ניתוח מודלים סטטיסטיים לתיאור תהליך   
ה-SHM בתאי דם לבנים

פרויקט שנה ד', במסגרת המחויבויות לתואר בהנדסת חשמל B.Sc.

הפקולטה להנדסה, אוניברסיטת בר אילן, רמת-גן

מגישים: גילעד אביב וידידיה שמעון

תחת הנחייתם של דניאל עיני ופרופ' גור יערי

שנה"ל תשפ"ב 2021-2022

תוכן עניינים

[תוכן עניינים 2](#_Toc114503587)

[תקציר ומטרת המחקר 3](#_Toc114503588)

[מבוא ורקע תיאורטי 4](#_Toc114503589)

[מערכת החיסון 4](#_Toc114503590)

[תאי B 4](#_Toc114503591)

[Somatic Hypermutation (SHM) 6](#_Toc114503592)

[מודלי מוטציות 8](#_Toc114503593)

[סקירה: מודלים שפותחו במחקר 10](#_Toc114503594)

[מודל ראשון: S5F (2013) 10](#_Toc114503595)

[מודל שני: (2020)N. Spisak et al. 10](#_Toc114503596)

[מודל שלישי: 2PM (דניאל עיני, 2023?) 10](#_Toc114503597)

[חלק א' של הפרויקט: השפעת אפקט הברירה על בניית מודל Substitution ל-SHM 12](#_Toc114503598)

[תהליך העבודה 12](#_Toc114503599)

[תוצאות 13](#_Toc114503600)

[השוואת התוצאות עבור shallow clones ורצפים עם מעט מוטציות 14](#_Toc114503601)

[חלק ב' של העבודה: השוואה בין מודלים שונים 16](#_Toc114503602)

[פיתוח שיטה להשוואה בין מודלים 16](#_Toc114503603)

[ולידציה לפרמטר ה-Score שהגדרנו 18](#_Toc114503604)

[שימוש ב-Score להשוואה על דאטה אמיתי 20](#_Toc114503605)

[סיכום ומסקנות 27](#_Toc114503606)

[כיוונים להמשך 29](#_Toc114503607)

[נספח: קוד המקור 30](#_Toc114503608)

[קוד א': חישוב מקדם המתאם של פירסון בין מודל ה-S5F שאומן על מוטציות שקטות לעומת מודל שאומן כלל המוטציות 30](#_Toc114503609)

[קוד ב': חישוב ה-score להשוואה בין המודלים השונים 33](#_Toc114503610)

# תקציר ומטרת המחקר

המחקר שלנו עוסק בניתוח תהליך ה-SHM בתאי B באמצעות מודלי מוטציות.

המחקר מתחלק ל-2 חלקים:

1. בשלב הראשון של עבודתנו, בחנו בעיה מרכזית איתה יש להתמודד כאשר ניגשים לניתוח תהליך ה-SHM באמצעות מודלי מוטציות: השפעת ה-Selection Bias ("הטיית הברירה", כתוצאה מתהליך ה-Affinity-Driven Selection על איכות המודלים שנבנים. זוהי שאלה מרכזית שמודלים שונים שפותחו במחקר ניסו להתמודד איתה בדרכים מגוונות. אנחנו בדקנו באיזו מידה הטיה זו אכן באה לידי ביטוי בפרמטרים של אחד המודלים, מודל ה-S5F.
2. בשלב השני של עבודתנו, פיתחנו פונקציית score מבוססת likelihood להשוואה בין מודלי SHM שונים. השתמשנו בשיטה זו כדי להשוות מודלים מסוגים שונים, וכאלה שאומנו על סוגים שונים של דאטה.

# מבוא ורקע תיאורטי

## מערכת החיסון

מערכת החיסון היא מכלול מורכב של איברים תאים ומולקולות, שתפקידם להגן על הגוף מפני גורמי מחלה למיניהם, כגון: חיידקים, נגיפים, פטריות וטפילים שונים, וכן מפני רעלנים והתפתחות סרטן.

מערכת החיסון מורכבת משתי מערכות: **מערכת החיסון המולדת**, ו**מערכת החיסון הנרכשת**.

**מערכת החיסון המולדת** קיימת בגוף החל משלב ההתפתחות העוברית. ככלל, מערכת זו מתאפיינת בכך שרכיביה השונים פועלים באופן בלתי-ספציפי, כלומר כל תא או מולקולה השייכים אליה פועלים כנגד מגוון רחב של פתוגנים (גורמי מחלה, כמו חיידקים, וירוסים וכו') רק על בסיס ההבחנה בין "עצמי" ל-"זר". היא תוקפת כל עצם שמזוהה על ידה כ-"זר" ונמנעת מלתקוף את מי שמזוהה על ידה כ-"עצמי". התגובה שלה איננה ספציפית לסוג הפתוגן שמזוהה על ידה.

**מערכת החיסון הנרכשת** מתפתחת במהלך חיי האורגניזם עקב חשיפתו לפתוגנים, והיא מתייחדת בכך שרכיביה פועלים באופן בררני וספציפי: כל תא או מולקולה המשתייכים אליה מותאמים לפעילות כנגד פתוגן מסוים אחד. המאפיין המרכזי המייחד את מערכת החיסון הנרכשת הוא היכולת לייצר זיכרון חיסוני. בניגוד לתגובה החיסונית המולדת, שהיא זהה באופן עקרוני בקרב כל הפרטים המשתייכים למין מסוים, התגובה החיסונית הנרכשת משתנה מפרט אחד באוכלוסייה למשנהו בהתאם לניסיון האישי שרכש, כלומר בהתאם לפתוגנים אליהם נחשף במהלך חייו. מערכת החיסון הנרכשת כוללת את תאי B ו-T ואת הנוגדנים.

## תאי B

תאי B הינם סוג של תאי דם לבנים, שמהווים חלק חשוב במערכת החיסון הנרכשת.

תאי B נלחמים בפתגונים באמצעות יצירה של נוגדנים (antibodies). נוגדנים הם חלבונים המבוטאים על פני קרום התא של תאי B, ומסוגלים להיקשר לפתוגנים שמזוהים ע"י התא. הנוגדנים שעל תאי B מכונים B Cell Receptors או בקיצור BCR. הדרך שבה נוגדנים נקשרים לפתוגנים היא ע"י היקשרות לאנטיגנים. אנטיגנים הם מולקולות שנמצאות על פני קרום התא של פתוגנים (וגם של תאי הגוף) ומשמשים כמעין "תעודת זהות" שלהם. הפתגונים משתמשים בהם לזיהוי עצמי, כדי שלא יתקפו תאים נוספים מסוגם, אולם בהקשר שלנו, הנוגדנים של תאי B משתמשים בהם כדי להיקשר לפתוגן.

כל תא B מייצר נוגדנים ספציפיים לאנטיגן של פתוגן מסוים, נוגדנים שהינם בעלי אפיניות (זיקה, או יכולת היקשרות) גבוהה לאותו האנטיגן. כאשר הנוגדנים נקשרים לאנטיגן, הם מביאים לרצף של תהליכים חיסוניים שמסתיים בהשמדת הפתוגן.

### התפתחות תאי B

לתאי B מס' שלבים בהתפתחות שלהם:

1. נולדים במוח העצם, שם עוברים תהליך שנקרא **VDJ Recombination** בו הגן Ig שאחראי על היצירה שלBCR**,** עובר רקומבינציה סומטית (תהליך בו גן ארוך נחתך, ומתקבל גן קצר יותר שמורכב מכמה חתיכות של הגן המקורי), מה שמוביל ליצירה של גן ייחודי לתא.
2. במפגש עם פתוגן, תא שמצליח להיקשר לאנטיגן של אותו הפתוגן, מעורר תגובה חיסונית.
3. במסגרת התגובה החיסונית, יחד עם סיוע של תאי T התא שהצליח להיקשר לאנטיגן מובל למרכזי רבייה המוכנים **Germinal Centers**,ועובר שלב של **Clonal Expansion**, תהליך של שכפול התא בצורה מואצת.
4. **Somatic Hypermutation (SHM)**: במהלך התפתחות המושבה (Clone) ב-Germinal Center, נוצרות מוטציות רבות בגן Ig, מה שמוביל ליצירה של רפרטואר גדול של נוגדנים.
5. **Affinity-Driven Selection**: תאים ב-Clone שנוצר שמציגים אפיניות גבוהה של הנוגדנים שלהם לאנטיגן שזוהה, מקבלים העדפה אבולוציונית, וכך נוצרת מושבה של תאים עם אפיניות הולכת וגדלה לאורך זמן. במקביל לכך, תאים שמציגים מוטציות אוטואימוניות עוברים המתה.
6. Diagram

   Description automatically generatedתאי ה-B שמשלימים את התהליך, עשויים להפוך בשלבים הבאים לתאי זיכרון או לתאי לפזמה. תאי הפלזמה מסוגלים לייצר כמויות גדולות של נוגדנים בעלי אפיניות גבוהה לאנטיגן. תאי הזיכרון יפעלו אם הגוף ייפגש שוב עם האנטיגן המדובר, ואז תאי הזיכרון יוכלו להתחלק גם הם לתאי לפזמה וכך להיאבק נגדו.

