

Analiza danych mikromacierzowych

Adrian Kania¹

¹Zakład Biofizyki Obliczeniowej i Bioinformatyki

2025/2026

Plan prezentacji

- 1 Elementy nauczania maszynowego
 - Metody klasteryzacji
 - Redukcja wymiaru

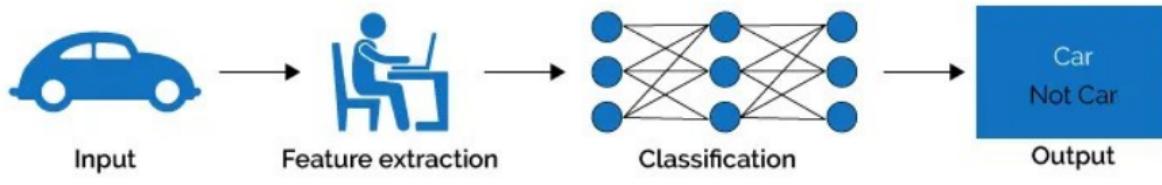
- 2 Mikromacierze
 - Cel i przebieg eksperymentu
 - Analiza statystyczna

- 3 Analiza funkcjonalna
 - Gene Ontology
 - KEGG PATHWAY

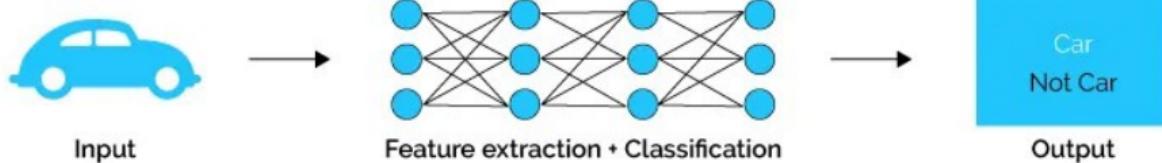
Metody nauczania maszynowego

- Uczenie nadzorowane
 - Klasyfikacja
 - Regresja
- Uczenie nienadzorowane
- Uczenie ze wzmocnieniem

Machine Learning



Deep Learning



Metody nauczania maszynowego

Po co budować?

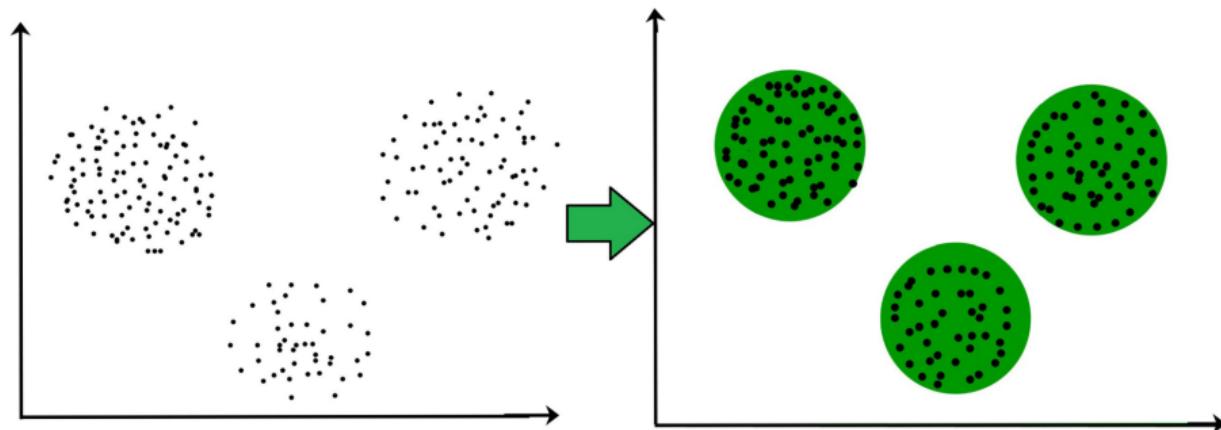
- ① opis danych (zrozumienie relacji)
- ② przewidywanie
- ③ i wiele wiele innych...

Przykłady

- ④ Modelowanie ekspresji genów w oparciu o aktywność czynników transkrypcyjnych (**Prediction of Gene Expression Patterns With Generalized Linear Regression Model**, *Front. Genet.*)
- ⑤ Przewidywanie miejsc karbonylacji białek (**iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features**, *Bioinformatics*)
- ⑥ Symulacja działania układu odpornościowego (**Understanding adaptive immune system as reinforcement learning**, *Phys. Rev.*)

Algorytmy grupowania - cel

Zasadniczym zadaniem algorytmów z tej grupy jest (jak nazwa wskazuje) podział danych na grupy (tzw. klastry).

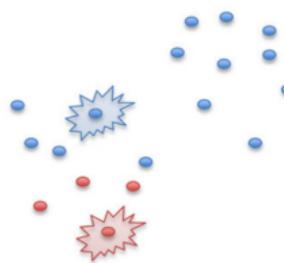


Na wejściu mamy więc dane bez etykiet, na wyjściu informację do której grupy należy każda z obserwacji.

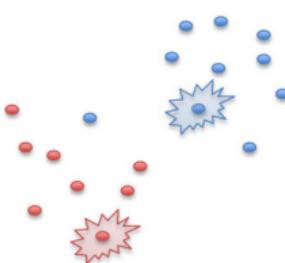
Algorytm k-średnich (k-means)

Startujemy z losowych punktów wybranych wśród danych (lub nie), a następnie przyporządkowujemy przynależność każdej z obserwacji do tego punktu, gdzie odległość od niego jest najmniejsza. Proces powtarzamy iteracyjnie.

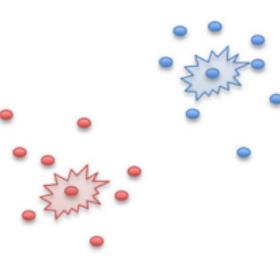
Initial Seeding



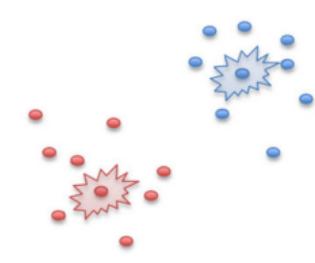
After Round 1



After Round 2



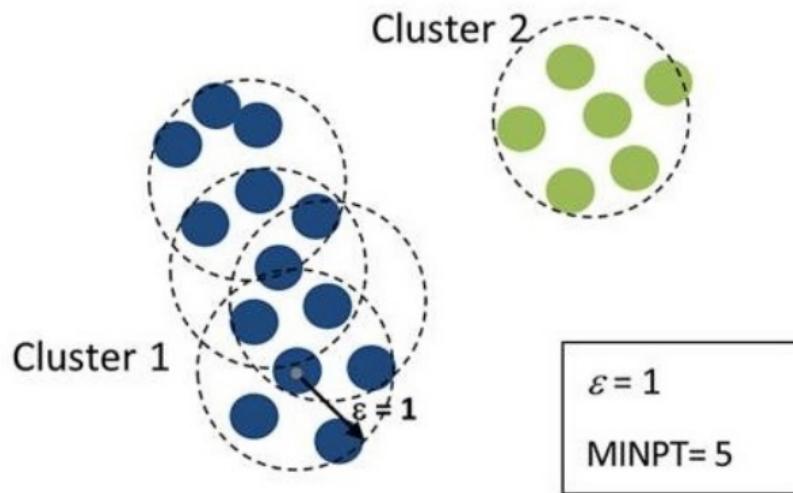
Final



Użytkownik musi zadać liczbę klastrów.

DBSCAN

Wokół każdej obserwacji rysujemy koło o promieniu ϵ . Obserwacje należą do tej samej grupy, jeśli są wystarczająco blisko siebie. W przeciwnym przypadku tworzona jest nowa grupa.



