

Bioinformatyka 2 – kurs mały | 2025/2026

Ćwiczenie4. Mikromacierze

Zadania wstępne

Zadanie1 Analizowano dane z mikromacierzy DNA dotyczące **palaczy** celem zbadania zmian w profilu ekspresji genów jakie wtedy występują. Porównano ekspresję **20 000 genów** między:

- a) 15 osobami niepalącymi (kontrola)
- b) 12 pacjentami palącymi

Dla wybranych genów uzyskano następujące wyniki analizy statystycznej (uwzględniono poprawkę na wielokrotne testowanie).

Gen	log2FC	p-value
G1	0.3	0.42
G2	1.8	0.001
G3	-2.1	0.004
G4	0.9	0.03
G5	0.1	0.77

gdzie

$$\log_2 FC = \log_2 \left(\frac{\text{ekspresja wśród palaczy}}{\text{ekspresja w kontroli}} \right)$$

- a) Które geny można uznać za **istotnie różnicujące** między grupami przy $\alpha = 0.05$?
- b) Dla każdego istotnego genu określ, czy jest **nadekspresjonowany** czy **obniżony** wśród palaczy.
- c) Wartość $\log_2 FC = -2.1$ z tabeli oznacza, że wśród palaczy ekspresja genu G3 jest _____ (niższa/wyższa) niż w kontroli i stanowi około _____ % kontroli.

Zadanie2 Przebadano grupę pacjentów z czerniakiem oraz kontrolę celem wytypowania grupy potencjalnych markerów chorobowych. W oparciu o metodę SVM zbudowano model postaci

$$f(A, B, C) = 0.6A - 0.02B - 0.4C$$

gdzie A, B, C to znormalizowane poziomy ekspresji trzech genów markerowych. Według modelu osoba o następującym profilu ekspresji genów jest chora

$$(A, B, C) = (0.8, 0.7, 0.2).$$

- a) Zdecyduj czy osoba o profilu ekspresji $(0.6, 0.9, 0.4)$ jest chora?
- b) Stosując metodę rekurencyjnej eliminacji cech (RFE) uprość model do dwóch genów. Czy przewidywanie modelu dla (a) zmieniło się?

Zadanie3 Wyznaczono poziom ekspresji genów dla 6 kolejnych chwil czasowych. Wyznacz współczynnik korelacji pomiędzy poziomem ekspresji *genu A* oraz *genu B*. Zinterpretuj uzyskany wynik.

Gen A	0.7	0.74	0.90	1.05	1.2	1.31
Gen B	1.5	1.32	1.1	1.2	0.72	0.60

Możesz użyć przygotowany kod napisany w języku Python (plik zad3.py). Ten można uruchomić lokalnie lub klikając na [Utwórz nowy program w języku Python – myCompiler - myCompiler](#), wklejając kod i klikając **Uruchom**.

Studium przypadku

NCBI (GEO DataSets) udostępnia dane z eksperymentów mikromacierzowych a także pozwala na ich analizę online. W tym kroku poddamy analizie wybrany taki eksperiment.

Zadanie4 Przypadek do analizy z NCBI GEO DataSets

- Wejdź w link poniżej:
<https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS810>
- Czego dotyczyło badanie?
- Gdzie badano ekspresje genów? (jaki materiał/tkanka)
- Ile było wszystkich próbek i na ile grup były one podzielone (**Experiment design and value distribution oraz Sample Subsets**)
- Według jakich kryteriów podzielono próbki na grupy? Opisz wykorzystane parametry. (**Experiment design and value distribution oraz Sample Subsets**)
- Podaj przykłady genów różnicujących? (**Find genes, Find genes that are up/down for this condition(s)**)
- Spójrz na przebieg ekspresji dla genów *SPARC*, *VSNL1* oraz *COL5A2* w kolejnych grupach. (**Expression Profiles** lub **Find genes**) Czy obserwujesz jakieś tendencje zmiany poziomu ich ekspresji w kolejnych grupach? Poszukaj w źródłach zewnętrznych informacji za odpowiadającą te geny. Czy badano ich związek z chorobą Alzheimera?
- Czym są *housekeeping genes*? Jaką pełnią rolę w eksperymencie mikromacierzowym? Wybierz dwa przykładowe geny tej kategorii (np. *ACTB*, *UBC*) i sprawdź ich ekspresje w kolejnych próbkach. Jaka ona jest?