איור 1: תהליך ההתפתחות של תאי B (מתוך: georgiou, g. *et al.* the promise and challenge of high-throughput sequencing of the antibody repertoire (2014))

## Somatic Hypermutation (SHM)

במהלך תהליך ההתבגרות של תאי B, אחד השלבים החשובים בפיתוח האפיניות של הנוגדנים לאנטיגן הספציפי שאליו הם צריכים להיקשר הוא ה-Somatic Hypermutation (SHM).

להבנתו נקדים: כידוע, החומר התורשתי בגופנו מורכב מ-DNA. כל מולקולת DNA בתאים השונים מורכבת מארבע תת-יחידות של בסיסים חנקניים (נוקלאוטידים), מעין ארבע אבני בניין יסודיות, החוזרות על עצמן לאורכה של המולקולה בצירופים שונים. ארבעת סוגי הנוקלאוטידים הללו הם אדנין (A), גואנין (G), תימין (T) וציטוזין (C).

בדרך כלל, במהלך תהליכים של מיטוזה (רבייה של תאים בתוך הגוף) ובתהליכים נוספים, ה-DNA של התא סופג שגיאות לאורך קטעי ה-DNA, שגורמות למוטציות ב-DNA של התא. בתא יש אנזימים האחראיים על תיקון שגיאות אלו, ובדרך כלל הן מגיעות לשיעור של מוטציות לנוקלאוטיד, בכל מעבר בין תא אב לתא בן שנוצר ממנו במיטוזה. אמנם, בתאי B מס' המוטציות שנוצרות מגיע לשיעור גבוה פי כמה: בין ל- יותר מוטציות מאשר בתא רגיל. המגוון הזה נוצר עקב פעילות שונה של מנגנוני תיקון השגיאות: כאשר האנזימים המעורבים בתהליך תיקון השגיאות ניגשים לתקן את המוטציות שנוצרות בתא באופן ספונטני, הם יוצרים בכל תיקון, כמה מוטציות נקודתיות במרחב שסביב לנוקלאוטיד שתוקן. תהליך זה, יחד עם גורמים נוספים גורם למס' גבוה של מוטציות בגן Ig, שאחראי על יצירת הנוגדן האופייני לתא.

התהליך של SHM תורם באופן מכריע להתפתחות התגובה החיסונית הנרכשת: המוטציות גורמות לגיוון גבוה בנוגדנים שתאי ה-B השונים מייצרים, ובשלבים הבאים של ה-Affinity Maturation, התאים שמציגים את האפיניות הגבוהה ביותר של נוגדניהם לאנטיגן, זוכים ליתרון אבולוציוני ומתרבים.

מחקרים שנעשו על תהליך הSHM הראו שהמוטציות בו אינן נוצרות בהתפלגות אחידה, אלא באופן שמשתנה כתלות בסוג הנוקלאוטיד, בנוקלאוטידים שמצדדיו, ובמיקומו על פני הגן (כמובן שיכול להיות שיש עוד פרמטרים שמשפיעים). במחקרים שנעשו (בעיקר על ידי בנייה של מודלי מוטציות, ראו להלן) התגלה שישנן **נקודות חמות** ו**נקודות קרות**. נקודות חמות הם צירופים של נוקלאוטידים בסדר מסוים, בהם הסיכוי למוטציה גדל משמעותית. נקודות קרותהם צירופים של נוקלאוטידים שבדרך כלל לא יעברו מוטציות בצורה משמעותית.

### המנגנון העומד מאחורי SHM

אחד ההסברים לאותם נקודות וחמות וקרות הגיע מזיהוי של האנזים המעורב בתהליך ה-SHM. התברר שבתהליך מעורב האנזים המכונה AID, שיוצר מוטציות שמחליפות את הנוקלאוטידים 𝐶→𝑈. אם המוטציה לא מתוקנת עד שכפול התא, מתקבלת בסופו של דבר מוטציית 𝐶→𝑇. זיהוי האנזים עם תהליך ה-SHM מצביע על גורם שיכול להביא למוטציות מוגברות בנוקלאוטידים של C, כפי שאכן זוהה במחקרים שבדקו נקודות חמות.

בנוסף למוטציות נקודתיות של AID, המוטציות הללו יכולות לגרום לתופעת שרשרת שגורמת למוטציות נוספות בסביבת הנוקלאוטיד C, זאת הודות ל-“**error prone repair**", מנגנון תיקון שגיאות מועד לשגיאות.

תהליך תיקון השגיאות יכול להמשיך באחד מארבעה דרכים (ראו תרשים להלן):

1. **Replication**: לאחר המעבר לא יהיה תיקון של השגיאה לפני השכפול הבא של התא. במקרה כזה ה-strand התחתון ישוכפל באופן תקין ליצירת צמד ואילו ל-strand העליון יותאם A ל-U ובשכפול הבא נקבל כבר רק T מול A. **תהליך זה יגרום למוטציה נקודתית**.
2. **UNG / BER**: האנזים UNG יכרות את הנוקלאוטיד U, וייווצר abasic site, מיקום ללא נוקלאוטיד. במקום הנוקלאוטיד שנפל ייכנס patch, נוקלאוטידים שיחליפו אותו, בעלי 1 מ-2 הצורות הבאות:
   1. **Short Patch** (הדוגמא שמופיעה בתרשים): נוקלאוטיד יחיד שיחליף את ה-abasic site. **תהליך זה יגרום למוטציה נקודתית**.
   2. **Long Patch** (לא מופיע בתרשים): הנוקלאוטיד החסר יגרום ליצירה של שרשרת נוקלאוטידים החל מאותו הבסיס והלאה במורד הזרם, והשרשרת הזו תחליף גם כן את הנוקלאוטידים התקינים שנמצאים במקומות הללו. **תהליך זה יגרום למוטציות בכל האזור.**
3. **MMR**: שרשרת תיקון שגיאות שגורמת למוטציות בכל האזור, כולל במרחק גדול ממנו. **תהליך זה יגרום למוטציות גם במרחקים גדולים ממוקד המוטציה הראשונית**.
4. **תהליך תקין של BER / MMR**: יגרום לתיקון המוטציה ללא מוטציות אזוריות אחרות.

Diagram

Description automatically generated

## מודלי מוטציות

בעבודת המחקר שלנו אנו בוחנים את תהליך ההתפתחות המוטציות בתאי B בשלב ה-SHM.

תהליך ה-SHM איננו תהליך דטרמיניסטי; המוטציות בו נוצרות באופן סטוכסטי, עם התפלגות מסוימת. במהלך השנים, המחקר פיתח עניין בפיתוח תיאור סטטיסטי מוצלח של התהליך. לתיאור כזה יכולה להיות חשיבות הן עבור ההבנה הביולוגית של התהליך, והן עבור יכולת זיהוי התנהגויות אנומליות במסגרת התהליך, ובאמצעות כך תרומה למחקר על מחלות מסוימות ופיתוח אפשרי של יכולת אבחון מוקדם עבורם. נוסף על כך, יתכן שיהיה ניתן לנצל את ההבנה של האלגוריתם על פיו הגוף שלנו מפתח מענה למחלות חדשות, כדי לפתח חיסונים בצורה מיטבית.

לצורך הבנת התהליך, פותחו מספר "**מודלי מוטציות**". אלו הם מודלים הסתברותיים, שמבוססים על high-throughput sequencing של mRNA או DNA מהגן Ig של תאי B. המודלים הללו מתארים בצורה סטטיסטית את התפלגות המוטציות שנוצרות ב-SHM, כתלות בגורמים השונים: מיקום בגן, רצף הנוקלאוטידים, ועוד.

מודלי המוטציות בדרך כלל מורכבים מ-2 חלקים:

1. מודל Targeting – היכן יש מוטציות.
2. מודל substitution– לאיזה נוקלאוטיד מתחלף הנוקלאוטיד בו הייתה המוטציה.

### מודלי מוטציות - אתגרים

ישנם כמה שיקולים שצריך לשים אליהם לב בבנייה של מודל מוטציות.

**שיקולים טכניים:** מודלים אלו בדר"כ מורכבים ממס' גבוה מאוד של פרמטרים, ולכן דורשים כמות גדולה של דאטה. בנוסף לכך, יש לסנן ריצופי mRNA/DNA בהם יש שגיאות או אי וודאות בריצוף כדי למנוע מצב בו סופרים שגיאות ריצוף כמוטציות.

