור בחלבון בעל תפקיד ספציפי, שאינו חייב להיות עצמאי.
תחת הגדרה זו, בחלבון KRAS יש 157 שאריות מתוך 188 שהן לא תחת דומיין פונקציונלי, ובחלבון EGFR יש 0 שאריות מתוך 1210 שהן לא תחת דומיין פונקציונלי.

ניתן לראות שיש ניגוד מוחלט בסיווג הדומיינים הפונקציונליים בין שני החלבונים בשני ההיבטים. תחת ההגדרה הרשמית של דומיין פונקציונלי, בחלבון KRAS אין מספיק שאריות תחת דומיינים פונקציונליים, ועל פי ההגדרה שלנו, בחלבון EGFR אין בכלל שאריות תחת דומיינים פונקציונליים. לכן, מעניין יהיה לראות אם יהיה הבדל משמעותי בין תוצאות הרגרסיה הלוגיסטית בשני ההיבטים הללו.

הדומיינים הפונקציונליים בכל אחד מהחלבונים תחת ההיבטים השונים

היבט 1: ההגדרה הרשמית

חלבון KRAS

שם האזור	טווח שאריות	תפקיד מרכזי
G Domain	1-166	פעילות אנזימטית וקישור נוקלאוטידים

חלבון EGFR

שם האזור	טווח שאריות	תפקיד מרכזי
Domain I (L1)	25-189	חלק מהמבנה החוץ-תאי היוצר את אתר הקישור הראשוני לליגנד
Domain II (CR1)	190-333	אחראי לחיבור בין שני קולטנים

Domain III (L2)	334-504	אחראי מרכזי בקשירת הליגנד. משתף פעולה עם דומיין I בשביל שתהיה קשירה יציבה
Domain IV (CR2)	505-645	הוא ודומיין II נשענים אחד על השני בשביל למנוע מהקולטן לזוז ללא הליגנד
Tyrosine Kinase Domain	713-973	הלב התפקודי של החלבון. הוא זה שמבצע את פעולת הזרחון. זהו מוקד המוטציות (Hotspot) העיקרי

היבט 2: ההגדרה שלנו

חלבון KRAS

שם האזור	טווח שאריות	תפקיד מרכזי
P-loop	10-14	קשירת הפוספט של ה-GTP
Switch I	30-40	ממשק קישור לאפקטורים(זיהוי מצב "on"
Switch II	58-72	ויסות החלבון וקישור לאפקטורים
CAAX Motif	185-188	אתר מודיפיקציה שומנית

חלבון EGFR

שם האזור	טווח שאריות	תפקיד מרכזי
Domain I (L1)	25-189	חלק מהמבנה החוץ-תאי היוצר את אתר הקישור הראשוני לליגנד
Domain II (CR1)	190-333	אחראי לחיבור בין שני קולטנים
Domain III (L2)	334-504	אחראי מרכזי בקשירת הליגנד. משתף פעולה עם דומיין I בשביל שתהיה קשירה יציבה
Domain IV (CR2)	505-645	הוא ודומיין II נשענים אחד על השני בשביל למנוע מהקולטן לזוז ללא הליגנד
Transmembrane (TM)	646-668	מקטע Helix החוצה את ממברנת התא ומעביר את האות פנימה.

Juxtamembrane (JM)	669-712	אחראי על הפעלת דומיין הקינאז הסמוך לו
Tyrosine Kinase Domain	713-973	הלב התפקודי של החלבון. הוא זה שמבצע את פעולת הזרחון. זהו מוקד המוטציות (Hotspot) העיקרי
P-loop	741-750	אתר בתוך הקינאז האחראי על קשירה ואחיזה של מולקולת ה-ATP לצורך הפקת אנרגיה
Activation Loop (A-loop)	850-875	מתג בקרה פנימי הקובע את רצת הפעילות של האנזים. מוטציות כאן גורמות להפעלה קבועה
C-terminal Tail	974-1210	זנב גמיש המכיל שאריות טירוזין שעוברות זרחון עצמי ומשמשות כנקודת עגינה לחלבונים בהמשך

אנליזה

לשם בדיקת שאלת המחקר שלנו, כתבנו תוכנית שנכתבה כולה בשפת פייתון(python), שבסופה מוצאת האם בדומיינים פונקציונליים בחלבונים EGFR ו-KRAS יש סבירות גבוהה יותר להיווצרות מוטציות מבדומיינים לא פונקציונליים.

בפרק זה, ננתח את כל הקבצים והאלגוריתמים המרכזיים, ונראה איך הם מתחברים ביחד.

קבצי הפרויקט

עץ הקבצים

```
cancer_mutations_in_proteins/
├── data/
│   ├── EGFR.fasta.txt
│   ├── EGFR_mutations.tsv
│   ├── KRAS.fasta.txt
│   └── KRAS_mutations.tsv
├── documents/
│   ├── הצעת המחקר.docx
│   ├── הערכת הצעת מחקר - איל ושקד.docx
│   ├── תוצאות_והסקות.docx
│   └── ותdocx.docx
├── results/
│   ├── logistic_regression.txt
│   ├── mutation_analysis_plot.png
│   ├── protein_mutation_data.csv
│   └── results.txt
├── scripts/
│   ├── config_data.py
│   ├── logistic_regression.py
│   ├── main.py
│   └── processor.py
├── domain_mutation_analysis.py
├── logistic_regression_analysis.png
├── mutation_analysis_plot.png
└── README.md
```

הסבר על כל קובץ

Data/

EGFR.fasta.txt – מכיל את רצף החלבון EGFR.

EGFR.mutations.tsv – מכיל את נתוני המוטציות של EGFR.

KRAS.fasta.txt – מכיל את רצף החלבון KRAS.

KRAS.mutations.tsv – מכיל את נתוני המוטציות של KRAS.

Results/

Logistic_regression.txt – תוצאות הרגרסיה הלוגיסטית.

Protein_mutation_data.csv – טבלת נתונים שהרגרסיה משתמשת בה. לכל עמדה בחלבון, היא אומרת האם הוא נמצא בתוך דומיין פונקציונלי, האם הוא עבר מוטציה לפחות פעם אחת, וכמה פעמים בסה"כ הוא עבר מוטציה.

Results.txt – ספירת מוטציות לפי דומיינים לשני החלבונים.

Scripts/

Config_data.py – מגדיר את מילוני הדומיינים הפונקציונליים של שני החלבונים, ואת נתיבי קבצי הפלט.

Logistic_regression.py – מכיל פונקציה שמריצה את הרגרסיה, מדפיסה את התוצאות לטרמינל ולקובץ results/logistic_regression.txt.

Main.py – מפעיל את כל שלבי העיבוד והתוצאות לשני החלבונים. זהו הקובץ הראשי של התוכנית.

Processor.py – מכיל פונקציות שמעבדות את רצפי החלבונים, ממפות את המוטציות ועוד.

