ДНІПРОВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ОЛЕСЯ ГОНЧАРА
Механіко-математичний факультет
Кафедра статистики й теорії ймовірностей

КУРСОВА РОБОТА

другий (магістерський) рівень вищої освіти

спеціальність 112 Статистика

РАНДОМІЗОВАНИЙ КРИТЕРІЙ ПОРІВНЯННЯ СЕРЕДНІХ ДВОХ ГРУП

Виконавець
студент групи МС-23м-1
Роман ЩЕРБАК

Керівник
зав. каф. МСТ, к. ф.-м. н.,
Євген КАРНАУХ

Завідувач випускової кафедри
к.ф.-м.н. Валерій ТУРЧИН

Дніпро
2024

ЗМІСТ

ВСТУП **3**

Актуальність теми дослідження **3**

Завдання роботи **3**

РОЗДІЛ 1. ТЕОРЕТИЧНА ЧАСТИНА **4**

1.1 Постановка задачі **4**

1.2 Лейкемії та їх типи **4**

1.3 Гени та рівень їх експресії **4**

1.4 Статистичні критерії порівняння середніх **6**

РОЗДІЛ 2. ПРАКТИЧНА ЧАСТИНА **11**

2.1 Аналіз даних **11**

2.2 Опис алгоритму **12**

ВИСНОВКИ **14**

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ **15**

ДОДАТОК А. Код програми мовою R **16**

ВСТУП

Актуальність теми дослідження

Гострий лімфобластний лейкоз (Acute lymphoblastic leukemia, ГЛЛ, ALL) – онкологічне захворювання клітин крови, за тип і перебіг якого відповідає генетична інформація організму. Серед типів захворювання виділяють два основні: BCR/ABL та NEG. Їхня відмінність полягає в тому, що перший має наявну генетичну абнормалію на 22 хромосомі, тоді як для другого жодної генетичної абнормалії не зафіксовано. На базі аналізу генетичної інформації пацієнтів із діагнозом ГЛЛ було сформовано дані для більш поглибленого вивчення відмінності між середніми експресіями генів.

У роботі розглянуто один статистичний критерій перевірки гіпотези щодо середніх значень багатовимірних розподілів на прикладі даних із бібліотеки ALL щодо рівнів експресії генів для осіб із Т-лейкемією та В-лейкемією.

Об'єкт дослідження – експресії генів для певного типу захворювань.

Предмет дослідження – статистична відмінність між середніми рівнями експресії генів для двох різновидів захворювання.

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

Завдання роботи

1.Огляд літератури щодо дослідження генної інформації.

2.Розгляд основних методів статистичного аналізу генетичної інформації.

3.Застосування Q_n-критерію для виявлення статистичної відмінности між експресією істотних генів.

РОЗДІЛ 1.ТЕОРЕТИЧНА ЧАСТИНА

1.1 Постановка задачі

Задано набір даних із бібліотеки ALL для мови R, що містить інформацію про 128 пацієнтів із Т-лейкемією та В-лейкемією та рівень експресії в них 12 625 генів. Назви цих лейкемій походять від типу лімфоцитів, які вражає цей вид захворювання: відповідно, Т-лімфоцити та В-лімфоцити.

Необхідно порівняти середній рівень експресії релевантних генів між пацієнтами з захворюванням В-лейкемії типу BCR/ABL та NEG та визначити, відсутня (гіпотеза H_0) чи наявна (гіпотеза H_1) в них відмінність. Усього в наборі даних наявно 37 пацієнтів із типом лейкемії BCR/ABL та 42 – з NEG.

1.2 Лейкемії та їх типи

В-лейкемії типу BCR/ABL та NEG відрізняються біологічною природою цього типу раку. Для типу BCR/ABL його появу зумовлено генетичною абнормалією 22 хромосоми (так званою філадельфійською хромосомою), що спричинено реципрокною транслокацією (взаємним обміном) ділянок із хромосомою 9, що призводить до того, що ген 9 хромосоми ABL1 опиняється та вступає в контрастну дію з геном BCR хромосоми 22, що призводить до появи гібридного сигнального протеїну для тирозинкінази. Оскільки цей протеїн постійно ввімкнено, то клітина з цією мутацією, починає ділитися безконтрольно, що є однією з передумов ракового захворювання. Для типу NEG неможливо виявити жодну схожу генетичну абнормалію, тому аналіз рівнів експресії генів може допомогти у виявленні та класифікації хвороби за типом.