**אפקט הברירה:** במהלך ההתבגרות, תאי ה-B עוברים שלבים של ברירת תאים ספציפיים שישרדו על פני אחרים שימותו, ע"פ האפיניות שהתא מציג לאנטיגן אליו נחשף הגוף וע"פ בדיקה של אוטואימוניות. התהליכים הללו משפיעים על פרופיל המוטציות שמתקבל לבסוף ברפרטואר שרוצף, ובדרך כלל מטרת מודלי המוטציות איננה למדל אותם. מידת ההשפעה של תהליך הברירה על איכות המודל איננה מובהקת, והיו מחקרים שניסו למנוע את השפעתו על המודל בצורות שונות:

1. **ניתוח רצפים לא פונקציונליים:** אלו רצפים שאינם מבוטאים לחלבון (אינטרונים) ולכן גם לא משפיעים על האפיניות. בעקבות כך, תהליך הברירה לא יוצר הטיות בהתפתחותם. סיבות אפשריות לאי-פונקציונליות הינם:
   1. Frameshift (כתוצאה מ-indels, תוספות או מחיקות של בסיסים בתווך שבין חלק ה-V וה-J של הגן)
   2. Stop Codon
   3. מוטציה ב-constant region

אתגר בשיטה זו הוא לשים לב מתי הרצף נהיה לא פונקציונלי – לפני או אחרי היווצרות המוטציות. לצורך כך ניתן להיעזר באלגוריתמים של עצי שושלות בתוך clone.

1. **מוטציות שקטות:**  מוטציות שקטות לא משפיעות על החלבון שנוצר, ולכן גם לא מושפעות מתהליך הברירה. החיסרון בשיטה זו היא שניתן לבחון כך רק מוטציות מסוג מסוים (צירופים מסוימים של נוקלאוטידים לעולם לא יעברו מוטציה שקטה).
2. **שימוש בכמות מספקת של דאטה**: ישנם חוקרים שהניחו שההשפעה של אפקט הברירה מתמצעת לאפס, אם משתמשים בכמות מספקת של דאטה.

# סקירה: מודלים שפותחו במחקר

## מודל ראשון: S5F (2013)

מודל שהציג המנחה שלנו, פרופ' גור יערי, במאמר משנת 2013. נחשב מודל חשוב ומצוטט. מכונה במודל הS5F היו 2 חידושים עיקריים:

1. המודל עשה ניתוח של פרופיל המוטציות כתלות ב-5-mer.

* **5-mer** הוא רצף של 5 נוקלאוטידים הכולל את הנוקלאוטיד שעבר מוטציה, יחד עם 2 הבסיסים שמימינו והשניים שמשמאלו. כך המודל בדק את השפעת נוקלאוטידים מימין ומשמאל, על מוטציה הנוקלאוטיד האמצעי.

1. ההתמודדות עם אפקט הברירה: המודל התבסס על חישוב ישיר של מוטציות שקטות, ובשיטות של Inferring כדי לנתח 5-mers (להלן: "מוטיבים") בהם לא הייתה מידה מספקת של מוטציות שקטות.

בבסיס המודל עמדה ההנחה שמוטציה בנוקלאוטיד כלשהו מושפעת מהנוקלאוטידים שמצדדיו. השימוש ב 5-merהביא ליצירה של מודל עם 1024 () פרמטרים בהתאם למס' המוטיבים (5-mers) האפשריים, עבור מודל ה-Targeting ופי 3 (3072) פרמטרים למודל ה-Substitution.

### השלמת נתונים

כאמור, המודל ניתח רק מוטציות שקטות. אמנם, כדי ליצור מודל נדרשה כמות מספקת של מוטציות לכל נוקלאוטיד (שתהיינה מוטציות של A הן ל-C, הן ל-G והן ל-T), ובנוסף לכך שהמוטציות הללו יהיו שקטות. הדרישות הללו גרמו לכך שבמאמר שפורסם, דווח על הגעה לערכים ישירים רק ל-307 מתוך 1024 מוטיבים אפשריים. כדי להתמודד עם הבעיה, עבור 5-mer חסרים, במודל ה-Substitution חושבה המוטביליות במידת הצורך על פי ה-3-mer הפנימי, ובמודל ה-Targeting על פי 3 הנוקלאוטידים השמאליים או הימניים מבין ה-5, באופן שהיה תלוי בנוקלאוטיד האמצעי.

## מודל שני: (2020)N. Spisak et al.

מודל נוסף פותח על ידי קבוצה של חוקרים מצרפת. מודל זה שילב את התלות ב-5-mer שהופיעה במודל ה-S5F יחד עם תלות במיקום לאורך הגן. הפרמטרים של המודל לא חושבו באמצעות "ספירה" של מוטציות כמו במודל ה-S5F אלא באלגוריתם של maximum likelihood ופתרון של בעיית קיצון בשיטה איטרטיבית.

## מודל שלישי: 2PM (דניאל עיני, 2023?)

המנחה שלנו בפרויקט, דניאל עיני, פיתח מודל מוטציות שנבנה באופן מתאים לידע הנוכחי על התהליך הביולוגי שעומד מאחורי ה-SHM. המודל ייחודי בהתאמה בינו לבין הידע הביולוגי, בדגש על המסלולים השונים של ה-SHM, בניגוד למודלים קודמים שלא הבדילו ביניהם.

המודל מורכב מ-2 שלבים:

1. הסיכוי למוטציה התחלתית ע"י האנזים AID (מוטציה נקודתית ).
2. התהליך אליו תוביל המוטציה הזו בהתאם לאחד מ-4 התהליכים שתוארו לעיל (ראו "המנגנון העומד מאחורי SHM").

המודל מנסה לשער הן את התפלגות הסיכויים לשלב ה-1 כתלות בהקשר הנוקלאוטידים, והן את ההתפלגות של המסלולים השונים של ה-SHM יחד עם הפרמטרים המאפיינים כל מסלול (כמו אורך ה-Long Patch).

המודל נבנה בשיטה של "למידה מבוססת גרדיאנט", ומנסה להביא ל-maximum likelihood של ה-observed mutations בדאטה עליו הוא מאומן.

# חלק א' של הפרויקט: השפעת אפקט הברירה על בניית מודל Substitution ל-SHM

מטרת עבודה זו הייתה למצוא את השפעת אפקט הברירה על מודל Substitution של הSHM במצב בו יש לנו הרבה דאטה, כדי לעשות זאת רצינו למצוא את מקדם הקורלציה בין תוצאות מודל ה substitution שנתון להשפעת אפקט הברירה, לבין תוצאות מודל substitution שאינו נתון להשפעת אפקט הברירה.

רצינו לבדוק זאת כדי לדעת האם נוכל להתעלם מאפקט הברירה, דבר זה יוכל לקרות אם האפקט מתמצע לאפס בכמויות דאטה שבבסיס הנתונים שלנו. קיווינו למקדם קורלציה של מעל 95% כדי שנוכל להרשות התעלמות מאפקט הברירה בהמשך הפרויקט, דבר שיוכל להקל מאוד על החישובים בהמשך.

כדי לעשות זאת השתמשנו בשני קונוטציות שונות של מודל הS5F:

1. שימוש במודל הS5F כאשר הוא מתייחס לכלל המוטציות הנצפות בדאטה.
2. שימוש במודל הS5F כאשר הוא מתייחס רק למוטציות השקטות הנצפות.

* ההנחה המקובלת היא שמוטציות שקטות לא נתונות לאפקט הברירה, ולכן תוצאות מודל הS5F בקונוטציה בה הוא מתייחס רק למוטציות שקטות, ייתן תוצאות שאינן נתונות להשפעת אפקט הברירה.

## תהליך העבודה

בניסוי השתמשנו בכלי “[createSubstitutionMatrix](https://shazam.readthedocs.io/en/stable/topics/createSubstitutionMatrix/)” שבחבילת “shazam” ב-R. עבדנו על דאטה שאסף המנחה שלנו, מר דניאל עיני, שהכיל למעלה מ-2.2 מיליון שורות של דאטה. המודל התבסס על השוואת המוטציות בין הגנים לבין האבות שהותאמו להם בעץ יוחסין שבנה המנחה שלנו גם כן.

לבסוף, חישבנו את מקדם הקורלציה בין הפרמטרים שחושבו בשני המודלים (התפלגות ה-substitution לבסיסים השונים עבור כל מוטיב, לכל מוטיב היו 3 פרמטרים שמייצגים את ההסתברות למוטציה (בנוקלאוטיד האמצעי ב5-mer) לכל נוקלאוטיד, למעט הנוקלאוטיד שהיה בהתחלה, ערך הפרמטרים הללו נע בין 0 ל-1 והסתכמו ל-1 ביחד). סיכמנו את תוצאות הקורלציה ע"פ ההרצות השונות (ראו בהמשך) בטבלה.