Tutaj zadajemy parametr ϵ czyli promień kuli (a nie liczbę klastrów).

k-means vs DBSCAN

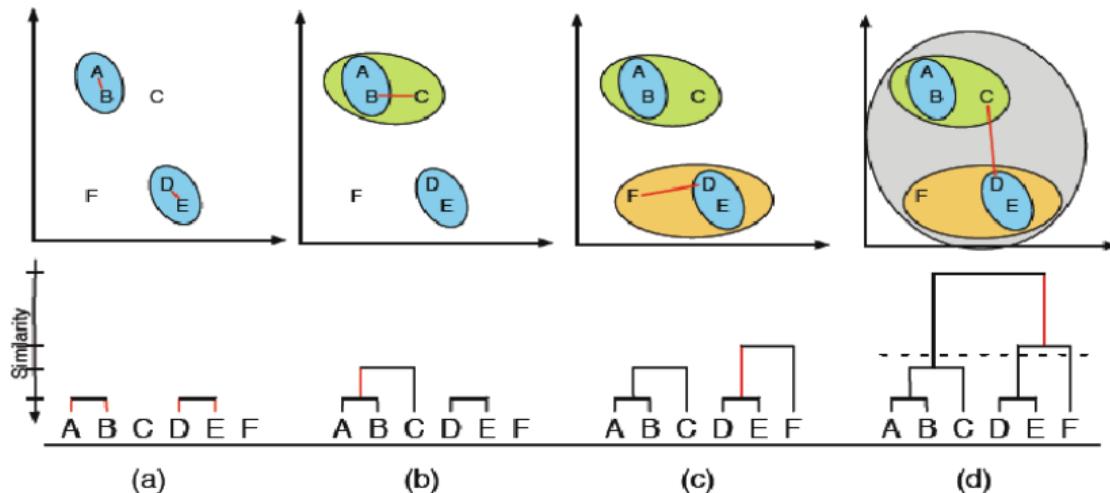
DBSCAN



k-means



Grupowanie hierarchiczne

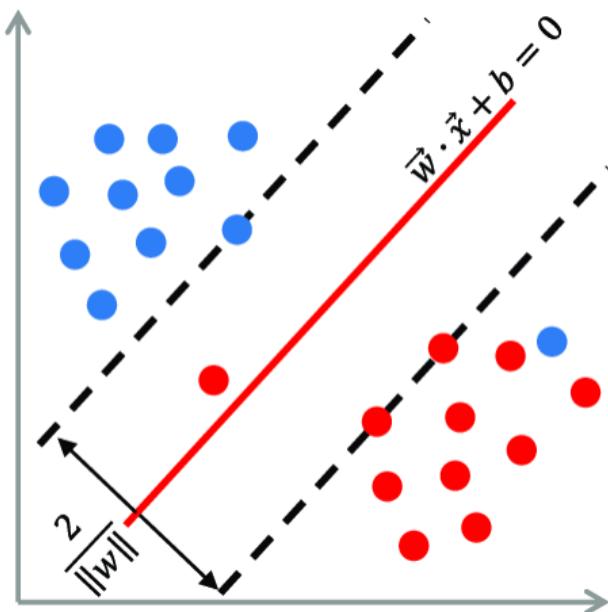


- Metoda połączenia minimalnego (Single Link Clustering), $d_{C1,C2} = \min_{x_i \in C_1, y_j \in C_2} d(x_i, y_j)$
- Metoda połączenia maksymalnego (Complete Link Clustering), $d_{C1,C2} = \max_{x_i \in C_1, y_j \in C_2} d(x_i, y_j)$
- Metoda połączenia średniego (Average Link Clustering), $d_{C1,C2} = 1/(|C1||C2|) \sum_{x_i \in C_1, y_j \in C_2} d(x_i, y_j)$

Maszyna wektorów nośnych (SVM)

Cel: poszukujemy wektora wag $w \in \mathbb{R}^n$ oraz składnia stałego b liniowej funkcji klasyfikującej

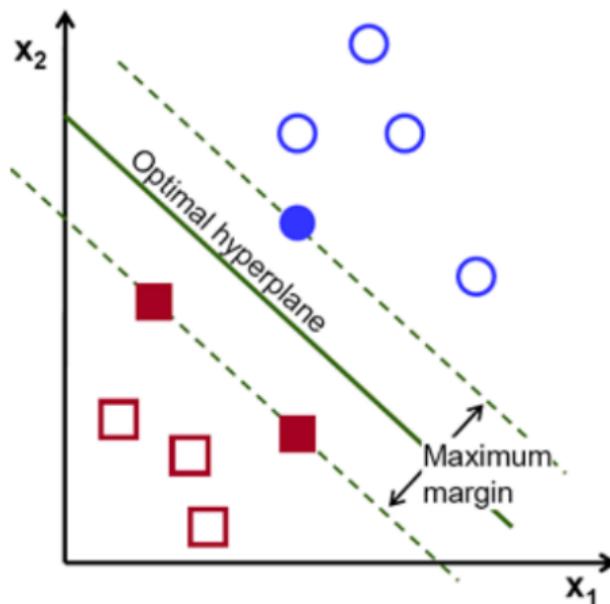
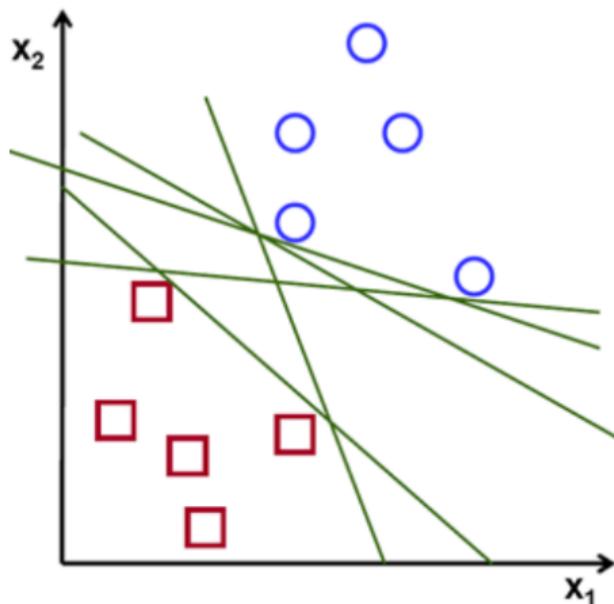
$$f(x) = w^T x + b$$



Funkcja ta wyznacza granicę pomiędzy dwoma klasami. Te obserwacje dla których $w^T x_i + b > 0$ należą do pierwszej klasy, z kolei te dla których $w^T x + b < 0$ należą do drugiej klasy.

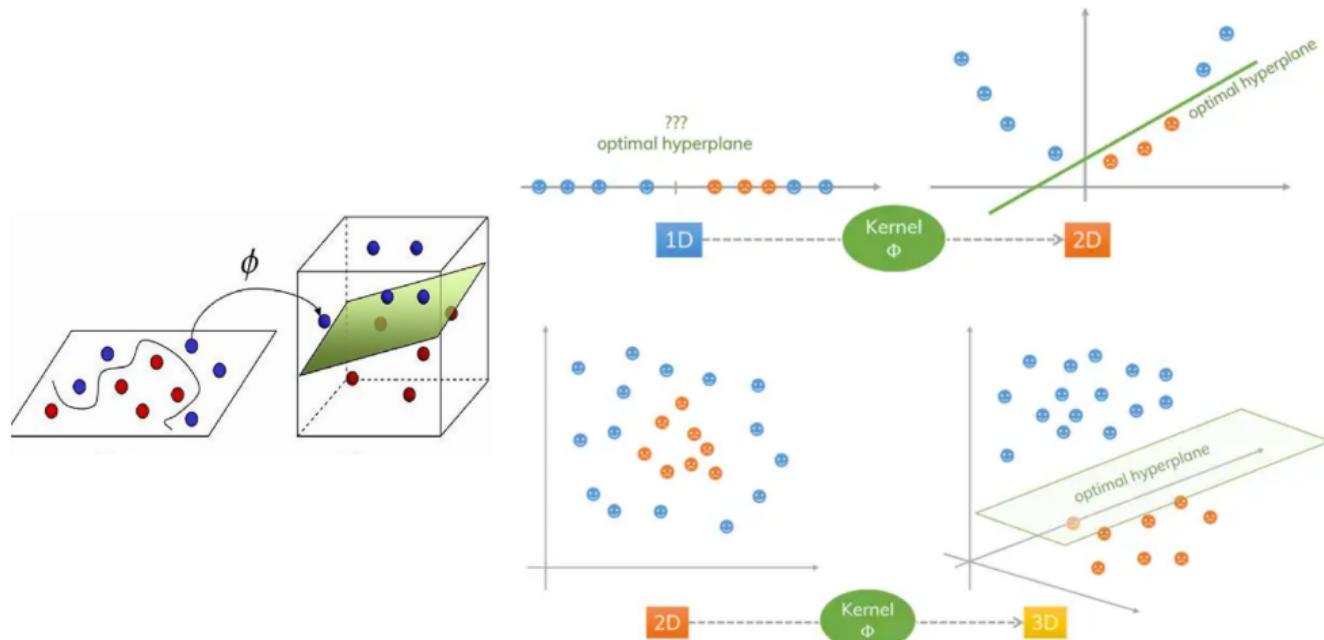
Maszyna wektorów nośnych (SVM)

Jak wybrać najlepszą (optymalną) prostą?