1.3 Гени та рівень їх експресії

Для відбору релевантних генів було застосовано процедуру, запропоновану Джентлменом та іншими [1], що вилучає гени, рівень експресії яких у кров'яних тільцях низький що у хворих на лейкемію типу BCR/ABL або NEG, що у здорових.

Також було вилучено гени з низькою мінливістю, з використанням міжквартильного розмаху (IQR).

За Келліс та інші [2] існує декілька різних методів вимірювання експресії генів, серед них описано, наприклад: ДНК-біочипи та РНК-послідовності. Для отримання експресії генів за методом біочипів беруться короткі ділянки ДНК, що звуться пробами. Їх прикріпляють до твердої поверхні, відомої як ДНК-біочип. Тоді отримана з клітини РНК-популяція інтересу зворотно транскрибується у кДНК (компліментарну ДНК). Тоді біочип промивають кДНК, що запускає процес гібридизації, який призводить до флуоресцентного світіння проб. Детекція цього світіння дозволяє визначити відносну кількість мРНК у пробі. РНК-послідовність є більш сучасною технологією визначення експресії генів. Її функція подібна до ДНК-біочипів, але з більшою точністю. Відмінність полягає в тому, що для методу ДНК-біочипів необхідне використання специфічних проб, і їх створення потребує знання про геном та розмір утвореного масиву даних. Технологія РНК-послідовности не має цих обмежень, вона дозволяє секвенувати всю кДНК, отриману в експериментах із біочипами за допомогою технології секвенування нового покоління. Цю техніку широко використовують, наприклад, для вивчення раку. Отримані з обох методів дані аналізуються однаковим чином за допомогою кластерингу.

Результати аналізу експресії генів часто подають у вигляді матриць та їхніх теплових карт. Для їх отримання за допомогою вищезазначених методів проводиться вимірювання генів за різних умов: часу, стадій розвитку, фенотипів, здоров’я чи хвороба тощо. За допомогою цього отримують значення рівня експресії генів у чисельній формі. Якщо було проведено багато експериментів, то можна побудувати матрицю значень, що відображає значення $log\left(\frac{T}{R}\right)$, де T – рівень експресії гена в тестовому зразку, R – рівень експресії гена в еталонному зразку. Такі матриці можна кластерувати за ієрархією, відображаючи відносини

між парами генів, парами пар тощо. Це утворює дендрограму, з колонками та рядками, упорядкованими за певним алгоритмом. Це дозволяє відкрити приховану структуру довгого сегмента геному та отримати розуміння про його функцію та, відповідно, краще розуміння причини певної хвороби. Такі теплові карти для розглянутого набору даних про ГЛЛ можна знайти, наприклад, у К'яретті та інші [3].

1.4 Статистичні критерії порівняння середніх

Стандартно для порівняння середніх застосовують критерій Готелінґа T^2 з фіксованим p, меншим за $n =: n_1 + n_2 - 2$ та за $\Sigma_1 = \Sigma_2 = \Sigma$, однак його сфера застосування не включає ситуацію, коли $p/n \to \infty$, тобто розмірність даних значно перевищує розмір вибірки. Баї та Сарандаса [4] показали також, що за $p \to c \in [0,1)$ його точність зменшується зі збільшенням c, зокрема через те, що матриця вибіркової коваріації S_n не збігається до популяційної коваріації коли p та n мають однаковий порядок. Їн, Баї та Крішная [5] показали, що за $p/n \to c$ максимум і мінімум власних значень вибіркової коваріації S_n не збігаються до відповідних власних значень Σ. За $p > n$ S_n може не бути оборотною, тому статистику T^2 не визначено.

Баї та Сарандаса запропонували обчислення статистики M_n, а Чен та Цін [6] розширили її до статистик T_n та Q_n.

Використання цих статистик полягає в наявності двох незалежних однаково розподілених випадкових вибірок із R^p, а саме

$$\{X_{i1}, X_{i2}, \cdots X_{in_i}\} \sim F_i, \qquad \text{для } i = 1,2, \tag{1.1}$$

де F_i – розподіл із R^p із середнім μ_i та коваріацією Σ_i. Одна з цілей багатовимірного аналізу – перевірка гіпотези про рівність середніх двох багатовимірних популяцій:

$H_0: \mu_1 = \mu_2$ проти $H_1: \mu_1 \neq \mu_2$.

Ця гіпотеза складається з p маргінальних гіпотез щодо середніх кожного з виміру даних.