### אתגרים

ישנן שתי סוגיות עליהן נזקקנו לתת את הדעת בהשוואה שביצענו:

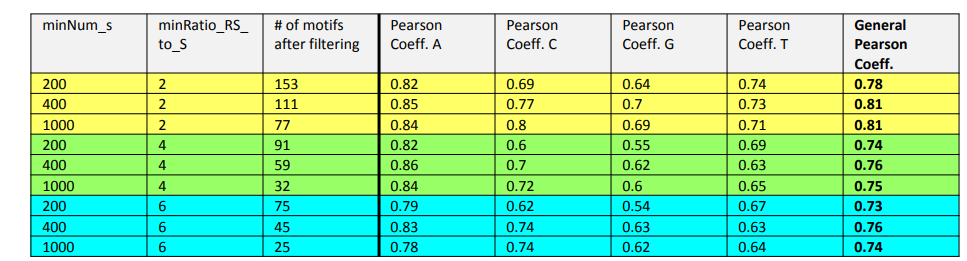
* כמות המוטציות השקטות בכל מוטיב: ראשית, בחלק מהמוטיבים לא היה מספיק דאטה של מוטציות שקטות. לעיתים אף לא היה דאטה בכלל. כדי להתמודד עם זה, השתמשנו בפונקציה של מודל ה-S5F שמייצאת את מספר המוטציות שנצפו עבור כל מוטיב. בקוד שכתבנו הפרמטר ששלט בתנאי זה הוא "minNum\_s", שתיאר את מספר המוטציות השקטות המינימלי שדרשנו בכל מוטיב. מוטיבים שלא עמדו בתנאי סוננו על ידנו.
* שקטות מול – "שקטות + לא שקטות": לא הייתה לנו אופציה להריץ את המודל כך שיתבסס על מוטציות שאינן שקטות בלבד, אלא רק כך שיתבסס על מוטציות שקטות בלבד, או שקטות ושאינן שקטות גם יחד. כאשר משווים בין שתי האפשרויות הללו, מקבלים **תלות מובנת** בין התוצאות, שכן המודל המבוסס על ה-"שקטות + לא שקטות" מבוסס בפרט על אותו הדאטה כמו המודל של ה-"שקטות", מה שגורם למקדם קורלציה גבוה יותר.   
  כדי להקטין את התלות סיננו את המוטיבים ע"פ דרישה שהמודל ה-"שקט + לא שקט" יורכב ברובו הגדול ממוטציות שאינן שקטות. לצורך כך הגדרנו פרמטר של **"**minRatio\_RS\_to\_S", ששווה ליחס המינימלי בין מס' המוטציות ה-"שקטות+לא שקטות" למס' המוטציות שקטות שנצפו בכל מוטיב. ככל שהפרמטר היה גבוה יותר, משמעות הדבר היא שרוב המודל ה-"שקט+לא שקט" התבסס על מוטציות שאינן שקטות, כפי שרצינו. כאשר הפרמטר היה נמוך, קיבלנו קורלציה גבוהה משמעותית מכפי שהיא באמת.

נציין כי דאגנו להתעלם מרצפים שהמודל איננו מחשב באופן ישיר אלא מבצע עבורם inferring, ע"פ התבססות על 3-mer, 1-mer וכו'. עשינו זאת ע"י הרצת המודל ללא תנאים מינימליים על מס' ההופעות של המוטיבים השונים, וסינון התוצאות בעצמנו. בנוסף התעלמנו ממוטיבים בהם לא נצפו מוטציות לכל אחד מ-3 הבסיסים האפשריים, כיוון שגם עבורם המודל עושה inferring.

לאחר מציאת המוטיבים שעומדים בתנאים, סיננו את תוצאות 2 המודלים על פיהם (כך שכל מוטיב שסונן על פי תנאים הקשורים למודל האחד, סונן גם בשני), והשוונו בין התוצאות עבור המוטיבים שנותרו.

## תוצאות

התוצאות עבור הרצות עם סינונים שונים מסוכמות בטבלה הבאה:



**הצבעים השונים** מתאימים לערך ה-minRatio\_RS\_to\_S. ככל שהערך של הפרמטר יותר גדול (צבע תכלת), ישנה פחות תלות מובנת בין 2 המודלים שהושוו, ומקדם הקורלציה מתאר יותר טוב את המתאם ביניהם.

העמודות: **Pearson Coeff. A, C, G, T** מתארות את מקדם הקורלציה בין הפרמטרים שהתאימו לרמת ה-substitution ל-A, C, G, T בהתאמה. **General Pearson Coeff** הוא מקדם המתאם של פירסון עבור ארבעת הפרמטרים.

אבחנות:

* **minNum\_s:** ראינו קודם כל שישנם הבדלים גדולים בין לקיחת דרישה של לפחות 200 מוטציות לבין לפחות 400 מוטציות בכל מוטיב (minNum\_s), אולם בין 400 ל-1000 ישנה כבר התכנסות, מה שמעיד על כך שלצורך הבדיקה שלנו מינימום של 400 מוטציות שקטות בכל מוטיב מספיק כדי לקבל תוצאה קרובה למדי.
* מקדם הקורלציה יורד עם עליית minRatio\_RS\_to\_S, כצפוי. הירידה הינה משמעותית בעיקר במעבר בין הערך 2 (לפחות 50% מוטציות לא שקטות במודל ה-"שקט+לא שקט") ל-4 (75%).
* הבדלים בין מקדמי הפירסון השונים: מקדם הפירסון של A הוא הגבוה ביותר בסביבות ה-0.8 בהרצות עם סינון גבוה, ושל G,T הנמוכים ביותר ברוב ההרצות המהימנות: בסביבות ה-0.6-0.65. הדבר אולי מצביע על כך שה-substitution ל-A מושפע הכי פחות מתהליך הברירה, וה-substitution ל-G,T הכי הרבה.

## השוואת התוצאות עבור shallow clones ורצפים עם מעט מוטציות

העלנו השערה שההשפעה של תופעת הברירה יכולה להיות מסוננת אם נבחן רצפים שעומדים בשני תנאים:

1. לאחר חלוקה ל-clones, רצפים החברים ב-clone עם מעט רצפים ב-clone (דרשנו לכל היותר 3)
2. רצפים בעלי מספר מוטציות נצפות גדול מ-10.

שיערנו שיתכן ו-clones כאלה הם יותר "צעירים", ולכן עברו פחות אפקטים של ברירה לאורך התפתחותם.

לאחר הרצת תהליך דומה לזה שביצענו בהרצות הקודמות, קיבלנו את התוצאות הבאות:

Table

Description automatically generated

לא הסתכלנו על תנאים יותר מחמירים כי אם היינו עושים זאת היינו מקבלים מספר קטן מדי של מוטיבים, ובמצב זה התוצאות שלנו לא היו נאמנות.

ניתן לראות לפי הטבלאות שהסינון שביצענו כן נתן קשר יותר חזק בין שני המצבים, אך לא נתן קשר מספיק חזק כדי שנוכל להניח שהברירה אינה משפיעה.

מסקנות

מקדם הקורלציה אליו הגענו לא היה גבוה מספיק, לא הצלחנו להגיע לקורלציה של 95%.  בנוגע לשימוש ב-clones קטנים וברצפים עם מעט מוטציות – ניתן לראות שיפור במקדם הקורלציה, אולם גם הוא לא עולה מכל 0.85 בתנאי סינון מחמירים.

הסיבה לכך שלא הגענו למקדם קורלציה טוב יותר יכולה להיות אחת משתיים:

* **לא השתמשנו במספיק דאטה**: ישנה אפשרות שכמויות גדולות יותר של דאטה יניבו שיערוך מדויק יותר של פרמטרים המודל השקט, ו-"ימצעו לאפס" את השפעות אפקט הברירה במודל שאיננו שקט.   
  כדי לדון באפשרות זו יש לבדוק את קצב ההתכנסות של פרמטרי המודל, ולראות האם הדאטה שלנו מספיק לכך. יש לציין, כי המודל הרועש הורכב מ**הרבה מאוד דאטה**. מדובר בכלל המוטציות שנצפו ב-2.2M רצפים. בנוגע למודל השקט – לא ראינו שדרישות מחמירות למס' המוטציות השקטות שנצפו (1000 בהשוואה ל-400) משפרות את מקדם הקורלציה, ונראה שהוא מתכנס במספרים אלו לערך קבוע.
* **אפקט הברירה לא "מתמצע לאפס"**: ככל הנראה, הפרמטרים של מודל שמושפע מאפקט הברירה אינם מתמצעים לפרמטרים של מודל שאיננו מושפע מאפקט הברירה, ומקדם המתאם בין שני המודלים לא יגיע לערכים קרובים ל-1.