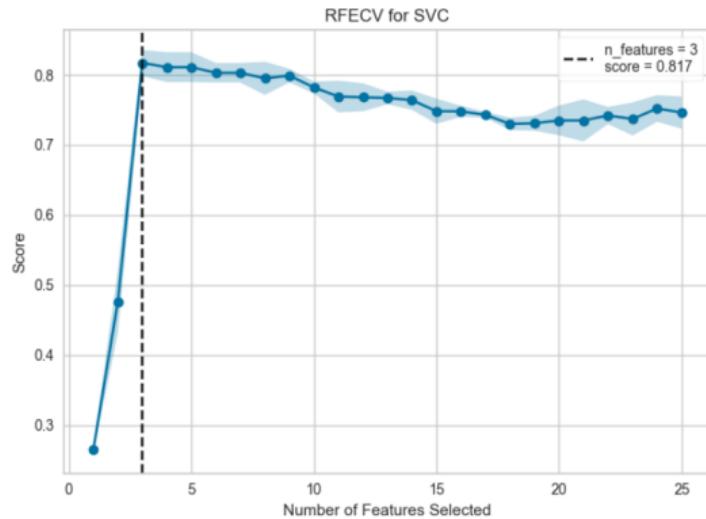
Maszyna wektorów nośnych (SVM)

Separacja liniowa nie zawsze jest możliwa - wtedy stosujemy przekształcenie danych do wyżej wymiarowej przestrzeni gdzie taka separacja już występuje (w tym celu stosujemy tzw. kernel).



Rekurencyjna eliminacja cech (RFE)

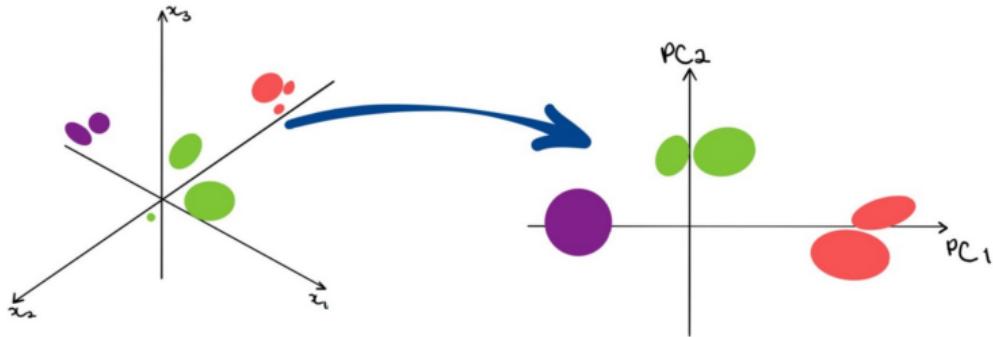
Metoda oparta o SVM. Startujemy z pewnego zbioru genów, następnie iteracyjnie usuwamy pojedyncze geny.



W kolejnej iteracji następuje dobranie optymalnej (w sensie SVM) liniowej funkcji klasyfikującej. Następnie eliminowana jest cecha, której odpowiada waga (element wektora w) o najmniejszej wartości bezwględnej.

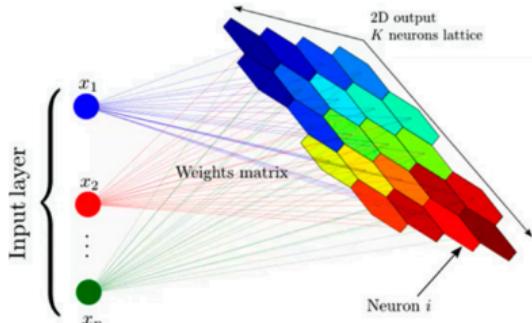
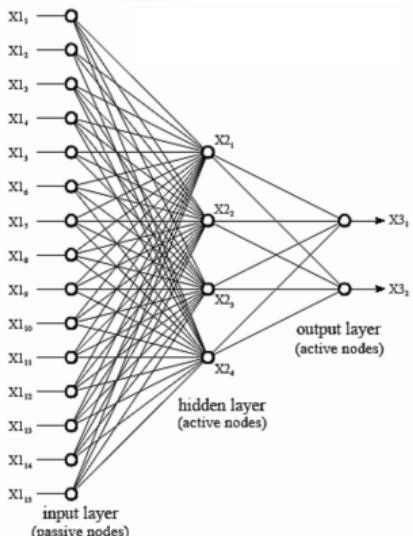
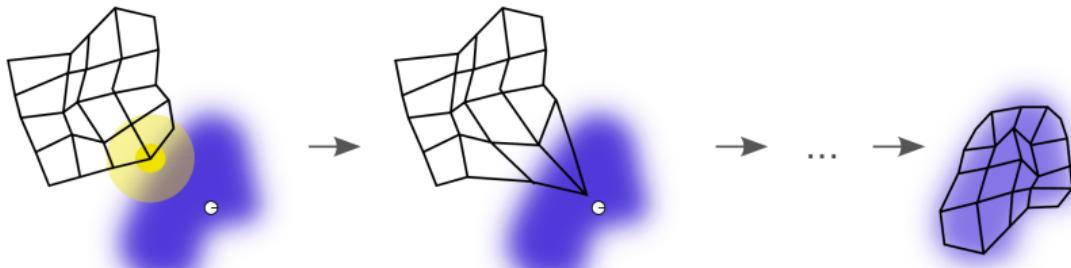
Analiza składowych głównych (PCA)

Algorytm służący redukcji wymiaru (ilości zmiennych).



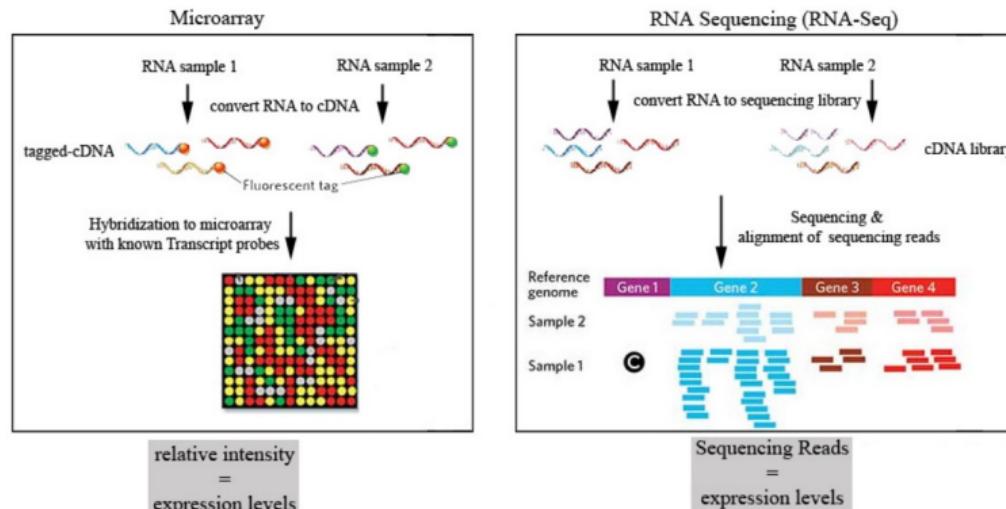
Idea: wyznaczamy kierunki największej zmienności i na nie rzutujemy nasze dane. Jak? Oryginalne cechy X_1, X_2, \dots, X_n zostają przekształcone na cechy wtórne F_1, F_2, \dots, F_n (kombinacje liniowe wejściowych cech), a następnie pozostawiane są te cechy/kierunki w których zmienność jest największa.

Sieci Kohonena (SOM)

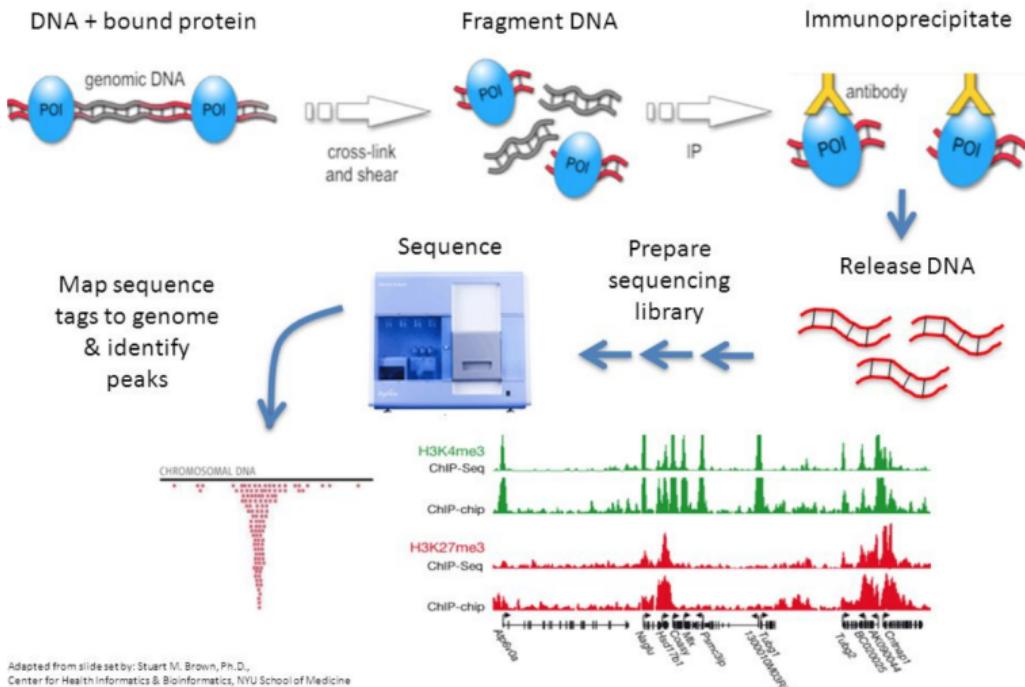


Jak zmierzyć poziom ekspresji genów w komórce?