STRING

Baza danych STRING ma na celu zbieranie, ocenianie i integrowanie wszelkich publicznie dostępnych źródeł informacji na temat interakcji między białkami, a także uzupełnienie ich o prognozy komputerowe. Jej celem jest stworzenie kompleksowej i obiektywnej globalnej sieci, obejmującej zarówno bezpośrednie (fizyczne), jak i pośrednie (funkcjonalne) interakcje białkowe.

Zadanie5 Wejdź na [STRING: functional protein association networks \(string-db.org\)](http://string-db.org) a następnie przeanalizuj zestaw potencjalnych genów markerowych dla prognozy raka piersi. Wybierz *Multiple proteins*, a jako organizm *Homo sapiens*.

EFNA1

EGFR

ERBB2

GATA3

GZMB

MST1

MYB

MYBL2

MYC

PLAT

SOX4

SOX9

SRF

XBP1

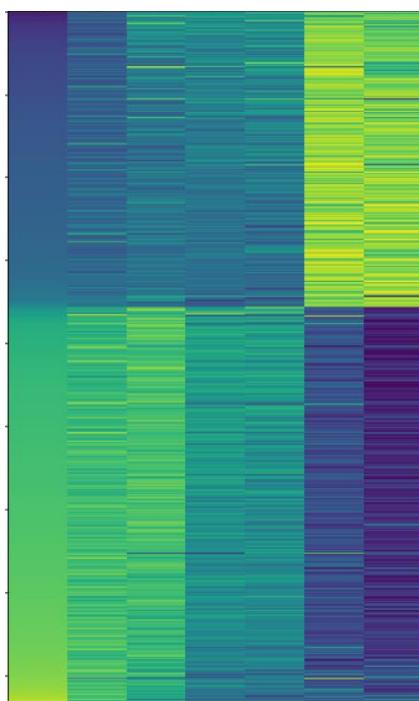
- Jakie 3 procesy biologiczne mają najmniejszy *FDR (false discovery rate)* w rozważanej grupie genów?
- Która funkcja molekularna ma najmniejszy *FDR (false discovery rate)* w rozważanej grupie genów?
- Która ścieżka KEGG ma najmniejszy *FDR (false discovery rate)* w rozważanej grupie genów?

Zadanie6 Poniżej zamieszczono dane dotyczące ekspresji genów w komórkach drożdży podczas procesu oddychania – fermentacji alkoholowej. Wyróżniamy dwa główne etapy tego procesu:

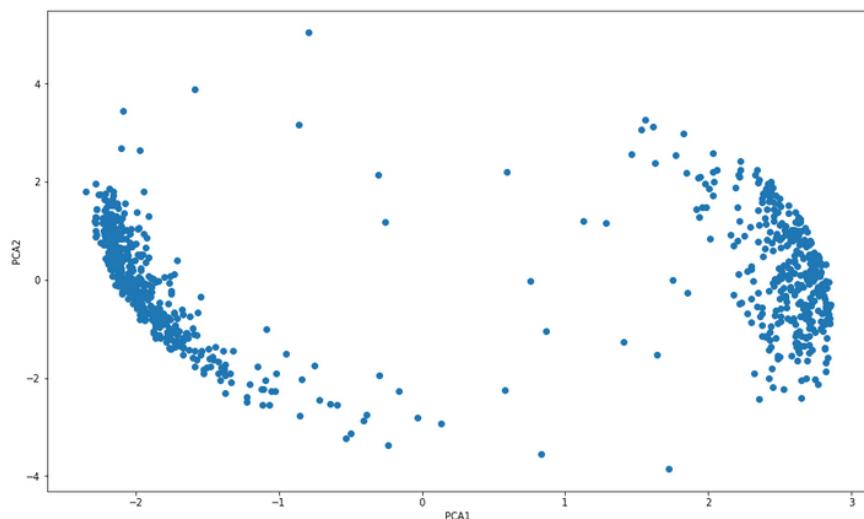
- rozkład glukozy do kwasu pirogronowego,
- przemianę kwasu pirogronowego do alkoholu.

Każdy z etapów kontrolowany jest przez 2 klasy genów odpowiedzialnych za te procesy. Dane pochodzą z 7 chwil czasowych (kolejne kolumny). Skomentuj i porównaj poniższe wyniki w kontekście powyższych informacji. W jaki sposób podzieliłbyś rozważane chwile czasowe?

