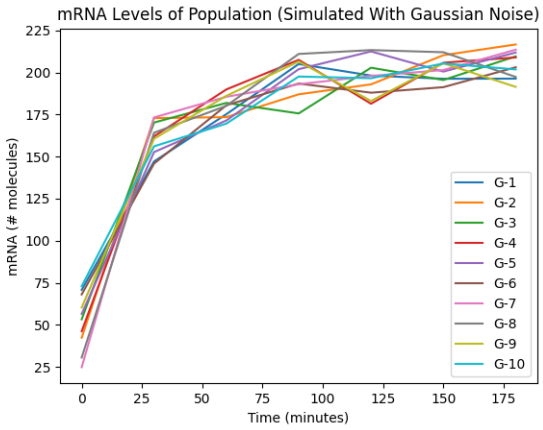
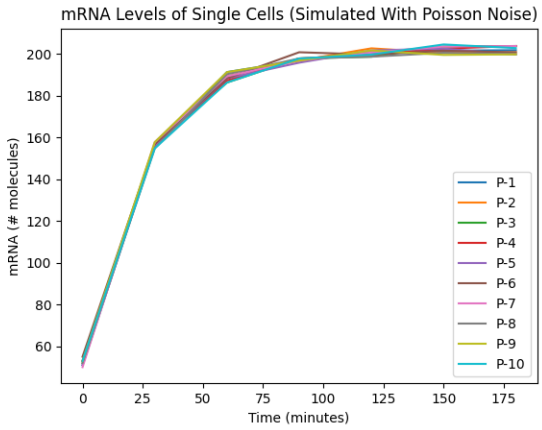
(88893) ביולוגיה מערכתית | תרגיל 4

שם: ניצן שלוי, רון מורן | ת"ז: 208649020, 206170920

שאלה 1

(א)