- Reakcja łańcuchowa polimerazy w czasie rzeczywistym (real time PCR),
- Hybrydyzacja northern (RNA blot),
- Sekwencjonowanie (Sanger, Maxam-Gilbert),
- Mikromacierze,
- Sekwencjonowanie nowej generacji (NGS/RNA-Seq).

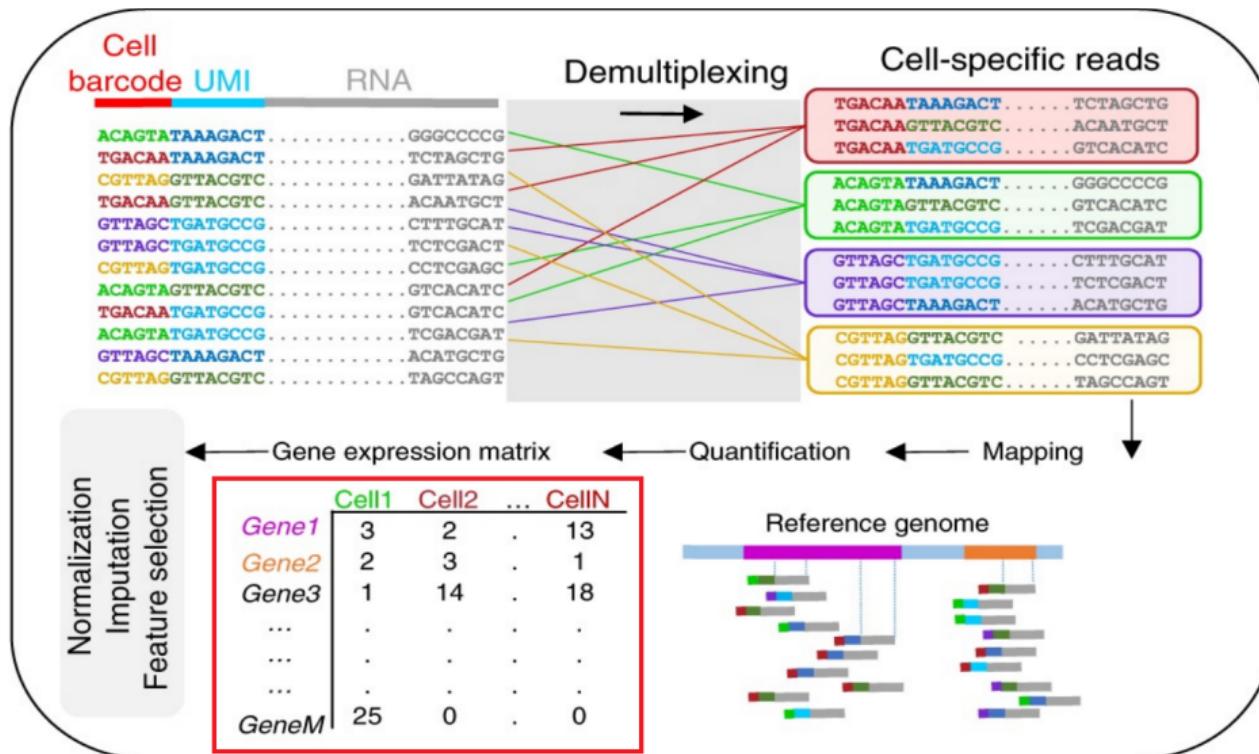


ChIP-seq



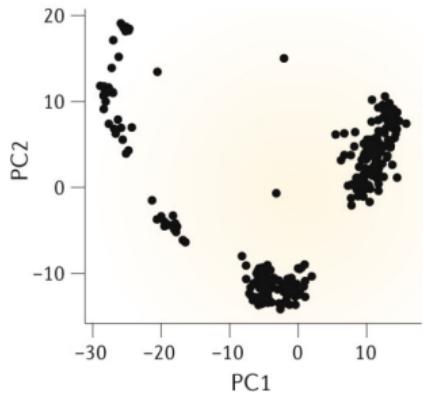
Otwarta chromatyna? ATAC-seq, FAIRE-seq.