Тестування гіпотези про рівність середніх за умови $\Sigma_1 = \Sigma_2 = \Sigma$ за статистикою Баї та Сарандаса базується на:

$$M_n = \left(\overline{X}_1 - \overline{X}_2\right)'\left(\overline{X}_1 - \overline{X}_2\right) - \tau tr(S_n), \quad (1.2)$$

де $S_n = \frac{1}{n}\left(X_{ij} - \overline{X}_i\right)\left(X_{ij} - \overline{X}_i\right)'$ та $\tau = \frac{n_1+n_2}{n_1 n_2}$. Пропозиція Баї та Сарандаса полягає у вилученні S_n^{-1} з тесту T^2, зважаючи на відсутність користі від нього за $p/n \to c > 0$. Також для отримання $E(M_n) = \left|\left|\mu_1 - \mu_2\right|\right|^2$ віднімають $tr(S_n)$. Для цього тесту припущено такі умови:

$$p/n \to c < \infty \quad \text{та} \quad \lambda_p = o\left(p^{1/2}\right) \quad (1.3)$$

$$n_1/(n_1 + n_2) \to k \in (0,1) \quad \text{та} \quad (\mu_1 - \mu_2)'\Sigma(\mu_1 - \mu_2) = o\{tr(\Sigma^2)/n\}, \quad (1.4)$$

де λ_p – найбільше власне значення Σ.

Статистика T_n має вигляд:

$$T_n =: \frac{\sum_{i\neq j}^{n_1} X_{1i}' X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2} X_{2i}' X_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{1i}' X_{2j}}{n_1 n_2} \quad (1.5)$$

після вилучення $\sum_{j=1}^{n_i} X_{ij}' X_{ij}$ для $i = 1$ та 2 з $\left|\left|\overline{X}_1 - \overline{X}_2\right|\right|^2$. Можна показати, що $E(T_n) = \left|\left|\mu_1 - \mu_2\right|\right|^2$.

Тож T_n є всім необхідним для тестування. За правильности гіпотези H_1 та умови

$$(\mu_1 - \mu_2)'\Sigma_i(\mu_1 - \mu_2) = o[n^{-1} tr\{(\Sigma_1 + \Sigma_2)^2\}] \quad (1.6)$$

маємо:

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

де за H_0 $o(1)$ зникає.

Нехай

$$X_{ij} = \Gamma_i Z_{ij} + \mu_i \quad \text{для} \quad j = 1, \cdots, n_i, \ i = 1 \text{ та } 2, \quad (1.7)$$

де кожне Γ_i – матриця $p \times m$ для деякого $m \geq p$ такого, що $\Gamma_i \Gamma_i' = \Sigma_i$, $\{Z_{ij}\}_{j=1}^n$ – m-вимірні незалежні однаково розподілені випадкові вектори, де $E(Z_{ij}) = 0, Var(Z_{ij}) = I_m$ – одинична матриця $m \times m$.

Якщо переписати $Z_{ij} = (z_{ij1}, \cdots, z_{ijm})'$, припускаємо, що $E(z_{ijk}^4) = 3 + \Delta < \infty$ та

$$E\left(z_{ijl_1}^{\alpha_1} z_{ijl_2}^{\alpha_2} \cdots z_{ijl_q}^{\alpha_q}\right) = E\left(z_{ijl_1}^{\alpha_1}\right) E\left(z_{ijl_2}^{\alpha_2}\right) \cdots E\left(z_{ijl_q}^{\alpha_q}\right) \tag{1.8}$$

для додатного цілого q такого, що $\sum_{l=1}^q \alpha_l \leq 8$ та $l_1 \neq l_2 \neq \cdots \neq l_q$.

$$n_1/(n_1 + n_2) \to k \in (0,1) \quad \text{за } n \to \infty \tag{1.9}$$

Якщо (1.6) не дійсна, то використовуємо

$$n^{-1} tr\{(\Sigma_1 + \Sigma_2)^2\} = o\{(\mu_1 - \mu_2)' \Sigma_i (\mu_1 - \mu_2)\} \quad \text{для } i = 1 \text{ або } 2 \tag{1.10}$$

На p ставиться умова:

$$tr(\Sigma_i \Sigma_j \Sigma_l \Sigma_h) = o[tr^2\{(\Sigma_1 + \Sigma_2)^2\}] \quad \text{для } i,j,l,h = 1 \text{ або } 2 \tag{1.11}$$

за $p \to \infty$.