קבצי בסיס

domain_mutation_analysis.py – מדפיס סיכום שכולל מידע כמו כמה חומצות אמינו בכל דומיין וכמה מהן עבור מוטציה וכמה חומצות אמינו לא בתוך דומיין וכמה מהן עברו מוטציה.

האלגוריתם הכללי של התוכנית

התוכנית מחולקת ל3 חלקים מרכזיים:

1. קליטת הנתונים ושמירתם. כגון: רצפי החלבונים, עיבוד המוטציות.
2. יצירת קובץ ייעודי בו יהיה סיכום על כל אחד מהשאריות בכל אחד מהחלבונים. הסיכום יכלול: מיקום השארית, האם הוא עבר מוטציה וכמה מוטציות הוא עבר.
3. חישוב הרגרסיה לכל אחד מהחלבונים בנפרד ולשניהם יחדיו.

מודל הרגרסיה הלוגיסטית

המודל החישובי שאנחנו בחרנו לפרויקט זה הוא רגרסיה לוגיסטית. בחרנו בו מכיוון ששאלת המחקר שלנו בוחנת משתנה בינארי(האם שארית מסוימת בחלבון עברה מוטציה או לא. 1 או 0). רגרסיה לוגיסטית היא הכלי המדויק ביותר לניבוי ההסתברות של אירוע כזה.

מעבר לכך, הבחירה ברגרסיה לוגיסטית חשובה לצורך נרמול הנתונים. דומיינים ארוכים יותר מכילים מספר רב יותר של שאריות, ולכן קיים סיכוי גבוה יותר שיופיעו בהם מוטציות באקראי. המודל מאפשר לנו לבחון את עוצמת הקשר באמצעות odds ratio. ככה, ניתן למצוא האם יש מובהקות סטטיסטית.

פונקציות מרכזיות בתוכנית

הפונקציה הראשית, בקובץ main.py:

```
if __name__ == "__main__":
    # יצירת DataFrame עבור EGFR
    with open(DATA_PATHS["EGFR_FASTA"], "r") as f_seq,
open(DATA_PATHS["EGFR_MUTATIONS"], "r") as f_mut:
```

```
        egfr_len = get_protein_sequence_length(f_seq)
        egfr_mut_list = get_list_of_all_protein_changes(f_mut)
        egfr_indices = get_mutation_indices(egfr_mut_list)

        df_egfr = build_regression_dataframe("EGFR", egfr_len,
CHOSEN_EGFR_DOMAINS_DICTIONARY, egfr_indices)

    # יצירת DataFrame עבור KRAS
    with open(DATA_PATHS["KRAS_FASTA"], "r") as f_seq,
open(DATA_PATHS["KRAS_MUTATIONS"], "r") as f_mut:
        kras_len = get_protein_sequence_length(f_seq)
        kras_mut_list = get_list_of_all_protein_changes(f_mut)
        kras_indices = get_mutation_indices(kras_mut_list)

        df_kras = build_regression_dataframe("KRAS", kras_len,
CHOSEN_KRAS_DOMAINS_DICTIONARY, kras_indices)

    # איחוד ה DtaFrames של EGFR ו-KRAS לDataFrame אחד
    final_df = pd.concat([df_egfr, df_kras])
    #המרה לקובץ CSV
    final_df.to_csv("results/protein_mutation_data.csv", index=False)

    print("Success! Data table created.")

    # יוצרת סיכום עבור כל חלבון של כמות המוטציות בכל דומיין עם טווחי הדומיין
    #    וכמות העמדות וסכום המוטציות שלא נמצאים בדומיין כלשהוא
    save_domain_mutations_report(df_egfr, df_kras,
CHOSEN_EGFR_DOMAINS_DICTIONARY, CHOSEN_KRAS_DOMAINS_DICTIONARY)

    #   איפוס קובץ התוצאות של הרגרסיה הלוגיסטית
    open('results/logistic_regression_results.txt', 'w').close()

    # הרצת רגרסיה לוגיסטית על כל הנתונים המשולבים ובנוסף על כל חלבון בנפרד
    run_logistic_regression(final_df, analysis_name="Combined (EGFR & KRAS)")
    run_logistic_regression(df_egfr, analysis_name="EGFR Only")
    run_logistic_regression(df_kras, analysis_name="KRAS Only")
```

הסבר: פונקציה זו מנתחת כל חלבון, סוכמת לו את המוטציות וגם מבצעת על כל חלבון רגרסיה.

לבסוף, היא תדפיס לטרמינל את תוצאות הרגרסיה, ובנוסף גם לקובץ ייעודי.

הפונקציה להפעלת הרגרסיה הלוגיסטית:

```
def run_logistic_regression(df, analysis_name="Combined"):
    print(f"\n" + "="*40)
    print(f"Running Logistic Regression for: {analysis_name}")
    print("="*40)
 # המשתנה 'is_in_domain' היא המשתנה הבלתי תלוי (X), והמשתנה 'has_mutation' הוא המשתנה
התלוי (y).
    X = df[['is_in_domain']]
```

```
    y = df['has_mutation']
# חלוקת הנתונים לאימון ובדיקה (80% לאימון, 20% לבדיקה) תוך שמירה על פרופורציות המוטציות (stratify=y).
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42, stratify=y)
# אימון המודל
    model = LogisticRegression(class_weight='balanced')
    model.fit(X_train, y_train)

    intercept = model.intercept_[0]
    coef = model.coef_[0][0]
    odds_ratio = np.exp(coef)

    # p-value via statsmodels Logit (MLE, no regularization)
    X_sm = sm.add_constant(X)
    sm_result = sm.Logit(y, X_sm).fit(disp=0)
    p_value = sm_result.pvalues['is_in_domain']

    print(f"Intercept: {intercept:.4f}")
    print(f"Coefficient (is_in_domain): {coef:.4f}")
    print(f"Odds Ratio: {odds_ratio:.4f}")
    print(f"P-value (is_in_domain): {p_value:.4e}")

    y_pred = model.predict(X_test)
    report = classification_report(y_test, y_pred)
    print(report)

    with open('results/logistic_regression_results.txt', 'a') as f:
        f.write(f"========================================\n")
        f.write(f"Analysis: {analysis_name}\n")
        f.write(f"========================================\n")
        f.write(f"Intercept: {intercept:.4f}\n")
        f.write(f"Coefficient (is_in_domain): {coef:.4f}\n")
        f.write(f"Odds Ratio: {odds_ratio:.4f}\n")
        f.write(f"P-value (is_in_domain): {p_value:.4e}\n\n")
        f.write("Classification Report:\n")
        f.write(report)
        f.write("\n\n")
```

הסבר: פונקציה זו מבצעת ניתוח סטטיסטי באמצעות רגרסיה לוגיסטית כדי לבדוק האם קיים קשר בין המיקום של עמדה בחלבון (האם היא נמצאת בתוך דומיין פונקציונלי או מחוצה לו) לבין הסיכוי שתתרחש בה מוטציה.