scRNA-seq

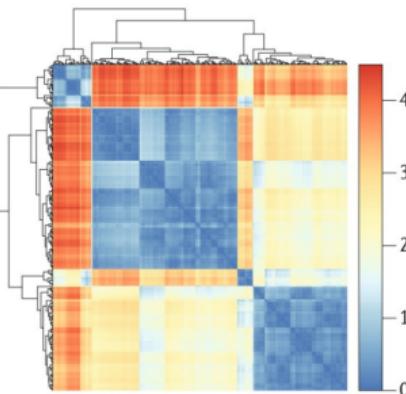


scRNA-seq

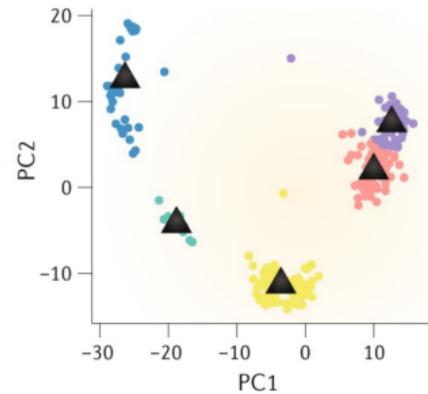
Dimensionality reduction



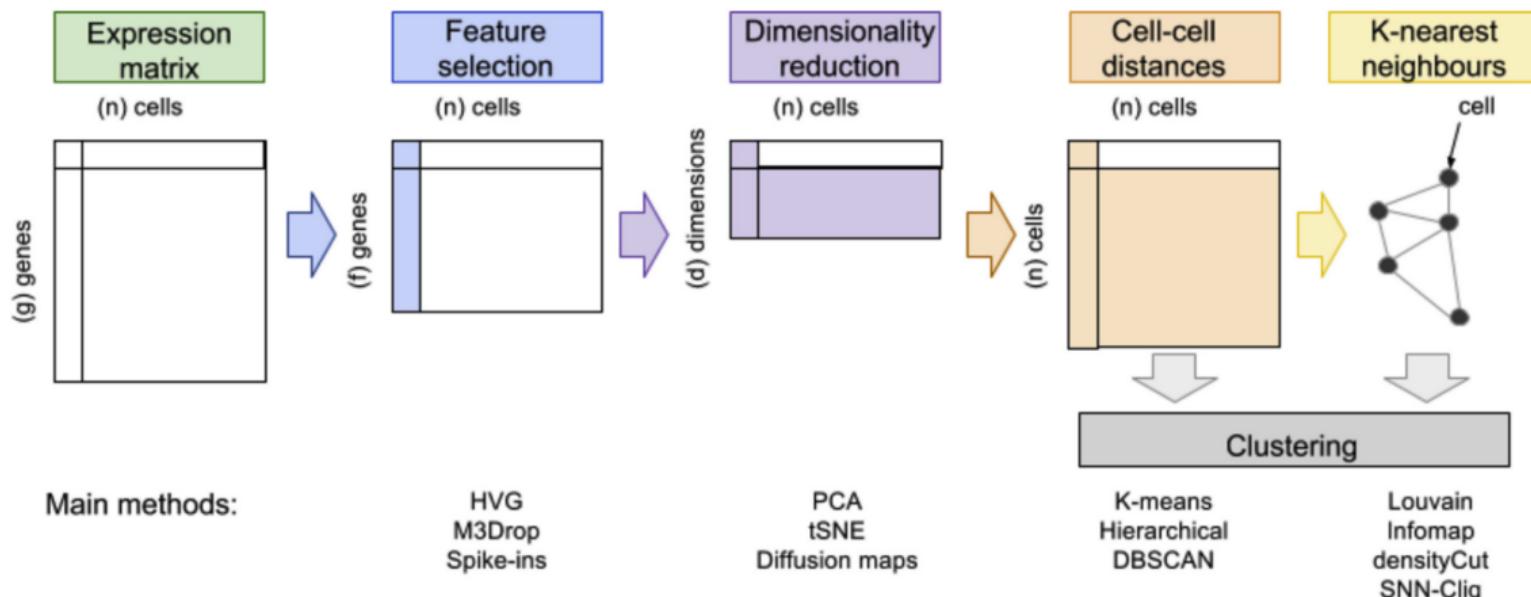
Cell-cell distances



Unsupervised clustering



scRNA-seq

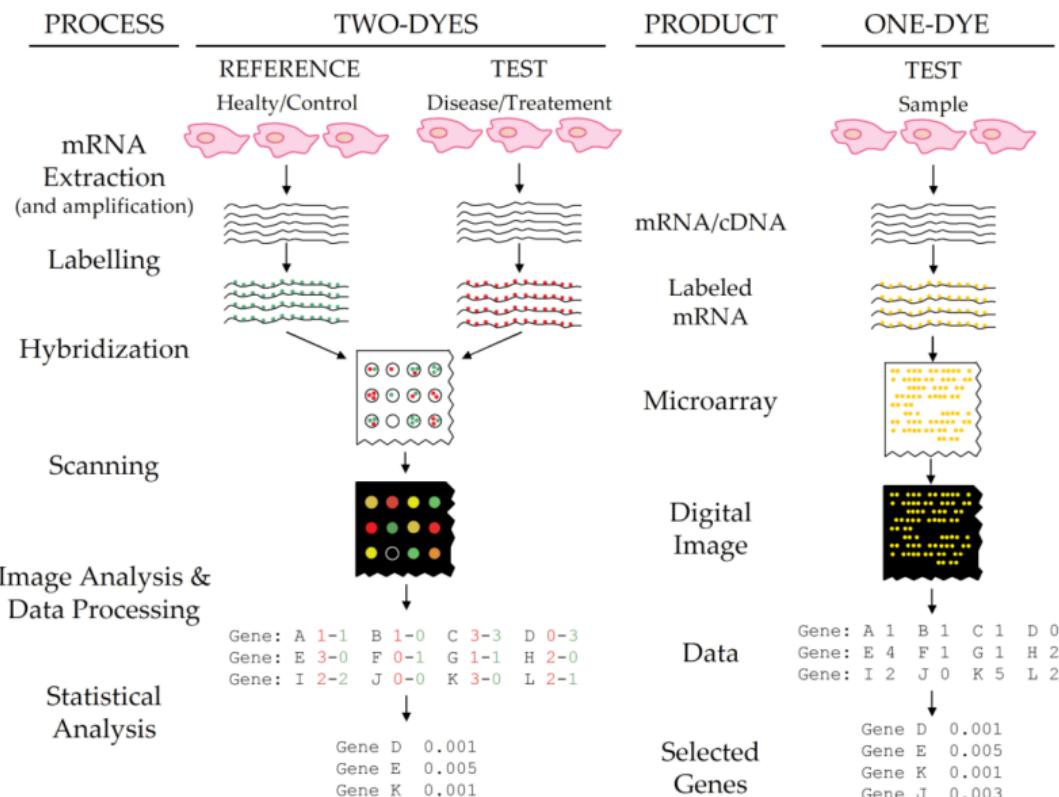


Zastosowanie mikromacierzy

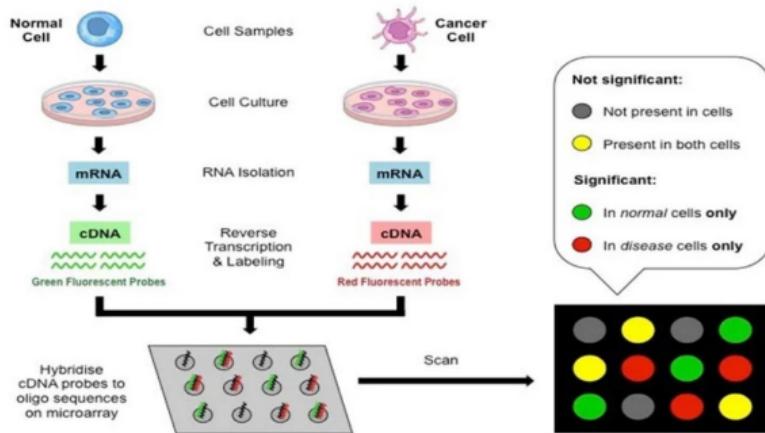
Wybrane zastosowania mikromacierzy

- Wyznaczanie profili ekspresji genów (w różnych tkankach, stanach chorobowych...).
- Badanie polimorfizmu/detekcja SNP.
- Badanie oddziaływania DNA-białko.
- Badanie nowych leków.
- Identyfikacja procesów komórkowych w które zaangażowane są geny.

Eksperyment mikromacierzowy



Eksperyment mikromacierzowy

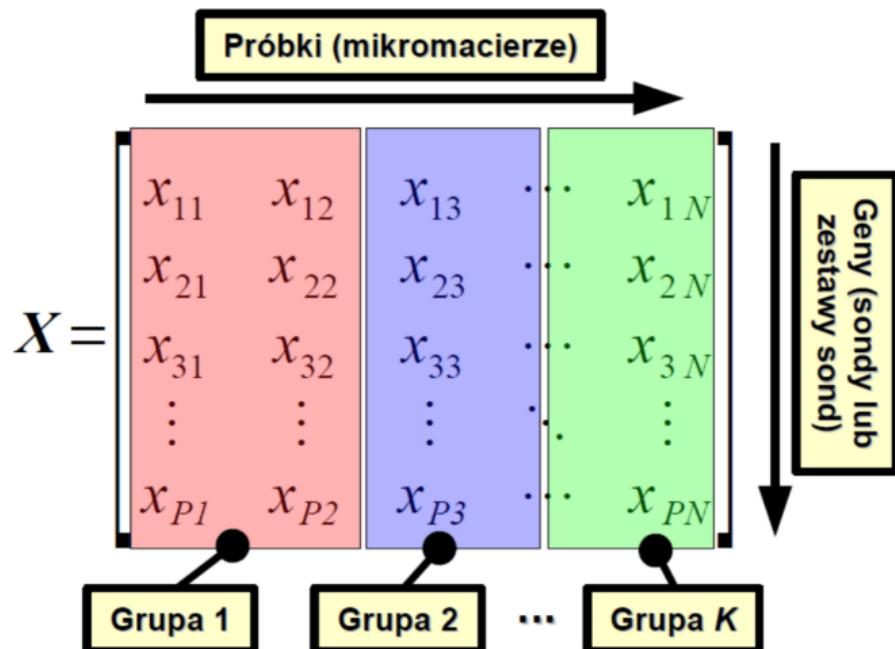


Wyznaczamy stosunek intensywności czerwonej i zielonej fluorescencji (czyli $x_i = R_i/G_i$) dla każdego genu.

- Jeżeli $x_i = 1$, to oznacza to, że poziom i -tego genu był taki sam w próbie kontrolnej i badanej.
- Jeżeli $x_i > 1$, to oznacza to, że poziom i -tego genu był wyższy w próbie badanej w porównaniu do próby kontrolnej.
- Jeżeli $x_i < 1$, to oznacza to, że poziom i -tego genu był niższy w próbie badanej w porównaniu do próby kontrolnej.

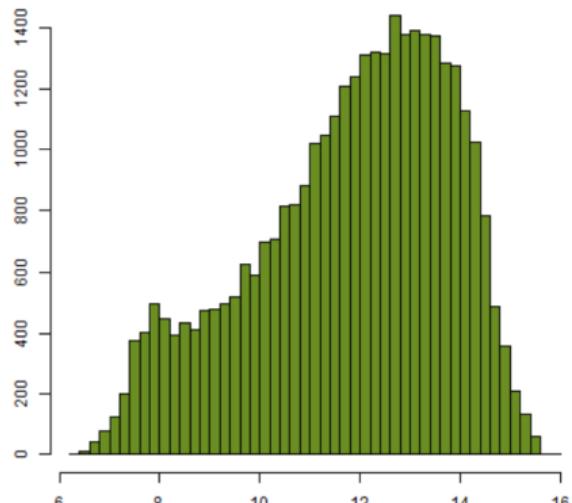
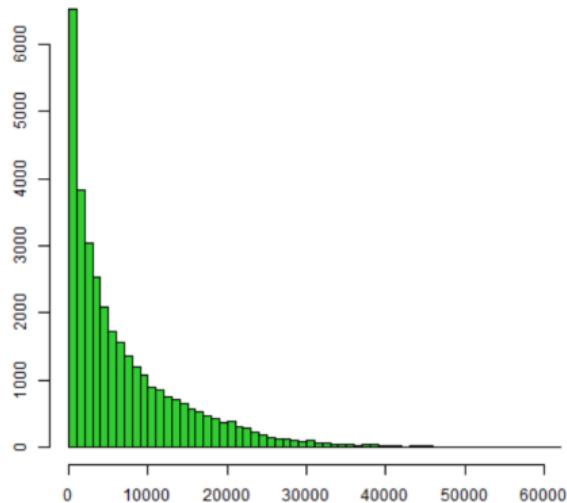
Eksperyment mikromacierzowy

Niech N oznacza liczbę eksperymentów mikromacierzowych. Każda mikromacierz dostarcza informacje ilościowe o ekspresji P genów. Formalnie możemy więc rozważane dane przedstawić w postaci zbiorczej macierzy $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{P \times N}$.