З (1.7), (1.8), (1.9), (1.11) та (1.6) або (1.10) дійсна така теорема:

ТЕОРЕМА 1. Має місце збіжність за розподілом:

$$\frac{T_n - ||\mu_1 - \mu_2||^2}{\sqrt{Var(T_n)}} \to N(0,1)$$

за $p \to \infty$ та $n \to \infty$. Асимптотичну нормальність витримано без певних прямих обмежень між p та n. Єдине обмеження на виміри дано в (1.11). Дисперсія цієї статистики має вигляд:

$$Var(T_n) = \sigma_n^2\{1 + o(1)\},$$

де за (1.6)

$$\sigma_n =: \sigma_{n1}^2 = \frac{2}{n_1(n_1-1)} tr(\Sigma_1^2) + \frac{2}{n_2(n_2-1)} tr(\Sigma_2^2) + \frac{4}{n_1 n_2} tr(\Sigma_1 \Sigma_2) \tag{1.12}$$

та за (1.10):

$$\sigma_n =: \sigma_{n2}^2 = \frac{4}{n_1}(\mu_1 - \mu_2)' \Sigma_1 (\mu_1 - \mu_2) + \frac{4}{n_2}(\mu_1 - \mu_2)' \Sigma_2 (\mu_1 - \mu_2) \tag{1.13}$$

Для формулювання тестової процедури, що базується на цій теоремі необхідно оцінити σ_{n1}^2 із (1.12).

Для виключення з T_n таких доданків, як $\sum_{j=1}^{n_i} X_{ij}' X_{ij}$ Чен та Цін запропонували такі оцінки $tr\left(\Sigma_i^2\right)$ та $\mathrm{tr}(\Sigma_1\Sigma_2)$:

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

та

$$\widehat{tr(\Sigma_1\Sigma_2)} = (n_1n_2)^{-1} tr\{\sum_{l=1}^{n_1}\sum_{k=1}^{n_2}\left(X_{1l}-\overline{X}_{1(l)}\right)X_{1l}'\left(X_{2k}-\overline{X}_{2(k)}\right)X_{2k}'\},$$

де $\overline{X}_{i(j,k)}$ – і-е вибіркове середнє після вилучення X_{ij} та X_{ik}, $\overline{X}_{i(l)}$ – і-е вибіркове середнє після вилучення X_{il}. Це подібне до ідеї кросвалідації, де під час утворення відхилень X_{ij} та X_{ik} від вибіркового середнього X_{ij} та X_{ik} виключаються з обчислення вибіркового середнього. Таким чином оцінки $\widehat{tr(\Sigma_i^2)}$ та $\widehat{tr(\Sigma_1\Sigma_2)}$ можна подати як слід сум добутків незалежних матриць.

За припущеннями (1.6)-(1.9) та (1.11) для $i = 1$ або 2 вище подані оцінки є спроможними за такою теоремою:

ТЕОРЕМА 2. Має місце збіжність за ймовірністю:

$$\frac{\widehat{tr(\Sigma_i^2)}}{tr(\Sigma_i^2)} \to 1 \quad \text{та} \quad \frac{\widehat{tr(\Sigma_1\Sigma_2)}}{tr(\Sigma_1\Sigma_2)} \to 1 \qquad \text{за} \ \ p \ \text{та} \ n \to \infty.$$

За H_0 оцінка σ_{n1}^2, спроможна:

$$\sigma_{n1}^2 = \frac{2}{n_1(n_1-1)}\widehat{tr(\Sigma_1^2)} + \frac{2}{n_2(n_2-1)}\widehat{tr(\Sigma_2^2)} + \frac{4}{n_1n_2}\widehat{tr(\Sigma_1\Sigma_2)}. \qquad (1.14)$$

Це спільно з теоремою 1 дає тестову статистику, збіжну за розподілом:

$$Q_n = T_n/\widehat{\sigma_{n1}} \to N(0,1) \qquad \text{за} \ \ p \ \text{та} \ n \to \infty$$

за H_0. Цей тест із рівнем статистичної значущости α відкидає гіпотезу H_0 якщо $Q_n > \xi_\alpha$, де ξ_α – верхній α квантиль розподілу $N(0,1)$.

За Менлі [7] рандомізований критерій перевірки нульової гіпотези полягає у виборі статистики S, що дозволяє виміряти, наскільки в заданих даних наявний

шуканий ефект. Наскільки значення s статистики S для наявних даних правдоподібні за умови, що дані отримані випадковою перестановкою. У випадку, якщо нульова гіпотеза дійсна, то будь-які варіанти даних могли статися з однаковою ймовірністю, і спостережувані дані є лише одним із однаково правдоподібних порядків, значення s має бути типовим значенням розподілу S. Якщо такий результат не спостережено, то s є значущою, тому більш правдоподібною є альтернативна гіпотеза. Рівень значущости s визначають як частку значень, що є так само чи більш екстремальними за його значення з рандомізованого розподілу. У такому разі якщо s менше за 5%, то є певні підстави вважати нульову гіпотезу неправильною, якщо s менше за 1%, то є досить сильні підстави вважати, що нульова гіпотеза неправильна, якщо s менше за 0.1%, то є дуже сильні підстави відкидати нульову гіпотезу.