# חלק ב' של העבודה: השוואה בין מודלים שונים

## פיתוח שיטה להשוואה בין מודלים

בחלק זה של העבודה, רצינו לבצע תהליך של train ו-test עבור המודלים שתיארנו לעיל, ולהשוות את ביצועיהם.

לצורך כך נדרשנו ראשית כל לשאלה: כיצד נכון לבחון את המודלים? הבחינה צריכה להיות זהה לכלל המודלים, כדי שלא לתת לאף מודל עדיפות. בפרט, שיטת הבחינה נדרשת להתאים את עצמה למודלים שמניחים הנחות שונות לגבי ההתפלגות של ה-SHM, שנכתבו בשפות תכנות שונות, ונותנים סוגי פלט שונים.

### גישה ראשונה: חישוב רמת המוטביליות לפי מיקום; גישה מקובלת במחקר

גישה אחת שנעשה שימוש בהיקף מצומצם במאמר של ה-S5F, ובהיקף רחב יותר במאמר של N. Spisak *et al.* הייתה בדיקת יכולת החיזוי של רמת המוטציות כתלות במיקום. בשיטה זו נלקח מערך נתונים שנועד ל-test, ופוצל ל-germlines שונים. לאחר מכן, נספרו המוטציות לאורך הרצפים שמקורם הוא באותו ה-germline, ונבנה גרף של המוטביליות הנצפית ("observed") כתלות במיקום. גרף זה הושווה לגרף אחר של רמת המוטציות הצפויה ("expected") לכל מיקום, ע"פ המודל שנבחן: עבור מודל S5F הדבר נמדד לפי הקשר ה-5-mer, עבור N. Spisak *et al* לפי ההקשר והמיקום ברצף, וכו'. בשלב האחרון נמדדה רמת המתאם הלינארי (מקדם פירסון) בין שני הגרפים.

דוגמא לגרף כזה ניתן לראות בתרשים הבא:

Chart, line chart, histogram

Description automatically generated

יש לשיטה זו מספר חסרונות: השיטה דורשת הפרדה לפי germline ולא יכולה להיות מחושבת על כל הדאטה ביחד. הפרדה זו גורמת לכך שהבחינה נעשית בכל פעם על חלק קטן של הדאטה (germline יחיד). בנוסף לכך, היא לוקחת השוואות עבור כל מיקום בפני עצמו, ומשלבת את ההשוואות יחד באמצעות מקדם פירסון. לשימוש במקדם פירסון כאן אין נימוק ביולוגי או סטטיסטי, שכן מקדם פירסון מניח לינאריות. לבסוף, הניתוח של רמת המוטציות לפי מיקום נראה טבעי יותר למודלים שמניחים תלות במיקום מאשר מודלים אחרים.

### גישה שניה: חישוב פרמטר Likelihood מנורמל לפי כל מודל; גישה שאנחנו פיתחנו

רצינו לבצע השוואה בין המודלים לפי ההסתברות שמודלים אלו נותנים לכלל המוטציות שנצפו בדאטה של ה-test.

התחלנו עם נוסחה כללית ל-Likelihood למוטציות שנצפו. היא חושבה ע"י מכפלת הסיכוי למיקום בו הייתה כל מוטציה, בהינתן שמוטציה התרחשה היכן שהוא לאורך הרצף:

*הערך הנ"ל חושב בתור רמת המוטביליות שהמודל נותן למקום בו הייתה מוטציה, חלקי סכום ערכי המוטביליות שהמודל נתן לכלל המיקומים ברצף:*

*זוהי הנוסחה ההסתברותית ל-Likelihood. בחרנו לשנות אותה מעט, מטעמים שנפרט להלן. את הסכום במכנה החלפנו בממוצע; כך ביטלנו תלות באורך הרצף. בנוסף לכך, כדי לנטרל תלות בכמות האיברים במכפלה (שקשורים למס' המוטציות שנצפו, ומס' הרצפים = גודל הדאטה), הוספנו שורש של מס' המוטציות הכולל בכלל הדאטה.*

בסה"כ, ה-Score שווה לממוצע הגיאומטרי של היחס בין רמת המוטביליות במיקומים בהם היו מוטציות, לרמה הממוצעת ברצף. מודל טוב ייתן ערך גדול מ-1, ומודל גרוע ייתן ערך קטן מ-1. הסף של Score = 1 מתקבל עבור "השערת האפס" – מודל שמניח רמת מוטביליות זהה לכל המיקומים ברצף.

מודלים שיקבלו ציון גבוה מאחרים על אותו דאטה סט בפרמטר Likelihood ללא הנירמולים שביצענו, יקבלו ציון גבוה יותר גם ב-Score המנורמל שהגדרנו. השינויים שביצענו שומרים על מונוטוניות.

\*נציין שהוספנו לפונקציית הscore שלנו גם confidens interval , חישבנו את הconfidens interval במישור הלוגריתמי, בגלל שבמישור הלוגריתמי מכפלה הופכת לסכום, ושורש הופך לחלוקה, ובכלל ממוצע גאומטרי הופך לממוצע חשבוני, ובגלל שאנחנו עובדים עם כמות גדולה של דאטה, אז לפי משפט הגבול המרכזי, הממוצע החשבוני מתפלג נורמלית וניתן למצוא לו confidens interval בצורה רגילה להתפלגות נורמלית.

קוד המקור לחישוב ה-Score הנ"ל עבור 3 המודלים אליהם התייחסנו, מופיע בנספח.

## ולידציה לפרמטר ה-Score שהגדרנו

כדי לבדוק את יכולתו של פרמטר ה-Score שפיתחנו לבצע השוואה הוגנת בין המודלים, ביצענו תהליך של ולידציה עבורו, שמבוסס על תהליך העבודה הבא:

* 1. את האימון הראשוני ביצענו על כ-2 מיליות רצפים. אימנו את המודלים: S5F-s, S5F-rs, 2PM-s,2PM-rs, כאשר s מסמן אימון רק על מוטציות שקטות (silent), ו-rs מסמן אימון על כלל המוטציות, שקטות ולא-שקטות (replacement+silent).
     + יכולנו להכניס למודלים פרמטרים אקראיים ולהמשיך איתם את הוולידציה. בחרנו אמנם להשתמש במודלים שאומנו כולם על דאטה סט אחיד, כיוון שכך מתקבלים מודלים שאמורים להיות "קרובים" במהותם זה לזה, ורצינו לראות האם גם בין מודלים כאלה ה-Score ידע לדרג בצורה נכונה.
  2. לאחר מכן, לקחנו דאטה סט של כ-100,000 רצפים, והחלפנו את המוטציות שנצפו בכל רצף, במס' שווה של מוטציות שנקבעו באופן סימולטיבי ע"פ המודלים: S5F-s או 2PM-s.
  3. בשלב השלישי חישבנו את ה-Score עבור כל אחד מן המודלים, על שני מערכים הנתונים שנבנו בשלב הקודם.

### תוצאות הוולידציה

Table

Description automatically generatedניתן לראות שכמו שציפינו, המודל עליו הדאטה סומלץ קיבל את הscore הכי גבוה לאותו הדאטה, דבר האומר שפונקציית הscore שלנו מהמנה ועובדת בצורה טובה.

## שימוש ב-Score להשוואה על דאטה אמיתי

ראשית כל, נציין שביצענו כמה סוגי השוואות, כל השוואה בעלת train data וtest data שונים.

מערכי הנתונים שעמדו לרשותנו הם:

* מערך נתונים של כ-2.2M רצפים שעובדו ע"י המנחה שלנו, דניאל עיני. רצפים אלו היו מסודרים למבנה של עצי שושלת, וספירת המוטציות בהם הייתה ביחס לאב שלהם בעץ. הרצפים היו מיושרים לפי IMGT. את מערך הנתונים פיצלנו לשתי יחידות: "train data 1", "test data 1", ביחס של 95:5
* מערך נתונים של כ-100K רצפים לא פונקציונליים (Out of Frame), מסודרים גם הם לעצים. היישור כאן לא היה לפי IMGT, ונעשה ע"י קבוצה החוקרים מצרפת (N. Spisak *et al*). מערך זה יכונה "oof\_data".

ביצענו כמה השוואות, בצירופים שונים של train ו-test

1. **אימון על train data 1, בחינה על test data 1:** בחינה זו מושפעת מהשפעת הברירה, שכן הבחינה הינה על כלל המוטציות, וברצפים פונקציונליים.
2. **אימון על train data 1, בחינה על oof data:** בחינה זו מנוטרלת מהשפעת הברירה, אולם המודל נבנה על דאטה שכן חשוף להשפעת הברירה.
3. **אימון על oof data ובחינה על test data 1:** בחינה זו מושפעת מהשפעת הבחירה אולם המודל נבנה על דאטה שאינו חשוף להשפעת הברירה.

