

ВИКОНАВЕЦЬ:
СТУДЕНТ ГРУПИ МС-23М-1
ЩЕРБАК РОМАН ОЛЕКСІЙОВИЧ

КЕРІВНИК
ДОЦЕНТ КАФ. МСТ, К. Ф.-М.Н.
КАРНАУХ ЄВГЕН ВОЛОДИМИРОВИЧ

РАНДОМІЗОВАНИЙ КРИТЕРІЙ ПОРІВНЯННЯ СЕРЕДНІХ ДВОХ ГРУП

ПОСТАНОВКА ЗАДАЧІ

Гострий лімфобластний лейкоз (Acute lymphoblastic leukemia, ГЛЛ, ALL) – онкологічне захворювання клітин крові (зокрема, Т-лімфоцитів та В-лімфоцитів).

Мета – застосувати методи статистичного аналізу для встановлення істотності різниці середніми рівнями експресії генів.

Завдання

- Огляд літератури щодо дослідження генної інформації.
- Розгляд основних методів статистичного аналізу генетичної інформації.
- Застосування T_n критерію для виявлення статистичної відмінності між експресією істотних генів.

НАБІР ДАНИХ ALL

Містить інформацію про 128 пацієнтів із Т-лейкемією та В-лейкемією та рівень експресії в них 12 625 генів.

Релевантних для розгляду – 2391 ген.

Вилучено гени з низьким рівнем експресії для обох типів хвороби та низькою мінливістю (за IQR).

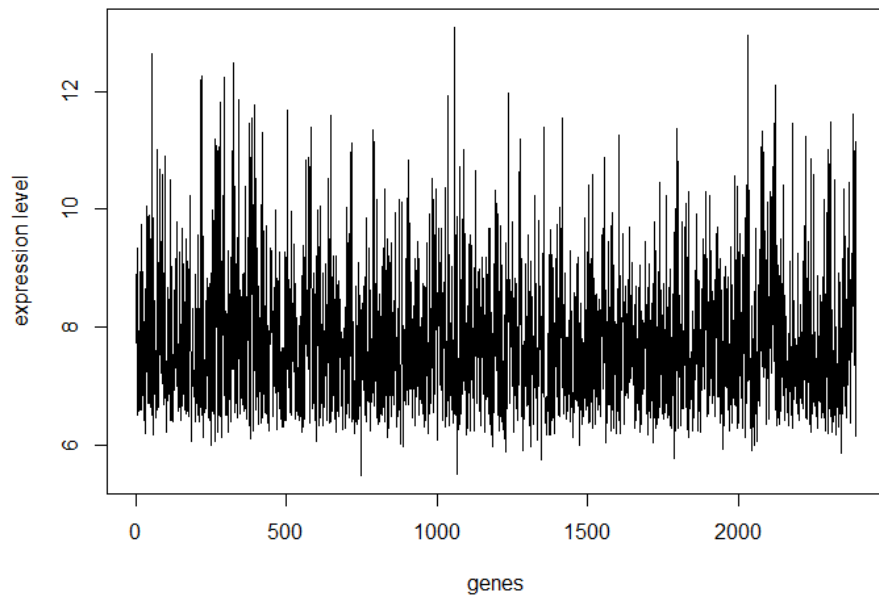
37 пацієнтів із BCR/ABL, 42 пацієнти з NEG (В-лейкемія).

BCR/ABL – генетична абнормалія 22 хромосоми (філадельфійська хромосома).

NEG – відсутність виявлених генетичних абнормалій.

НАБІР ДАНИХ ALL

Gene expression level means BCR/ABL



Gene expression level means NEG

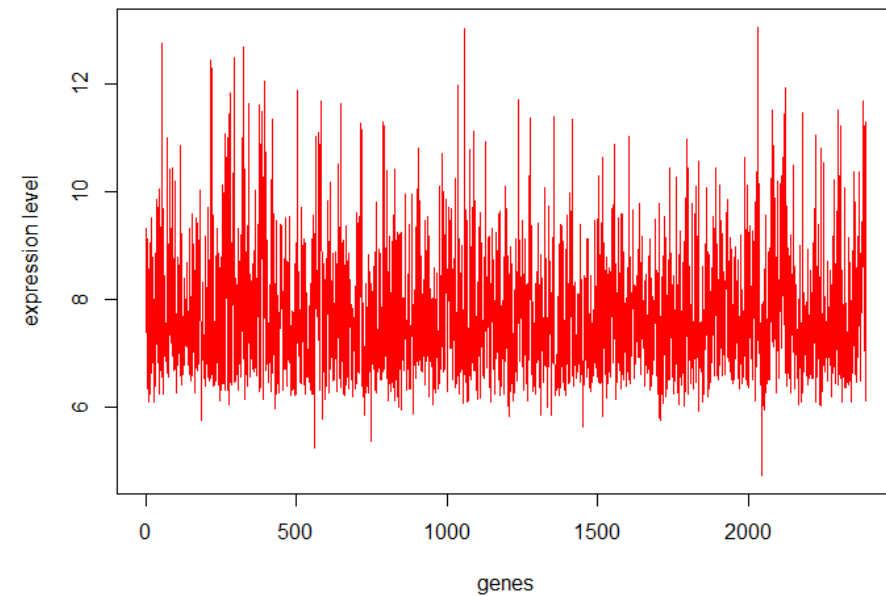


Рис 1. Графік середніх рівнів експресії генів для захворювання типу BCR/ABL (чорна лінія) та NEG (червона лінія)

СТАТИСТИКА M_n

Стандартно для порівняння середніх застосовують критерій Хотелінга T^2 з фіксованим p .