РОЗДІЛ 2.ПРАКТИЧНА ЧАСТИНА

2.1 Аналіз даних

Набір даних ALL для мови R містить інформацію про 128 пацієнтів із Т-лейкемією та В-лейкемією та рівень експресії в них 12 625 генів. Після фільтрування за вищезазначеною процедурою з Джентлмен та інші було знайдено 2391 ген, істотний для розгляду в цій задачі.

Зі 128 пацієнтів у 79 наявна лейкемія типів BCR/ABL або NEG. Для обчислення статистик за процедурою, запропонованою Джентлменом та іншими з оригінального набору даних вибрано цих пацієнтів та розподілено на дві матриці, що містять, відповідно, лише пацієнтів із лейкемію типу BCR/ABL (позначено як X1) та лейкемією типу NEG (позначено як X2).

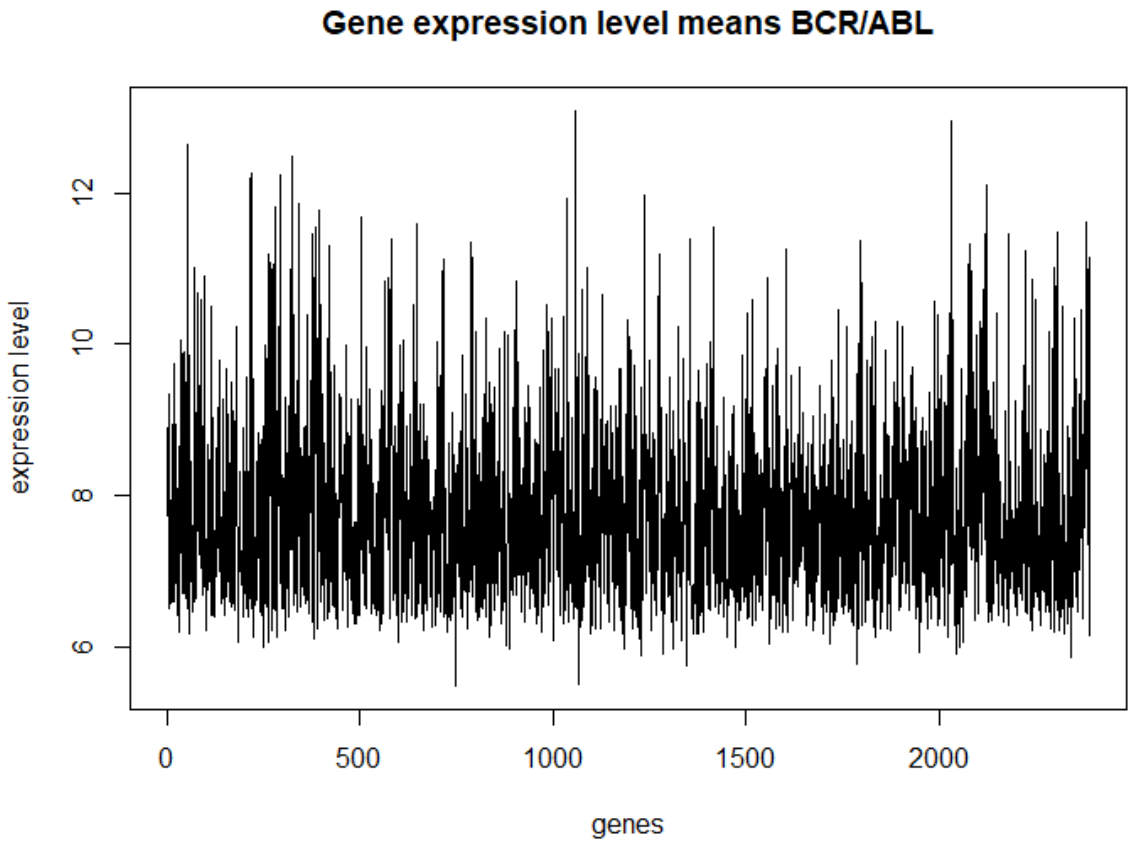

Рис 2.1. Графік середніх рівнів експресії генів для захворювання типу BCR/ABL

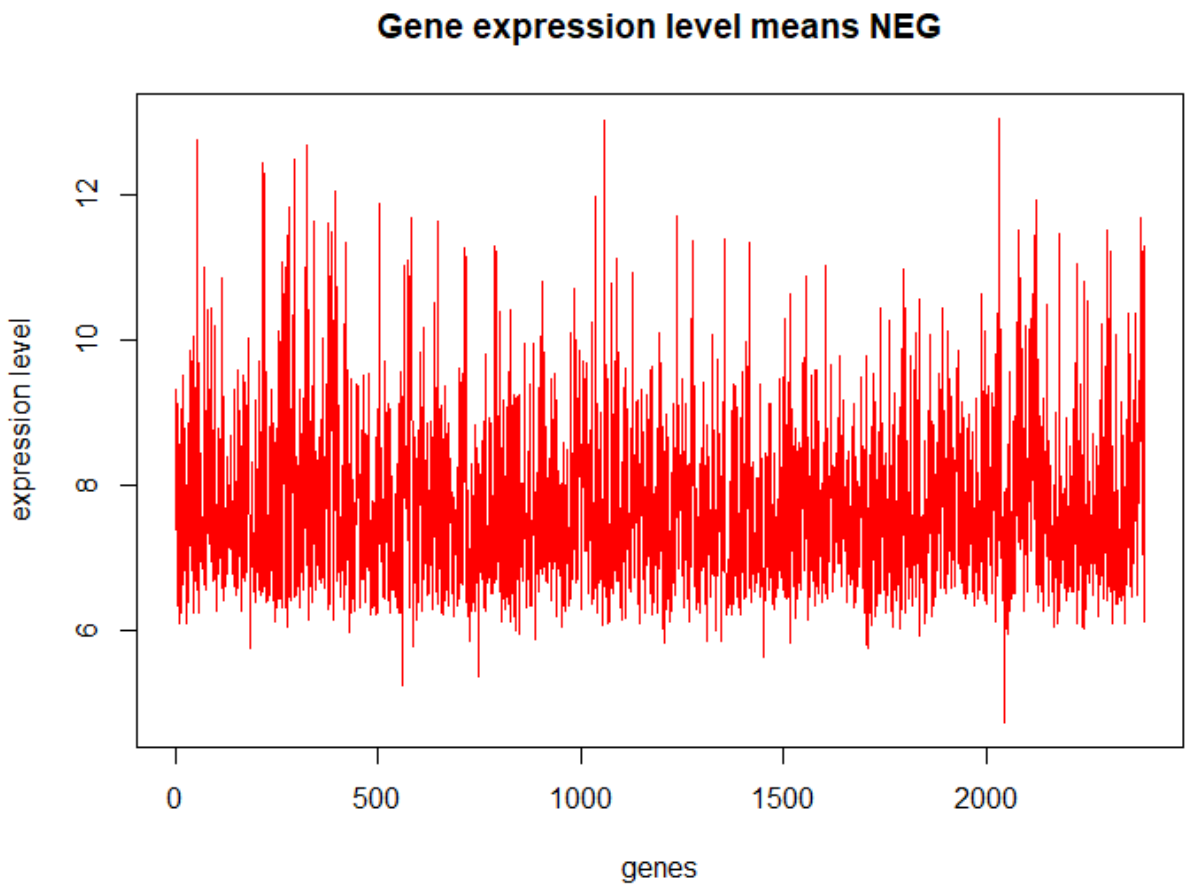

Рис 2.2. Графік середніх рівнів експресії генів для захворювання типу NEG

2.2 Опис алгоритму

Робота програми починається з задання змінної рівня значущости sign_level як 0.05. З використанням функції sum_for_T з модуля sum_for_T_stat проводиться обчислення за формулою для статистики T_n (1.5) для вектору X1, для вектору X2 та для їх комбінації, після чого три отримані значення підсумовуються для отримання статистики T_n.

Наступним кроком за допомогою функції trace_for_Q з модуля trace_for_Q_stat проводиться обчислення оцінки сліду матриці за заданими формулами для статистики Q_n (1.14) для вектору X1, для вектору X2 та для їх

комбінації, після чого три отримані значення множаться на відповідні дроби, а результати підсумовуються для отримання оцінки $\widehat{\sigma_{n1}^2}$.