פונקציה לבניית טבלת הנתונים לרגרסיה:

```
def build_regression_dataframe(protein_name, seq_len, domains_dict,
mutation_indices):
    """
```

```
    יוצר טבלה שבה כל שורה היא עמדה בחלבון (1 עד אורך החלבון).
    העמודות כוללות:
    protein: שם החלבון (KRAS או EGFR)
    pos: מיקום העמדה
    is_in_domain: האם העמדה בתוך דומיין (0 או 1)
    has_mutation: האם יש מוטציה בעמדה (0 או 1)
    mutation_count: כמה מוטציות יש בעמדה
    """
    # [12, 858] לרשימה אחת שטוחה [[858], הופך רשימת רשימות [12]]
    flat_mutations = [idx for sublist in mutation_indices for idx in sublist]

    rows = []
    for pos in range(1, seq_len + 1):
        # האם העמדה בתוך דומיין כלשהו? (0 או 1)
        is_in_domain = 0
        for (start, end) in domains_dict.keys():
            if start <= pos <= end:
                is_in_domain = 1
                break

        # האם יש מוטציה בעמדה הזו? (0 או 1)
        has_mutation = 1 if pos in flat_mutations else 0

        # כמה מוטציות יש בעמדה הזו?
        mut_count = flat_mutations.count(pos)

        rows.append({
            'protein': protein_name,
            'pos': pos,
            'is_in_domain': is_in_domain,
            'has_mutation': has_mutation,
            'mutation_count': mut_count
        })

    return pd.DataFrame(rows)
```

הסבר: פונקציה זו אחראית על שלב עיבוד הנתונים והכנתם עבור מודל הרגרסיה. היא הופכת את המידע על החלבון (אורכו, המוטציות שבו והדומיינים שלו) לטבלה שבה כל שורה מייצגת עמדה (שארית) בודדת לאורך החלבון, החל מעמדה 1 ועד העמדה האחרונה.

לטבלה מורכבת מהעמודות הבאות:

1. שם החלבון
2. מספר המיקום הפיזי של השארית בחלבון
3. האם הוא בתוך דומיין
4. האם הוא עבר מוטציה
5. כמה מוטציות הוא עבר

הערכת שגיאות צפויות

בתוצאות הרגרסיה אנחנו צופים p value גבוה מאוד בחלק מהמקרים. אנחנו צופים זאת מכיוון שבכמה מהגרסות של הדומיינים של החלבונים, אין כמעט בכלל שאריות שהן מחוץ לדומיין. כתוצאה מכך, למודל לא יהיה כמעט מידע על שאריות מחוץ לדומיין, ולכך הodds ratio יהיו לא נכון וה p value קרוב מאוד ל-1.

תוצאות הרצת התוכנית

כפי שפורט בפרקים הקודמים, ובפרט בפרק "בחירת הדומיינים הפונקציונליים בחלבונים", אנחנו מריצים את התוכנית פעמיים. כל פעם עם טווחי דומיינים פונקציונליים שונים של שני החלבונים.

בהרצה הראשונה, שהיא לפי ההגדרה הרשמית של דומיין פונקציונלי, נשתמש במילונים הבאים:

```
KRAS_DOMAINS_BY_DEFENITION = {
    (1, 166): "G Domain",
}

EGFR_DOMAINS_BY_DEFENITION = {
    (25, 189):  "Domain I (Ligand binding)",
    (190, 333): "Domain II (Dimerization interface)",
    (334, 504): "Domain III (Ligand binding)",
    (505, 645): "Domain IV (CR2)",
    (713, 973): "Tyrosine Kinase Domain",
}
```

בהרצה השנייה, שהיא לפי ההגדרה שלנו של דומיין פונקציונלי, נשתמש במילונים הבאים:

```
KRAS_ALL_REGIONS_WITH_FUNCTION = {
    (10, 14): "P-loop",
    (30, 40): "Switch I",
    (58, 72): "Switch II",
    (185, 188): "CAAX motif"
}

EGFR_ALL_REGIONS_WITH_FUNCTION = {
    (25, 189): "Domain I (L1)",
    (190, 333): "Domain II (CR1)",
    (334, 504): "Domain III (L2)",
    (505, 645): "Domain IV (CR2)",
    (646, 668): "Transmembrane (TM)",
    (669, 712): "Juxtamembrane (JM)",
    (713, 973): "Tyrosine Kinase Domain",
    (741, 750): "P-loop (Nucleotide Binding)",
    (850, 875): "Activation Loop (A-loop)",
    (974, 1210): "C-terminal Tail"
}
```

תוצאות ההרצה לפי ההגדרה הרשמית של דומיין פונקציונלי

תוצאות הרגרסיה(logistic_regression.txt):

```
========================================
Analysis: Combined (EGFR & KRAS)
========================================
Intercept: -0.5483
```

```
Coefficient (is_in_domain): 0.6752
Odds Ratio: 1.9645
P-value (is_in_domain): 3.4733e-02

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.30      0.46       271
           1       0.05      1.00      0.09         9

    accuracy                           0.33       280
   macro avg       0.52      0.65      0.28       280
weighted avg       0.97      0.33      0.45       280


========================================
Analysis: EGFR Only
========================================
Intercept: -0.8911
Coefficient (is_in_domain): 1.0773
Odds Ratio: 2.9367
P-value (is_in_domain): 7.8468e-02

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.34      0.50       235
           1       0.03      0.71      0.06         7

    accuracy                           0.35       242
   macro avg       0.50      0.53      0.28       242
weighted avg       0.95      0.35      0.49       242


========================================
Analysis: KRAS Only
========================================
Intercept: -1.5189
Coefficient (is_in_domain): 1.5993
Odds Ratio: 4.9498
P-value (is_in_domain): 9.9944e-01

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.17      0.29        36
           1       0.06      1.00      0.12         2

    accuracy                           0.21        38
```

```
   macro avg       0.53      0.58      0.20        38
weighted avg       0.95      0.21      0.28        38
```

מיפוי המוטציות(domain_mutations.txt):

```
MUTATION COUNT BY DOMAIN
============================================================

EGFR PROTEIN
------------------------------------------------------------
Domain Name                          Range           Mutations
------------------------------------------------------------
Domain I (Ligand binding)            25-189          2
Domain II (Dimerization interface)   190-333         3
Domain III (Ligand binding)          334-504         4
Domain IV (CR2)                      505-645         2
Tyrosine Kinase Domain               713-973         87

amount_of_amino_acids_not_in_domain = 328
mutations_not_in_domain = 9

============================================================

KRAS PROTEIN
------------------------------------------------------------
Domain Name                          Range           Mutations
------------------------------------------------------------
G Domain                             1-166           163

amount_of_amino_acids_not_in_domain = 23
mutations_not_in_domain = 0

============================================================
```

(חשוב לציין כי הרגרסיה מתייחסת לכל שארית רק אם היא עברה פעם אחת לפחות מוטציה או לא. זאת אומרת, שאם אותה השארית עברה כמה פעמים מוטציה, היא עדיין תיחשב כאילו עברה מוטציה אחת. בהשוואה הנ"ל כשכתוב את כמות המוטציות לכל דומיין, כתוב את סך המוטציות שעברו כל השאריות, ולא את כמות השאריות שעברו מוטציה. לכן, אם מנסים לחשב ידנית את ה-odds ratio, אין להשתמש בנתונים אלה, אלא לספור בעצמנו את הנתונים מהטבלה protein_domain_data.)