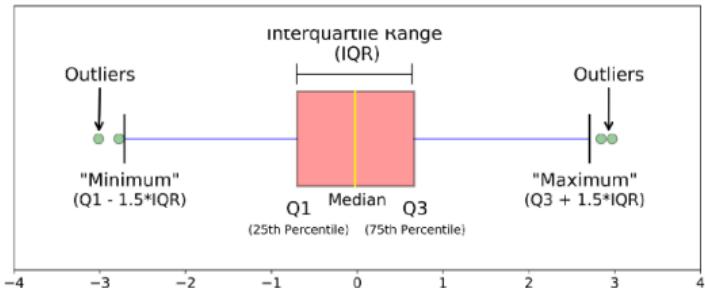
Transformacja logarytmiczna

Stosujemy transformację $x_i = \log(x)$ aby wykres intensywności przybrał bardziej symetryczną (Gaussowską) postać.

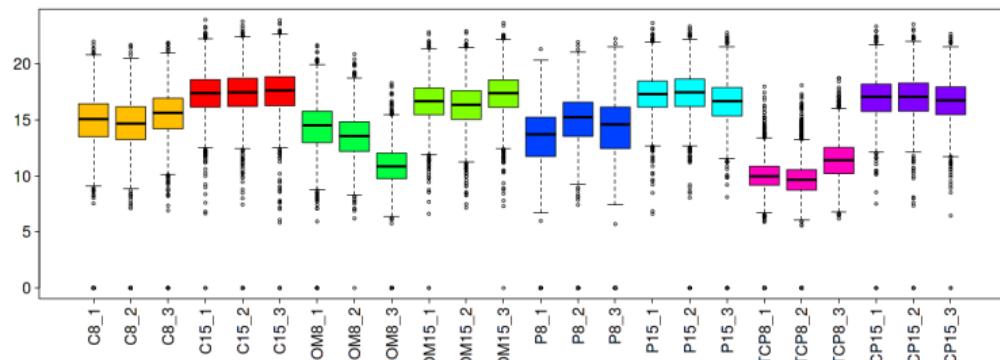


Wykresy pudełkowe

Zamiast histogramów, do prezentacji wyników mikromacierzowych, częściej stosuje się wykresy pudełkowe (boxplot).



Przykładowe zestawienie wyników z kilku eksperymentów mikromacierzowych.



Normalizacja

Po co stosujemy normalizację?

- różna ilość materiału w kolejnych eksperymentach
- różna wydajność: ekstrakcji RNA, odwrotnej transkrypcji, znakowania, fotodetekcji

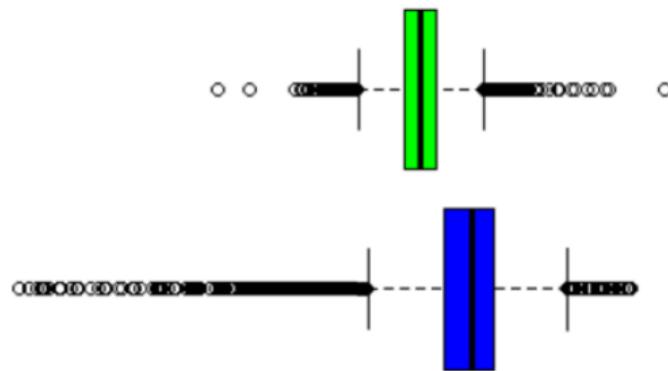
Rodzaje normalizacji

- globalna - wszystkie geny biorą udział w normalizacji
- lokalna - używamy niewielkiej puli genów (np. housekeeping genes)

Normalizacja

Mówimy tu o dwóch ważnych kwestiach:

- lokalizacja
- skalowanie



$$x_{norm} = \frac{x - \mu}{\sigma}$$

Lokalizacja i skala dla różnych mikromacierzy powinna być na ogół (prawie) taka sama.

Gdzie można znaleźć dane z eksperymentów mikromacierzowych?

The screenshot shows the NCBI GEO DataSets homepage. At the top, there is a navigation bar with links for NCBI Resources, How To, and Sign in to NCBI. Below the navigation bar, there is a search bar with the placeholder "GEO DataSets" and a "Search" button. A "Help" link is also present. A prominent banner at the top of the page displays COVID-19 information, including links to Public health information (CDC), Research information (NIH), SARS-CoV-2 data (NCBI), Prevention and treatment information (HHS), and Español. The main content area features a large image of DNA sequence data with various biological terms overlaid, such as "expression", "genomic", "gene", "microarray", and "molecule". The title "GEO DataSets" is displayed in large white text. Below the title, a text block explains that the database stores curated gene expression DataSets and original Series and Platform records from the Gene Expression Omnibus (GEO) repository. It encourages users to enter search terms to locate experiments of interest. The page is divided into three main sections: "Getting Started", "GEO Tools", and "More Resources", each with a list of links.

GEO DataSets

This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.

Getting Started

- [GEO Documentation](#)
- [GEO FAQ](#)
- [About GEO DataSets](#)
- [Construct a Query](#)
- [Download Options](#)

GEO Tools

- [Submit to GEO](#)
- [Advanced Search](#)
- [DataSet Browser](#)
- [Programmatic Access](#)
- [GEO2R](#)

More Resources

- [GEO Home](#)
- [GEO Profiles](#)
- [SRA](#)

Specyfika danych z mikromacierzy

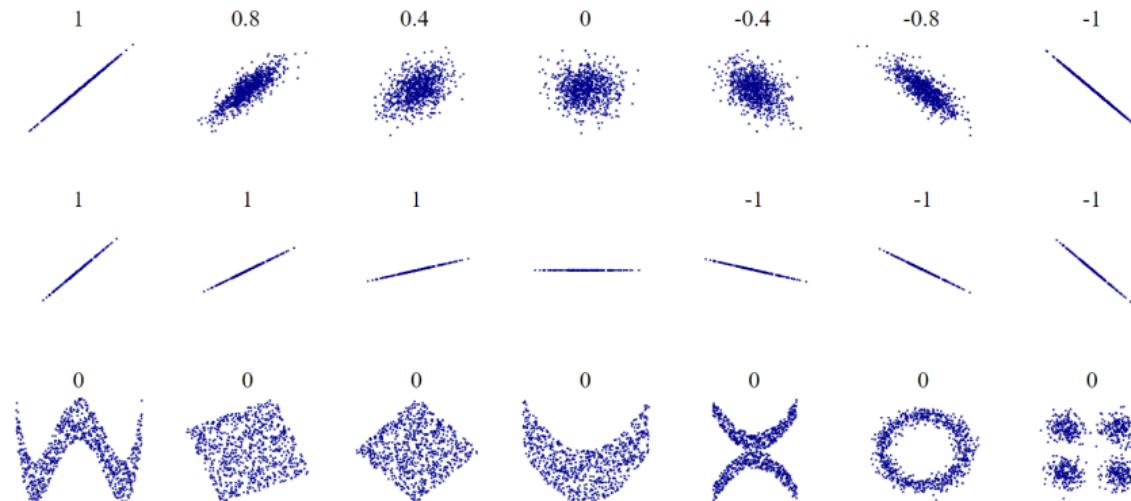
Przy analizie danych mikromacierzowych należy zwrócić szczególną uwagę na następujące zagadnienia:

- Zasadniczy problem: $P \gg N$ (liczba genów znacznie większa niż liczba mikromacierzy).
- Duży zakres zmienności ekspresji genów (1% genów odpowiada za połowe masy mRNA w komórce).
- Możliwość braku danych - spowodowana m.in. lokalnymi defektami mikromacierzy.
- Duża podatność na zakłócenia i błędy w obróbce laboratoryjnej.