Тоді значення отриманої статистики T_n ділиться на значення оцінки $\widehat{\sigma_{n1}^2}$, що дає значення статистики Q_n. Для проведення тесту використовується значення ξ_α, отримане за допомогою функції qnorm(1-sign_level/2).

У результаті обчислення було отримано значення Q_n – 0.4584315, тоді як значення ξ_α – 1.959964, що дані не суперечать гіпотезі H_0 про рівність середніх.

ВИСНОВКИ

У роботі було розглянуто набір даних про гострий лімфобластний лейкоз, що містить інформацію про 128 пацієнтів та значення рівнів експресії в них 12 625 генів. З них відібрано для подальшого розгляду 2391 релевантний ген для 79 пацієнтів, що мають лейкемію одного з двох визначених типів.

Розглянуто методи статистичного аналізу генетичної інформації Зокрема, для визначення наявности відмінностей між середніми рівнями експресій генів між двома типами захворювання було перевірено гіпотезу про рівність середніх за допомогою Q_n −критерію, використовуючи асимптотику для $p/n \to \infty$.

Було встановлено, що дані не суперечать нульовій гіпотезі про відсутність різниці між середніми рівнями експресій генів між двома захворюваннями. Статистичної відмінности між середніми експресіями істотних генів за допомогою зазначеного критерію виявлено не було.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Robert Gentleman et al. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer Science+Business Media, 2005. 473 p.
2. Manolis Kellis et al. Computational Biology - Genomes, Networks, and Evolution. Massachusetts Institute of Technology : LibreTexts, 2024. 565 p. URL : https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)
3. Sabina Chiaretti et al. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood, 1 April 2004. Vol. 103, No. 7. P. 2771-2778. DOI : https://doi.org/10.1182/blood-2003-09-3243 URL : https://ashpublications.org/blood/article/103/7/2771/18318/Gene-expression-profile-of-adult-T-cell-acute
4. Z. Bai, H. Saranadasa. Effect of high dimension: By an example of a two sample problem. Statist. Sinica 6, 1996. 311–329 p.
5. Y. Yin, Z. Bai, P. R. Krishnaiah. On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. Probab. Theory Related Fields 78, 1988. 509–521 p.
6. Song Xi Chen, Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. The Annals of Statistics Vol. 38, No. 2, 2010, 808-835 p. URL: 10.1214/09-AOS716
7. Bryan F. J. Manly. Randomization Bootstrap and Monte Carlo Methods in Biology. Boca Raton: Chapman & Hall/CRC, 2007. 450 p.

ДОДАТОК А. Код програми мовою R

- Модуль main.R

```
library(ALL)
library(geneset)
library(genefilter)
library(tidyverse)

import::from(sum_for_T_stat.R, sum_for_T)
import::from(trace_for_Q_stat.R, trace_for_Q)

data(ALL)

# кількість генів
# genes amount
print("genes:")
print(length(row.names(exprs(ALL))))

# кількість пацієнтів
# patients amount
print("patients:")
print(length(colnames(exprs(ALL))))

# from Gentleman et al. "Bioinformatics and
Computational Biology Solutions Using R and Bioconducto
subset <- intersect(grep("^B", as.character(ALL$BT)),
```

```
                        + which(ALL$mol %in% c("BCR/ABL",
"NEG")))
    ALL <- ALL[, subset]

    # from Gentleman et al. "Bioinformatics and
Computational Biology Solutions Using R and Bioconductor
    # прибрати некеспресовані та низькомінливі гени
    # remove not expressed and low variability genes
    f1 <- pOverA(0.25, log2(100))
    f2 <- function(x) (IQR(x) > 0.5)
    ff <- filterfun(f1, f2)
    selected <- genefilter(ALL, ff)
    sum(selected)
    ALL <- ALL[selected, ]

    # кількість генів після прибирання
    # genes amount after removal
    print("genes:")
    print(length(row.names(exprs(ALL))))

    # кількість пацієнтів після відбору
    # patients amount after selection
    print("patients:")
    print(length(colnames(exprs(ALL))))

    ALL1 <- data.frame(ALL)

    ALL_BCRABL = ALL1[ALL$mol.biol == 'BCR/ABL', ]
```

```
n_1 = length(ALL_BCRABL$X1005_at);

print("BCR/ABL patients:")
print(n_1)

ALL_NEG = ALL1[ALL$mol.biol == 'NEG', ]

n_2 = length(ALL_NEG$X1005_at);

print("NEG patients:")
print(n_2)

X1 = ALL_BCRABL
X2 = ALL_NEG

# прибирає негенні колонки
# removing non-gene columns
X1 = X1[,1:(ncol(X1)-21)]
X2 = X2[,1:(ncol(X2)-21)]

X1_means = lapply(X=X1, FUN=mean)

X2_means = lapply(X=X2, FUN=mean)

sample_genes_amount = 10

set.seed(1)
```

```
    samplegenes = sample(colnames(X1), sample_genes_amount)

    colors = c('black', 'red')

    plotsymbol = 19

    x = seq(1, length(row.names(exprs(ALL))))

    plot(x, X1_means, pch = plotsymbol, col = colors[1], cex = 1.5, xlab = 'genes', ylab = 'expression level', type = 'l')

    title(main = 'Gene expression level means BCR/ABL')

    plot(x, X2_means, pch = plotsymbol, col = colors[2], cex = 1.5, type = 'l', xlab = 'genes', ylab = 'expression level')

    title(main = 'Gene expression level means NEG')


    sign_level = 0.05

    start <- Sys.time()

    sum_X1 <- sum_for_T(X1, X1)
```

```
    sum_X2 <- sum_for_T(X2, X2)

    sum_X1X2 <- sum_for_T(X1, X2)

    T_n <- sum_X1 + sum_X2 + sum_X1X2

    trace_X1 <- trace_for_Q(X1, X1)

    trace_X2 <- trace_for_Q(X2, X2)

    trace_X1X2 <- trace_for_Q(X1, X2)


    sigma_n <- (2/(n_1*(n_1-1)))*trace_X1 + (2/(n_2*(n_2-
1)))*trace_X2 + (4/(n_1*n_2))*trace_X1X2

    Q_n <- T_n / sqrt(sigma_n)

    print("T_n = ")

    print(T_n)

    print("Q_n = ")

    print(Q_n)

    Q_norm = qnorm(1-sign_level/2)
```

```
print("Q_norm")

print(Q_norm)

print("Execution time")

print(Sys.time()-start)

if(Q_n > Q_norm)
{
  print("Means are not equal")
} else
{
  print("Means are equal")
}
```