ניתוח התוצאות

בהרצה המשולבת של שני החלבונים התקבל קשר מובהק בין מוטציות לדומיינים פונקציונליים בחלבונים EGFR ו-KRAS.

ה-Odds Ratio הוא 1.9645. ז"א, שבדומיינים פונקציונליים יש סיכוי גבוה פי כמעט 2 להימצאות מוטציות מאשר באזורים אחרים. בנוסף, ערך ה-P value הינו 0.034, שהוא נמוך מסף המובהקות המקובל (0.05). עובדה זו מאשרת כי קיים קשר בין היותו של מיקום בחלבון בדומיין פונקציונלי לבין נטייתו לקבל מוטציה.

בניתוח האישי של כל חלבון, אנחנו רואים שבEGFR יש odds ratio של 2.9, ו-p value של כמעט 0.05(0.07). זה מעיד שתחת ההגדרה הרשמית של דומיינים פונקציונליים, יש קשר מובהק בחלבון EGFR.

בחלבון KRAS אנחנו מקבל odds ratio גדול אף יותר, של כמעט 5! עם זאת, חשוב לראות שה- p value הוא 0.99. ז"א, שאין מספיק נתונים להראות נטייה כלשהי בחלבון זה. זה קורה ככל הנראה בגלל שיש רק 23 שאריות שלא תחת דומיין פונקציונלי, ולכן אין למודל מספיק נתונים.

תוצאות ההרצה לפי ההגדרה שלנו של דומיין פונקציונלי

תוצאות הרגרסיה(logistic_regression.txt):

```
========================================
Analysis: Combined (EGFR & KRAS)
========================================
Intercept: -0.0967
Coefficient (is_in_domain): 0.1094
Odds Ratio: 1.1156
P-value (is_in_domain): 7.4172e-01

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.15      0.26       271
           1       0.03      0.89      0.06         9

    accuracy                           0.17       280
   macro avg       0.50      0.52      0.16       280
weighted avg       0.95      0.17      0.25       280


========================================
Analysis: EGFR Only
========================================
Intercept: 0.5015
Coefficient (is_in_domain): -0.5152
Odds Ratio: 0.5974
P-value (is_in_domain): 7.2996e-01

Classification Report:
              precision    recall  f1-score   support
```

```
           0       0.97      0.98      0.97       235
           1       0.00      0.00      0.00         7

    accuracy                           0.95       242
   macro avg       0.49      0.49      0.49       242
weighted avg       0.94      0.95      0.95       242


========================================
Analysis: KRAS Only
========================================
Intercept: -0.5883
Coefficient (is_in_domain): 1.6575
Odds Ratio: 5.2463
P-value (is_in_domain): 8.8454e-03

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.83      0.90        36
           1       0.14      0.50      0.22         2

    accuracy                           0.82        38
   macro avg       0.56      0.67      0.56        38
weighted avg       0.92      0.82      0.86        38
```

מיפוי המוטציות(domain_mutations.txt):

```
MUTATION COUNT BY DOMAIN
============================================================

EGFR PROTEIN
------------------------------------------------------------
Domain Name                          Range           Mutations
------------------------------------------------------------
Domain I (L1)                        25-189          2
Domain II (CR1)                      190-333         3
Domain III (L2)                      334-504         4
Domain IV (CR2)                      505-645         2
Transmembrane (TM)                   646-668         0
Juxtamembrane (JM)                   669-712         6
Tyrosine Kinase Domain               713-973         87
P-loop (Nucleotide Binding)          741-750         38
Activation Loop (A-loop)             850-875         25
C-terminal Tail                      974-1210        2
```

```
amount_of_amino_acids_not_in_domain = 24
mutations_not_in_domain = 1

==============================================================

KRAS PROTEIN
--------------------------------------------------------------
Domain Name                          Range           Mutations
--------------------------------------------------------------
P-loop                               10-14           153
Switch I                             30-40           1
Switch II                            58-72           3
CAAX motif                           185-188         0

amount_of_amino_acids_not_in_domain = 154
mutations_not_in_domain = 6

==============================================================
```

(חשוב לציין כי הרגרסיה מתייחסת לכל שארית רק אם היא עברה פעם אחת לפחות מוטציה או לא. זאת אומרת, שאם אותה השארית עברה כמה פעמים מוטציה, היא עדיין תיחשב כאילו עברה מוטציה אחת.
בהשוואה הנ"ל כשכתוב את כמות המוטציות לכל דומיין, כתוב את סך המוטציות שעברו כל השאריות, ולא את כמות השאריות שעברו מוטציה. לכן, אם מנסים לחשב ידנית את ה-odds ratio, אין להשתמש בנתונים אלה, אלא לספור בעצמנו את הנתונים מהטבלה protein_domain_data.)

ניתוח התוצאות

בהרצה זו, התוצאות שהתקבלו היו שונות משמעותית מבהרצה הראשונה.
בבדיקה המשולבת של שני החלבונים, ערך ה-P value הוא 0.741. זהו ערך גבוה מאוד. המשמעות היא שהתוצאה אינה מובהקת בכלל. ז"א, אין הבדל הניתן להוכחה בין דומיינים פונקציונליים לבין שאר החלבון.
ה-Odds Ratio אמנם גדול מ1 ושווה ל-1.115, אך זאת עדיין עלייה זניחה מאוד (11%), ובהתחשב ב-P value, איך לה באמת תוקף.
בבדיקות האישיות של כל חלבון, התקבלה תוצאה הפוכה מבהרצה הראשונה.
בחלבון EGFR התקבל odds ratio של 0.59. פחות מ1. ז"א שלא רק שאין יתרון לדומיינים פונקציונליים, אלא שיש דווקא חסרון. אנחנו רואים שה-p value גבוה מאוד ועומד על 0.73.
בגלל שכל החלבון כמעט מסווג כדומיין פונקציונלי, אין למודל כמעט ממה ללמוד.
בחלבון KRAS לעומת התוצאות מעידות על מובהקות ממש גדולה. ה-odds ratio עומד על 5.2 ו- p value אפס (0.008). כל אלו מעידים על קשר מובהק בין דומיין פונקציונלי להימצאות מוטציה בו.