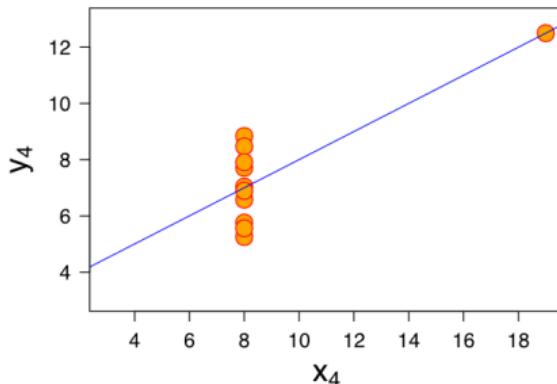
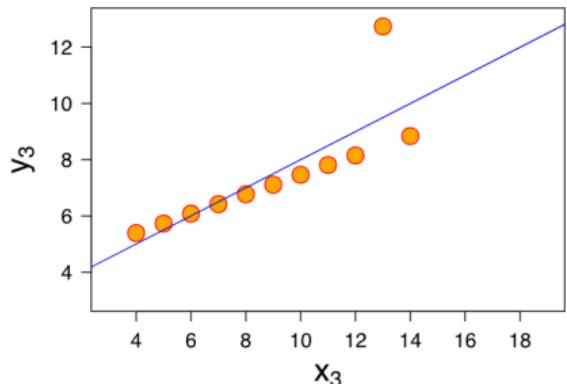
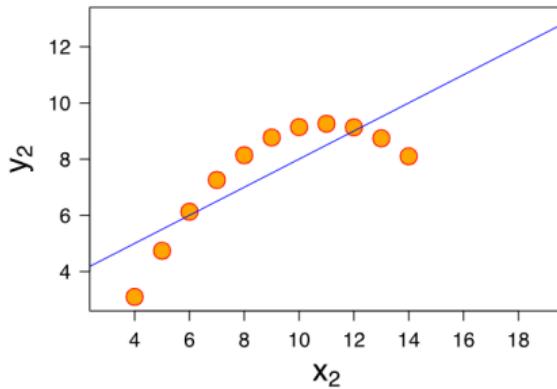
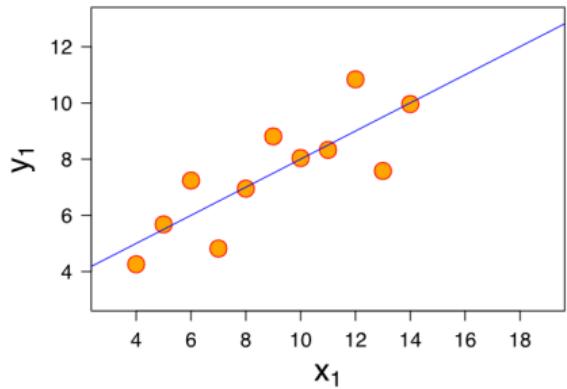
Analiza korealacji

Jednym z celów przeprowadzenia analizy macierzowej może być wytypowanie genów o podobnym profilu ekspresji w kolejnych grupach (ta sama monotoniczność). Do oceny zależności liniowej między dwoma zmiennymi (tutaj: profilami genów) służy współczynnik korealacji. Najczęstsze stosowane to:

- współczynnik korelacji r Pearsona,
- współczynnik korelacji rang Spearmana.



Kwartet Anscombe'a



Problem testowania wielu hipotez

Założymy, że badamy poziom ekspresji 9000 genów aby sprawdzić skuteczność działania nowego leku. Mamy do dyspozycji grupę kontrolną i grupę badaną oraz wykonujemy test statystyczny dla każdego genu. Niech

- H_0 : gen nie uległ zróżnicowanej ekspresji
- H_1 : różnica w ekspresji genu jest znacząca

Przyjmijmy $p_{value} = 0.01$ (jest 1% szansy na zaobserwowanie zróżnicowanej ekspresji przez przypadek). W przypadku badania 9000 genów, nawet gdyby lek nie miał żadnego wpływu, to spodziewamy się że dla 90 genów ich p_{value} będzie mniejsze niż 0.01.

Problem testowania wielu hipotez

Przyjmijmy $p_{value} = 0.01$ (jest 1% szansy na zaobserwowanie zróżnicowej ekspresji przez przypadek). W przypadku badania 9000 genów, nawet gdyby lek nie miał żadnego wpływu, to spodziewamy się że dla 90 genów ich p_{value} będzie mniejsze niż 0.01.

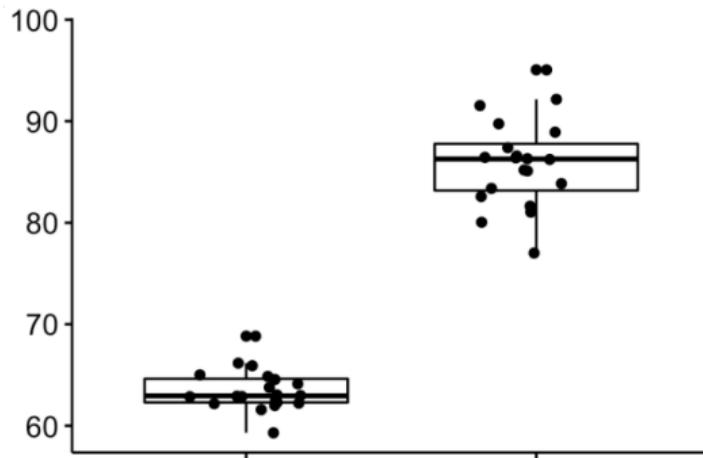
Poprawka Bonferroniego

- wyznaczamy wartości p_i dla każdego genu, $i = 1, 2, \dots, n$.
- wyznaczamy $p'_i = \min(np_i, 1)$
- wybieramy te geny, dla których $p'_i < \alpha$, gdzie α - założony poziom istotności.

Testy statystyczne

- Testy parametryczne (dla danych z rozkładem normalnym) - t test dla dwóch prób
- Testy nieparametryczne (mają mniej założeń) - test Wilcoxona dla par obserwacji, test Manna-Whitneya.
- Metoda Bootstrap - pozwala ominąć założenie o rozkładzie normalnym.
- eBayes - gdy mamy mało powtórzeń i nie można wyliczyć wariancji.
- Test Anova - gdy mamy więcej niż dwa warunki

t-test



Etapy:

- Dla każdej grupy i dla każdego genu wyznaczamy średnią i odchylenie standardowe.
- Wyznaczamy parametr t .

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

- Odczytujemy wartość parametru p z tablic (o rozkładzie t-studenta).

Jak można popełnić błąd?

	Odrzucenie hipotezy zerowej	Nie-odrzucenie hipotezy zerowej
Hipoteza zerowa (H_0) jest prawdziwa	False positive Błąd I typu	True positive poprawne
Hipoteza zerowa (H_0) jest fałszywa	True negative poprawne	False negative Błąd II typu

Ocena klasyfikatorów

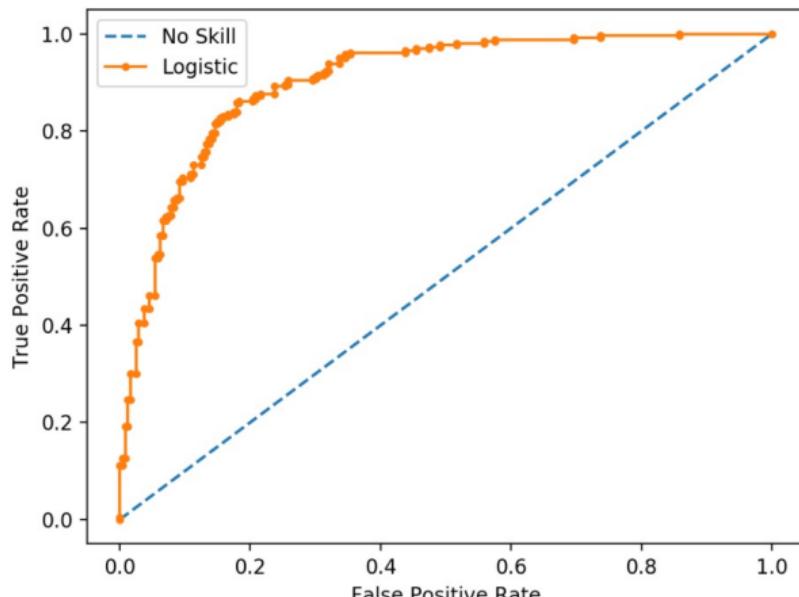
- czułość (prawdopodobieństwo, że test wskaże na wynik pozytywny jeśli faktycznie taki ma być)

$$TPR = \frac{TP}{TP+FN}$$

- specyficzność (prawdopodobieństwo, że test wskaże wynik negatywny jeśli faktycznie taki powinien być)

$$TNR = \frac{TN}{TN+FP}$$

- krzywa ROC (zależność TPR od 1-TNR) oraz pole pod krzywą AUC



Co dalej z wytypowanym zbiorem genów?