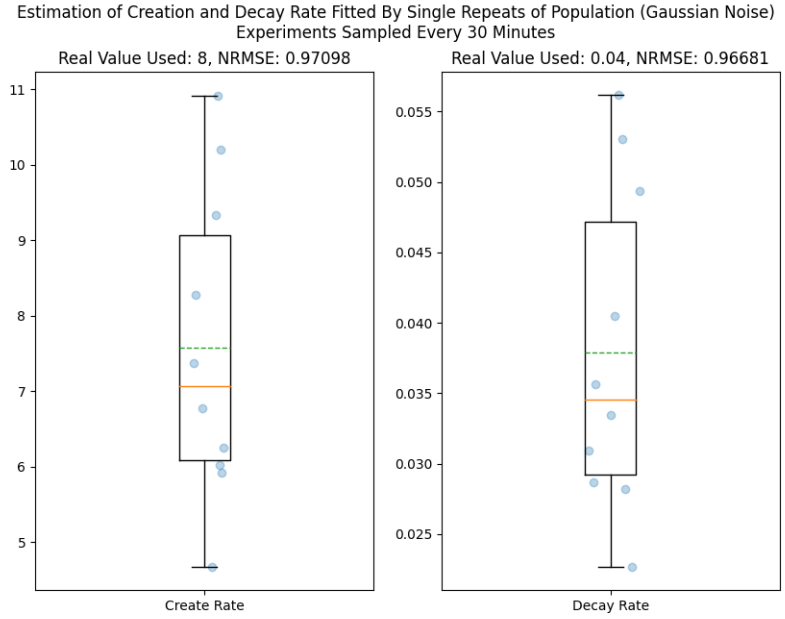


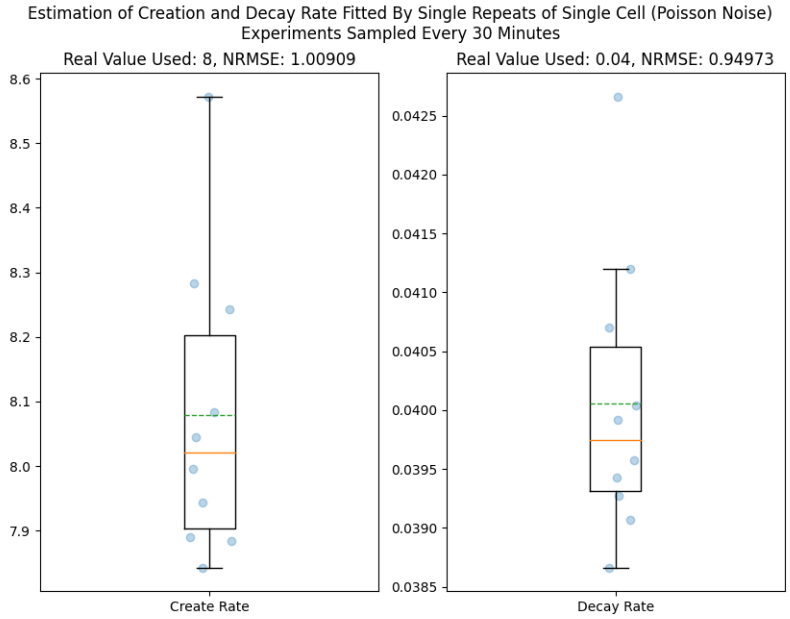


(ב)

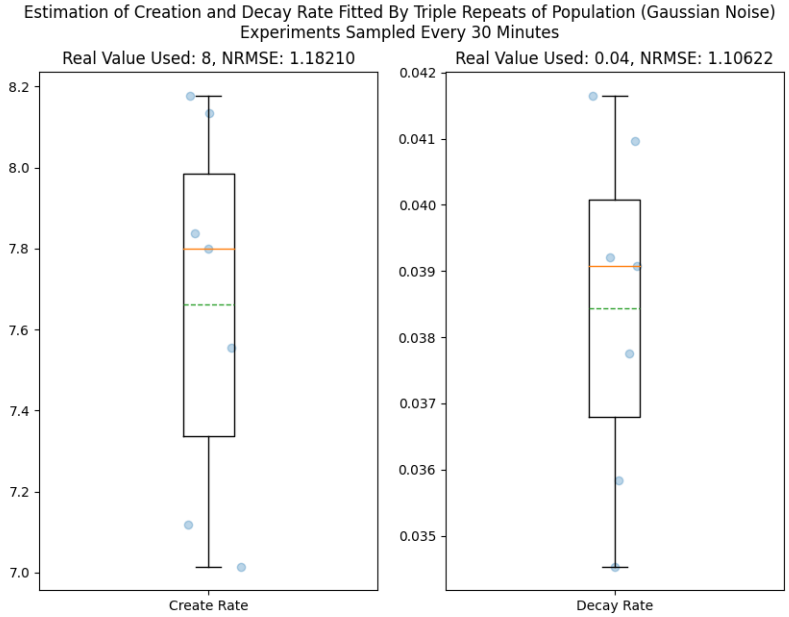
שערכנו את קצב הייצור לכל אחד מהמקרים בקוד המצורף וקיבלנו מספר אומדים לכל מקרה.
בגרפים הבאים אנחנו מציגים את התפלגות אומדי קצב הייצור והפירוק שקיבלנו ב-boxplot, בו הקו הכתום מסמל את החציון והירוק המקוקו את הממוצע. בחרנו להציג את ה-NRMSE, נרמול של שורש ה-MSE לפי סטיית התקן (מוצג יותר בפירוט בשאלה 2), מכיוון שערך ה-MSE של האומדים לקצב הפירוק יוצאים קטנים עד כדי זניחים (כי הערך עצמו כבר קטן, לכן גם המרחק של כל אמדן ממנו קטן וריבוע של מספר קטן בהרבה מ-1 מקטין אותו עוד). בנוסף יש הבדל של מספר סדרי גודל בין הקצבים אותם אנו אומדים כך שהשוואה בין ה-MSE שלהם חסרת משמעות.
המקרים הם:

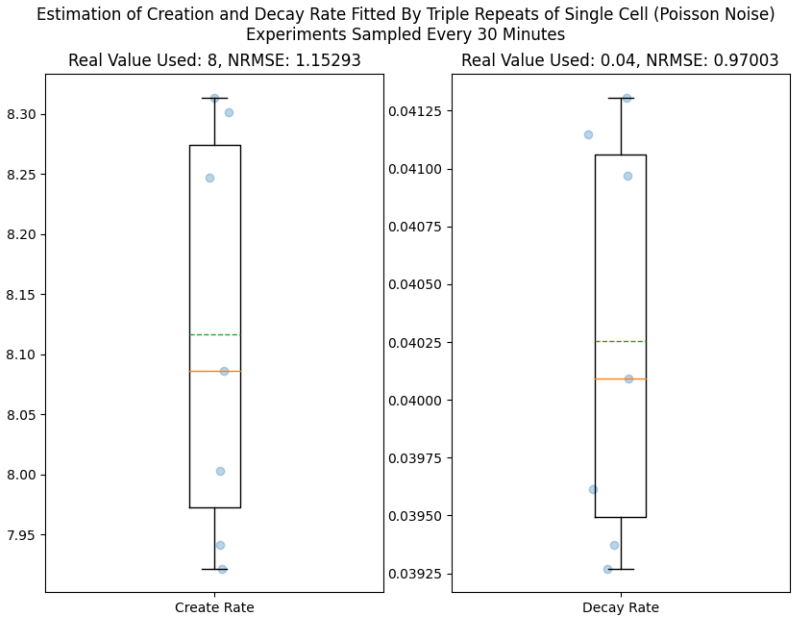
- שערוך אמדן לפי כל ניסוי בנפרד בו מודדים כל חצי שעה (סה"כ 10):





- שערוך אמדן לפי ממוצעים של שלשות של ניסויים בו מודדים כל חצי שעה (שלשות עוקבות, סה"כ 7):





(ג)

נביט בהבדלים בין האומדים (הממוצע של האומדים בכל אחד מהמקרים):

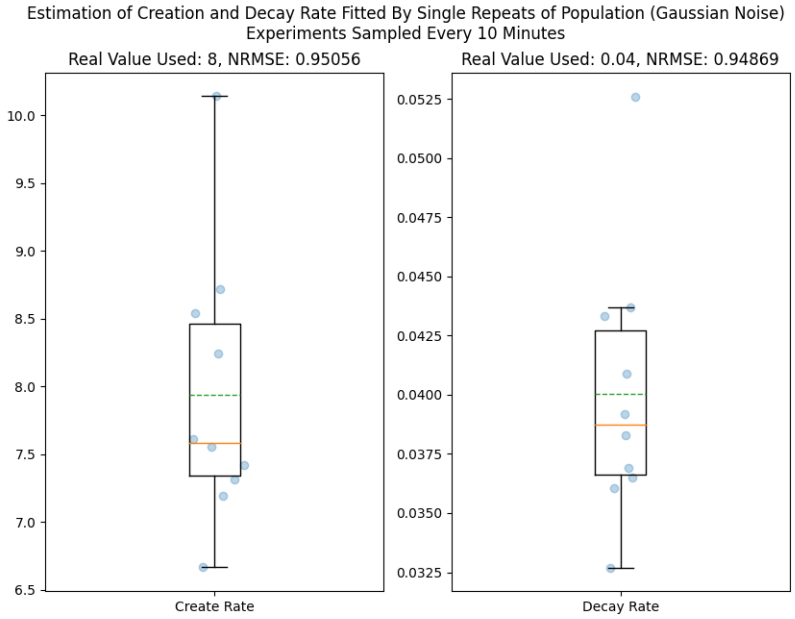
סוג ניסוי	אוכלוסייה (רעש גאוסיאני)		תאים בודדים (רעש פואסוני)	
סוג מדידה	מדידות יחידות	ממוצע שלשות	מדידות יחידות	ממוצע שלשות
קצב ייצור	7.577	7.662	8.078	8.116
קצב פירוק	0.0379	0.0384	0.0401	0.0403