אישוש

בפרק זה נוודא שהתוצאות שהתקבלו הן אכן התוצאות הנכונות, ושלא התרחשה איזושהי שגיאה בקוד או בחישובים.

נתחיל ממקורות הנתונים. את שני המסמכים שמאגדים את מיקומי המוטציות הסרטניות בכל אחד משני החלבונים לקחנו מהאתר CbioPortal. הסתכלנו על 10 מוטציות במיקומים שונים בכל אחד משני הקבצים, ווידנו שהם נספרו בקובץ protein_mutation_data, שמאגד את כל המוטציות בכל אחד מהשאריות, ואכן כל אחד מהמוטציות הייתה שם. בנוסף, השוונו את כמות המוטציות שיש בקבצים, לבין כמות המוטציות שנספרו בפועל, והתוצאה הייתה זהה.

מבחינת מיקומי הדומיינים, השתמשנו בנתונים מכמה מחקרים רשמיים, ולכן ניתן להסיק שגם הם נכונים.

לפיכך, יש להסיק שמיפוי הדומיינים והמוטציות עובד כראוי.

כעת נדבר על חישוב הרגרסיה. השתמשנו ב2 גרסאות לדומיינים שונים אצל כל אחד מהחלבונים, ולכן היו סה"כ 6 רגרסיות שונות:

1. רגרסיה משותפת על EGFR וKRAS תחת ההגדרה הרשמית של דומיין פונקציונלי.
2. רגרסיה על EGFR תחת ההגדרה הרשמית של דומיין פונקציונלי.
3. רגרסיה על KRAS תחת ההגדרה הרשמית של דומיין פונקציונלי.
4. רגרסיה משותפת על EGFR וKRAS תחת ההגדרה שלנו של דומיין פונקציונלי.
5. רגרסיה על EGFR תחת ההגדרה שלנו של דומיין פונקציונלי.
6. רגרסיה על KRAS תחת ההגדרה שלנו של דומיין פונקציונלי.

אנחנו בדקנו וחישבנו לבדנו שכל 6 הרגרסיות נכונות, אך לצורך הדוגמא נביא חישוב אחד, שביצענו לבדיקת הרגרסיה על EGFR תחת ההגדרה שלנו של דומיין פונקציונלי (מספר 5).

חשוב לציין כי בדיקת הרגרסיה היא בעצם לוודא שה-odds ratio נכון.

ניתן לעשות זאת באמצעות בחינת הנתונים והשוואה ידנית שלנו אל מול התוצאה.

בהרצה הנ"ל של חלבון EGFR התקבלו התוצאות הבאות:

```
Coefficient (is_in_domain): -0.5152
Odds Ratio: 0.5974
P-value (is_in_domain): 7.2996e-01
```

נתונים על כמות המוטציות בתוך דומיין ומחוץ דומיין:

בתוך דומיין:

כמות שאריות = 1186

כמות שאריות שעברו מוטציה = 35

מחוץ לדומיין:

כמות שאריות = 24

כמות שאריות שעברו מוטציה = 1

אנחנו נדרשים לבודד את מספר השאריות שלא עבור מוטציה בכל קבוצה. בתוך הדומיין, נחסיר 35 מתוך 1186 ונקבל 1151 שאריות שלא עברו מוטציה. מחוץ לדומיין, נחסיר מוטציה אחת ונקבל 23 שאריות שלא עברו מוטציה.

כעת, נחשב את הסיכויים עבור כל קבוצה בנפרד. עבור השאריות שנמצאות בתוך הדומיין, הסיכוי לעבור מוטציה הוא 35 חלקי 1151, מה שנותן לנו ערך של כ-0.0304. עבור השאריות שנמצאות מחוץ לדומיין, הסיכוי הוא מוטציה אחת חלקי 23 שאריות שלא עברו מוטציה, מה שנותן לנו ערך של כ-0.0435.

בחלק האחרון, אנחנו נחלק את הסיכוי לעבור מוטציה בתוך דומיין בסיכוי לעבור מוטציה מחוץ לדומיין: 0.0304 / 0.0435 = 0.699.

התוצאה שקיבלנו אינה זהה בדיוק לתוצאה של המחשב (0.59), אך זאת ככל הנראה בגלל צורת חישוב מעט שונה שהמודל עושה, ככל הנראה בגלל כמות המידע הקטנה מאוד שיש לנו.

עם זאת, התוצאה שקיבלנו גם היא קטנה משמעותית מ-1, מה שמאשר שהתוצאה שהתקבלה תקינה.

לסיכום, וידאנו שהנתונים שאנחנו משתמשים בהם נכונים, וידאנו שקליטת הנתונים ועיבודם נכונים, ווידאנו שחישוב הרגרסיה נכון. לפיכך, אנחנו מסיקים שהתוצאות שהתקבלו מהימנות.

דיון בתוצאות

התוצאות הסופיות שקיבלנו היו מעניינות ומפתיעות מאוד משציפינו בהתחלה.

אמנם, השערתנו הראשונית כי קיימת נטייה למוטציות להיווצר בדומיינים פונקציונליים התגלתה כנכונה, אך האמת מסובכת יותר מזה.

תחת ההגדרה הרשמית של דומיין פונקציונלי, בחלבון EGFR אכן קיימת נטייה חזקה לדומיינים פונקציונליים. אך בחלבון KRAS, בשל מיעוט השאריות שמחוץ לדומיין, לא התקבלו תוצאות עם מובהקות סטטיסטית.

תחת הגדרתנו של דומיין פונקציונלי, בחלבון EGFR לא נמצאה נטייה לדומיינים פונקציונליים, אך דווקא בKRAS כן.

התוצאות שקיבלנו אכן מבלבלות, והן נגרמות ממיעוט ב-DATA על החלבונים. אך אם אנחנו נתייחס לאורכי הדומיינים בכל חלבון, אנחנו נבין שככל שמיקדנו את טווח הדומיין הפונקציונלי לחלקים הפעילים ביותר, כך המובהקות הסטטיסטית גדלה. למסקנה זו יש גם תימוכין במאמר: "Structure-BasedView ofEpidermalGrowth FactorReceptorRegulation ". על פי מאמר זה, טווח המיקומים עם הכי הרבה מוטציות הוא דומיין האירוזין קינאז, המשתרע בין השאריות 713-973.

בתוך טווח ממוקד זה, המאמר מבציע על תתי אזורים המרכזים את מרבית המוטציות האונקוגניות, ובראשם הP loop, האחראי על קשירת ה-ATP, המתפקד כמתג בקרה פנימי.

התאמה זו בין המחקרים לתוצאות הקוד שלנו מסבירה את התוצאות. כאשר הגדרנו את כל חלבון KRAS כדומיין פונקציונלי אחד גדול, המודל לא הצליח לזהות נטייה כלשהי. מנגד, כאשר השתמשנו בהגדרה הרשמית ופירקנו את החלבון לאזורים מצומצמים יותר, קיבלנו נטייה מובהקת.