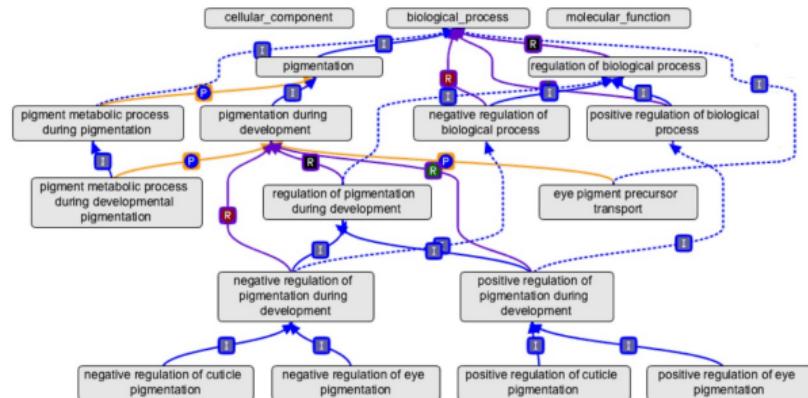
Analiza funkcjonalna - nadanie interpretacji biologicznej uzyskanym wynikom. W kontekście mikromacierzy takim wynikiem jest zbiór genów różnicujących. Przydatne w tym celu mogą być klasyczne bazy danych jak GenBank czy DDBJ czy te bardziej specjalistyczne (Gene Ontology, KEGG PATHWAY) zawierające informacje o wzajemnych powiązaniach między obiektami.



Co dalej z wytypowanym zbiorem genów?

Ontologia genowa i trzy główne domeny (subontologie)

- Funkcja molekularna,
- Proces biologiczny,
- Składnik komórkowy.



Każda z domen zorganizowana jest w sposób hierarchiczny jako **acykliczny graf skierowany** (węzły = kategorie GO, krawędzie = relacje). Oprócz ontologii, konieczne są jeszcze **adnotacje** (określają powiązania produktów ekspresji genów ze zdefiniowanymi wcześniej obiekty i procesami).

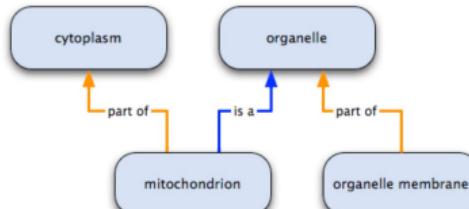
Ontologia genowa

Rekordami są tzw. kategorie GO (GO terms), z których każdy definiowany jest poprzez: unikalny identyfikator (id), nazwę (name), opis słowny (def), subontologia (namespace), rodzaj relacji z innymi kategoriami GO (is_a, part_of, regulates).

Przykładowy rekord bazy GO (format OBO)	
[Term]	
id:	GO:0032452
name:	histone demethylase activity
namespace:	molecular_function
def:	"Catalysis of the removal of a methyl group from a histone." [GOC:mah]
is_a:	GO:0032451 ! demethylase activity

Relacja

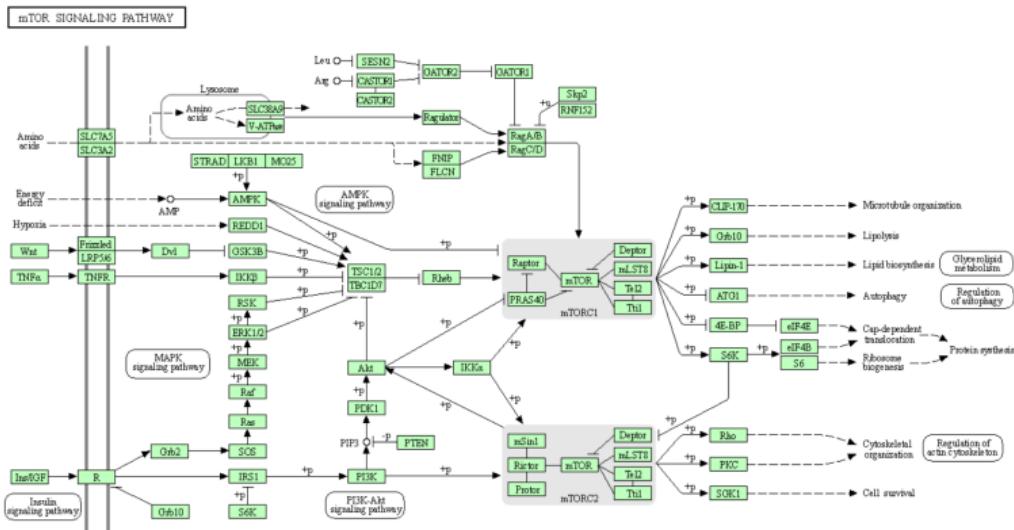
- A is_a B oznacza, że kategoria A jest podtypem kategorii B.
- A part_of B oznacza, że A jest zawsze częścią B.
- A regulates B oznacza, że A zawsze bezpośrednio wpływa na B.



KEGG PATHWAY

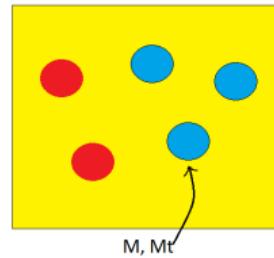
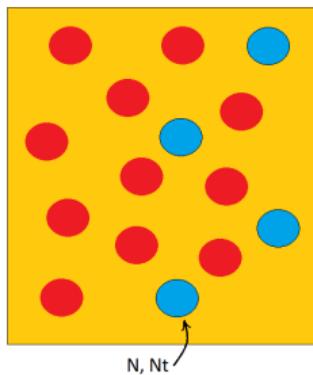
KEGG PATHWAY

- część zbioru biologicznych baz danych KEGG,
- informacja na temat sieci interakcji pomiędzy produktami ekspresji genów (także dla leków)



Badanie nadreprezentacji grupy genów

Pytanie: Czy w wytypowanym przez nas zbiorze genów występuje większa grupa powiązana np z tym samym procesem biologicznym?

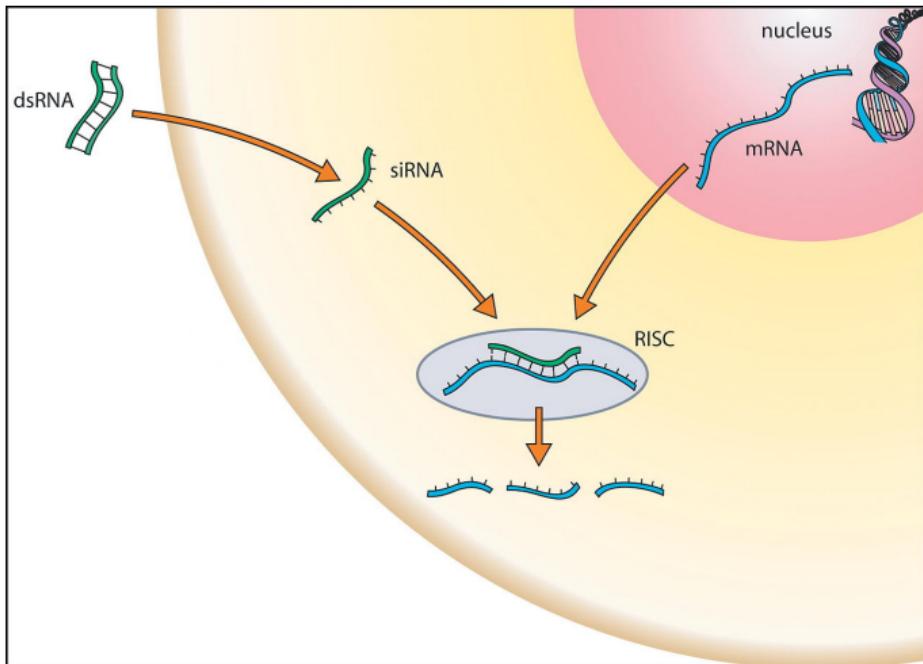


Porównujemy oczekiwane prawdopodobieństwo wystąpienia interesującej nas grupy genów z obserwowanym.

$$P(x = M_t) = \frac{\binom{N_t}{M_t} \binom{N - N_t}{M - M_t}}{\binom{N}{M}}$$

Interferencja RNA

Do tej pory zakładaliśmy niejawnie, że: poziom ekspresji genu \approx poziom białka



To sprawia, że wyniki z mikromacierzy (podobnie NGS) należy traktować z pewną rezerwą...