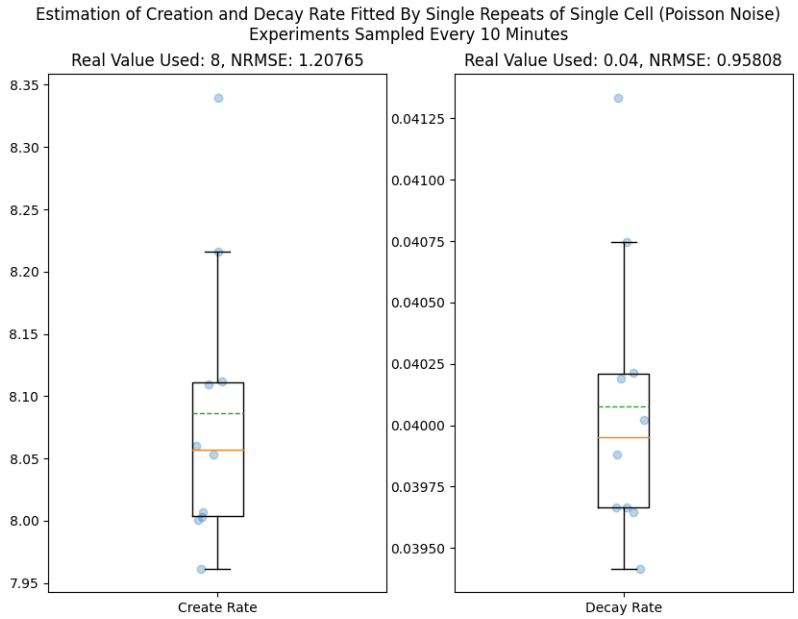
במקרה הזה ממוצעי האומדים לפי מדידות של האוכלוסייה יצאו מעט נמוכים מהערך האמיתי, וממוצעי האומדים לפי מדידות של תאים בודדים יצאו מעט גבוהים מהערך האמיתי. בנוסף, ממוצע האומדים לפי מדידות יחידות יוצא מעט קרוב יותר לערך האמיתי מאשר ממוצע האומדים שחושבו לפי ממוצעים של שלוש מדידות, שזו תוצאה קצת מפתיעה. גם לפי מדד ה-NRMSE ישנה עדיפות לאומדים המבוססים על מדידות יחידות, וגם עדיפות קטנה לאומדים שנבנו לפי מדידות של האוכלוסייה. עם זאת, יוצא שממוצעי האומדים לפי מדידות של תאים בודדים קרוב יותר לערכים האמיתיים מאשר ממוצעי האומדים לפי מדידות של האוכלוסייה. בנוסף, באופן לא מפתיע השונות של האומדים שמקורם בממוצע של שלוש מדידות קטן יותר.

נציין שההבדלים הם קטנים ונוטים להשתנות בין הרצה להרצה אז קשה להסיק מסקנות חד משמעיות מהתוצאות.

(ד)

כעת נוסיף את המקרה בו משערכים לפי כל ניסוי בנפרד בו מודדים כל 10 דקות (סה"כ 10):





סוג ניסוי	אוכלוסייה		תאים בודדים	
סוג מדידה	שלשות כל 30 דק'	יחידות כל 10 דק'	שלשות כל 30 דק'	יחידות כל 10 דק'
קצב ייצור	7.662	7.940	8.116	8.086
קצב פירוק	0.0384	0.04002	0.0403	0.04007

בכל המקרים האומדים שמקורם בממוצע של אומדים לפי מדידות יחידות כל 10 דקות קרובים לערכים האמיתיים, וגם במדד ה-RMSE יש עדיפות לאומדים ממדידות יחידות כל 10 דקות (פרט לקצב ייצור בתאים בודדים, שם מדד ה-NRMSE מעט טוב יותר במקרה של שלשות כל 30 דק). גם פה ההבדלים לא גדולים ומשתנים מהרצה להרצה.