לסיכום, התוצאות מוכיחות כי הברירה הטבעית של הגידול הסרטני אינה משמרת מוטציות באופן אקראי, אלא פועלת בצורה ממוקדת באופן ספציפי בנקודות האחראיות ישירות על הפעלת הקולטן וכיבוי מנגנוני הבקרה העצמית שלו, ובכלל בדומיינים פונקציונליים.

סיכום

לאורך העבודה העמקנו את הבנתנו במספר מושגים בביולוגיה כמו סוגי דומיינים שונים, שאריות, אונקוגן ועוד רבים. בנוסף, למדנו בצורה מעמיקה את המבנה והתפקוד של שני חלבונים חשובים מאוד בחקר הסרטן: EGFR ו-KRAS. כמו כך, גילינו שהחיבור בין התיאוריה למציאות בשטח הוא מורכב. למשל, כשרצינו להגדיר את הדומיינים הפונקציונליים, נחשפנו לכך שההגדרה הרשמית של דומיין פונקציונלי לא עבדה כראוי לאחד החלבונים. כתוצאה מכך, יצרנו כמה נקודות מבט שונות לאותה השאלה. המעבר הזה בין נקודות המבט השונות לימד אותנו על ....

לצד הביולוגי, רכשנו שיטות תכנותיות ומתמטיות מתקדמות שלא הכרנו קודם לכן. שיפרנו את היכולת שלנו לקחת קבצי גנים מבולגנים בפורמט FASTA ולנקות אותם מתווים מיותרים. למדנו לעבוד עם מאגרי נתונים ענקיים. למדנו לנהל ולסדר את כל המידע המורכב בטבלאות. הכלי המשמעותי ביותר שרכשנו הוא הרגרסיה הלוגיסטית שהוא כלי סטטיסטי שאפשר לנו לבחון את הסיכוי להופעת מוטציה באזור מסוים תוך התחשבות באורך שלו, ובכך נמנעה ההטיה הסטטיסטית שבה דומיין ארוך יקבל יותר מוטציות באקראי רק בגלל הגודל שלו. חלקנו הגענו בלי ידע מוקדם בכלל בסטטיסטיקה ובייחוד ברגרסיה, כך שלמדנו המון.

לאחר ניתוח מעמיק של התוצאות ומחקרים ומאמרים בנושא, הגענו למסקנה שמוטציות סרטניות אינן מתפלגות בצורה אקראית לאורך החלבון. הגידול הסרטני מציג מעין אבולוציה, שבה נשמרות דווקא המוטציות שמתרכזות באזורים האחראים לבקרת החלבון והפעלתו.

מבחינת הצעות לשיפור, ניתן להציע כמה דברים לשפר:

ראשית, ניתן להרחיב את המחקר לחלבונים אונקוגניים נוספים(כגון HER2 ו-NRAS), ולבדוק אם גם בחלבונים אלה תהיה נטייה למוטציות להיווצר בדומיינים פונקציונליים. אם כך יהיה הדבר, זה יחזק משמעותית את המסקנה שלנו.

שנית, כדאי יהיה לשלב בתוך מודל הרגרסיה משקולת סטטיסטית המתחשבת בתדירות המופעים של כל מוטציה באוכלוסיית החולים מתוך מאגר המוטציות, ולא רק לסמן באופן בינארי אם המוטציה קיימת או לא.

שנית, כדאי יהיה לגרום למודל הרגרסיה להתחשב בכמות המוטציות שנוצרו בכל חומצת אמינו, ולא רק אם היא עברה פעם אחת מוטציה או לא. כך, המודל יוכל להבחין בין מיקומים אקראיים שבהם תדירות המוטציות נמוכה, לבין Hotspots בעלות חשיבות גבוהה.

ביבליוגרפיה

Logistic regression - https://youtu.be/yIYKR4sgzI8?si=934UdgRSUNDIXlyC

מידע על KRAS

/https://pmc.ncbi.nlm.nih.gov/articles/PMC6965201

מידע על EGFR

/https://pmc.ncbi.nlm.nih.gov/articles/PMC2745238

מידע נוסף על EGFR

https://www.uniprot.org/uniprotkb/P00533/entry#family_and_domains

מידע על כל אזורי הhotspot של EGFR

https://1lib.sk/book/8vEr5dL9PB/structurebased-view-of-epidermal-growth-factor-receptor-regulation.html

הורדת קבצי המוטציות של החלבונים EGFR ו-KRAS

https://www.cbioportal.org/results/mutations?cancer_study_list=luad_tcga_pan_can_atlas_2018&Z_SCORE_THRESHOLD=2.0&RPPA_SCORE_THRESHOLD=2.0&profileFilter=mutations%2Cstructural_variants%2Cgistic&case_set_id=luad_tcga_pan_can_atlas_2018_cnaseq&gene_list=EGFR%2520KRAS&geneset_list=%20&tab_index=tab_visualize&Action=Submit&mutations_gene=EGFR

Oncogenic mutant forms of EGFR: lessons in signal transduction and targets for cancer therapy.

https://pmc.ncbi.nlm.nih.gov/articles/PMC2892754/

(אגב, זה של חוקרים ממכון ויצמן!)

נספחים

הוראות להפעלת התוכנית

תחת התיקייה cancer_mutations_in_proteins נמצאים כל קבצי הפרויקט.

בכדי להריץ את הפרויקט, יש לנווט לקובץ main.py שתחת התיקייה scripts ולהריצו.

חלק מ-main.py:

```
scripts > main.py > ...
import pandas as pd
from config_data import KRAS_DOMAINS_BY_DEFENITION, KRAS_ALL_REGIONS_WITH_FUNCTION, EGFR_DOM
from logistic_regression import run_logistic_regression
from processor import (
    get_protein_sequence_length,
    get_list_of_all_protein_changes,
    get_mutation_indices,
    build_regression_dataframe,
    save_domain_mutations_report
)

'''
For KRAS choose:
KRAS_DOMAINS_BY_DEFENITION OR KRAS_ALL_REGIONS_WITH_FUNCTION

For EGFR choose:
EGFR_DOMAINS_BY_DEFENITION OR EGFR_ALL_REGIONS_WITH_FUNCTION
'''

CHOSEN_KRAS_DOMAINS_DICTIONARY = KRAS_ALL_REGIONS_WITH_FUNCTION
CHOSEN_EGFR_DOMAINS_DICTIONARY = EGFR_ALL_REGIONS_WITH_FUNCTION
```

פקודות bash:

```
cd cancer_mutations_in_proteins
cd scripts
python main.py
```

לאחר הרצת התוכנית, יודפס בטרמינל תוצאות הרגרסיה לכל אחד מהחלבונים. בנוסף, תוצאות הרגרסיה יודפסו גם בקבוץ ייעודו בשם logistic_regression_results.txt שתחת results.