**אימון על train data 1, בחינה על test data 1:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Score | UPR.CI | LWR.CI | Test data | Train data | model |
| 1.1314 | 1.1339 | 1.1289 | test data 1 | Train data 1 | 2PM-s |
| 1.2185 | 1.2214 | 1.2157 | test data 1 | Train data 1 | 2PM-rs |
| 1.2816 | 1.2849 | 1.2783 | test data 1 | Train data 1 | S5F-s |
| 1.5882 | 1.5928 | 1.5836 | test data 1 | Train data 1 | S5F-rs |
| 1 | 1 | 1 | test data 1 | Train data 1 | Null |

ובbar plot:

Chart, bar chart

Description automatically generated

ניתן לראות שהמודל שנתן את ה-score הגבוה ביותר הוא מודל ה-S5F-rs. בנוסף לכך, בשני המודלים מודלי ה-rs (כלל המוטציות) קיבלו תוצאה טובה יותר ממודלי ה-s (מוטציות שקטות). תוצאה זו הגיוניות, משום שגם ה-train data וגם ה-test data מושפעים מהשפעת הברירה, ולכן אין עדיפות לשימוש במוטציות שקטות לבדן באימון המודל לצורך נטרול השפעתה.

**אימון על train data 1, בחינה על oof data:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Score | UPR.CI | LWR.CI | Test data | Train data | model |
| 1.0081 | 1.0109 | 1.0052 | Off data | Daniel data | Daniel-s |
| 1.0392 | 1.0422 | 1.0361 | Off data | Daniel data | Daniel-rs |
| 1.0778 | 1.0813 | 1.0742 | Off data | Daniel data | S5F-s |
| 1.2314 | 1.2266 | 1.2218 | Off data | Daniel data | S5F-rs |
| 1 | 1 | 1 | Off data | -- | Null |

ובbar plot:

Chart, bar chart

Description automatically generated

ניתן לראות כי קיבלנו שדווקא המודלים שאומנו על כלל המוטציות בדאטה, קיבלו score גבוה יותר ממודלים שאומנו רק על מוטציות שקטות. תוצאה זו מאוד מפתיעה משום שה-test data בבחינה זו הינו OOF, ולכן אינו נתון להשפעת הברירה בשעה שה-train data מורכב מדאטה פונקציונלי שכן נתון לה. כיוון ששימוש במוטציות שקטות באימון אמור לנטרל את השפעת הבחירה, התוצאות אמורות לצאת טובות יותר במודלים שאומנו רק על מוטציות כאלו. למרות זאת קיבלנו שדווקא המודלים שאומנו על דאטה הנתון להשפעת הברירה חזו טוב יותר דאטה שאינו נתון להשפעת הברירה.

יכולים להיות לכך כמה הסברים:

* אימון על מוטציות שקטות בלבד מוריד את כמות הדאטה עליו המודל מאומן.
* במודל ה-S5F, אימון על מוטציות שקטות דורש inferring עבור מוטיבים בהם אין מוטציות שקטות.
* אפשרות נוספת: (אפשרות שהועלתה ע"י פרופ' פופובצר) – רמת המוטביליות, כבר בשלב ה-SHM הינו שונה במוטציות שקטות.

**אימון על oof data ובחינה על test data 1:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Score | UPR.CI | LWR.CI | Test data | Train data | model |
| 1.1578 | 1.1588 | 1.1558 | Daniel data | Off data | Daniel-rs |
| 1.4334 | 1.437 | 1.4298 | Daniel data | Off data | S5F-rs |
| 1.2977 | 1.3009 | 1.2945 | Daniel data | Off data | N.Spisak et al |

ובbar plot:

Chart, bar chart

Description automatically generated

כאן ההשוואה הינה בין שלושת המודלים, כאשר הם אומנו על כלל המוטציות (בגלל שהtrain data הוא דאטה לא פונקציונלי אין סיבה לאמן רק על מוטציות שקטות כי כלל המוטציות בדאטה אינן נתונות להשפעת הברירה). אפשר לראות שמודל הS5F קיבל את הscore הכי גבוה, למרות שהוא המודל הראשוני וה"פשוט" ביותר.

# סיכום ומסקנות

החלק הראשון של עבודתנו בדק שאלה חשובה ויסודית: האם מודלים שמסננים רק מוטציות שקטות מקבלים תוצאות בעלות דמיון גבוה למודלים שלא עושים זאת. המסקנה בהקשר זה, לפחות בנוגע למודלי Substitution היא שיש שוני לא מבוטל בין התוצאות, גם כאשר מבטלים את השפעת ה-inferring. אמנם – התוצאה לא מכומתת בצורה סטטיסטית מלאה, ויתכן שכדי לקבל תמונה מלאה יש להשוות את התוצאות למידת הדמיון שמתקבלת בין דגימות מאינדבידואלים שונים. בכל אופן, התוצאה הראתה שבמחקר שלנו – לא ניתן להניח ששימוש רק במוטציות שקטות יביא את אותם התוצאות.

תוצאה זו עשויה לנבוע כתוצאה מכמות הדאטה שיורדת כאשר מסננים מוטציות שקטות. אפשרות נוספת, ומעט יותר "מהפכנית" היא שהדבר קורה משום שהגוף משנה את הסיכוי למוטציות שקטות ביחס למוטציות שאינן שקטות. נהוג להניח שמנגנון ה-SHM בלתי תלוי בשאלת השינוי שיהיה בחלבון שייווצר בפועל, אלא רק בהקשר ה-DNA או המיקום. דרך בה השפעת היות המוטציה שקטנה יכולה לבוא לידי ביטוי הוצעה ע"י פרופ' רחלה פופובצר במהלך הצגת עבודתנו בפני הסגל האקדמי של מסלול ביו הנדסה: יתכן שמיקומים מסוימים בהם מוטציות נוטות להיות שקטות, מקבלות עדיפות במהלך ה-SHM, וכך כאשר מנתחים את פרופיל המוטציות כתלות בהקשר 5-mer בלבד, יתקבלו הבדלים שקשורים לתלות במיקום.

בחלק השני של עבודתנו ביצענו תהליכים של train ו-test למודלים שונים. ההשוואות שלנו הראו 2 תוצאות עיקריות:

* 1. בכל המקרים, קיבלנו שדווקא מודלים שאומנו על כלל המוטציות, קיבלו score גבוה יותר מזה של מודלים שאומנו על מוטציות שקטות. הדבר עשוי לנבוע מהבדלים בכמויות הדאטה, מתהליכים של הסקה עבור פרמטרים שאין עבורם דאטה של מוטציות שקטות, ואולי גם מהבדלים בהתנהגות ה-SHM בנוגע למוטציות שקטות, כפי שהוצע בפסקה הקודמת.
  2. מודל הS5F קיבל את התוצאות הטובות ביותר, דבר שכנראה אומר שלמיקום ברצף אין השפעה על המוטביליות בהינתן ה5-mer (או לחילופין שהמודל של N.Spisak et al לא מכמת בצורה טובה השפעה זו). יתכן שמודל ה-2PM קיבל ציון יותר נמוך או בגלל שהידע הביולוגי שבידנו איננו שלם וממילא המודל לא מהווה תיאור שלם, בגלל אתגרים בהתכנסות של המודל (בייחוד: האפשרות שהמודל לא התכנס דווקא למינימום הגלובלי), או בגלל שהשתמשנו בגרסה של מודל ה-2PM שמכילה מעט פרמטרים.

# כיוונים להמשך

1. בעבודה שלנו, השתמשנו בגרסה בסיסית של המודל 2PM, בהמשך אפשר להשתמש בפונקציית הscore שבנינו כדי להשוות גם גרסאות מתקדמות יותר של המודל לבין המודלים האחרים (ואולי גם להשוות בין גרסאות המודל, כדי לראות שגרסה מתקדמת יותר נותנת תוצאות טובות יותר).
2. פונקציית הscore שבנינו מתאימה למודל mutability, ניתן להרחיב אותה שתתאים גם למודלים של substitution וtargeting.
3. להבין טוב יותר למה מודלים שהסתמכו על מוטציות שקטות, קיבלנו תוצאות פחות טובות. האם זה בגלל הבדלים בכמויות הדאטה? האם ניתן לאתר מוטיבים ספציפיים / מיקומים ספציפיים בהם מוטציות שקטות מופיעות בהסתברות שונה ממוטציות שאינן שקטות?
4. להשתמש בפונקציית הscore שבנינו כדי להבחין בהתנהגויות לא שגרתיות של התפלגות המוטציות בפציינטים ספציפיים, ובצורה כזו לבצע אבחון מוקדם של מחלות שונות במערכת החיסון. זאת על ידי השוואה של ה-score שיקבל מודל שאומן על דאטה שמגיע מפציינטים בריאים לעומת מודל שאומן על פציינטים המאובחנים למחלה, וקבלת תוצאה מובהקת סטטיסטית.