- Модуль sum_for_T_stat.R

```
library(zoo)
library(dplyr)

sum_for_T <- function(X1, X2)
{
  sum = 0

  n1 = length(rownames(X1))
  n2 = length(rownames(X2))
```

```
for(i in rownames(X1))
{

  for(j in rownames(X2))
  {
    X1_vector = as.numeric(c(X1[i,]))

    X2_vector = as.numeric(c(X2[j,]))

    if(!(identical(X1, X2)) | i != j)
    {
      sum = sum + t(X1_vector)%*%X2_vector
    }
  }
}

if(identical(X1, X2))
{
  sum = sum / (n1*(n1-1))

}
else
{
  sum = -2*sum/(n1*n2)
}

return(sum)
```

```
}
```

- Модуль trace_for_Q_stat.R

```
library(zoo)
library(dplyr)
library(psych)

trace_for_Q <- function(X1, X2)
{
  trace = 0

  n1 = length(rownames(X1))
  n2 = length(rownames(X2))

  X1 <- as.data.frame(X1)

  X2 <- as.data.frame(X2)

  for(i in rownames(X1))
  {
    for(j in rownames(X2))
    {
      X1_vector = as.numeric(c(X1[i,]))

      X2_vector = as.numeric(c(X2[j,]))

      if(identical(X1,X2))
```

```
          {
            X1_copy = data.matrix((X1[!(row.names(X1) %in%
c(i,j)),]))

            X2_copy = data.matrix((X2[!(row.names(X2) %in%
c(i,j)),]))
          }
          else
          {
            X1_copy = data.matrix((X1[!(row.names(X1) %in%
c(i)),]))

            X2_copy = data.matrix((X2[!(row.names(X2) %in%
c(j)),]))
          }

          X1_copy = na.spline(X1_copy)

          X2_copy = na.spline(X2_copy)

          sample_mean_1 = mean(X1_copy)

          sample_mean_2 = mean(X2_copy)

          if(!(identical(X1, X2)) | i != j)
          {
            X1_vector = data.matrix(X1_vector)
```

```
            X2_vector = data.matrix(X2_vector)

            result_matrix          =          (X1_vector-
sample_mean_1)%*%t(X1_vector)%*%(X2_vector-
sample_mean_2)%*%t(X2_vector)

            trace = trace + tr(result_matrix)
          }
        }
      }

      if(identical(X1, X2))
      {
        trace = trace / (n1*(n1-1))

      }
      else
      {
        trace = trace/(n1*n2)
      }

      return(trace)

    }
```