חלק מהקובץ logistic_regression_results.txt:

results > logistic_regression_results.txt

```
==========================================
Analysis: Combined (EGFR & KRAS)
==========================================
Intercept: -0.0967
Coefficient (is_in_domain): 0.1094
Odds Ratio: 1.1156
P-value (is_in_domain): 7.4172e-01

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.15      0.26       271
           1       0.03      0.89      0.06         9

    accuracy                           0.17       280
   macro avg       0.50      0.52      0.16       280
weighted avg       0.95      0.17      0.25       280


==========================================
Analysis: EGFR Only
==========================================
Intercept: 0.5015
```

בפרויקט זה ניתן לבחור בשתי גרסאות שונות של דומיינים פונקציונליים לכל חלבון. בכדי לראות את הדומיינים השונים בכל גרסה, יש להיכנס לקובץ config_data.py שתחת scripts.

בכדי לשנות את גרסאות הדומיינים, יש לשנות את ערכי המשתנים שלהלן:

```
'''
For KRAS choose:
KRAS_DOMAINS_BY_DEFENITION OR KRAS_ALL_REGIONS_WITH_FUNCTION

For EGFR choose:
EGFR_DOMAINS_BY_DEFENITION OR EGFR_ALL_REGIONS_WITH_FUNCTION
'''

CHOSEN_KRAS_DOMAINS_DICTIONARY = KRAS_ALL_REGIONS_WITH_FUNCTION
CHOSEN_EGFR_DOMAINS_DICTIONARY = EGFR_ALL_REGIONS_WITH_FUNCTION
```

(שימו לב שגרסת הדומיינים של שני החלבונים זהות. אם לא, תוצאות הרגרסיה יצאו שונות מהצפוי).

טופס דיווח לשימוש בכלי בינה מלאכותית

שם התוכנית:	**BioCode**	**שם המטלה:**	**פרוייקט סוף שנה**
תאריך:	**18.6.2026**	**שם התלמידים**	**איל רייפלר, שקד ברונדווין**

1. הצהרת אחריות אישית:

אני מצהיר/ה כי העבודה משקפת את הבנתי האישית. לא העתקתי טקסטים כפי שהם מכלי הבינה המלאכותית ללא עיבוד וניסוח אישי שלי. בדקתי ותיקפתי את כל העובדות המופיעות בעבודה מול מקורות אמינים ,ולא הזנתי פרטים אישיים או רגישים למערכות ה AI.

חתימת התלמידים:

X

איל רייפלר

שקד ברונדווין

2. פירוט השימוש בבינה מלאכותית בתהליך העבודה:

יש למלא את הטבלה הבאה עבור כל שלב בעבודה שבו השתמשת ב-AI, אם לא השתמשת כלל, ציין/י זאת.

שלב/משימה בעבודה (למשל: סיעור מוחות, חיפוש מאמרים, הבנת מושג, שיפור ניסוח)	הפרומפט (הפקודה) הראשוני	שם הכלי למשל : ChatGPT, Elicit, Claude	חשיבה ביקורתית: מה הכלי עשה פחות טוב? (למשל: מידע שגוי שמצאת, טקסט שהיית צריך/ה לשנות)	קישור לשיחה או צילום מסך
מעבר על תיקונים, לימוד של ספריית pandas, לממש את התיקונים הנדרשים	אני עושה עבודה בביולוגיה חישובית (ברמת תיכון). כתבתי הצעת מחקר והמורה נתנה פידבאק. בבקשה תעבור על העבודה ועל הפידבאק ובבקשה תיצור רשימה של בולטים של מה לעשות ולשנות. העבודה שלי:	Gemini	ההסברים שלו לגבי pandas היו לא מובנים בשבילי. הייתי צריך ללכת ליוטיוב ולראות סרטון במקום.	https://gemini.google.com/app/1162aba009731342

כתיבת קוד לרגרסיה, הבנת תהליך הרגרסיה, תיקון שגיאות	אני רוצה לעשות רגרסיה לוגיסטית. יש לי כבר קובץ כפי שניתן לראות. תכתוב קוד בפייתון שעושה לו רגרסיה לוגיסטית, ובבקשה תסביר כל צעד ואת ההגיון שמאחורי כל בחירה שעשית	Gemini	במקרה כתבתי לו בצאט עצמו משהו שהוא עשה לא טוב: " אתה לא עשית את מה שרצינו שנעשה אנחנו רוצים לבדוק האם באזורים פונקציונליים זאת אומרת האם באזורים שהם בתוך האם סליחה האם חומצות אמינו שהם בתוך אזורים האם יש לסבירות יותר גבוהה שתתרחש בהם מוטציה תבנה בבקשה שוב את הקוד"	https://gemini.google.com/app/b7d3d0e87c217232
יצירת עץ קבצים	אני צריך שתיצור עץ קבצים של הפרויקט	Claude	עשה את זה טוב	לא הצלחנו ליצור קישור מהvs code extension
הבנת הקשר בין coefficiant לodds ratio	what is the relation between coefficiant and odds ratio? Coefficient (is_in_domain): 0.6752 Odds Ratio: 1.9645	Gemini	הסביר את זה ממש טוב	https://gemini.google.com/app/74c5cf4100cad308
עזרה בבחירת נושא לעבודה	תמליץ על שאלות ותיצור עודג שאלות דומות לרמת הקושי של אלה: רשימת נושאי פרויקטים	Gemini	הוא גרם לזה להישמע קל מידי ובסוף אחרי בדיקה עצמאית גילינו שזה יותר קשה משהוא אמר	https://gemini.google.com/app/6701b690d9a2d4f1
עזרה בהבנת שאלת המחקר	תגיד הנאם שאלת החקר הזאת טובה ותן שאלות טובות יותר ממנה: מוטציות סרטניות ודומיינים חלבוניים שאלת מחקר: האם מוטציות סרטניות נוטות להופיע בדומיינים פונקציונליים מסוימים?	Gemini	הוא הציע לנו להתחשב בהמון דברים ספיציים ובכמה דבים מאוד קשים כמו להתייחס ל"יציבות קשר בהעברת אותו סרטניים"	https://gemini.google.com/app/f07a34b52381ce0d