# נספח: קוד המקור

כלל הקודים הנדרשים להרצת המודלים יחד עם שימוש בפונקציית ה-Score, נמצאים ב-GitHub, בכתובת: <https://github.com/yedidyaShimon/shm_project_2022>. קובץ README רלוונטי מכיל הוראות לשימוש בקודים.

## קוד א': חישוב מקדם המתאם של פירסון בין מודל ה-S5F שאומן על מוטציות שקטות לעומת מודל שאומן כלל המוטציות

setwd("~/data/2")

Sub\_Merged\_Mut\_s<-read.table("Sub\_Merged\_Mut\_s\_all\_t.tsv",header = TRUE)

Sub\_Merged\_Mut\_rs<-read.table("Sub\_Merged\_Mut\_rs\_all\_t.tsv",header = TRUE)

#setting the parameteres for the filtering

minNum\_S<-200

minRatio\_RS\_to\_S<-6

print(paste("requiring ",(minNum\_S)," mutations, and ",minRatio\_RS\_to\_S," times larger amount of silent + non-silent mutations than silent alone. Number of 5-mers that comply with requirements: ",sep=""))

Sub\_Merged\_Mut\_s\_measured<-Sub\_Merged\_Mut\_s[Sub\_Merged\_Mut\_s$fivemer.every & Sub\_Merged\_Mut\_s$fivemer.total>minNum\_S & Sub\_Merged\_Mut\_rs$fivemer.total>minRatio\_RS\_to\_S\*Sub\_Merged\_Mut\_s$fivemer.total,]

Sub\_Merged\_Mut\_rs\_measured<-Sub\_Merged\_Mut\_rs[Sub\_Merged\_Mut\_s$fivemer.every & Sub\_Merged\_Mut\_s$fivemer.total>minNum\_S & Sub\_Merged\_Mut\_rs$fivemer.total>minRatio\_RS\_to\_S\*Sub\_Merged\_Mut\_s$fivemer.total,]

print(nrow(Sub\_Merged\_Mut\_s\_measured))

print(nrow(Sub\_Merged\_Mut\_rs\_measured))

list\_s\_A<-Sub\_Merged\_Mut\_s\_measured[,1:4]$A

list\_s\_A<-list\_s\_A[complete.cases(list\_s\_A)]

list\_rs\_A<-Sub\_Merged\_Mut\_rs\_measured[,1:4]$A

list\_rs\_A<-list\_rs\_A[complete.cases(list\_rs\_A)]

list\_s\_T<-Sub\_Merged\_Mut\_s\_measured[,1:4]$T

list\_s\_T<-list\_s\_T[complete.cases(list\_s\_T)]

list\_rs\_T<-Sub\_Merged\_Mut\_rs\_measured[,1:4]$T

list\_rs\_T<-list\_rs\_T[complete.cases(list\_rs\_T)]

list\_s\_C<-Sub\_Merged\_Mut\_s\_measured[,1:4]$C

list\_s\_C<-list\_s\_C[complete.cases(list\_s\_C)]

list\_rs\_C<-Sub\_Merged\_Mut\_rs\_measured[,1:4]$C

list\_rs\_C<-list\_rs\_C[complete.cases(list\_rs\_C)]

list\_s\_G<-Sub\_Merged\_Mut\_s\_measured[,1:4]$G

list\_s\_G<-list\_s\_G[complete.cases(list\_s\_G)]

list\_rs\_G<-Sub\_Merged\_Mut\_rs\_measured[,1:4]$G

list\_rs\_G<-list\_rs\_G[complete.cases(list\_rs\_G)]

list\_s<-c(list\_rs\_A,list\_s\_C,list\_s\_G,list\_s\_T)

list\_rs<-c(list\_rs\_A,list\_rs\_C,list\_rs\_G,list\_rs\_T)

print("The correlation coefficient of the substitution to A rates is:")

print(cor(list\_rs\_A,list\_s\_A))

print("The correlation coefficient of the substitution to C rates is:")

print(cor(list\_rs\_C,list\_s\_C))

print("The correlation coefficient of the substitution to G rates is:")

print(cor(list\_rs\_G,list\_s\_G))

print("The correlation coefficient of the substitution to T rates is:")

print(cor(list\_rs\_T,list\_s\_T))

print("The general correlation coefficient rate is")

print(cor(list\_rs,list\_s))

## קוד ב': חישוב ה-score להשוואה בין המודלים השונים

שלב ראשון: לבחור בתחילת הקוד את ה-“modelID”, בהתאם לסוג המודל עבורו רוצים לבחון. הספרה המתאימה לכל מודל מופיעה בתחילת הקוד.

לאחר מכן יש להזין בתוך תנאי ה-if המתאים למודל שנבחר את כתובת הקובץ/קבצים הנדרשים:

* עבור מודל מסוג S5F יש צורך להזין קובץ csv שמכיל את הפרמטרים של ה-mutability לכל 5-mer. המבנה שלו יהיה כדלהלן:

|  |  |
| --- | --- |
| x | X |
| 0.001455 | AAAAA |
| 0.001158 | AAAAC |
| 0.001242 | AAAAG |
| 0.000907 | AAAAT |
| 0.001005 | AAACA |
| 0.001019 | AAACC |
| 0.001153 | AAACG |
| 0.000519 | AAACT |

כאשר העמודה הימנית היא ה-5-mer, והשמאלית המוטביליות עבורו. שמות העמודות צריכות להיות בהתאם לדוגמא.

* עבור מודל מסוג Daniel (הכוונה ל-2PM, שפותח ע"י המנחה שלנו דניאל), עבדנו בצורה הבאה:
  + השתמשנו בקובץ קוד בשפת python, בשם calc\_mut\_vec\_for\_score.py בקובץ הנ"ל הזנו ידנית את הפרמטרים הסופיים של המודל יחד עם קובץ ה-test data שלנו, ויצרנו קובץ המכיל ווקטורים של הסתברויות לכל מיקום על פני כל אחד מן הרצפים שב-test, לפי המודל. את הקובץ הנ"ל יש להריץ כחלק מהפרויקט “two-pahse-model” של המנחה שלנו, מר דניאל עיני.
* עבור המודל (2020)N. Spisak et al , המכונה בקוד “French”, לקחנו את הקבצים הבאים: model\_all\_context.tsv, model\_all\_position.tsv. הללו מגיעים מהמחקר של N.Spisak.

#### choosing modelID ####

# S5F\_s = 1.1, S5F\_rs = 1.2, French = 2, Daniel\_s = 3.1 daniel\_rs = 3.2, NULL = 0

modelID<-1.1

#### choosing dataID ####

# data for test: 1 = 100k, 2 = french data, 3 = v-gene daniel-data-simulated-on s5f-rs,

# 4 = v-gene daniel-data-simulated-on s5f-s, 5 = v-gene daniel-data-simulated by daniel model -s,

#6 = v-gene daniel-data-real data-s

dataID = 6

#### just for checking ####

father<-"ACNTTCTTACNGGGGG"

son<-"AGATTCTGANGGGGGG"

#### libraries ####

require(stringr)

require(seqinr)

require("psych")

require("DescTools")

#### reading datasets ####

if(modelID == 1.1 | modelID == 0) {

  targeting1<-read.csv("mut\_all\_s.csv", header = TRUE)

}

if(modelID==1.2) {

  targeting1<-read.csv("mut\_french\_rs.csv", header = TRUE)

}

if(modelID==2) {

  context<-read.table(file = "model\_all\_context.tsv",header = TRUE, sep = "\t")

  position<-read.table(file =  "model\_all\_position.tsv", header = TRUE, sep = "\t")

  context<-context[,c("Motif","Mutability")]

  position<-position[order(position$Position),]

  for(i in 1:nrow(position)) {

    if(position$Mutability[i] == 0) {

      position$Mutability[i]<-min(position$Mutability[position$Mutability>0])

    }

  }

}

if(modelID==3.1 )

{

  daniel\_model\_prob<-scan("~/data/12/daniel\_probs\_for\_franch/probs\_daniel\_s\_on\_oof\_data\_daniel\_s\_simulate.csv",what="",sep="\n")

  daniel\_model\_prob<-strsplit(daniel\_model\_prob,",")

  daniel\_model\_prob<-lapply(daniel\_model\_prob,as.numeric)

  print("nrow probs: ")

  print(length(daniel\_model\_prob))