שאלה 2

(א)

המידע הגולמי נקרא, לכל גן בוצע מיצוע והערכים הממוצעים הם אלו ששימשו לחיפוש הפרמטרים.
משוואת המודל הבסיסי הינה:

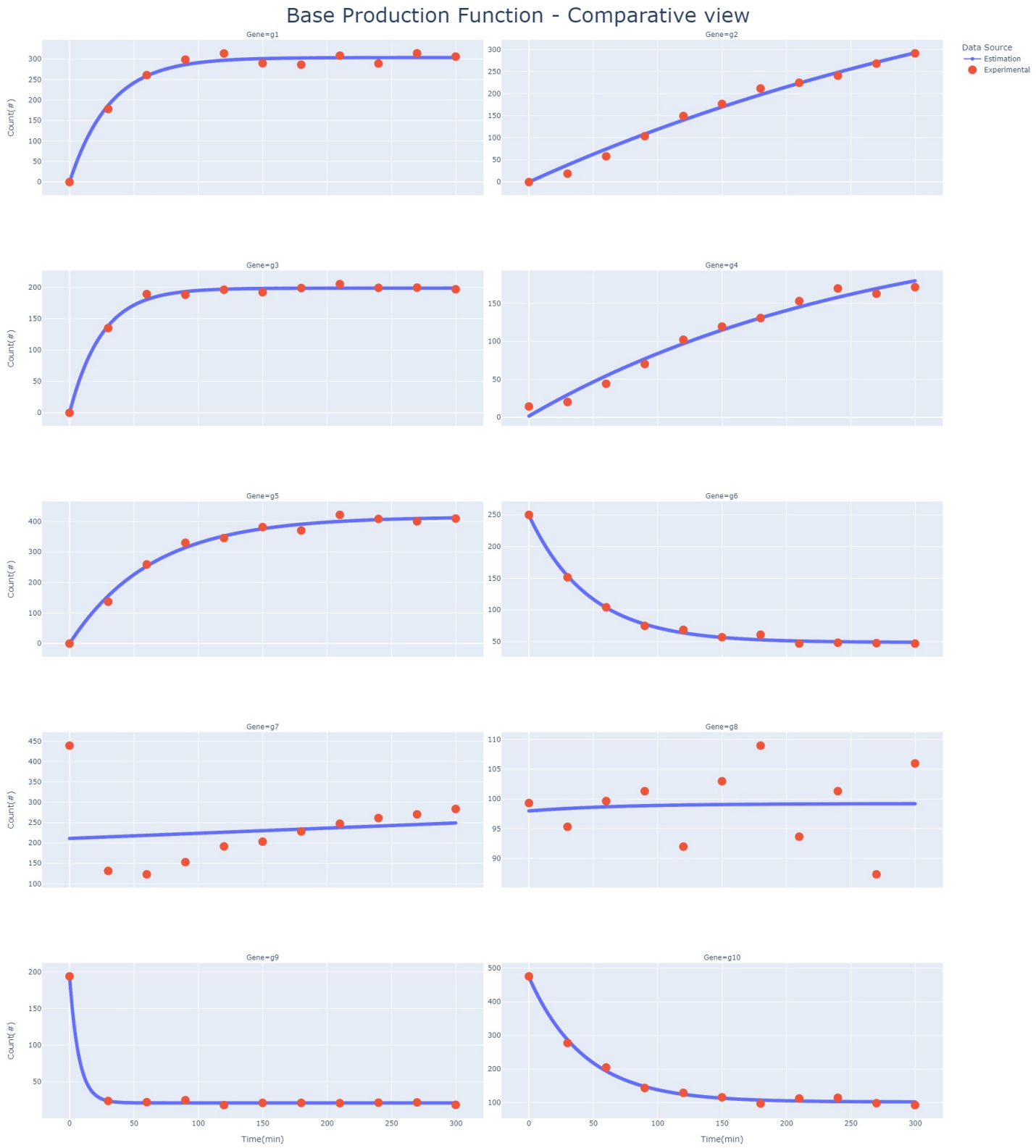
$$\dot{X} = \alpha - \beta X \Rightarrow X(t) = \frac{\alpha}{\beta}\left(1 - e^{-\beta t}\right) + X_0 e^{-\beta t}$$

כאשר $\alpha\left[\frac{\#}{min}\right]$ קצב הייצור, $\beta\left[\frac{1}{min}\right]$ קצב הפירוק ו $X_0\left[\#\right]$ הכמות ההתחלתית ברגע $t = 0$. הזמן כאמור נמדד בדקות.

