מעבדה לסטטיסטיקה: אמידת מספר עותקים מדגימה בודדת (א)

אנחנו רוצים להשתמש בפונקציית ה GC שלמדנו ובערכים הנצפים Y_k כדי לאמוד את מספר-העותקים (Copy number) של הגנום.

בסיכום הזה נציג שיטת אמידה שדורשת רק אמידה של פונקציית ההשפעה של GC. לעיתים, קוראים לשיטות אמידה אלו "תיקון", משום שהן מתקנות את ההטיה הנוצרת מאפקט הGC.

מודל המתאר קשר בין מספר עותקים למספר פרגמנטים

נשתמש ב k=1…K לתאר את מספר התא, וב Y_k לתאר את הכיסוי הנצפה בתא.

נסמן את התוחלת של התא הk בλ_k

$$\lambda_k = E[Y_k].$$

המודל עבור התוחלת צריך לסמן את הרכיבים שמרכיבים את התוחלת, ולאפשר לנו לתאר איך היינו מייצרים נתונים כאלו. נשים לב ל3 רכיבים שאנחנו יודעים שמשפיעים על הכיסוי:

- כמות העותקים הוא מספר בכפולות שלמות של חצי (0, ½, 1, 1 ½ , 2…). תוחלת הכיסוי צריכה להיות פרופורציונלית למספר העותקים.
- סה"כ כמות הפרגמנטים (ע"פ כל התאים) יכולה להשתנות מניסוי לניסוי, ואינה קשורה למספר העותקים אלא למכונות הריצוף, נסמנה בJ.
- לכמות הGC בתא (או נגזרות כמו כמות הGC בתתי התאים) יש השפעה על הכיסוי. תיארנו תלות זו בעזרת מודלים של רגרסיה.

אם כן, נוח לתאר את התוחלת כפונקציה כפלית של שלושה גורמים <u>בלתי תלויים,</u>

כפול גורם לא ידוע η_k:

$$\lambda_k = N \cdot a_k \cdot f(gc_k) \cdot \eta_k,$$

לצורך זיהוי הפרמטרים צריך להניח הנחות לגבי הממוצעים שלη_k, $f(gc_k)$ ושל a_k. נבחר לדרוש מהחציון של a_k להיות 1, והממוצעים של הגורמים האחרים להיות 1:

$$med_k\{a_k\} = 1, \quad avg_k\{f(seq_k)\} = 1, \quad avg_k\{\eta_k\} = 1,$$

ואז N הקבוע שמתקן בהתאם, כלומר

$$N \approx med_k\{\lambda_k\}$$

אמידת מספר עותקים מהמודל

כעת נניח שיש לנו אומדן לפונקציה $\hat{f}(seq_k)$ ואת מספרי הפרגמנטים הנצפים $y_1, \dots, y_k$, ואנחנו רוצים לאמוד את מספר העותקים a_k.

(הפונקציה $\hat{f}(gc_k)$ כפי שאמדנו בכיתה לא תקיים $avg_k\{\hat{f}(gc_k)\} = 1$ אלא $avg_k\{\hat{f}(seq_k)\} = avg_k\{\lambda_k\}$ ולכן נצטרך לשים לב לקבועים בהמשך.)

ראשית נחלץ את a_k מהמודל:

$$a_k = \frac{\lambda_k}{f(gc_k)} \cdot \frac{1}{N} \cdot \frac{1}{\eta_k},$$

וכעת נשתמש באומד ההצבה כאשר נציב y_k במקום λ_k ואת $\hat{f}(gc)$ במקום $f(gc_k) \cdot N$.

$$\widetilde{a_k} = \frac{y_k}{\hat{f}(gc_k)}.$$

הערות

1. ניתן להכפיל בעוד קבוע M[1] שידאג שהחציון של האומדים יהיה 1:

$$\hat{a}_k = \frac{y_k}{\hat{f}(gc_k)} \cdot \widehat{M}, \qquad \widehat{M} = \frac{1}{med_k\{\widetilde{a_k}\}}.$$

2. ייתכן שיש עוד גורמים שתלויים ברצף הגנומי שמשפיעים על λ_k, כמו לדוגמה הסיכוי למפות פרגמנט, ואולי דרכי מדידה יותר עדינות ממספר הgc בתא. כדי לסמן מצב כזה, ניתן להשתמש ב $f(seq_k)$ בתיאור המודל במקום $f(gc_k)$. ככל שנכניס יותר גורמים ל$f(seq_k)$, כך המשמעות של η_k תשתנה ובתקווה השונות שלו תקטן.

---

[1] M מגלם גם את 1/N וגם את הממוצע של $1/\hat{f}(seq_k)$