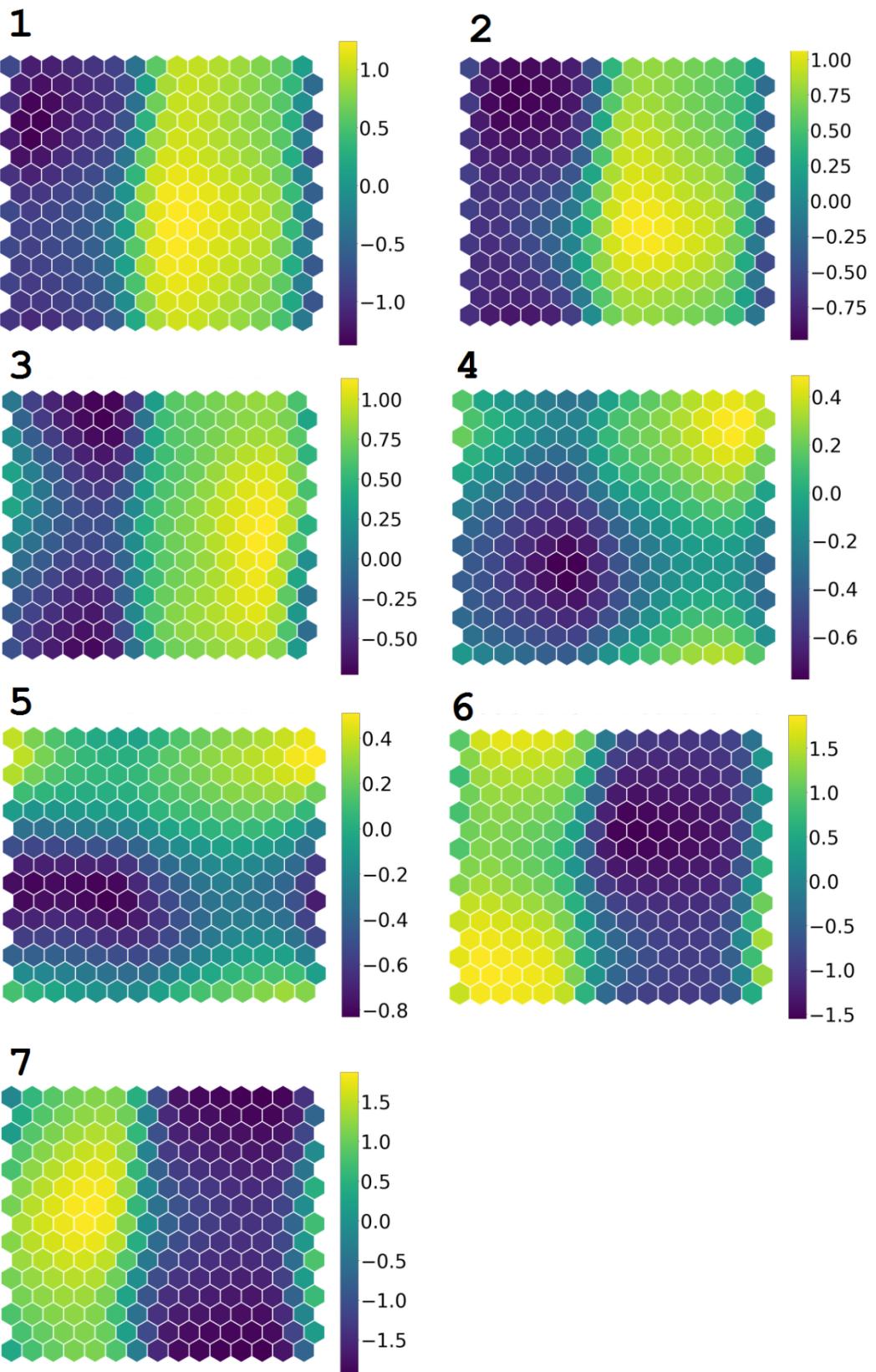
- rozważana mikromacierz (geny zostały posortowane względem pierwszej chwili czasowej):



- po zastosowaniu algorytmu PCA, gdzie jako kolejne obserwacje wybrano wiersze (geny) z powyższej macierzy:



- mapy Kohonena dla kolejnych chwil czasowych. Kolor wskazuje na poziom ekspresji określonej grupy genów:



Zadanie uzupełniające

Zadanie7 Wejdź na stronę https://singlecell.broadinstitute.org/single_cell/. Znajdują się tam dane z eksperymentów scRNA-seq. Wyszukaj eksperiment: Single-cell transcriptomics of the spinal cord of a severe SMA mouse i odpowiedz na poniższe pytania.

- Czego dotyczył eksperiment? (**Summary**)
- Ile komórek i genów rozważano?
- Ile typów komórek rozważano?

Dodatkowo, sprawdź ekspresję genów: hemoglobiny (HBB) oraz SPARC wśród rozważanych typów komórek. W której grupie występuje największa ich ekspresja? (**Explore, search genes**)