	Production (#/min)	Removal (1/min)	Initial (#)
g1	9.7	0.03	0
g2	1.33	~0	0
g3	7.94	0.04	0
g4	0.99	~0	1.72
g5	6.53	0.02	0
g6	1.05	0.02	249.41
g7	0.13	~0	211.24
g8	1.37	0.01	98
g9	2.94	0.14	194
g10	2.38	0.02	474

(ב)

גרף המציג את התוצאות המשוערכות למול מיצוע התוצאות הניסויות:



את השגיאה נמדוד באמצעות מדד NRMSE:

$$NRMSE\left(Y,\hat{Y}\right)=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2}}{\widehat{\sigma_y}},$$
$$\widehat{\sigma_y}=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\bar{Y}\right)^2}{n-1}}$$

כאשר Y סדרה נתונה בת n תצפיות ו $\widehat{Y}$ נתוני המודל האומד את Y.
המדד מורכב מ RMSE, שורש ה-MSE, מדד מוכר להערכת שגיאה, אך כדי להשוות בין הגנים צריך להתייחס לשונות בביטוי של כל אחד מהם. לכן מחלקים בהערכת סטיית התקן של התצפיות.

	NRMSE
g9	0.0323
g6	0.0540
g10	0.0658
g3	0.0701
g5	0.0926
g2	0.0979
g1	0.1134
g4	0.1398
g7	0.9438
g8	0.9517

בפער גדול למדי $g7, g8$ הם הגנים שהמודל מתאים להם בצורה הקטנה ביותר.

(ג)

הסיבה המרכזית לקושי בהתאמה היא מספר החזרות המועט שבוצע. ישנן 3 חזרות בלבד לכל גן. מחוק המספרים הקטנים, כל רעש בביטוי הגנים יגרום לשינוי גדול בממוצע הדגימה.
ככל שיש מספר דגימות גדול יותר כך ההסתברות לקבל דגימה רחוקה מהממוצע קטן יותר (החוק החלש של המספרים הגדולים). הבעיה הזו ניכרת הרבה יותר עבור הגנים $g7, g8$ שהמדידות שלהם נמצאות בטווח ערכים צר מאוד ונראה שהדגימה אכן רועשת.

(ד)

למדנו מודל נוסף לביטוי גנים, כזה שמתייחס למצב של פרומוטור פעיל ושאינו פעיל:

$$\dot{X} = \frac{k_{on}}{V\left(k_{on}+k_{off}\right)} \cdot \alpha_{active} + \frac{k_{off}}{V\left(k_{on}+k_{off}\right)} \cdot \alpha_{incactive} - \beta X$$

באמצעות Wolfram Alpha קיבלנו את הפתרון הסגור:

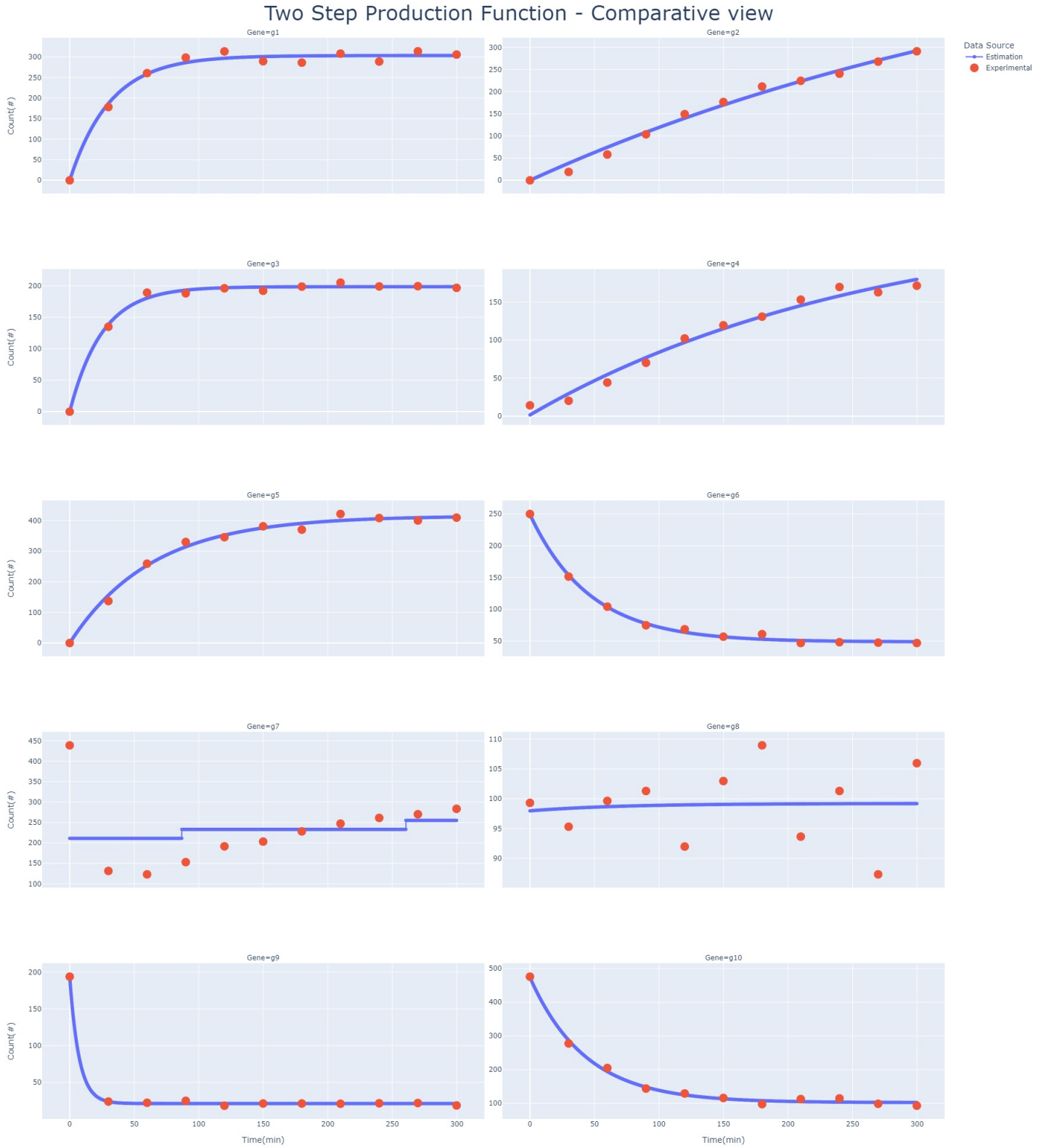
$$X\left(t\right) = \left(\frac{k_{on} \cdot \alpha_{active} + k_{off} \cdot \alpha_{incative}}{\beta V\left(k_{on}+k_{off}\right)}\right)\left(1-e^{-\beta t}\right) + X_0 e^{-\beta t}$$

נתונים "הגיוניים" מהמודל יהיו $\alpha_{active} >> \alpha_{inactive}$, הרי קצב הייצור של פרומוטור פעיל גבוה בהרבה מזה של פרומוטור שאינו פעיל. נזכור שמדובר ב"הרחבה" של המודל הישן, הרי כאשר $\alpha_{inactive} = \alpha_{active}$ לכל ערכי k_{on}, k_{off}, V חוקיים נקבל את מודל הבסיס.

בהרצת המודל החדש, נראה שרק $g7$ קיבל קבועי ביטוי שונים בצורה משמעותית ($\alpha_{active} = 5.27\frac{\#}{min}, \alpha_{inactive} = 0.08\frac{\#}{min}$), אך גם הם לא הביאו לשינוי במדד ה NRMSE.

בגרף בעמוד הבא ניתן לראות שרק עבור $g7$ נגרם שינוי קל (מאוד) במודל, עבור כל האחרים מדובר בגרפים זהים למעט סטיות זעירות.

יש לציין שעבור הגן $g7$ המדידה הראשונה גבוהה בצורה משמעותית מהשאר והיא אחראית ל"מירב" השונות. עם זאת, נראה שאינה outlier שכן היא מופיעה בערכים דומים עבור שלוש המדידות. ייתכן שביטוי הגן מתאים למודל אחר מהשניים שהוצעו לעיל.