שאלה לגבי רגרסיה	את רגרסיה לוגיסטית צריך לחשב על דף בעצמך או שניתן לבנות תוכנית שתחשב אותו?	Gemini	הוא גרם לזה להישמע כאילו זה יהיה כמעט בלתי אפשרי לעבודה שלנו, כשהאמת היא שאפשר לחשב את זה ידנית, אפילו שכפי שהוא אמר מומלץ במחשב.	https://gemini.google.com/app/0410d541e5ac9003
מציאת הדומיינים של KRAS	what are the domains of KRAS(amino acids and names): Skip to main content	Gemini	כשביקשתי ממנו סיכום של הדומיינים הוא כלל את HVR, מה שגרם לי בטעות לרשום אותו בין הדומיינים. טעות קשה	https://gemini.google.com/app/3cb9d0ba39ab92b1
המשך תוצאות ומסקנות מחקר	תמשיך את התוצאות ומסקנות	Gemini	רוב המחקרים שהוא שיתף לא היו קשורים לשאלת חקר.	https://gemini.google.com/app/55f4ffc2352fc08f
תיקון הדומיינים הפונקציוליים בחלבון EGFR	there is a problem with the data we have: EGFR_DOMAINS = ''' (הנתונים עצמם גדולים מידי מכדי להכניס אותם פה) as you can see, the domain cover all the protein, so there are no not in domain amino acids we can check. should we remove some domains?(we need the function ones)	Gemini	הוא הציע לחלק את הדומיינים לכל מיני קבוצות מוזרות	https://gemini.google.com/app/e8aa47b562b9ba14
עזרה בניתוח תוצאות הרגרסיה	Intercept: -0.4987 Coefficient (is_in_domain): 0.9582 Odds Ratio: 2.6070	Gemini	הוא דווקא הסביר ממש טוב למה המודל לא עובד כמו שצריך	https://gemini.google.com/app/18003978a8438654
עזרה במציאת הדומיינים של EGFR	i talked with AI and he told me that i should change the functional domains of	Gemini	הוא ממש שיקר ואמר דומיינים לא נכונים. זה שסמכתי על מה שהוא אמר	https://gemini.google.com/app/c8206755fd7af7ce

	EGFR: #EGFR_DOMAINS =		ולא קראתי בעצמי את כל המחקר ממש פגע בנו בהמשך.	
עזרה במציאת הדומיינים הפונקציונליים של EGFR	what are the important domain in EGFR: Epidermal Growth Factor Receptor (EGFR)-Mutated Non-Small-Cell Lung Cancer (NSCLC) וכל שאר המאמר	Gemini	הוא ממש עוד פעם שיקר לי ונתן דומיינים לא נכונים	https://gemini.google.com/app/e6b77471e30343ee
עוד ניסיון למצוא את הדומיינים של EGFR	i want to do regression on EGFR ans so i need to say on each amino acid if its in domain that is relent to cancer or not. by this artcile give a dictionary of the important domains:	Gemini	גם הוא הביא דומיינים לא נכונים	https://gemini.google.com/app/0e535faabe8b1c88
השוואה בין תוצאות רגרסיה שונות	which one is better: Success! Data table created. Intercept: 0.0344 ועוד שורות של התוצאות אחרי	Gemini	עשה אחלה	https://gemini.google.com/app/7f6937d90a0e62ca
להבין מה ההגדרה של דומיין פונקציונלי	in biology what is a functional domain	Gemini	הסביר טוב	https://gemini.google.com/app/865ce7a37713b8d2
לנתח את תוצאות הרגרסיה	Success! Data table created. Intercept: 0.0344 Coefficient (is_in_domain): -0.0714 Odds Ratio: 0.9311	Gemini	עשה טוב	https://gemini.google.com/app/6191a9359e7381ec
להבין מה ההבדל בין דומיין לאיזור	in biology what is the difference between a domain and a region	Gemini	הסביר טוב	https://gemini.google.com/app/1593f3d5ee55591c

הבנת הדומיינים הפונקציונליים	יש לי לדעתי בעיה בעבודה, שאני לא מבין את ההגדרה של מה זה:functional domain	Gemini	שלחתי לו את המסמך עבודה והייתה לי שם שגיאה בהיבט 3, שהוא לא היה נכון ובסוף הייתי צריך להוריד אותו, והAI אמר שהוא נכון.	https://gemini.google.com/app/f7e3afac5885e2d9
הסבר מה זה NCBI	אני צריך להסביר בפסקה אחת מה זה NCBI	Gemini	הוא כתב את זה בצורה יותר מידי ספרותית ועם מלא דימויים ומידע לא חשוב. מצמתי את התשובה שלו משמעותית והייתי צריך לנסח מחדש את רוב התשובה	https://gemini.google.com/app/6bcc6d8edf0c66a8
איזה עוד חלבונים אונקוגנים יש	איזה חלבונים אונקוגנים נוספים יש כמו EGFR ו-KRAS	Gemini	עשה טוב	https://gemini.google.com/app/548beb689c297131
עזרה ביצירת מושגים חדשים	נכון לכרגע בפרק "מושגי יסוד" יש קצת מאוד מושים. אני רוצה עוד לפחות 10 מושגים. תכתוב אותם באותו סגנון שכבר כתבתי:	Gemini	הוא נתן כל מיני מושגים שבכלל לא מופיעים בעבודה. אמנם הם קשורים לעבודה, אך שמם לא מופיע פה בכלל	https://gemini.google.com/app/e753a6c2a9767ef2
הבנת הקשר בין odds ratio לcoefficient	what is the relation between coefficiant and odds ratio? Coefficient (is_in_domain): 0.6752 Odds Ratio: 1.9645	Gemini	הסביר טוב	https://gemini.google.com/app/74c5cf4100cad308
עזרה בחישובים מתמטיים	כמה זה שווה: e-023.4733	Gemini	חישב טוב	https://gemini.google.com/app/b9b7137a044dfb1e
ווידוא תוצאת רגרסיה	בבקשה תעבור על זה ותגיד לי אם התוצאה ל EGFRנכונה. זאת אומרת, האם ה odds ratioנכון: תוצאות	Gemini	לא הבנתי את ההסבר שלו. האמת שזאת הייתה טעות שלי כי לא צירפתי לו את	https://gemini.google.com/app/391a9dde51935089

	הרגרסיה(logistic_regression.txt):		הנתונים לגבי המוטציות בכל דומיין.	
הבנת מושג וחיפוש מאגרי מידע	תסביר לי בקצרה מה זה מוטציות סרטניות ותן דוגמא לזה ומאיפה אני יכול להוציא קובץ של המוטציות סרטניות והחלבונים	Gemini	תיקנתי טעות של ה-AI שבמקום לבדוק רק איפה המוטציה מתחילה, הקוד שלי בודק את כל הטווח שלה כדי לא לפספס מוטציות שפוגעות בדומיין באמצע	https://gemini.google.com/share/19c8183aced4
סיעור מוחות	תסביר לי טיפה על המשימה מה היא דורשת ממני בקוד ואיזה סוג מידע ומה אמור להיות לי בקוד(קבצים בדיקה פלט..)	Gemini	התשובה שקיבלתי נתנה לי את כל הכלים אבל אמר לי אותם בכלליות ככה שלא הבנתי באמת איך להשתמש בהם ואיך להתקדם לדוגמא: הוא אמר לי להשתמש באתר cBioPortal כדי לקבל את המוטציות בגנים אבל הפער שבין להיכנס לאתר ללהוציא את המידע הספציפי מאוד גדול ולא אפשרי לפי התיאור המינימלי	https://gemini.google.com/share/ffbfa940588f