За $p/n \rightarrow \infty$ критерій Хотелінга незастосовний, а за $p/n \rightarrow c \in [0,1)$ його потужність зменшується зі збільшенням c

Альтернативні статистики:

$$1) M_n = (\bar{X}_1 - \bar{X}_2)' (\bar{X}_1 - \bar{X}_2) - \tau \text{tr}(S_n), \text{ де}$$

$$S_n = \frac{1}{n} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)'$$

$$\tau = \frac{n_1 + n_2}{n_1 n_2}$$

СТАТИСТИКА T_n

$$2) T_n =: \frac{\sum_{i \neq j}^{n_1} X'_{1i} X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i \neq j}^{n_2} X'_{2i} X_{2j}}{n_2(n_2-1)} - 2 \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} X'_{1i} X_{2j}}{n_1 n_2}$$

Вилучено $\sum_{j=1}^{n_i} X'_{ij} X_{ij}$ для $i = 1$ та 2 з $\left| \bar{X}_1 - \bar{X}_2 \right|^2$

$$E(T_n) = \left| \mu_1 - \mu_2 \right|^2$$

Якщо правильна гіпотеза H_1 :

$$\text{Var}(T_n) = \left\{ \frac{2}{n_1(n_1-1)} \text{tr}(\Sigma_1^2) + \frac{2}{n_2(n_2-1)} \text{tr}(\Sigma_2^2) + \frac{4}{n_1 n_2} \text{tr}(\Sigma_1 \Sigma_2) \right\} \{1 + o(1)\}$$

СТАТИСТИКА Q_n

3) $Q_n = T_n / \widehat{\sigma}_{n1} \rightarrow N(0,1)$, $p \rightarrow \infty$, $n \rightarrow \infty$, де

$$\sigma_{n1}^2 = \frac{2}{n_1(n_1 - 1)} \widehat{tr}(\Sigma_1^2) + \frac{2}{n_2(n_2 - 1)} \widehat{tr}(\Sigma_2^2) + \frac{4}{n_1 n_2} \widehat{tr}(\Sigma_1 \Sigma_2)$$

$$\widehat{tr}(\Sigma_i^2) = \{n_i(n_i - 1)\}^{-1} \widehat{tr}\left\{\sum_{j \neq k}^{n_i} (X_{ij} - \bar{X}_{i(j,k)}) X'_{ij} (X_{ik} - \bar{X}_{i(j,k)}) X'_{ik}\right\}$$

$$\widehat{tr}(\Sigma_1 \Sigma_2) = (n_1 n_2)^{-1} \widehat{tr}\left\{\sum_{l=1}^{n_1} \sum_{k=1}^{n_2} (X_{1l} - \bar{X}_{1(l)}) X'_{1l} (X_{2k} - \bar{X}_{2(k)}) X'_{2k}\right\},$$

$\bar{X}_{i(j,k)}$ – i -е вибіркове середнє після вилучення X_{ij} та X_{ik} , $\bar{X}_{i(l)}$ – i -е вибіркове середнє після вилучення X_{il} .

Якщо $Q_n > \xi_\alpha$, де ξ_α – верхній α квантиль розподілу $N(0,1)$, гіпотезу H_0 відкидають.

ПРОГРАМА

Рівень значущості було встановлено як 0.05.

Містить три модулі:

- `main` – завантаження та аналіз даних, ділення сум для статистики T_n , обчислення оцінки σ_{n1}^2 та статистики Q_n , обчислення `qnorm(1-sign_level/2)`.
- `sum_for_T_stat` – обчислення сум для статистики T_n для векторів $X1, X2$ та для їх комбінації.
- `trace_for_Q_stat` – обчислення оцінки сліду матриці для статистики Q_n для векторів $X1, X2$ та для їх комбінації.

У результаті обчислення було отримано значення $Q_n = 0.458$, тоді як значення $\xi_\alpha = 1.959$, що не дозволяє відкидати гіпотезу H_0 про рівність середніх.

ВИСНОВКИ

Було розглянуто набір даних про гострий лімфобластний лейкоз, відібрано релевантні гени та пацієнтів із типами захворювання BCR/ABL та NEG

Розглянуто методи статистичного аналізу генетичної інформації, перевірено гіпотезу про рівність середніх за допомогою Q_n -критерію, використовуючи асимптотику для $p/n \rightarrow \infty$.



ДЯКУЄМО!