}

if(modelID==3.2)

{

  daniel\_model\_prob<-scan("daniel\_probs\_for\_franch/probs\_daniel\_rs\_train\_franch\_test\_daniel.csv",what="",sep="\n")

  daniel\_model\_prob<-strsplit(daniel\_model\_prob,",")

  daniel\_model\_prob<-lapply(daniel\_model\_prob,as.numeric)

  print("nrow probs: ")

  print(length(daniel\_model\_prob))

}

if(dataID == 1) {

  data1<-read.csv("100k\_data.csv",header = TRUE)

  data1$ALIGNMENT<-data1$sequence\_alignment

  data1$ancestorseq<-data1$ancestor\_alignment

}

if(dataID == 2) {

  data1<-read.csv("data\_all\_alignment\_updated.csv",header =TRUE)

}

if(dataID == 3) #simulation rs

{

  data1<-read.csv("test\_data\_3\_S5F\_rs\_simulate\_final.csv",header=TRUE)

  data1$ALIGNMENT<-data1$simulate\_alignment

  data1$ancestorseq<-data1$ancestor\_alignment

}

if(dataID == 4) #simulation s

{

  data1<-read.csv("test\_data\_3\_S5F\_s\_simulate\_final.csv",header=TRUE)

  data1$ALIGNMENT<-data1$simulate\_alignment

  data1$ancestorseq<-data1$ancestor\_alignment

}

if(dataID == 5) #simulation daniel s

{

  data1<-read.csv("test\_data\_3\_daniel\_s\_simulate\_final.csv",header=TRUE)

  data1$ALIGNMENT<-data1$simulated\_sequence

  data1$ancestorseq<-data1$ancestor\_alignment

}

if(dataID == 6) #real data from daniel (test\_3)

{

  data1<-read.csv("test\_data\_3\_v\_gene\_final.csv",header=TRUE)

  data1$ALIGNMENT<-data1$sequence\_alignment

  data1$ancestorseq<-data1$ancestor\_alignment

}

if(dataID == 7) #real data out-of-frame (from natanael spisak)

{

  data1<-read.csv("/home/bcrlab/giladaviv/data/shm\_oof\_french\_research/\_alignment\_updated/data\_all\_alignment\_updated\_2.csv",header=TRUE)

  data1<-data1[!is.na(data1$ancestorseq),]

}

data2<-data1[,c("ALIGNMENT","ancestorseq")]

#### filtering data ####

#data3<-data2[data2$ALIGNMENT != data2$ancestorseq, ]

data3<-data2

data4<-data3[complete.cases(data3), ]

#str\_replace

aligment<-str\_replace\_all(data4$ALIGNMENT,"[^ACGTN]","N")

anc<-str\_replace\_all(data4$ancestorseq,"[^ACGTN]","N")

datafinal<-data4

datafinal$ALIGNMENT<-aligment

datafinal$ancestorseq<-anc

#sanity check

datafinal<-datafinal[nchar(datafinal$ALIGNMENT)==nchar(datafinal$ancestorseq),]

print("nrow data: ")

print(nrow(datafinal))

#### changing to NULL HYPOTHESIS model (if modelID == 0) ####

if(modelID == 0) {

  targeting1$x<-rep(1/1024,nrow(targeting1))

}

#### Functions ####

FindTargeting<-function(seq) {

  #initializing the output vector

  g<-vector(mode="numeric",length=nchar(seq))

  g[1]=NA

  g[2]=NA

  g[length(g)]=NA

  g[length(g)-1]=NA

  #filling the rest of g

  for (i in 3:(nchar(seq)-2)){

    if(grepl("N",substr(seq,i-2,i+2)) == FALSE) {

      g[i]<-targeting1$x[targeting1$X==substr(seq,i-2,i+2)]

    }

    else {

      g[i]<-NA

    }

  }

  return(g)

}

FindTargeting\_french<-function(seq) {

  #initializing the output vector

  g<-vector(mode="numeric",length=nchar(seq))

  g[1]=NA

  g[2]=NA

  g[length(g)]=NA

  g[length(g)-1]=NA

  #filling the rest of g

  for (i in 3:(nchar(seq)-2)){

    if(grepl("N",substr(seq,i-2,i+2)) == FALSE) {

      g[i]<-context$Mutability[context$Motif==substr(seq,i-2,i+2)] \* position$Mutability[i]

    }

    else {

      g[i]<-NA

    }

  }

  return(g)

}

likelihood\_S5F<-function(son,father) {

  # calculating the expected mutation rate along father sequence

  temp<-FindTargeting(father)

  # finding mutations positions

  indexes<-((s2c(son) != s2c(father)) & (s2c(son) != "N") & (s2c(father) != "N")) %>% which()

  indexes<-indexes[(indexes>=3) & (indexes<=nchar(son)-2)]

  indexes<-indexes[!is.na(temp[indexes])]

  # creating a list of expected mutation probabilities at observed mutations locations

  thevector<-c()

  for (i in indexes)

  {

    if(!is.na(temp[i])) {

      thevector<-append(thevector,temp[i])

    }

  }

  # {???????? ????????}

  if(length(thevector)>0 & length(indexes)<20)

  {

    OurLikelihood<-thevector/mean(temp[complete.cases(temp)])

  } else {

    OurLikelihood<-NA

  }

  return(list(OurLikelihood,length(indexes)))

}

likelihood\_french<-function(son,father) {

  # calculating the expected mutation rate along father sequence

  temp<-FindTargeting\_french(father)

  # finding mutations positions

  indexes<-((s2c(son) != s2c(father)) & (s2c(son) != "N") & (s2c(father) != "N")) %>% which()

  indexes<-indexes[(indexes>=3) & (indexes<=nchar(son)-2)]

  indexes<-indexes[!is.na(temp[indexes])]

  # creating a list of expected mutation probabilities at observed mutations locations

  thevector<-c()

  for (i in indexes)

  {

    if(!is.na(temp[i])) {

      thevector<-c(thevector,temp[i])

    }

  }

  # {???????? ????????}

  if(length(thevector)>0 & length(indexes)<20)

  {

    OurLikelihood<-thevector/mean(temp[complete.cases(temp)])

  } else {

    OurLikelihood<-NA

  }

  return(list(OurLikelihood,length(indexes)))

}

likelihood\_daniel<-function(son,father,j) {

  temp<-unlist(daniel\_model\_prob[j])

  for(i in 1:length(temp))

  {

    if(grepl("N",substr(father,i-2,i+2)) == TRUE) {

      temp[i]<-NA

    }

  }

  # finding mutations positions

  indexes<-((s2c(son) != s2c(father)) & (s2c(son) != "N") & (s2c(father) != "N")) %>% which() #not N

  indexes<-indexes[(indexes>=3) & (indexes<=nchar(son)-2)] #not in the edges

  indexes<-indexes[!is.na(temp[indexes])] # no N's in the 5-mer

  # creating a list of expected mutation probabilities at observed mutations locations

  thevector<-c()

  for (i in indexes)

  {

    if(!is.na(temp[i])) {

      thevector<-append(thevector,temp[i])

    }

  }

  if(length(thevector)>0 & length(indexes)<20)

  {

    OurLikelihood<-thevector/mean(temp[complete.cases(temp)])

  } else {

    OurLikelihood<-NA

  }

  return(list(OurLikelihood,length(indexes)))

}

confidence\_interval <- function(vector, interval) {

  # Standard deviation of sample

  vec\_sd <- sd(vector)

  # Sample size

  n <- length(vector)

  # Mean of sample

  vec\_mean <- mean(vector)

  # Error according to t distribution

  error <- qt((interval + 1)/2, df = n - 1) \* vec\_sd / sqrt(n)

  # Confidence interval as a vector

  result <- c("lower" = vec\_mean - error, "upper" = vec\_mean + error)

  return(result)

}

#### running the test ####

s<-vector("list")

if(modelID == 1.1 | modelID == 1.2 | modelID == 0) {

  for (j in 1:nrow(datafinal)){

    output<-likelihood\_S5F(datafinal$ALIGNMENT[j],datafinal$ancestorseq[j])

    s<-append(s,output[[1]])

  }

}

if(modelID == 2) {

  for (j in 1:nrow(datafinal)){

    output<-likelihood\_french(datafinal$ALIGNMENT[j],datafinal$ancestorseq[j])[[1]]

    s<-append(s,output)

  }

}

if(modelID == 3.1 | modelID == 3.2)

{

  for (j in 1:nrow(datafinal)){

    output<-likelihood\_daniel(datafinal$ALIGNMENT[j],datafinal$ancestorseq[j],j)[[1]]

    s<-append(s,output)

  }

}

s<-unlist(s)

score1<-Gmean(s,conf.level = 0.95, na.rm = TRUE)

print(score1)