Two Step Production Function - Comparative view
Gene=g1
Gene=g2
Gene=g3
Gene=g4
Gene=g5
Gene=g6
Gene=g7
Gene=g8
Gene=g9
Gene=g10
Count(#)
Time(min)
Data Source
Estimation
Experimental

שאלה 3

(א)

הסתברויות לאותיות בכל עמדה במוטיב:

PSSM:

FOXH1

$$\begin{bmatrix} A & C & G & T \\ 0.108333 & 0.375000 & 0.183333 & 0.333333 \\ 0.158333 & 0.383333 & 0.191667 & 0.266667 \\ 0.083333 & 0.358333 & 0.425000 & 0.133333 \\ 1.000000 & 0.000000 & 0.000000 & 0.000000 \\ 0.891667 & 0.000000 & 0.025000 & 0.083333 \\ 0.000000 & 0.000000 & 0.000000 & 1.000000 \\ 0.333333 & 0.666667 & 0.000000 & 0.000000 \\ 0.108333 & 0.891667 & 0.000000 & 0.000000 \\ 1.000000 & 0.000000 & 0.000000 & 0.000000 \\ 0.000000 & 1.000000 & 0.000000 & 0.000000 \\ 0.883333 & 0.000000 & 0.000000 & 0.116667 \end{bmatrix}$$

SMAD2

$$\begin{bmatrix} A & C & G & T \\ 0.008333 & 0.933333 & 0.041667 & 0.016667 \\ 0.000000 & 0.966667 & 0.025000 & 0.008333 \\ 0.908333 & 0.058333 & 0.033333 & 0.000000 \\ 0.000000 & 0.000000 & 1.000000 & 0.000000 \\ 1.000000 & 0.000000 & 0.000000 & 0.000000 \\ 0.000000 & 1.000000 & 0.000000 & 0.000000 \end{bmatrix}$$

והלוגואים:

FOXH1 **SMAD2**

Information logo:

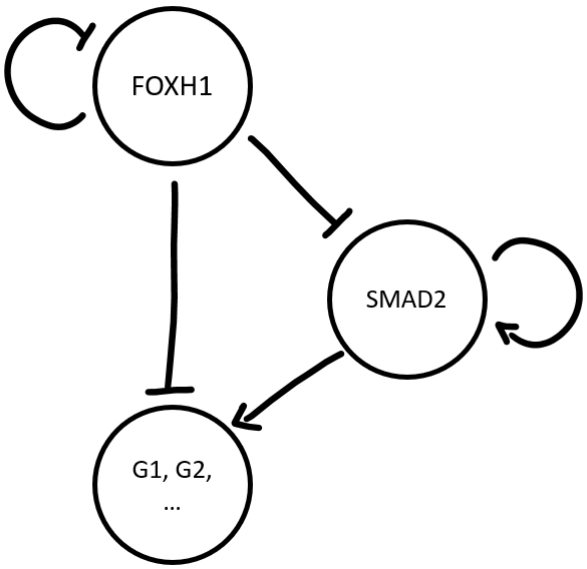
Nucleotide frequency logo:

(ב)

רשימות הגנים שקיים בהם מוטיב הקישור של SMAD2, FOXH1 נמצאות בהתאמה בקבצים המצורפים: fimo_SMAD2.tsv, fimo_FOXH1.tsv. נמצאים בהם מוטיבים הן על הגדיל הנתון והן על המשלים. נחוץ מידע על פעילות הגנים כדי לדעת את מידת השפעתם בציס/בטרנס.

נמצאו 32 גנים שהפרומוטורים שלהם מכילים הן רצפי בקרה ל-SMAD2 והן ל-FOXH1.

(ג)



לפי רצפי ההכרה בפרמוטורים של הגנים אפשר לנסות להסיק אילו מהם מושפעים על ידי FOXH1 ו-SMAD2.
כך ניתן לזהות מוטיב של אוטורגולציה שלילית ב-FOXH1 ואוטרגולציה חיובית ב-SMAD2.
בנוסף, כאמור לעיל, ב-32 מהפרומוטורים לגנים ברשימה ישנה נוכחות של רצפי ההכרה של שני הפרומוטורים. כל אחד מהם יוצר עם FOXH1 ו-SMAD2 מוטיב של Coherent FFL type 2.